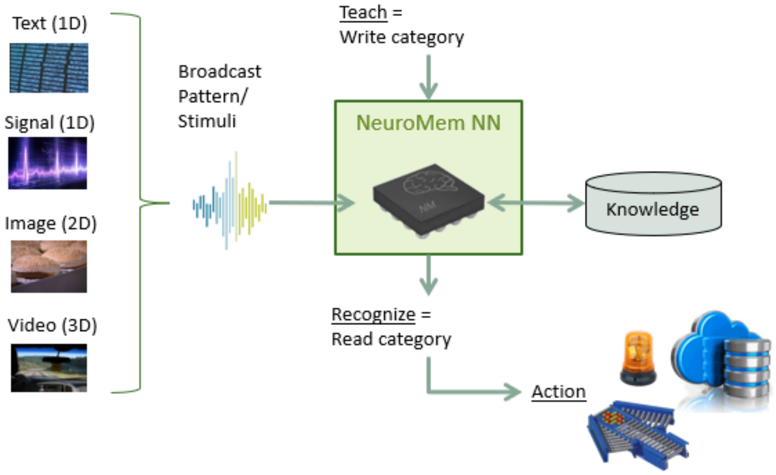
NeuroMem Smart!™

Neuromorphic on Silicon

for practical Artificial Intelligence

A NeuroMem chip is a bank of identical neuromorphic memory cells (neurons) which react to digital stimuli and can learn and recognize in real-time. They are addressed in parallel and have their own "genetic" material to learn and recall patterns without running a single line of code and without reporting to any supervising unit. This is made possible through a patented parallel bus which allows the neurons to fully collaborate with each other and is the key to accuracy, trainability, and speed performance.



A neuron integrates information from the other neurons into its own learning and recognition logic. This interconnectivity allows three mechanisms essential for Artificial Intelligence:
(1) Always retrieving the response of the most confident neurons first,
(2) Learning immediately upon request and without duplication,
(3) Reporting novelty as well as potential uncertainty or conflict,
Other resulting achievements of the parallel architecture of a NeuroMem network are:
(4) its low-power requirement (in Mhz),
(5) its deterministic latency to learn and recognize regardless of the number of connected neurons,
(6) its expendability by cascading chips.

The NeuroMem neurons can learn and recognize digital signatures extracted from any data types such as text, measurements, time series, bio-signals, audio files, images, and videos, etc. NeuroMem can benefit a wealth of AI applications requiring high speed and low-power classification along with life-long learning capabilities.

NEUROMEM KEY FEATURES

PARALLEL BROADCAST MODE

- As a new input pattern is broadcasted to the NeuroMem network, all the neurons update their distance simultaneously. They are ready to respond to a query as soon as the last component is received.

CHOICE OF CLASSIFIER: KNN OR RCE

- A **Restricted Coulomb Energy (RCE) classifier** uses **Radial Basis Function** as activation function. It is capable of complex nonlinear mappings and widely used for function approximation, time series prediction, and image recognition.
- A **K-Nearest Neighbor algorithm** (*K*NN) is a method for classifying objects based on closest models. The parallel architecture of the NeuroMem chip makes it the fastest candidate to retrieve the K closest neighbors of a vector among ANY number.

REACTIVE RECOGNITION WITH WINNER-TAKES-ALL

- The neurons reacting to an input pattern autonomously order themselves per decreasing confidence. This unique feature pertains to the parallel architecture of a NeuroMem network which allows a winner-takes-all among the reacting neurons.
- Neurons can report conflicting responses or cases of uncertainty.
- The absence of reacting neurons allows to detect anomaly or novelty which are essential for many applications in predictive maintenance, quality control and security.

FIXED LATENCY

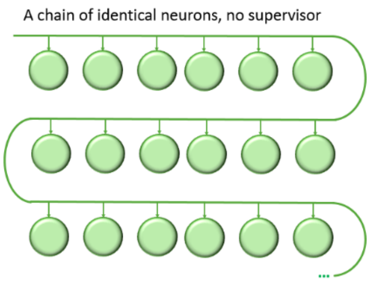
- The time necessary to read the response of the network to a new input pattern is independent of the number of committed neurons.
- At each query, only the neuron with the highest confidence responds and outputs its distance after 19 clock cycles, or its category in 37 clock cycles
- If an application requires an RBF classification reading the response of the N closest neurons if applicable with N=3, the categories of the 3 closest neurons are read in 3 * 37 clock cycles.
- If an application requires the use of KNN with K equal to 50, the distance values of the 50th closest neurons are read in 50 * 19 clock cycles.

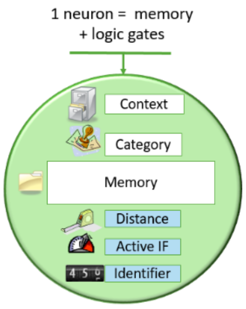
AUTONOMOUS MODEL GENERATOR

- The model generator built-in the NeuroMem chip makes it possible to learn examples in real-time when they drift from the knowledge residing in the committed neurons.
- Deduplication is intrinsic, since neurons only learn novelties
- The knowledge built by the neurons is cloneable since their content can be saved and restored.

MULTIPLE CONTEXTS OR NETWORK DYNAMIC SEGMENTATION

- The ability to assign the neurons to different contexts or sub-network allows building hierarchical or parallel decision trees between sub-networks. This leads to advanced machine learning with uncertainty management and hypothesis generation.





For a better understanding of the functionality and interactions of these modules, you can refer to the NeuroMem Technology Reference Guide.

NEUROMEM ICs

TECHNICAL COMPARISON CHART

ANN Attributes	CM1K	QuarkSE/Curie	NM500
Manufacturer	General Vision	Intel	General Vision/nepes
Neuron capacity	1,024	128	576
Memory capacity per neuron	256 bytes	128 bytes	256 bytes
Categories	15 bits	15 bits	15 bits
Distances	16 bits	16 bits	16 bits
Contexts	7 bits	7 bits	7 bits
Radial Basis Function (RBF)	X	X	X
K-Nearest Neighbor (KNN)	X	X	X
Distance Norm L1 (Manhattan)	X	X	X
Distance Norm LSUP	X	X	X
Daisy chaining	X		X
Other specifications	CM1K	QuarkSE/Curie	
Clock	27 Mhz (16 Mhz for chain of chips)	32 Mhz	37 Mhz (20 Mhz for chain of chips)
Package size	TQFP 16x16 mm	BGA 10x10 mm	WCSP64 4x4mm
Geometry	130 nm	22 nm	110 nm
Cost	$$	$	$
Other features	Recognition stage with digital input bus, I2C controller	CPU, Flash, RAM, Sensor, subsystem (DSP), UARTs, USB	

NEUROMEM IP

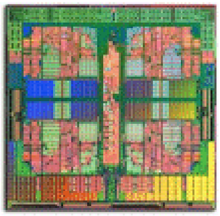
The NeuroMem IP is available for licensing in multiple formats and under different contractual terms.

- IP for Evaluation on FPGA
- IP for Production On FPGA
- IP for SoC design

WHY IS NEUROMEM DIFFERENT?

Traditional Von Neuman Architecture

Data is captured and put into a storage device – hard drives, flash memory, DRAM, SRAM, etc. The microprocessor is responsible for accessing and processing the data to determine a course of action.



Example of a multicore processor surrounded by DMA and SDRAM controllers

Serial process
Multi-core processors can operate in parallel, but with sequential access to memory

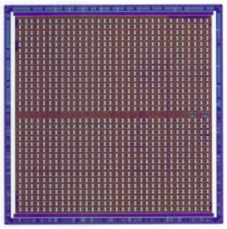
Costly high performance and scalability
The larger the data set, the longer the latency. Costs rise significantly with increased data sizes to achieve manageable latencies, especially in programming

Complex programming
Gains in performance come at increasing costs in programming complexity

High power consumption
Runs at Ghz, provisions for fans and heat sink are mandatory.

NeuroMem Technology

Data is stored through a learning process ensuring deduplication and novelty detection. Data is recognized within memory in a time independent from the size of the database.



NeuroMem CM1K chip showing 1024 identical neuromorphic memories, all interconnected through a patented architecture

Content addressability
Memory and processing logic reside in each memory cell. All cells are interconnected and work in parallel.

Intrinsic high performance and scalability
Deterministic latency regardless of size of data set. Scalability is possible thanks to a low and fixed number of I/Os independent of the number of cells.

Trainability
The nature of NeuroMem is to adapt or 'learn by examples. Duplicates are automatically prevented.

Low power consumption
Thanks to its parallel architecture NeuroMem can deliver GigaOps while running at clock frequencies in the order of Mhz

Contact Information

General Vision Inc. , 1150 Industrial Avenue, #A, Petaluma, CA 94952
www.general-vision.com, +1 707 765-6150