

The Bottlenecks: Where Value Accrues

Shay O'Kelly · August 2026

The first three essays made an argument for scale. Intelligence has become a manufactured good, the factories cost a trillion dollars a year, and the economic models underestimating what comes next are broken in identifiable ways. But scale is not an investment thesis. Growth and returns are different things, and the history of technology is a graveyard of investors who confused them. Airlines transformed civilization and destroyed capital for a century. The question that matters for money is never whether something is important. It is where the scarcity sits, because value does not accrue to importance. It accrues to whatever stays scarce while everything around it grows.

The lesson everyone half-remembers here is fiber. In the late 1990s, telecom companies raised hundreds of billions and buried tens of millions of miles of fiber optic cable on the thesis that internet traffic would explode. The thesis was right. The traffic came, and the investors were wiped out anyway, because supply outran demand so badly that bandwidth prices collapsed more than 90%¹ and most of the fiber sat dark for a decade. The value did not disappear. It fled up the stack, to the companies that consumed the glutted resource, Google, Netflix, and everyone else who built businesses on top of bandwidth that had become too cheap to meter. This is the template in every consensus forecast today. Infrastructure commoditizes, applications win, so fade the buildout and buy the software. Whether that template applies is, I think, the single most important investment question of the decade. And I think it does not, at least not on the timeline consensus assumes, for a reason you can verify layer by layer.

There is a structural reason to think the template inverts this time, and it comes back to the first essay. The internet economy was defined by operating leverage. Software cost essentially nothing to replicate, bandwidth glutted, and so value pooled in the one layer where marginal cost was zero, the application. Infrastructure was a commodity input to somebody else's margin. Intelligence breaks that logic, because intelligence is the first digital product with a real marginal cost. Every token is paid for in silicon, memory, and electricity. The thing being sold is not code that copies for free, it is the metered output of physical capital, which is exactly what the first essay meant by intelligence becoming a manufactured good. When the product itself consumes scarce physical capacity with every unit sold, operating leverage migrates down the stack, away from the software that spends tokens and toward whoever owns the capacity that makes them. The internet moved value up the stack because infrastructure was abundant. AI moves it down, because for the first time since the industrial era, the binding constraint on the digital economy is physical.

A glut requires supply to outrun demand. The demand side we covered in the last essay. Token consumption is compounding at several hundred percent a year, already double what equipment makers forecast for 2028². And even that understates it, because what we are watching is two exponential curves stacked on top of each other, running on different clocks. Model capability is already well up the steep part

of its curve, compounding through scale and now through models that think longer and act as agents. Diffusion, the share of the economy actually using this stuff, is still at the flat beginning of its own curve. Most companies are running pilots, most workflows are untouched, and agents barely exist in production. Adoption curves always look flat right before they go vertical. When diffusion hits its ramp while capability is still compounding, the two multiply, and today's token numbers, the quadrillions that already broke every forecast, will look like the quiet part. Inference is on its way to becoming the largest market in the world. So the question reduces to supply, and the receipts say supply is rationed at every layer of the stack. Walk it from silicon to power.

Start with silicon. Every leading-edge chip flows through ASML's EUV machines, roughly 65 shipping this year³. TSMC's advanced packaging capacity, the CoWoS lines that bolt GPU dies to their memory, is sold out through 2026 with lead times stretching deep into 2027, despite capacity roughly quadrupling in two years, from the mid 30,000s of wafers a month to a planned 130,000 by the end of this year, and still falling short of demand⁴. High-bandwidth memory is the same story. Samsung, SK Hynix, and Micron have sold out their entire output for the year ahead, and SK Hynix concedes 2026 is almost fully allocated⁵. Sit with what that means. The capacity expansions are selling out before they are built. In fiber terms, it is as if every mile of cable had been leased before it was buried. That is not how gluts form. That is how shortages persist.

Within silicon, memory deserves its own paragraph, because the market is mispricing a structural change as another cycle. Model architecture is doing something specific to hardware. Mixture-of-experts designs and relentless algorithmic efficiency keep cutting the compute needed per token, while total parameter counts keep climbing as models get more precise and have more training compute behind them, and context windows keep stretching as agents hold more in working memory. Less compute per token, with more parameters and context to hold, is a formula for inference becoming memory-bound. The AI server is quietly turning into a memory product with a logic chip attached, and HBM4, ramping now for the next accelerator generation, pushes memory's share of the bill of materials higher again⁶. Yet the memory makers are still priced like the pure commodity cyclical they were for thirty years, valued as if today's demand is a peak that mean-reverts on schedule. Given what the architecture curve is doing to memory intensity, I think those valuations are laughable. The market is pricing the most structurally advantaged layer of the stack off its most traumatic historical pattern.

Now power, which everyone, including the government, calls the great bottleneck. The numbers are real. GE Vernova's turbine backlog hit 116 gigawatts, production slots are sold out through 2029⁷, grid transformers that took a year before the pandemic now average nearly three, with worst cases toward five⁸, and of the twelve gigawatts of US data center capacity announced for 2026, only five are actually under construction⁹. But here I hold a minority view. Power is a softer bottleneck than the market believes, because it has workarounds and the hard bottlenecks do not. Power is a modest share of a data center project's cost, perhaps a tenth. And there are many ways around it. You can site campuses where stranded power already exists, build behind-the-meter generation and skip the interconnection queue entirely, restart nuclear plants, deploy solar and storage in the desert, or shift flexible training workloads to off-peak hours. Every one of those paths is being used right now, and the turbine backlog is partly just the receipt

for the workarounds being bought. Power is a real constraint that raises costs and adds years. It is not a wall. There is no equivalent list for lithography. Nobody sites around ASML. Nobody builds behind-the-meter TSMC. The semiconductor chain is a bottleneck in the strict sense, single points of failure, sold out for years, with expansion gated by physics and decades of accumulated process knowledge that no amount of capital can shortcut. When I rank where value accrues, I rank by substitutability, and the scarcest thing in the world right now is not a megawatt. It is a CoWoS slot.

This points to the principle that organizes the whole investment landscape. The bottleneck migrates, and value migrates with it. In 2023 and 2024 the constraint was GPUs themselves, and essentially all the economics of the boom pooled at Nvidia. By this year it had spread to memory and packaging, and businesses Wall Street priced as commodities for thirty years started printing record numbers with multi-year visibility. Power equipment is having its moment now, with backlogs stretching past 2029. But migration is only half the story, because each layer keeps its pricing power only as long as it stays scarce, and layers differ in how long they can stay scarce. Constraints with substitutes get engineered around in a few years. Constraints without substitutes hold for as long as demand holds. That is why I weight the semiconductor chain over energy even as energy gets the headlines. The returns are made by owning the constraint that cannot be routed around, before consensus stops pricing it as a cycle.

It is just as important to be clear about what is not scarce, and the uncomfortable answer is intelligence itself. The price of a token of a given capability has been collapsing continuously, open-weight models trail the frontier by months, and the labs compete each other's margins away even at extraordinary revenue scale. The biggest of them run rates in the tens of billions and still lose money on every dollar. The product is miraculous and the economics of producing it are brutal, which is the oldest combination in technology. On a decade horizon I would push this further. I believe everything at the application layer eventually commoditizes and gets priced toward the cost of tokens, because any workflow built on rented intelligence can be rebuilt by anyone else renting the same intelligence. And I believe the models themselves end the same way. Once models are doing the research and building the next generation of models, the advantage of any one lab's researchers stops compounding, frontier capability converges, and open-source implementations run everything toward the marginal cost of inference. The endgame of manufactured intelligence is intelligence priced like a utility. The one scenario that stops it is the big one. If AI becomes a first-order geopolitical concern, governments wall off frontier capability the way they wall off enriched uranium, and commoditization stops at national borders. I think that concern is real, and the next essay argues it is closer to the base case than the exception. Either way, the conclusion for where value accrues is the same. The durable rents are not in the layer where competition is a price war between geniuses, and not in applications being priced toward token cost. They are in the layers underneath, where supply is rationed by machine counts, backlogs, and permits, and where no amount of brilliance lets a competitor conjure a turbine slot before 2030.

Which brings me to the mispricing, because a bottleneck everyone can see should already be in the price, and the reason I think it is not comes down to how Wall Street models these businesses. Turbine makers, memory companies, electrical equipment suppliers, and utilities are all priced as cyclical. The models assume today's demand is a peak, margins mean-revert, and backlogs are a moment in a cycle rather than a

change in regime. That assumption is exactly what the first essay argued against. If model capability and diffusion keep outrunning consensus, then compute demand estimates are meaningfully too low, and what looks like the top of a cycle is actually the early innings of a permanent step-change in demand for power and silicon. The gap between those two readings, priced as a cycle, behaving like a regime, is where the returns live. It is the gap I am positioning for, long the bottlenecks, without leverage, because the thesis needs time more than it needs timing.

One more thing follows from all of this, and it opens the next essay. When a handful of chokepoints, one lithography company, one leading foundry, a few turbine makers, and the electrical grid, decide which nations get to manufacture intelligence and how fast, those chokepoints stop being industrial assets and become strategic ones. Governments have noticed. The buildout is already reorganizing alliances, export rules, and industrial policy, and that collision between the bottlenecks and geopolitics is where we go next.

SOURCES

1. Dot-com fiber overbuild and >90% bandwidth price collapse. Widely documented telecom history (McKinsey and FCC retrospectives on the 1996-2002 buildout).
2. Token consumption ~2.4x Dell's 2028 forecast, and Goldman Sachs projecting ~70x token growth 2025-2030. Via IO Fund and Goldman Sachs Research.
3. ASML EUV shipments (~65 in 2026). ASML guidance via TechPowerUp.
4. TSMC CoWoS sold out through 2026 with lead times of 52-78 weeks into 2027, and capacity from ~35k wafers/month (end-2024) to ~125-130k planned by end of 2026. TSMC shareholder meeting statements via TrendForce and DigiTimes, and SemiAnalysis.
5. HBM output sold out at Samsung, SK Hynix, and Micron, with SK Hynix 2026 allocation. Via AI CERTs, TrendForce, and company earnings commentary.
6. HBM4 ramp and rising memory share of accelerator bill of materials. Via SemiAnalysis and Wing VC, "The Memory Triopoly."
7. GE Vernova 116 GW turbine backlog, slots sold out through 2029, ~10 GW remaining across 2029-30, 5-7 year waits. Via Turbomachinery Magazine and Power Engineering.
8. Grid transformer lead times averaging ~2.5-3 years vs ~12 months pre-pandemic, worst cases of 4-5 years. Wood Mackenzie survey via T&D World, and The Conversation.
9. 5 GW under construction of 12 GW announced 2026 US data center capacity. Sightline Climate via Tech Fund.

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.