1.1 TMACS/mW Load-Balanced Resonant Charge-Recycling Array Processor

Rafal Karakiewicz, Roman Genov
Department of Electrical and Computer Engineering
University of Toronto
Toronto, ON M5S 3G4 Canada
Email: {rafal,roman}@eecg.toronto.edu

Gert Cauwenberghs
Division of Biological Sciences
University of California San Diego
La Jolla, CA 92093 USA
Email: gert@ucsd.edu

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

I. Introduction

Despite dramatic improvements in energy efficiency, today's solid-state circuit computing technology still operates far from fundamental kT/bit energy limits. In this paper we combine adiabatic CMOS circuit approaches with stochastic encoding techniques to achieve record-level energy efficiency in massively parallel array computation. 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].

Adiabatic and reversible computing have been introduced as a means to overcome the $CVdd^2$ dynamic energy dissipation in digital CMOS circuits [2], [3] and further in mixed-signal VLSI [4]. Adiabatic drivers slowly ramp the supply voltage from 0V 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 $\frac{1}{2}CVdd^2$, 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 0V.

Generating the linear voltage ramps 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 [3], [5]. The increased energy dissipation in the pull-up network, due to the non-optimal 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, 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 [5] over their non-adiabatic modes.

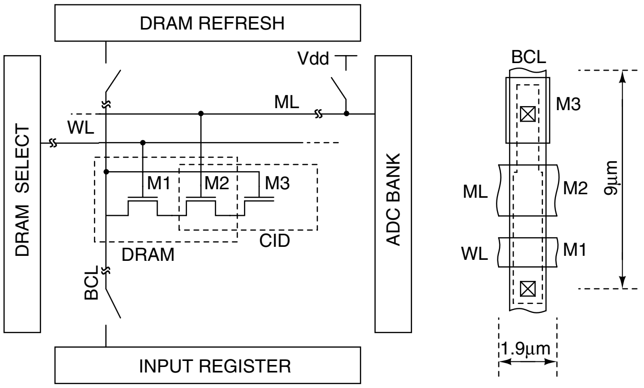

Fig. 1. CID/DRAM computational array, with 3-T unit cell and with peripheral functions (left). Compact, single-pitch layout of the cell (right).

We report an 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 [6]. Applications include pattern recognition [6], data compression [7] and CDMA matched filters [8].

II. 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

1-4244-1623-X/07/$25.00 ©2007 IEEE

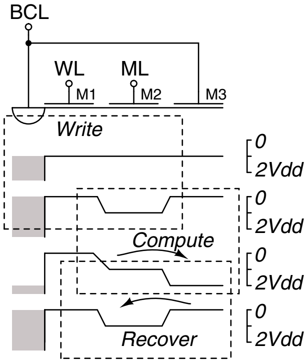

Fig. 2. Charge transfer diagram for *write* and *compute* operations with logic-one (left) and logic-zero (right) stored in the CID/DRAM cell.

diagram are shown in Fig. 1 (left). The unit cell in the analog array combines a charge injection device (CID) computational element with a DRAM storage element. Each cell performs a one-quadrant binary-binary multiplication between the bit stored in the DRAM cell and the input bit on its bit/compute-line (BCL). 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 [6].

Capacitive coupling of all cells in a single row into a single match-line (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. The analog output vector is quantized by row-parallel analog to digital converters (ADCs) to provide convenient digital VMM result, in a mixed-signal architecture that directly extends to multi-bit data [9].

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 word-line (WL) as depicted in Fig. 2. An active charge transfer from M2 to M3 can only occur if there is non-zero charge stored, and if the potential on the gate of M3 rises above that of M2 as shown in Fig. 2 (right). The cell performs non-destructive computation since the transferred charge is sensed capacitively on the MLs. Once a computation is performed the charge is shifted back into the dynamic random access memory (DRAM) cell.

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 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. 3 (a). A simplified ERL driver is shown.

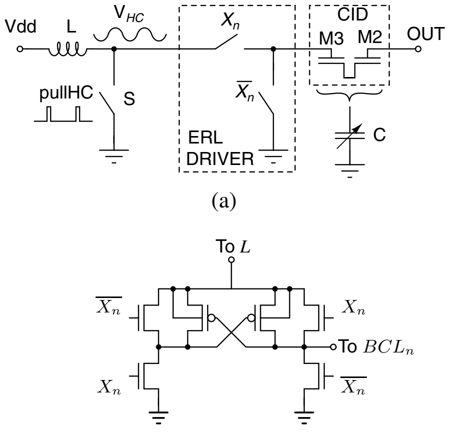

Fig. 3. (a), *LC* tank oscillator driving an array of charge-recycling CID cells. One cell is shown, with two charge-coupled MOS transistors M2 and M3 according to Fig. 1 and Fig. 2. (b), double-range input-enabled energy recovery logic (ERL) driver.

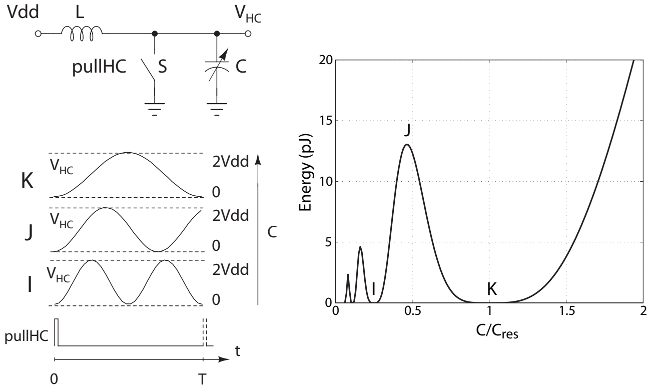

Fig. 4. Lossless *LC* oscillator with variable load capacitance *C* and examples of three corresponding $V_{HC}(t)$ waveforms (left). Energy dissipation in switch *S* of a lossless varying-capacitance *LC* oscillator (right).

The circuit diagram of a modified ERL driver shown in Fig. 3 (b) is utilized [6]. 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. 3 (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. 4, 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

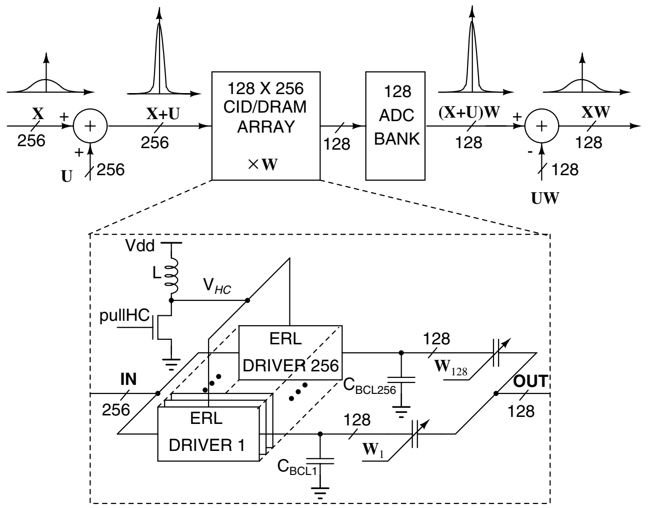

Fig. 5. Mixed-signal array and stochastic modulation architecture.

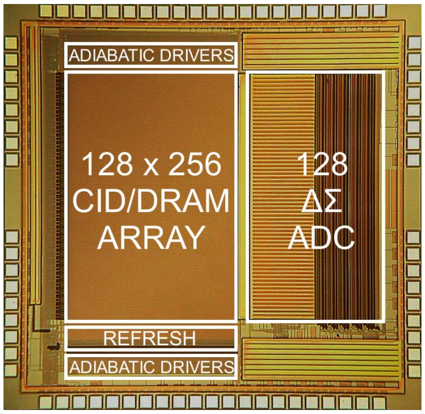

Fig. 6. Adiabatic VMM processor micrograph and floor plan. The die area is 2.2 × 2.2mm^2.

active CID 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. 4. The energy dissipated in each computation is plotted on the right of Fig. 4. 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. 4.

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 pre-processing 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 [6]. 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. 5.

A bit/compute line has approximately constant capacitance, C_{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 percent of the time the array capacitance is typically (95-percent of cases) within a range $C_{BCL}N^{1/2}$ (twice the standard deviation) from the mean, a factor $N^{1/2}$ closer than the full allowable range from 0 to $C_{BCL}N$.

To randomize the bit-level distribution of a non-random digital input, a randomly chosen but fixed integer U_n in the range $[0, 2(N^{1/2} - 1)]$ is added to each input component X_n [10]. 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. 5. The modulated inner products in **X** are demodulated by digitally subtracting the inner products obtained for **X+U** and **U**. The random dithering vector **U** is chosen once, so its inner product with the templates is pre-computed 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 **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 [10].

V. Experimental Results

The 0.35-μm CMOS integrated prototype of the adiabatic VMM processor shown in Fig. 6 contains 32,768 CID/DRAM cells and 128 row-parallel 8-bit $\Delta\Sigma$ algorithmic ADCs [9], as well as pipelined input shift registers, sense amplifiers, refresh logic, and scan-out logic. Various cell configurations are included. All of the supporting digital clocks and control signals are generated on-chip.

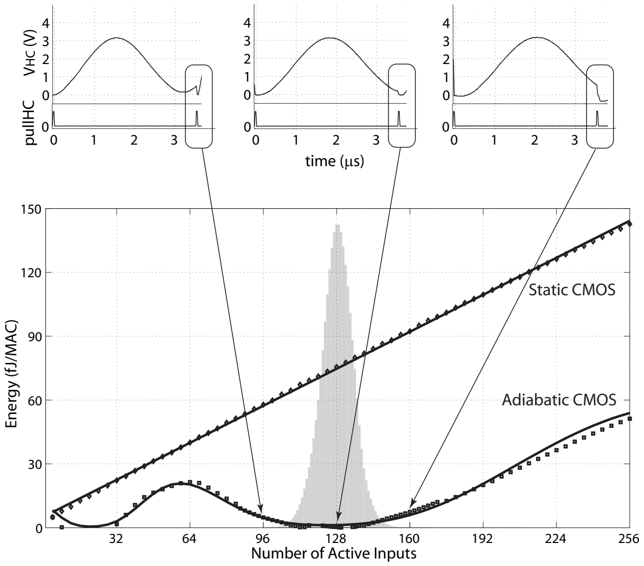

Fig. 7. Experimentally measured (dotted line) and theoretical (solid line) array dynamic energy dissipation as a function of input data statistics in adiabatic resonant mode and static CMOS mode. Three corresponding experimentally measured hot clock waveforms are shown. The probability density function of modulated input data is shown in gray.

Fig. 7 shows 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) experimentally measured at 22 kHz oscillator clock frequency. 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 $C_{BCL}N^{1/2}$ of its mean for 95 percent of the data. As shown in Fig. 7, this 95 percent interval is positioned to coincide with a broad region of resonance, minimizing energy 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 0.59μW power from 1.6V DC supply. A summary of the computing array characteristics is given in Table I.

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 15W of power, although the MACS measure provides a very incomplete metric of neural computation.

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

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 outweighted 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, A. Chandrakasan, "A 180-mV subthreshold FFT processor using a minimum energy design methodology," *IEEE JSSC*, Vol. 40, No. 1, pp. 310-319, Jan. 2005.

[2] H. Yamauchi, H. Akamatsu, T. Fujita, "An asymptotically zero power charge-recycling bus architecture for battery-operated ultrahigh data rate ULSI's," *IEEE JSSC*, Vol. 30, No. 4, pp. 423 - 431, April 1995.

[3] Y. Moon, D. K. Jeong, "An efficient charge recovery logic circuit," *IEEE JSSC*, Vol. 31, No. 4, pp.514-522, April 1996.

[4] J. Ammer, M. Bolotski, P. Alvelda, T.F.Jr Knight, "160x120 pixel liquid-crystal-on-silicon microdisplay with an adiabatic DAC," *IEEE ISSCC*, pp. 212 - 213, 1999.

[5] W. Athas, N. Tzartzanis, W. Mao, L. Peterson, R. Lal, K. Chong, Joong-Seok Moon; L. Svensson, M. Bolotski, "The design and implementation of a low-power clock-powered microprocessor," *IEEE JSSC*, Vol. 35, No. 11, pp. 1561 - 1570, Nov. 2000.

[6] R. Karakiewicz, R. Genov, A. Abbas and G. Cauwenberghs, "175 GMACS/mW Charge-Mode Adiabatic Mixed-Signal Array Processor," *Proc. IEEE 2006 Symp. VLSI Circuits*, Honolulu HI, Jun 13-17, 2006.

[7] A. Nakada, T. Shibata, M. Konda, T. Morimoto, and T. Ohmi, "A Fully Parallel Vector-Quantization Processor for Real-Time Motion-Picture Compression", *IEEE JSSC*, Vol. 34, No. 6, pp. 822-830, June 1999.

[8] T. Yamasaki, T. Nakayama, and T. Shibata, "A low-power and compact CDMA matched filter based on switched-current technology", *IEEE JSSC*, Vol. 40, No. 4, pp. 926-932, Apr. 2005.

[9] R. Genov, G. Cauwenberghs, G. Mulliken, F. Adil, "A 5.9mW 6.5GMACS CID/DRAM Array Processor," *Proc. European Solid-State Circuits Conference (ESSCIRC'2002)*, Florence Italy, Sept. 24-26, 2002.

[10] R. Genov and G. Cauwenberghs, "Stochastic Mixed-Signal VLSI Architecture for High-Dimensional Kernel Machines," *Adv. Neural Information Processing Systems (NIPS'2001)*, Cambridge: MIT Press, vol. 14, 2002.