Energy-Efficient Dot Product Computation using a Switched Analog Circuit Architecture

Ihab Nahlus[†], Eric P. Kim[†], Naresh R. Shanbhag[†], and David Blaauw[*]
[†]Department of Electrical and Computer Engineering, University of Illinois, Urbana, IL 61801, USA
{nahlus2, epkim2, shanbhag}@illinois.edu
[*]Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA
blaauw@umich.edu

ABSTRACT

In this paper, we present switched analog circuit (SAC), a new circuit architecture, to implement an energy-efficient mixed-signal dot product (DP) kernel for machine learning and signal processing applications. SAC operates by fast switching the analog inputs to output via variable width digital pulses. The output accuracy and energy consumption of SAC is analyzed and verified for an average and Gaussian blur filter. Simulations in a commercial 130 nm process for a 120×120 image show energy savings of 19×-to-32× compared to a digital implementation for signal-to-noise ratios (SNRs) of 30 dB-to-24 dB, respectively.

Categories and Subject Descriptors

B.2 [**Hardware**]: Arithmetic and logic structures

Keywords

Switched analog circuit; Low-power; Dot product; Mixed-signal

1. INTRODUCTION

The demand for ubiquitous computing with learning and decision making capabilities has grown in the past several years. These applications need to process large amounts of data acquired by sensing the surrounding environment and are subject to strict energy demands. Figure 1 depicts a typical sensory data processing chain. Sensors such as CMOS image sensors acquire analog data, which is then processed by a digital processor, or an actuator. The dot product (DP) kernel within the processor implements a variety of functions including but not limited to vector inner products, correlators, filters, convolutions, multiply-accumulate, L-1 and L-2 norms, which are extensively used in classifiers such as support vector machines (SVMs), in deep learning networks, image processors, and communication receivers.

Conventionally, these kernels are designed using digital logic. This approach enables complex algorithms requiring high precision to be implemented reliably, but at a high energy cost. Feature size scaling has reduced energy costs significantly, making digital design the favorable choice. However, as the number of sensors increase, the analog-to-digital converter (ADC) overhead can be quite large, especially if an analog output is required to drive an actuator. Analog processing has been reported to be more energy-efficient at low precision [4]. By operating in the analog domain, the overhead of ADCs and digital-to-analog converters (DACs) can be eliminated. Energy efficient designs have been proposed that use current summing, but their application has been limited to ultra high speed applications [2]. A mixed-signal approach that utilizes switched capacitors has been reported to give large energy savings [1]. However, these designs are susceptible to process, voltage and temperature (PVT) variations, and do not scale well with process technology which makes it challenging for implementing in sensory chains.

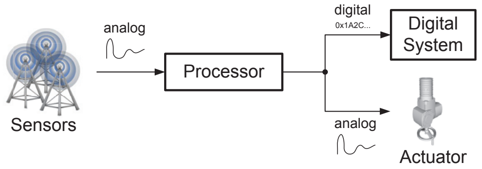


Figure 1: Block diagram of a sensory data processing chain.

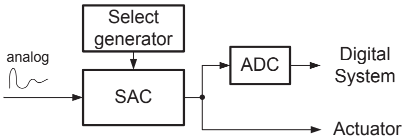
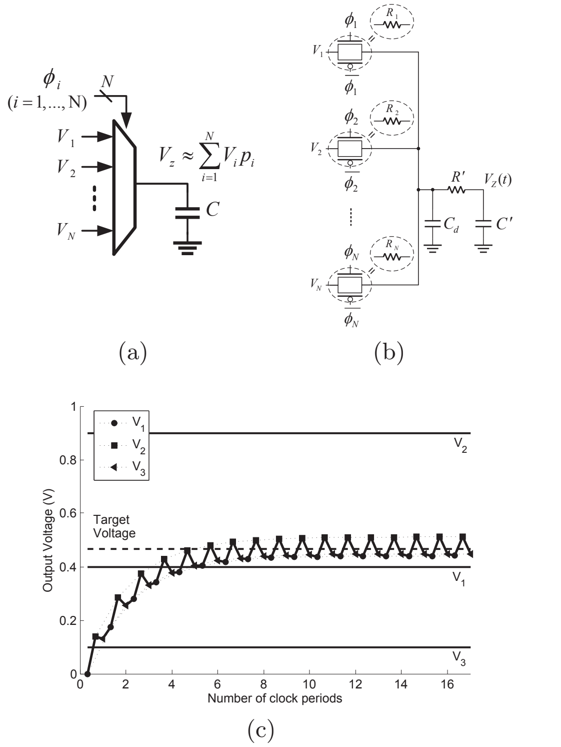


Figure 2: Switched analog circuit (SAC) implementation for processor kernel.

In this paper, we present switched analog circuit (SAC) (Fig. 2), which is an energy-efficient mixed-signal circuit architecture. SAC implements the DP kernel by fast switching the analog input to the output via variable width digital pulses. The input analog voltages are passed through an N input MUX with N select signals (Fig. 3(a)). By having only one select signal active at a time, and switching among the inputs at a high frequency, the output voltage is obtained as the weighted sum of the input voltages. An example operation with $N = 3$ is depicted in Fig. 3(c).

In this paper, we implement a SAC based average and

ISLPED '14 August 11 - 13 2014, La Jolla, CA, USA. Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-2975-0/14/08 ...$15.00.
http://dx.doi.org/10.1145/2627369.2627664.



(a)

(b)

(c)

Figure 3: Switched analog circuit (SAC)-based DP kernel: (a) conceptual operation, (b) circuit implementation, and (c) output waveform for $N = 3$, $V_1 = 0.4$, $V_2 = 0.9$, $V_3 = 0.1$, and $p_1 = p_2 = p_3 = 1/3$.

Gaussian blur filter in a commercial 130 nm process. When applied to a 120 × 120 image, 19×-to-32× energy savings can be achieved compared to a digital implementation at a signal-to-noise ratio (SNR) of 30 dB-to-24 dB, respectively.

The remainder of the paper is organized as follows. Section 2 describes SAC in detail. Section 3 presents the behavioral model of SAC. Section 4 presents simulation results and Section 5 concludes the paper.

2. SWITCHED ANALOG CIRCUIT (SAC)

2.1 SAC-based DP kernel

A length N SAC-based DP kernel, as shown in Fig. 3(a), takes input voltages $V_1, \ldots, V_N$, and computes the output voltage $V_o = \sum_{i=1}^{N} V_i p_i$. The weights p_i are implemented by a set of non-overlapping pulses ϕ_i (i^{th} element of Φ_v) of period T and duty cycle p_i. The switch can be viewed as a switched resistor [3] network with its effective resistance divided by p_i. By designing T to be significantly smaller than the time constant of the RC network, V_z will converge to $\sum_{i=1}^{N} V_i p_i$, with accuracy increasing at an exponential rate with the number of cycles until it settles.

The circuit implementation of the SAC kernel is shown in Fig. 3(b). Transmission gates are used to implement the switches to allow full swing at the output. A detailed analysis of this kernel is given in Section 3. A series resistance R' and a capacitor C' is added to suppress the effect of variation in path resistances R_i.

2.2 Select generation

Generation of the duty-cycled clocks (select signals) with period T is needed for proper SAC operation. The select signals are generated by a *multi-phase clock generator* (MPCG) that provides clock inputs to the combinational logic. The MPCG is designed using a length M ring counter operating at a frequency $f_{CLK} \triangleq \frac{1}{T_{CLK}} = \frac{M}{T}$. The combinational logic generates pulses with variable width $\frac{x}{M}T$ with a phase offset of $\frac{y}{M}T$. For large M, the ring counter becomes expensive. An alternative would be to use a counter at the expense of complexity/energy.

2.3 Energy Consumption

For the SAC-based DP kernel, the total energy consumption per DP computation can be written as:

$$E_{tot}[n] \triangleq E_{SAC}[n] + \frac{nE_{MPCG}}{V}$$

where $E_{SAC}[n]$ is the energy consumption of the SAC DP kernel and combinational logic over n clock cycles, E_{MPCG} is the energy consumption of the MPCG per clock cycle, and V is the number of SAC DP kernels sharing the same MPCG. E_{MPCG} depends largely on the topology used and hence will be obtained through simulations. The energy dissipated in the combinational logic, gate and drain capacitors of the kernel is linear in n and dominates the energy dissipated in C', when C' and C_d are of the same order.

3. BEHAVIORAL MODEL

The transient response of the SAC computation kernel, a switched RC circuit, can be obtained using linear constant-coefficient difference equations. Let $\tau_{max} \triangleq \max_{i=1..N}(R_i)C_d$ be the largest time constant of the circuit when $R' = 0$ and $C' = 0$. Two conditions are imposed on the values of R' and C': C1) $R'C' \gg \tau_{max}$, and C2) $R' \gg \max_i(R_i)$. C1 ensures that the new time constant will be dominated by $R'C'$ while C2 will have an impact on the output accuracy as will be shown. Note: C1 is automatically satisfied if C2 is, by ensuring C' is of the same order as C_d. First, we make the following two claims:

Claim 1. If $T \ll R'C'$,

$$\tilde{V}_z \triangleq \lim_{n\to\infty} V_z[n] \approx \sum_{i=1}^{N} V_i p'_i, \qquad (1)$$

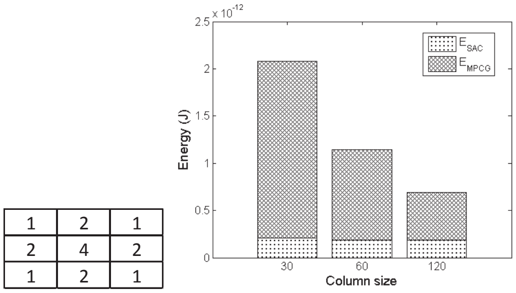
where $V_z[n]$ is the DP kernel's output after n clock cycles, and $p'_i = \frac{\frac{p_i}{R'+R_i}}{\sum_{j=1}^{N} \frac{p_j}{R'+R_j}} \approx p_i$ (since $R' \gg R_j, j = 1, ..., N$).

Now, define the error at the output after n clock cycles to be $e[n] \triangleq V_0 - V_z[n]$, where $V_o = \sum_{i=1}^{N} V_i p_i$ is the ideal output.

Claim 2. The mean-square error $J[n, K]$ after n clock cycles is:

$$J[n,K] \triangleq E\{e[n]^2\} = E\{(V_0 - \tilde{V}_z)^2\} + K^n \alpha_A + K^{2n} \beta_A \quad (2)$$

where $K \triangleq e^{-\frac{T}{C'}\sum_{j=1}^{N}\frac{p_j}{R'+R_j}} \approx e^{-\frac{T}{R'C'}}$, and α_A and β_A are constants that depend largely on the input's $1st$ and $2nd$ order statistics. Note that as n increases, K^n will go to zero. Proofs for the above two claims have been omitted due to space limitations. Circuit simulation results in Section 4.3 support these claims.



1	2	1
2	4	2
1	2	1

(a) (b)

Figure 4: (a) Coefficients of a Gaussian blur filter with $\sigma^2 = 0.85$, and (b) Energy per computation breakdown of the Gaussian filter applied to 30×30, 60×60, and 120×120 images.

Let $H = \frac{R'C'}{T} = \frac{-1}{ln(K)}$. Then, decreasing H results in faster convergence but greater inaccuracy as $\tilde{V}_z$ will no longer equal V_0 since the approximation in (1) is no longer valid. As energy consumption has a strong dependence on the number of clock cycles, H and n become two important variables in the design of a SAC-based DP kernel. We finally note that the exact value of the resistor R' is unimportant as long as C2 is satisfied. A polysilicon resistor of around 1 MΩ in a commercial 130 nm process with minimum width corresponds to area of 600 minimum sized transistors. This presents a large area overhead for small N ($N = 2$, e.g.) but for larger values of N ($N = 9$ in our implementation), we obtain large area savings (compared to digital implementation) since the same resistor will be shared by all N paths.

4. SIMULATIONS AND RESULTS

4.1 Simulation setup

A SAC-based DP kernel (Fig. 3(b)) is used to implement an image filter. Circuit simulations in a commercial 130 nm CMOS process at the nominal corner were performed for image sizes 30×30, 60×60, and 120×120. We note that the image is processed on a per-row basis and hence $V = 30, 60$ and 120 for the different image sizes. The MPCG was designed as a ring counter that operates at $T_{CLK} = 400$ ps and gives $T = MT_{CLK}$, where M is the length of the ring counter. Three average filters of lengths $M = 9$, 25, and 49 were implemented which correspond to a 3×3, 5×5 and 7×7 window, respectively. These filters do not require any combinational logic at the output of the ring counter since the different phases already represent the coefficients. A 3×3 Gaussian blur filter with $\sigma^2 = 0.85$ (Fig. 4(a)) has also been implemented. For this filter, M is chosen to be 16 which is the sum of all coefficients. The D-flipflops in the ring counter are implemented using true single-phase clocking (TSPC). Select signals are generated using static-CMOS based NOR and NAND gates. Select signals corresponding to coefficient of unit value do not need a logic stage but are still passed through inverters to match the delay of other select signals. A single ring counter was shared for all parallel SAC units, while the combinational logic may be duplicated to ensure sharp rise and falls of the select signals. In our simulations, one set of logic gates for the 30×30 image, two sets for the 60×60 image, and four sets for the 120×120 image were used. For our design, simulations show that $E_{MPCG} \gg E_{SAC}$, and energy per single computation is dominated by $\frac{E_{MPCG}}{V}$. As more computation kernels can share the same ring counter, more energy benefits can be obtained (Fig. 4(b)).

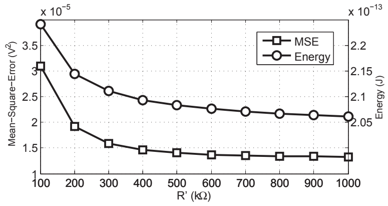
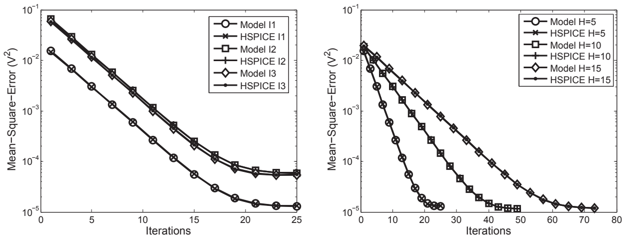


Figure 5: Circuit simulation of a SAC-based Gaussian blur filter with $H = 5$ ($n = 5H$).



(a) (b)

Figure 6: Accuracy comparison between the behavioral model and circuit simulation for: (a) a Gaussian filter with different images and (b) Gaussian filter with varying H.

4.2 Choice of R′

The auxiliary resistor R' and capacitor C' were designed to satisfy condition C1 and C2 (Section 3). To obtain a good value for R', the circuit was simulated at different values of R' while keeping $H = \frac{R'C'}{T}$ at a fixed value by adjusting C' accordingly. C1 was satisfied by choosing $R'C'$ large enough to dominate τ_{max}. Figure 5 shows the plot of the MSE and energy consumption of the circuit vs. R' with $H = 5$. The same trend was observed for different values of H. A total of $5H$ iterations were performed. It can be seen that MSE and energy consumption reduce as R' increases until $R' \approx 700k\Omega$ for the MSE and $R' \approx 900k\Omega$ for energy. Hence, in all simulations, R' was chosen to be $1M\Omega$, and the value of C' was set to obtain a specific value for H.

4.3 Validation of behavioral model

Comparison of circuit simulations and behavioral model for mean-square error (MSE) vs. iterations are shown in Fig. 6. Figure 6(a) shows the Gaussian filter applied to three different 30×30 images (I1,I2 and I3). I2 and I3 were chosen

Table 1: Fitted accuracy parameter values of a Gaussian filter with $H = 5$ ($C' = 32$ fF).

Image	α_A	β_A	$E\{(V_0 - \tilde{V}_z)^2\}$
I1	-3.07×10^{-4}	2.3×10^{-2}	1.4×10^{-5}
I2	-1.11×10^{-3}	9.83×10^{-2}	6.17×10^{-5}
I3	-1.44×10^{-3}	8.9×10^{-2}	5.96×10^{-5}

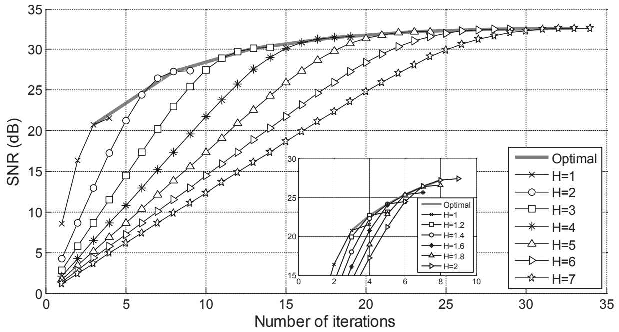


Figure 7: SNR vs. number of iterations of a Gaussian filter for a 30 × 30 image.

to have similar statistics while being different from those of I1. Weighted least-squares fitting was applied to obtain the parameters α_A, β_A and $E\{(V_0 - \tilde{V}_z)^2\}$ in Section 3. The fitted parameters obtained for $H = 5$ ($C' = 32\,\mathrm{fF}$) are tabulated in Table 1. It can be seen that the parameters for I2 and I3 are similar in value as expected.

In Fig. 6(b), H was varied for I1 to see its effect on MSE. As expected, the parameters α_A and β_A in this model are a weak function of H and depend largely on the input statistics. The asymptotic MSE value $E\{(V_0 - \tilde{V}_z)^2\}$ decreases as H increases as expected from Section 3. However, the decrease in MSE is minimal due to the overall accuracy being dominated by the imperfection of the select-signal generation block.

4.4 Design optimization

A large number of options exist for choosing the values of H and n to minimize the energy consumption of the SAC-based DP kernel for a given SNR. HSPICE simulations were performed on a Gaussian filter for I1, by sweeping over H at various iterations (Fig. 7). The number of iterations for a given SNR should be minimized since the linear component in E_{SAC} dominates ($\alpha_E \gg \beta_E$). From the close up view of $1 \leq H \leq 2$, it can be seen that the minimum iterations for a given SNR occur at $n = 5H - 2$. The optimal curve is obtained by joining these minimizing points. SNR improvements saturate around $H = 5$ due to the overall accuracy being dominated by the imperfection of the select-signal generation block.

4.5 Comparison to digital implementations

The SAC-based DP kernel is compared against a digital logic implementation using Baugh-Wooley multipliers (BWM) and ripple carry adders (RCA). To estimate the energy consumption of the adders and multipliers, the energy for a 1-bit full adder (E_{FA}) using a mirror-adder structure loaded with $FO4$ inverters was simulated. In a 130 nm process, $E_{FA} = 18.63\,\mathrm{fJ}$. Energy consumption of a B_x bit RCA is then estimated to be:

$$E_{RCA}[B_x] = \alpha_{0\to1} B_x E_{FA} \tag{3}$$

where $\alpha_{0\to1}$ is the activity factor of the RCA. We assume the inputs are uniformly distributed and hence $\alpha_{0\to1}$ is 0.25. The energy consumption of a B_x bit BWM is lower-bounded [5] by:

$$E_{BW}[B_x] \geq E_{FA}(B_x^2 - 2B_x + 2)$$

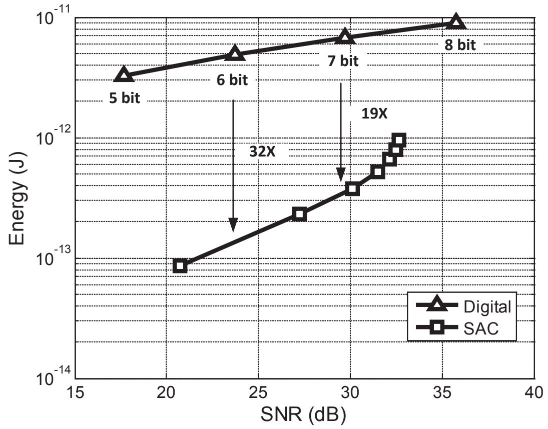


Figure 8: Comparison of SNR vs. energy per DP computation for a Gaussian filter.

The SNR vs. energy per DP computation is shown in Fig. 8 for a 120 × 120 image. For $SNR \approx 24\,\mathrm{dB}$, energy savings are approximately 32× whereas for $SNR \approx 30\,\mathrm{dB}$, the energy savings are approximately 19×. These savings are pessimistic as E_{BW} was based on a lower bound.

5. CONCLUSION

In this paper, we have presented a new energy-efficient mixed-signal DP kernel that can achieve large energy savings for the same level of accuracy. This work opens up the possibility of employing SAC to design inference kernels for various emerging applications.

6. ACKNOWLEDGEMENTS

This work was supported by Systems on Nanoscale Information fabriCs (SONIC), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA.

7. REFERENCES

[1] M. Duppils and C. Svensson. Low power mixed analog-digital signal processing. In *Proc. of Int. Symp. on Low Power Elect. and Design*, pages 61–66, 2000.

[2] Y. Lu and E. Alon. Design techniques for a 66 Gb/s 46 mW 3-tap decision feedback equalizer in 65 nm CMOS. *IEEE J. Solid-State Circuits*, 48(12):3243–3257, Dec. 2013.

[3] M. H. Perrott, S. Pamarti, E. Hoffman, F. S. Lee, S. Mukherjee, C. Lee, V. Tsinker, S. Perumal, B. Soto, N. Arumugam, et al. A low-area switched-resistor loop-filter technique for fractional-N synthesizers applied to a MEMS-based programmable oscillator. In *IEEE Int. Solid-State Circuits Conf. (ISSCC)*, pages 244–245, 2010.

[4] R. Sarpeshkar. Analog versus digital: extrapolating from electronics to neurobiology. *Neural computation*, 10(7):1601–1638, 1998.

[5] J. H. Satyanarayana and K. K. Parhi. A theoretical approach to estimation of bounds on power consumption in digital multipliers. *IEEE Trans. Circuits Syst. II*, 44(6):473–481, 1997.