Ultra-low Power Electronics with Si/Ge Tunnel FET

Amit Ranjan Trivedi, Mohammad Faisal Amir, and Saibal Mukhopadhyay
Department of Electrical and Computer Engineering
Georgia Institute of Technology, Atlanta, GA, USA

Abstract—Si/Ge Tunnel FET (TFET) with its subthermal subthreshold swing is attractive for low power analog and digital designs. Greater $I_{on}/I_{off}$ ratio of TFET can reduce the dynamic power in digital designs, while higher $g_{m}/I_{on}$ can lower the bias power of analog amplifier. However, the above benefits of TFET are eclipsed by MOSFET at a higher power/performance point. Ultra low power scalability of the key analog and digital circuits, SRAM and operational transconductance amplifier (OTA), with TFET is demonstrated. Analyzing a TFET based cellular neural network, this work shows the feasibility of ultra-low-power neuromorphic computing with TFET.

Keywords—Tunnel FET, Low power design, SRAM, Operational transconductance amplifier, Cellular neural network.

I. INTRODUCTION

In the quest of achieving higher performance/functionality even at lower power, Tunnel FET (TFET) as an alternative to the conventional transistor, MOSFET, is being explored [1-17]. Unlike thermionic injection of carriers over the channel barrier as in a MOSFET, in TFET, the current conducts by band-to-band-tunneling (BTBT) of carriers. This paper reviews the potential of TFET in reducing the power of current MOSFET based circuits. Unlike MOSFET, TFET can achieve subthermal subthreshold swing (SS). A steeper SS in TFET enables achieving the same $I_{on}/I_{off}$ ratio as in a MOSFET, although, at an even lower supply voltage, $V_{DD}$ [15]. Thus, the supply voltage of TFET based digital circuits can be scaled down resulting in quadratic reduction in the dynamic power. Due to its diode like built-in barrier, TFET can also achieve very low off current as compared to a MOSFET. Thus, the standby power of TFET based digital circuits will also be very low. TFET with much lower power-delay-product compared to a MOSFET has been demonstrated [2, 4]. A steeper SS of TFET also benefits analog circuits, where for a current bias, $I_{DS}$, a higher transconductance, $g_{m}$ than the MOSFET can be achieved [18]. Therefore, the bias current/power of TFET based analog circuits can be scaled down while achieving the same transconductance as in a MOSFET based design. Furthermore, very low off current of TFET enables ultra-low bias current of designs and exploration of ultra-low power analog circuits [18]. In addition, the suppressed temperature dependence of current in TFET, especially as compared to the exponential dependence as in a subthreshold MOSFET, enhances the robustness of TFET based analog designs. Hence, the application of TFET to low power circuits is attractive for both the digital as well as analog designs.

However, a quantum-mechanical tunneling of carriers, as compared to the thermionic injection in a MOSFET, significantly limits the on-current of TFET. In Si-TFET, the on-current is limited ~nA/μm [19]. A lower bandgap material at the source/channel junction can reduce the barrier height, and increase the BTBT probability and on-current. Hence, the alternative channel materials of lower bandgap such as Ge [11], InAs [2, 12], InSb [20] have been investigated to achieve better on-current from TFET. On current as high as ~300μA/μm was achieved in Ge-TFET [7]. However, the on-current improvement in TFET with the lower bandgap material also comes at the expense of increased off current and worse short-channel effect [7, 21].

Despite its limited on-current a CMOS compatible silicon channel TFET can be of significant interest. Besides the cost effectiveness of silicon TFET, co-integration of TFET with CMOS can bring forth a variety of efficient circuit designs. For example, integrating TFET with CMOS, an efficient bandgap reference design was demonstrated [10]. Application of TFET as an ultra cut-off sleep transistor was suggested [16]. Efficient chip multiprocessor with co-integrated TFET and CMOS based cores were studied [22]. Efficient detection of bio-molecules with TFET was suggested [23]. Ultra-low power neural amplifier based on TFET was studied [18].

Ensuing discussion, in this work, will review the electrical characteristics of silicon channel Si/Ge TFET, compare/contrast features from an equivalent MOSFET, and review the challenges and recent improvements in the structure. We also explore two key designs, SRAM and operational transconductance amplifier, to exploit various unique features.

Fig. 1: (a) TFET schematic, and (b) BTBT of electrons in n-TFET.
of the TFET to improve the efficiency of these circuits for low power applications. Finally, feasibility of ultra-low power neuromorphic computing with the TFET is discussed.

II. TUNNEL-FET

A schematic of TFET is shown in Fig. 1a. Similar to a MOSFET, a TFET is also comprised of source, channel, and drain regions; however, unlike MOSFET, the doping of source and drain regions is different from each other. In an n-TFET, the source is doped as P+ type and the drain doping is N+, while in a p-TFET, the source is N+ doped and the drain is P+ doped. The channel in TFET is intrinsic. In an n-TFET, current conducts by the electron tunneling from the valence band in source to the conduction band in drain [Fig. 1b], while in a p-TFET, the holes tunnel from the conduction band in source to the valence band in drain. BTBT depends on the barrier height and width as shown in Fig. 1b. The barrier height is determined by the material bandgap at the source/channel junction, while the barrier width can be controlled by the gate electrode. As shown in Fig. 1b, in the off state (VGS = 0V), the barrier width is significantly large, and thereby, BTBT probability is very low, and the off-current in TFET is significantly low. Off-current as low as ~fA/μm has been measured in Si-TFET [24]. However, as the gate voltage increases, a significant band bending occurs at the source/channel junction and the barrier width decreases [Fig. 1b]. Hence, with the decreasing barrier width at the high gate voltage, a significant charge injection from source to channel and current conduction to drain occurs in TFET.

However, due to the higher bandgap of silicon, the barrier height of BTBT in silicon TFET is high, and the on-current is very low. A variety of techniques have been explored to enhance the on-current in silicon TFET. Boucart et al. suggested utilizing stress to lower the bandgap at the source/channel junction resulting in the on-current enhancement in TFET [25]. A thin Si/Ge layer was embedded at the source/channel junction, where due to the lower bandgap of Si/Ge layer, the on-current in TFET increases [26]. Bi-layer silicon TFET was investigated, where the p-i-n band bending of TFET is directly aligned with the gate electric field, and the improved gate electrostatics increases the on-current in TFET [27]. Source silicidation was investigated to increase the junction steepness at the source/channel junction to improve the on-current in TFET [1], while abrupt doping profile in axially doped silicon TFET was achieved [9]. Halo-doping at the source/channel junction [28] was utilized to improve the on-current in silicon TFET. Raised Ge source, silicon channel TFET with the higher on-current was shown [13].

A silicon channel vertical nanowire TFET with the gate all around and embedded Si/Ge layer at the source/channel junction was demonstrated in [29]. Here, we compare the electrical characteristics of this TFET with the equivalent channel length and oxide thickness FinFET. FinFET is a state-of-the-art MOSFET structure, which achieves superior characteristics than its bulk counterpart due to the better gate electrostatics of wrap around gate over the channel. The schematic of nanowire TFET is shown in Fig. 2, and various specifications are listed in Table I. The FinFET schematic is shown in Fig. 3, and design specifications are listed in Table II. TFET is also an ambipolar device which means the conduction can also occur at the reverse polarity gate bias due to the BTBT of carriers from the drain to source. To suppress this deleterious ambipolar behavior, especially for the digital designs, an asymmetry in the source and drain regions is utilized. Here, the drain region is lightly doped with gradual junction gradient to suppress BTBT at the drain/channel junction [29]. An underlap at the drain/channel junction was also proposed to suppress the ambipolar behavior [14], which also reduces the gate/drain parasitic capacitance. Note that these techniques to suppress ambipolarity are applied in the TFET nanowire studied in Fig. 2.

The electrical and capacitive characteristics of TFET and FinFET are extracted using TCAD simulator Sentaurus Device [30]. For the TFET, non-local BTBT model [30] with default parameters and trap assisted tunneling (TAT) model with the experimentally calibrated parameters [31] were used. A non-local BTBT model dynamically searches for the tunneling path at varying bias conditions along the valence band gradient [30]. For the FinFET, unified mobility models, scattering models for mobility degradation due to high-κ dielectric, and quantum confinement models were used [30]. The IDSS-VGS characteristics of these transistors are compared.
in Fig. 4a. The on-current in TFET is significantly limited as compared to FinFET. However, TFET also achieves much lower off-current (~fA/μm). Embedding a Si/Ge layer at the source/channel junction improves the on-current of Si/Ge-TFET when compared to Si-TFET. Also, as shown in [29], since the dopant diffusion is lower in Ge, a Si/Ge layer can improve the junction steepness and further boost the on-current in Si/Ge-TFET. However, due to the lattice mismatch between Si/Ge and silicon, stress imparted by the Si/Ge-layer limits the Ge mole-fraction (\(x\)) and thickness of Si/Ge-layer [32]. Si/Ge-TFET with Ge molefraction ~0.5 were shown in [29], and a higher defect density with the increasing molefraction was suspected.

As shown in the figure, TFET achieves sharper subthreshold swing (SS) than even the ideal MOSFET limit (in ideal MOSFET SS = 60mV/dec at room temperature). However, SS in TFET degrades at the increasing gate voltage. Therefore, higher \(I_{on}/I_{off}\) ratio of TFET over MOSFET is limited to the low supply voltages, and a MOSFET eventually outperforms TFET at a higher \(V_{GS}\) performance point. Trap-assisted-tunneling (TAT) current limits the off-current of TFET. Especially, for Si-TFET, the steeper part of \(I_{DS}-V_{GS}\) characteristic through BTBT is masked by the TAT current, and a sharper SS is not achieved. TAT was also shown to arise significant temperature dependence in the off-current [29]. The \(I_{DS}-V_{DS}\) characteristics of TFET and FinFET are compared in Fig. 4b. At an equivalent on-current, while a subthreshold FinFET shows saturation at \(V_{DS} > 100mV\), in TFET the saturation is delayed [33]. A higher saturation voltage of TFET will be a limitation to the output swing of TFET based analog amplifier.

Small signal transconductance, \(g_{m}\), of TFET and FinFET is compared at the varying current bias, \(I_{DS}\), in Fig. 4c. Since, the off-current in FinFET is limited to ~pA/μm, the transconductance of FinFET drops at such bias levels. Meanwhile, TFET is scalable to much lower bias. Furthermore, the \(g_{m}/I_{DS}\) of TFET increases at the lower bias current, since the steeper part of \(I_{DS}-V_{GS}\) is exposed. Due to the thermionic injection based conduction, \(g_{m}/I_{DS}\) in MOSFET has a fundamental limit to \(q/kT\). On the other hand, TFET with its BTBT based conduction doesn’t have such limitations, and at sub-pA bias \(g_{m}/I_{DS}\) in Si/Ge-TFET exceeds even the ideal MOSFET limit. Higher \(g_{m}/I_{DS}\) of TFET can reduce the power in analog designs where an equivalent \(g_{m}\) as a MOSFET can be achieved at a lower bias power. However, due to the increasing SS at a higher gate voltage, benefits of the higher \(g_{m}/I_{DS}\) are limited to a lower bias current [18]. Drain transconductance \(g_{ds}\) of the TFET and FinFET is compared in Fig. 4d, where a lower \(g_{ds}\) implies a higher output resistance \(r_0\). Since the current conduction in TFET is controlled by the source/channel junction rather than the channel/gate junction as in a MOSFET, the drain influence to current conduction in saturation reduces, and TFET can exhibit higher output resistance. In Fig. 4d, although, \(g_{ds}\) of TFET and FinFET is equivalent, also note that the diameter of TFET (= 70nm) is much larger than the fin thickness of FinFET (= 10nm). And, as shown in the figure, the TFET design with an equivalent lower diameter (here, 20nm) shows much higher \(r_0\). A higher output resistance with TFET can avoid complexity such as the cascode implementation and associated biasing power in analog designs.

The gate to source capacitance, \(C_{GS}\), of TFET and FinFET is compared in Fig. 4e, while, the gate to drain capacitance, \(C_{GD}\), is compared in Fig. 4f. Note that in contrast to a MOSFET, in TFET, \(C_{GD}\) constitutes majority of the gate capacitance. Unlike MOSFET, the band bending in TFET occurs at the source/channel junction, and thus, the drain voltage continues to significantly influence the channel.
charge, which reflects in the higher C_{GD} [33]. However, a higher C_{GD} of TFET is a challenge, since it increases the dynamic power dissipation, and it can also result in a higher overshoot in the transience of a digital circuit [33].

III. SILICON CHANNEL TFET BASED ULTRA LOW POWER CIRCUIT DESIGNS

Section II discusses the unique characteristics of TFET such as much lower off-current, lower SS, and higher g_m/IDS. The Si/Ge-TFET shows superior characteristics. The Si/Ge-TFET has higher on-current than the Si-TFET, yet it maintains equivalent off current. In this section, we explore a few key circuits, SRAM and operational transconductance amplifier, where the characteristic of Si/Ge-TFET can enhance the energy efficiency of the designs. A Verilog-A based table model of TFET is utilized for the circuit explorations [18]. Here, at first the electrical characteristics of a single transistor is obtained at finely varying operating condition, and a table model based on Verilog-A interpolates over the table to create a spice compatible compact model.

A. SRAM

The low leakage current of TFET makes it an ideal candidate for memories where transistors remain in the off-state for a large fraction of time. We have studied SRAM cell designed with TFETs (Fig. 5). Although the underlying read mechanism of this cell is similar to the standard 6T-SRAM cell with MOSFET; due to the asymmetry of TFET, an anti-parallel implementation of pass transistor becomes necessary. In Fig. 5b, we compare the maximum clock frequency of SRAM at varying power considering a read operation of the design. The (leakage + dynamic) power of an SRAM array (64×32 cells) is obtained at the varying supply voltage, VDD, while considering an activity factor of 1% for the array (i.e. 1 read operation out of 100 clock cycles). A bit line capacitance of 25fF is assumed, and read time is defined as the duration to discharge the bit line voltage by 50mV when the SRAM cell is activated through high RWL. In Fig. 5b, due to the higher leakage of FinFET, the FinFET based design is not scalable to sub-nW power, and performance suddenly drops. On the other hand, the TFET based SRAM can be scaled to the very low power (~pW). However, at a higher power/performance point, the efficiency of FinFET is superior. In Fig. 5c, we show the static noise margin (SNM) of the design at the varying SRAM power. SNM is extracted using the methodology described in [34]. At an equivalent power, due to the lower leakage, TFET enables sustaining the SRAM at a higher VDD, hence, the TFET based SRAM has a better SNM than the FinFET design. Therefore, a better SRAM yield can be obtained with TFET even under the limited power budget.

B. Operational transconductance amplifier

TFETs can reduce the power of an analog amplifier where the same transconductance as in a MOSFET based design can be achieved even at a lower power through higher g_m/IDS of TFET. In this section, we study the application of TFET in designing an OTA. OTA finds widespread applications in areas such as in neural amplifier [35], cellular neural network (CNN) [36], and non-linear functional synthesis [37]. The schematic of OTA is shown in Fig. 6a. In response to a differential input voltage at the nodes IN' and IN; the transconductance current, id_1 and id_2, is generated through the
transistors $M_1$ and $M_2$. This transconductance current is mirrored at the current-mirrors, $M_3$-$M_8$, and $M_9$-$M_{18}$, and supplied to the load connected at OUT. The DC transconductance of OTA ($G_m$) is defined in Fig. 6a, and in Fig. 6b, we plot $G_m$ generated by the TFET and FinFET based OTA at varying biasing power. The biasing power of OTA is varied by varying the bias voltage $V_{BB}$. Transistor parameters are the same for the TFET and FinFET based OTA. Due to its very low off current, the TFET based OTA such scaling is limited due to the higher off current in FinFET. Moreover, as the power, while in the FinFET based OTA such scaling is limited to very low power. Moreover, as the efficiency of TFET based OTA increases at the lower power [Fig. 6b], the TFET design is scalable to ~pW level. TFET OTA at ~pW power will be useful for the ultra-low-power scalability of TFET based OTA is attractive for the neuromorphic computing through cellular neural networks (CNN). A CNN consists of identical cells interacting through local interconnects (synapse) [39]. Architecture of a CNN of neighborhood radius (NR) = 1 is shown in Fig. 7a. A higher NR design of CNN is similarly synthesized by interconnecting more cells together. For example, for NR = 2, all cells within the dotted lines will be connected to the cell C33. The schematic of a CNN cell is shown in Fig. 7b [36], along with the considered design parameters. In CNN, the interaction between neighboring cells is governed through OTA. In response to the input and output voltage of the neighboring cells, the dynamics of CNN cell evolves, and under the stability constraints (such as in [39]), the output settles to the saturating limit of the activation function generator, $V_{TP}$ or $V_{TN}$. OTA becomes the most dominant CNN component, where for NR=1, each cell consists of 18 OTAs, and for NR=2, each cell has 50 OTAs. The inter-cell synaptic interaction is determined by programming the OTA transconductance. Edge detection operation on CNN were studied [39], and in Fig 7c, the edge detected output on a 64×64 CNN array is demonstrated. Here, the black and white pixel of the patterns correspond to the activation function saturation limits, $V_{TP}$ and $V_{TN}$. When the inputs of the cells are excited with the input pattern as shown in Fig. 7c, the output of cells evolve to the edge detected output pattern.

In Fig. 8, the efficiency of CNN is compared at varying CNN power considering TFET and FinFET based OTA. The efficiency of CNN is expressed as the Giga operations per second (GOPS) per power. With predominance of OTAs, for simplicity we ignore the power contribution from the other cell components, opamp and saturation function generator. The power of CNN is scaled by scaling down the OTA power while proportionally scaling up the cell resistance (R). This approach ensures the stability of CNN across power variation. The power-Gm traces of OTAs for TFET and FinFET [Fig. 6b] are utilized for CNN power estimation. Due to the ultra-low-power scalability of TFET OTA, the TFET based CNN is scalable to very low power. Moreover, as the efficiency of TFET OTA increases at the lower power [Fig. 6b], the efficiency of TFET based CNN is also enhanced. Similar to the characteristics in Fig. 6b, however, the TFET design is inferior to FinFET at a higher power/performance.

![Fig. 7: (a) CNN architecture with neighborhood radius = 1, (b) CNN cell schematic, and (c) edge detection operation on a 64×64 CNN.](image-url)

![Fig. 8: For edge detection operation CNN power vs. efficiency.](image-url)
V. CONCLUSIONS

Silicon TFET has several unique characteristics to it than MOSFET, which will require distinct attention from the designers. The distinct characteristics of TFET are: low on current, subthermal SS, delayed saturation, greater $g_m/Ids$, higher output resistance, and higher miller capacitance. Studying key circuit designs, SRAM and OTA, it is shown that silicon TFET, even with its low on-current, finds important applications in ultra-low power electronics. Very low off current of TFET significantly lowers the power of SRAM. Lower SS of TFET attributes to the greater $g_m/Ids$ and higher efficiency of TFET based OTA. Sub-nW power range enabled by TFET in both these designs is unattainable in MOSFET due to the higher MOSFET leakage. However, MOSFET outperforms TFET at a higher performance/power condition. Therefore, silicon TFET presents favoring opportunities for the ultra-low-power designs/applications. Utilizing the TFET based ultra-low-power OTA, TFET also opens opportunity of ultra-low power neurocomputing even for a reasonably larger array size.

Acknowledgement: This work was supported in part by National Science Foundation and Office of Naval Research Young Investigator Award

VI. REFERENCES


