PitchBook

PitchBook Data, Inc.

John Gabbert Founder, CEO

Nizar Tarhuni Senior Director, Institutional Research & Editorial

Paul Condra Head of Emerging Technology Research

Institutional Research Group

Analysis

Brendan Burke
Senior Analyst, Emerging Technology
brendan.burke@pitchbook.com

Data

Matthew Nacionales
Data Analyst

pbinstitutionalresearch@pitchbook.com

Publishing

Designed by **Chloe Ladwig**

Published on November 10, 2022

Contents

EMERGING TECH RESEARCH

Inferring the Future of AI Chips

Use cases and architectures driving neural network processor adoption

PitchBook is a Morningstar company providing the most comprehensive, most accurate, and hard-to-find data for professionals doing business in the private markets.

Key takeaways

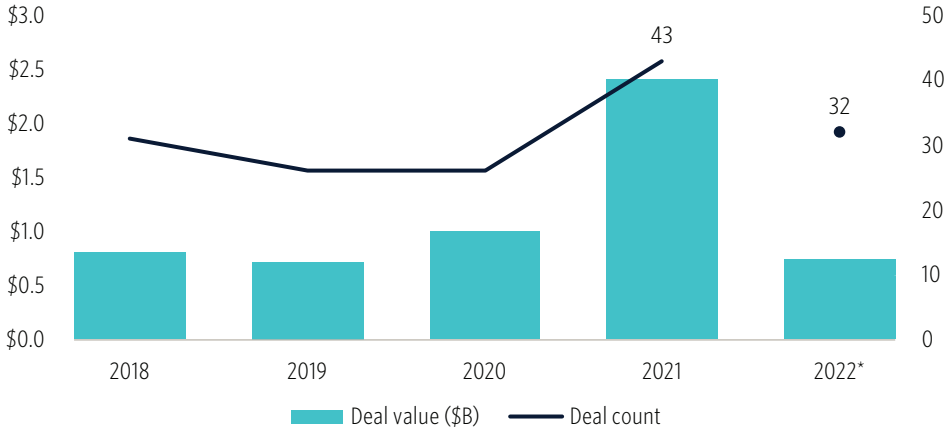
- A 69.0% decline in year-over-year VC funding for artificial intelligence & machine learning semiconductor startups outside of China may encourage M&A for some startups that align with the product needs of incumbents.
- Both the PC and automotive AI semiconductor markets are growing faster than the datacenter AI semiconductor market at over 30% each, suggesting they will surpass datacenter's market size by 2025.
- Amid the financial downturn of 2022, inference-focused VC funding has exceeded the funding of all training-focused companies, breaking a historical trend.

Overview

Artificial intelligence (AI) chip deal value fluctuates with market cycles to a greater degree than the rest of technology. Shifting market conditions have led to both rapidly declining valuations for public companies and rapidly declining deal values for private companies. 2021 was a boom year for both public and private companies, with 16 companies raising the VC mega-deals necessary to commercialize novel chip designs. 2022 has inhibited continued growth with economic decline, a whiplash effect from demand spikes in 2021, and the fracturing of US-China cooperation on chip fabrication. As a result of these geopolitical tensions and the recent policy change from US export controls, we view the AI semiconductor market as bifurcating into companies in China and those outside of China. Previously, China-based companies could make claims to global expansion, but the recent regulations have prevented China's collaboration with the global community. AI computing remains a major growth driver for the semiconductor industry, and at $43.6 billion in 2022, the market remains large enough to support large private companies.

Excluding China, VC deal value for AI startups has declined 69.0% year-over-year through October 2022. While some companies raised sufficient cash to achieve product development milestones in 2021, they may require commercial traction to raise successive large rounds. Coupled with robust competition from industry leaders, this may cause some companies to face pressure to achieve exits earlier than planned. Outside of China, AI chip startups have achieved outstanding outcomes only via M&A, as evidenced by Annapurna Labs' $370.0 million exit to Amazon and Habana Labs' $1.7 billion exit to Intel. Limited VC funding may encourage M&A for some startups that align with the product needs of incumbents.

AI & ML semiconductor VC deal activity



Source: PitchBook | Geography: Global (excluding China)
*As of October 31, 2022

Market segments driving adoption

While highly funded startups focus on enhancing the efficiency of cloud-native model training, other segments of the market are growing more quickly with the demand for AI chips. The datacenter market has historically controlled the lion's share of AI semiconductors, but edge-based use cases are on pace to surpass the category in end-user spending by 2025. We estimate the market for datacenter AI chips will reach $8.2 billion in 2022, followed by PC and automotive AI chips at $6.4 billion and $6.8 billion, respectively. Growing more than 30% in 2022, both the PC and automotive markets are growing faster than datacenter servers, suggesting they will surpass the datacenter market by 2025. Together, these three segments contribute 90.7% of the AI chip market outside of smartphones and smartwatches as of 2022. We exclude the contributions of smartphones and smartwatches due to the dominance of Apple and Samsung in those categories. The size of these categories suggests that innovation can yield commercial traction in each.

AI semiconductor (excluding smartphones and smartwatches) market size estimate ($B)*

Source: PitchBook | Geography: Global
*As of October 31, 2022

This trend suggests that the datacenter market is becoming saturated for third-party vendors. Six vendors earned 98.9% of the $10.0 billion datacenter chip market in 2021, including new entrant AWS and recently acquired Xilinx.[1] The cohort of vendors outside of these six grew more slowly than Nvidia, AMD, and AWS, demonstrating the barriers to entry in this market. Intel lost market share in 2021, leaving room for disruption, though scaled competitors such as Nvidia benefit primarily from this weakness due to the superior performance of their new chips, including its DGX systems and the upcoming Hopper GPU. Furthermore, Microsoft and Google have competitive architectures for their internal offerings, limiting the market size overall.

1: "Worldwide Datacenter Processing Semiconductor Market Shares, 1Q22: Datacenter Semiconductor Vendors Evolving Toward Solutions-Based End-Market-Driven Demand Models," IDC, Shane Rau, August 2022.

Automotive and edge computing demands are driving more commercial agreements for inference-focused chips than for cloud training chips. In Q2 2022, edge AI chip startup Hailo announced a partnership with leading automotive chipmaker Renesas for self-driving applications. Renesas is on pace to generate $3.0 billion in automotive revenue in 2022 and was only the third-largest automotive chipmaker in the world in 2021, demonstrating the scale of this market.[2,3] The company plans to package Hailo neural network processors with its advanced driver assistance system chips to facilitate autonomous driving features. Startups focusing on datacenter training remain reliant on contracts with research laboratories for specific tasks. Multiple advanced datacenter startups list Argonne National Laboratory as a leading client, despite the federal entity's low budget for procurement each year relative to hyperscalers. We believe these research-based contracts continue to take the place of significant revenue-generating contracts with hyperscalers, thus limiting revenue growth for startups. Pilot projects completed with hyperscalers in 2021 have not evolved into publicly disclosed commercial agreements, leaving startups behind on their commercial milestones for future fundraises while the datacenter AI market decelerates from its high-growth period.

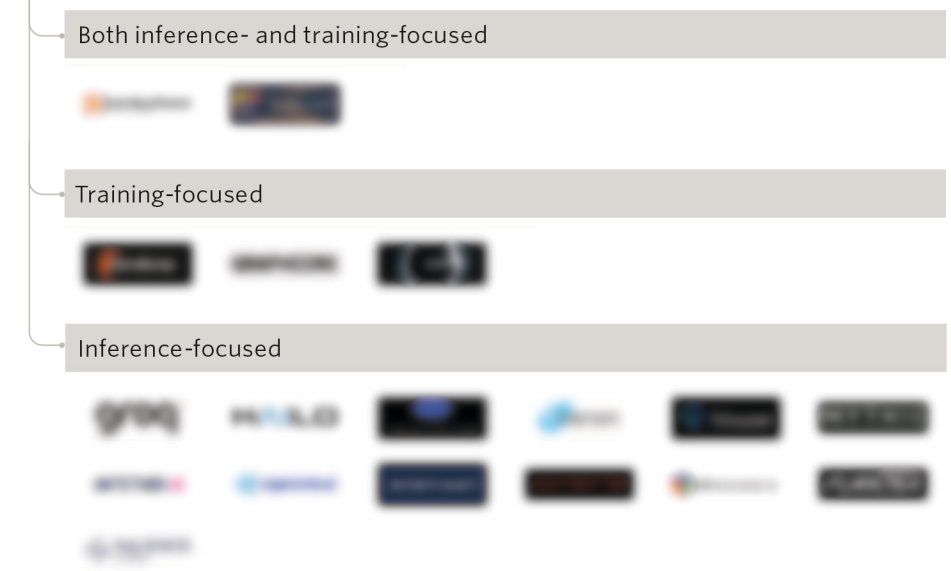
Architectures driving adoption

Inference-based architectures are starting to take precedence in VC deals. AI chips address two critical phases of the AI lifecycle: the training of new models based on existing data and the inference of those models on live data in production. Training-focused startups establish benchmarks based on the time required to train and the efficiency of common benchmark models. For example, Cerebras claims to train the language transformer model BERT-large 15x faster than Nvidia's DGX A100 GPU cluster, and Graphcore claims to improve the efficiency of training convolutional neural network EfficientNet-B4.[4,5] These startups compete primarily with Nvidia GPUs, given their state-of-the-art training results. Inference startups, in comparison, stand out for their speed and efficiency of inference on standard models. These startups test themselves on a wider array of models at varying levels of complexity and benchmark against various chipmakers, including Nvidia, for deep neural nets, competing primarily on energy efficiency for comparable performance. Startups also focus on specific use cases, including speech detection, object detection, and facial recognition, creating competitive advantages in specific areas. As a result, we are tracking 43 inference-focused companies and only 24 companies with a training focus.

2: "3Q 2022 Presentation," Renesas Electronics Corporation, October 26, 2022.
3: "Worldwide Automotive Semiconductor Market Shares, 2021: Established Vendors Continue Dominance," IDC, Nina Turner, July 27, 2022.
4: "Train Large BERT Models Faster with Cerebras Systems," Cerebras Systems, May 2021.
5: "The WoW Factor: Graphcore Systems Get Huge Power and Efficiency Boost," Graphcore, Nigel Toon and Simon Knowles, March 3, 2022.

The full market map is available in the AI & ML Analyst Workspace.

AI & ML semiconductor VC ecosystem market map

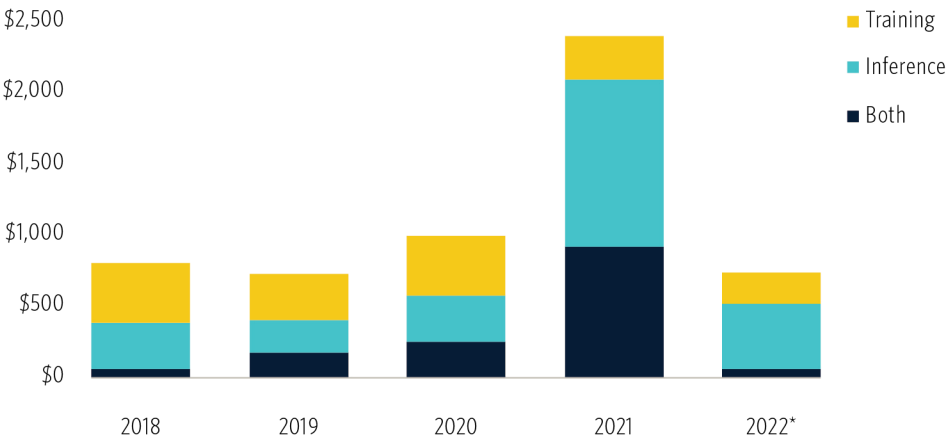


Source: PitchBook | Geography: Global (excluding China)

For further analysis of end user AI infrastructure priorities relating to foundation model training, see our analyst note Ai4 Conference Focuses on Rightsizing AI Infrastructure.

Companies that address AI training times, including those that address both training and inference, have historically led the market in VC funding. In each of the previous four years, the combined funding totals of training-focused and dual-focused startups have exceeded those of inference-only startups. Amid the financial downturn of 2022, inference-only funding has exceeded the funding of all training-focused companies. There is a cyclical component to this pattern, given the limited need for training companies to raise funding each year, yet we observe that inference-focused companies are achieving significant commercial partnerships during the economic recession. A prominent example of this is Kneron, which raised a $48.0 million Series B in October 2022 on the back of numerous contracts

AI & ML semiconductor VC deal value ($M) by subtask



Source: PitchBook | Geography: Global (excluding China)
*As of October 31, 2022

for smart city and transportation applications. Leading datacenter startups claim outstanding results in training for foundation models, including BERT and GPT, yet can face reduced demand from increased spending discipline during a recession.

Incumbent M&A priorities

Without benefiting from significant revenue growth, startups will require M&A to achieve positive VC exits. Incumbent appetite for M&A is limited due to the success of internal innovation and antitrust scrutiny. Nvidia's bid for Arm was denied by regulators and has resulted in outstanding outcomes in datacenter innovation. AMD made a horizontal purchase for Xilinx that addresses field-programmable gate array applications in the datacenter, automotive, and industrial settings. Intel has launched a competitive AI processor line, including Gaudi for training and Greco for inference, via its Habana Labs acquisition, and it is spinning out its primary automotive business in Mobileye after acquiring the company for $15.3 billion. All leading vendors are relatively underexposed to the automotive and industrial themes despite claims to have high addressable markets in both.

Leading chipmaker performance in key AI categories*

	Data center		Automotive		PC		Industrial	
Chipmaker	Current FY revenue run rate ($B)	Most recent FQ growth rate	Current FY revenue run rate ($B)	Most recent FQ growth rate	Current FY revenue run rate ($B)	Most recent FQ growth rate	Current FY revenue run rate ($B)	Most recent FQ growth rate
Nvidia	$15.1	42.4%	$0.6	5.6%	$2.2	6.5%	N/A	N/A
AMD	$6.0	83.0%	N/A	N/A	$8.8	25.0%	$5.2	28.0%
Intel	$16.8	-27.0%	$1.8	38.0%	$32.4	-17.0%	$9.2	14.0%
Amazon	$5.0	N/a	N/A	N/A	N/A	N/A	N/A	N/A

Source: Company disclosures | Geography: Global
*As of October 31, 2022

We believe that automotive may be a key focus area for chipmakers going forward. Nvidia's automotive results lag behind its ambitions in the category, which are focused primarily on the autonomous driving opportunity. Nvidia previously made an automotive software acquisition for high-definition mapping startup DeepMap and may continue to build its platform via M&A. AMD lacks scale in the category even after its Xilinx acquisition and may make a standalone acquisition in the category to foster growth. Intel retains control of Mobileye after its spinout, yet may make further investment in the high-growth area with support from the CHIPS and Science Act. The size of the market encourages large bets to capture market share from the range of automotive chipmakers led by Infineon, NXP, and Renesas.

Outlook

Inference-focused companies in niche applications can generate large businesses. Even without beating Nvidia at foundational AI tasks, large businesses can be created at the leading edge of emerging use cases. For example, Renesas has developed a sustainable market position in automotive semiconductors with a market cap of around $15 billion. We are seeing inference startups carve out commercial traction in large and high-growth markets outside of datacenters, giving them the potential to create standalone businesses.

Datacenter training startups may be subject to talent acquisition or intellectual property sales in an economically challenging environment. Nvidia's continued rate of innovation leaves datacenter training struggling to achieve superior results in test settings, let alone in customer environments. Furthermore, startups' focus on optimizing specific model types, including foundation models, leaves them exposed to shifts in AI paradigms as transformer models make room for diffusion models for multimodal learning. The serviceable market for foundation model training will likely remain too small to support large companies, thereby resulting in relatively low acquisition offers from research & development-focused innovators and hyperscalers.