Application of Silicon-Germanium Source Tunnel-FET to Enable Ultralow Power Cellular Neural Network-Based Associative Memory

Amit Ranjan Trivedi, *Student Member, IEEE*, Suman Datta, *Fellow, IEEE*, and Saibal Mukhopadhyay, *Senior Member, IEEE*

***Abstract*— This paper studies the application of tunnel FET (TFET) in designing a low power and robust cellular neural network (CNN)-based associative memory (AM). The lower leakage, steeper switching slope, and higher output resistance of TFET are exploited in designing an ultralow-power TFET-based operational transconductance amplifier (OTA). A TFET-OTA is utilized as a programmable synaptic weight multiplier for CNN. The ultralow-power of TFET-OTA enables a higher connectivity network even at a lower power, and thereby improves the memory capacity and input pattern noise tolerance of CNN-AM for low power applications. The TFET-based higher connectivity CNN also exploits the unique characteristics of TFET to improve the throughput efficiency of CNN-AM.**

***Index Terms*— Associative memory (AM), cellular neural network (CNN), tunnel FET (TFET), ultralow-power computing.**

I. Introduction

THE Hopfield associative memory (AM) [1] provides a computation paradigm based on collective computing by a large number of equivalent computing elements interfaced by programmable interconnects. The unique ability of AM is to remember the relationship between two patterns as shown in Fig. 1(a). Applications of AM have been investigated in solving problems, such as character/face recognition, pattern classification, database search, and understanding/replicating cerebral activities as listed in Table I [2]–[5]. While performance requirement for the problems such as recognition and classification are moderate, an ultralow power of AM can enable solving these complex problems in a low power platform, such as a mobile system-on-a-chip. Ultimately, an ultralow-power of AM can also enable an ambitious goal of a very large scale AM computing, such as in a mammalian brain with 10^{10} neurons (biological computing elements) and 10^{14} synapses (biological interconnects), at sustainable operating power. Hence, a critical requirement for an AM computing platform is to minimize its power while meeting the throughput and performance demands. While operating at lower power, AM should be able to correctly identify the correlation (referred to as a successful recall), even under noise in the input pattern and maximize the total number of stored associations (defined as memory-capacity).

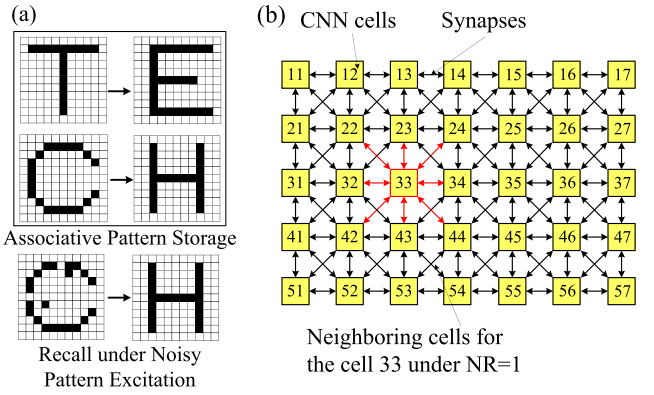

Fig. 1. (a) Heteroassociative storage of patterns in an 11 × 11 CNN array, and successful recall with noisy input C. (b) CNN array with identical locally interconnected cells, NR = 1.

TABLE I
Applications of AM

Applications	Typical performance	Ref
Face recognition	~Hz – KHz	[2]
Character recognition	~Hz – KHz	[3]
Classification	~Hz – KHz	
Replicating and understanding cerebral activities	~Hz	[4]
Database search	~KHz – MHz	[5]

Cellular neural network (CNN) has been investigated for AM applications [6], [7]. A CNN is composed of a set of identical computing elements (cells) organized as a 2-D-array where each cell is interfaced to its local neighbors using programmable synaptic-weight multipliers [8] [Fig. 1(b)]. This local connectivity makes CNN an attractive hardware platform, particularly in advanced nanometer nodes where interconnect scaling is challenging [9]. The algorithmic analysis of CNN-AM has shown that increasing the cell-to-cell connectivity, i.e., having more connections per cell, improves

Manuscript received May 5, 2014; revised July 29, 2014; accepted September 2, 2014. Date of publication September 30, 2014; date of current version October 20, 2014. The review of this paper was arranged by Editor Y.-H. Shih.

A. R. Trivedi and S. Mukhopadhyay are with the Department of Electrical and Computer Engineering, Georgia Institute of Technology, Atlanta, GA 30332 USA (e-mail: amitrt@gatech.edu; saibal@ece.gatech.edu).

S. Datta is with the Department of Electrical and Computer Engineering, Pennsylvania State University, University Park, PA 16801 USA (e-mail: sdatta@engr.psu.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/TED.2014.2357777

successful recall and memory-capacity [10]. A higher cell-to-cell connectivity implies a nonlinear increase in the number of synaptic-weight multipliers per cell. Therefore, the ability to design ultralow-power synaptic-weight multiplier becomes critical for low power AM.

A variety of CMOS-based CNN implementations have emerged [11]–[14]. However, power reduction of CMOS-CNN is constrained due to the limited switching slope and higher leakage current of MOSFET. The tunneling devices have been explored as potential alternatives to MOSFET for implementing low-power CNN cells. For example, reduced complexity CNN cells with resonant tunneling diodes have been studied [15], [16]. More recently, CNN cells with TFETs have been reported [17], [18]. While these prior works have primarily studied CNN cells; in a CNN-AM, due to the high connectivity requirement reducing the power dissipation in synaptic-weight multipliers becomes equally, if not more, crucial.

In our prior work, we presented the benefits of TFET for ultralow power analog design [19]. Very low OFF-current of TFET enables ultralow power operation of analog design, and steeper switching slope of TFET enables greater transconductance (GM) even at a low power. Specifically, a TFET based ultralow power operational transconductance amplifier (OTA) was demonstrated. This paper utilizes a TFET-OTA as a synaptic-weight multiplier for CNN, and explores a TFET-based low power, robust, and high performance CNN-AM platform. We focus on silicon (Si) channel TFET with Si_xGe_{1-x} ($x = 0.5 - 0.7$) tunnel junction near the source region, due to their excellent ON-OFF ratio over small gate voltage swing. It should be noted that Si-based TFETs are less suitable for digital applications due to their low ON-current as compared with heterojunction TFETs with alternate (e.g., InAs [20]) channel materials. However, in this paper, we show that silicon-based TFET, even with its low digital performance, can realize a high performance and robust AM computing platform.

The rest of this paper is organized as follows. Section II presents CNN-AM. Section III discusses properties of Si_xGe_{1-x} TFET. Section IV presents the simulation results for TFET-based CNN-AM. Section V discusses various technological aspects of a TFET-CNN-AM. Finally, the conclusions are drawn in Section VI.

II. CNN-Based AM

Fig. 1(b) shows a CNN platform. The CNN consists of an array of cells. An analog implementation of a CNN cell is shown in Fig. 2 [11]. Each cell (C_{ij}) consists of three nodes: the input (u_{ij}), state (x_{ij}), and output (y_{ij}), and OTA is used to enable intercell interaction. Underlying dynamics of the CNN cell C_{ij} is given by [8]

$$C\frac{dV(x_{ij})}{dt} = -\frac{V(x_{ij})}{R} + \sum_{kl\epsilon S_{ij}} A_{kl,ij}V(y_{kl}) + \sum_{kl\epsilon S_{ij}} B_{kl,ij}V(u_{kl}) + I_{ij} \quad (1a)$$

$$V(y_{ij}) = f_{\text{act}}(V(x_{ij})) \quad (1b)$$

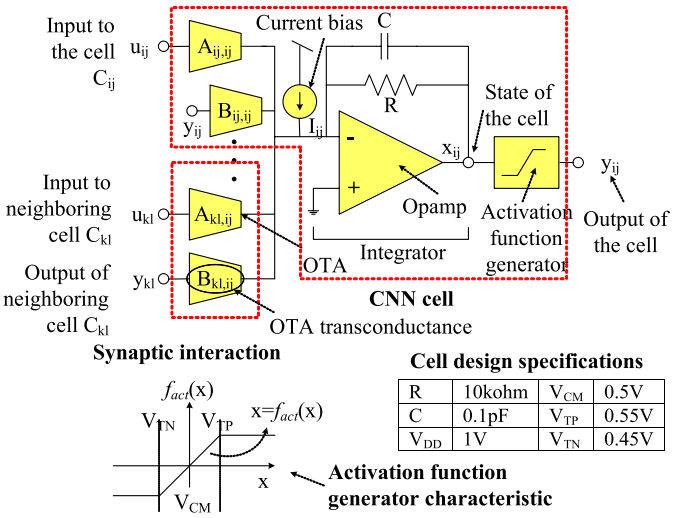

Fig. 2. CNN cell schematic and inset showing the output characteristics of activation function generator, and the cell design parameters.

where S_{ij} is the set of neighboring cells directly connected to the cell C_{ij}. Various other notations of the equation above were shown in Fig. 2. In CNN dynamics, at equilibrium the output saturates to the limits of the activation function generator, V_{TP} or V_{TN} [8]. The parameters $A_{kl,ij}$ and $B_{kl,ij}$ represent the feedback and feed-forward templates, and I_{ij} represents the bias term. The templates define the synaptic-weights and these weights are realized by GM values of a set of OTAs as shown in Fig. 2. On the other hand, the CNN cell is implemented using an integrator including RC elements, a bias generator, and a saturating function generator.

The AM operation is shown in Fig. 1(a), where association of letters T, C has been established to E, H, respectively. Various pixels correspond to the CNN cells, and black and white color corresponds to V_{TN} and V_{TP} voltages. When the input of CNN cells are excited with a distorted pattern of C, the output of cells accurately evolve to the respective pattern H at equilibrium. The design of CNN-AM involves the algorithmic synthesis of synaptic weights, $A_{kl,ij}$, $B_{kl,ij}$, and bias I_{ij}. For this synthesis, the Hebbian learning method from [7] is utilized in this paper. In general, the AM synaptic weights are space variant, i.e., the feed-forward/feed-back templates of CNN cell can vary from cell-to-cell. Distribution and storage of space varying synaptic weights is a critical challenge in CNN-AM.

A CNN architecture with a neighborhood-radius (NR) of one (NR = 1) is shown in Fig. 1(b). A higher NR design interconnects more cells together. For example, for NR = 2, all CNN cells in the dotted box will be directly connected to the cell C_{33}. Due to the challenges in distribution and storage of space variant synaptic-weights, several works consider low resolution or quantized implementation of the weights [21]. In Fig. 3(a), we compare the recall probability among various NR CNN-AM and at varying degree of quantization. A least square quantization is performed. It is observed that while a higher NR design is tolerant to quantization; in a low NR design, the recall probability degrades with decreasing quantization bits. Thus, a higher NR design can mitigate complications of accurate distribution/storage of space variant

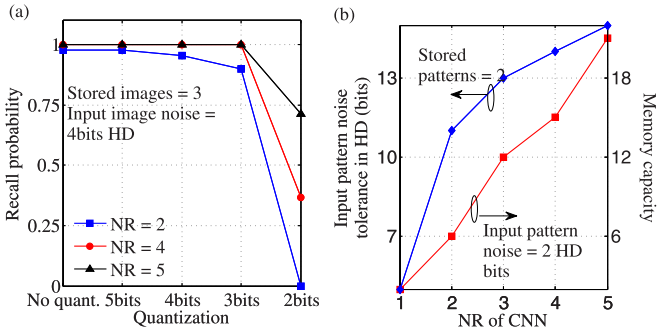

Fig. 3. (a) Recall probability at varying degree of quantization and for varying NR CNN-AM. (b) Memory capacity and input noise tolerance (in HD bits), at varying NR. Results are for 11 × 11 CNN-AM.

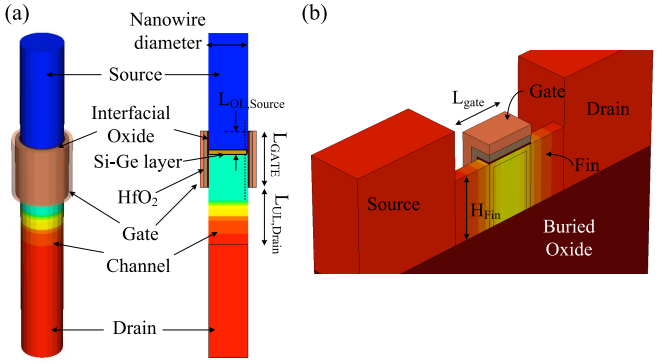

Fig. 4. (a) TFET schematic with Si_xGe_{1-x} tunnel junction at the source. (b) FinFET schematic.

synaptic-weights due to the amenability toward quantization. Fig. 3(b) shows the memory-capacity and input pattern noise tolerance [in Hamming distance (HD) bits] of CNN-AM for varying NR designs. A five bit quantization on synaptic-weights is considered. The input pattern noise tolerance is defined as the maximum number of corrupted bits in the input pattern while a successful recall is still performed. The memory-capacity is defined here as the maximum number of stored associations when the AM can successfully recall the stored associated pattern even under noise in the input pattern. As observed with increasing NR, both the memory-capacity and pattern noise tolerance of AM are significantly enhanced. Therefore, a higher NR CNN-AM, apart from its amenability to imprecise implementation, also exhibits a greater algorithmic quality.

However, a higher NR CNN-AM, due to its greater connectivity, also requires an increasingly higher number of synaptic-weight multipliers (i.e., OTAs) while the number of cells remains constant. Hence, a critical requirement to realize a robust but low-power AM with high algorithmic quality is the ability to scale down the OTA-power, so that a higher NR CNN can still operate under lower power constraint. In addition, while reducing the CNN power, the throughout efficiency (number of operations per s per power) of the AM platform also needs to be maintained or improved.

In the subsequent discussion, we study TFET for its utility in low-power high NR CNN-AM design. We observe that several unique electrical characteristics of TFET make it suitable for such a computing platform.

III. Si_xGe_{1-x} TFET

A vertical nanowire TFET with Si_xGe_{1-x} tunnel junction at source, similar to the one fabricated in [22], [Fig. 4(a)] is studied in this paper. Current conduction in TFET occurs through band-to-band-tunneling (BTBT), and the gate voltage controls the BTBT barrier width. We compare the electrical characteristics of TFET with FinFET [Fig. 4(b)]. The studied TFET and FinFET transistors are of similar gate length (45 nm), equivalent fin/wire perimeter, and oxide thickness [Table II]. The FinFET geometric specifications correspond to the 22 nm channel length technology, however, a larger channel length is considered here for analog applications.

TABLE II

TFET/FinFET SPECIFICATIONS: GEOMETRY AND SIMULATION MODELS

Specification	**n(p)-TFET**	**Specification**	**n(p)-FinFET**
L_{GATE}	45nm	L_{GATE}	45nm
Diameter	25nm	W_{FIN}	10nm
$L_{UL(OL),Drain(Source)}$	10nm	H_{FIN}	25nm
Source doping	$10^{20}/cm^3$	Source doping	$10^{20}/cm^3$
Drain doping	$5\times10^{18}/cm^3$	Drain doping	$10^{20}/cm^3$
High-κ Oxide	2nm	High-κ Oxide	2nm
Source (Drain) grad.	3(7)nm/dec.	Unified mobility model	[25]
Ge mole	0.5 (0.3)	Quantum confinement model	[23]
Si/Ge layer	5(8) nm		
Non-local BTBT	[23]	Mobility degradation at high-κ	[23]
TAT tunneling	[24]		

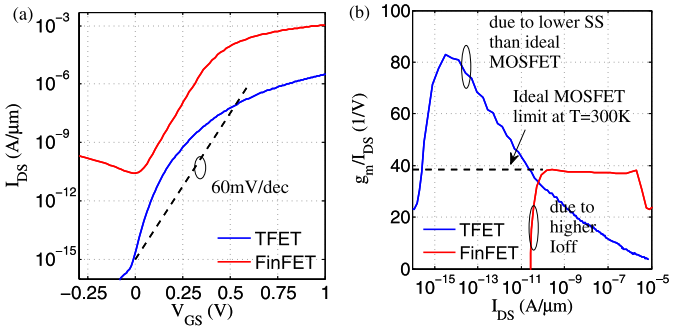

Fig. 5. Comparison of the electrical characteristics of n-TFET and n-FinFET. (a) I_{DS}–V_{GS}. (b) gm/I_{DS}–I_{DS} (V_{DS} = 1 V).

The characteristics of TFET and FinFET were extracted using Sentaurus device with the simulation models listed in Table II [23]–[25].

Fig. 5(a) shows that the ON-current of TFET is several orders of magnitude lower compared with FinFET, however, TFET also achieves much lower OFF-current and steeper switching slope. The minimum achievable OFF-current in FinFET is limited through subthreshold leakage and the other mechanisms such as gate-induced-drain-leakage [26]. The TFET achieves much lower OFF-current due to its built-in

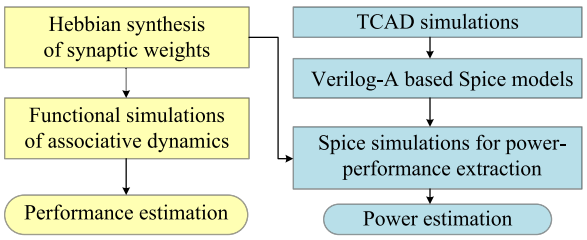

Fig. 6. Cohesive simulation methodology, integrating TCAD, SPICE, and functional simulations to extract CNN-AM characteristics at different technologies, TFET and FinFET.

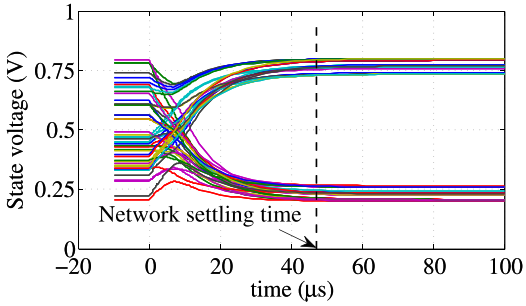

Fig. 7. Transient evolution of various cell state voltages, and the network settling time.

barrier (p-i-n). A much lower OFF-current of TFET enables a necessary power scaling of TFET-OTA for a high NR CNN.

Steeper switching-slope (SS) of TFET improves g_m per bias current (g_m/I_{DS}) for TFET [$g_m/I_{DS} = \log(10)/SS$]. As shown in Fig. 5(b), g_m/I_{DS} of TFET can exceed the thermal limit of MOSFET, q/kT. A sub-60 mV/decade steeper switching slope was also experimentally demonstrated in the similar nanowire TFET structures [27]. However, at a higher bias current as the switching slope of TFET degrades, g_m/I_{DS} of TFET drops as well. Hence, benefits of greater g_m/I_{DS} in TFET are limited to a smaller bias current. At very low I_{DS}, g_m/I_{DS} drops since I_{DS} is dominated by the trap assisted tunneling. Higher g_m/I_{DS} of TFET will facilitate a lower power in OTA, where an equivalent GM of the MOSFET-based design can be achieved at a lower bias power.

IV. TFET-Based CNN-AM

A. Simulation Methodology

We have developed an integrated methodology connecting device simulations using TCAD, circuit simulations using SPICE, and functional simulations using MATLAB Fig. 6. In CNN-AM, for a given cell resistance R, the Hebbian learning algorithm [7] determines the synaptic-weights (OTA-GMs) $A_{ij,kl}$, $B_{ij,kl}$, and bias, I_{ij}, to store correlation between the desired patterns. For these parameters, we use MATLAB to solve (1) and estimate the CNN-AM performance (recall speed, input pattern noise tolerance, and memory-capacity). Recall speed is defined by the duration when various state voltages of the CNN array saturate to 95% of their equilibrium values as shown in Fig. 7.

For a given synaptic GM specification, we use power-performance trace from OTA circuit simulation to estimate the OTA biasing power, P_{OTA} (GM). To simulate TFET- and FinFET-based OTA, electrical characteristics of TFET/FinFET are first extracted using Sentaurus-based TCAD simulations [as shown in Section III]. A Verilog-A-based table model interpolates these electrical characteristics across finely varying gate and drain bias voltages [19].

Using these Verilog-A table models, SPICE-based circuit simulation is performed to predict the power versus performance characteristics of TFET/FinFET-OTAs. The cell biasing current, I_{ij}, obtained through the Hebbian algorithmic synthesis [7], determines the power due to cell biasing. The voltage trace across cell capacitance, C, obtained through the functional simulations of (1), determines the dynamic power dissipation, $P_{dyn}(Vx_{ij})$, across C. Cell capacitance, C, is much larger than the other parasitic capacitances, hence, we ignore the effect of other parasitics. Due to predominance of OTAs in CNN, and for simplicity, we ignore the bias power contribution from the other cell components, integrator, and activation function generator. The net CNN power is given by

$$P_{CNN} = \sum_{ij}\left(P_{dyn}(V_{x,ij}) + V_{DD}\times I_{ij} + \sum_{kl\in S_{ij}} P_{OTA}(A_{kl}) + \sum_{kl\in S_{ij}} P_{OTA}(B_{kl})\right). \quad (2)$$

In (1), while scaling R and inversely scaling down the GMs, $A_{ij,kl}$, $B_{ij,kl}$, and bias, I_{ij}, does not affect the equilibrium states [i.e., $x_{ij}(t)\ \forall\ dx_{ij}(t)/dt = 0$]. Thereby, with invariant equilibrium states, the algorithmic quality of AM operation is retained. Hence, while retaining the algorithmic quality, increasing the cell resistance, R, enables reducing the cell bias current, I_{ij}, and OTA-GMs, and in turn, the OTA bias power, P_{OTA}(GM). However, with increasing R, the cell time constant, $R\times C$, increases, and therefore, recall time also increases. The above power scaling approach is applied to explore CNN-AM across power-performance while retaining its algorithmic quality.

B. TFET-Based OTA

Schematic of an OTA is shown in Fig. 8(a), where OTA consists of the transconductance generator (TG) and current summer, and generates an output current, I_{OUT}, proportional to its input voltage, V_{IN}. The OTA-GM is controlled by its bias current, I_{BIAS}. A cross coupled configuration at the TG stage, as shown in the figure, expands linearity of OTA to the threshold limits, V_{TN} to V_{TP}, of the activation function. The ratio of the OTA-GM to its bias power, P_{OTA}, can be expressed as

$$\frac{GM}{P_{OTA}} = \frac{1}{2}\times\frac{1}{V_{DD}}\times\sum_{i=1}^{2}\left(\frac{g_m(M_{ia})}{I_{DS}(M_{ia})}\frac{K}{K+1} + \frac{g_m(M_{ib})}{I_{DS}(M_{ib})}\frac{1}{K+1}\right)\times\frac{MR}{MR+1}. \quad (3)$$

Here, $g_m(M_i)$ and $I_{DS}(M_i)$ are the realized GM and bias current in transistor, M_i, respectively; K is the transistor width ratio for M_{1a} to M_{1b} (M_{2a} to M_{2b}), and MR is the mirror ratio of current summation stage, i.e., the ratio of width for M_{4b} to M_{4a} (M_{4d}–M_{4c}).

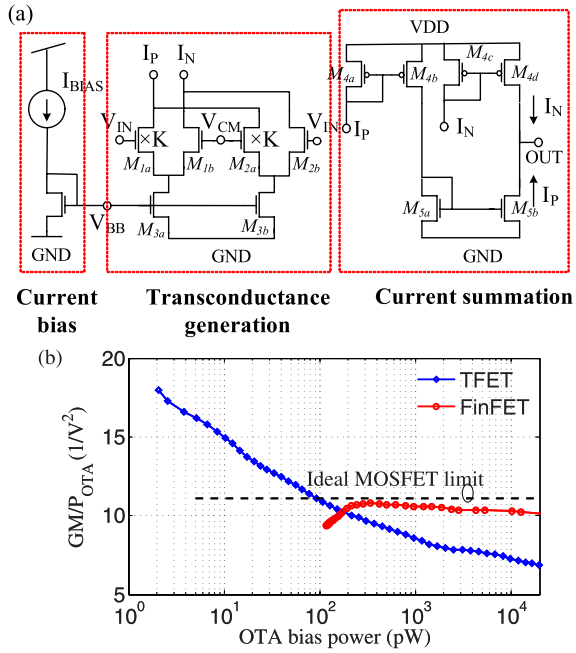

Fig. 8. (a) OTA schematic. (b) Power efficiency (GM/P_{OTA}) and power scaling comparison between TFET- and FinFET-OTA.

TABLE III
OTA DESIGN PARAMETERS

Transistor	# of Fins (Nanowires) in FinFET (TFET)
M_{1b} (M_{2b})	4
M_{3a} (M_{3b})	20
K	3 (FinFET) * 2-6 (TFET)
M_{4a} (M_{4c})	4
M_{4b} (M_{4b})	20
$M_{5a\text{-}d}$	**

* For TFET due to varying SS, K is varied accordingly to maintain linearity range = 100mV

** Designed to match output resistance with $M_{4a\text{-}d}$ in TFET and FinFET

In Fig. 8(b), considering the parameters in Table III, we compare GM of TFET-OTA and FinFET-OTA across bias power by varying I_{BIAS}. From (3), higher g_m/I_{DS} of OTA input transistors improves the GM/P_{OTA} of OTA. At a lower bias current, since g_m/I_{DS} of TFET is higher than MOSFET [Fig. 5(b)], TFET-OTA has higher GM/P_{OTA} than FinFET-OTA under low power operation. Below 100 pW bias power, TFET-OTA has higher GM/P_{OTA} than even an ideal MOSFET-based OTA. The ideal MOSFET-OTA characteristics are extracted considering switching slope = 60 mV/decade. With its higher GM/P_{OTA}, the TFET-OTA can operate at a lower power than MOSFET-OTA while still achieving an equivalent GM. Furthermore, due to an ultralow OFF-current of TFET, TFET-OTA can operate down to <5 pW power, while FinFET-OTA is scalable only till ∼200 pW. Power scaling of FinFET-OTA is limited due to the higher OFF-current in FinFET, and below 200 pW, the leakage current overwhelms the GM current. However, at a higher bias condition as g_m/I_{DS} of TFET degrades, FinFET-OTA achieves higher GM/P_{OTA} than TFET-OTA for such high power/performance operation. Therefore, TFET-OTA achieves higher GM/P_{OTA} than FinFET-OTA only under low-power operation. Studying a TFET-OTA design for neural amplifier, we earlier reported similar benefits of TFET-OTA [19].

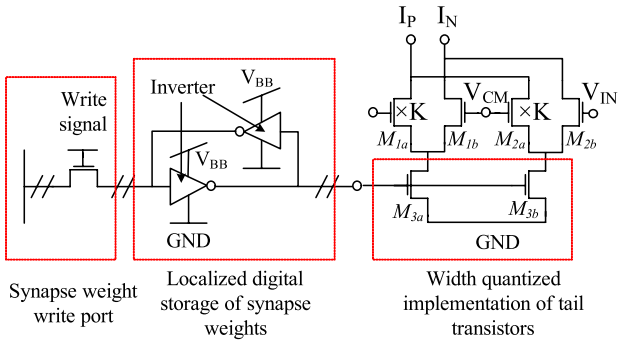

Fig. 9. Circuit scheme to locally store and implement quantized synaptic weights.

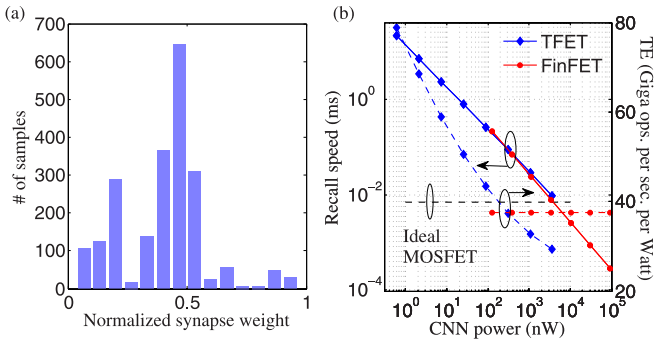

Fig. 10. (a) Distribution of synapse weights. (b) Recall speed and TE between TFET- and FinFET-CNN-AM across CNN power.

C. TFET-Based CNN-AM

We compare the power-performance characteristics of TFET and FinFET-based CNN-AM with NR of one. The test case of Fig. 1(a) is considered. A five bit least square signed quantization of the synaptic-weights is considered due to the complications of accurate analog weight storage. With the quantized synaptic-weights, the circuit schematic of Fig. 9 will enable digital storage/distribution of space varying analog synaptic-weights. Here, the tail transistors of OTA, M_{3a}, and M_{3b}, are implemented with quantized widths. And, a number of instances of M_{3a} and M_{3b} are selected to implement the tail biasing of OTA depending on the localized storage of synaptic-weights. Simulation methodology as described in Section IV-A is utilized. Best-fit power-performance characteristics of TFET and FinFET-OTA as shown in Fig. 8(b) are utilized. In Fig. 10(a), a distribution of quantized synaptic-weights is shown. The power-scaling approach scales each of these coefficients inversely proportional to R, and therefore, the bias power to realize corresponding GMs also reduces. However, as noted in Fig. 8(b), the minimum power and minimum realizable GM in TFET- and FinFET-OTA is limited; this, in turn, limits the power-scaling through the earlier approach and the minimum operational power for CNN. In Fig. 10(b), the TFET-CNN-AM operates down to ∼nW power due to the ultralow power scalability of TFET-OTA, and ∼ms recall time at this power will still be

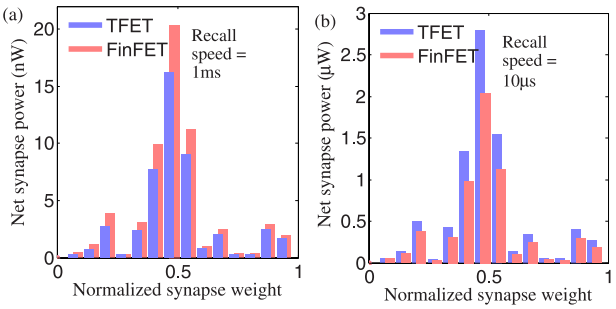

Fig. 11. Synapse power distribution and comparison between TFET- and FinFET-CNN-AM at net iso-power cases for (a) low and (b) high performance application.

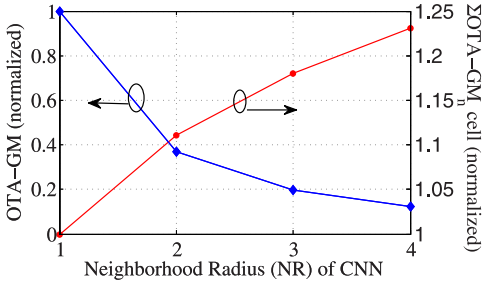

Fig. 12. OTA-GM and net CNN cell GM across NR for iso-powered CNN designs.

useful for recognition and classification applications in low power systems (Table I).

The net OTA bias power distribution at varying synaptic-weights for TFET- and FinFET-CNN-AM is shown in Fig. 11(a) for low performance operation (recall speed = 1 ms). A higher GM/P_{OTA} of TFET-OTA at low power attributes to lower bias power across synaptic-weights in TFET-CNN-AM than FinFET-CNN-AM. However, at a higher power since GM/P_{OTA} of TFET-OTA is lower, the TFET-CNN-AM has higher power than FinFET-CNN-AM for such high performance operation [Fig. 11(b), recall speed = 10 μs]. Therefore, a TFET-CNN-AM can only achieve higher throughput efficiency (TE) than FinFET-CNN-AM as long as the performance constraints are low.

In Fig. 10(b), the TE of TFET-CNN-AM and FinFET-CNN-AM is compared across recall speed and CNN power. Since, GM/P_{OTA} of TFET-OTA improves at lower bias current [Fig. 8(b)], the TE of TFET-CNN-AM also improves at lower power. Below 200 nW, TFET-CNN-AM has better TE than even an ideal MOSFET (i.e., with switching slope = 60 mV/decade)-based CNN-AM. However, note that under subthreshold operation, due to a constant switching slope of FinFET, GM/P_{OTA} of FinFET-OTA is invariant across power [Fig. 8(b)], hence, the TE of FinFET-CNN-AM is also invariant across power under subthreshold operation of FinFET-OTA.

D. Improving TFET-CNN-AM by High NR Design

The TE of a TFET-CNN-AM can be improved by a higher NR design. In a higher NR design, there are more OTAs per cell. Thus, at a given total power for CNN, in higher NR design, the power allocated for each of the OTA is lesser, and the realized OTA-GM is lower [Fig. 12]. However, as

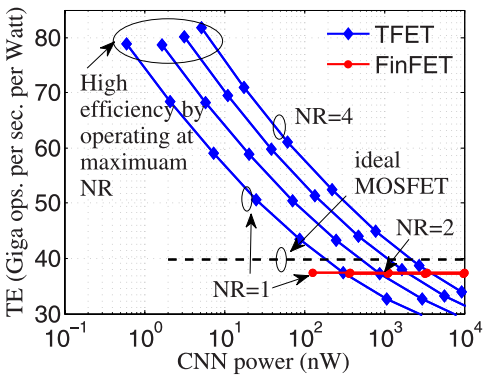

Fig. 13. TE of TFET- and FinFET-CNN-AM at varying NR.

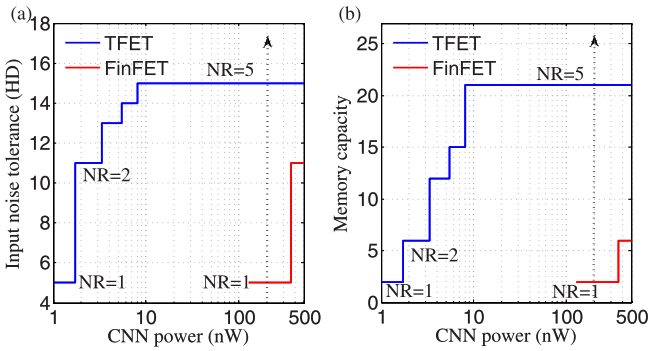

Fig. 14. For maximum NR operation between TFET and FinFET CNN-AM at varying power. (a) Input pattern noise tolerance. (b) Memory capacity.

shown in Fig. 8(b), GM/P_{OTA} of TFET-OTA increases at a lower power. Therefore, at the same total power, in a higher NR design, TFET-OTAs operate at a more energy-efficient point. Due to this unique characteristic of TFET, and with higher OTA count in a higher NR CNN, a higher net GM ($\sum$OTA-GM$_n$) can be realized for a higher NR TFET-CNN cell at the similar power [Fig. 12]. In Fig. 13, TE characteristics of TFET-CNN-AM are demonstrated for higher NR designs. Although, a higher NR TFET-CNN-AM design is less scalable in power, it achieves better TE at the similar power. Also, note that these TE benefits of high NR design are unique to TFET-CNN-AM. Due to a constant GM/P_{OTA} of FinFET-OTA, the TE in FinFET-CNN-AM is limited, and a higher NR (NR = 2) does not improve the TE.

Therefore, an optimal approach to exploit the higher GM/P_{OTA} of TFET-OTA is by operating the TFET-CNN-AM at the maximum NR given the CNN power scaling limitations as shown in Fig. 13. Furthermore, a higher NR operation can also achieve the higher algorithmic quality [Fig. 3]. In Fig. 14, utilizing a maximum NR TFET- and FinFET-CNN-AM, the algorithmic quality is shown at varying power. For Fig. 14(a), the test case of Fig. 1(a) is considered. In Fig. 14(b), for the memory-capacity test, apart from the patterns of Fig. 1(a), varying count of additional random binary patterns are considered for synthesis and recall. The TFET, by enabling a higher NR CNN-AM even at a lower power, enables a higher algorithmic quality. At 200 nW, while TFET-based NR = 5 CNN-AM enables HD noise tolerance of 15 bits and memory-capacity of 21 patterns, the FinFET-based NR = 1 CNN-AM only achieves 5 bits and two patterns, respectively.

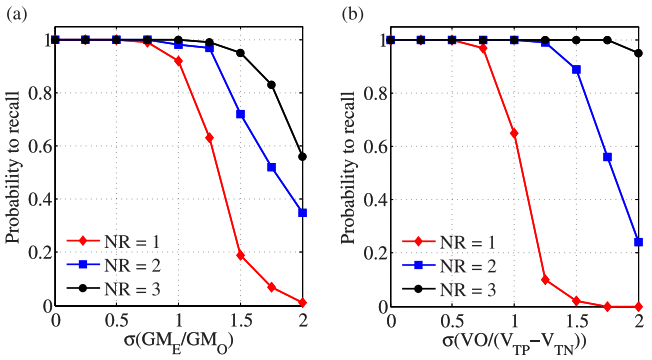

Fig. 15. Probability to recall across NR architectures with increasing variability in (a) GM and (b) VO (normalized by saturation limit of f_{sat}).

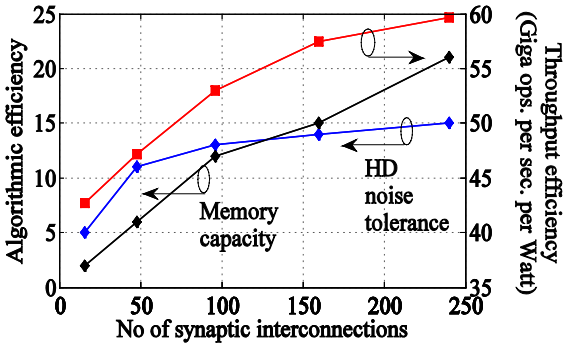

Fig. 16. Saturating HD noise tolerance and throughput efficiency at increasing number of synaptic interconnections.

E. Impact of Process Variations

We study the impact of transistor variability on AM quality considering the effect of higher connectivity through functional simulations of CNN-AM. Transistor variability induces non-idealities in OTA, as offset voltage (VO) and GM error (GM_E rather than the designed GM_O). Considering a normal distribution on VO and a log-normal distribution on GM_E/GM_O, we introduce the corresponding distribution to synaptic-weights, $\text{A}_{ij,kl}$ and $\text{B}_{ij,kl}$, in the functional simulations of (1), and study the impact of transistor variability. A log-Normal distribution for GM_E/GM_O is used due to the exponential sensitivity of current to gate voltage in both subthreshold FinFET and TFET. Higher transistor variability induces a greater variability in VO and GM. In Fig. 15, a higher NR CNN shows significant resiliency against increasing $\sigma(\text{GM}_O/\text{GM}_E)$ and $\sigma(\text{VO})$. A higher NR CNN is more robust in AM operation to begin with (Fig. 3), and the design softly fails with the higher unreliability of OTAs. Hence, a higher NR CNN design, apart from greater TE and algorithmic quality, will also be resilient to process defects.

V. Discussion

A. Implementation Complexity in a Higher NR TFET-CNN

The TFET enables implementation of a higher NR CNN even at a low power, and thus, provides opportunities for a low power AM design with higher algorithmic quality, process variation resiliency, and higher TE. However, a higher NR implementation of CNN will also increase the complexity and area of implementation. In Fig. 16, we demonstrate algorithmic and TE of TFET-CNN-AM with increasing number of synaptic interconnections at varying NR. Note that HD noise tolerance and TE increment with higher number of synaptic interconnections begins to saturate, meanwhile, the network complexity increases proportionally to the count of interconnections. Hence, due to the area and complexity constraints in an AM design, the optimal NR will be limited.

Various technological innovations such as a vertical and low footprint implementation of TFET [22] can reduce the network complexity and area requirement, and enhance the optimal NR of design. Simultaneously, novel circuit techniques can be explored to mitigate the impact of higher interconnect capacitance at higher NR. For example note that, in Fig. 2, a higher interconnect capacitance does not significantly affect the functionality of CNN, since the output node of OTAs is regulated through integrator, and thus, the effect of parasitic capacitances is suppressed.

B. TFET CNN for Image Processing Versus AM

Notably, CNN-AM exploits the TFET characteristics differently, and perhaps more effectively, than in simple CNN-based image processing applications [17], [18], [28], [29]. In our prior work [28], for TFET-CNN-based image processing applications, TFETs are exploited to increase the number of nodes in the network for a given power (by reducing the OTA power) and keeping the connectivity (NR = 1) constant. More nodes lead to higher parallelism, and hence, higher TE by exploiting parallelism in CNN-based image processing. A lower OFF-current of transistors becomes more critical in expanding the network size and exploiting the parallelism benefits.

On the other hand, a TFET-CNN-AM is significantly benefited both from the low OFF-current as well as higher g_m/I_{DS} of TFET as presented in this paper. It is shown that lower OFF-current, and hence, the low-power of TFET-OTAs are better exploited by increasing the NR in CNN-AM (under power constraints). From an algorithmic perspective, higher NR improves algorithmic quality (noise tolerance and memory-capacity). Interestingly, from an algorithmic perspective, TE is expected to be independent of NR; as observed in case of the FinFET-CNN-AM (Fig. 13). However, as explained in Fig. 12, the higher g_m/I_{ds} of TFET changes the cell dynamics at higher NR leading to higher TE in TFET-CNN-AM.

Furthermore, unlike space invariant templates in CNN-based image processing, synapse weights in AM depend on the pattern itself, are space variant, and can vary widely in magnitude (e.g., $\sim$15$\times$ variation between the largest and smallest synapse weight in Fig. 11). Therefore, the interactions between the CNN-AM power and the variable switching slope in TFET becomes significantly more important to consider as explained in Fig. 11.

VI. Conclusion

This paper presents the potential of $\text{Si}_x\text{Ge}_{1-x}$-TFET in designing efficient and robust CNN-AM. A lower OFF-current of $\text{Si}_x\text{Ge}_{1-x}$-TFET enables lower power operation of CNN

synaptic-weight multiplier enabling a higher NR CNN for a given power. A higher NR CNN-AM improves algorithmic quality of AM. In addition, higher NR also improves TE of TFET-CNN-AM at a constant power, thanks to the steeper switching slope (higher g_m per unit bias power) of TFET. Increasing performance benefits along with the increasing tolerance against process variability at higher NR indicates building higher NR CNN-AM as the suitable approach to build large scale higher performance and robust AM implementation. However, increasing implementation area along with interconnect complexity can ultimately limit the NR of implementation. The application of TFET in CNN-AM also reveals more involved device-algorithm interactions, than what observed in TFET-CNN-based image processing [28]. Increasing NR, as performed here for CNN-AM, more effectively exploits unique TFET and AM characteristics for quality, noise tolerance, speed, power, and TE. Future work needs to consider the design challenges of higher NR CNN-AM including area and interconnect. The vertical orientation of Si_xGe_{1-x} nanowire TFET alleviates the area constraints, feature size scaling of the nanowire will be important. The local connectivity in CNN help to partially mitigate the challenge of global interconnects. The analog communication also reduces the density; however, the requirements of higher local interconnect density for higher NR design will still be a challenge.

REFERENCES

[1] J. Hopfield, "Neurons with graded response have collective computational properties like those of two-state neurons," *Proc. Nat. Acad. Sci. United States Amer.*, vol. 81, no. 10, pp. 3088–3092, 1984.

[2] M. Namba and Z. Zhang, "Cellular neural network for associative memory and its application to Braille image recognition," in *Proc. Int. Joint Conf. Neural Netw.*, 2006.

[3] B. Baird, M. W. Hirsch, and F. Eeckman, "A neural network associative memory for handwritten character recognition using multiple Chua characters," *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.*, vol. 40, no. 10, pp. 667–674, Oct. 1993.

[4] Y. V. Pershin and M. Di Ventra, "Experimental demonstration of associative memory with memristive neural networks," *Neural Netw.*, vol. 23, no. 7, pp. 881–886, 2010.

[5] H. J. Mattausch, W. Imafuku, A. Kawabata, T. Ansari, M. Yasuda, and T. Koide, "Associative memory for nearest-Hamming-distance search based on frequency mapping," *IEEE J. Solid-State Circuits*, vol. 47, no. 6, pp. 1448–1459, Jun. 2012.

[6] N. N. Aizenberg and I. N. Aizenberg, "CNN based on multi-valued neuron as a model of associative memory for grey scale images," in *Proc. Int. Workshop Cellular Neural Netw. Appl.*, Oct. 1992, pp. 36–41.

[7] P. Szolgay, I. Szatmari, and K. Laszlo, "A fast fixed point learning method to implement associative memory on CNN's," *IEEE Trans. Circuits Syst. I, Fundam. Theory Appl.*, vol. 44, no. 4, pp. 362–366, Apr. 1997.

[8] L. O. Chua and L. Yang, "Cellular neural networks: Applications," *IEEE Trans. Circuits Syst.*, vol. 35, no. 10, pp. 1273–1290, Oct. 1988.

[9] A. Ceyhan and A. Naeemi, "Cu interconnect limitations and opportunities for SWNT interconnects at the end of the roadmap," *IEEE Trans. Electron Devices*, vol. 60, no. 1, pp. 374–382, Jan. 2013.

[10] C. M. Newman, "Memory capacity in neural network models: Rigorous lower bounds," *Neural Netw.*, vol. 1, no. 3, pp. 223–238, 1988.

[11] L. Wang, J. P. de Gyvez, and E. Sanchez-Sinencio, "Time multiplexed color image processing based on a CNN with cell-state outputs," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 6, no. 2, pp. 314–322, Jun. 1998.

[12] E. P. Santana, R. C. S. Freire, and A. I. A. Cunha, "A CMOS analog four-quadrant multiplier for CNN synapses," in *Proc. 8th Int. Caribbean Conf. Devices, Circuits Syst.*, 2012, pp. 1–4.

[13] F. A. Khanday, C. Psychalinos, and N. A. Shah, "Square-root-domain realization of single-cell architecture of complex TDCNN," *Circuits, Syst., Signal Process.*, vol. 32, no. 3, pp. 959–978, 2013.

[14] C.-S. Poon and K. Zhou, "Neuromorphic silicon neurons and large-scale neural networks: Challenges and opportunities," *Frontiers Neurosci.*, vol. 5, p. 108, Sep. 2011.

[15] S.-R. Li, P. Mazumder, and L. O. Chua, "Cellular neural/nonlinear networks using resonant tunneling diode," in *Proc. 4th IEEE Conf. Nanotechnol.*, Aug. 2004, pp. 164–167.

[16] A. Khitun and K. L. Wang, "Cellular nonlinear network based on semiconductor tunneling nanostructure," *IEEE Trans. Electron Devices*, vol. 52, no. 2, pp. 183–189, Feb. 2005.

[17] I. Palit, X. S. Hu, J. Nahas, and M. Niemier, "TFET-based cellular neural network architectures," in *Proc. IEEE Int. Symp. Low Power Electron. Design*, Sep. 2013, pp. 236–241.

[18] I. Palit, B. Sedighi, A. Horvath, X. S. Hu, J. Nahas, and M. Niemier, "Impact of steep-slope transistors on non-von Neumann architectures: CNN case study," in *Proc. Design, Autom. Test Eur. Conf. Exhibit.*, Mar. 2014, pp. 1–6.

[19] A. R. Trivedi, S. Carlo, and S. Mukhopadhyay, "Exploring tunnel-FET for ultra low power analog applications: A case study on operational transconductance amplifier," in *Proc. 50th ACM/EDAC/IEEE Design Autom. Conf. (DAC)*, Jun. 2013, pp. 1–6.

[20] S. Mookerjea, R. Krishnan, S. Datta, and V. Narayanan, "Effective capacitance and drive current for tunnel FET (TFET) CV/I estimation," *IEEE Trans. Electron Devices*, vol. 56, no. 9, pp. 2092–2098, Sep. 2009.

[21] T. Lehmann, E. Bruun, and C. Dietrich, "Mixed analog/digital matrix-vector multiplier for neural network synapses," *Analog Integr. Circuits Signal Process.*, vol. 9, no. 1, pp. 55–63, 1996.

[22] A. Vandooren, D. Leonelli, R. Rooyackers, K. Arstila, G. Groeseneken, and C. Huyghebaert, "Impact of process and geometrical parameters on the electrical characteristics of vertical nanowire silicon n-TFETs," *Solid-State Electron.*, vol. 72, pp. 82–87, Jun. 2012.

[23] (2013). *Sentaurus Device*. [Online]. Available: http://www.synopsys.com/tools/tcad/Pages/default.aspx

[24] G. A. M. Hurkx, D. B. M. Klaassen, and M. P. G. Knuvers, "A new recombination model for device simulation including tunneling," *IEEE Trans. Electron Devices*, vol. 39, no. 2, pp. 331–338, Feb. 1992.

[25] D. B. M. Klaassen, "A unified mobility model for device simulation—I. Model equations and concentration dependence," *Solid-State Electron.*, vol. 35, no. 7, pp. 953–959, 1992.

[26] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits," *Proc. IEEE*, vol. 91, no. 2, pp. 305–327, Feb. 2003.

[27] R. Gandhi, C. Zhixian, N. Singh, K. Banerjee, and L. Sungjoo, "CMOS-compatible vertical-silicon-nanowire gate-all-around p-type tunneling FETs with $\leq$50-mV/decade subthreshold swing," *IEEE Electron Device Lett.*, vol. 32, no. 11, pp. 1504–1506, Nov. 2011.

[28] A. R. Trivedi and S. Mukhopadhyay, "Potential of ultralow-power cellular neural image processing with Si/Ge tunnel FET," *IEEE Trans. Nanotechnol.*, vol. 13, no. 4, pp. 627–629, Jul. 2014.

[29] A. R. Trivedi, M. F. Amir, and S. Mukhopadhyay, "Ultra-low power electronics with Si/Ge tunnel FET," in *Proc. Design, Autom. Test Eur. Conf. Exhibit. (DATE)*, Mar. 2014, pp. 1–6.

Amit Ranjan Trivedi (S'10) is currently pursuing the Ph.D. degree with the Georgia Institute of Technology, Atlanta, GA, USA.

He was a Research Internship Student with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA, in 2012, and the Intel's Circuit Research Laboratory, Hillsboro, OR, USA, in 2014. His current research interests include ultralow-power devices and circuits, and applications of emerging technologies in neuromorphic computing.

Suman Datta (F'13) received the B.S. degree in electrical engineering from IIT Kanpur, Kanpur, India, in 1995, and the Ph.D. degree in electrical and computer engineering from the University of Cincinnati, Cincinnati, OH, USA, in 1999.

He was the Joseph Monkowski Associate Professor of Electrical Engineering with Pennsylvania State University, University Park, PA, USA, in 2007, where he is currently a Professor of Electrical Engineering.

Saibal Mukhopadhyay (SM'11) received the Ph.D. degree in electrical and computer engineering from Purdue University, West Lafayette, IN, USA.

He was with the IBM Thomas J. Watson Research Center, Yorktown Heights, NY, USA, as a Research Staff Member. He joined the faculty of the Georgia Institute of Technology, Atlanta, GA, USA, in 2007. His current research interests include low-power circuit design at advanced nanometer nodes.