Too many AI chips, too few customers

A white paper from JPR

Suppliers up 267%. More than $31 billion invested. The customer list never followed.



JPR
Jon Peddie Research

AI processors began development in 2012 and appeared in 2014 as video post-processors and conditioners for image capture systems in AVs and high-end cameras. They were also used as audio pre-processors. When LLMs and transformers were developed, the GPU was (and still is) the best processor for that work. Since then, there has been a proliferation of companies offering specialized AI processors for inference and training.

Too many AI chips, too few customers

Five segments, 267% supplier growth, $31 billion

Here is a market that broke its own math. Since 2012, AI processor suppliers grew 267%, investors put more than $31 billion behind them, and new processor types kept arriving. The customer list never grew to match. Most founders admitted they would not catch Nvidia and turned toward device inference and IoT. Here is the short history of how that happened, where those bets sit today, and what it means for anyone buying or building on this silicon.

The AI processor (AIP) market is crazy. Why?

- Too many suppliers
- Too much money
- Too many processor types
- Too few customers

The AIP market got started in 2014 with image processor chips, single-function, limited programmability, and limited market size. In 2012, Dr. Fei-Fei Li's ImageNet showed cats could be discovered without being told what a cat was (her initial work started in 2007 at Princeton). In 2017, transformers and the first LLM programs came out. (Dr. Fei-Fei Li is now referred to as the godmother of AI.)

2012 kicked off the crazy rise of AIP companies.

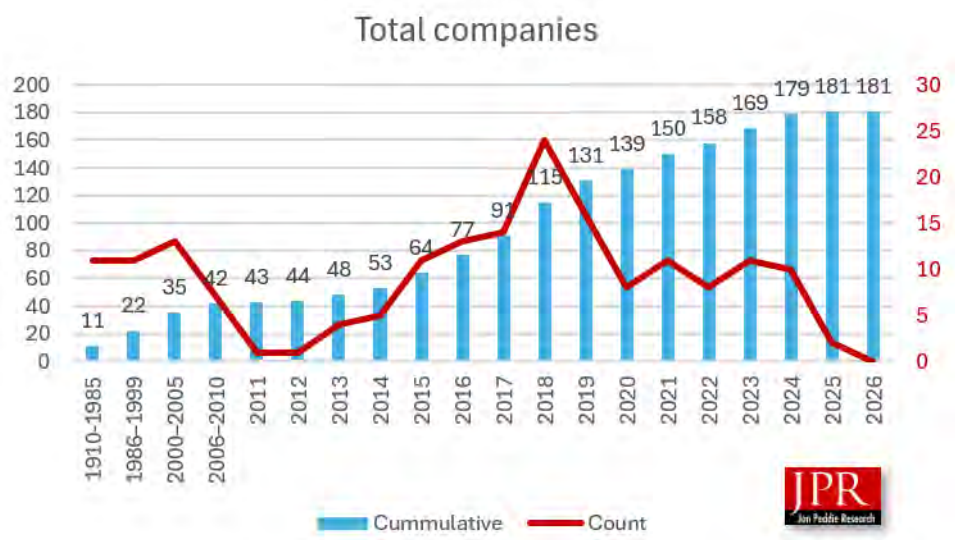

Figure 1. 2012 to 2018 had the most dramatic growth in AIP companies.

ImageNet demonstrated the value of data at scale. AlexNet demonstrated the value of GPU-accelerated deep learning. Unsupervised experiments showed that neural networks could discover useful representations without explicit labeling. Transformers provided an architecture that could exploit increasingly large datasets and compute resources. Scaling those ingredients produced large language models, which turned deep learning from a collection of specialized recognition systems into general-purpose generative computing.

That is also why 2012 and 2017 mark the inflection points in AI processor history: 2012 established the GPU/deep learning relationship; 2017 provided the architecture that would eventually consume GPU compute at data center scale.

How was it possible to expand the AIP supplier base by 267% in just seven years? Money. Lots and lots of money. So far, there has been over $31 billion dollars. And the volumes of money invested accelerated after 2017.

© Copyright Jon Peddie Research 2026

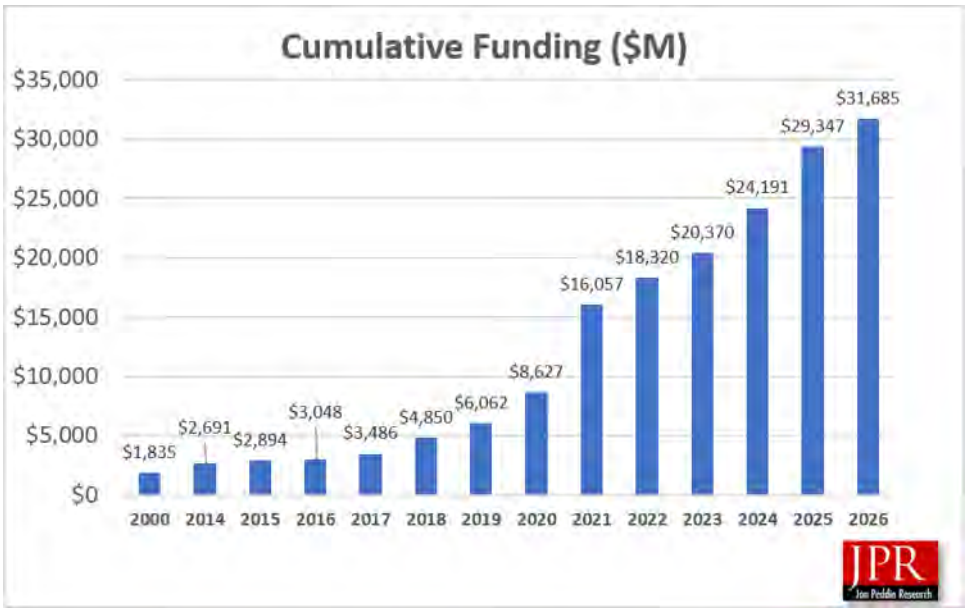

Figure 2. The amount of venture capital has never been this high before.

The crazy money followed the types of AIPs being introduced, most of them YANKs (yet another Nvidia killer). By 2019, Nvidia was stealing all the headlines as it became a trillion-dollar company in May 2023.

As the market expanded, clever new processor types emerged, following the expansion of new AIP companies.

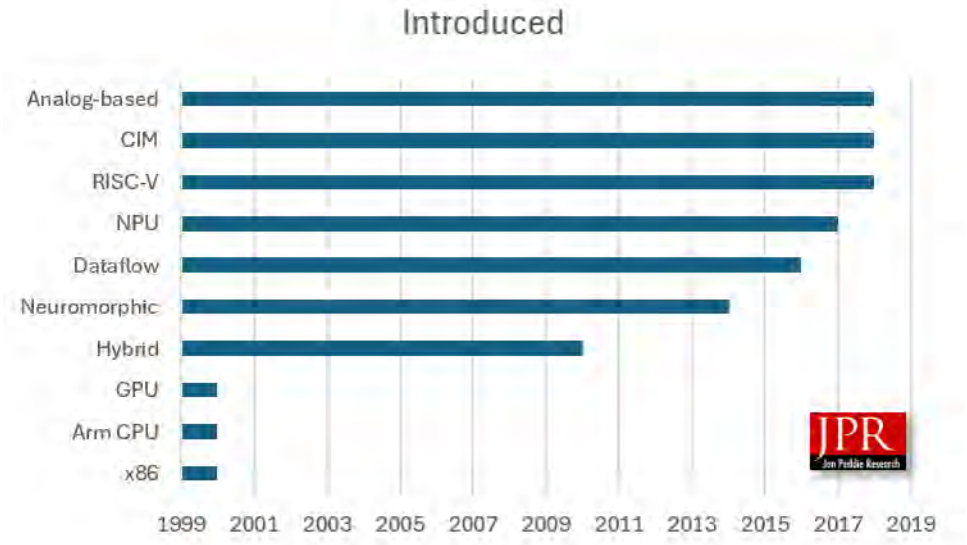

Figure 3. Introduction dates of nuevo AIPs.

As the number of suppliers and the types of AIPs expanded, many of the start-ups realized they didn't stand a snowball's chance in hell of displacing Nvidia or even catching up with them, so they diverted into other market segments.

© Copyright Jon Peddie Research 2026

The AIP market is segmented into five primary categories, with device inference and IoT being the fastest-expanding and potentially largest sectors. As a result, 67% of them have targeted those two markets. And, in AI training, the sector Nvidia dominates has the smallest number of suppliers.

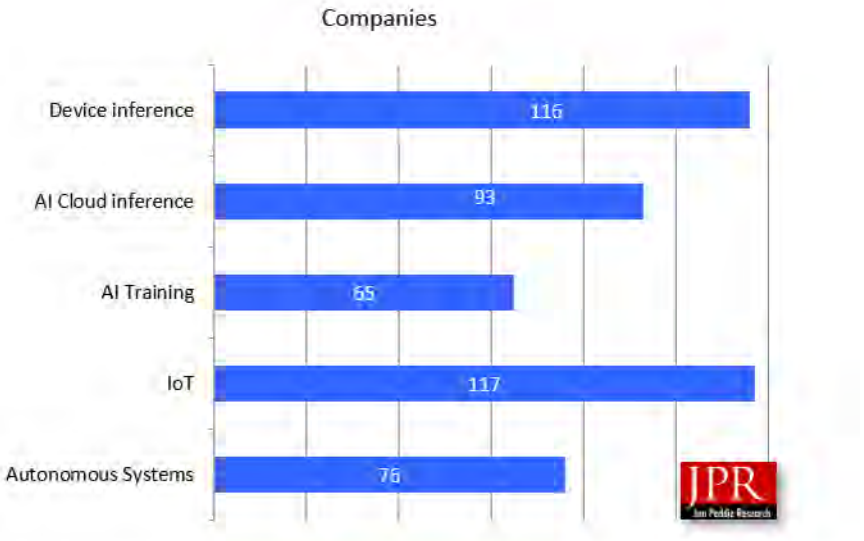

Figure 4. Number of companies in AIP market segments.

AI has taken over the news. You'd be challenged to find something that isn't AI now.

Chipmakers are racing to keep up with demand.

Data center developers are racing to secure the power and land to run those chips.

A fast-growing set of generative AI companies is racing to turn all of that infrastructure into products, revenue, and market share.

And it all rests on AI processors—the little engines that could, are.

The pattern behind all four complaints is the same one: Capital moved faster than demand, and architecture moved faster than software. Silicon teams now compete on compiler coverage and system integration as much as on TOPS, and buyers hold more leverage than the headlines suggest. Device inference and IoT will absorb most of the surviving suppliers, training stays concentrated, and consolidation writes the next chapter. Count the customers before counting the chips.

Supplier count tells you about venture appetite. Customer count tells you about the market. A field this crowded resolves through acquisition, quiet shutdown, and retreat into embedded sockets over years. ISVs should back platforms with funded software teams. Buyers should ask every vendor for a five-year support commitment in writing, then treat the answer as the real specification.

© Copyright Jon Peddie Research 2026

The 2012 and 2017 inflection points came from data and architecture. The next inflection point arrives from economics: A market with 267% supplier growth, $31 billion invested, and a narrow customer base cannot hold that shape. Consolidation moves value toward whoever owns the software stack and the interconnect, and device inference becomes the volume story, with training still concentrated. Watch funding rounds convert to acquisitions, and watch how many suppliers still ship silicon in 2028.

© Copyright Jon Peddie Research 2026

© Copyright Jon Peddie Research 2026. All rights reserved.

Reproduction in whole or in part is prohibited without written permission
from Jon Peddie Research.

This report is the property of Jon Peddie Research (JPR) and made available to a restricted number of clients only upon these terms and conditions.

Agreement not to copy or disclose. This report and all future reports or other materials provided by JPR pursuant to this subscription (collectively, "Reports") are protected by: (i) federal copyright, pursuant to the Copyright Act of 1976; and (ii) the nondisclosure provisions set forth immediately following.

License, exclusive use, and agreement not to disclose. Reports are the trade secret property exclusively of JPR and are made available to a restricted number of clients, for their exclusive use and only upon the following terms and conditions. JPR grants site-wide license to read and utilize the information in the Reports, exclusively to the initial subscriber to the Reports, its subsidiaries, divisions, and employees (collectively, "Subscriber"). The Reports shall, at all times, be treated by Subscriber as proprietary and confidential documents, for internal use only. Subscriber agrees that it will not reproduce for or share any of the material in the Reports ("Material") with any entity or individual other than Subscriber ("Shared Third Party") (collectively, "Share" or "Sharing"), without the advance written permission of JPR.

No material from JPR may be used for AI training. The term "AI training" refers to the process of teaching a machine learning model to make predictions or decisions based on data. That includes but is not limited to: 1. Supervised Learning: Training with labeled data (input-output pairs), such as Image classification, spam detection, sentiment analysis. 2. Unsupervised Learning: Training with unlabeled data to find patterns or groupings, such as Clustering (e.g. K-means), dimensionality reduction (e.g. PCA). 3. Self-Supervised Learning: A hybrid where the system generates labels from the data itself, such as Contrastive learning, masked language modeling (used in GPT, BERT). 4. Reinforcement Learning (RL): Agents learn by taking actions and receiving rewards or penalties, such as Game-playing AIs (e.g. AlphaGo), robotics, navigation. 5. Unsupervised Reinforcement Learning: Agents explore environments without a defined task or reward, such as Pretraining agents for general skills before fine-tuning. 6. Multi-Task Learning: Training one model to handle multiple tasks simultaneously, such as A model that can translate and summarize text. 7. Transfer Learning: Using a pre-trained model on a new, related task with minimal retraining, such as Fine-tuning GPT or BERT on a specific domain. 8. Few-Shot / One-Shot / Zero-Shot Learning, Few-shot: Trained on a very small number of examples. One-shot: Trained on just one example. Zero-shot: Performs tasks without any task-specific training. 9. Meta Learning ("Learning to Learn"): Training models that can adapt to new tasks with very little data, such as Model-Agnostic Meta-Learning (MAML). 10. Curriculum Learning: Training models by gradually increasing task difficulty. 11. Online Learning: Model updates continuously as new data arrives, such as Stock prediction systems. 12. Federated Learning: Training across decentralized devices without sharing raw data, such as Mobile keyboard prediction models. 13. Contrastive Learning: Learns by comparing positive and negative pairs of data. Used In: Self-supervised vision models like SimCLR, CLIP. Specialized AI Training Terms, Fine-tuning: Adjusting a pre-trained model on a new dataset. Pretraining: Initial broad training, usually on large datasets (e.g. GPT pretraining). Prompt Tuning / Instruction Tuning: Adjusting models using instructions or task-specific prompts. RLHF (Reinforcement Learning from Human Feedback): Used in GPT models to align behavior with human intent. Distillation: Training a smaller model to mimic a larger model. Active Learning: Selecting the most informative data points to label next. Semi-supervised Learning: Mix of labeled and unlabeled data for training.

Subscriber shall be liable for any breach of this agreement and shall be subject to cancellation of its subscription to Reports. Without limiting this liability, Subscriber shall be liable for any damages suffered by JPR as a result of any Sharing of any Material, without advance written permission of JPR.

Important Note re Sharing of Material. If Subscriber wishes to Share any Material with any Shared Third Party, it shall contact JPR and request permission to do so. JPR shall forthwith prepare an appropriate Sharing agreement, which shall set forth the name and contact information of the Shared Third Party, the terms and conditions of such Sharing, the compensation to JPR for such Sharing, and the agreement by Subscriber and the Shared Third Party that the Shared Third Party shall not, under any circumstances, Share any of the Material with any other entity or person.

Disclaimers. JPR Reports contain "reviews" of various products and information. The content of these Reports represent the interpretation and analysis of statistics and information that are either generally available to the public or released by entities or individuals deemed responsible. The information in the Reports is believed to be reliable but is not guaranteed as to its accuracy or completeness. The Reports do not endorse or attempt to sell any products and should not be interpreted differently. JPR assumes no responsibility for the correct or incorrect usage of any trademarks or service marks.

© Copyright Jon Peddie Research 2026