A CNN-Inspired Mixed Signal Processor based on Tunnel Transistors

Behnam Sedighi, Indranil Palit, X. Sharon Hu, Joseph Nahas, and Michael Niemier
Department of Computer Science and Engineering, University of Notre Dame
Notre Dame, IN 46556, USA, Email: {bsedighi, ipalit, shu, jnahas, mniemier}@nd.edu

Abstract—Novel devices are under investigation to extend the performance scaling trends that have long been associated with Moore's Law-based device scaling. Among the emerging devices being studied, tunnel FETs (or TFETs) are particularly attractive, especially when targeting low power systems. This paper studies the potential of analog/mixed-signal information processing using TFETs. The design of a highly-parallel processor – inspired by cellular neural networks – is presented. Signal processing is performed partially in the time-domain to better leverage the unique properties of TFETs, i.e., (i) steep slopes (high g_m/I_{DS}) in the subthreshold region, and (ii) high output resistance in the saturation region. Assuming an InAs TFET with feature sizes comparable to the 14 nm technology node, a power efficiency of 10,000 GOPS/W is projected. By comparison, state-of-the-art hardware assuming CMOS technology promises a power efficiency only close to 1,000 GOPS/W.

I. Introduction

Tunnel FETs (TFETs) are a promising candidate for realizing energy efficient digital circuits in the post-CMOS era. Their on-current (I_{on}) to off-current (I_{off}) ratio [1]–[3] can be made large. Subthreshold swings as low as 21 mV/dec have been observed experimentally [3]. Moreover, TFETs could provide excellent saturation behavior [4]–[6], which is important for the design of analog circuits. Previous publications have examined the merits of TFET-based digital circuits, and have shown that at low supply voltages, TFET-based digital circuits have a better energy efficiency compared with conventional CMOS designs [7]–[10]. Turning to analog applications/circuits, researchers have also considered how the higher g_m/I_D of TFETs in the subthreshold region could be employed to design low-power amplifiers [11]–[13]. (However, the impact of process variations – which can have a significant impact on deep-sub-micron technologies – is often not addressed [11], [12].) More recently, researchers have also begun to consider RF-powered systems based on TFET devices [14], [15]. To the best of our knowledge, the merits of using TFETs for analog/mixed-signal *computation* has not yet been studied.

In this paper, we examine the potential of TFET-based processing engines to pre-process/condition analog signals and output digital signals. For many applications, this pre-processing step is critical to reduce the amount of data sent on to digital processors, and hence the overall system energy consumption. The premise of this work is on highly parallel processing platforms similar to single-instruction-multiple-data (SIMD) processors [16]–[18], cellular neural networks (CNNs) [19]–[21], or vision chips [22]–[25]. The performance comparisons for many of the CMOS processors referenced above will be summarized and discussed later in this paper in order to place our work in the proper context. Specifically, we will highlight mega-operations per cell (MOPS) as a measure of performance efficiency, and giga-operations per second per Watt (GOPS/W) as a measure of power efficiency. (In both instances, higher numbers are more desirable).

The primary contribution of this paper is a *new architecture for mixed signal computation that specifically exploits the unique characteristics of TFET devices*. This approach is fundamentally different from prior efforts which mainly focused on how TFETs might be used to simply duplicate the functionality of existing hardware – e.g., TFET-based SRAM [26], [27], TFET-based multi-core architectures [10], etc. Specifically, we present a CNN-inspired processor that eliminates the need for voltage controlled current sources (VCCSs) (which are needed to realize feedback and feedforward templates in CNNs [28]), and are the dominant source of power consumption in a CNN array [29]. Instead, VCCSs are replaced with simple comparators – that can be efficiently realized with TFETs given their high intrinsic gain. Initial (and conservative) projections suggest that power efficiencies of more than 10,000 GOPS/W could be achievable with our approach. This represents an improvement of more than 10X over state-of-the-art architectures that assume MOSFET/FinFET technology and seek to accomplish similar information processing tasks.

As a case study, we consider the calculation of weighted sums of analog inputs – a task at the heart of many signal processing circuits (in CNNs, etc.). TFET-based hardware that can *simultaneously* perform analog computation and analog-to-digital conversion is presented. We convert input voltage to pulse-width, and measure pulse width with the aid of a high frequency clock. (There are some similarities to single-slope A/D converters (ADCs) that have been partially explored in other efforts [21], [30].) Also, we present an offset cancellation scheme to address the impact of device variations that are sure to exist, and that have often been ignored in prior work on TFET-based circuits. We also discuss differential measurements of pulse-width, which significantly reduces the signal activity of the counters that are employed in our architecture (to quantitatively determine a weighted sum of inputs) and lowers energy dissipation. Lastly, we present a methodology for adjusting the weight of different inputs in the desired weighted sum using a direct-digital frequency synthesizer.

II. Background

Here, we review the CNN architecture, and articulate why we believe TFETs are particularly amenable to our work.

A. CNN

The conventional CNN architecture [28] is an $M \times N$ array of identical cells, where each cell has identical synaptic connections with all the adjacent cells in a predefined neighborhood N, which typically includes only the immediate neighbors. A cell consists of one resistor, one capacitor, a number of linear voltage controlled current sources (VCCSs), one fixed current source, and one non-linear voltage controlled voltage source. The node voltages u_{ij}, x_{ij}, and y_{ij} correspond
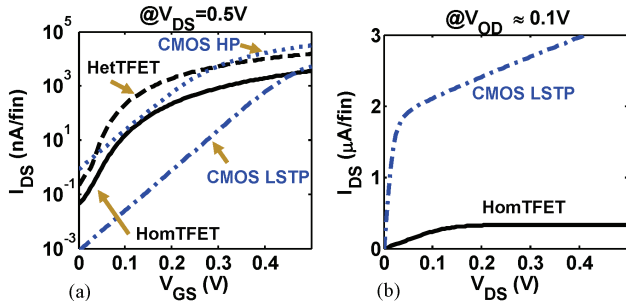
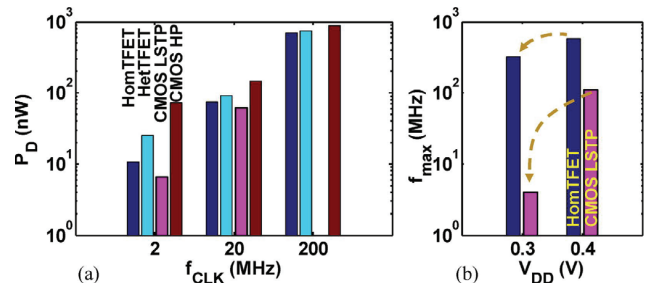Fig. 1. Characteristics of TFET and CMOS ($V_{OD} = V_{GS} - V_{TH}$).

Fig. 2. (a) Simulated power dissipation of an 8-bit counter at $V_{DD} = 0.4$ V (LSTP CMOS counter can not work at 200MHz). (b) Maximum clock frequency of the counter.

to the input, state, and output of a given cell C_{ij}, respectively. The input and output voltages of each neighbor contribute a feedback, and a control current to a given cell via the VCCSs (i.e., affect cell state x via the VCCSs). The dynamics of the cell C_{ij} can be expressed by Eq. 1. To ensure fixed binary outputs, a cell typically employs a non-linear sigmoid-like transfer function (Eq. 2) at the output.

$$C\frac{dx_{ij}(t)}{dt} = -\frac{x_{ij}(t)}{R} + \sum_{C_{kl} \in \mathbb{N}_{ij}} a_{ij,kl}y_{kl}(t) + \sum_{C_{kl} \in \mathbb{N}_{ij}} b_{ij,kl}u_{kl} + Z$$

$$\tag{1}$$

$$y_{ij}(t) = \frac{1}{2}(\mid x_{ij}(t) + 1 \mid - \mid x_{ij}(t) - 1 \mid) \tag{2}$$

The parameters $a_{ij,kl}$, and $b_{ij,kl}$ act as weights for the feedback and control currents from cell C_{kl} to cell C_{ij}. Due to their space invariant nature, these parameters are often denoted by two 3×3 matrices, namely the feedback template A and the control template B. By setting the values of A, B, and Z, it is possible to solve a wide range of problems [31].

B. Device Technology

1) Devices: We employ an InAs homo-junction TFET (HomTFET) [8] as it is one of the most mature and well-studied TFETs. Hetero-junction TFETs (HetTFETs) (e.g., a GaSb-InAs HetTFET [32] – with a higher I_{on}) could also be used. We compare the InAs HomTFET with both GaSb-InAs HetTFET and CMOS transistors in Fig. 1. (These, and all results in this paper are based on SPICE simulation – see [33] for model details). The $I_{DS} - V_{GS}$ plots clearly illustrate the "steep slopes" of the TFETs. Note that leakage current of the HomTFET is between that for HP and LSTP CMOS. Also, the HomTFET has a higher current drive than LSTP CMOS for a V_{DD} smaller than 0.4 V. As such, TFET-based digital circuits would be faster given said supply voltages.

2) Preliminary Circuit Analysis: Digital counters are extensively used in our presented architecture (see Sec. III). To justify what device technology is the "best fit" for our applications, we use an 8-bit counter as a testbench. As seen in Fig. 2(a), for the HetTFET and HP CMOS devices, the leakage current becomes dominant at frequencies below 2 MHz. In our work, clock frequencies of ~ 100 MHz are of interest, but as clock gating is utilized, acceptable effective clock rates can be as low as just a few MHz. Given this, HomTFETs and CMOS LSTP (low standby power) are more desirable given lower leakage. As LSTP transistors work in the subthreshold region when $V_{DD} < 0.4$ V, device speed is exponentially affected by the supply voltage; if V_{DD} is reduced from 0.4 to 0.3 V, the

maximum operating frequency of the CMOS LSTP counter is reduced by more than 20X (see Fig. 2(b)). In contrast, when V_{DD} of the HomTFET-based counter is reduced to 0.3 V, the maximum clock frequency remains above 300 MHz – which suggests that a HomTFET-based counter should be robust with respect to process/voltage/temperature (PVT) variations. As such, for *digital circuits* we will also assume HomTFETs (with V_{DD}=0.4 V).

The threshold voltage of a HomTFET is 120 mV, which facilitates the design of low-voltage *analog circuits*. Moreover, if TFETs are biased in the subthreshold region, they present a higher transconductance (g_m) than a MOSFET biased at a similar drain current, because of their steep I_{DS}/V_{GS} slope. Another advantage of HomTFET for analog designs is higher output resistance (constant I_{DS} in the saturation region) per Fig. 1(b). Thus, HomTFET-based amplifiers will have higher intrinsic voltage gain.

III. A MIXED SIGNAL PROCESSOR ARCHITECTURE

In a conventional CNN, VCCS's, which may be implemented by operational transconductance amplifiers, suffer from several non-ideal effects. For example, transistor mismatches prevent having well-defined gains in a given VCCS and introduce offsets. Mismatches and process variation have been exacerbated in deep-sub-micron technologies. Moreover, at small supply voltages, it has become more difficult to make circuits linear in a large input range.

TFETs provide the possibility of building high-gain amplifiers, and the ultimate high-gain amplifier is a comparator. The non-ideal effects in a comparator cause input-referred offset (V_{offset}). Since gain-error and nonlinearity are not relevant in a comparator, robustness is relatively easier to guarantee.

We present a CNN-inspired processor that eliminates the need for VCCS's and replaces them with comparators. TFET technology helps us design a simple comparator because of the high intrinsic gain of transistors. Power dissipation of the comparator will be lower than its CMOS equivalent since the input differential pair of the comparator are biased in subthreshold region where TFETs have a higher g_m/I_{DS}. Finally, additional processing tasks can be transferred to the digital domain, where robust, low-voltage circuits can be designed given the low threshold voltages of TFETs.

The architecture of the processor is shown in Fig. 3(a). It consists of a homogeneous array of processing cells. Each cell receives an analog input, communicates with its neighbors, and produces a digital output. A frequency synthesizer generates
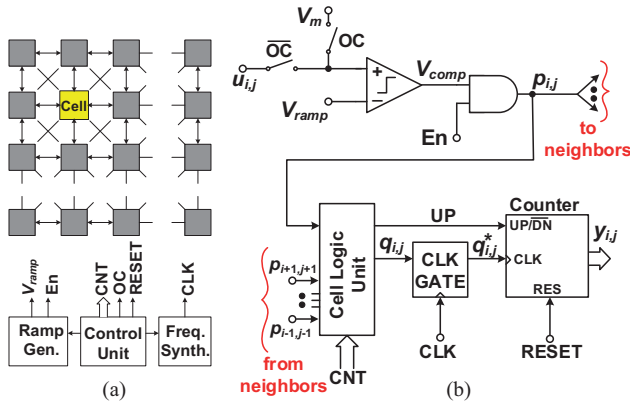
Fig. 3. (a) Architecture of processor. (b) A cell in ith row and jth column.

Fig. 4. (a) Voltage to pulse-width conversion. (b) Pulse-width measurement. (c) Calculating a weighted sum by utilizing TDM and changing CLK.

a variable clock frequency. A ramp signal is generated and applied to all cells. A control unit sets the frequency of the synthesizer, starts and stops the ramp, and applies proper settings to all cells. We avoid placing any multipliers or adders inside cells, in order to keep cell size as compact as possible.

A. Cell Design

Each cell, shown in Fig. 3(b), has three main components: a comparator, a small logic block (cell logic unit), and a gated counter. The counter can be initially reset (using the $RESET$ signal) by the control unit. One input of the comparator is connected to either an input voltage ($u_{i,j}$) or a reference voltage (V_m). The latter indicates the minimum input value and is shared by all cells. A voltage ramp (V_{ramp}) is applied to the other input of the comparator. The ramp is shared by all comparators in the network and its slope is not changed during operation. The comparator is connected to $u_{i,j}$ or V_m in two consecutive ramp cycles as shown in Fig. 4. Each time a ramp is applied, a logic signal called En is set to a high-level by the ramp-generation circuitry. When both En and the output signal of the comparator (V_{comp}) are high, it means that the ramp is active, and V_{ramp} is smaller than the input. During this time, the signal $p_{i,j}$ will be high. As a result, at each ramp cycle, a pulse is generated at $p_{i,j}$, the width of which carries information about the magnitude of the input.

In this design, alternation between V_m and $u_{i,j}$ is an offset-cancellation mechanism. The signal OC, generated by the control unit, determines which input is applied to the comparator. Consider a comparator that has an offset voltage of V_{offset} with a fixed timing skew T_{skew} between En and V_{comp}. T_{skew} is caused by the delay of the comparator, digital circuits, or inter-cell wirings. Also, assume that the difference between the rise and fall times of the AND gate that follows the comparator is ΔT_{rf}. When V_m is applied to the comparator, the pulse-width of $p_{i,j}$ can be easily found as

$$T_m = (V_m + V_{offset})/s_{ramp} + T_{skew} + \Delta T_{rf}/2 \quad (3)$$

where s_{ramp} is the slope of the ramp signal (in V/s). Similarly, when $u_{i,j}$ is applied to the comparator, the pulse-width of $p_{i,j}$ is

$$T_{i,j} = (u_{i,j} + V_{offset})/s_{ramp} + T_{skew} + \Delta T_{rf}/2. \quad (4)$$

The difference between $T_{i,j}$ and T_m will be

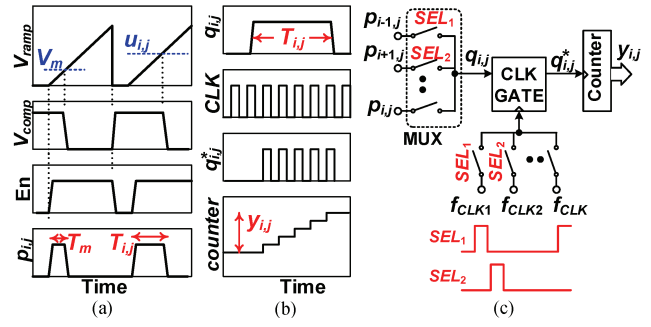$$\Delta T_{i,j} = T_{i,j} - T_m = (u_{i,j} - V_m)/s_{ramp}. \quad (5)$$

Note that offset and timing skew do not affect $\Delta T_{i,j}$. It can also be shown that low frequency noise (i.e., the flicker noise) of the comparator will be diminished by this method if the noise frequency is much smaller than the ramp frequency.

The next step is measuring $\Delta T_{i,j}$. For now, assume that the cell logic unit in Fig. 3(b) simply passes the input pulse to the output (i.e., $q_{i,j} = p_{i,j}$). Since $q_{i,j}$ is used to gate the clock signal (CLK) as shown in Fig. 4(b), the counter value $y_{i,j}$ will change by $y_m = T_m/T_{CLK}$ and $y_{i,j} = T_{i,j}/T_{CLK}$ during the two subsequent ramp cycles. (The clock period is $T_{CLK} = 1/f_{CLK}$.) Furthermore, the direction of counting can be reversed using the signal UP, which is generated by the control unit. If the counter counts down during the first ramp cycle, and up in the second cycle, the final change in the counter value is

$$\Delta y_{i,j} = (u_{i,j} - V_m)f_{CLK}/s_{ramp}. \quad (6)$$

$\Delta y_{i,j}$ is the digital representation of the difference between the input voltage and V_m, amplified (attenuated) by a weighting factor $w = f_{CLK}/s_{ramp}$. w can be adjusted by altering either the slope of the ramp as in [30], or the clock frequency. The former requires a digital-to-analog converter (DAC) in the ramp generation block, whereas the latter requires a frequency synthesizer. In this work, a direct-digital frequency synthesizer (DDS), shared by all cells, is used to set f_{CLK} as it is fully-digital, robust and scalable. The overhead of having a DDS with respect to the total area and power dissipation of the chip is small, if the processor has a large number of cells.

B. Time-Division Multiplexing

The next step is producing the sum of multiple inputs. We perform the summation in subsequent ramp cycles in a time-division multiplexed (TDM) manner. The rationale behind this is that one level of parallelism already exists in the system and, as the circuits are fast enough, the internal operation of the cells can be performed serially. Thus, the output of the comparator (more precisely $p_{i,j}$) is used in a cell in the first two ramp cycles, then it is used in another neighboring cell in the next two ramp cycles, and so on.

The TDM concept is shown in Fig. 4(c), wherein a different weight translates into a different clock frequency. Two multiplexers are used to sequentially route each input pulse and its corresponding clock frequency to the clock gating circuit. For example, when SEL_1 is high, a ramp is applied to all comparators, $p_{i-1,j}$ gates f_{CLK1}, and the counter value

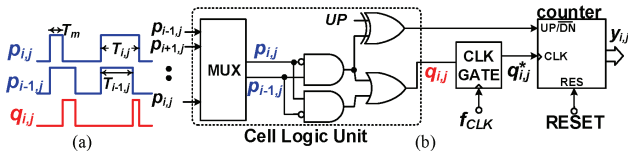Fig. 5. (a) Difference of two pulses. (b) Details of the cell logic unit for differential measurements.

changes by $\Delta T_{i-1,j} f_{CLK1}$. Next SEL_2 is set to high and the process is repeated. After all the neighboring pulses are digitalized, the counter holds a value that is equal to the weighted sum of the inputs:

$$y_{i,j} = \sum_{k,l \in N_{i,j}} \frac{S_{k,l}(u_{k,l} - V_m) f_{CLK_{k,l}}}{s_{ramp}}, \qquad (7)$$

where $f_{CLK_{k,l}}$ refers to the clock frequency used for measuring each pulse width, and $N_{i,j}$ refers to the set of all neighbors of the cell in ith row and jth column. The sign of each term $S_{k,l}$ ($= \pm 1$) is controlled by the signal UP. Note that in actual implementation, the multiplexer used to switch the clock frequency is not needed since the DDS generates one frequency at a time. In other words, altering the clock frequency is done by applying a new input to the frequency synthesizer, and not by switching f_{CLK} as conceptually illustrated in Fig. 4(c).

C. Differential Measurement

Using TDM, the logic unit of each cell is reduced to a MUX. However, more complicated logic circuits are possible, and might offer a more efficient implementation. Specifically, in many applications, the difference of two analog inputs needs to be calculated. To do so, the subtraction can be performed in time-domain as in Fig. 5(a). Fig. 5(b) shows how two pulses can be subtracted using simple logic gates. Here, $w(u_{i,j} - u_{i-1,j})$ is calculated in two ramp cycles as opposed to four without differential calculation. The logic circuit ensures that the clock (f_{CLK}) is not gated when $p_{i,j}$ and $p_{i-1,j}$ are equal. Moreover, if $p_{i,j}$ is low and $p_{i-1,j}$ is high, UP is inverted, i.e., if $u_{i,j} - u_{i-1,j} < 0$, then the counting direction is reversed.

From the above, differential measurement could seemingly improve performance by a factor of two. In practice, energy efficiency may be even more significant. In many applications, neighboring inputs are likely to be in close proximity – e.g., in an image, the neighboring pixels will have large intensity differences only for pixels on edges. Thus, when performing differential measurement the resulting waveform ($q_{i,j}$) is likely to consist of narrow pulses, and clock will be gated for most of the time (see Fig. 5(a)). Instead of counting for two long time intervals, the counter is only active (consuming dynamic power) for the difference of the two time intervals.

IV. CIRCUIT DESIGN

Here, we present circuit designs for key components discussed in Sec III. Unique properties of TFETs are exploited.

A. Comparator

The schematic of the comparator is shown in Fig. 6(a). It has a well-known topology with a differential input stage and a common-source output stage. The only non-traditional aspect of the design is the use of T5-6 and T12-13. To have a current mirror with good precision, transistors should have a
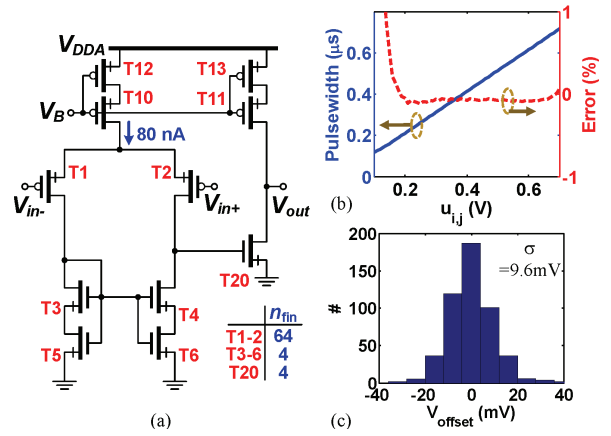

Fig. 6. (a) Simplified comparator circuit; number of fins (n_{fin}) of few transistors is given. (b) Relationship between input voltage of the comparator and the output pulse width (c) Histogram of the random offset after 500 runs.

large area and, at the same time, transistors should be biased in the saturation region (i.e., near- or sub-threshold operation should be avoided). Only increasing transistor length (L) can satisfy these requirements. The current TFET model does not allow for changing L. Therefore, in places that a transistor with a larger L is needed, resistor degeneration is utilized. For example, T5 has a small drain-source voltage and operates in the ohmic region. It acts as resistive source degeneration for T3, which is operating in the saturation region. The degeneration lowers the total transconductance of the T3-T5 pair (or the T4-T6 pair) which reduces the impact of threshold voltage mismatch on the output current of the mirrors.

For correct operation of the system, the voltage-to-pulse-width conversion should be linear. The comparator was tested in the setup of Fig. 3(b). The relationship between pulse width and input voltage ($u_{i,j}$) for a ramp signal with $S_{ramp} = 1$ V/μs is found as shown in Fig. 6(b). For input voltages close to ground potential, input transistors (T1,2) go out of the saturation region, whereas for inputs close to V_{DDA}, T10 will go out the saturation region. Both these phenomena change the speed of the circuit and cause errors. Nevertheless, the output pulse-width remains a linear function of input voltage in a large input ranges. To have sufficient margins an input range of 320 mV is considered, wherein the error is below 0.2%. This leaves more than 50 mV from either size of the input range as buffer zones against variations caused by offset. The input range of 320 mV and 8-bit accuracy translates into an equivalent least-significant bit (LSB) of 1.25 mV.

While the mismatch data for 14-nm TFET technology is not available, we considered mismatch coefficients $A_{TH} = 1$mV.μm and $A_\beta = 0.01\mu$m based on technology trends and the ITRS roadmap [34]. Due to the offset cancellation scheme, the exact value of these parameters is not important, but it is necessary to make sure that comparator offset is not excessively large. The histogram of the offset in Fig. 6(c), derived from Monte-Carlo simulation in SPICE, reveals that $|V_{offset}|$ is well below 50 mV. Hence, offset will not degrade the input common-mode range of the comparator and can be diminished by the offset cancellation scheme.

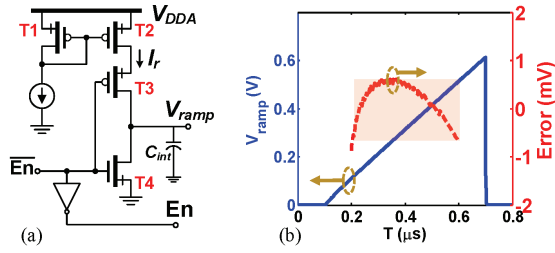The voltage gain of the comparator remains higher than 1200 V/V given mismatches. With a $V_{DDA}/3$ (=0.27 V) differ-

Fig. 7.   (a) Ramp generator. (b) Simulated ramp voltage and its error.

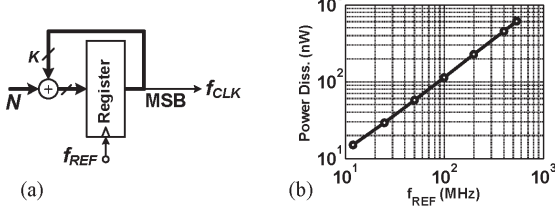Fig. 8.   (a) DDS. (b) Power dissipation of a 6-b DDS (N=13).

Fig. 9.   Two example simulation results for the optimal edge detection problem. Correct operation is observed.

ence between high and low output levels, the input sensitivity of the comparator is 0.22 mV. This is much smaller than one LSB and the gain of this simple comparator is sufficient for 8-b resolution due to the high output resistance of TFETs. Simulated comparison time and power dissipation are 10 ns and 0.11 μW, respectively, at V_{DDA} =0.8 V.

B. Ramp generation

Due to a large TFET output resistance, it is easy to build a ramp generator (Fig. 7(a)), simply by charging a capacitor by a fixed current. When En is high, T3 is on, and I_r charges C_{int}, generating a ramp. When En is low, T4 discharges C_{int}, and V_{ramp} will become zero. The slope of the ramp signal is equal to I_r/C_{int}. I_r and needs to be adjustable such that the slope of V_{ramp} could be fine-tuned. Fig. 7(b) shows a simulated output ramp. The error in the nonlinearity of the ramp, caused by the finite output resistance of T2, remains below ±0.5 LSB.

The capacitance C_{int} consists of the parasitic capacitances of the interconnects, as the ramp signal should be routed to all cells. (If there are 1000 cells, and each cell contributes 5 fF to C_{int}, the total capacitance will be 5 pF.) I_r should be 5 μA to achieve an s_{ramp}=1 V/μs. With a 0.8 V supply, this leads to a 4 μW power dissipation in the ramp generator, contributing 4 nW to the per-cell power dissipation. Per Sec. V, this is negligible compared to the power dissipation of the cell itself.

C. Digital Circuits

Digital parts of a circuit are conventional and use a topology similar to static CMOS. An important digital block is the DDS (Fig. 8(a)). Each time, before a ramp start, a new value of N is applied to the DDS. The DDS output (the most-significant bit (MSB) of the accumulator), goes to all cells. The output frequency is $f_{REF}N/2^K$, where K is the word-length of the accumulator. N can be any integer in the [0, $2^K - 1$] range. K depends on the desired accuracy of the calculations and the ratio between the largest and smallest output frequencies (i.e., the largest and smallest weights) is $2^K - 1$. The simulated power dissipation of a 6-bit DDS is shown in Fig. 8(b). When divided by the number of cells, DDS power dissipation is negligible. The DDS can operate up to f_{REF} =0.54 GHz.
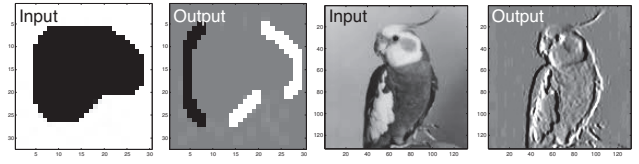
V.   CASE STUDY

We consider system performance via a case study of the optimal edge-detection task [35] – where edges are identified horizontally by assigning: (i) a 'black' color if an edge separates a darker region to its right side from a lighter region to its left, or (ii) a 'white' color if it separates a darker region to its left side from a lighter region to its right. With CNN terminology, the template for the task is expressed as [35]:

$$A = 0, \quad Z = 0, \quad B = \begin{bmatrix} -0.11 & 0 & 0.11 \\ -0.28 & 0 & 0.28 \\ -0.11 & 0 & 0.11 \end{bmatrix} \quad (8)$$

Architectural functionality (with differential measurement) is verified via simulation for the *optimal edge detection* problem (Fig. 9). Each pixel of the input images are provided to the corresponding CNN cells for processing. Each cell performs three passes of differential measurements (i.e., 6 ramp cycles) to calculate the final counter value. All the counters are initially reset. In the first pass, the operations in the first row of the matrix B (plus offset cancellation) are performed. The second and third passes account for the remaining rows. At the end, the counter of each cell holds the final result as an 8-bit binary number. In the first and last pass, the DDS is loaded with N =11, whereas in the second pass the DDS is loaded with 28. At each pass, (i) output frequency of the DDS $f_{CLK} = f_{REF}N/2^K$, is calculated for the corresponding N, (ii) a weighting factor $w = f_{CLK}/s_{ramp}$ that is indicative of how many times the counter should count for the corresponding row is calculated, and (iii) the counter counts $w(u_{k,j+1} - u_{k,j-1})$ times accordingly, where $k \in \{1, 2, 3\}$ accounts for the three rows of template B.

Additional details are given in Table I. A ramp of 1 V/μs is used. A faster ramp requires GHz clock frequencies for similar accuracy (8-bit output), whereas a slower ramp duration improves the accuracy but lowers the throughput. Although the presented processor loses some flexibility when compared to a digital processor, it has the advantage of having a built-in A/D conversion and compact hardware. Quantitatively, when compared with other CMOS-based architectures from the literature (see [16]–[25] for details), our processor has modest processing ability (MOPS), and superior power efficiency (GOPS/W). Per Fig. 10 (which plots GOPS/W as a function of MOPS for [16]–[23] and our work), the TFET-based comparator approaches the desired corner of the graph – where both performance and power efficiency are maximized. (Data was missing/non-competitive for other work – e.g., [25].)

VI.   CONCLUSION

We have introduced a robust and power-efficient mixed-signal processor using HomTFETs. The proposed processor is designed to exploit the unique properties of HomFETs and could attain power efficiencies of at least 10,000 GOPS/W (despite the fact that I_{on} is not large). TFETs with higher I_{on}'s

TABLE I. SPECIFICATIONS OF A CELL FOR EDGE DETECTION

DDS	6-bit with f_{REF} =360 MHz
Ramp	s_{ramp} =1.0 V/μs, ramping frequency=1.6 MHz
Total time	3.75 μs (for edge detection)
Total operations	11 (6 multiplications + 5 additions) per cell
Input	Analog (full-scale range = 0.32 V)
Output	8-bit (digital)
Power supply	Analog: 0.8 V; Digital: 0.4 V
Power dissipation†	0.23 μW per cell
Throughput	2.9 MOPS/cell; 192 GOPS in 256x256 network
Power efficiency	12,600 GOPS/W

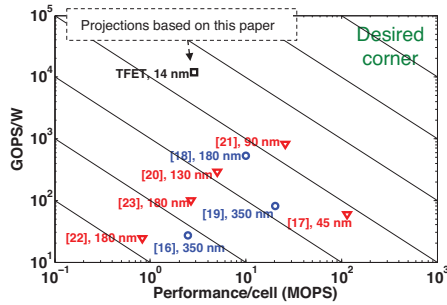†: average dissipation of a cell when inputs have uniform distributions.

Fig. 10. Comparisons of CMOS digital and analog processors for pre-processing/conditioning analog signals (GOPS/W vs. MOPS). ∇ and ○ indicate existing digital and analog implementations, respectively.

have been experimentally measured, however, such devices do not yet exhibit small sub-threshold swings and have larger leakage. Of course, device work is ongoing. If efforts to increase I_{on} (without degrading I_{off}) succeed, clock frequency and throughput improve, and a power efficiency much higher than 10,000 GOPS/W is expected. Finally, we would also like to point out that the proposed architecture is not limited to TFETs. For example, it can also be implemented with CMOS technology and may lead to better power efficiency than existing CMOS implementations. However, better power efficiency cannot be readily achieved by simply replacing TFETs with MOSFETs, and greater effort is needed. We leave this to future work.

ACKNOWLEDGMENT

REFERENCES

[1] A.C. Seabaugh, Q. Zhang, "Low-voltage tunnel transistors for beyond CMOS logic," *Proc. IEEE*, vol. 98, no. 12, pp. 2095-2110, Dec. 2010.

[2] H. Lu, A. Seabaugh, "Tunnel Field-Effect transistors: state-of-the-art," *IEEE J. Electron Devices Society*, vol. 2, no. 4, pp. 44-49, Jul. 2014.

[3] K. Tomioka, *et al.*, "Steep-slope tunnel field-effect transistors using III-V nanowire/Si heterojunction," in *VLSI Symp. Tech. Dig.*, 2012, pp. 47-48.

[4] S. Mookerjea, *et al.*, "Effective capacitance and drive current for tunnel-FET (TFET) CV/I estimation," *IEEE TED*, 56(9), pp. 2092-8, 2009.

[5] P. Ghedini der Agopian, *et al.*, "Experimental comparison between trigate p-TFET and p-FinFET analog performance as a function of temperature," *IEEE Trans. Electron Devices*, vol. 60, no. 8, pp. 2493-2497, Aug. 2013.

[6] S.O. Koswatta, *et al.*, "Performance comparison between pin tunneling transistors and conventional MOSFETs," *IEEE TED*, 56(3), pp. 456-65, 2009.

[7] V. Saripalli, *et al.*, "An energy-efficient heterogeneous CMP based on hybrid TFET-CMOS cores," *DAC*, pp. 729-734, 2011.

[8] U.E. Avci, *et al.*, "Comparison of performance, switching energy and process variations for the TFET and MOSFET in logic," in *VLSI Symp. Tech. Dig.*, pp. 124-125, 2011

[9] Y. Lee, *et al.*, "Low-Power circuit analysis and design based on hetero-junction tunneling transistors (HETTs)," *IEEE TVLSI*, 21(9), pp. 1632-43, 2013.

[10] K. Swaminathan, *et al.*, "An Examination of the Architecture and System-level Tradeoffs of Employing Steep Slope Devices in 3D CMPs'," *ISCA*, p. 241-252, 2014.

[11] B. Senale-Rodriguez, *et al.*, "Perspectives of TFETs for low power analog ICs," in *IEEE Subthreshold Microelectronics Conf.*, 2012, pp. 1-3.

[12] A.R. Trivedi, S. Carlo, S. Mukhopadhyay, "Exploring tunnel-FET for ultra low power analog applications: a case study on operational transconductance amplifier," *Design Automation Conf.*, 2013, pp. 1-6.

[13] H. Liu, *et al.*, "Tunnel FET-Based Ultra-Low Power, Low-Noise Amplifier Design for Bio-signal Acquisition," *ISLPED*, p. 57-62, 2014.

[14] H. Liu, *et al.*, "Tunnel FET-based ultra-low power, high sensitivity UHF RFID rectifier," *ISLPED*, p. 157-62, 2013.

[15] X. Li, *et al.*, "RF-Powered Systems Using Steep Slope Devices," *IEEE NEWCAS*, 2014.

[16] P. Dudek, and P. Hicks, "A general-purpose processor-per-pixel analog SIMD vision chip," *IEEE TCAS 1*, 52(1), pp. 13-20, Jan. 2005.

[17] R. Pawlowski, *et al.*, "A 530mV 10-lane SIMD processor with variation resiliency in 45nm SOI," *ISSCC*, 2012, pp. 492–494.

[18] S. Carey, *et al.*, "A 100000 fps vision sensor with embedded 535GOPS/W 256x256 SIMD processor array," in *Proc. Symp. VLSI Circuits (VLSIC)*, 2013, pp. C182-183.

[19] A. Rodriguez-Vazquez, *et al.*, "ACE16k: The third generation of mixed-signal SIMD-CNN ACE chips toward vSoCs," *IEEE TCAS-1*, 51(5), pp. 851-863, May 2004.

[20] S. Lee, *et al.*, "24-GOPS 4.5-mm digital cellular neural network for rapid visual attention in an object recognition SOC," *IEEE Trans. Neural Netw.*, vol. 22, no. 1, pp. 64-73, Jan. 2011.

[21] M. Di Fednferico, *et al.*, "SCDVP: A Simplicial CNN Digital Visual Processor," *IEEE TCAS-I*, 61(7), p. 1962-9, 2014.

[22] W. Miao, *et al.*, "A programmable SIMD vision chip for real-time vision applications," *IEEE J. Solid-State Cir.*, 43(6), pp. 1470-9, Jun. 2008.

[23] W. Zhang, *et al.*, "A programmable vision chip based on multiple levels of parallel processors," *IEEE JSSC*, 46(9), pp. 2132-47, 2011.

[24] S. Lee, *et al.*, "A 345mW heterogeneous many-core processor with an intelligent inference engine for robust object recognition," *ISSCC*, pp. 332–333, 2010.

[25] N. Cottini, *et al.*, "A 33uW 64x64 pixel vision sensor embedding robust dynamic background subtraction for event detection and scene interpretation," *IEEE JSSC*, 48(3), pp. 850-863, 2013.

[26] Y. Lee, *et al.*, "Low-Power Circuit Analysis and Design Based on Heterojunction Tunneling Transistors (HETTs)," *IEEE T. on VLSI*, 21(9), p. 1632-1643, 2013.

[27] J. Singh, *et al.*, "A Novel Si-Tunnel FET based SRAM Design for Ultra Low-Power 0.3V VDD Applications," *ASP-DAC*, p. 181-6, 2010.

[28] L. Chua and L. Yang, "Cellular Neural Networks: Theory," *IEEE TCAS*, 35(10), p. 1257-1272, 1988.

[29] A. Horvth, *et al.*, "Architectural Impacts of Emerging Transistors," *IEEE NEWCAS*, 2014.

[30] M. Nagata, *et al.*, "A smart CMOS imager with pixel level PWM signal processing," *VLSI Symp. Tech. Dig.*, 1999, pp. 141–144.

[31] K.R. Crounse and L. Chua, "Methods for image processing and pattern formation in Cellular Neural Networks: a tutorial," *IEEE T. on CAS*, 42(10), p. 583-601, 1995.

[32] G. Zhou, *et al.*, "Novel gate-recessed vertical InAs/GaSb TFETs with record high Ion of 180 A/m at VDS = 0.5 V," *IEEE Int. Electron Devices Meeting (IEDM)*, 10-13 Dec. 2012, pp.32.6.1-32.6.4.

[33] B. Sedighi, *et al.*, "Analog Circuit Design Using Tunnel-FETs," *IEEE TCAS-I*, 2014. DOI: 10.1109/TCSI.2014.2342371

[34] (2013). *The International Technology Roadmap of Semiconductors* [Online]. Available: http://www.itrs.net.

[35] S. Ando, "Consistent Gradient Operations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 3, pp. 252-265, Mar. 2000.