PitchBook
a Morningstar company

Institutional Research Group

Dimitri Zabelin
Senior Research Analyst,
AI and Cybersecurity
dimitri.zabelin@pitchbook.com

Oscar Allaway
Senior Data Analyst

pbinstitutionalresearch@pitchbook.com

Published on October 20, 2026

Contents

Key takeaways	1
Introduction	3
VC deal activity	4
Regional breakdown	6
Valuations	8
Exits	9
Industry insights	10
Due diligence guide	17
Appendix	22
References	28

EMERGING TECH RESEARCH

Inference Chip Startups Are Just Starting Their Bull Run

Out of the bullpen and into the datacenter

Key takeaways

1. **Bullish on inference chip startups building novel approaches to memory and energy usage:** We are bullish on companies such as Tensordyne and d-Matrix, which address the energy and memory constraints limiting inference. The former's Pareto chip replaces power-intensive multiplications with simpler additions, while the latter performs calculations directly within memory, reducing the need to move data between memory and processors. Both approaches could lower inference costs, but commercial success will depend on reliable performance, compatibility with customers' existing software, and the ability to manufacture and deploy at scale.

2. **Diligence roadmap for VCs investing in inference chips:** VCs diligencing early-stage chip startups should look for startups that bring memory and chips closer together, integrate into existing rack and networking architectures, build robust software components, and have deployed or high-certainty LOIs. Please see our diligence guide below for key questions and focus areas for VCs.

3. **Demand for faster memory–chip communication is driving inference chips:** Inference is set to drive roughly 75% of compute demand by 2030, pulling capital toward purpose-built silicon over general-purpose GPUs. Memory constraints are creating openings for inference chip startups. Both d-Matrix and MemryX identify the time and energy required to move data between memory and processors as a key limitation that additional compute alone cannot resolve. Agentic AI compounds this pressure: Longer contexts, larger models, and repeated inference calls increase the volume of data chips must store and move. We believe these demands will continue to favor firms developing architectures that bring memory and computation closer together. Commercial durability, however, will depend on more than architectural advantages.

4. **Compatibility with existing compute systems, strong software offering, and powerful and certain supply chain program are key to success:** Across our interviews, software compatibility, reliable deployment, and manufacturing readiness emerged as prerequisites for scaling. Startups must demonstrate that their hardware fits into customers' existing infrastructure and that their suppliers can support large orders. Investors should therefore assess performance on representative workloads, accounting for power consumption, integration costs, and the trade-off between total throughput and speed per user. Tensordyne, MemryX, and d-Matrix all caution against relying on isolated benchmarks that reveal little about commercial viability.

5. **Cost per token and compute-type specific hardware becoming important:** The shift toward inference is also changing purchasing behavior. Cirrascale reports that customers are making longer capacity commitments as AI moves into ongoing production use. This creates a more persistent source of infrastructure demand, with purchasing decisions increasingly determined by cost per token and the hardware's suitability for the workload.

6. **Corporate investment is important for winning:** Venture growth captured 51% of funding through August 28, 2026, but strategic corporate investors can offer chip startups more than capital. They can help secure customers, validate products, and provide a path to acquisition. We see the strongest acquisition prospects in startups with technology that buyers cannot easily develop themselves. However, higher entry valuations mean these startups must sell for more to deliver strong returns to VC investors.

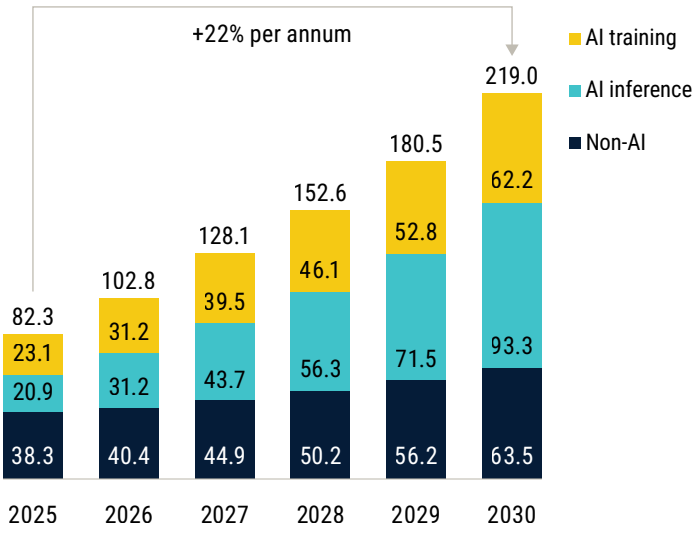
7. **US dominates inference chip capital raises but Europe is making a concerted effort:** North America has captured roughly 85% of inference AI VC dollars since 2025 and will likely continue to widen its lead over China, and certainly over Europe. Deep venture capital markets and proximity to hyperscalers provide chip developers with funding, customers, development partners, and potential acquirers. Asian deal counts are rising, led by China, but US restrictions on advanced chips, semiconductor manufacturing equipment, and outbound investment constrain Chinese firms' access to parts of the global supply chain and capital base. Europe faces a different challenge: limited late-stage funding and the absence of domestic hyperscalers comparable in scale to those in the US. Nevertheless, sovereign AI initiatives and demand for locally controlled infrastructure could create opportunities for European chip developers. For investors, geography therefore shapes both the resources available to build a chip business and the commercial routes available to scale or exit it.

Introduction

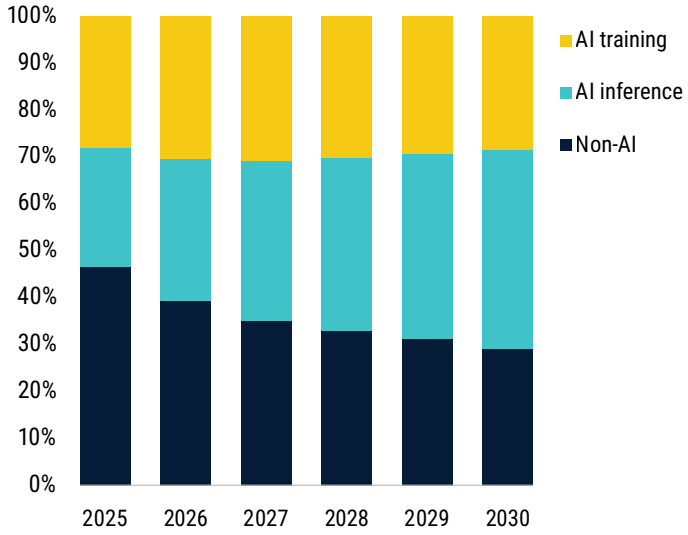
VC-backed, hardware-inference startups sit at a key layer of the global AI infrastructure build-out. Datacenter capital expenditure surged 57% in 2025,[1] is projected to surpass $1 trillion in 2026,[2] and to reach $1.7 trillion by 2030.[3] Chips are the biggest cost input of datacenters, and 50% to 60% of that capital expenditure will go into them. By the end of the decade, an estimated 75% of AI compute demand will come from inference rather than training.[4] McKinsey points in the same direction,[5] projecting inference power demand from 20.9 GW in 2025 to 93.3 GW by 2030, a 35% CAGR that outpaces every other datacenter workload.

The future of AI workloads: Global data center demand by workload 2025-2030

By gigawatts



By share



Source: Exhibit from "The future of AI workloads," February 2026, McKinsey & Company, Copyright (c) 2026 McKinsey & Company. All rights reserved. Reprinted by permission. • Geography: Global • As of February 2026

Agentic AI is the primary catalyst. Gartner projects 40% of enterprise applications will embed task-specific AI agents by the end of 2026,[6] up from under 5% a year earlier, and agentic workflows consume 5x to 30x more tokens per task than single-turn queries.[7] Goldman Sachs expects global token consumption to climb from roughly 5 quadrillion per month today to 120 quadrillion by 2030,[8, 9] while the agentic market expands from $9 billion in 2026 to more than $200 billion by 2033.[10, 11] The result is a structural pull for purpose-built silicon that can sustain low-latency, high-throughput inference at scale.

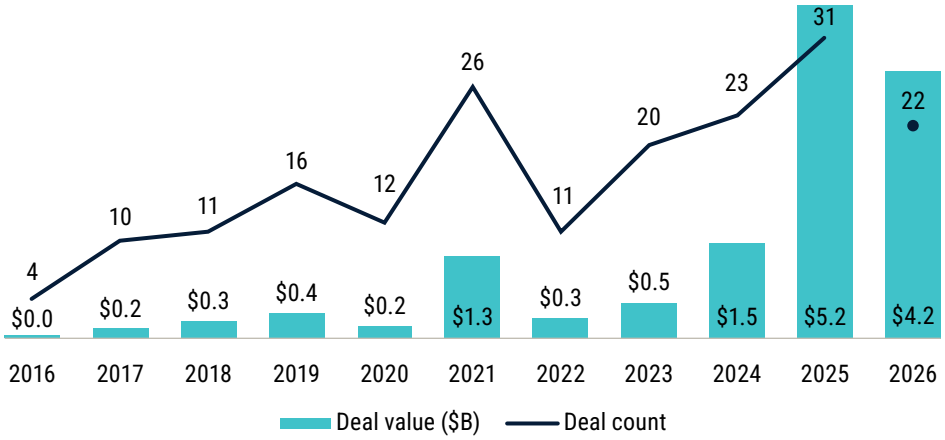
This paper maps the semiconductor companies designing, manufacturing, or licensing that silicon. Chips that also support training are included only where inference is the primary design intent; notable entrants since being acquired or taken public are retained, while general-purpose GPUs, module integrators, and interconnect fabric are excluded. Companies fall into five segments: datacenter inference ASICs, edge inference silicon, compute-in-memory and novel architectures, photonic and optical compute, and semiconductor IP licensing.

VC deal activity

2026's VC-backed AI inference deals are on pace to surpass 2025's $5.2 billion across 31 transactions. YTD through August 28, 2026, activity stands at roughly $4.2 billion across 22 deals. On a quarterly basis, Q3 2026 already ranks second all-time at $2.2 billion behind Q3 2025's nearly $2.5 billion, with four months still left in 2026.

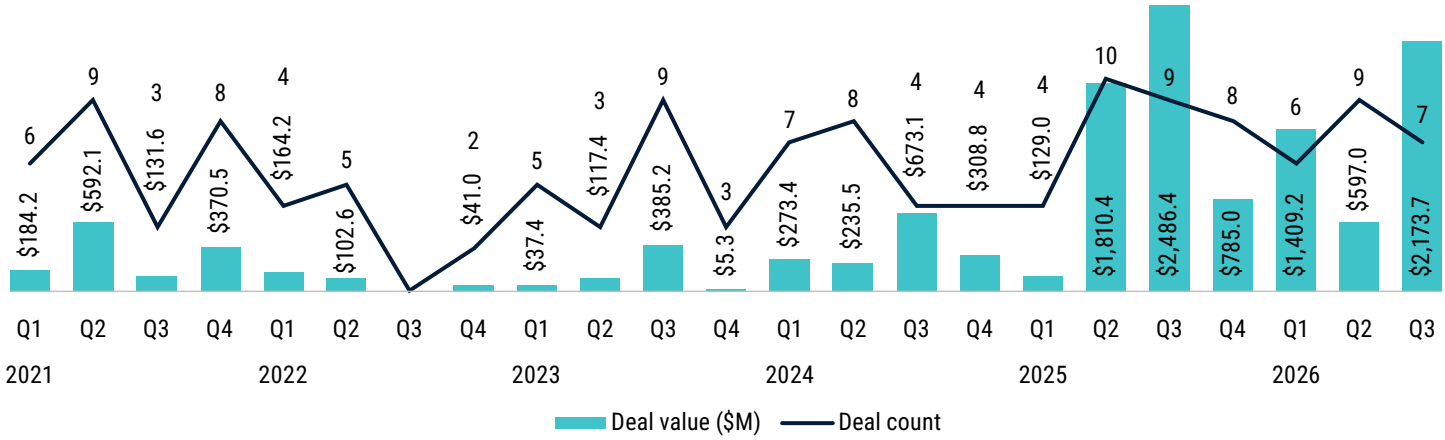
In 2025 and 2026, venture growth accounted for 65% and 51% (so far) of deal value, respectively.

Inference AI VC deal activity



Source: PitchBook • Geography: Global • As of August 28, 2026

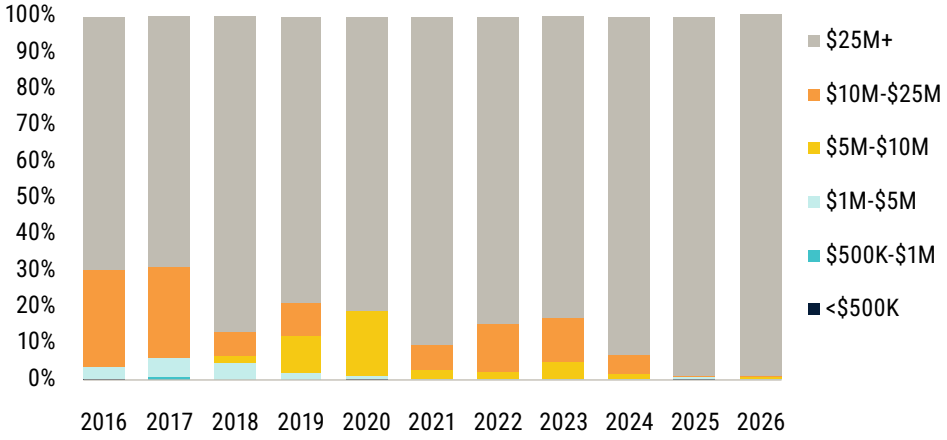
Inference AI VC deal activity by quarter



Source: PitchBook • Geography: Global • As of August 28, 2026

Investors will likely continue rewarding firms that are operating at the infrastructure level and offer novel chip architectures that address key pain points. Later-stage and venture growth firms have asymmetrically magnetized most of the capital, particularly in the past two years; $25 million+ rounds have accounted for 83% to 100% of inference AI VC dollars every year since 2016, and effectively 100% in both 2025 and 2026 YTD.

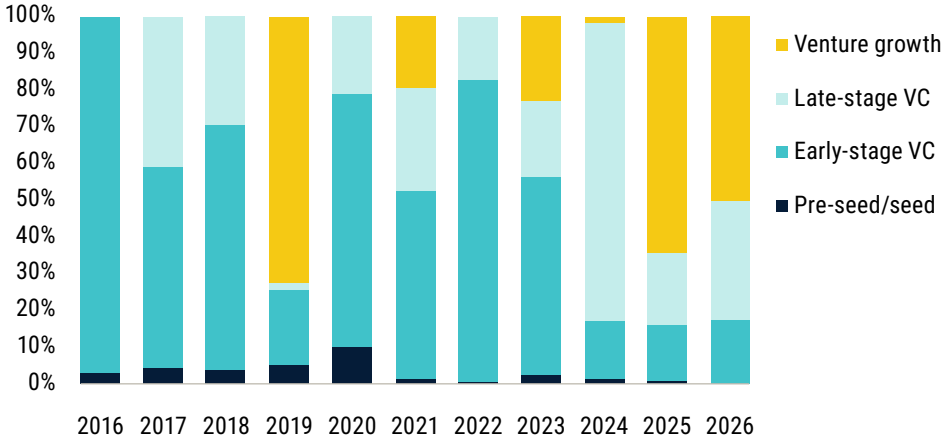
Share of inference AI VC deal value by size



Source: PitchBook • Geography: Global • As of August 28, 2026

Pre-seed and early-stage deal activity remains muted relative to the higher rungs of the capital ladder. But benchmarked within its own league, early-stage funding is quietly beating its own records. For most of the past decade—apart from one standout year—annual deal activity had been rangebound between $0.1 billion and $0.3 billion. 2025 quadrupled to $0.8 billion from $0.2 billion in 2024, and 2026 is sitting at a comfortable $0.7 billion, on pace to beat last year.

Share of inference AI VC deal value by type



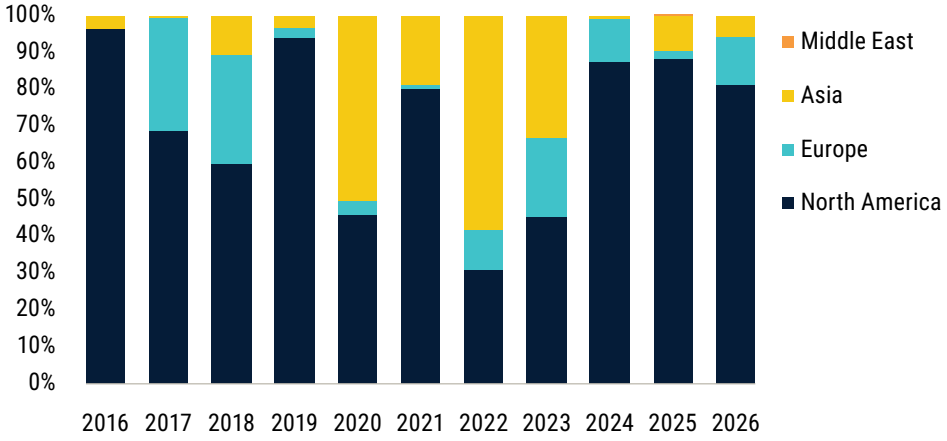
Source: PitchBook • Geography: Global • As of August 28, 2026

We believe there is still fertile ground for chip design firms to blossom. Structural bottlenecks at key points in the AI value chain, such as memory bandwidth and inference energy efficiency, combined with sustained compute demand, are creating openings for firms innovating around GPU limitations.

Regional breakdown

US technological hegemony in both the hardware and software stacks extends to inference-focused chips: since 2025, North America has captured about 85% of all inference AI VC dollars. The VC ecosystem and deeply rooted hyperscaler network in the US enable frontier-carving startups to research and scale operations, thereby further cementing North American supremacy.

Share of inference AI VC deal value by region



Source: PitchBook • Geography: Global • As of August 28, 2026

Deal counts in Asia have been rising, mostly propelled by the region's powerhouse, China. However, the regional giant's strategic rivalry with the US has resulted in regulators in Washington cracking down on Beijing's access to advanced hardware, software, and capital. For macro context, a major theme impacting cross-border investments and trade is the silicon curtain being drawn between the US and China. Washington began restricting China's access to advanced chips and semiconductor manufacturing equipment with the October 2022 BIS export controls,[12] which the BIS tightened again in October 2023.[13] It tightened the controls once more in December 2024.[14] The Treasury then added a second layer through its outbound investment rule,[15] finalized October 28, 2024, and effective January 2, 2025,[16] restricting US citizens from investing in Chinese advanced semiconductor companies.

Marquee US venture firms with China exposure have restructured in response; for example, Sequoia spun out its China business as HongShan in 2023,[17] and GGV Capital split into Notable Capital (US) and Granite Asia in 2024.[18] We expect US-China rivalry and the ripple effects that it generates to continue affecting the investment landscape in private and publicly traded markets. We believe US-China rivalry will increase the strategic value of chipmakers that relieve memory, power, and other compute constraints, as both countries seek to expand AI capacity and reduce dependence on foreign technology. National-security funding has previously targeted these bottlenecks. DARPA's 2017 Electronics Resurgence Initiative included a program explicitly aimed at overcoming the time and energy costs of moving data between memory and processors,[19] supporting research into architectures that bring

the two functions closer together. This provides a precedent for public investment in the technical challenges addressed by companies in this report, although securing commercial customers will remain essential.

For US and European investors, we favor developers with access to Western customers and capital whose growth does not depend on continued access to China. Chinese suppliers may benefit from domestic substitution and state support, but government-linked ownership can complicate cross-border transactions. In 2017, Washington blocked Canyon Bridge's proposed acquisition of Lattice Semiconductor,[20] citing potential intellectual property transfers, semiconductor supply-chain integrity, and the Chinese government's role in financing the transaction. In 2022, the UK ordered Nexperia to divest at least 86% of its ownership of Newport Wafer Fab on national-security grounds.[21] These cases demonstrate that investors can lose a prospective buyer or face the reversal of a completed acquisition, making ownership and exit planning material considerations alongside chip performance.

Pressure can also come from Beijing. Following Didi's US IPO in 2021, Chinese authorities initiated a cybersecurity review and required the company to suspend new user registrations.[22] In May 2022, shareholders approved delisting from the NYSE to better cooperate with the review and corrective measures.[23] Although Didi operates outside the semiconductor sector, its experience illustrates how national-security intervention can disrupt overseas listings and investor liquidity. Therefore, we view restrictions on financing and exits as more immediate risks than outright nationalization.

NVIDIA had already ordered H20 chips from manufacturers for sale to Chinese customers when new US export restrictions disrupted those sales in April 2025. Left with unsold chips and outstanding production commitments, it recorded a $4.5 billion expense,[24] illustrating how policy changes can leave chipmakers unable to recover manufacturing costs. For inference startups with smaller balance sheets, similar disruptions could threaten operations and force additional fundraising. Therefore, we favor companies whose commercial viability does not depend on continued access to Chinese customers.

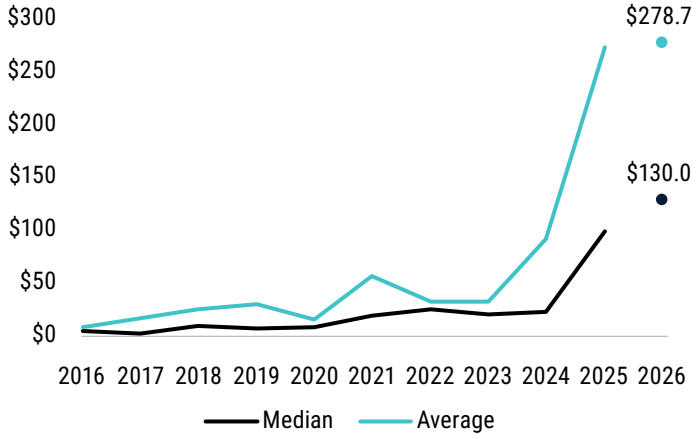
European activity remains subdued amid persistent structural challenges, including bureaucratic constraints, the absence of domestic hyperscalers, and limited late-stage capital sources. Europe also sits in an awkward position: It has neither the market forces of the US nor the political structure of China that enables it to quickly deploy targeted, state-directed capital. However, it would be foolish to label the continent irrelevant. The EU AI Act creates commercial demand for jurisdiction-controlled inference by making it harder for European enterprises to run high-risk AI on US hyperscaler infrastructure. The European Chips Act provides foundry access via TSMC Dresden and Intel Magdeburg, plus design-stage funding through national programs. The European AI Factories initiative, together with national sovereign cloud efforts, is building a publicly funded buyer base with an explicit mandate to procure diverse, non-NVIDIA silicon. Together, these forces create a commercial pull for European-designed inference infrastructure, even where deal counts remain modest.

PitchBook data identified just two seed-stage, VC-backed European startups working on hardware for AI inference that raised capital in 2026: AeroSilicon and Lamb Labs. AeroSilicon develops custom designs for programmable chips, with applications including on-device inference, while Lamb Labs is developing dedicated inference chips. AeroSilicon secured $200,000 in pre-seed funding through SAFE notes from Davidovs Venture Collective. Lamb Labs raised an undisclosed amount on September 18 following a $500,000 accelerator/incubator round on September 10. Broadening the taxonomy to include inference software and cloud infrastructure increases both deal count and capital raised. Callosum develops software that coordinates AI workloads across different types of chips and raised $100 million in June 2026. TensorX provides private inference services using GPUs hosted in the EU, securing $8 million that same month. Meanwhile, Cyanis AI is developing software to coordinate inference across distributed computing sites, bringing processing closer to users to reduce delays. The company raised $290,000 in the same month.

Valuations[25]

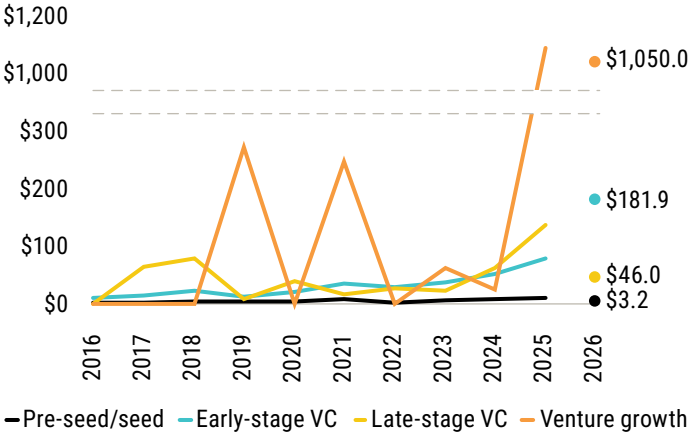
Since 2023, median deal sizes have surged from $22.2 million to $130 million in 2026, while the average was catapulted by megadeals, climbing from $34.1 million to nearly $280 million over the same period. This was largely driven by headline deal rounds in the venture-growth rung, surging from $25 million in 2024 to more than $1 billion in 2026. We attribute the valuation spike to the sense of urgency that developed around compute capacity. ChatGPT's November 2022 debut accelerated the AI arms race both inside and outside the US, driving a surge in demand for the hardware to run the software expected to catapult civilization into a new digital age.

Median inference AI VC deal value ($M)



Source: PitchBook • Geography: Global • As of August 28, 2026

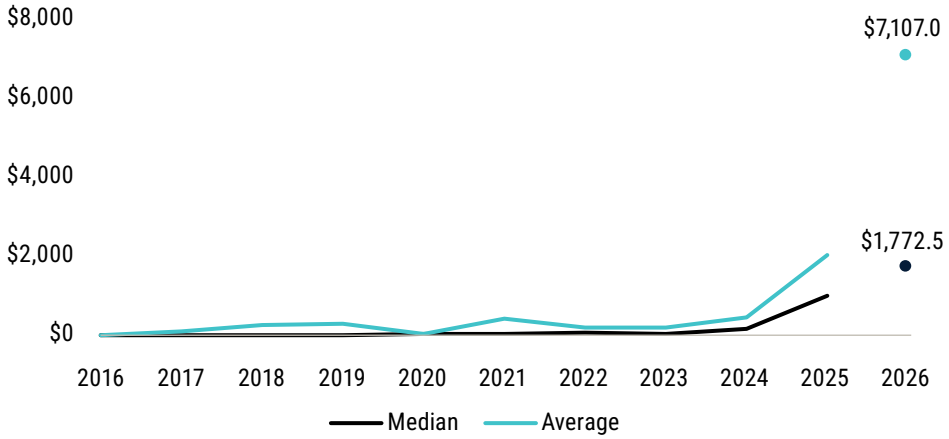
Median inference AI VC deal value ($M) by type



Source: PitchBook • Geography: Global • As of August 28, 2026

As models grew more sophisticated, particularly with the rise of agentic AI, firms optimizing for inference-heavy workloads began to capture investor attention. Median pre-money valuations exemplified this dynamic even more acutely, with startup valuations surging from $194 million in 2024 to more than $1.7 billion in 2026.

Median and average inference AI VC deal pre-money valuation ($M)



Source: PitchBook • Geography: Global • As of August 28, 2026

Early-stage VC median deal sizes have climbed from $37.4 million in 2023 to $181.9 million in 2026 YTD, a nearly 5x increase in three years. Even "early" rounds now command what used to be growth-stage checks, reflecting the capital intensity of silicon startups. Late-stage VC medians followed a similar trajectory, rising from $22 million in 2023 to $62 million in 2024 and $139.4 million in 2025, before dropping to $46 million in 2026 YTD. The pullback likely reflects a small-sample phenomenon rather than a real reversal, given that only seven late-stage deals have closed YTD.

Venture growth medians tell the most dramatic story, moving from $62.4 million in 2023 to $25 million in 2024, then vaulting to $1.1 billion in 2025 and $1.1 billion in 2026 YTD. These figures rest on just two to three deals per year, meaning the medians are essentially reporting the mega-rounds raised by Groq, Cerebras, and their peers directly. Pre-seed/seed medians have grown more modestly, from $5 million in 2023 to $8 million in 2024 and $10 million in 2025, before falling to $3.2 million in 2026 YTD.

Exits

Exit activity is quiet in volume but loud in value. The IPO of Cerebras Systems and NVIDIA's acquisition of Groq together created $34.3 billion in exit value in 2026 alone. We expect most exit activity to come from acquisitions by capital-rich hyperscalers and established chipmakers looking to strengthen their competitive positioning. Public listings will continue to drive exit value in proportion to their ability to address hardware pain points. Durability of operations and proprietary IP will be necessary moats to survive the colosseum of public markets.

Industry insights

d-Matrix

To bring an industry perspective into the analysis, PitchBook interviewed Sid Sheth from d-Matrix. The company is a US-based firm that develops inference accelerator chips with an integrated software stack, enabling efficient deployment of transformer and agentic AI workloads. Its architecture is designed to reduce latency and power requirements, positioning the company well for sovereign AI programs where inference at scale is a critical bottleneck.

When you started the company, what was the specific bet on inference that you were making, and how has that thesis held up or shifted?

When we founded d-Matrix in 2019, the entire industry was focused on training. Our bet was contrarian: that inference, not training, would become the dominant AI workload and that the hardware built for training would be fundamentally mismatched to it. Specifically, we bet that the bottleneck in inference wouldn't be raw compute, but memory—the cost, latency, and energy of moving data between memory and processor.

That thesis has held up better than we could have imagined. What we couldn't have predicted was the volume and velocity. The rise of generative AI, and now agentic AI, has pushed inference demand far beyond what anyone forecast. And we still don't have a full handle on what it might be, with every forecast continually being replaced by yet higher ones. What has shifted is the framing. In 2019 we were explaining why inference would matter. Today, no one questions that. The conversation has moved to how you serve inference economically at scale—which is exactly the problem we built the company to solve. The bet was right. The market just arrived faster and bigger than we expected.

Memory bandwidth is a key bottleneck for LLM inference. How does that constraint reshape chip design over the next generation, and what ripple effects does it have across the industry?

For a decade, chip design has been organized around compute—more FLOPS [floating-point operations per second], more cores, more density. Memory was an afterthought, bolted on through HBM [high bandwidth memory]. Inference flips that. For instance, in the decode phase of AI coding, where most inference time is spent, the chip is starved for memory bandwidth, not compute. That means the next generation of inference silicon will be designed memory-first, not compute-first.

And the ripple effects are significant. If memory is the constraint, then the winners are the companies that rethink the relationship between memory and compute—whether that's bringing compute closer to memory, or, as we've done, building a memory-centric silicon architecture that brings compute directly into memory. It also reshapes the supply chain. The industry's dependence on HBM has created a bottleneck that's now showing up in server pricing. Companies today that can reduce

> "A team that talks about replacing GPUs rather than working within the ecosystem—because the new reality includes heterogeneous workloads, and any thesis that requires every customer to rip and replace its entire infrastructure is fighting physics and economics. And any company that hasn't thought seriously about how it manufactures, deploys, and supports at datacenter scale. The graveyard of chip startups is full of great architectures that couldn't cross that chasm."
>
> SID SHETH FROM D-MATRIX

or eliminate HBM dependence can gain both a performance and a supply chain advantage. And it changes the economics of the datacenter—because memory is now a much higher portion of the bill of materials for a high-end AI rack and rising.

If you were an investor evaluating a hardware-first inference chip startup today, what would you want to see, and what would be the red flags?

What I'd want to see: silicon in production, not slides. A clear, defensible architectural thesis that isn't just "We're faster." Evidence of commercial pull—quality customer engagements, revenue trajectory, not just benchmarks. I'd also want to see a strong software story because hardware without a well-tuned supporting software stack will one day be remembered as a science project. And a supply chain strategy because the best architecture in the world doesn't matter if you don't have the right partners to help you manufacture and go to market at scale. Too many businesses have failed because leadership wants and thinks it can do everything itself and doesn't understand how extremely difficult that is, [as well as] the power of a strong ecosystem. The red flags: a company selling a benchmark instead of a business. Performance claims based only on internal assumptions with no independent validation. A team that talks about replacing GPUs rather than working within the ecosystem—because the new reality includes heterogeneous workloads, and any thesis that requires every customer to rip and replace its entire infrastructure is fighting physics and economics. And any company that hasn't thought seriously about how it manufactures, deploys, and supports at datacenter scale. The graveyard of chip startups is full of great architectures that couldn't cross that chasm.

Inference workloads have shifted quickly—longer contexts, MoE [mixture of experts], reasoning models. How does that shape what you're building, and where has the workload evolved in ways you didn't originally design for?

The shift toward reasoning models and longer contexts has been a tailwind for our architecture because all these trends are memory-bound. Longer context windows explode the size of the KV cache. Mixture-of-experts models stay memory-bound even at modest batch sizes. Reasoning models generate dramatically more tokens per request. Every one of these trends makes memory bandwidth more important, not less—which is exactly what we optimized for. Where the workload evolved in ways we didn't fully anticipate: the sheer token intensity of agentic workflows. When we started, we were thinking about serving a model efficiently. We weren't fully modeling a world where a single coding task might fire thousands of sequential inference calls, each dependent on the last. That's led us to understand the power of disaggregation—separating the phases of inference and mapping each to the hardware best suited for it. The workload got more complex than we designed for, but it got more complex in the direction of our thesis, not against it.

What are underappreciated risks and tailwinds that analysts and press often misunderstand about the global datacenter build-out?

The most underappreciated risk is the social license to build. The press covers the capital and the chips, but the real constraint emerging is community opposition.

> “Seventy-one percent of Americans oppose datacenter construction in their area.”
> SID SHETH

Seventy-one percent of Americans oppose datacenter construction in their area. Projects worth $100 billion+ have been disrupted. You can have all the capital and all the silicon in the world, but if communities won’t let you build, and grids can’t supply the power, the build-out stalls. The industry treats this as a PR problem. But, at its heart, it’s actually an engineering problem—it means efficiency isn’t optional; it’s existential. The most underappreciated tailwind is the efficiency curve. There’s a persistent narrative that AI’s energy consumption grows linearly with demand—that more AI inevitably means proportionally more power and more datacenters. That ignores the pace of architectural innovation. The industry is getting dramatically more inference per watt every year. The companies that win won’t be the ones that build the most infrastructure—they’ll be the ones that do more with less. Analysts model the demand curve carefully but underestimate the efficiency curve that runs against it.

MemryX

How do you frame the total inference opportunity, and what share of that workload do you expect to be served at the edge versus in centralized infrastructure?

We see inference as the long-term volume driver for AI. A model may be trained periodically, but it can be run millions or even billions of times across datacenters, factories, vehicles, cameras, and intelligent devices.

The balance between the edge and centralized infrastructure depends on how it is measured. Centralized infrastructure will continue to account for most aggregate compute and accelerator spending, particularly for the largest generative AI models and workloads that benefit from batching. The edge, however, will account for far more deployment points and an increasing share of the real-time decisions made by AI systems.

Where a workload runs will depend on latency, bandwidth, privacy, data sovereignty, reliability, and cost. For applications such as video analytics, industrial inspection, robotics, and transportation, it often makes little sense to send continuous streams of data to the cloud just to determine whether something important happened.

Most AI systems will ultimately be hybrid, which is why MemryX is building for both markets. MX3 supports scalable inference across edge devices and edge servers today. MX3+ will extend that capability to larger models and higher-density deployments, while MX4 is being developed for datacenter inference. The same fundamental challenge exists in both environments: Moving data between memory and compute consumes power, increases cost, and limits performance. Our architecture was designed to address that challenge efficiently, whether the power envelope is a few watts at the edge or significantly larger in the datacenter.

Which verticals are adopting edge inference fastest, which are lagging, and what explains the divergence?

We are seeing the fastest adoption in video analytics, security, industrial manufacturing, and robotics. These markets already generate large amounts of local

data, and the business case is usually clear. A manufacturer can measure the value of reducing defects or automating inspection. A security or transportation operator can quantify the benefit of monitoring more cameras without adding more people.

Transportation infrastructure, logistics, and smart-city applications are also moving quickly when the use case is well defined. We are already seeing customers move beyond evaluations into production deployments for applications such as manufacturing inspection, worker safety, and large-scale video analytics.

Automotive OEM programs and healthcare have significant long-term potential, but they tend to move more slowly because of qualification, safety, regulatory, and product development requirements. Some retail and general enterprise applications also take longer because deployments are fragmented and the return on investment is not always well defined.

The difference is generally not the level of interest in AI. Adoption moves fastest when there is a specific operational problem, usable data, a clear path to deployment, and someone in the organization who can measure the economic benefit.

Memory has become the binding constraint across inference. How does that constraint manifest differently in power- and thermally bounded edge deployments than in rack-scale systems?

The underlying issue is not just the amount of memory. It is the time, energy, and cost required to move data between memory and the compute engines.

At the edge, there is very little room to work around that problem. A system may have only a few watts available for AI, limited memory bandwidth, and little or no active cooling. It may also need to process several camera streams at the same time while sharing the host processor with the rest of the application. Adding more compute does not help if the system cannot keep that compute supplied with data.

Rack-scale systems have more options. They can use HBM, high-speed interconnects, and advanced cooling, but those solutions are expensive and consume significant power. Memory capacity and bandwidth can determine how many accelerators are required, which models can be served, and whether the overall system is economically practical.

This is a problem MemryX was designed to address. Our architecture distributes memory alongside the compute engines and processes neural networks as streaming dataflow pipelines. By keeping active model parameters close to the compute and minimizing unnecessary off-chip data movement, we can maintain high utilization without simply adding more raw compute. That is especially important in power-constrained edge systems, but the same principle becomes increasingly valuable as inference systems grow.

What is the primary friction in scaling edge AI deployment?

The biggest challenge is moving from a successful demonstration to a production deployment that can be maintained and scaled. It is relatively easy to show that a model can run on a piece of hardware. Production requires integration with the customer's existing system, consistent performance across multiple workloads, acceptable accuracy, remote management, and predictable power and thermal behavior. The platform also needs to support new models as the application evolves.

That last point is often underestimated. Models can change several times during the life of a hardware platform. If every change requires retraining, extensive quantization work, or major changes to the application, engineering costs rise quickly and deployments slow down. The companies that solve this problem will combine efficient hardware with software that fits into the customer's existing development process. At MemryX, we have focused on supporting models from the frameworks developers already use without requiring retraining, while offering a scalable product family that includes the PCIe-based Cascade 100P, M.2-based Cascade 100M, USB-based Cascade 100U, and Raspberry Pi-based Cascade 100R. Ultimately, customers are not buying benchmark results. They are buying a reliable and economical way to put AI into production.

If you were assessing a startup in your domain, what would you look for, and what are red flags that would turn you off as an investor?

I would first look for a real architectural advantage that shows up at the system level. That means performance on representative customer workloads, latency, power consumption, memory efficiency, and total system cost—not just peak TOPS [trillions of operations per second]. Software would be equally important. The hardware needs to support the models and frameworks customers already use and provide a practical path for adopting new models as they emerge.

Beyond that, I would want to understand whether the architecture can scale into additional markets, whether the roadmap is credible, and [whether] the company can build an attractive business without becoming dependent on custom engineering. The red flags would be an excessive focus on peak TOPS, benchmark results based on unrealistic batch sizes, performance claims that ignore host and memory power, or a software stack that creates significant additional work for the customer. I would also be cautious if the company's differentiation depended primarily on a future product rather than technology that can be demonstrated on representative workloads today.

Cirrascale

We also spoke with Dave Driggers from Cirrascale Cloud Services. The company is a US-based neocloud, headquartered in San Diego, that provides dedicated bare-metal GPU infrastructure and managed services for private AI workloads spanning training, fine-tuning, and inference at scale.

How are you thinking about the shift from training-dominated compute to inference-dominated compute?

With inference, the economics shift toward a total cost of ownership model focused on the cost per token rather than simply how quickly a model can be trained. This radically changes the hardware requirements. The infrastructure needs to be appropriately sized to both the model itself—where size matters—and the way the model is being used, whether that is real-time, near-time, or batch inference.

How are customer contracts evolving in terms of length and flexibility, and what does that shape tell you about where the market is heading and the broader implications—if any—for the datacenter build-out?

As workloads shift toward production inference, customer contracts are getting longer. Training can be a one-and-done workload, whereas inference is an ongoing requirement. Inference is effectively forever, which is driving customers toward longer-term capacity commitments.

Tensordyne

PitchBook spoke with co-founder Gilles Backhus.

Matrix multiplication is the dominant energy cost in AI inference, and Pareto shifts that work into additions using a logarithmic format. After the conversion overhead, what is the net energy saving?

According to co-founder Gilles Backhus, net energy saving we see on a system level (the real deal) compared to other systems of similar system performance—3x to 20x. It massively depends on how fast you run the token generation, what model, etc. I think ballpark 5x to 8x is a good range.

Most datacenters being built now are designed for liquid cooling. How does that impact the architecture of the chips and what are the implications—if any—for retrofitting?

Our current gen is built for air cooling. However—it is entirely possible to put our racks entirely into a so-called “rear door heat exchanger,” which is an enclosure that itself is liquid-cooled. Other than that, we can also redesign our system to be liquid-cooled. Not impossible. For now, we don’t see that necessity. Nobody seems to have a problem with it. Our next-gen will be entirely liquid-cooled.

If you were a VC conducting due diligence on an inference-focused chip startup, what are three or four questions you should be asking about the founder and their company?

- What does their Pareto curve (total throughput per MW or per dollar vs. speed per user) look like on a long-context (agentic) large MoE, and how does it stack up against other systems? That's the hardcore value right there. And how did they actually model or measure it?

- What exactly does integration into a multivendor datacenter look like from a SW point of view? Can I manage their system with the same SW stack (at least to 90%) that I use for my AMD and NVIDIA systems? Can I unify those systems into a well-operable heterogeneous system landscape?

- Who are their partners in manufacturing, memory, silicon, [and] system assembly—are they able to cope with a 1,000+ system order? Are they ready for scale?

- What is their GTM strategy? Direct hardware sales? Hardware lease? Joint deployment with a neocloud? Hyperscaler? Frontier lab? TaaS company? Who do they partner with on the GTM side?

What is your outlook on regional inference demand and supply in North America, Asia, and Europe? Do you see the latter catching up anytime soon?

The absolute current and future demand for Europe is huge. As it is for Asia and [North America]. Right now, all regions need more build-out. It's a rising market, and everybody needs more capacity. Because of latency and sovereignty reasons, local gigawatt AI factories will have to be built in almost any country anyway. Every country needs its own AI infrastructure, just as every country needs its own electricity, water, and food production. At least partially. It's just good government hygiene to do that. Just as Spain doesn't want to import all its electricity from, say, France, it also shouldn't import all its tokens. Of course, there will be more autarkic and less autarkic nations. Some nations will own far more of the stack than others (e.g., from chip to frontier models to datacenters to energy). In terms of relative numbers, I don't think Europe will catch up, relatively speaking, anytime soon. But I don't think that matters as much as just making sure the EU/Europe makes sure that more and more supply and infra is built at an increasing pace, and that capital is being unlocked through new [channels] (e.g., pension funds). Currently, the US [accounts for] 65% of all tokens consumed. Asia is 20%. Europe is 10%.

Hyperscalers and in-house chip designers order in bulk and often get priority at Taiwan Semiconductor Manufacturing Company (TSMC), making capacity fiercely competitive. How does Tensordyne secure its allocation against players like Apple and NVIDIA?

That's right. And that's exactly where relationships go very, very far. We've had relationships with Broadcom, Micron, [and] TSMC longer than even Tensordyne. We're leveraging those, and we're choosing our other partners carefully too (e.g., HPE on the system side right now) to make sure we can cater to customers and cover their demand.

Due diligence guide

Section 1: Silicon and technology

1. **Is the chip actually working, and how mature is the silicon?**

Rationale: There's a big difference between a chip that's been simulated, one that has come back from the fab and boots, and one that's running real customer workloads at production yield. First silicon almost always has bugs, and a respin (a corrected version of the chip) can cost tens of millions of dollars and six to 12 months.

> **Ask:** "Which version of the silicon are you on, what bugs did you find in bring up, what are your yields, and can I see it running a real model live, not a recorded demo?"

Benchmarks on simulated or hand-tuned hardware don't count. If the answer is "next quarter," treat the company as pre-product.

2. **What bets has the chip made about how AI models will look, and what happens if those bets are wrong?**

Rationale: Every specialized chip assumes things about future workloads: model size, numeric precision, attention patterns, context length, or how much memory sits on the chip versus off it. Those assumptions are what make it fast, but they also make it fragile. Models change faster than chips do, as the shifts to mixture-of-experts, very long contexts, and low-precision formats have shown.

> **Ask:** "What model trends would make your architecture less competitive, and how long does it take you to support a brand-new open model after release?"

A good answer names the risks openly and points to a fast compiler path, ideally days rather than months. A team that says its chip handles everything equally well hasn't thought this through or isn't being straight with you.

3. **How does the system perform across the full throughput-versus-speed trade-off on long-context mixture-of-experts workloads, and how did they measure it?**

Rationale: Any accelerator can look impressive at one carefully chosen point: maximum total throughput when it is serving huge batches, or blazing speed for a single user. Customers don't buy a point; they buy the overall trade-off between the two, because that trade-off determines the cost per token they can sell at a given level of responsiveness. Long-context agentic workloads on large mixture-of-experts models are where that trade-off is hardest: Memory demand grows with every token of context, the experts have to be spread across many chips, and chip-to-chip communication becomes the bottleneck. Designs that shine on small models at short context can fall apart here.

> **Ask:** "Show me your full Pareto curve and total throughput per megawatt and per dollar against speed per use on a large mixture-of-experts model at long context. Did you measure it on real silicon or model it, and exactly which competing system and software version did you compare against?"

A single hero number, or a comparison against last-generation GPUs running unoptimized software, is a red flag. Measured, reproducible results, ideally ones a customer or third party has checked, are what count.

4. **Is this a one-time bet, or does the company have a real generational roadmap?**

Rationale: A single fast chip is a product. A company that can convincingly plan its next two or three chips (with a credible timeline and improvements that build on real customer feedback rather than just bigger specs on paper) is a business.

> **Ask the founder:** "What does your next generation look like, when does it tape out, and what did you learn from this generation that's actually shaping it?"

A team that's vague on what comes after the current chip or treats the roadmap as an afterthought is signaling it may not have thought past the first product. That matters a lot in a market where NVIDIA ships a new generation every year and customers expect the same discipline from challengers.

Section 2: Operations and scale

5. **Will this hardware fit into the datacenters customers actually have?**

Rationale: Many new accelerators need very high power per rack, liquid cooling, or special networking that most existing facilities can't support without major retrofits. That can quietly add months and millions to every deployment or limit the company to a few new-build sites.

> **Ask:** "What power density and cooling does a rack need, how many of today's datacenters can take it as-is, and who pays for the facility changes?"

A chip that wins on benchmarks but needs a custom build to run in is a much smaller market than it looks on paper.

6. **Can their system run inside a multivendor datacenter on the same software stack as NVIDIA and AMD systems?**

Rationale: Operators don't run one chip; they run a fleet, and the real cost of adding a new accelerator is the operational burden it brings along. Scheduling, monitoring, model-serving frameworks, driver and firmware updates, and security tooling are already built around NVIDIA and, increasingly, AMD. If the new system needs its own separate stack, it becomes an island that requires dedicated engineers, can't share workloads with the rest of the fleet, and quietly erodes whatever cost advantage the chip offered.

> **Ask:** "Can I manage your system with at least 90% of the stack I already use for my NVIDIA and AMD systems? Which parts work unchanged (orchestration, serving framework, monitoring, model formats), and which require your proprietary tools? Can my scheduler route the same workload across your systems and my GPUs, and is a customer running that mixed setup today?"

A team that answers "just use our SDK" hasn't lived inside a real datacenter. Upstream support in the open-source frameworks customers already use and standard interfaces into existing tooling are the signs of a vendor that understands how infrastructure is actually operated.

7. **Who is the memory provider, and what happens if that supply gets tight or expensive?**

Rationale: The specialized memory these chips need (the fast kind that feeds data to the chip) is in short supply industry-wide, and pricing/availability can swing hard. Some startups design around this by avoiding that memory type altogether; others depend on it and need a real supply relationship to back that up.

> **Ask:** "Where does your memory come from, do you have that locked in, and what's your plan if it gets more expensive or scarce?"

Vague answers here are a red flag. This is a real cost and supply-chain exposure, not a nice-to-have detail.

8. **Are their manufacturing, memory, silicon, and system-assembly partners ready for a 1,000-system order?**

Rationale: Building 10 systems for a pilot and building 1,000 are completely different problems. Every system depends on a chain of partners: foundry wafers, advanced packaging, memory, boards, system assembly, rack integration, and testing. Several of those links, especially leading-edge packaging and high-bandwidth memory, are capacity-constrained across the industry. A startup is a small customer to these

suppliers, and they often push orders to the back of the line when capacity gets tight. A large order also means committing to parts many months ahead and funding inventory long before the customer pays.

> **Ask:** "Who are your foundry, packaging, memory, and system-assembly partners, how much capacity have they actually committed to you in writing, and what is the lead time and cash requirement from purchase order to delivery on a 1,000-system order?"

"We have a great relationship with TSMC" is not an answer. Named partners, signed capacity reservations, an assembly partner with a track record at volume, and a plan to finance the inventory are what separate a company that is ready for scale from one that will stall on its first big win.

Section 3: Commercial and team

9. **Is the traction real revenue or promises of future revenue?**

Rationale: Startups in this space often announce eye-catching numbers like "$1 billion in signed contracts." Often these are non-binding letters of intent or conditional pre-orders, not cash collected or even contractually guaranteed.

> **Ask:** "Of the deals you're citing, how much is binding, how much has actually been paid, and what would it take for a customer to walk away?"

This is one of the easiest places for a compelling story to outrun the underlying reality.

10. **What is their go-to-market strategy, and who are their partners in it?**

Rationale: How a chip reaches customers shapes almost everything else about the business: revenue timing, margins, how much capital it consumes, and who carries the risk. Selling hardware outright brings in cash up front but requires a real sales and support organization and asks customers to own unfamiliar equipment. Leasing or selling tokens as a service reduces customer risk but puts the hardware on the startup's balance sheet, turning it into a capital-hungry cloud business. Partnering with a neocloud provides reach but shares margin and ties the company's fate to that partner, while hyperscalers and frontier labs offer huge volume but concentrate risk in a few buyers who often have their own in-house chips and plenty of negotiating leverage.

> **Ask:** "Is your model direct hardware sales, leasing, joint deployment with a neocloud, a hyperscaler or frontier-lab deal, or tokens-as-a-service? Who are your first 10 customers, how exactly do they pay you, who owns the hardware, and which partner gets you in the door?"

An answer of “all of the above” usually means no channel has actually been proven. A focused primary channel, matched to the company’s funding model and backed by named partners, is a far stronger signal than a broad menu of options.

11. **Does the company have a history of selling and supporting hardware, not just designing it?**

Rationale: A lot of founders in this space come from brilliant chip-design backgrounds. That’s necessary, but it’s not the whole job. Winning a customer means more than a faster chip. It means a software stack that doesn’t force customers to rebuild everything, real support when something breaks, and a shipping cadence customers can plan around.

> **Ask:** “Has this team shipped and supported real hardware in someone else’s datacenter before?”

A team that only talks architecture and never talks about the customer’s side of the problem requires more digging.

Appendix

Top inference AI VC-backed companies by capital efficiency

Company	VC raised to date ($M)	Post-money valuation ($M)	Capital efficiency	Segment	Country	Success probability	IPO probability	M&A probability
Etched	$1,710.4	$21,000.0	12.3x	Datacenter inference ASICs	US	98%	77%	21%
Element Labs	$128.1	$1,092.9	8.5x	Datacenter inference ASICs	Israel	78%	13%	65%
Luminous Computing	$115.0	$701.4	6.1x	Photonic/optical compute & interconnect	US	38%	8%	30%
Expedera	$39.4	$215.4	5.5x	Semiconductor IP licensing	US	77%	4%	73%
RaiderChip	$1.1	$5.4	5.0x	Semiconductor IP licensing	Spain	0%	0%	0%
d-Matrix	$446.3	$2,000.0	4.5x	Compute-in-memory/ analog/novel architecture	US	98%	90%	8%
Mentium Technologies	$7.7	$33.3	4.3x	Compute-in-memory/ analog/novel architecture	US	55%	1%	54%
Fractile	$321.2	$1,348.8	4.2x	Compute-in-memory/ analog/novel architecture	UK	92%	15%	77%
SuperMEM	$1.8	$7.5	4.2x	Edge inference silicon	US	23%	1%	22%
Quadric	$84.3	$300.0	3.6x	Semiconductor IP licensing	US	96%	11%	85%
Myrtle.ai	$12.0	$42.5	3.5x	Semiconductor IP licensing	UK	15%	1%	14%
Positron AI	$314.0	$1,064.9	3.4x	Datacenter inference ASICs	US	87%	32%	56%
Aril	$9.5	$29.5	3.1x	Semiconductor IP licensing	US	5%	1%	4%
Furiosa	$260.7	$730.4	2.8x	Datacenter inference ASICs	South Korea	98%	72%	26%
SiMa.ai	$355.3	$960.0	2.7x	Edge inference silicon	US	93%	30%	63%
CogniFiber	$13.5	$35.0	2.6x	Photonic/optical compute & interconnect	Israel	65%	2%	62%
Sunrise	$139.4	$355.7	2.6x	Edge inference silicon	China	0%	0%	0%
Tensordyne	$208.9	$510.0	2.4x	Datacenter inference ASICs	US	90%	51%	39%

Source: PitchBook • Geography: Global • As of September 30, 2026

Note: Probability data is based on PitchBook VC Exit Predictor methodology. Model does not include any China-based companies.

Top inference AI VC-backed companies by capital efficiency (continued)

Company	VC raised to date ($M)	Post-money valuation ($M)	Capital efficiency	Segment	Country	Success probability	IPO probability	M&A probability
MemryX	$63.0	$136.0	2.2x	Compute-in-memory/ analog/novel architecture	US	83%	8%	74%
Poco-Apoco Networks	$0.8	$1.7	2.2x	Semiconductor IP licensing	Japan	0%	0%	0%
Sagence AI	$50.5	$97.3	1.9x	Compute-in-memory/ analog/novel architecture	US	53%	2%	50%
Kneron	$521.0	$1,000.0	1.9x	Edge inference silicon	US	96%	83%	14%
NeuReality	$114.6	$219.5	1.9x	Datacenter inference ASICs	Israel	68%	9%	59%
Rain AI	$66.9	$126.0	1.9x	Compute-in-memory/ analog/novel architecture	US	80%	40%	40%
EdgeCortix	$37.0	$62.0	1.7x	Edge inference silicon	Japan	91%	56%	35%

Source: PitchBook • Geography: Global • As of September 30, 2026
Note: Probability data is based on PitchBook VC Exit Predictor methodology. Model does not include any China-based companies.

Top inference AI VC-backed companies by IPO probability, only companies predicted successful IPO

Company	VC raised to date ($M)	Post-money valuation ($M)	Capital efficiency	Segment	Country	Success probability	IPO probability	M&A probability
d-Matrix	$446.3	$2,000.0	4.5x	Compute-in-memory/ analog/novel architecture	US	98%	90%	8%
Kneron	$521.0	$1,000.0	1.9x	Edge inference silicon	US	96%	83%	14%
Etched	$1,710.4	$21,000.0	12.3x	Datacenter inference ASICs	US	98%	77%	21%
Furiosa	$260.7	$730.4	2.8x	Datacenter inference ASICs	South Korea	98%	72%	26%
EdgeCortix	$37.0	$62.0	1.7x	Edge inference silicon	Japan	91%	56%	35%
Tensordyne	$208.9	$510.0	2.4x	Datacenter inference ASICs	US	90%	51%	39%

Source: PitchBook • Geography: Global • As of September 30, 2026
Note: Probability data is based on PitchBook VC Exit Predictor methodology. Model does not include any China-based companies.

Q4 2026 inference silicon market map

Compute-in-memory / analog / novel architecture

d-Matrix · FRACTILE · memryX AI · analogAI · Qernel AI — v1.0.0

Datacenter inference ASICs

POSITRON AI · groq · TENSORDYNE · FURIOSA

Edge inference silicon

Kneron · SiMa.ai · Acrab · AXELERA · UNTETHER AI

Photonic/optical compute & interconnect

LUMINOUS · arago · COGNIFIBER Pure Photonic Computing · LiteEdge.ai · Topolog IQ

Semiconductor IP licensing

quadric · expedera · myrtle.ai · Aril Inc. · RaiderChip Accelerating Artificial Intelligence

Source: PitchBook • Geography: Global • As of August 28, 2026

Note: Habana was historically training-focused prior to Intel's 2019 acquisition; the current Gaudi generation is positioned for both training and inference. It is included here as a notable historical entrant.

AI and cloud infrastructure M&A and VC investments since 2021

Deal date	Acquirer/ investor	Portfolio company	Category	Deal value ($M)	Portfolio company description
July 29, 2026	Alphabet	Eliyan	VC	$145	Chip interconnect/chiplet IP for AI infrastructure
July 21, 2026	Microsoft	Aligned Data Centers	M&A	$40,000	Hyperscale datacenter operator
July 14, 2026	Alibaba	AIsphere Technology	VC	$499	Video generation/world foundation models
July 2, 2026	Tencent	Kling AI	VC	$3,000	AI video generation platform
June 24, 2026	SpaceX	AI Insiders	M&A	N/A	AI intelligence/capability integration
June 16, 2026	Nebius	Eigen AI	M&A	$643	AI company
June 16, 2026	Tencent	DeepSeek	VC	$7,522	AI foundation model lab
June 16, 2026	SpaceX	Cursor	M&A	$60,000	AI-powered code editor/developer tool
June 10, 2026	Amazon	NEURA Robotics	VC	$1,500	AI/robotics
May 28, 2026	Alphabet	Anthropic	VC	$70,950	AI foundation model lab (also backed by Amazon)
May 28, 2026	Amazon	Anthropic	VC	$73,000	AI foundation model lab (also backed by Alphabet)

Source: PitchBook • Geography: Global • As of September 24, 2026

Note: Table includes cloud-, AI-, chip-, and datacenter-related deals from 2021 onward across all nine hyperscaler and neocloud companies. Companies with earlier and later rounds from the same investor show total capital across all rounds (including any pre-2021 rounds) to keep the total accurate, with the date reflecting the most recent round. The $65 billion Anthropic round (from May 28, 2026) appears under both Alphabet and Amazon as a shared syndicated round.

AI and cloud infrastructure M&A and VC investments since 2021 (continued)

Deal date	Acquirer/ investor	Portfolio company	Category	Deal value ($M)	Portfolio company description
May 27, 2026	CoreWeave	Tensormesh	VC	$20	AI infrastructure tooling
May 25, 2026	Tencent	TianJi Intelligence	VC	$147	AI hardware/robotics scale-up
May 18, 2026	Alphabet	AI cloud company	Other	N/A	Cloud infrastructure company (Alphabet/Blackstone joint venture)
May 8, 2026	CoreWeave	Buf	M&A	N/A	API/protocol tooling for developers
March 24, 2026	Amazon	Fauna Robotics	M&A	N/A	Robotics
March 12, 2026	CoreWeave	Standard Kernel	VC	$20	AI infrastructure/systems
March 11, 2026	Alphabet	Wiz	M&A	$32,000	Cloud security platform (2024 agreed-upon deal at $23B did not close; final 2026 acquisition closed at $32B)
March 4, 2026	Alibaba	VAST Technology	VC	$50	AI 3D/world models
February 19, 2026	Alphabet	MonetizationOS	Other	$6	Ad/monetization infrastructure platform
February 10, 2026	Alibaba	Corise Semiconductor	VC	N/A	Chip company
February 10, 2026	Nebius	Tavily	M&A	$275	AI search/retrieval API
February 3, 2026	SpaceX	xAI	M&A	$250,000	AI foundation model lab (formerly Elon Musk owned; folded into SpaceX)
January 20, 2026	Tencent	Beijing Moonshot Technology (Kimi AI)	VC	$1,500	AI foundation model lab
January 9, 2026	Alibaba	MiniMax AI	Other	$619	AI infrastructure/LLM (also backed by Tencent)
November 5, 2025	CoreWeave	Monolith AI	M&A	N/A	AI platform for industrial/manufacturing
November 5, 2025	SpaceX	EchoStar (AWS-3 licenses, Englewood, CO)	Other	$2,600	Wireless spectrum license purchase (network/ connectivity infrastructure)
October 30, 2025	CoreWeave	Marimo	M&A	N/A	Developer tooling for AI app deployment
October 21, 2025	Nebius	Avride	Other	$375	Autonomous vehicle/robotics (Nebius co-investor)
October 7, 2025	Nebius	ClickHouse	VC	$650	Cloud-native database/analytics platform
September 22, 2025	CoreWeave	Salt AI	VC	$10	AI engineering for biopharma/healthcare
September 7, 2025	SpaceX	EchoStar (AWS-4 and H-Block spectrum licenses)	Other	$17,000	Wireless spectrum license purchase (network/ connectivity infrastructure)
September 3, 2025	CoreWeave	OpenPipe	M&A	N/A	AI model fine-tuning/optimization
July 2, 2025	CoreWeave	Chai Research	VC	$53	Conversational AI
May 27, 2025	SpaceX	Akoustis Technologies	M&A	$76	Radiofrequency filters/wireless communications components
May 16, 2025	CoreWeave	Moonvalley AI	VC	$53	AI video generation

Source: PitchBook • Geography: Global • As of September 24, 2026

Note: Table includes cloud-, AI-, chip-, and datacenter-related deals from 2021 onward across all nine hyperscaler and neocloud companies. Companies with earlier and later rounds from the same investor show total capital across all rounds (including any pre-2021 rounds) to keep the total accurate, with the date reflecting the most recent round. The $65 billion Anthropic round (from May 28, 2026) appears under both Alphabet and Amazon as a shared syndicated round.

AI and cloud infrastructure M&A and VC investments since 2021 (continued)

Deal date	Acquirer/ investor	Portfolio company	Category	Deal value ($M)	Portfolio company description
May 11, 2025	Alphabet	AI21	VC	$300	LLM/enterprise AI
May 7, 2025	Nebius	Toloka AI	VC	$72	AI data labeling/annotation platform (Nebius originated)
May 5, 2025	CoreWeave	Weights & Biases	M&A	$1,700	AI & ML model management and monitoring
January 28, 2025	Microsoft	VEIR	VC	$75	Power transmission tech for datacenters
December 30, 2024	Tencent	Enflame Technology	VC	$1,291	AI chip designer
October 2, 2024	Amazon	Perceive	M&A	$80	Edge AI/custom silicon
September 30, 2024	Microsoft	OpenCyber AI	VC	N/A	AI venture funding platform
September 4, 2024	CoreWeave	W.AI	VC	$9	AI-powered content creation
September 1, 2024	Tencent	Manus AI	VC	$10	AI agent platform
August 16, 2024	Microsoft	Armada	VC	$40	Modular AI datacenters
June 28, 2024	Alibaba	OceanBase	VC	N/A	AI database services
May 14, 2024	Amazon	Alchip Technologies	VC	$16	ASIC/custom chip design
March 31, 2024	Microsoft	Group 42 Holding	Other	$1,500	AI/cloud partnership and infrastructure investment
March 15, 2024	Alibaba	MiniMax AI	VC	$654	AI infrastructure/LLM (also backed by Tencent)
February 29, 2024	Amazon	Figure	VC	$675	Humanoid robotics
February 26, 2024	Microsoft	Mistral AI	VC	$431	AI foundation model
October 17, 2023	Tencent	Baichuan Group	VC	$300	AI foundation model lab
September 6, 2023	Microsoft	d-Matrix	VC	$154	Datacenter inference chips for hyperscale customers
August 18, 2023	Microsoft	Arm	VC	N/A	Chip IP company
May 1, 2023	Tencent	Light Years Beyond	VC	$230	LLM company (later acquired by Meituan)
March 31, 2023	Tencent	Horizon Quantum Holdings	VC	$18	Quantum computing
January 9, 2023	Tencent	MiniMax AI	VC	$257	AI infrastructure/LLM (also backed by Alibaba)
January 1, 2023	CoreWeave	Conductor Technologies	M&A	N/A	Cloud rendering for visual effects/animation
December 9, 2022	Microsoft	Lumenisity	M&A	N/A	Hollow-core fiber optics
November 2, 2022	Tencent	JV (China United Network Comm./Tencent)	Other	N/A	5G lab focused on internet datacenters, content delivery networks, and edge computing
October 1, 2022	Alphabet	InfinyOn	VC	$7	Cloud-native data streaming infrastructure
September 12, 2022	Alphabet	Mandiant	M&A	$6,101	Cybersecurity platform
February 11, 2022	Alibaba	CXMT Corporation	VC	$2,927	DRAM/memory chip manufacturer

Source: PitchBook • Geography: Global • As of September 24, 2026

Note: Table includes cloud-, AI-, chip-, and datacenter-related deals from 2021 onward across all nine hyperscaler and neocloud companies. Companies with earlier and later rounds from the same investor show total capital across all rounds (including any pre-2021 rounds) to keep the total accurate, with the date reflecting the most recent round. The $65 billion Anthropic round (from May 28, 2026) appears under both Alphabet and Amazon as a shared syndicated round.

AI and cloud infrastructure M&A and VC investments since 2021 (continued)

Deal date	Acquirer/ investor	Portfolio company	Category	Deal value ($M)	Portfolio company description
February 2, 2022	CoreWeave	GooseAI	Other	N/A	AI inference as a service (CoreWeave/NovelAI joint venture)
December 20, 2021	Alibaba	Vastai Technologies	VC	$251	Cloud GPU/edge computing AI chips
November 3, 2021	SpaceX	Swarm	M&A	$524	Satellite Internet of Things connectivity network
October 6, 2021	Alphabet	Trusted cloud offering	Other	N/A	Sovereign cloud offering (Alphabet/Thales joint venture)
September 15, 2021	Tencent	Beijing Momenta Technology	VC	$1,200	Autonomous driving AI
July 21, 2021	Microsoft	CloudKnox	M&A	N/A	Cloud security posture management
June 30, 2021	Microsoft	AT&T	M&A	N/A	Telecom cloud infrastructure unit
June 1, 2021	Microsoft	ReFirm Labs	M&A	N/A	Firmware/chip-to-cloud security
March 16, 2021	Microsoft	The Marsden Group	M&A	N/A	Cloud/edge/AI consulting

Source: PitchBook • Geography: Global • As of September 24, 2026

Note: Table includes cloud-, AI-, chip-, and datacenter-related deals from 2021 onward across all nine hyperscaler and neocloud companies. Companies with earlier and later rounds from the same investor show total capital across all rounds (including any pre-2021 rounds) to keep the total accurate, with the date reflecting the most recent round. The $65 billion Anthropic round (from May 28, 2026) appears under both Alphabet and Amazon as a shared syndicated round.

Land investments (2021-present)

Deal Date	Hyperscaler	Site / property	Deal value ($M)	Deal count	Details
May 26, 2026	Amazon	Amazon Land (Mumbai, India)	$13	1	Data center land purchase
February 27, 2026	Amazon	Virginia Science and Technology Campus	$427	1	Land/campus acquisition to expand AI infra
February 26, 2026	Alphabet	Alphabet Land (Vaala, Finland)	N/A	1	900-hectare data center land purchase
September 10, 2025	Microsoft	Microsoft Land (Middenmeer, NL)	N/A	1	50-hectare data center land purchase
June 20, 2025	Alphabet	Alphabet Land (Botetourt County, VA)	$14	1	312-acre data center land purchase
February 7, 2025	Microsoft	Microsoft Land (Pulai, Malaysia)	$27	1	Data center land purchase
November 1, 2024	Microsoft	Microsoft Land (Pataskala, OH)	$11	1	Data center land purchase
September 13, 2024	Microsoft	Microsoft Land (Pune, India)	$54	1	Data center land purchase
August 13, 2024	Microsoft	Microsoft Land (Mount Pleasant, SC)	$13	1	Data center land purchase
February 29, 2024	Microsoft	Microsoft Land (Palmetto)	$6	1	Data center land purchase
June 15, 2023	Microsoft	Microsoft Land (Vihti, Finland)	$13	1	Data center land purchase
November 23, 2022	Amazon	State Street Data Center (Grafton, MA)	$15	1	Data center property purchase
October 21, 2022	Microsoft	Microsoft Land (Pimpri-Chinchwad, India)	$40	1	Data center land purchase
March 21, 2022	Microsoft	Microsoft Land (Newport, UK)	N/A	1	Data center land/building purchase
March 8, 2022	Microsoft	Microsoft Land (Hyderabad, India)	$37	1	Data center land purchase (3 parcels)
January 26, 2021	Alphabet	Heidelberg Materials Property (Germany)	N/A	1	Data center land purchase

Source: PitchBook • Geography: Global • As of September 24, 2026

References

1: "Data Center Capex Surges 57 Percent in 2025 as AI Deployments Accelerate, According to Dell'Oro Group," March 17, 2026.

2: Ibid.

3: AI Boom Drives Data Center Capex to $1.7 Trillion by 2030, According to Dell'Oro Group," Dell'Oro Groupj, February 11, 2026.

4: "Building the Backbone of AI," Brookfield, August 2025.

5: "The Future of AI Workloads," McKinsey & Company, February 24, 2026.

6: "Agentic AI Market Trends 2025 – 2026: Adoption Rates, and What Lies Ahead," Svitla, April 27, 2026.

7: "The Bill Arrives: How to Manage Agentic AI Costs at Scale," Cockroach Labs, Quentin Packard, June 10, 2026.

8: "AI Agents Forecast to Boost Tech Cash Flow as Usage Soars," Goldman Sachs, May 20, 2026.

9: Ibid.

10: "How Big Is the Agentic AI Market and What Is the Adoption Reality In 2027?" Pulse, June 14, 2026.

11: "Agentic AI Market," Marketsandmarkets, August 2026.

12: "Foreign-Produced Direct Product Rule Additions, and Refinements to Controls for Advanced Computing and Semiconductor Manufacturing Items, Federal Register, December 5, 2024.

13: "BIS Updated Public Information Page on Export Controls Imposed on Advanced Computing and Semiconductor Manufacturing Items to the People's Republic of China (PRC)," Bureau of Industry and Security, October 28, 2022.

14: "Commerce Strengthens Export Controls to Restrict China's Capability to Produce Advanced Semiconductors for Military Applications," Bureau of Industry and Security, December 2, 2024.

15: "Addressing United States Investments in Certain National Security Technologies and Products in Countries of Concern," Federal Register, August 11, 2023.

16: "Treasury Issues Regulations to Implement Executive Order Addressing US Investments in Certain National Security Technologies and Products in Countries of Concern," US Department of the Treasury, October 28, 2024.

17: "Sequoia Capital to Split Into Three Parts, Separating Its China and India Businesses, CNBC, Rohan Goswami, June 6, 2023.

18: "Another Top Silicon Valley Investor Is Splitting Off Its China Business as Pressure Mounts," CNN Business, Michelle Toh, September 22, 2023.

19: "DARPA Rolls Out Electronics Resurgence Initiative," DARPA, September 13, 2017.

20: "Statement on the President's Decision Regarding Lattice Semiconductor Corporation," US Department of the Treasury, September 13, 2017.

21: "Newport Wafer Fab," Hansard, November 21, 2022.

22: "Form 20-F: DiDi Global Inc.," US Security and Exchange Commission, 2023, accessed October 12, 2026.

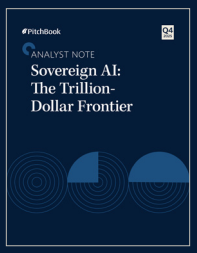
23: "DiDi Announces Results of Extraordinary General Meeting," DiDi, May 23, 2022.

24: "Form 10-Q: NVIDIA Corporation, US Securities and Exchange Commission, May 23, 2025.

25: Data-point counts for this segment are low and therefore steepen the rate of change in valuations.

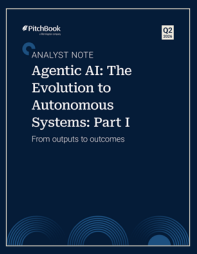
PitchBook provides actionable insights across the global capital markets.

Additional research:



Sovereign AI: The Trillion-Dollar Frontier

Download the report here



Agentic AI: The Evolution to Autonomous Systems: Part I

Download the report here

PitchBook Insights is an online compendium of in-depth data, news, analysis, and perspectives that shape the private capital markets.

PitchBook subscribers enjoy exclusive access to a comprehensive suite of private market insights, including proprietary research, news, data, tools, and more on the PitchBook Platform.

COPYRIGHT © 2026 by PitchBook Data, Inc. All rights reserved. No part of this publication may be reproduced in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, and information storage and retrieval systems—without the express written permission of PitchBook Data, Inc. Contents are based on information from sources believed to be reliable, but accuracy and completeness cannot be guaranteed. Nothing herein should be construed as any past, current or future recommendation to buy or sell any security or an offer to sell, or a solicitation of an offer to buy any security. This material does not purport to contain all of the information that a prospective investor may wish to consider and is not to be relied upon as such or used in substitution for the exercise of independent judgment.

Nizar Tarhuni
Executive Vice President of Research and Market Intelligence

Paul Condra
Senior Director, Global Head of Private Markets Research

James Ulan
Director, Industry & Technology Research

Report created by:

Dimitri Zabelin
Senior Research Analyst, AI and Cybersecurity

Oscar Allaway
Senior Data Analyst

Josie Doan
Graphic Designer

Learn more about PitchBook's Institutional Research team.

Click here for PitchBook's report methodologies.