1.1 TMACS/mW Fine-Grained Stochastic Resonant Charge-Recycling Array Processor

Rafal Karakiewicz, *Senior Member, IEEE*, Roman Genov, *Member, IEEE*, and Gert Cauwenberghs, *Fellow, IEEE*

***Abstract*—We present a resonant adiabatic mixed-signal 128 $\times$ 256 array processor that achieves the energy efficiency of 1.1 TMACS (10^{12} multiply accumulates per second) per mW of power operating from a 1.6 V DC supply. The $1.9\ \mu\mathrm{m} \times 9\ \mu\mathrm{m}$ 3T NMOS unit cell with a single-wire pitch multiplexed bit/compute line provides charge-conserving 1b-1b multiplication and single-node charge-domain analog accumulation. A stochastic data modulation scheme minimizes on-chip capacitance variability maintaining sinusoidal clock oscillations near resonance.**

***Index Terms*—Adiabatic, charge-recycling, matrix-vector multiplication, mixed-signal and charge-mode.**

I. Introduction

LOW-POWER dissipation is a dominant objective in the design of integrated circuits for consumer devices. Despite dramatic improvements in energy efficiency, today's solid-state circuit technology still operates far from the fundamental kT/bit energy limits. CMOS process and voltage scaling have contributed remarkable savings in power, but are approaching physical limits as silicon technology enters the nano-regime and voltages approach thermal noise limits [1].

Array-based mixed-signal integrated circuits such as integrated sensor arrays, memories, and parallel processors share many unique architectural and circuit-level design solutions necessitated by their massive parallelism. These features can be exploited in order to reduce their power dissipation. A class of such circuits known as mixed-signal VLSI computing arrays take advantage of such features. Various implementations both with on-chip sensors such as computational CMOS imagers [2]–[4] and integrated acoustic array sensors [5], and as standalone sensory data processors [6]–[11] have been reported, often with unsurpassed energy efficiency characteristics.

Manuscript received September 15, 2010; revised January 23, 2011; accepted February 03, 2011. Date of publication February 10, 2011; date of current version February 08, 2012. This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). The associate editor coordinating the review of this paper and approving it for publication was Dr. Alexander Fish.

R. Karakiewicz was with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada. He is now with Vsemi, Toronto, ON M3C 3E5, Canada.

R. Genov is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: roman@eecg.utoronto.ca).

G. Cauwenberghs is with Department of Bioengineering, Jacobs School of Engineering, and Institute for Neural Computation, University of California San Diego, La Jolla, CA 92093 USA (e-mail: gert@ucsd.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSEN.2011.2113393

Analog or mixed-signal VLSI signal processing often allows for increases in integration density, computational throughput, and energy efficiency beyond what is achievable by digital processors, typically at the expense of reduced accuracy. High integration density is achieved by compact analog circuits such as those operating in charge domain. Computational throughput is enhanced by larger dimensions of computing arrays with compact cells and by the low-cost nature of some of analog operations such as zero-latency addition in charge domain. Energy efficiency is increased as clocking is reduced. Lower accuracy of computation is a result of nonidealities of analog components such as inherent nonlinearity and mismatches and is typically only weakly dependent on the dissipated power for a given implementation. A detailed quantitative analysis of the analog-versus-digital tradeoff is given in [12]. In applications such as sensory pattern recognition and sensory data classification a modest accuracy of under 8 bits is often sufficient.

In this paper, we combine adiabatic CMOS circuit approaches with stochastic encoding techniques to achieve record-level energy efficiency in massively parallel array computation. These design solutions are well suited to other array based integrated circuits.

Reversible and adiabatic computing have been introduced as a means to overcome the $CVdd^2$ dynamic energy dissipation in digital CMOS circuits [13], [14] and further in mixed-signal VLSI [15]. Adiabatic drivers slowly ramp the supply voltage from 0 V during the pull-up phase to reduce the voltage drop across the pull-up network. The voltage drop is made arbitrarily small by keeping the ramp period sufficiently longer than the time constant of the driver. For long ramp periods, the voltage across C is approximately equal to the supply ramp and the energy taken from the voltage source is $(1/2)CVdd^2$, the minimum required to charge C to Vdd. In the pull-down phase, the energy stored on C is discharged back into the supply voltage source by slowly ramping Vdd back to 0 V.

Linear voltage ramps generation to provide constant charging and discharging currents requires energy dissipation in the supply generator. This is often impractical for low-power applications and an oscillatory waveform, or hot-clock, from a resonator is typically used instead [14], [16]. The increased energy dissipation in the pull-up network, due to the nonoptimal sinusoidal shape, is offset by the low energy dissipation and simplicity of resonant hot-clock generation. Resonant adiabatic circuits recycle charge energy through transfer between electrostatic and inductive power in resonant LC circuits. In theory, the adiabatic energy consumption per unit computation approaches zero as the computation cycle extends to infinity, so that distributing the computation in parallel architecture (more but slower silicon) leads to net energy savings. In practice,

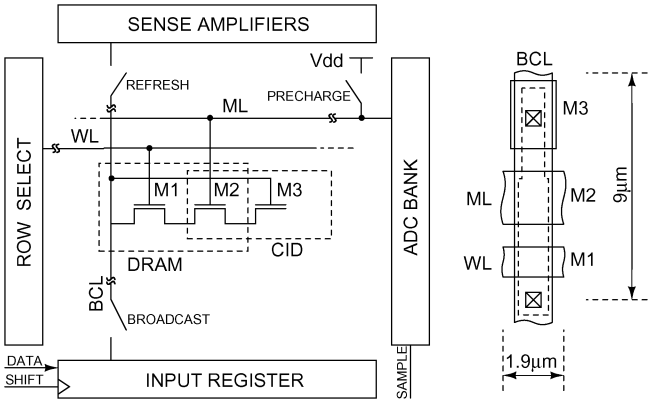

Fig. 1. CID/DRAM computational array, with 3-T unit cell and with peripheral functions (left). Compact, single-pitch layout of the cell (right).

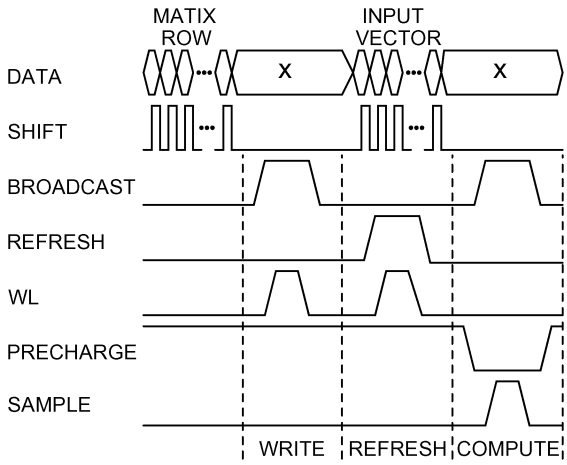

Fig. 2. Simplified timing diagram illustrating the three operation types of one row of the CID/DRAM computational array. The write operation takes place once for a given matrix to be stored. Many compute operations are performed for each refresh operation.

the energy efficiency is limited by the conversion efficiency between static and AC power in the resonant clock generator. Previously reported adiabatic processors achieve power gains of up to seven [16] over their nonadiabatic modes.

We present a 128 $\times$ 256-cell adiabatic charge-mode computing array achieving 85-fold improvement over the energy efficiency obtained when resonant drivers are replaced with CMOS drivers. The massive-parallel nature of the architecture yields high computational throughput at low clock frequency significantly reducing resistive losses. In order to maintain approximately constant resonant frequency, low load capacitance variability is achieved by a simple input data stochastic encoding and decoding scheme. The array yields twice the integration density and six times the energy efficiency of our previously reported prototype [17], [18]. Applications include pattern recognition [18], data compression [19], and CDMA matched filters [9], where 128 dimensions of the input vector are often sufficient. The array dimensions can be increased by tiling multiple arrays either in the column space or the row space or both on the same die, subject to conventional CMOS fabrication yield constraints [18]. This paper extends on an earlier report of the principle and demonstration in [20], and offers a more detailed analysis and experimental results characterizing the power saving performance. The rest of this paper is organized as follows. Section II describes the architecture and circuit implementation of the charge-mode computing array. In Section III, a resonant adiabatic clock generator is introduced in order to achieve high energy efficiency of the array-based computation. Limitations of the resonant clock generator are formulated and analyzed. Section IV describes a stochastic data modulation scheme that overcomes these limitations. Section V presents experimental results from the adiabatic array processor prototyped in a 0.35-μm CMOS technology.

II. Array Architecture and Circuit Implementation

The mixed-signal array computes linear transforms in the general form of vector-matrix multiplication (VMM) $\mathbf{Y} = \mathbf{W} \cdot \mathbf{X}$, with N-dimensional input vector $\mathbf{X}$, M-dimensional output vector $\mathbf{Y}$, and $M \times N$ matrix elements $\mathbf{W}$ ($N = 256$, $M = 128$). The array architecture and the cell circuit diagram are shown in Fig. 1 (left). The analog array is interfaced with a bank of on-chip row-parallel analog to digital converters (ADC) to provide convenient digital outputs as needed in some general purpose applications as well as in the array experimental testing and demonstration.

Similarly to the design in [18], the unit cell in the analog array shown in Fig. 1 (left) combines a charge injection device (CID) computational element [21], [22] with a dynamic random access memory (DRAM) storage element [23]. Each CID cell performs a one-quadrant binary-binary multiplication between the bit stored in it and the input bit on its bit/compute-line (BCL). In contrast to [18], the bit line and compute line are multiplexed on a single minimum wire-pitch (BCL) yielding a compact layout shown in Fig. 1 (right) with twice cell packing density of that reported in [17] and [18].

As depicted in the timing diagram in Fig. 2, during the *write* operation the data to be stored are broadcast from the input register onto the vertical (BCL), which extend across the array. A row to be written into is selected by activating its world-line (WL) turning transistor M1 on. The output match-line (ML) is held at Vdd during the *write* phase. The *refresh* operation is similar to the *write* operation, but in this case the charge in the DRAM cell is sensed and rewritten by a column-parallel sense amplifier. During the *compute* operation, the input data are broadcast on the BCLs and applied to the gate of transistor M3, while MLs, previously precharged to Vdd are now left floating.

Capacitive coupling of all cells in a single row into a single ML implements zero-latency analog accumulation along each row. An array of cells thus performs analog multiplication of a binary matrix with a binary vector.

During the write operation, the data to be stored are broadcast on the vertical folded multiplexed BCLs, and written into the row with the active WL, as depicted in Fig. 3. An active charge transfer from M2 to M3 can only occur if there is nonzero charge stored, and if the potential on the gate of M3 rises above that of M2, as shown in Fig. 3 (right). The cell performs nondestructive computation since the transferred charge is sensed capacitively on the MLs. Once a computation is performed the

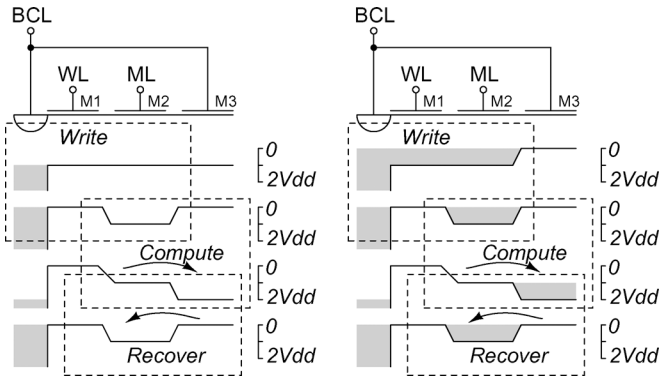

Fig. 3. Charge transfer diagram for *write* and *compute* operations with logic-**zero** (left) and logic-**one** (right) stored in the CID/DRAM cell.

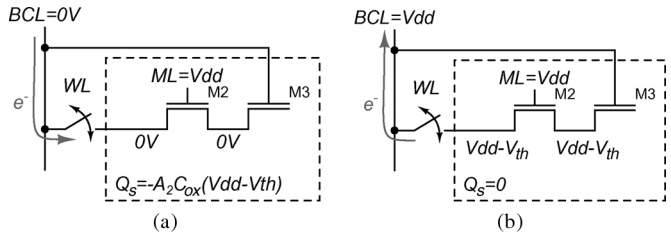

Fig. 4. Writing (a) logic-one and (b) logic-zero into the BCL multiplexed CID/DRAM cell.

charge is shifted back into the DRAM cell. The *write*, *refresh*, and *compute* operations are analyzed next in more detail.

A. Write

During *write* operation, the BCL is charged to 0 V (logic-one) or Vdd (logic-zero) and the ML is held at Vdd. When the WL goes high, electrons are stored under the gate of M2 if the BCL is at 0 V, as shown in Fig. 4(a), or are swept away if BCL is at Vdd, as shown in Fig. 4(b). When the WL goes low, the bit is stored as either trapped charge

$$Q_s = -A_2 C_{OX}(Vdd - V_{th}) \tag{1}$$

or lack of charge

$$Q_s = 0 \tag{2}$$

under M2, where A_2 is the gate area of M2, C_{OX} is the unit area gate capacitance, and V_{th} is the threshold voltage of a NMOS transistor. This *write* operation is functionally similar to that of the cell in [18]. The difference is that here when storing logic-zero in a cell, the gates of *all* M3 transistors in that column are also driven to Vdd. This has no effect on the final state of the cell being written into since either way all the charge is removed. In the remaining rows, the MLs can be left floating to ensure no charge is lost. The significant advantage of the design is that multiplexing the bit-line and the compute-line on a single BCL reduces the cell area by a factor of two.

B. Refresh

The stored bits in the cells are periodically refreshed to compensate for subthreshold leakage to the BCL, and junction leakage to the substrate. During *refresh*, the BCL is precharged to the trip-point, V_p, of the sense amplifier and the ML is held at Vdd, as shown in Fig. 5(a). When the access transistor, M1, is turned on by pulsing its WL, the charge stored in the cell is shared with the charge on the BCL, as shown in Fig. 5(b). In the case of a logic-one stored in the CID/DRAM cell, when charge is stored under the gate of M2, and ignoring body effect, the change in voltage on the BCL when M1 turns on is given by

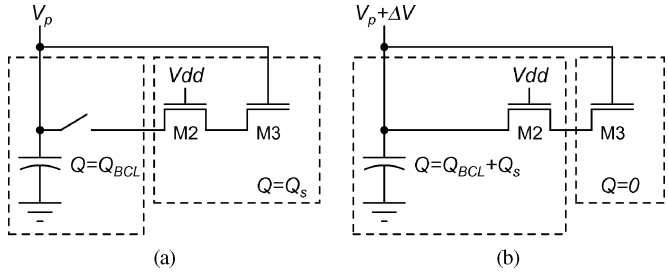

Fig. 5. During *refresh*, the BCL is first (a) precharged to the sense amplifier trip point before (b) the stored charge Q_s is shared with the BCL.

$$\Delta V = -\frac{A_2 C_{OX}}{C_{\mathrm{BCL}} + A_2 C_{OX}} V_p \tag{3}$$

where A_2 is the gate area of M2, V_p is the trip-point voltage of the sense amplifier, and C_{BCL} is the total capacitance of the BCL. In the case of a logic-zero stored in the CID/DRAM, when no charge is stored under the gate of M2, and ignoring the body effect, the voltage change on the BCL when M1 turns on is given by

$$\Delta V = +\frac{A_2 C_{OX}}{C_{\mathrm{BCL}} + A_2 C_{OX}}(Vdd - V_p - V_{th}). \tag{4}$$

In differential configuration, the voltage on the two BCLs connected to the sense amplifier change in opposite directions as given by (3) and (4). The differential voltage change on the two BCLs is thus

$$\Delta V = \frac{A_2 C_{OX}}{C_{\mathrm{BCL}} + A_2 C_{OX}}(Vdd - V_{th}). \tag{5}$$

The WL is de-asserted once the sense amplifier has replenished or removed the charge in the cell, restoring the bit.

The BCL capacitance C_{BCL} is approximately given by

$$C_{\mathrm{BCL}} = \frac{N}{2} A_J C_J(V_p) + N A_3 C_{OX} + C_{\mathrm{BCL},P} \tag{6}$$

where N is the number or rows in the array, A_J is the area of the junction connecting M1 to the BCL, $C_J(V_p)$ is the junction capacitance at V_p reverse bias voltage, A_3 is the gate area of M3, and $C_{\mathrm{BCL},P}$ is the metal line parasitic capacitance of BCL. The cost of increased cell packing density compared to a design with separate bit-line (BL) and compute-line (CL) [18] is a higher BL capacitance due to transistor M3. However, this additional capacitance can be made very small by choosing the trip-point voltage of the sense amplifier to be below the threshold voltage of transistor M3. In this case, when sensing is initiated transistor M3 is below threshold. There is no channel under its gate and its gate capacitance is the parallel plate capacitance between the gate and the substrate which is small. This results in only

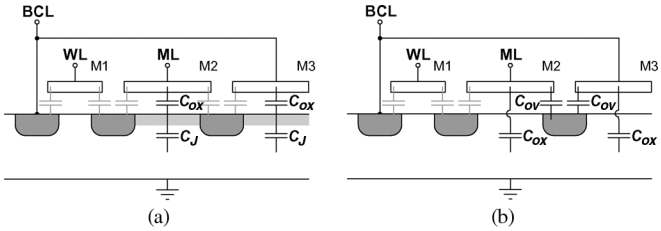

Fig. 6. Computing with (a) logic-one and (b) logic-zero stored in the multiplexed cell.

a minor increase in the total BCL capacitance above that in [18] or a conventional DRAM architecture. For a given technology, sense amplifier sensitivity, and size of array, the required area for the storage transistor M3 is obtained from (3) and (4).

C. Compute

The cell performs a one-quadrant binary-binary multiplication (AND operation) between its stored value and its BCL value, and accumulates the result on the ML across cells in the same row. The computation in each cell occurs as follows. First, the BCL is driven low and the ML is precharged to Vdd and left floating. Then, if the input to the cell is a logic-zero, the BCL remains low and the ML voltage remains unchanged. If the input to the cell is a logic-one, the BCL is driven to $2Vdd$ by peripheral energy recovery logic (ERL) drivers from a resonant supply generator as described in Section III.

If a logic-one is stored in the cell and the BCL is driven high, a channel is formed under both M2 and M3 once the BCL voltage reaches the threshold voltage of M3. Further increase in the BCL voltage is strongly coupled through the channels of M2 and M3 to the ML, as shown in Fig. 6(a). The change in the ML voltage in a single cell not connected to other cells is given approximately by

$$\Delta V_{\mathrm{ML}} = \frac{A_3 C_{OX}}{A_3 C_{OX} + A_3 C_J + A_2 C_J} \Delta V_{\mathrm{BCL}}. \qquad (7)$$

Given that ΔV_{BCL} is a digital signal, the voltage on the floating gate of transistor M2, when not connected to anything else, raises above Vdd by a fixed value depending on the capacitances of transistors M2 and M3 (which in turn depend on the presence of charge in the cell). When the gate of transistor M2 in a given cell is connected to the gates of M2 in all other cells in the same row, the change in its voltage depends on the number of active cells in that row. In fact, as all cells are capacitively coupled to ML, its voltage represents the normalized accumulation of all output voltages produced by each cell. The output of a row is a discrete analog quantity reflecting the number of active cells coupling into the ML of that row.

When a logic-zero is stored in the cell and the BCL is driven high, there is no channel present to capacitively couple the BCL to the ML, as shown in Fig. 6(b), and the ML voltage remains unchanged. In practice, the parasitic overlap capacitance, C_{OV}, causes some of the BCL voltage to feedthrough to the ML causing a change in the ML voltage given approximately by

$$\Delta V_{\mathrm{ML}} = \frac{W_2 C_{OV} || W_3 C_{OV}}{A_2 C_{OX} + W_2 C_{OV} || W_3 C_{OV}} \Delta V_{\mathrm{BCL}} \qquad (8)$$

where W_2 and W_3 are the widths of transistors M2 and M3, respectively. This parasitic feedthrough is minimized by careful layout or its effect is removed altogether by using differential computation where it appears as an input independent offset.

The added diffusion capacitance of all the access transistors is also charged during the *compute* operation if the corresponding BCL goes high. This has a negligible effect on the overall power dissipation due to the small area of the junction, which is shared with the adjacent cell, and the use of adiabatic computing which recycles energy back into the voltage supply source, as described next.

D. ADC

In order to validate the computing array functionality in general purpose applications, a 128-channel row-parallel ADC bank was implemented on the same chip. Each ADC channel is an 8-bit hybrid $\Delta\Sigma$-algorithmic ADC. Its architecture is detailed in [17]. In the algorithmic ADC mode its sampling rate is matched to the resonant frequency of the computing array (e.g., 20 ksps). In the oversampling ADC mode its effective sampling rate is equal to the resonant frequency of the computing array divided by the oversampling ratio. Combination of algorithmic ADC and first-order incremental $\Delta\Sigma$-modulated ADC techniques results in a sampling rate within this range [17].

III. Resonant Power Generation

Power in the array is dissipated by charging and discharging the total capacitance of all active BCLs. The capacitance of a BCL is approximately equal to the sum of capacitances of all charge injection device (CID) cells in the corresponding array column. CID cells perform reversible computation by shifting a charge between two potential wells without destructing it. This reversibility allows for recovery of the energy spent in a computation by using a resonant clock generator. Instead of conventional static CMOS logic, a resonant clock generator drives all active BCLs. It is implemented by coupling all active BCLs to an external inductor through a bank of energy recovery logic (ERL) drivers, as shown in Fig. 7(a). A simplified ERL driver is shown.

The circuit diagram of a modified ERL driver shown in Fig. 7(b) is utilized [17], [18]. When the input vector component bit, X_n, is logic-one, the corresponding BCL, BCL_n, is connected to the inductor through a pass gate. The maximum voltage on the inductor is $2Vdd$, while the logic-one level of X_n is Vdd. A cross-coupled PMOS transistor pair ensures that the pass gate is turned off completely when X_n is low.

To compensate for resistive losses in the tank in Fig. 7(a), the signal $pullHC$ is pulsed at the LC resonant frequency. The energy dissipation approaches zero when $pullHC$ is pulsed at the minima of the hot clock voltage, $V_{HC}(t)$. The capacitance of all active BCLs varies as a function of input data and stored data. Variations in this load capacitance C cause energy losses.

A simplified model of a variable-capacitance LC oscillator is shown in Fig. 8, where C has a mean value of C_{res} and resistive losses are assumed to be zero for simplicity. The tank capacitor C represents the capacitive load of all active CID

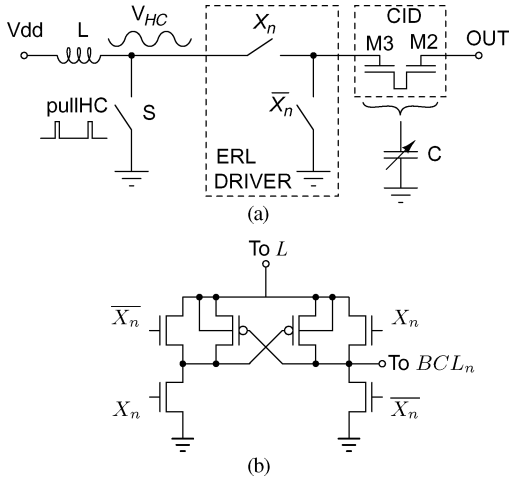

Fig. 7. (a) LC tank oscillator driving an array of charge-recycling CID cells. One cell is shown, with two charge-coupled MOS transistors M2 and M3 according to Figs. 1 and 3. (b) Double-range input-enabled energy recovery logic (ERL) driver.

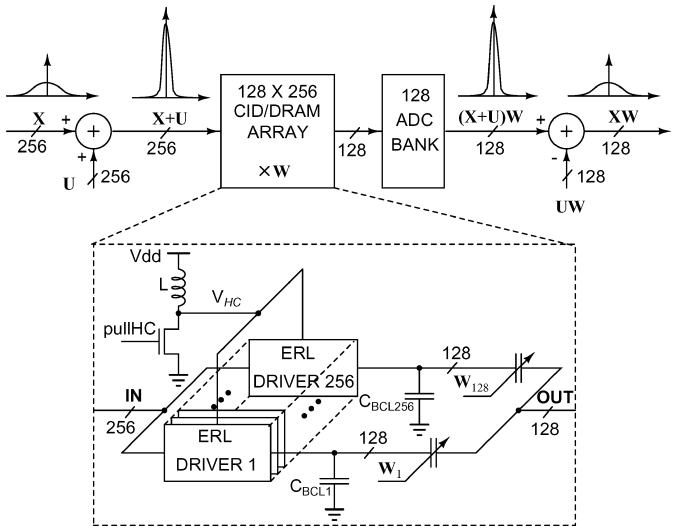

Fig. 9. Mixed-signal array and stochastic modulation architecture.

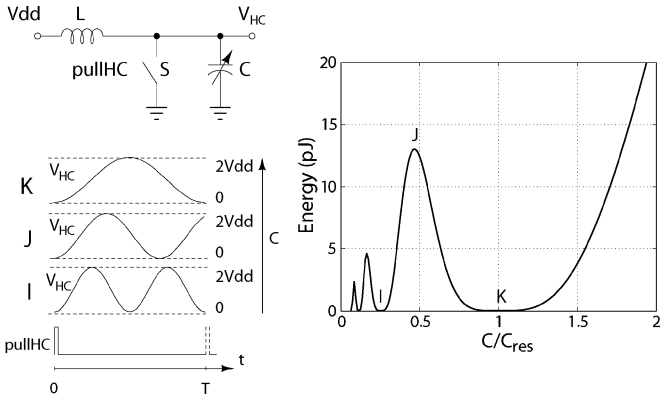

Fig. 8. Lossless LC oscillator with variable load capacitance C and examples of three corresponding $V_{HC}(t)$ waveforms (left). Energy dissipation in switch S of a lossless varying-capacitance LC oscillator (right).

cells and implies that they are being driven by the resonant clock. The signal $pullHC$ is pulsed at the LC mean-capacitance resonant frequency. The dependence of the LC tank instantaneous resonant frequency on variations in load capacitance C causes the power dissipation to fluctuate depending on the relative timing of switch S, as illustrated in the special cases I, J and K shown in Fig. 8. The energy dissipated in each computation is plotted on the right of Fig. 8. When $C = Cres$ (case K) or $C = Cres/4$ (case I), $V_{HC}(t)$ completes one or two full oscillation(s), respectively, before $pullHC$ is pulsed. At its resonance condition, the voltage difference across the switch S is minimum upon closure. The energy dissipation approaches zero over a wider range of capacitance at point K in Fig. 8.

A more detailed analysis of energy dissipation in the tank includes a careful consideration of the resistive losses [18]. It can be shown that theoretically, for a fixed tank capacitance, energy dissipation in the tank can be made arbitrarily small by increasing L. This corresponds to lowering the frequency of $pullHC$. In practice, the dynamic energy dissipation asymptotically approaches a finite value, limited by the quality of the inductor. Thus increasing L beyond a certain level may not be justifiable as it yields diminishing reduction in energy dissipation. This result also implies that increasing the frequency of $pullHC$ leads to an increase in energy dissipation due to resistive losses in the tank. This suits well with a parallel computing architecture performing very many operations in each clock cycle and maintaining a high overall throughput.

IV. Stochastic Data Modulation

Variations in the number of active inputs in the multiplication imply a variable capacitive load, leading to variations in power consumption. With appropriate preprocessing of data in a typical pattern recognition application, the bit-level statistics of the array input and stored matrix can be made sufficiently random such that the spread in number of actively charge-coupling cells in the array during each computation can be narrowed down significantly, so that the resonance condition can be maintained for minimum power loss in the clock generator [17], [18]. However, this balance of capacitive load is highly sensitive to the distribution of the data, which cannot always be controlled. To minimize resonant power supply energy dissipation due to data-dependent array capacitance variability, we investigate a coding scheme where the input data are stochastically modulated such that in every clock cycle half or near half of all bit/compute lines are active, as shown in Fig. 9.

A bit/compute line has approximately constant capacitance, $C_{\rm BCL}$, independent of the data stored in its DRAM cells, as transistor M3 is biased either in inversion or accumulation. For input bits that are equally probable zero or one, the number of compute lines that are connected to the inductor follows a Gaussian distribution with the mean halfway the range at $0.5N$, and variance $0.25N$, where N is the input dimension $(N = 256)$. Thus, for random inputs in high dimensions N, 95% of the time the array capacitance is typically (95% of cases) within a range $C_{\rm BCL}N^{1/2}$ (twice the standard deviation) from the mean, a factor $N^{1/2}$ closer than the full allowable range from 0 to $C_{\rm BCL}N$.

To randomize the bit-level distribution of a nonrandom digital input, a randomly chosen but fixed integer U_n in the

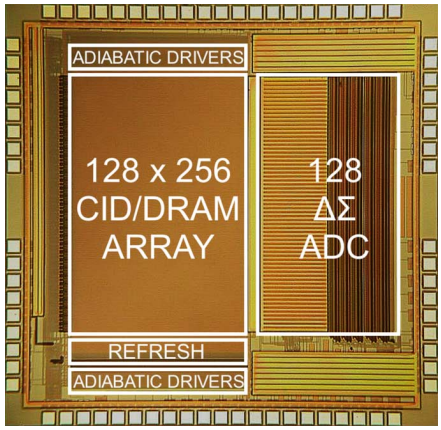

Fig. 10. Adiabatic VMM processor micrograph and floor plan. The die area is 2.2×2.2 mm^2 in 0.35-μm CMOS.

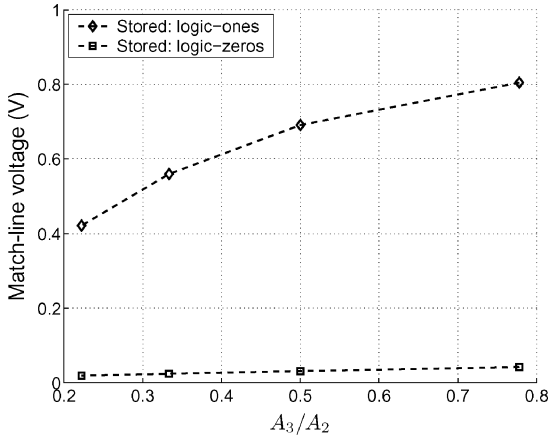

Fig. 11. Experimentally measured ML voltage as a function of A_3/A_2 for the two cases of logic-one and logic-zero stored in all CID/DRAM cells. The array is operated in differential mode with BCL voltage of 3.3 V.

range $[0, 2(N^{1/2}-1)]$ is added to each input component X_n [24]. While this adds an additional $\log_2 N/2$ bits to the input, it produces pseudo-random coded bits at the input of the array leading to a narrow load distribution, as depicted in Fig. 9. The modulated inner products in $\mathbf{X}$ are demodulated by digitally subtracting the inner products obtained for $\mathbf{X}+\mathbf{U}$ and $\mathbf{U}$. The random dithering vector $\mathbf{U}$ is chosen once, so its inner product with the templates is precomputed upon initializing or programming the array. The implementation cost is thus limited to one component-wise vector addition and one component-wise vector subtraction, achieved using $N+M$ one-bit full adder cells and one-bit registers, as well as $N+M$ external ROM cells to store $\mathbf{U}$ and $\mathbf{W}_m \cdot \mathbf{U}$, respectively. All of the overhead circuits scale linearly with the array dimensions and operate at the low clock frequency of the array (kHz-range) and thus have small power dissipation overhead. This stochastic modulation scheme also presents significant benefits in reduced requirements on linearity of analog accumulation and on resolution of row-parallel ADCs [24].

V. Experimental Results

The 0.35-μm CMOS integrated prototype of the adiabatic VMM processor shown in Fig. 10 contains 32 768 CID/DRAM cells and 128 row-parallel 8-bit $\Delta\Sigma$ algorithmic ADCs [23], as well as pipelined input shift registers, sense amplifiers, refresh logic, and scan-out logic. All supporting digital clocks and control signals are generated on-chip.

Various size cell configurations are included to experimentally validate the theoretical model discussed in Section II-C. Fig. 11 shows experimentally measured ML voltage as a function of A_3/A_2. The array is operated in the differential mode. In the first case, logic-one is stored in every CID/DRAM cell and the BCLs are driven to 3.3 V causing a large voltage change on the ML due to the strong capacitive coupling between the BCLs and MLs. In the second case, logic-zero is stored in every CID/DRAM cell and the BCLs are driven to 3.3 V causing only a small voltage change on the ML due to the weak parasitic coupling between the BCLs and MLs. In both cases, as the ratio A_3/A_2 increases so does the ML voltage when the BCLs are driven high as predicted by (7) and (8). The ML voltage can be further increased by operating the array in the single-ended

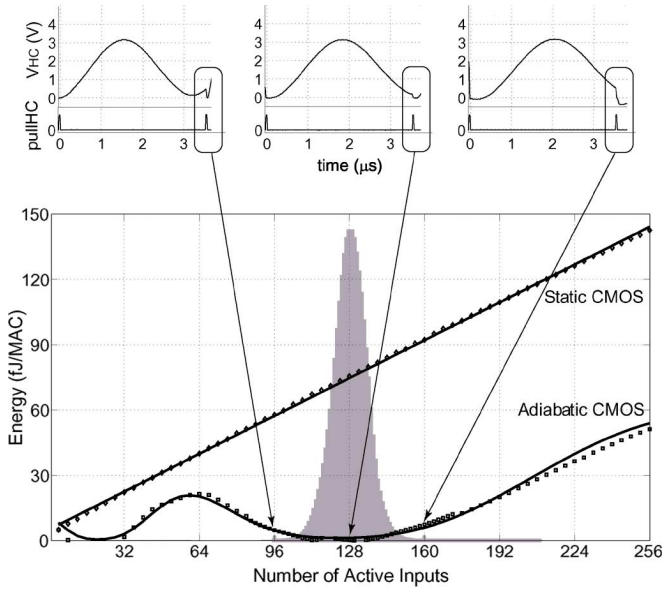

Fig. 12. Experimentally measured (dotted line) and theoretical (solid line) array dynamic energy dissipation as a function of input data statistics in adiabatic resonant mode and static CMOS mode. Three corresponding experimentally measured hot clock waveforms are shown. The probability density function of modulated input data is shown in gray.

mode instead of differentially, doubling the number of cells coupling into the ML. Also, ERL drivers can be used to approximately double the ML voltage by driving the peak BCL voltage to $2Vdd$.

Fig. 12 shows experimentally measured energy consumption per computation cycle of the CID/DRAM array in the static CMOS driver mode and in the adiabatic mode as a function of the number of active inputs (number of logic-"1" bits in the input vector). The adiabatic power measured through the Vdd pin includes all losses in the inductor, energy recovery logic (ERL) drivers and CID/DRAM array. Worst-case uniformly distributed 8-bit data is stochastically modulated producing 12-bit pseudo-random inputs to maintain array capacitance within $C_{\mathrm{BCL}}N^{1/2}$ of its mean for 95 percent of the data. As shown in Fig. 12, this 95 percent interval is positioned to coincide with a broad region of resonance, minimizing energy

TABLE I
EXPERIMENTALLY MEASURED ARRAY CHARACTERISTICS

Technology	0.35 μm CMOS
Supply Voltage Vdd	1.6V
Die Area	2.2×2.2 mm^2
Array Area	1.34×0.94 mm^2
CID/DRAM Cell Area	9×1.9 μm^2
CID/DRAM Cell Count	32,768
Throughput	640 MMACS at 20 kHz
Energy Efficiency	12.9 GMACS/mW Static CMOS 1.1 TMACS/mW Adiabatic CMOS (worst-case modulated data)
Output Resolution	8 bits

loss across S at activation of $pullHC$. This broad tuning of resonance and load balancing yields an improvement in energy efficiency from 12.9 GMACS/mW in static CMOS mode to 1.1 TMACS/mW in adiabatic mode, independent of the input distribution. This corresponds to 6.4×10^8 multiply accumulates per second at 0.59 μW power from 1.6 V DC supply. A second region of resonance in Fig. 12 should ideally occur when the number of active inputs is 32. This corresponds to point I in Fig. 8. In practice, the second minimum occurs at a number of inputs less than 32. This is likely due to parasitic capacitances in the LC tank which are more prominent when the tank capacitance is small. A summary of the computing array characteristics is given in Table I.

The stated energy efficiency figures do not include the power dissipated in the ADCs and other peripheral circuits. Common digital peripheral circuits such as shift registers can be efficiently implemented using conventional digital adiabatic design techniques. In the present prototype, the ADCs were included for convenience of characterization. Applications in pattern classification, such as vector quantization or nearest neighbor classification, call for winner-take-all (WTA) or rank-ordered selection of best template matches. WTA selection is efficiently implemented using a cascade of comparators potentially operating adiabatically in the charge domain [25], or using other low-power charge-based circuits [26]. In non-WTA classifiers such as support vector machines (SVM) computing the kernel and the linear weighting may also be implemented using similar adiabatic circuits.

By virtue of the parallel architecture and the favorable properties of the stochastic modulation for large N, this adiabatic mixed-signal computing technology scales to large throughput at a moderate and linear cost in power. Interestingly, the TMACS/mW performance of the array exceeds the power efficiency of the human brain which attains roughly 10^{16} synaptic connections per second at 15 W of power, although the MACS measure provides a very incomplete metric of neural computation.

VI. CONCLUSION

An adiabatic array processor for general purpose VMM operations has been presented. A single wire pitch multiplexed bit/compute line cell design yields twice cell packing density. A simple stochastic scheme allows to minimize losses in the resonant power generator due to capacitive load variability. Input data are modulated such that the array capacitance follows a small-variance Gaussian distribution. The overhead of modulating the inputs and demodulating the outputs is out-weighted by a boost in energy efficiency. The 0.35-μm CMOS adiabatic computational array integrated prototype delivers 1.1 TMACS for every mW of power for worst case data distribution. This constitutes a 85-fold improvement in energy efficiency in its resonant adiabatic mode, compared to its already intrinsically low energy efficiency when operating in the static CMOS driver mode.

REFERENCES

[1] A. Wang and A. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, Jan. 2005.

[2] A. Olyaei and R. Genov, "Focal-plane spatially oversampling CMOS image compression sensor," *IEEE Trans. Circuits and Systems I*, vol. 49, no. 1, pp. 26–34, 2007.

[3] A. Nilchi, J. Aziz, and R. Genov, "Focal-plane algorithmically-multiplying CMOS computational image sensor," *IEEE J. Solid-State Circuits*, vol. 44, no. 6, pp. 1829–1839, Jun. 2009.

[4] Y. M. Chi, A. Abbas, S. Chakrabartty, and G. Cauwenberghs, "An active pixel CMOS separable transform image sensor," in *Proc. IEEE Int. Symp. Circuits Syst.*, Taipei, Taiwan, May 24–27, 2009, pp. 1281–1284.

[5] M. Stanacevic and G. Cauwenberghs, "Micropower gradient flow acoustic localizer," *IEEE Trans. Circuits Syst. I: Regular Papers*, vol. 52, no. 10, pp. 2148–2157, Oct. 2005.

[6] R. Genov and G. Cauwenberghs, "Charge-mode parallel architecture for matrix-vector multiplication," *IEEE Trans. Circuits Syst. II: Analog and Digital Signal Process.*, vol. 48, no. 10, pp. 930–936, Oct. 2001.

[7] R. Genov and G. Cauwenberghs, "Kerneltron: Support vector machine in silicon," *IEEE Trans. Neural Networks*, vol. 14, no. 5, pp. 1426–1434, Sep. 2003.

[8] A. Kramer, "Array-based analog computation," *IEEE Micro*, vol. 16, no. 5, pp. 40–49, Oct. 1996.

[9] T. Yamasaki, T. Nakayama, and T. Shibata, "A low-power and compact CDMA matched filter based on switched-current technology," *IEEE J. Solid-State Circuits*, vol. 40, no. 4, pp. 926–932, Apr. 2005.

[10] A. Basu, C. M. Twigg, S. Brink, P. Hasler, C. Petre, S. Ramakrishnan, S. Koziol, and C. Schlottmann, "A floating-gate-based field-programmable analog array," *IEEE J. Solid-State Circuits*, vol. 45, no. 9, pp. 1781–1794, Sep. 2010.

[11] S. Chakrabartty and G. Cauwenberghs, "Sub-microwatt analog VLSI trainable pattern classifier," *IEEE J. Solid-State Circuits*, vol. 42, no. 5, pp. 1169–1179, 2007.

[12] R. Sarpeshkar, "Analog versus digital: Extrapolating from electronics to neurobiology," *Neural Computation*, vol. 10, pp. 1601–1638, 1998.

[13] H. Yamauchi, H. Akamatsu, and T. Fujita, "Anasymptotically zero power charge-recycling bus architecture for battery-operated ultrahigh data rate ULSI's," *IEEE JSSC*, vol. 30, no. 4, pp. 423–431, Apr. 1995.

[14] Y. Moon and D. K. Jeong, "An efficient charge recovery logic circuit," *IEEE J. Solid-State Circuits*, vol. 31, no. 4, pp. 514–522, Apr. 1996.

[15] J. Ammer, M. Bolotski, P. Alvelda, and T. F. Knight, Jr, "160 × 120 pixel liquid-crystal-on-silicon microdisplay with an adiabatic DAC," in *Proc. IEEE Solid-State Circuits Conf.*, 1999, pp. 212–213.

[16] W. Athas, N. Tzartzanis, W. Mao, L. Peterson, R. Lal, K. Chong, J.-S. Moon, L. Svensson, and M. Bolotski, "The design and implementation of a low-power clock-powered microprocessor," *IEEE J. Solid-State Circuits*, vol. 35, no. 11, pp. 1561–1570, Nov. 2000.

[17] R. Karakiewicz, R. Genov, A. Abbas, and G. Cauwenberghs, "175 GMACS/mW charge-mode adiabatic mixed-signal array processor," in *Proc. IEEE 2006 Symp. VLSI Circuits*, Honolulu, HI, Jun. 13–17, 2006, pp. 100–101.

[18] R. Karakiewicz, R. Genov, and G. Cauwenberghs, "480-GMACS/mW resonant adiabatic mixed-signal processor array for charge-based pattern recognition," *IEEE J. Solid-State Circuits*, vol. 42, no. 11, pp. 2573–2584, Nov. 2007.

[19] A. Nakada, T. Shibata, M. Konda, T. Morimoto, and T. Ohmi, "A fully parallel vector-quantization processor for real-time motion-picture compression," *IEEE J. Solid-State Circuits*, vol. 34, no. 6, pp. 822–830, Jun. 1999.

[20] R. Karakiewicz, R. Genov, and G. Cauwenberghs, "1.1 TMACS/mW load-balanced resonant charge-recycling array processor," in *Proc. IEEE Custom Integr. Circuits Conf.*, Sep. 2007, pp. 603–606.
[21] C. Neugebauer and A. Yariv, "A parallel analog CCD/CMOS neural network IC," in *Proc. IEEE Int. Joint Conf. Neural Networks*, Seattle, WA, 1991, vol. 1, pp. 447–451.
[22] V. Pedroni, A. Agranat, C. Neugebauer, and A. Yariv, "Pattern matching and parallel processing with CCD technology," in *Proc. IEEE Int. Joint Conf. Neural Networks*, Jun. 1, 1992, vol. 3, pp. 620–623.
[23] R. Genov, G. Cauwenberghs, G. Mulliken, and F. Adil, "A 5.9 mW 6.5 GMACS CID/DRAM array processor," in *Proc. Eur. Solid-State Circuits Conf.*, Florence, Italy, Sep. 24–26, 2002.
[24] R. Genov and G. Cauwenberghs, "Stochastic Mixed-Signal VLSI Architecture for High-Dimensional Kernel Machines," in *Advanced Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2002, vol. 14.
[25] K. Kotani and T. Ohmi, "Feedback charge-transfer comparator with zero static power," in *Proc. IEEE Solid-State Circuits Conf.*, Feb. 1999, pp. 328–329.
[26] R. Genov and G. Cauwenberghs, "Dynamic MOS sigmoid array folding analog-to-digital conversion," *IEEE Trans. Circuits and Syst. I*, vol. 51, no. 1, pp. 182–186, Jan. 2004.

Rafal Karakiewicz (SM'03) received the B.A.Sc. and the M.A.Sc. degrees in electrical engineering from the University of Toronto, Toronto, ON, Canada, in 2003 and 2006 respectively.

He is currently a Design Engineer at Synopsis, Canada.

Roman Genov (S'96–M'02) received the B.S. degree (Hon) in electrical engineering from the Rochester Institute of Technology, Rochester, NY, in 1996, and the M.S. and Ph.D. degrees in electrical and computer engineering from The Johns Hopkins University, Baltimore, MD, in 1998 and 2002, respectively.

He held engineering positions at Atmel Corporation, Columbia, MD, in 1995 and Xerox Corporation, Rochester, in 1996. He was a Visiting Researcher in the Laboratory of Intelligent Systems, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, in 1998, and in the Center for Biological and Computational Learning at the Massachusetts Institute of Technology, Cambridge, in 1999. Currently, he is an Associate Professor in the Department of Electrical and Computer Engineering at the University of Toronto, Toronto, ON, Canada.

Dr. Genov received the Canadian Institutes of Health Research (CIHR) Next Generation Award in 2005, the Brian L. Barge Award for Excellence in Microsystems Integration in 2008, as well as the Best Paper Award on sensors and Best Student Paper Award at the IEEE International Symposium on Circuits and Systems in 2009. He is an Associate Editor of the IEEE TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS and IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS II: EXPRESS BRIEFS.

Gert Cauwenberghs (S'89–M'94–SM'03–F'11) received the M.Eng. degree in applied physics from the University of Brussels, Brussels, Belgium, in 1988, and the M.S. and Ph.D. degrees in electrical engineering from the California Institute of Technology, Pasadena, in 1989 and 1994, respectively.

He is a Professor of Bioengineering and Biology at the University of California San Diego, and Co-Director of the Institute for Neural Computation, La Jolla. Previously, he held positions as Professor of Electrical and Computer Engineering at The Johns Hopkins University, Baltimore, MD, and as a Visiting Professor of Brain and Cognitive Science at the Massachusetts Institute of Technology, Cambridge.

Prof. Cauwenberghs is a Francqui Fellow of the Belgian American Educational Foundation, and received the National Science Foundation Career Award in 1997, the Office of Naval Research Young Investigator Award in 1999, and the Presidential Early Career Award for Scientists and Engineers in 2000. He serves on the Technical Advisory Board of GTronix, Inc., Fremont CA. He was a Distinguished Lecturer of the IEEE Circuits and Systems Society in 2003–2004, and chaired its Analog Signal Processing Technical Committee in 2001–2002. He currently serves as Editor-in-Chief of the TRANSACTIONS ON BIOMEDICAL CIRCUITS AND SYSTEMS, Associate Editor for the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS I, the IEEE TRANSACTIONS ON NEURAL SYSTEMS AND REHABILITATION ENGINEERING, and the IEEE SENSORS JOURNAL.