Edge AI:
The Cloud-Free Future is Now

brainchip
Essential AI

GSA

Contents

Introduction

For decades, Moore's Law saw the number of transistors in densely integrated circuits (ICs) double approximately every two years. This allowed semiconductor companies to design increasingly powerful server chips for centralized data centers.

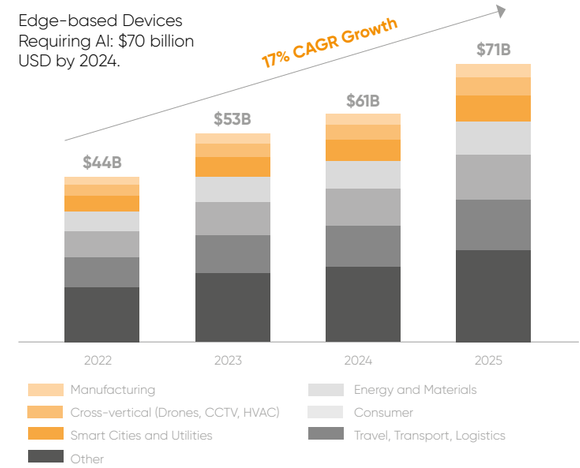
The subsequent rise of cloud computing prompted data centers to shift to a more distributed and scalable model—and optimize heavy and complex workloads by running them concurrently on multiple clusters of CPUs, GPUs, and TPUs.

This distributed cloud-based computing approach enabled artificial intelligence (AI) and machine learning (ML) applications to effectively overcome the von Neumann bottlenecks that once limited data throughput on conventional systems.

With enormous amounts of targeted compute power and memory available in the cloud, AI/ML training and inference models continue to increase in both size and sophistication. However, cloud-based data centers also create a new set of obstacles and concerns for AI applications at the edge including latency, power, and security.

Although shifting AI capabilities to the edge on scaled-down hardware addresses some of these issues, it is arguably insufficient to support a new generation of advanced multi-modal use cases that demand independent learning and inference capabilities, faster response times, and less power.

These include self-driving cars that personalize cabin settings for individual drivers, automated factories and warehouses, advanced speech and facial recognition applications, and robots that use sophisticated sensors to see, hear, smell, touch, and even taste.



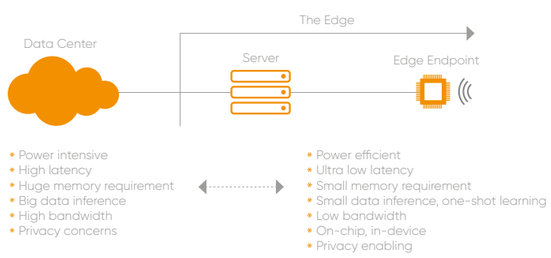
Highlighting the growing market for edge-based devices requiring AI

With smart sensors proliferating and the number of edge-enabled IoT devices expected to hit 7.8 billion by 2030, the semiconductor industry needs to more effectively address the unique learning and performance requirements of edge AI. Neuromorphic edge computing is one solution to this challenge.

Chapter 1:
It's time to cut the cord

Untethering edge AI from the cloud is an important first step to designing faster and smarter endpoints. According to Grace Lewis, principal researcher at the Carnegie Mellon University Software Engineering Institute, edge computing enables devices to achieve faster response times by independently performing localized processing.

As Gartner analysts note, edge devices are also more intelligent when decision-making happens closer to the original source of information. Jennifer Cooke, research director for edge strategies at IDC, expresses similar sentiments, pointing out that a smart edge is needed to make sense of data generated by endpoints.



Differentiating intelligent endpoint requirements

Untethering edge AI from the cloud:

- **Unlocks advanced AI capabilities with incremental and one-shot learning:** Learning independently, locally, and continuously with edge AI eliminates the need for costly retraining.
- **Reduces latency:** Minimizing data movement between endpoints and the cloud reduces latency and accelerates service delivery. This is especially important for time-sensitive applications and use cases such as autonomous vehicles, smart home devices, mobile phones, factory automation, and smart farming.
- **Decreases carbon footprints:** Shifting certain tasks and functions to more efficient edge silicon reduces data center workloads, decreases power consumption, and lowers costs. Because of limited energy budgets and passive cooling constraints, smart edge devices only consume microwatts to milliwatts of power.
- **Enhances privacy and improves security:** Processing, storing, and analyzing sensitive data locally—instead of uploading to the cloud—reduces the overall attack surface.

Perhaps most importantly, untethering edge AI from the cloud creates opportunities for companies to design new products with smarter sensors, devices, and systems. For example, autonomous vehicles are beginning to leverage edge AI learning at high speeds to continuously update and define safety parameters that make it easier for onboard systems to detect anomalous structural vibrations and engine sounds.

Edge AI also enables gesture control with faster response times, allowing doctors and therapists to help people with disabilities interact with sophisticated robotic assistance devices. In addition, edge AI improves applications relying on object and facial recognition. Moreover, edge AI is helpful for many in-field use cases where maintaining constant connectivity is challenging.

These include smart farms in remote areas that help lower food prices by efficiently increasing crop yields with intelligent soil sensors, irrigation systems, and autonomous drones. In the future, field hospitals in disaster zones can deploy medical robots with advanced edge AI capabilities, even if connectivity is limited.

Chapter 2:
Start from the edge

Sales of smart edge devices and systems continue to increase at a rapid pace across multiple verticals. However, advanced edge applications are already hitting the limits of conventional AI silicon and standard learning models.

Limits of Conventional AI

Many chips used in edge applications today are still general-purpose processors such as GPUs that consume approximately 1,000 times more power than purpose-built edge silicon.

While some smart devices are equipped with low-power digital signal processors (DSPs), these single-purpose chips typically lack advanced learning and analytics capabilities. This limitation prevents DSPs from performing more sophisticated tasks such as image classification and responding to gestures and visual wake words.

Although scaling down edge AI hardware to meet the requirements of intelligent endpoints is an important and necessary step, it is arguably insufficient. Moreover, most training and inference models do not function effectively at the edge where resources are limited.

As an example, training a robotic hand to manipulate a Rubik's cube with conventional methods once required approximately 2.8 GWh of electricity—enough to power hundreds of homes for a year!

While training techniques have improved, conventional deep learning models typically demand massive amounts of power and compute resources to support backpropagation and process extremely large dataset workloads. In addition, conventional AI models can't learn or react to new stimuli without retraining entire datasets—along with the devices and systems they are on.

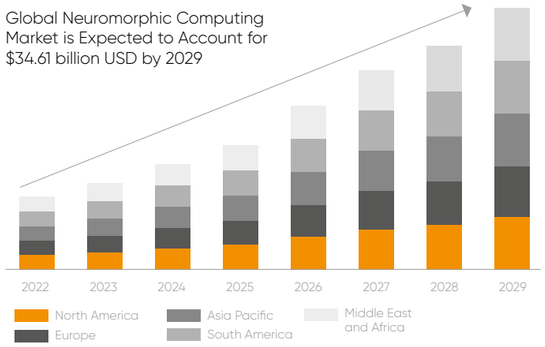
Unique requirements of edge AI endpoints

These real-world limitations make it difficult for the semiconductor industry to meet the unique AI requirements of intelligent endpoints. Without new AI architectures and learning models at the edge, for instance, automotive companies will struggle to design fully autonomous vehicles. Robots that can see, hear, smell, taste, and touch will require a new approach for AI sensor processing.

Medical applications that identify and diagnose diseases by analyzing and comparing thousands of images will take longer to process higher-resolution scans. And factories will find it challenging to cut costs and increase efficiency without the rapid automation of equipment and processes.

Chapter 3:
The human brain can show us the way

According to Gartner, traditional computing technologies will hit a digital wall in 2025 and force a shift to new paradigms such as neuromorphic computing. With neuromorphic computing, endpoints can create a truly intelligent edge by efficiently identifying, extracting, analyzing, and inferring only the most meaningful data.



Highlighting the growing neuromorphic computing market

Neuromorphic computing systems digitally mimic the neuro-biological architectures of the human nervous system. As inventor Carver Mead notes, neuro-biological (neuromorphic) compute architectures learn to understand their environment and are often "many orders of magnitude more effective" than conventional systems for certain use cases.

Edge AI Learning Models

In recent years, neuromorphic computing has helped unlock new learning models and methodologies for edge computing such as incremental learning. Also referred to as continual or lifelong learning, this technique enables devices and systems to learn new tasks without retraining.

Like the human brain, neuromorphic systems leverage incremental learning to build bigger pictures from the basics. By only processing meaningful spikes that represent new events or relevant data, neuromorphic edge silicon can perform up to trillions of operations per second while consuming minimal power.

One-shot learning is another methodology used to train neuromorphic edge systems against very small datasets. One-shot learning techniques can improve the hearing, smelling, tasting, and tactile abilities of smart sensors using a single sample instead of thousands.

Bob Gill, a research vice president in Gartner's Infrastructure Strategies group, says machine learning inferencing is expected to account for over 60% of edge use cases by 2027, while machine learning training at the edge will include 20% of use cases in the same timeframe. As machine learning models evolve, says Gill, more training will happen at the edge.

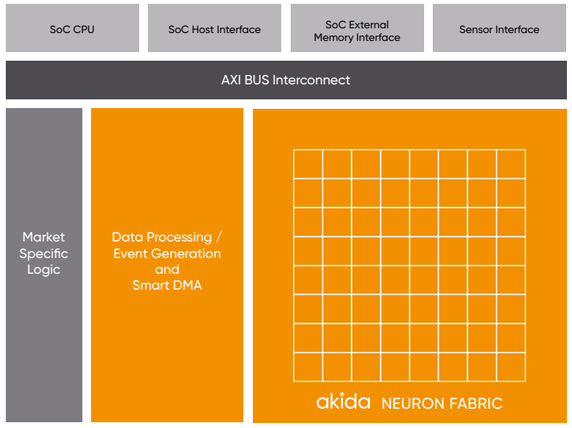
New Edge AI Architectures

Neuromorphic computing has enabled the development of edge AI silicon that processes data with efficiency, precision, and economy of energy. Untethered from the cloud, neuromorphic edge AI silicon can operate and learn independently with an on-chip processor, multiple neural processing units (NPUs), scalable memory, as well as pixel and data spike converters.

By keeping machine learning on the device, neuromorphic edge silicon dramatically reduces latency, minimizes power consumption, and improves security.

From our perspective, AI-enabled edge silicon should follow four primary design principles:

* **Distributed computation:** Multiple NPUs should include dedicated compute and memory to reduce data movement.
* **Event-based processing:** Multiple NPUs should only perform computation on events and spikes (non-zero values).
* **Event-based communication:** Events should be sent over the mesh network without a host CPU.
* **Event-based learning:** Algorithms should enable continuous on-chip learning without retraining.



Optimized Edge AI SoC

Neuromorphic computing effectively addresses the unique learning and performance requirements of intelligent endpoints and offers the semiconductor industry a viable alternative to simply scaling down hardware at the edge. In recent years, the concept of neuromorphic AI chips has gained significant traction, with Intel debuting Loihi and IBM introducing TrueNorth.

Chapter 4:
The new AI: Autonomous

Neuromorphic computing will help build a more unified and heterogenous edge with new AI learning models and architectures. A range of edge devices and applications will leverage these learning models and architectures, including self-driving cars, advanced medical robots, smart home devices, the Industrial Internet of Things (IIoT), and smart farming.

All will infer maximum meaning from a minimal amount of information. The goal is that roads will be safer, smart homes smarter, farms more productive, healthcare more affordable, factories more efficient, and, overall, access to everyday services will improve.

Concurrently, cloud data centers will continue to host and run heavy non-edge AI/ML workloads, including the large-scale training models, inference, and deep analytics that require powerful accelerators such as CPUs, GPUs, and TPUs. Together, cloud and edge AI will form a new, efficient, and faster model of distributed computing optimized to fit the specific requirements of each application.

Gartner's Bob Gill defines this heterogenous approach to the edge as "edge-in." With an edge-in model, companies build edge applications that are optimized for low-latency and low-bandwidth autonomous connections and tap the cloud for other tasks. As Gill explains, some edge implementations now include the use of Google Cloud, AWS, or Microsoft Azure on the backend—and a cloud-independent platform at the edge itself.

Michael McCarthy, assistant professor of information systems at Carnegie Mellon University's Heinz College of Information Systems and Public Policy, also sees cloud computing and edge computing working well together, with cloudlets potentially providing more localized services to edge devices.

Conclusion

The Future's Not Only Bright, It's Essential

Moore's Law and distributed cloud computing have enabled artificial intelligence and machine learning applications to effectively overcome von Neumann bottlenecks that once limited data throughput on conventional systems. With enormous amounts of targeted compute power and memory available in the cloud, AI/ML training and inference models continue to increase in both size and sophistication.

However, cloud-based data centers can also create a new set of bottlenecks for AI applications at the edge such as latency, power, and security. Although shifting AI to the edge on scaled-down hardware is an important first step, this approach does not effectively address the fundamental limits of conventional AI silicon and standard learning models.

In recent years, neuromorphic computing has unlocked new learning models and architectures for edge AI. Smart edge silicon that follows the principles of essential AI—doing more with less—now supports a new generation of advanced multimodal use cases with independent learning and inference capabilities, faster response times, and a lower power budget.

These include cars that personalize cabin settings for individual drivers, smart farms, automated factories and warehouses, advanced speech and facial recognition applications, and robots that use sophisticated sensors to see, hear, smell, touch, and taste.

At BrainChip, we believe edge AI presents both a challenge and opportunity for the semiconductor industry to look well beyond scaling down conventional chip architectures. Neuromorphic edge AI silicon is already enabling people to seamlessly interact with smarter devices that independently learn new skills, intelligently anticipate requests, and instantly deliver services.

Specific strategies to unlocking the full potential of edge AI will undoubtedly vary, which is why it is important to explore a future in which semiconductor companies play a collaborative role in helping to design and implement new neuromorphic architectures.

About The Author:
The worldwide leader in edge AI on-chip processing and learning, BrainChip's commercially deployed technology mimics the human brain to analyze only essential sensor inputs at the point of acquisition, processing data with unparalleled efficiency, precision, and economy of energy. Keeping machine learning local to the chip, independent of the cloud, also dramatically reduces latency while improving privacy and data security. In enabling effective edge compute to be universally deployable across real world applications such as connected cars, consumer electronics, and industrial IoT, BrainChip is proving that on-chip AI, close to the sensor, is the future, for its customers' products, as well as the planet.

brainchip
Essential AI