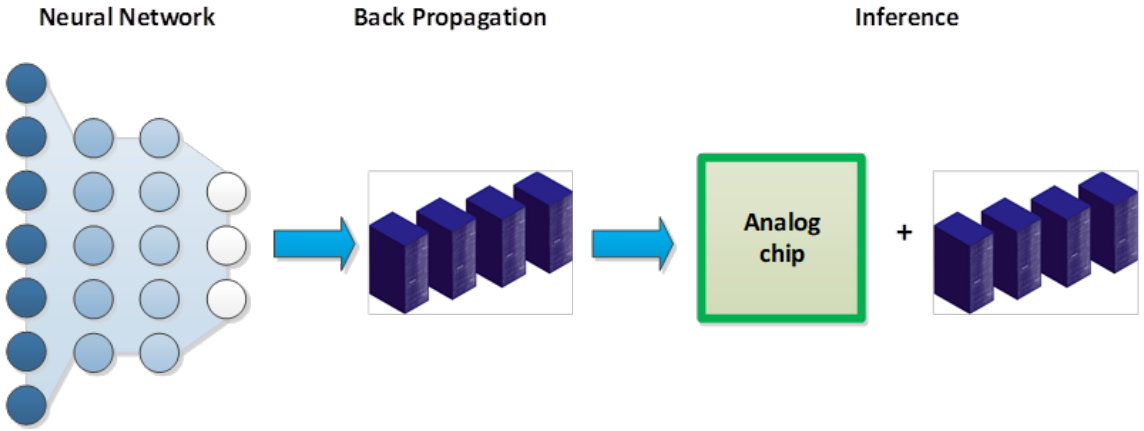
ANABRID's AI Hardware Market Opportunities



Analog Computing in AI

Unlike traditional digital computing, which uses discrete values (bits) to represent data, analog computing processes information in a continuous range of values, resembling the way biological brains process information. This allows neuromorphic chips to efficiently handle the complex, noisy, and dynamic signals inherent in real-world AI applications.

AI Hardware as Bottleneck for Future AI Applications

AI models have enormous compute requirements running in real-time on cost-effective, low-power hardware, therefore we work on a new hardware platform that employs analog compute-in-memory computing.

Today, most neural network inferences are carried out on servers using Graphics or Tensor Processing Units (GPUs or TPUs), specifically designed for parallel matrix calculations. These accelerators process thousands of coefficient "parameters," akin to synapses, across each "node," similar to neurons. These networks are structured in layers, with each layer containing thousands of nodes, each connected to thousands of others in preceding and subsequent layers. In large language models (LLMs), these nodes correspond to tokens, which are textual language elements and symbols. The model uses the history of previously generated tokens, such as a prompt and its response, to calculate probabilities and determine the most likely next token. For generative AI focusing on speech-to-text, the demand for tokens—comparable to words or symbols—is expected to surpass 10 trillion by the end of 2023 (Tirias Research), with more than 400 million monthly active users predominantly in developed markets. Projections for the end of 2028 anticipate over 6 billion users, accounting for roughly 90% of the smartphone market, and an annual demand of more than 1 quadrillion tokens, representing a 100-fold increase.

→ By 2028, it is predicted that the power consumption for Cloud GenAI will exceed 66 billion kilowatt-hours.
→ By 2028, it is expected that cloud-based generative AI will annually use as much power as 19 billion flagship smartphones.

AP3003 Neuromorphic Training and Inference Processor Product Brief

Market Drivers and Value Potential

Research predicts that by 2028, the combined infrastructure and operational costs for data centers running generative AI will surpass $76 billion. This figure is double the projected annual operational expenses of Amazon's cloud service, AWS.

- The prognosis for cost and scale of Generative AI will require innovation in optimizing Neural Networks and is likely to push the computation workload out from data centers to single devices like PCs and smartphones.
- Replacing current TPU and GPU technology and move GenAI to the edge even in part could save up to $ 16 billion Dollars
- The anticipated expansion of practical GenAI application services is being propelled by research and development in both academic and corporate settings.
- Neuromorphic computing is poised to significantly enhance embedded systems in edge devices, impacting various industries profoundly. Central to its vast potential is its capacity for low power consumption, which is fundamental to its utility and appeal. As these services are introduced to the market, there is an expected increase in demand from businesses and consumers.
- Additionally, the total operating costs of the hardware that powers these services in the cloud are also projected to rise.
- Turner and Townsend Survey explains that 88 percent of their data center roster have a need for data centre capacity for AI and machine-learning projects due to rapid increase in demand. AI is already resulting in higher power density requirements for data centers.

Use Cases/ Benefit potential	**Market Size in Billion $**	**Market Share in %**	**Benefit Potential in % of Operating Profit**	**Applications**
Edge Server	100	2-4	10-20	keyword spotting, machine learning and object detection, context-aware inference
MedTec and Pharma	60-110	3-5	15-25	Research, Drug Design, Documentation
Banking and Finance	200-340	3-5	10	Risk Modelling, Legacy Code Migration, Custom Banking
Consumer Electronics Device				Stress Level Estimation; Speech to Text or Voice Control, Trigger Word Detection
Embedded control systems	240-460	45-10	10	Hybrid Motion Control, IoT on Device, Industry 4.0

Benefits for Data Centers

Energy Efficiency: Analog neuromorphic chips require significantly less power than their digital counterparts.

Speed Improvements: By processing data in a manner akin to human neurons and synapses, neuromorphic chips can speed up the processing of neural network algorithms.

Scalability: Analog components can be compactly integrated, providing a denser and potentially more scalable architecture suitable for the expansive needs of large data centers.

Reduced Latency: The direct processing capabilities of analog neuromorphic chips enable faster data throughput and lower latency up to 500 times with 100 times shorter latency compared to traditional CPU.

Target Applications: Edge Server , Security/Surveillance, Industrial Machine Vision, Consumer Electronics, Smart Home, UAV/Drone, Embedded control systems, Wearables and Medtec

Overview

The AP3003 is a variant of the AP3000 Series of Anabrid General Purpose Analog Processors (Anabrid™ G-PAC) architected for high-performance AI training and inference at the edge. The AP3003 follows the industry's first AP3001 with specialization for wave based computing blocks (WBC Paradigm). The AP3003 is a dataflow processor configured as an array of tiles each powered by Anabrid CMOS Computational Core IP (CCIP) to deliver the AI compute performance of a desktop GPU at 1000x the speed and at the same time 10,000x lower power – all in a single chip. It is ideal for processing complex deep neural networks (DNNs) for applications with power, form factor and thermal constraints.

The AP3003 is available as a chip or PCIe card form factor. The AP3003 hardware comes with the novel anabrid analog compiler for adopting of existing libraries/infrastructures such as TensorFlow, PyTorch and HuggingFace, delivering pre-qualified DNNs that bring new level of performance and power efficiency to demonstrator applications such as always-on voice recognition and life in-situ video stream comprehension.

Anabrid CMOS Computational Core IP

Anabrid is the global word market leader in future pointing analog computing. The Core IP drives every AP3003 processor blocks. Anabrid CMOS IP integrates a memory array and analog circuits that store DNN weight parameters and perform low-power, high-performance equilibrium propagation and optimisation problem solving without the need of external DRAM or a traditional digital ALU.

Product Features

- 1.000 Anabrid CMOS Oscillator Nodes
- Capacity for up to 1M DNN weight parameters
- 4-lane PCIe 2.1 high speed serial interface with 4GB/s/lane or up to 2GB/s of bandwith
- Available I/Os – GPIOs, QSPI, I2C and UART
- -40°C to 150°C operating temperatur (junction)

Key Benefits

Performance and Power Efficiency

- Power of 3W when running typical models
- 10,000x lower power than comparable digital solutions

Single-Chip Execution

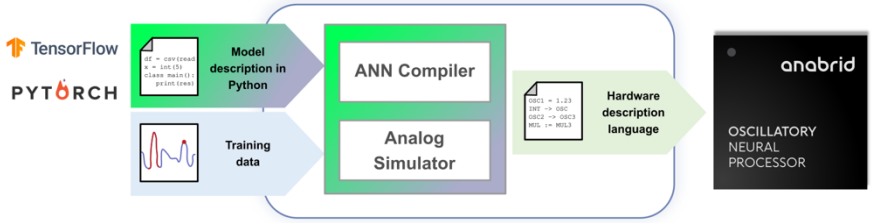
- Model parameters are stored and executed entirely on-chip. No external DRAM is required.
- Multiple DNNs can be executed concurrently
- Efficient dataflow architecture provides predictable and deterministic execution of DNN workloads

Extensions

- CMOS IP can be integrated neatless on-chip as a co-processor for any digital design
- CMOS-based circuits: existing semiconductor technologies configured to emulate neuron and synapse functions.
- Anabrid chips can be interconnected without bounds to allow any size of network

The architecture is optimized for rapid response and real-time processing, essential for applications requiring immediate action, like safety monitors or real-time communication systems.

Target Applications: Data integration from different sensor types; Trigger Word Detection in Always on Devices; Stress Level Estimation; Speech Pattern Transcription for Voice Control; Deep-Learning and ML Acceleration for future AI workloads



think hybrid