Pattern recognition is based on the comparison of an incoming vector with other reference vectors. This simple operation has become an everyday dilemma for the growing number of applications involved in data mining and search engines. The compromise between speed performance, power-consumption and cost is an endless challenge for board designers and it may be time to rethink native parallel architectures with new components such as the CogniMem neural network chip allowing infinite connectivity.

This paper demonstrates how CogniMem can outperform DSPs and multi-core processors when vector matching is involved. Indeed its parallel architecture allows calculating the distances between one input vector and any number N of models stored in its neurons in a constant amount of time. CogniMem stands for Cognitive Memory, but the chip can also be qualified as a **pattern recognition co-processor** and a very high-speed K-Nearest Neighbor chip.

NATIVE PARALLELISM IS KEY TO SPEED PERFORMANCE

Traditional computers have known limitations for pattern recognition. With respect to speed, the instructions are executed sequentially and the search time increases with the number of references to compare. Also the memory access through a single bus is a bottleneck. The new processors featuring dual and quad core CPUs are an improvement, but at the expense of simple data access and synchronization protocols. They must run at high clock frequency (100 MHz to 4GHz) leading to high power consumption and limiting miniaturization. With respect to cost, DSPs and CPUs can be expensive and also require access to RAM memory and non-volatile memory for the data storage.

Compared to the above, the CogniMem chip (CM1K) has a very simple and self-contained architecture consuming very little power: it is a chain of identical neurons operating in parallel. A neuron is a cognitive memory which can autonomously compare an incoming pattern with its reference pattern or prototype. This process takes a number of clock cycles equal to the length of the pattern and takes place in all the neurons at once. Upon termination, they communicate briefly with one another

(for 16 clock cycles) to identify which one of them holds the smallest distance value and therefore the pattern with the closest match.
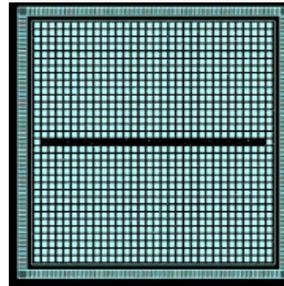
The parallel architecture of the CM1K can be seen in this plot of the ASIC: 1024 identical cells, or neurons, interconnected through a bus of 28 lines.

BENCHMARK DESCRIPTION

Definitions:
- V is an incoming pattern of length L
- P_k is a prototype stored in the memory of the neuron #k
- N is the number of committed neurons (holding a prototype)

The L1 or Manhattan distance between the vector V and N prototypes P_k is calculated with the following simple, yet very iterative, formula:

```
0:    for( n=0; n< N ; n++; MinDist = 0xFFFF)
         {
1:       for( i=0 ; i< L ;i++; L1Dist[n]=0)
2:         { if ( V[i] < N[n][i] ) then L1Dist[n] += N[n][i] - V[i] ; else
               L1Dist[n] += V[i] - N[n][i] ; }
3:       if (L1Dist[n] < MinDist) MinDist = L1Dist[n] ;
         }
```

The native parallelism of CM1K suppresses the need for the loop in Line 0 and all neurons execute Line 2 at the same time in L+1 clock cycles. Line 4 is a single command executed in 16 clock cycles for any value N. In conclusion, the number of clock cycles to find the shortest L1 distance between a vector V and N prototypes is 275, equivalent to 10.18 microseconds at 27 MHz. This performance is achieved with no need for an external controller or program by simply broadcasting the vector data to the neurons over the input data bus.

As a courtesy, Prof. Pierre Raymond from the Institut St Louis (ISL) has developed a program in Assembler optimized for a DSP Tiger-Shark TS101 from Analog

CogniMem Technologies, Inc.
81 Blue Ravine Road, Suite 240, Folsom, CA 95630
916.358.9483 F 916.358.9485

*www.cognimem.com*
May 2009
P a g e | 1

Devices and executing the L1 distance between a vector V and N prototypes. The DSP reaches a number of clock cycles equal to 284,706, or 0.95 milliseconds at 300 MHz.

The comparative results are presented below:

N=1024 L =256 bytes	CM1K	DSP Tiger SHARK
Clock frequency	27 Mhz	300 Mhz
Clock cycle (ns)	37	3.33
Number of instructions	275	N*278+34
Total cycles	275	284,706
Total time (usec)	10.18	949.02
Ratio		93

CM1K is 93 times faster than the DSP Tiger Shark when N=1024. Since the recognition time is independent from the number N, we can also say that CM1K will be twice as fast if N=2048, triple if N=3072 and so on.

With a recognition cycle of 10.18µsec at 27 MHz, CM1K can deliver 98,231 recognitions per second. The same speed performance on a DSP would require the execution of 278 instructions * 98,231 times per second or 27.3 giga instructions per second.

L1 distance calculation
CM1K speed / DSP speed
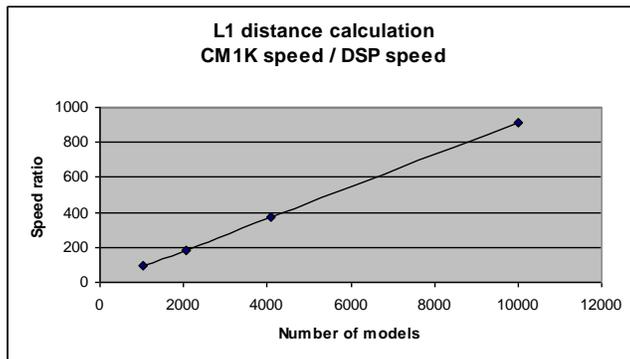
The above results are limited to a vector length L equal to 256 because the neurons of the CM1K have a memory of 256 bytes. If an application manipulates longer data sets with a resolution greater than 1 byte per component, there are alternatives to calculate their distance per segment taking advantage of a CM1K feature called the neuron context. For example, a vector longer than 256 can be divided into adjacent segments of length 256, with the last segment padded with zeros if needed. Similarly, if a vector has more than 8-bits of resolution, it can be segmented into upper and lower bytes. In these cases, the distance calculation of each segment must be preceded by the selection of a context value which takes one clock cycle. The category of the neurons can be used to re-assemble the segments belonging to the same vector and calculate the sum of

their distances. A second alternative more costly but delivering the best speed performance is the manufacturing of a new CogniMem chip with larger neuron cells (such as 512 * 12-bit components for example).

PARALLELISM INTER CHIPS FOR TRUE DATA MINING

Now that it is demonstrated that the CM1K recognition time depends only on the operating clock and not on the number of models stored in the neurons, let's introduce a second key feature of its parallel architecture: The CM1K chips can be interconnected to build a neural network of any capacity per increment of 1024 neurons and less than 1 watt per chip.

Network of N CM1K, or N*1024 neurons in parallel
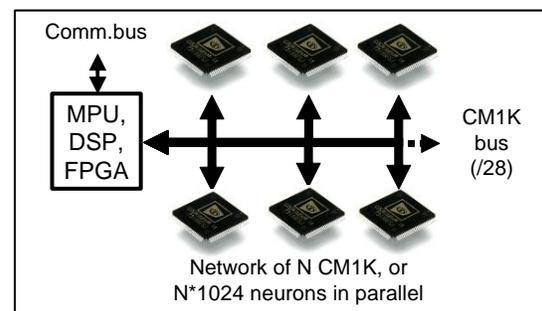
The CM1K control and data bus is simple and composed of only 28 lines. Adding more chips to increase the network size is totally transparent to the controller since the neurons include their own learning and recognition logic. The access time to read and write the neuron registers remains the same. Depending on the number of CM1Ks needed for a system design, it might be of interest to consider the manufacturing of a new chip. Indeed its architecture on silicon is easy to modify to add more neuron cells on a die.

PATTERN RECOGNITION BEYOND DISTANCE CALCULATION

For readers familiar with classification, the test bench described in this paper demonstrates that the CM1K is the fastest KNN (**K-Nearest Neighbor**) classifier available on silicon. The chip can also be used as an RBF classifier (**Radial Basis Function)** derived from the Restricted Coulomb Energy model. In this case, the neurons make use of a feature called their influence field which they autonomously adjust as new models are learned by other neurons. The RBF classifier introduces the notion of decision space with zones of unknown and

CogniMem Technologies, Inc.
81 Blue Ravine Road, Suite 240, Folsom, CA 95630
916.358.9483 F 916.358.9485

www.cognimem.com
May 2009
P a g e |2

uncertainty, which goes beyond the simple pattern matching and opens doors to pattern classification with confidence levels and hypothesis generation. In this mode, the distance and category of the neurons becomes important information. They are read respectively in 16 and 18 clock cycles or less than 2.5μsecs with a 27 MHz clock. With such performances, CM1K can be considered more than a pattern recognition co-processor but rather a **high-speed recognition engine for artificial intelligence**.

In 1988, DARPA published a "Neural Network Study" and following is one conclusive extract from the book: "As the number of neurons and interconnects increases with regard to the size of the application, the amount of memory required to store the interconnect values increases. If that memory cannot be stored locally with every processor, then the processor must access memory external to itself – and that slows the overall speed of the simulator." Twenty years later, CogniMem addresses this concern and it is not a software solution nor a hardware simulation, but a real and affordable chip.

CONCLUSION

The CogniMem chip is not a newcomer since its first production batch was released in January 2008. It is actually a descendant of the **ZISC (Zero Instruction Set Computer)** chip manufactured by IBM until 2001. To date, a few companies have integrated the CM1K chip in their product designs, mostly to add recognition capabilities to embedded sensor boards. The recently founded European Laboratory for Sensory Intelligence is presently attempting the design of a system featuring 100 CM1Ks totaling 102,400 neurons. With advances in semi-conductor technologies, CM1K might be the first enabler for affordable and portable data mining appliances or even USB keys!

ABOUT THE AUTHOR

Anne Menendez is one of two designers of the CogniMem chip and one of the founders of both General Vision Inc., a company applying the CogniMem technology of image recognition, and Cognimem Technologies, Inc.
Email: anne.menendez@cognimem.com

RELATED LINKS

http://www.cognimem.com
http://elsi.isl.eu

CogniMem Technologies, Inc.   www.cognimem.com
81 Blue Ravine Road, Suite 240, Folsom, CA 95630   May 2009
916.358.9483 F 916.358.9485   P a g e |**3**