NASP technology for near-sensor Tiny-AI

Many applications could benefit from AI, and especially from the neural network paradigm, but practical implementation of this formidable mathematical method suffers from excess power consumption when performed in the traditional way on standard CPUs or GPUs. If an application uses a large amount of data, and accesses the memory very often, it causes a bottleneck within Von Neuman architecture. For cases with a continuous signal flow, special purpose processors are more efficient.

A good example is wearables with Heart Rate (HR) tracing and Human Activity Recognition (HAR), where PPG/IMU sensors constantly generate data, whose processing consumes a lot of battery power.

For devices that perform truly always-on measurements, Neuromorphic Analog Signal Processing (NASP) is an ideal solution, with ultra-low 100uW power consumption and accuracy doubled compared to traditional algorithms. The increased accuracy also enables simplification of the entire system, and reduction of related costs.

Another power-hungry application is the Predictive Maintenance (PDM) sensor node. Industrial IoT (IIoT) utilizes IoT devices and sensors to monitor machines and environments, to ensure optimal performance of the equipment and processes. PDM monitoring the health of machines to identify (a.k.a. predict) a probable failure of components is an IoT technique receiving a lot of attention lately. To achieve effective PDM, massive amounts of data are collected, processed, and analyzed by Machine Learning (ML) algorithms. If all this data must be sent to a center for analysis, the data communication and processing would be more trouble than it's worth. On-sensor data pre-processing could significantly reduce the amount of data sent to the cloud, saving money, and improving latency.

The NASP addresses all these situations, and many other uses with smart optimization (pre-processing) of raw data directly on-sensor. Not only could it solve problems for existing applications, but it could also open new opportunities for the whole industry.

On-sensor data optimization

NASP is a true Tiny AI solution targeted for raw data optimization and reducing the CPU load and amount of data forwarded to the cloud. The NASP chip is located right next to a sensor, forming the Tiny AI logical layer. It is an inference solution that uses already trained machine-learning models to make predictions.

In the NASP concept, the data processing chip is synthesized from already trained neural networks by NASP automation tools.

Based on POLYN's years of expertise, the "inference-only" approach is highly efficient for applications such as voice extraction, sound/vibration processing, measurements on wearables and more. It provides a huge advantage in power, accuracy, and latency.

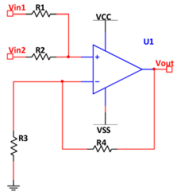
Neuroscience behind the NASP

The main advantage of neural network computing is parallel operation. The top advantage is neuromorphic computing, especially geared for maximum parallelism through hardware and software design that strives to mimic the human brain and achieve its compute-to-power-consumption efficacy. Besides low power consumption and improved performance of computing workloads, neural networks provide fault tolerance, which means the system can still produce results if a sensor data is inconsistent.

All sensor signals entering the input layer of the NASP chip at the same time are transmitted to the successive layers in parallel. There are no execution cycles, and no instructions directed to/from memory.

The human brain is not only an ultra-low-power parallel operating system, but also an analog system, processing a variety of signals without converting them into a binary format. For tasks such as signal perception, analog systems are preferable. According to Semiconductor Research Corporation, the analog signal deluge is expected in the coming decade, demanding fundamental breakthroughs in hardware to generate smarter world-machine interfaces.

The NASP is precisely one of these breakthroughs, built to perceive analog signals as well as digital ones and, most importantly, to add "intelligence" to various sensors.



The NASP chip contains artificial neurons (nodes performing computations) and axons (connections with weights between the nodes) implemented using circuitry elements: neurons are implemented using operational amplifiers, and axons by using thin-film resistors.

The NASP chip design embodies the approach of a sparse neural network, with only the necessary connections between neurons required for inference, which means the solution reduces the neural connections significantly and efficiently. In contrast to in-memory designs, where each neuron is connected to each neighboring neuron, the NASP approach simplifies the chip layout. Such a design especially suits for Convolutional Neural Networks (CNN), where connections are very sparse, as well as RNN, Transformers, and Autoencoders.

A neural network adjustment to the chip design is a significant part of every neural network-on-chip solution. Programmable solutions available today in the market have architectural

restrictions that impose additional transformation on a neural network. Sometimes, the original neural network undergoes an almost 100% transformation during porting, which is a costly approach.

To address this issue, the NASP model includes the chip design automation tools, namely POLYN's T- Compiler and Synthesis tools, that convert any trained neural network into an optimal math model for further chip layout generation, while completely preserving compliance with POLYN's customer neural network, and saving related efforts and costs.

The digital transformation that the industry is going to embrace will not be possible without an analog computing renaissance for several reasons.

One is the concept of energy saving. Excessive power consumption is incompatible with data computations in sensory systems.

The next trend is that AI is moving more and more towards the edge, and being applied today to sensor nodes. It is required to optimize communications between billions of IoT devices, and to offload data processing from the cloud, improving TCO and efficiency.

Like the human brain, which excels at processing information that is complex, and changing dynamically in time, the Neuromorphic Analog Signal Processors excel in real-time computing, thus contributing to the beneficial meshing of digital and analog tech worlds.