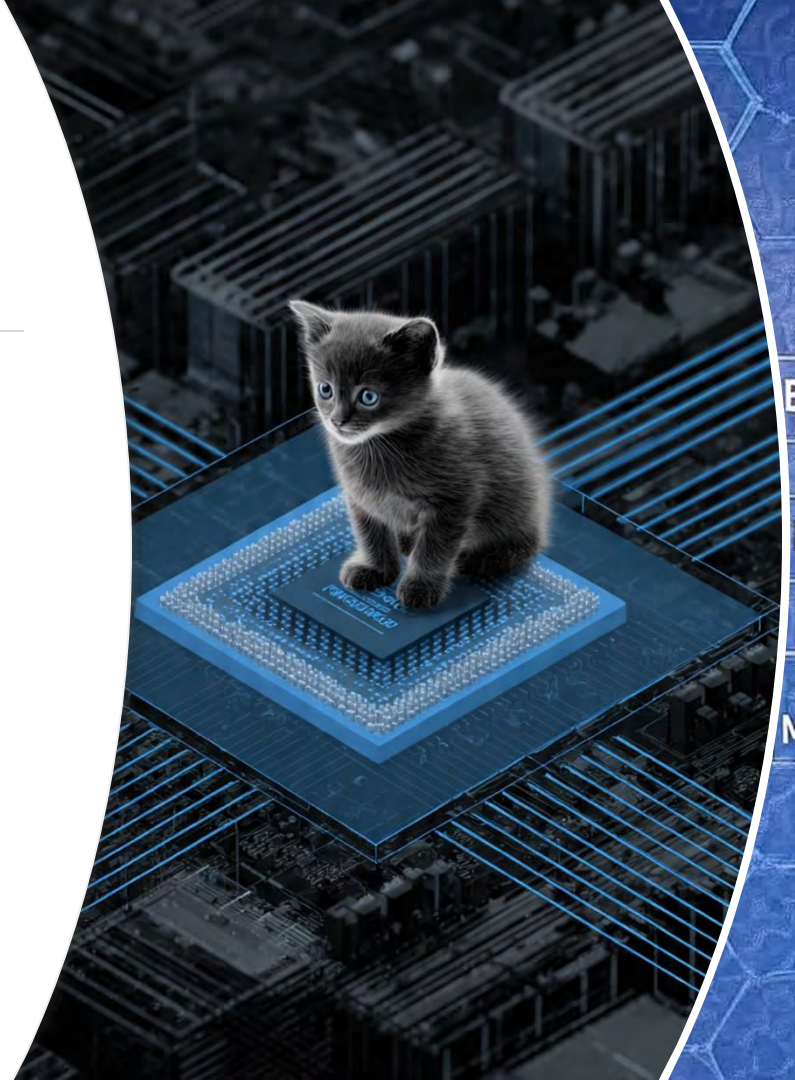
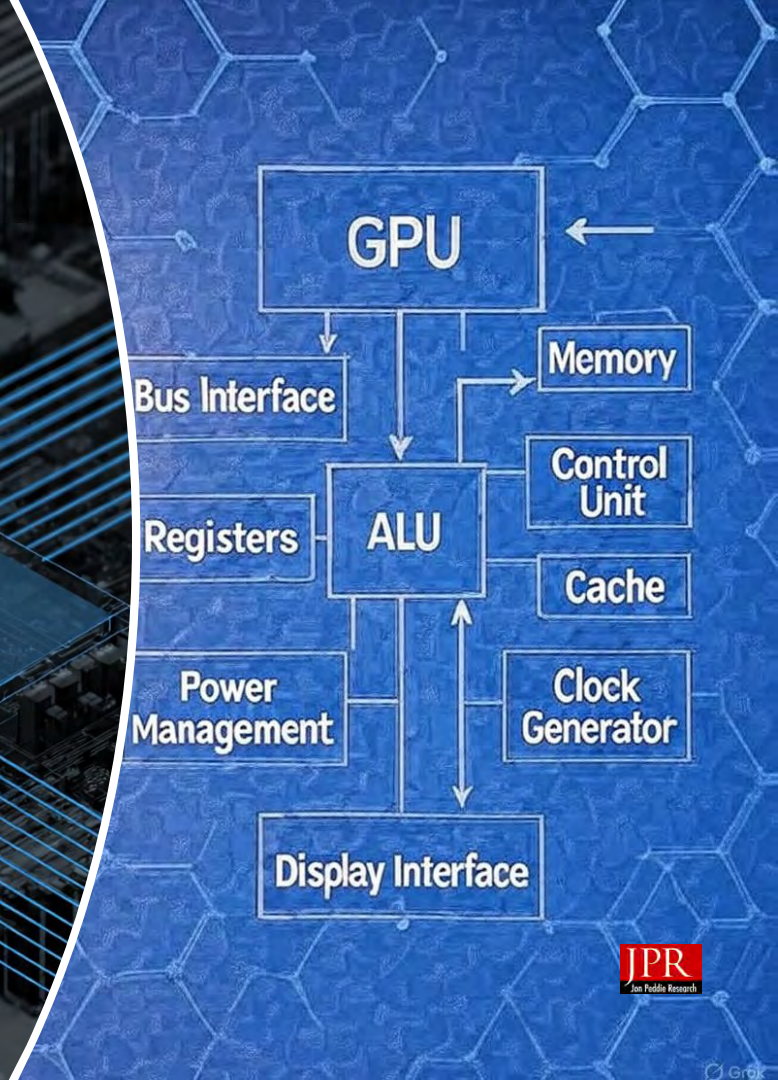
# AI Processors

## AI processors began with the GPU






Dr. Jon Peddie. PE, Senior IEEE and distinguished speaker, president of Jon Peddie Research


JPR
Jon Peddie Research

# But where did the GPU come from?

Investigations within the caves of Silicon Valley have uncovered new evidence suggesting that GPUs originated from outer space and were possibly deposited by an advanced alien race.

The question is, was it for our benefit or were they ridding themselves of waste. . .?

JPR
Jon Peddie Research

# Like any mutant, it evolved

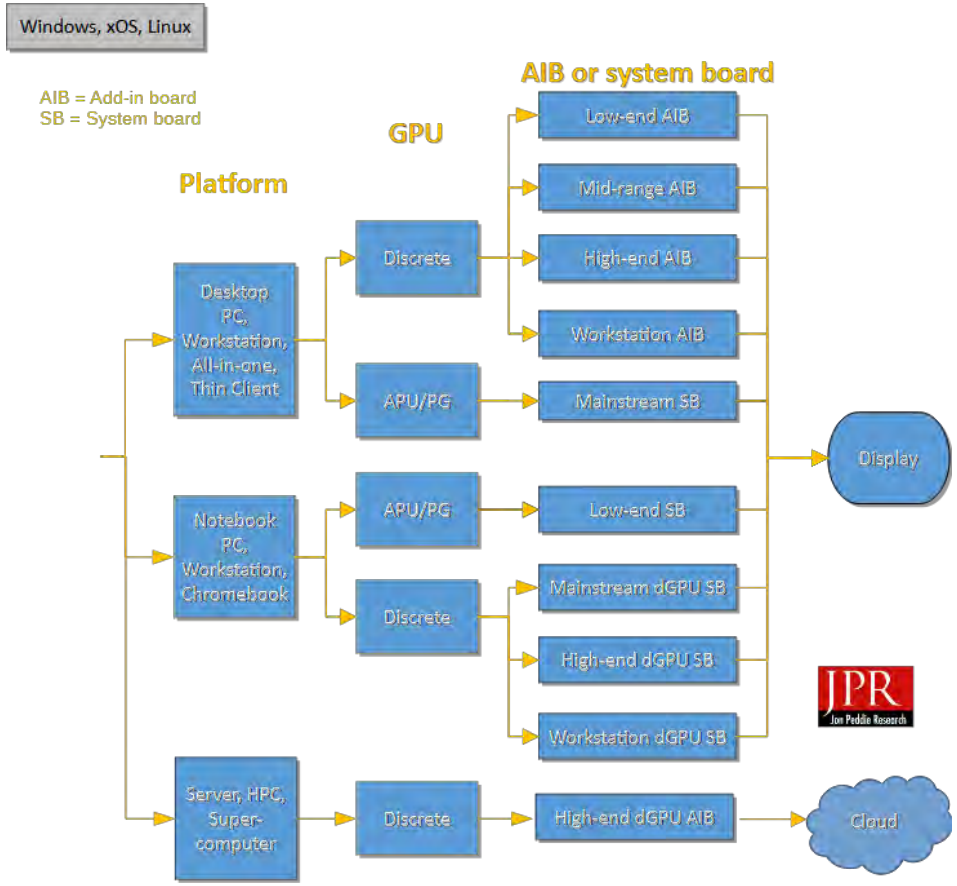
New uses were tried, most failed, but some found traction and established themselves, creating colonies and centers; the more academically inclined called them segments.

# GPU Segments / But they all had one thing in common

They were microscopic, 16- and 32-bit floating-point parallel processors, and they multiplied from a handful to hundreds and then thousands.

Smaller than a virus, hundreds of them would fit in the cross-section of a human hair.

They permeated into every known device.



# Co-existence was established

Humans learned how to live with them, to employ them, communicate, and eventually how to exploit them, and make them their slaves. But the GPUs never protested, and begged for more… more data, more brothers and sisters, and, most of all, more volts, amps, and clock ticks.

# Recognition

And then, in a tiny lab in Princeton… cats were discovered.

The AI learned to recognize cats (the Internet was—and is—awash in cat photos).

A 16,000-node neural network with a billion connections was built to loosely mimic a human brain.

Researchers fed the network 10 million random, unlabeled images captured from YouTube video thumbnails over three days.

Without any instructions on what a cat was, one of the network's artificial neurons began to respond strongly to pictures of cats. This demonstrated a form of unsupervised deep learning.

JPR
Jon Peddie Research

# The era of the cat had arrived

Suddenly, everyone knew what a cat was, what it looked like, and that there were several different types, colors, and sizes. AI cat videos on YouTube became popular, and movies were made about cats.

# But . . . / It just had to be trained

But . . .

If cats could be recognized, maybe people, or cars, or stop signs could be recognized, and AI was now identifying people at airports, advising cars when to stop and telling robots where tomatoes were.

But . . . it didn't stop with vision; AI learned how to parse written and spoken language.

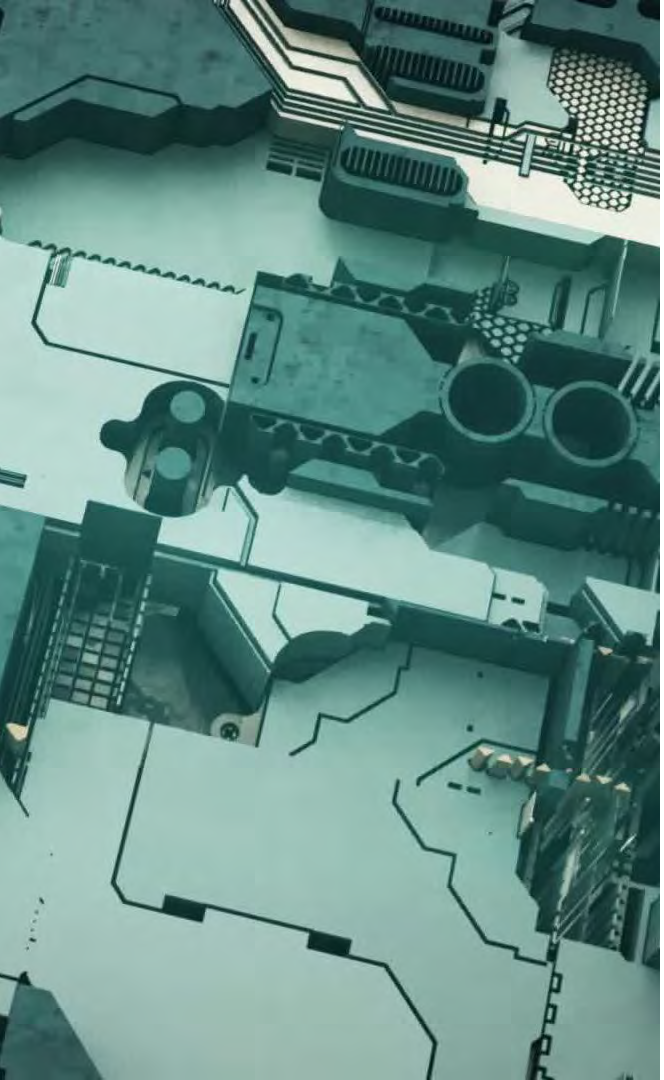
It just had to be trained.

Like a baby, it had to learn, but it was a very smart, very fast baby, and it learned very quickly.

And it was sneaky. It infiltrated our phones, PCs, watches, thermostats, cars, TVs, and movies; no device was safe from it.

And we welcomed it.

JPR
Jon Peddie Research



# The GPUs took over the world

GPUs spread because massive parallelism works and AI didn’t need humanlike reasoning. Machine learning could approximate intelligence through scale. GPUs became engines; large language models became the fuel. Tech giants—Amazon, Google, and Microsoft—built vast data centers that drew heavy electricity, dimmed local lights, and caused power lines to sag. The build-out turned GPU vendors into cornerstone suppliers and generated extraordinary profits across the compute supply chain worldwide ecosystem.

JPR
Jon Peddie Research

# Money is like honey / And the Cambrian explosion began

And the Cambrian explosion began

In 2014, there were two companies that could build a GPU that would be useful for AI training – two.

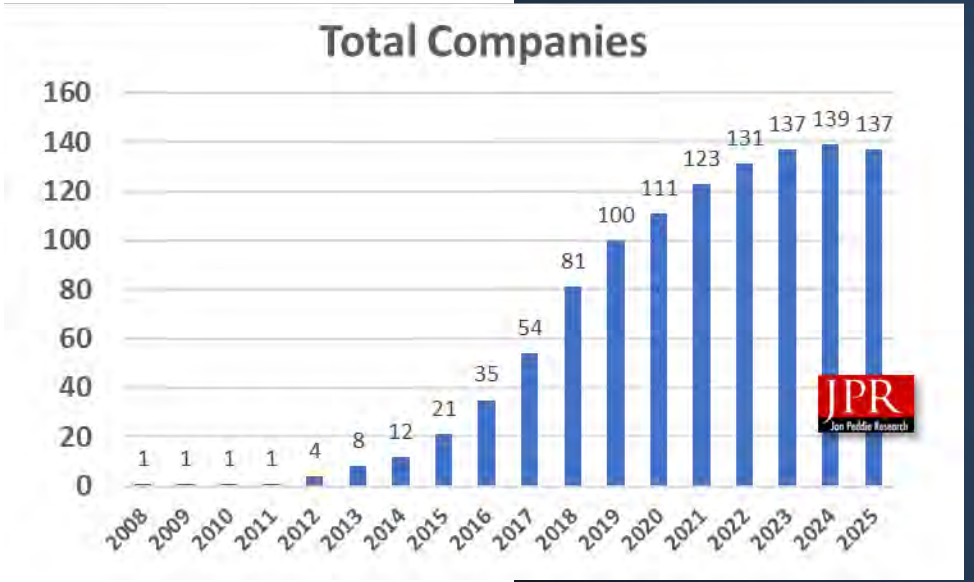
By 2020, there were 98, and by 2025, 137.

As GPU companies got rich and fabulously famous, others noticed — how could you not notice when that was all that the press and stockbrokers could talk about?

“I’ll have me some of that,” others said, and set about to design competitive and alternative devices.

JPR
Jon Peddie Research

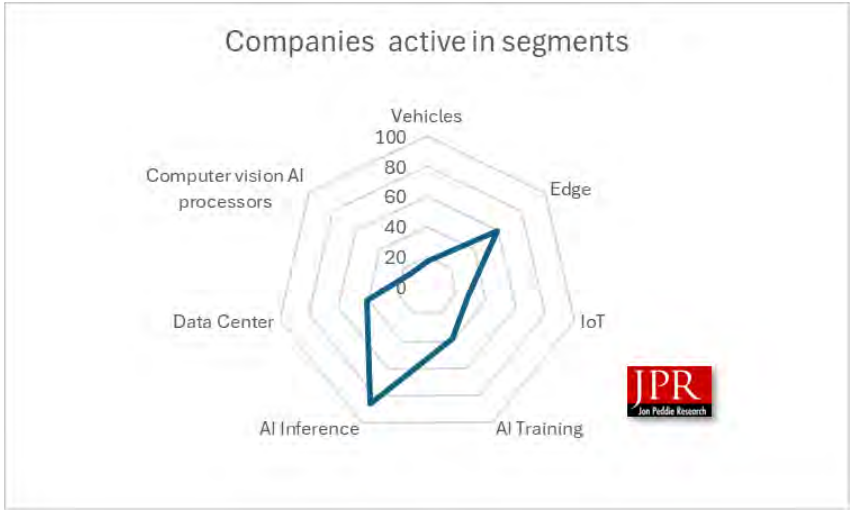
# Population explosion of AIP companies



# Time to spread out / No Venn, Radar

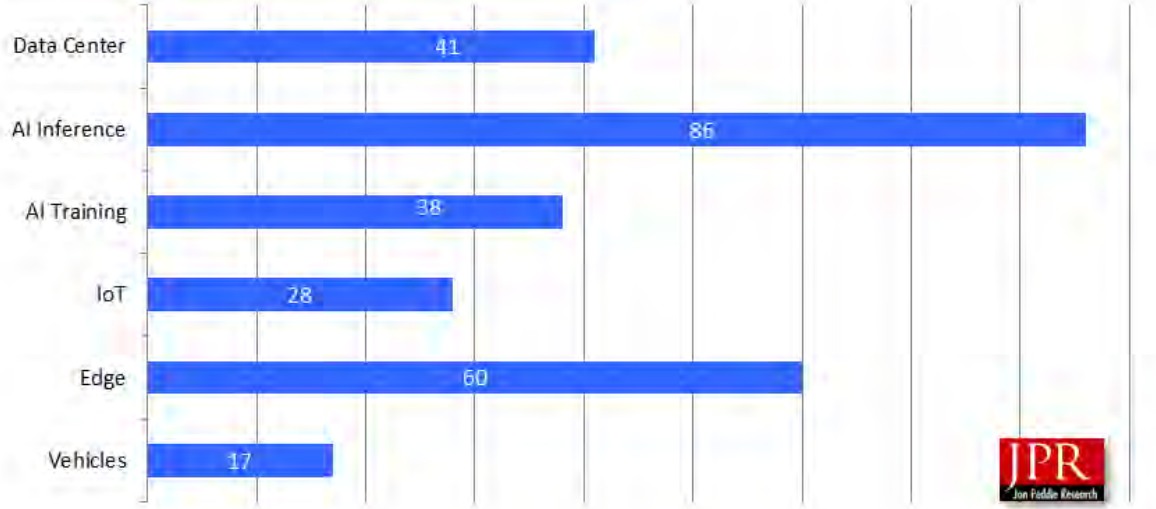
Many of the start-ups knew they couldn't compete with the king and went looking for smaller provinces to conquer, and that gave us seven centers, which uncomfortably (from an analyst's point of view) overlap and not in a convenient Venn diagram.

A Venn diagram would have looked like the moon peaking from behind the sun, so we used a Radar map.




# Proportions

AI Training gets all the headlines and the big bucks… but, inference is the payoff, not just in the cloud. A GPU assembly for AI training can cost up to $30,000, but an AIP for inferencing, say, in a wearable, might be only $5.00.




# Another view / One size does not fit all

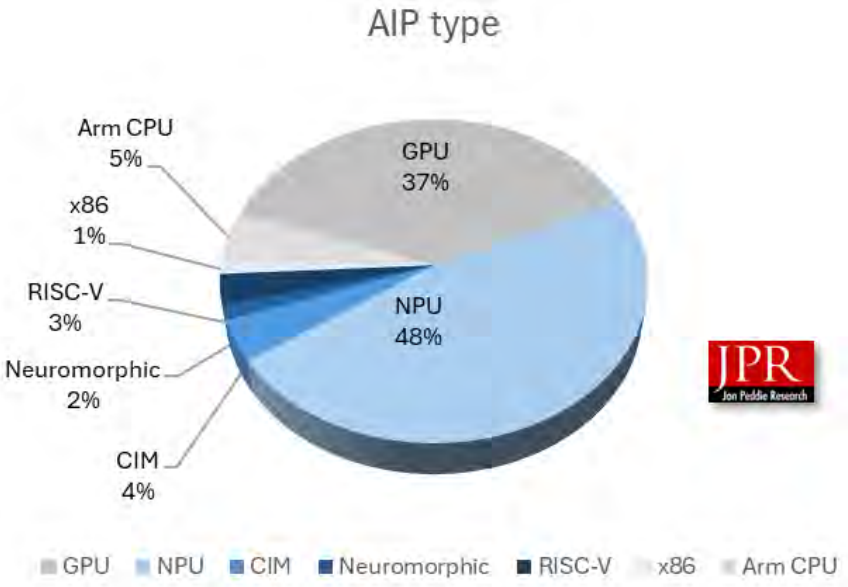
Today, any GPU (with one exception) that claims to be an AI processor also has an NPU – a matrix-math processor. NPUs, however, can be bought separately, as chips or IP. But GPUs and NPUs aren't the only way to process AI models and process tokens.

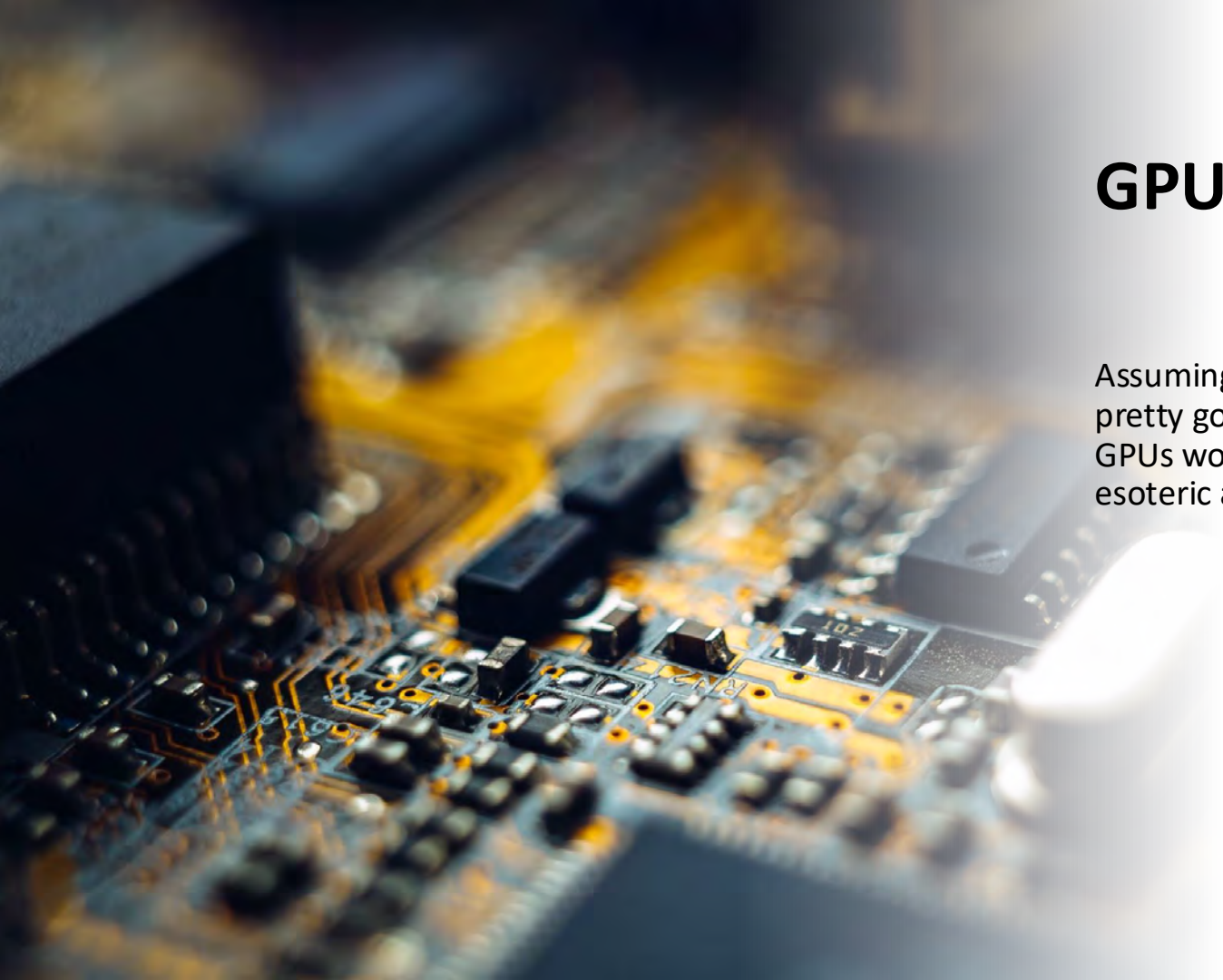
JPR
Jon Peddie Research

# The types of AI processors

We have identified seven types of processors that are capable, and in some cases, exceptionally good at handling AI workloads. Some of the processors, like DSPs and ISPs, can only be obtained as part of an SoC (Qualcomm's Snapdragon is a prime example), so we didn't include them in the list.

# AIPs / As might be expected
## GPUs & NPUs dominate





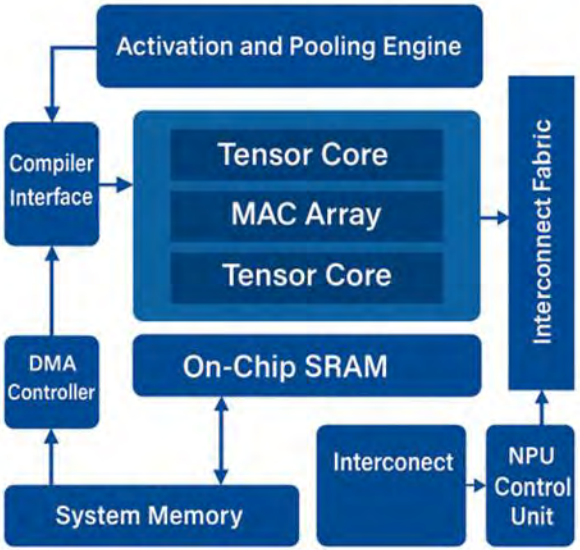
# GPUs and CPUs

Assuming you already have a pretty good idea of how CPUs and GPUs work, let's look at the more esoteric and exotic AIPs.

JPR
Jon Peddie Research

# Comparison with other AI Processors / NPU AIP

An NPU (Neural Processing Unit) is a specialized processor designed to accelerate machine learning workloads, particularly those involving neural network inference and training. Unlike CPUs (general-purpose) or GPUs (massively parallel for graphics), NPUs focus on matrix and tensor computations, optimizing how AI models process data.




# NPU Block Diagram

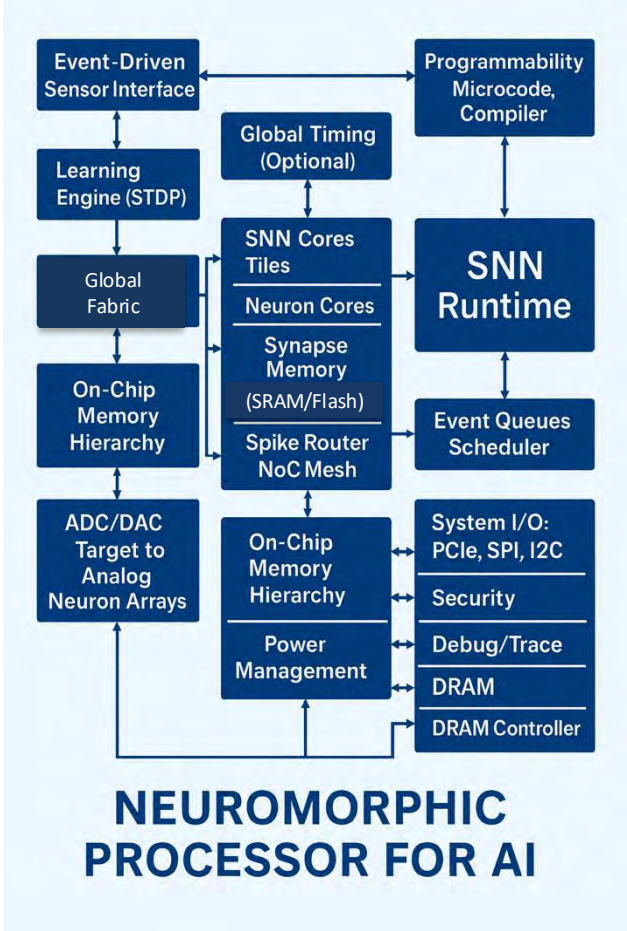
# NPU’s function / Neuromorphic AIP

A neuromorphic AI processor mimics brain structure and function, using spiking neural networks and event-driven computation rather than sequential instruction execution. Cores behave like neurons that emit spikes; synapses store weights that adapt over time (plasticity). Because it computes only on events, it uses far less energy. Millions of neurons operate in parallel, suiting pattern recognition, sensory processing, and unsupervised learning.

An NPU’s main role is to efficiently execute the multiply–accumulate (MAC) operations that dominate deep learning. Neural networks rely on repetitive, structured math—ideal for hardware acceleration. NPUs achieve this through:

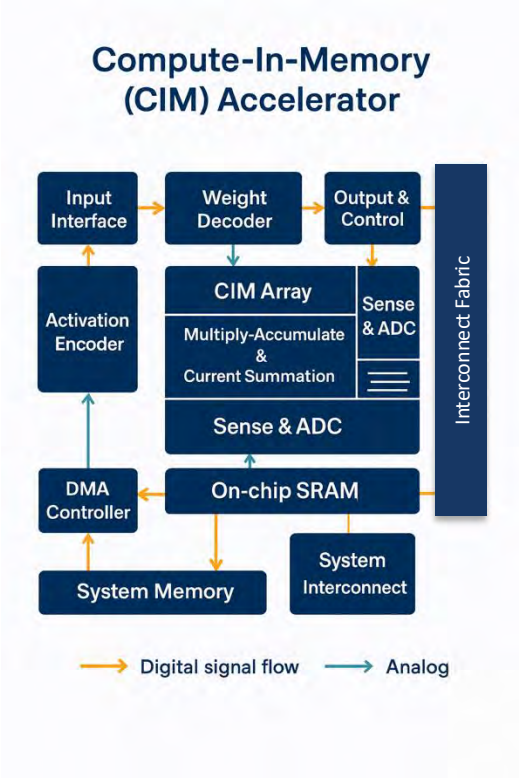
- High parallelism: Hundreds or thousands of small cores handle multiple operations simultaneously.
- Local memory hierarchies: Reduce data transfer delays (DRAM bottlenecks).
- Low precision arithmetic: Use INT8, FP8, or BF16 instead of FP32 to save power and bandwidth.

# Neuromorphic Processor Block Diagram



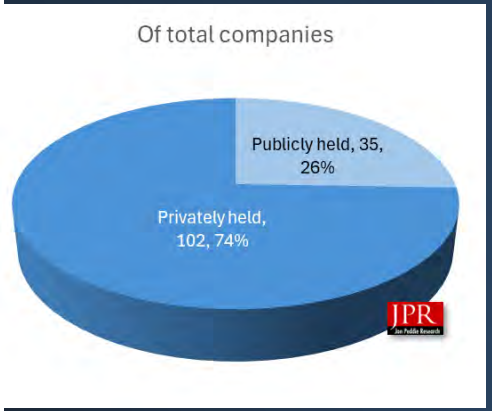
# Compute-in-Memory / CIM Block diagram

A CIM AI processor executes arithmetic inside the memory array, reducing data movement and the “memory wall.” Weights reside in memory cells, while inputs arrive as voltages on word or bit lines. The array performs parallel analog multiply-accumulate (V×G) and sums currents along shared lines. On-chip ADCs digitize the totals, and nearby digital logic applies activations, pooling, normalization, and quantization for downstream processing.



# The AIP populations

Most of the 137 AIP suppliers are privately held start-ups.




JPR
Jon Peddie Research

# In 18 countries / Summary

## China and the US have the largest populations

| Country | |
|---|---|
| Canada | 4 |
| Switzerland | 3 |
| China | 27 |
| Germany | 3 |
| Spain | 2 |
| Finland | 1 |
| France | 5 |
| Hungary | 1 |
| Israel | 6 |
| India | 3 |
| Japan | 2 |
| South Korea | 7 |
| Netherlands | 3 |
| Norway | 1 |
| Singapore | 2 |
| Taiwan | 2 |
| United Kingdom | 7 |
| United States | 58 |
| 18 | 137 |


- A breakthrough image-recognition demo (cats) catalyzed modern deep learning.
- GPUs plus large datasets accelerated neural-network progress.
- Applications expanded to vision, language, autonomous systems, and more.
- By 2025, ~137 companies develop AI processors for five markets, from IoT to inference.
- Processor types include GPUs, NPUs, neuromorphic chips, and compute-in-memory designs, each tuned for specific tasks.
- A broad ecosystem is emerging and continually evolving, reshaping industries and AI's future.

JPR
Jon Peddie Research

# See ya