WORKING PAPER

When the chips are down:
Interdependence and concentration in the artificial intelligence supply chain

Pietro Benedetti

July 2026

Abstract. Public discussion of artificial intelligence is framed as a contest between independent firms — OpenAI against Anthropic, Microsoft against Google and so on. This paper argues that this framing is misleading. Using a descriptive, evidence-based approach, it documents two overlapping forms of structural interconnection in the AI industry: technological and infrastructural dependence on a shared hardware base; and value-chain and financial entanglement, including the circular financing arrangements that bind suppliers to their own customers. Drawing on company filings, market-research estimates, and financial reporting, the analysis shows that the sector's apparent plurality rests on a small number of indispensable nodes — above all an axis running through Nvidia, TSMC, and ASML — and that the same structure which makes the industry efficient also makes it fragile. The appropriate unit of analysis, the paper concludes, is not the individual firm but the network in which it is embedded.

Keywords: *artificial intelligence, supply chain, market concentration, interdependence, systemic risk, semiconductors*

1. Introduction

The public story of artificial intelligence is a story of competition. OpenAI races Anthropic; Microsoft races Google; a fresh crop of startups and models arrives every few months, and the coverage frames the whole thing as a contest of rivals straining to outbuild one another. This paper starts from a different premise: that the contest, though real, is played out on top of a base so concentrated and so widely shared that the industry is better understood as a single interdependent system than as a field of independent firms. The unit of analysis that makes sense of it is not the company but the network in which the company sits.

The claim is not that competition is an illusion. It is narrower, and, I believe, more useful: beneath the visible rivalry lies a small set of nodes on which every competitor depends at once, and these nodes recur across two different dimensions of the industry — the hardware it runs on and the money that funds it. Follow either and the same usual suspects appear.

The approach is descriptive rather than inferential. The paper is an attempt to assemble, from public evidence, an accurate picture of how the pieces of the AI industry connect, and so to shed light on what the public believes to be the internet of the third millennium. The evidence for this purpose is mostly quantitative — market shares, manufacturing capacity, dollar commitments.

Section 2 describes the sources and how they are used. The analysis then proceeds in two parts. The first concerns the hardware: Section 3 examines the technology stack and shows that every model, whoever builds it, funnels down through the same chokepoints — foundry, memory and lithography — while Section 4 places those chokepoints on a map and asks what their physical

concentration means for the industry's resilience. The second follows the money: Section 5 shows capital circulating among the same firms, in arrangements where suppliers finance the customers who then buy from them. Section 6 draws the two together, and Section 7 concludes.

2. Sources and methods

Because this is a descriptive study, its method is largely a question of where the numbers come from and how much weight they can bear. The evidence falls into three kinds. The first is disclosure by the companies themselves — earnings reports, investor presentations, and the announcements through which the major deals of 2025 and 2026 were made public. The second is estimates from market-research and industry-analyst sources, which supply the market-share and capacity figures that firms do not publish directly: the share of advanced chips fabricated by TSMC, of accelerators sold by Nvidia, of high-bandwidth memory shipped by SK Hynix, and so on. The third is financial reporting from the business press, used chiefly for the investment and infrastructure commitments described in Section 5.

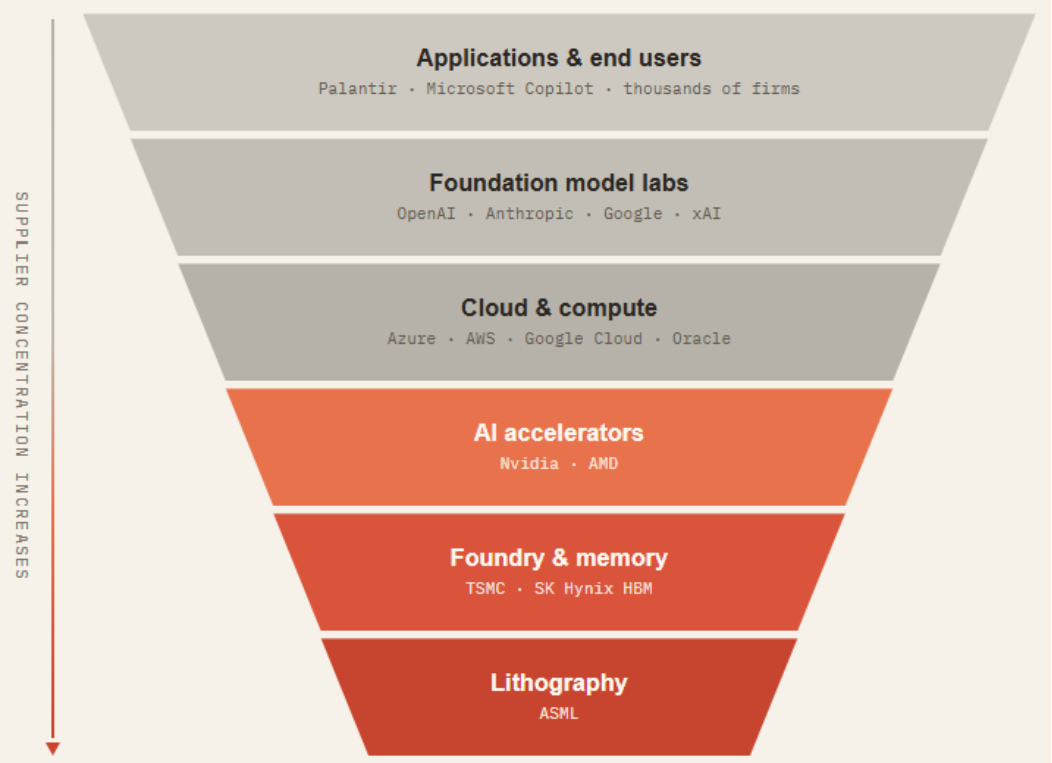
Where the same fact could be sourced in more than one way, I have preferred company disclosure to secondary reporting. Numbers are generally given as approximations or ranges, both because that is how the underlying sources report them and because false precision would mislead: this is a fast-moving industry, and a market share or a valuation quoted to the decimal point would be stale within months.

Three limitations of the evidence should be stated at the outset, since they shape how the later chapters ought to be read. First, several of the largest figures — the trillion-dollar infrastructure commitments in particular — are announced intentions rather than realised spending, are spread over many years, and cannot simply be added together; they convey the scale and shape of the arrangements, not cash that has changed hands. Second, some of the relevant numbers, especially the private valuations and terms of the model labs, are only partially disclosed. Third, market-share estimates vary with the metric used — revenue, units, or shipments — and with the date, so the figures here are best read as orders of magnitude rather than audited accounts.

3. The technological stack: anatomy of dependence

The biblical contrasts of the AI world — Altman against Amodei, Nadella against Pichai, and a steady churn of new models and startups competing with one another — make the industry look like a crowded and competitive field. That impression is not wrong, but it describes only the top of a structure that narrows sharply the further down one looks. It helps to picture the industry as an inverted pyramid: at the top sit thousands of applications and firms; below them, a handful of laboratories that build the models; below those, a few cloud providers with the machines to train them; and below those, the physical layers — the chips, the memory, the equipment that etches the silicon. The number of suppliers shrinks at every step until, at the base, it reaches one, so that each wide layer rests on a narrower one beneath it. The revealing question is not who competes at the top but how few firms hold up the bottom.

Figure 1. The AI technology stack as an inverted pyramid



Before descending into those layers, it helps to sketch the chain in the simplest possible terms. The businesses and applications that use AI need models to run on; the models need vast amounts of cloud computing to be trained and served; the clouds need specialised accelerator chips to do that computing; the chips must be manufactured in advanced foundries; the foundries need the lithography machines that print the circuits; and those machines depend, in turn, on a few irreplaceable components and materials. Demand flows downward, from the visible top to the hidden base — and, as the rest of this section shows, the further down the chain one goes, the fewer the suppliers become.

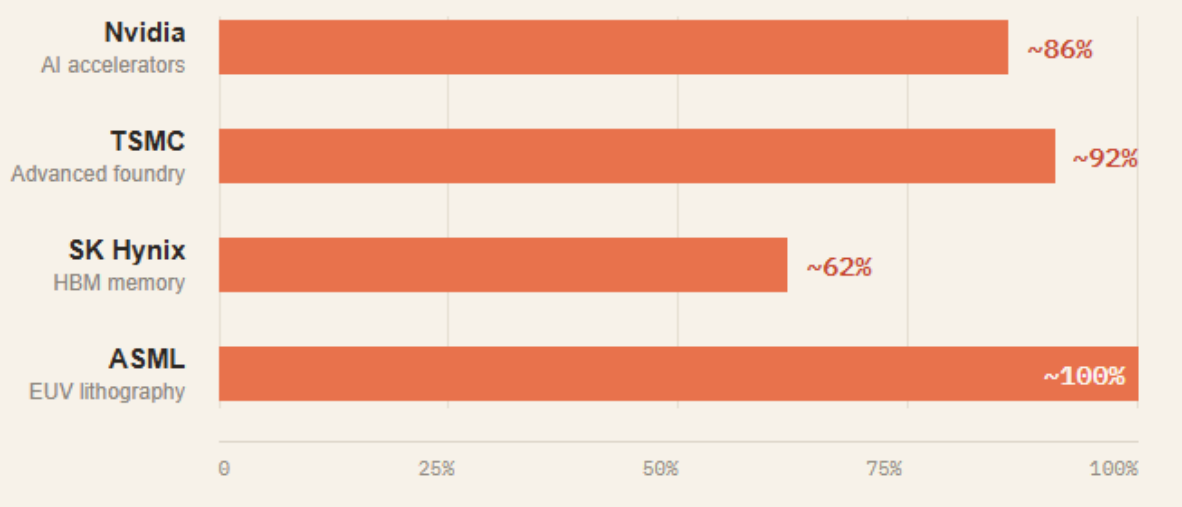
Follow the pyramid down and the room quickly empties. Start with Wall Street's golden boy, Nvidia. The company designs the accelerators that run the models and holds something like 80–90% of that market by revenue — more than $190 billion in data-centre revenue in its 2026 fiscal year (Silicon Analysts, 2026; FinancialContent, 2025). Yet Nvidia does not make its own chips. Every advanced processor in AI, whoever designs it, has to be fabricated, and that work is concentrated in one company in one country: TSMC in Taiwan produces roughly 92% of the world's advanced chips (MLQ.ai, 2026; Big Data Supply, 2026). To complicate things further, fabrication is not the whole story. A finished accelerator is not a single chip but a processor die bonded onto a module alongside its stacks of high-bandwidth memory, and the dominant process for that final assembly — TSMC's CoWoS packaging — is a bottleneck in its own right, booked so far ahead that Nvidia alone holds about 60% of the capacity and the three largest customers more than 85% (Silicon Analysts, 2026). Owning a chip design and the memory to pair with it means little without a place in that queue.

For readers less at home with the technology, the point in plainer terms is this: designing a chip is like drawing the blueprint for an engine, but the blueprint still has to be cast in a foundry — that is TSMC — and the finished parts still have to be bolted together on a specialised assembly line

— that is packaging. In the whole world there is essentially one foundry and one assembly line capable of doing this at the frontier, and both are booked well in advance.

The memory is a third narrow point. High-bandwidth memory, which feeds data to the GPUs fast enough to keep them busy, comes almost entirely from South Korea, where SK Hynix supplies around 62% of shipments and has sold out its production until at least the first quarter of 2027 (SK Hynix, 2026; Tom's Hardware, 2025). Of all the bottlenecks this is the least concentrated — three producers rather than one — yet it is already spoken for years ahead, a reminder that scarcity can come from rigid supply as much as from a single dominant name. And beneath all of it lies the narrowest point of the entire structure: the machines that etch the chips. The most advanced logic can only be printed using extreme-ultraviolet (EUV) lithography, and just one company in the world builds those machines — ASML, in the Netherlands — which in turn depends on a still smaller circle of German suppliers for its optics and lasers (Hardware Busters, 2026; CommonWealth Magazine, 2025; The Generalist, 2025). At the very bottom of an industry defined by American and Asian giants sit a few European firms most people have never heard of, and without them nothing above can be built at all.

Figure 2. Share held by the single largest supplier at critical layers of the AI stack

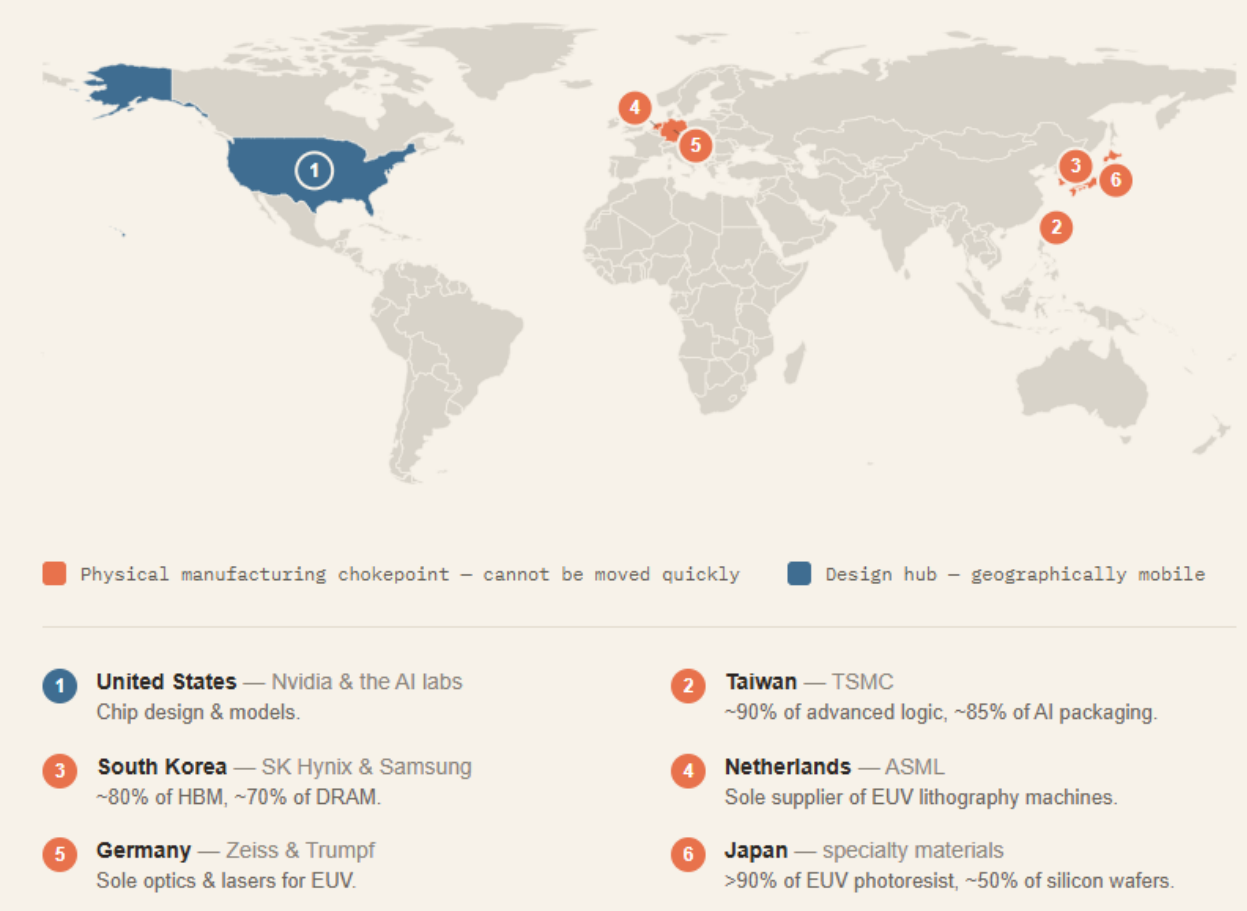


Move down the pyramid and the alternatives disappear one by one. At the very top there is real plurality: Nvidia designs the great majority of AI chips, but AMD competes with it, the largest cloud companies design their own silicon — Google's TPUs, Amazon's Trainium — and the labs increasingly split their workloads across several at once, as Anthropic does in training Claude on AWS Trainium, Google TPUs and Nvidia GPUs alike (Anthropic, 2026). Yet every one of those designs still narrows to a single foundry, then to a single packaging process, and finally to a single lithography supplier. That is the difference between a competitive sector and an interdependent one: in a competitive market a substitute exists at every level, whereas here the many alternatives at the top all converge on the same few nodes at the bottom — which is exactly why a disruption at the base would strike the whole system at once.

4. The geography of fragility

It is worth placing these players on a world map. Market concentration can, in principle, be diluted by new entrants, but geography cannot be moved without rebuilding, over many years and at vast cost, some of the most complex factories ever constructed. When a single layer depends on a single company, and that company depends on a single site, market power turns into geopolitical exposure. Mapped out, almost the entire physical base of the AI industry sits in just five places — a narrow band of East Asia and two clusters in Europe — while only the design layer, the one part that could in principle be relocated, is anchored in the United States.

Figure 3. The geography of the AI supply chain



The heart of it is East Asia. Taiwan comes first: TSMC fabricates roughly 90% of the world's most advanced logic chips and, by one estimate, close to 99% of the chips used to train frontier AI models, and the island also handles about 85% of advanced packaging (Statistics of the World, 2026; The Hilltop, 2026; Silicon Analysts, 2026). The dependence is so complete that Taiwan's own economy has become an outlier, with chips making up roughly 78% of its exports (Statistics of the World, 2026). This is the origin of the “silicon shield”: the idea that the world cannot afford a disruption to Taiwan because it cannot function without what Taiwan makes. Now let’s stop for a second for a gentle reminder of what Taiwan is — a much-contested island barely 130 km from mainland China, which Beijing claims as its own territory. The shield, in other words, is also a

measure of how much of the world's technology rests on one disputed island. Next door lies the second East-Asian node, and it carries its own single-country dependency: memory. High-bandwidth memory comes almost entirely from South Korea, where SK Hynix and Samsung together hold an estimated 80% of the HBM market and around 70% of all DRAM (Silicon Analysts, 2026; Counterpoint Research, 2026). Every Nvidia accelerator is fed by these Korean stacks, so a disruption on the peninsula would starve the GPUs even if every fab in Taiwan kept running.

The other two nodes lie in the older industrial economies of Europe and Japan, which supply the tools and materials on which even Taiwan and Korea depend. Europe holds the deepest chokepoint of all: the Netherlands is home to ASML, the sole maker of EUV lithography machines, and those machines in turn rely on Germany, taking their optics only from Carl Zeiss and their lasers only from Trumpf (CommonWealth Magazine, 2025; The Generalist, 2025) — a small “deep-tech” triangle on which the world's ability to make advanced chips ultimately rests. The point is easy to miss: when people worry about chip dependence they picture Taiwan and China, yet the single deepest bottleneck of all — the one machine without which no advanced chip can be made — sits not in Asia but in Europe, which makes the world's exposure both wider and less watched than the usual focus on Taiwan suggests. Japan supplies the layer beneath even that: the specialty chemicals and substrates without which no fab can run, including more than 90% of the photoresist used in EUV lithography and roughly half of the world's silicon wafers (Silicon Analysts, 2026). That this is a real chokepoint and not a theoretical one became clear in 2019, when Japan restricted photoresist and hydrogen-fluoride exports to South Korea in a trade dispute and the move was read at once as a threat to the entire memory industry (Silicon Analysts, 2026).

The one real exception is the United States, and it is an exception precisely because it is the layer that can move. The American node — chip design, together with the model labs themselves — is the least physically fragile of all, because design is intellectual work: it can be done almost anywhere, and a designer who loses access to one fab can, in principle, port to another. That is why the map marks the United States differently from the manufacturing chokepoints. American leadership in AI is real, but it is leadership of the one layer that can move, resting on layers that cannot.

The upshot is a supply chain that runs smoothly in normal times and is acutely exposed to shocks. The recent past offers a string of near-misses — the 2019 Japan–Korea export dispute, the 2021 Texas winter storm that idled Samsung's Austin fab for six weeks, and the 2022 disruption to neon supply, half of it semiconductor-grade, after Russia's invasion of Ukraine (Silicon Analysts, 2026) — against which most fabs hold only one to three months of critical materials. And the concentration has proved very hard to unwind: TSMC is spending some $165 billion on fabs and packaging in Arizona and SK Hynix is building HBM packaging in Indiana, yet Taiwanese officials have publicly called the goal of moving even 40% of chip supply to American soil “impossible,” a project measured in decades rather than years (The Hilltop, 2026). For the foreseeable future, the geography of the AI industry is also the geography of its fragility.

5. The value chain and circular finance

We are used to reading a supply chain in one direction, upstream from the customer who pays to the suppliers who are paid — models, chips, capacity, memory. In the AI industry the reverse flow is just as interesting, because over the course of 2025 the suppliers began financing their own customers. In September 2025 Nvidia and OpenAI announced that Nvidia would invest up to $100 billion in OpenAI, which would in turn deploy at least ten gigawatts of Nvidia systems and buy the

chips to fill them (CNBC, 2026). The money was meant to travel in a loop — from Nvidia into OpenAI, and from OpenAI back to Nvidia as hardware orders — and what made the announcement notable was less its size than its nature: the largest supplier in the industry underwriting the demand for its own product. It also proved less solid than the headline implied. The deal was only a letter of intent; by early 2026 no contract had been signed and no money had moved, and the commitment was being restructured into a smaller equity stake of roughly $30 billion inside a broader OpenAI funding round (TechTimes, 2026; CNBC, 2026).

Nvidia's arrangement was one strand of a much larger web, with OpenAI near its centre. Days after the Nvidia announcement, OpenAI signed a firmer agreement with AMD — this one with signatures — to deploy six gigawatts of AMD accelerators, and in exchange AMD granted OpenAI warrants for up to 160 million shares, close to a tenth of the company, vesting as OpenAI actually installs the hardware (CNBC, 2026). Around these sit further commitments to Oracle, Microsoft, Amazon, Broadcom and the specialist cloud provider CoreWeave, adding up on paper to more than a trillion dollars over the coming decade (Tunguz, 2025). The exact total is debatable, and in truth beside the point. What matters is the shape of the arrangement: a supplier lending its customer the money to buy the supplier's own products. Stripped of the financial language, it is a little like a fireman who quietly starts fires so that he will always be called to put them out — some of the demand that keeps the industry growing is boosted by the very firms that profit from meeting it.

The same pattern surrounds OpenAI's chief rival. Amazon has invested about $33 billion in Anthropic, which has committed to some $100 billion of AWS spending and five gigawatts of Amazon's Trainium chips, while Microsoft and Nvidia together put in up to $15 billion in early 2026, against which Anthropic committed roughly $30 billion of Azure capacity (Forbes, 2026). The striking part is not that investors spread their bets — everyone diversifies — but that Nvidia and Microsoft turn up as backers of both OpenAI and Anthropic, the two direct rivals that are also their customers. This is less a neutral investor hedging across an industry than a supplier holding a stake in both sides of the race that buys its products: whichever lab wins, its chief supplier wins too, and neither lab is left much room to loosen its dependence on it.

The loops tighten one layer further down. Nvidia holds an equity stake in CoreWeave and has agreed to buy billions of dollars of its capacity, and CoreWeave in turn rents that Nvidia-filled capacity to the very hyperscalers — Microsoft, Oracle — that serve OpenAI (IDC, 2026; Bloomberg, 2026). Microsoft's own investment in OpenAI has been made partly in Azure credits, so that a portion of OpenAI's compute spending simply returns to Microsoft as revenue (CraftedCharts, 2026). By 2026, analysts estimated that arrangements of this kind exceeded $800 billion across the sector (Alatirok, 2026).

Whether all of this is prudent or precarious is genuinely contested, and the strongest version of each case deserves stating. Defenders argue that when advanced chips are scarce and fabs take years to build, pairing long-term purchase commitments with financing is simply how a company secures supply; the asset manager Janus Henderson has described the deals as a “virtuous circle” that lines up suppliers, builders and buyers to meet real demand (Bloomberg, 2026). And the demand does appear to be partly real, since enterprises and consumers are paying for these products.

The opposing case rests on what circularity can conceal. When a supplier funds a customer that spends the money back on the supplier, revenue can look robust and demand organic even when much of it is the same capital moving around a small ring. The underlying economics are not yet reassuring: OpenAI is reportedly on course to lose around $14 billion in 2026, even as it projects $100 billion in annual revenue by 2029 (Alatirok, 2026). Some of the deals generate margins thin enough to resemble a subsidy — Oracle has been reported to earn roughly 14% gross margin on

its Nvidia-powered cloud sales to OpenAI, against about 70% elsewhere (Finomics Edge, 2025). Strain of this kind tends to surface first on the infrastructure providers' balance sheets, in rising debt and lease commitments, which is why the report in February 2026 that the Nvidia–OpenAI deal had stalled was felt at Oracle — a company borrowing heavily to build the data centres OpenAI has promised to fill.

> **A historical parallel.** History offers a close precedent, and not an encouraging one. In the late 1990s the great equipment makers of the telecom boom — Cisco, Lucent and Nortel — lent heavily to their own customers so those customers could keep buying equipment; Cisco alone extended some $2.4 billion and briefly became the most valuable company in the world, worth around $450 billion. When the demand proved to have been inflated by the financing itself, the structure gave way: dozens of carriers went bankrupt between 2000 and 2003, the vendors wrote off billions in unpaid loans, and Cisco's shares fell nearly 90%, never again reaching their 2000 peak (Hansen Zheng, 2025; Ventures Edge, 2025). The lesson is not that vendor financing always ends this way — railroads and earlier computing build-outs used it to bootstrap genuine industries — but that it makes a sector's reported growth an unreliable guide to its real health.

Seen this way, the financial structure of the AI industry rhymes with its physical one. Just as the hardware of every model funnels down through the same foundry, the same memory and the same lithography, the money that funds the build-out circulates through the same small group of chipmakers, clouds and labs. The interdependence traced in the previous chapters is not only a matter of silicon; it is also a matter of capital, and the two reinforce each other.

6. Convergence: the single organism

The previous chapters each followed a different thread, and each arrived at the same place. The hardware of every model funnels down through one foundry, one packaging process, one memory duopoly and one lithography supplier. The capital that pays for the build-out circulates among a handful of chipmakers, clouds and labs, often looping from a supplier to a customer and back again. Two threads that might have been independent instead point to a single conclusion: at its base, the AI industry is not many firms but a single tightly coupled system, resting on a few shared organs.

Those organs can almost be named individually. An axis running through Nvidia, TSMC and ASML — chip design, fabrication, and the machines that make fabrication possible — sits beneath essentially everything. Every strand of the analysis converges here: it is where the market shares are highest, where the geography is most concentrated, and where a disruption would do the most damage. It is the industry's spine, and a double-edged one: the same integration that makes the system efficient — capital, capacity and expertise pooled in the hands of the few players able to operate at the frontier — is precisely what makes it fragile. In a loosely coupled market the failure of one firm is absorbed by its competitors; in a tightly coupled system a shock to a shared node passes through everything that rests upon it. The interdependence is not a side effect to be managed but the defining feature of the structure.

A small illustration arrived while this paper was being written. In February 2026 the reported stalling of the Nvidia–OpenAI investment — a financial arrangement, not a physical failure — was enough to unsettle Oracle, a company that had been borrowing to build the data centres OpenAI had promised to fill. One loosened tie in the financial layer was transmitted, within days, to a firm two steps away. It is not hard to extrapolate. A financial wobble is the mildest shock the system can receive; a physical one — an earthquake or a blockade around Taiwan, an export restriction

on ASML, a fire at a Korean memory fab — would strike the spine directly, and there is no second supplier waiting to absorb it.

7. Conclusions

This paper set out to examine a simple proposition: that in artificial intelligence the individuality of the firm has largely dissolved, and that the industry is better understood as a single interdependent system than as a collection of competitors. The evidence assembled here supports that proposition on two fronts. The hardware of every model passes through the same handful of makers, and the capital that funds the industry circulates among the same handful of backers. Each, taken on its own, points to the same small set of nodes — an axis through Nvidia, TSMC and ASML — on which the whole structure rests.

What follows is the double character the paper has returned to throughout. The concentration that makes the AI industry extraordinarily efficient is the very thing that makes it fragile, because a system with few shared organs has few places to absorb a shock. That is a statement about structure, and it holds regardless of how the current boom resolves.

Only time will tell how the tension will evolve: whether the circular financing proves a “virtuous circle” or a bubble, and whether the physical chokepoints are diversified away or struck by a shock before they can be.

It is easy to be dazzled by the shiny new toys that artificial intelligence keeps handing us; it is harder to notice that all of their power hangs from a very thin thread.

References

Alatirok (2026). AI circular financing explained: the Nvidia–OpenAI–Oracle money loop. alatirok.com.

Anthropic (2026). Expanding our use of Google Cloud TPUs and services; Anthropic expands partnership with Google and Broadcom for next-generation compute. anthropic.com.

Big Data Supply (2026). 15 leading AI hardware companies dominating the market in 2026. bigdatasupply.com.

Bloomberg (2026). AI circular deals: how Microsoft, OpenAI and Nvidia keep paying each other. bloomberg.com.

CNBC (2026). OpenAI chip deal with Cerebras adds to roster of Nvidia, AMD, Broadcom; Nvidia, OpenAI appear stalled on their mega deal. cnbc.com.

CommonWealth Magazine (2025). Taiwan enters the angstrom era with ASML's High-NA EUV (E. Huang). english.cw.com.tw.

Counterpoint Research (2026). Global DRAM and HBM market share (quarterly). counterpointresearch.com.

CraftedCharts (2026). AI circular financing map: Nvidia, Microsoft & OpenAI. craftedcharts.com.

FinancialContent / PredictStreet (2025). NVIDIA: powering the AI revolution and navigating a trillion-dollar future. markets.financialcontent.com.

Finomics Edge (2025). Circular financing in AI: Nvidia funds OpenAI to buy its own chips. medium.com.

Forbes (2026). Amazon's $33 billion Anthropic deal and the limits of AI infrastructure (J. Markman). forbes.com.

The Generalist (2025). ASML: a monopoly on magic. generalist.com.

Hansen Zheng (2025). Circular vendor financing in the AI sector. medium.com.

Hardware Busters (2026). The quiet monopoly powering every advanced chip in the world. hwbusters.com.

The Hilltop (2026). Taiwan Strait tensions push countries to diversify semiconductor supply chains (citing the U.S. International Trade Administration). thehilltoponline.com.

IDC (2026). Circular financing has muddied the AI story; watch the application layer instead. idc.com.

MLQ.ai (2026). AI chips & accelerators. mlq.ai.

Silicon Analysts (2026). NVIDIA AI accelerator market share 2024–2026; TSMC foundry allocation 2026 (CoWoS sold out); semiconductor supply-chain chokepoint map. siliconanalysts.com.

SK Hynix (2026). 2026 market outlook: focus on the HBM-led memory supercycle. news.skhynix.com.

Statistics of the World (2026). Taiwan economy 2026: the world's most concentrated boom. statisticsoftheworld.com.

TechTimes (2026). Nvidia–OpenAI investment shrinks from $100B to $30B. techtimes.com.

Tom's Hardware (2025). SK hynix to build first U.S. packaging plant for HBM. tomshardware.com.

Tunguz, T. (2025). OpenAI's $1 trillion infrastructure spend, 2025–2035. tomtunguz.com.

Ventures Edge (2025). AI roundtripping: Nvidia, OpenAI, Oracle and the circular financing debate.