Ultra-Low-Power Analog Associative Memory Core Using Flash-EEPROM-Based Programmable Capacitors

A. Kramer*, R. Canegallo, M. Chinosi, D. Doise,
G. Gozzini, P.L. Rolandi, M. Sabatini, P. Zabberoni

Neural Network Design Group - Central R&D
SGS-Thomson Microelectronics
Agrate Brianza (Milan), ITALY

Abstract

Analog techniques can lead to ultra-efficient computational systems when applied to the right applications. The problem of associative memory is well suited to array-based analog implementation. The architectures which result can be ultra efficient both in terms of high density and low power consumption. We have implemented a small (16x512) analog associative memory array which uses programmable nonlinear capacitors based on flash EEPROM technology for both analog storage and analog Manhattan Distance computation. The core circuit involved is based on only two of these novel devices. Preliminary results from this test circuit indicate that we can achieve a computing precision of more than 8 digital-equivalent bits in a chip which is capable of performing 128 Giga absolute-value-of-difference-accumulate operations per second at a power consumption of less than 150 mW. Performance of this level is more than an order of magnitude more efficient than the best low-power digital techniques and demonstrates the potential advantages analog implementation has to offer when applied to certain applications.

Introduction - Associative Memory

The function of an associative memory, or content-addressable memory, is more or less the inverse of that of a random access memory: when presented with a partial or complete data vector, the memory should return the row address of the internally stored data vector which best "matches" the input data vector. The matching function is typically a distance function; in standard digital implementations Hamming distance is usually used. Associative memory lends itself to array-based parallel implementation. A typical architecture consists of a 2-dimensional distance-computing / memory array, and several 1-dimensional arrays including an accumulator array for accumulating distances, a comparator array for finding the smallest distance, a priority encoder array for selecting rows one at a time, and a ROM array for presenting outputs [5].

* This work has been partially sponsored by U. C. Berkeley where Mr. Kramer is completing a Ph.D.

Analog Associative Memory

We are exploring an analog implementation of this type of architecture for Associative Memory. The result is an analog associative memory in which both stored memory rows and inputs consist of analog-valued vectors (5-bit equivalent precision). The goal is to achieve an ultra-efficient design in terms of both density and power consumption. Our target is an associative memory containing 4K lines of 64-dimensional memory vectors and capable of performing nearest neighbor match based on Manhattan Distance in less than 2uS at a power consumption of less than 150mW.

Computation of 4K 64-dimensional Manhattan Distances requires 256K 5-bit absolute-value-of-difference-accumulate computations, thus achieving a cycle time of 2uS requires performing 128G of these operations per second. Performing this much computation on a single chip at a power consumption of less than 150mW represents an increase in efficiency both in terms of density and power consumption of more than an order of magnitude over the best low-power digital techniques [1]. Practical realization of computing systems based on analog techniques may provide a viable alternative for ultra-efficient system design if the design generality lost can be justified by the added efficiency gained.

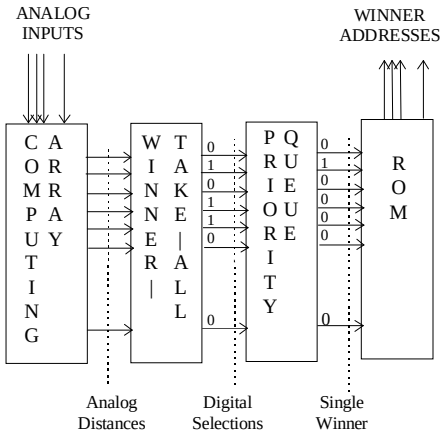

Fig 1:Block Diagram of Analog Associative Memory Architecture. ROM address of best-matching row (minimum Manhattan Distance) is output.

The block diagram of our analog architecture is shown in figure 1. It is a fairly direct mapping of the typical digital **a**chitecture described earlier and contains an analog memory / Manhattan Distance-computing array, a charge integrator (accumulator) array, a winner-take-all (comparator) array, and a digital output path consisting of a priority queue and a ROM. The core of the architecture is the analog memory/computation array which is based on novel programmable nonlinear capacitors implemented in flash-EEPROM technology. Each device in this array can store an analog value with 5-bit equivalent precision and, when given an analog input which is column-driven, can compute the absolute value of the difference between the stored value and the input value. The way this operation is performed is the central concept in our computing system and is presented in detail in the following sections.

A sketch of the data flow through the architecture is as follows: Analog inputs are presented to the computing array; each element of the array has a local analog value stored and computes the *absvaldiff* (absolute-value of difference) of the difference between its stored value and its input; the charge integrator array sums the individual absvaldiffs along each row into an analog Manhattan Distance; the winner-take-all array compares all of the computed analog distances and selects the one (or few) which has the smallest distance (best match); and the digital output path then prioritizes the winners selected and sends their row addresses off chip through the ROM. Control circuitry allows the winners selected to be disabled following output through the ROM; in this way a sorted list of rows in order of distance from the input can be read from the chip.

Analog Storage and Computation with Flash EEPROMs

As in a digital implementation, the power budget in an analog associative memory is dominated by the energy needed to compute the distances between each row of the memory array and the common input vector. The way this array is implemented and these distances are computed is the central novelty of our architecture and will be a focus of this paper. The distance computing array we present is highly efficient both in terms of density and power consumption: a circuit consisting of only two novel programmable nonlinear capacitors is able both to store an analog value and to perform an absvaldiff computation at an energy consumption of less than 1pJ. This is done by making use of floating-gate technology and the MOS physics controlling the channel charge of these devices.

The use of floating gate technology for efficient long-term analog storage is well explored, especially in neural network implementations [2,3,4]. Typically, these devices are employed only for storage, providing input to larger analog computational circuits such as multiplying amplifiers [2]. In this work we extend the use of these devices by using a single Flash-EEPROM based device for both analog storage and analog computation, resulting in a large increase in computational efficiency. This is done by making use of the MOS physics controlling the charge in the channel of a floating gate transistor to perform a nonlinear difference operation. The charge in the channel of an MOS transistor is nonlinear: below Vt the channel charge is effectively 0, while above Vt it is linear in (Vg - Vt) (fig 2). In the case of a floating gate device, this non-linearity is programmable. By storing one analog value as the threshold of a floating gate device, applying a second analog value on the gate of the device and measuring the channel charge with a charge integrator, it is possible to efficiently compute the amount by which the gate voltage exceeds the threshold voltage (fig 3).

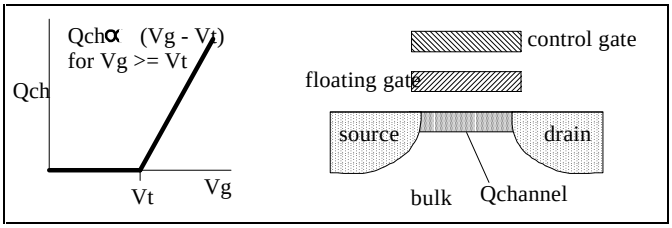

Fig 2: Channel charge versus gate voltage for a floating gate MOSFET. The curve shows the idealized function.

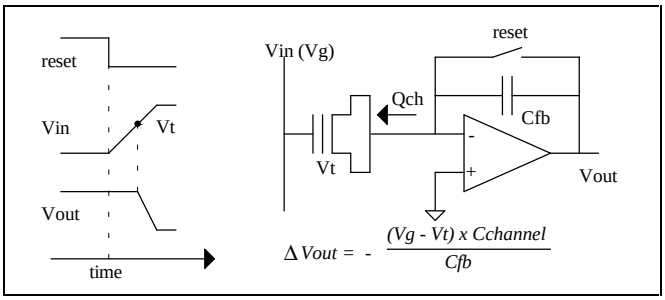

Fig 3: Channel charge to voltage conversion. After resetting the charge integrator, the input is applied

Distance Computation

The use of differential signaling allows a pair of these devices to be programmed so that their combined channel charge represents the absolute-value of the difference between two analog values (absvaldiff). The two devices have their threshold voltages programmed and their gate voltages applied in a differential manner (fig 4). The computation of the Manhattan Distance between two vectors requires the sum of the absolute values computed in each dimension, and conservation of charge allows many absvaldiff-computing circuits to be row-connected to a single common charge integrator which can then efficiently compute the Manhattan distance between the vector stored on the gates of the devices in the row, and the

vector applied on the gates. In addition, many such rows can be accessed by the columnar gates in parallel, allowing for a highly-efficient array-based architecture for the parallel computation of the Manhattan distances between a set of row-stored vectors and a single column-applied gate vector (fig 5).

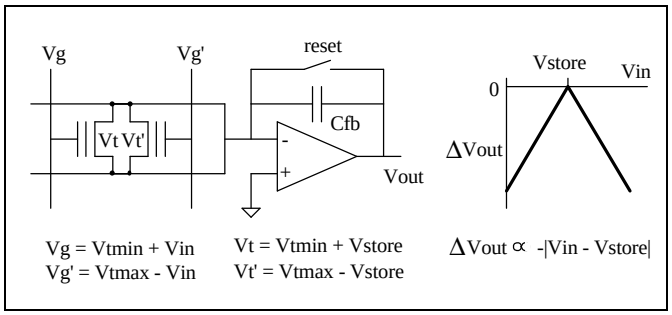

Fig. 4: Absolute Value of Difference Circuit. Using differential signaling for the input Vin and the stored value Vstore, (the value stored on the pair of floating gates), a second device can provide the other "half curve" needed for the absvaldiff function.

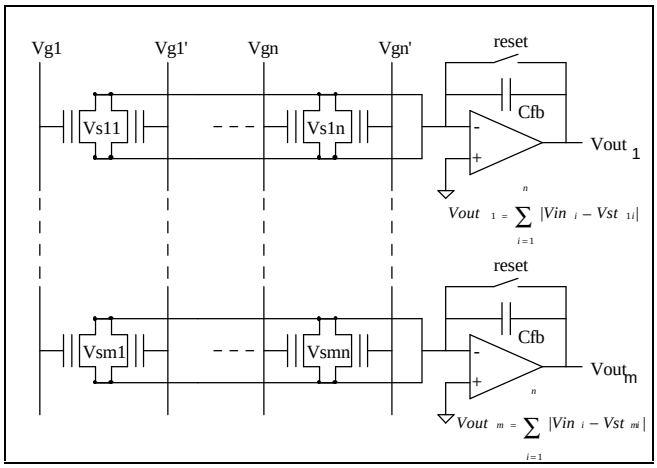

Fig 5: Architecture of the Manhattan Distance Computing Array. The inputs are applied simultaneously to the gates. Distances are computed in parallel.

CAPFLASH Device

The standard flash cell presents several limitations for use in our architecture. The first of these is that to prevent charge sharing between rows, we must have a cell with parallel source/drain access (compared to the standard cell layout which is parallel source/gate). The second limitation involves the parasitic capacitances of the standard device which modify the charge-domain properties of the idealized device in a way which adversely affects computational precision by introducing a large common-mode signal which must be compensated [7]. To overcome these limitations a new programmable nonlinear capacitor based on flash EEPROM technology has been designed, fabricated and tested [7]. This new structure, which we call the CAPFLASH device, allows parallel source/drain access and provides an effective channel capacitance which is almost a factor of 10 greater than that of the standard device while increasing cell size by less than 40% over that of the minimum parallel source-drain layout.

Analog programmability of the CAPFLASH device has been successfully tested, though at somewhat higher voltages than for "standard" devices (because device length is more than double) [7]. We have programmed these devices to an analog precision of better than 8 bits (8mV) confirming typical results [2,3]. In addition, charge retention has been characterized and the preliminary results are encouraging: charge loss in a maximally programmed (Vt=2V above virgin) or maximally erased (Vt=2V below virgin) device following corresponds to a retention of more than 5 bits for more than 10 years at 125°C [2,7]. At the lower temperatures where we expect to operate our circuit retention times should be even greater.

We have characterized the charge-domain properties of the CAPFLASH cell and the results from the new device are a much improved Q-V characteristic representing a signal-to-error ration which makes the device usable for our application [7]. We have also made a preliminary characterization of the mismatch in channel capacitance among 1024 CAPFLASH devices in an array of 512 rows and found the worst-case mismatch in capacitance to be less than 2%. This represents a precision of $5^1/_2$ bits and is encouraging for our goal of 5-bit overall precision. A circuit for the computation of Manhattan Distance based on 2 CAPFLASH and a charge integrator (fig 4) has been tested (fig 6). The results demonstrate the viability of using a single floating-gate device for both analog storage and ultra-efficient analog computation.

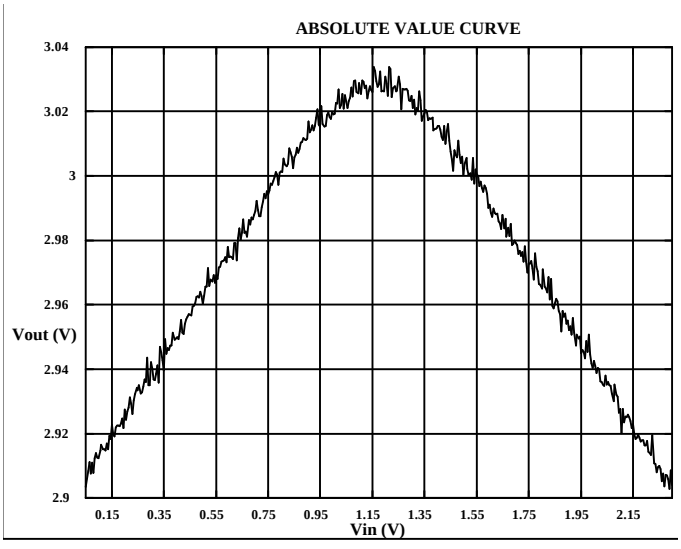

Fig 6: Computation of Absolute Value of Difference using two CAPFLASH devices (circuit shown in figure 4). Charge integrator output is shown.

Array core test circuit

We have realized in silicon a circuit to test the core array of our target architecture. The circuit implemented has the full input width of 64 input pairs (128 columns of CAPFLASH devices), though only 16 of the inputs are connected to pads (the rest are tied to ground), and contains 512 computing rows which consist of the CAPFLASH devices, charge integrator, winner-take-all, and output path. The circuit has been realized in an 0.7 um CMOS-flash process and overall size is 7mm x 3.5mm (fig 7).

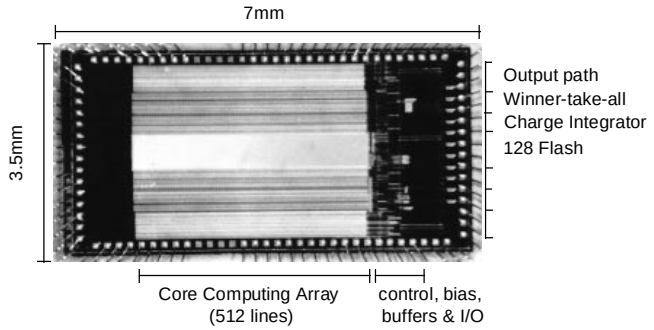

Fig 7: Photograph of 16x512 core array test chip. Flash is < 25% of array.

As can be seen from the photograph, the flash array is very dense: it consumes less than 25% of the core array area. This circuit has allowed us to test the full functionality and characterize the computing precision of our architecture. The circuit is fully functional as an analog associative memory and preliminaryresults indicate that the circuit is able to compute the absvaldiff function with a precision of 5 digital-equivalent bits, to accumulate these absolute difference into a Manhattan Distance measure with a precision of more than 8 bits, and to compare and select the smallest of these Manhattan Distances also with a precision of more than 8 bits. The details of how this was measured are the subject of the following sections.

Analog probing

An analog probe was included in the to allow analog testability. This analog probe consists of a common probe wire which has a CMOS switch connecting it to each individual charge integrator output and is also output off-chip through an analog buffer connected in follower configuration (fig 8). A shift register on the periphery of the array allows a digital control pattern to be shifted into the array. Every row of the computing array has a row (bit) in the shift register which controls the corresponding probe switch. By shifting a single "1" through the shift register, each line of the charge integrator array can be sequentially connected to the probe for measurement. "Scanning" the array in this way allows the relative outputs of the charge integrators to be read with high precision (better than 2mV).

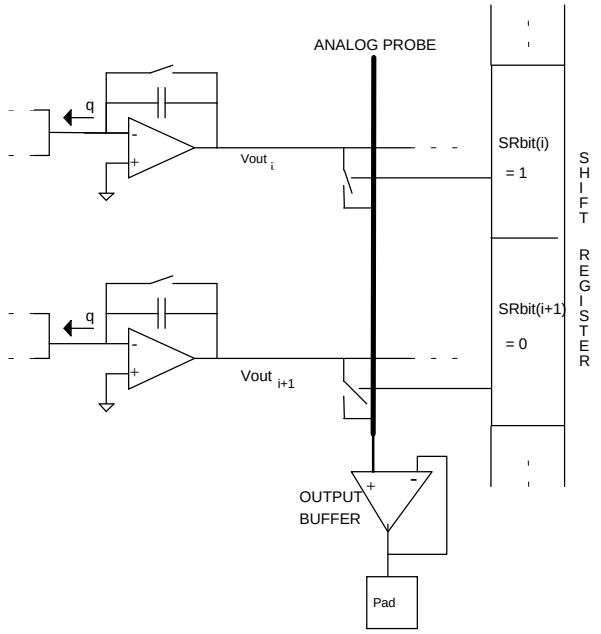

Fig 8: An analog probe was built into the array to allow charge integrator outputs to be scanned. A shift register provides the control to CMOS switches.

Charge Integrator

The charge integrator we have used is a low-power version of an offset-compensated instrumentation amplifier [6] connected in charge-integrator configuration. In the same reset cycle we both compensate the amplifier offset and reset the charge integrator. The precision of this reset operation is a source of additive noise in our system which must be small to achieve high computational precision. Using the analog probe, we have measured the output of a single charge integrator following each of 1000 reset cycles. The histogram of the 1000 reset values is plotted in figure 9. The absolute precision of the charge integrator reset is better than 8mV (sigma = 1.21mV).

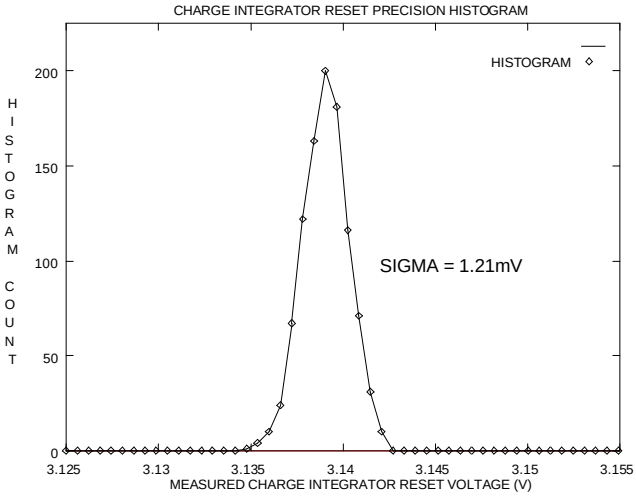

Fig 9: Histogram of 1000 charge integrator reset values (< 8mV noise)

For our computing system, relative reset precision is in fact more important than absolute reset precision (the winner-take-all circuit will self-compensate for any correlated offset common to all the charge integrators). Using the difference between two neighboring charge integrator outputs as our signal (measure-shift-measure-subtract), relative reset precision has been measured at less than 2mV. Relative reset precision among charge integrators at larger distances from each other in the array should be characterized as well, but these preliminary measurements indicate that the charge integrator reset operation reaches a relative precision within our target of 8mV.

Winner-Take-All

Winner-take-all is the operation of converting an analog-valued vector into a binary valued vector with one "1" (or at most a few) corresponding to the input vector element with the largest analog value. In our computing architecture, the winner-take-all serves as the parallel sense amplifier which finds the charge integrator with the highest output value, corresponding to the row with the smallest measured analog distance, and selects that row to be output from the chip. The circuit we have employed for the winner-take-all operation is the subthreshold analog circuit designed by Lazzaro and Mead [8]. Theory and operation of the circuit will not be described here. We have used the standard configuration except that we have employed two stages to increase gain (we want digital outputs and a single stage cannot guarantee this), we have used a simple parallel voltage-to-current converter at the input to each stage, and we have used large input transistors to improve matching and reduce offsets through the array.

Offsets in the Vts of the winner-take-all input transistors are another source of additive noise in our system, and more importantly, the computing precision of the winner-take-all circuit itself will critically limit the overall computing precision we achieve. We have characterized the winner-take-all circuit in our test array and preliminary results indicate a precision of better than 6mV over a 3V input range. This corresponds to a digital-equivalent computing precision of better than 8 bits. This measurement has been done by using the full functionality of the associative memory test circuit and involves several steps.

The first step is to program the flash array such that a given input to the array results in a desired analog vector to be present on the charge integrator outputs. To test 6mV precision, the analog vector chosen was a 6mV "staircase" in which every successive row has an output 6mV higher than the previous even row (the odd rows were unprogrammed). 10 staircase "steps" were programmed in this way; probe measurement of the resulting staircase is shown in figure 10.

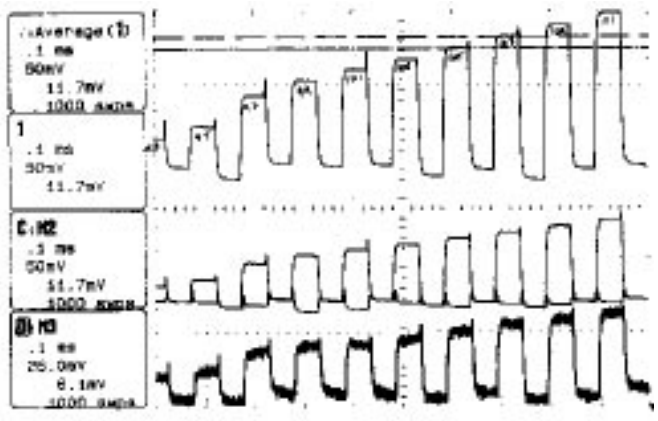

Fig 10: Oscilloscope traces showing 6 mV staircase. Bottom trace shows a single-sweep scope trace of a partial charge integrator array scan. Two middle traces are averaged sweeps showing charge integrator outputs before and after flash programming. Top trace is a zoom showing staircase to have 6mV steps.

The bottom trace is a single-sweep charge integrator scan. To get the trace, the charge integrators are first reset, the reference input (a 2-to-4 volt step to the programmed column) is then given to the flash array, and finally the scope is triggered as the analog probe is activated and scans the charge integrator outputs. Each square-wave-like transition in the trace represents a shift of the controlling shift register which connects the analog probe to the next successive charge integrator output. The next trace up is actually a pair of averaged tracks showing the same charge integrator scan before and after programming. The random charge integrator offsets before programming are caused by small differences in the virgin Vt of the unprogrammed flash column. The final (top) trace is a zoom of the averaged scan of the staircase showing the precision to be 6 mV.

The next step in the computing precision measurement involves using the full functionality of the associative memory to read a sequence of 10 winning rows to determine if the order of winners output from the circuit correspond correctly to programmed staircase. Digital control circuitry allows winning rows to be disabled so that a sequence of successive winners can be read in order of selection. Full functionality means:

1) Charge integrator reset

2) Reference input applied to flash array (staircase now present on charge-integrator outputs)

3) Winner-take-all enabled

4) ROM address of selected winning row read from chip

5) winning row disabled, allowing next successive winner ROM address to be output, etc...

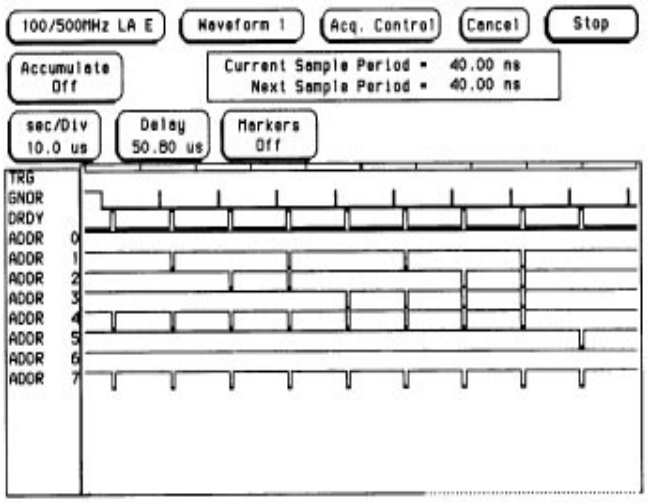

Fig 11: ROM output addresses read out following flash programming and input that produced charge integrator output levels of fig 10. Digital addresses descending by twos demonstrate full functionality of circuit at 6mV (> 8 bit) precision and at 100kHz (10 us per output).

Figure 11 shows the sequence of 10 ROM addresses read from the chip in this manner. The first address corresponds to the last row of the staircase (highest analog output) and the sequence of ROM addresses which follow counts down by twos and corresponds to descending down the analog staircase scanned in figure 10. This output pattern is stable over 1000 complete cycles and provides a preliminary indication that the winner-take-all circuit we have implemented is capable of resolving 6mV differences in its 512 inputs. As the charge integrator outputs have an output range of 3 volts, 6mV sensitivity by the winner-take-all circuit corresponds to a computing precision of more than 8 digital-equivalent bits.

Power, precision, and speed.

We have successfully tested the full functionality of this chip at speeds of 100kHz, the limitation imposed by our current test setup. Simulation results show correct operation at speeds in excess of 500kHz, but of course the effect of the faster operating frequency on computational precision must be determined. Power consumption of the chip is dominated by the charge integrator bias current. Each of the 512 charge integrators draws a bias current of 6uA, giving a total current of 3mA for this test circuit. Remaining analog components such as the winner take all and buffer circuits draw bias currents in the 10 uA range and do not contribute significantly to power consumption. Digital switching power is also negligible on this scale. The equivalent digital function performed by each row of the circuit consists of 64 5-bit absvaldiff and 8-bit accumulate operations. At the measured period of 10us, each 5-bit-absvaldiff-8-bit-accumulate operation is performed using less than 5 pJ of energy. Successfully testing the chip at 500kHz will mean we have achieved less than 1 pJ per each of these complex operations, as our simulations demonstrate is possible. While we have only 16 inputs actually connected to pads, we have programmed the 16 inputs in a way which allows the integrated charge to be roughly equivalent to what it will be when all 64 inputs are driven; we have seen that this does not change observed precision, power consumption or speed in any appreciable way. The preliminary results we have achieved with this small test circuit have been encouraging enough that we are now developing a full-size circuit containing a 64-input x 4k rows version of this architecture.

Conclusions

We have developed an analog associative memory based on the analog computation of Manhattan distance using pairs of novel programmable nonlinear capacitor devices for both analog storage and retention. A small test circuit containing 32k of these device pairs has been successfully tested at speeds of 100kHz, at a power consumption of 15mW. Preliminary measurements indicate that the circuit is capable of computing Manhattan Distances to a precision of 8 digital-equivalent bits based on 16 dimensional analog-valued vectors with an equivalent precision of 5-bits. This small test circuit has been measured at a throughput of nearly 1 Giga 8-bit accumulate-5-bit-absolute-value-of-difference operations per second. The full-size full-speed version of this architecture on which we are now working should be capable of performing 128 Giga ops of this type at an energy of less than 1 pJ per complex operation.

References:

(1) A. P. Chandrakasan, S. Sheng, R. W. Broderson, "Low-Power CMOS Digital Design," IEEE Journal of Solid State Electronics, Vol. 27, pp. 473-484, 1992.

(2) M. Holler, S. Tam, H. Castro, and R. Benson, "An Electrically Trainable Neural Network Chip (ETANN) with 1024 'Floating Gate' Synapses," in Proc. IJCNN, June 1989, pp 2.191-2.196.

(3) A. Kramer, V. Hu, C. K. Sin, B. Gupta, R. Chu, and P. K. Ko, "EEPROM Device as a Reconfigurable Analog Element for Neural Networks," IEDM Tech. Dig., pp. 10.3.1-10.3.4, Dec., 1989

(4) A. Kramer, C. K. Sin, R. Chu, and P. K. Ko, "Compact EEPROM-based Weight Functions," in *Neural Information Processing Systems 3*, R. P. Lippmann, J. E. Moody, and D. S. Touretzky, Eds., San Mateo CA: Morgan Kaufmann Publishers, Inc., 1991, p. 1001 - 1007.

(5) J. P. Wade and C. G. Sodini, "A Ternary Content Addressable Search Engine," IEEE Journal of Solid State Circuits, 1989, pp. 1003 - 1013.

(6) M. Degrauwe, E. Vittoz and I. Verbaouwhede, "A Micropower CMOS-Instrumentation Amplifier," IEEE Journal of Solid State Circuits, vol. SC-20, 1985, pp.805 - 807.

(7) A. Kramer et. al. "Flash-Based Programmable Nonlinear Capacitor for Switched-Capacitor Implementations of Neural Networks," IEDM Tech. Dig., pp. 17.6.1-17.6.4, Dec., 1994.

(8) Lazzaro, J., Ryckenbusch, S., Mahowald, M. A., and Mead, C. (1988). .Winner-take-all-networks of O(n) complexity. In Tourestzky,D. (ed), *Advances in Neural Network Information Processing Systems 1*. San Mateo, CA: Morgan Kaufmann Publishers, pp. 703-711