Comparison of TFETs and CMOS using optimal design points for power-speed trade-offs

Juan Núñez and María J. Avedillo

Abstract— Tunnel transistors are one of the most attractive steep subthreshold slope devices currently being investigated as a means of overcoming the power density and energy inefficiency limitations of CMOS technology. In this paper, the evaluation and the comparison of the performance of distinct fan-in logic gates, using a set of widely accepted power-speed metrics, are addressed for five projected tunnel transistor (TFET) technologies and four MOSFET and FinFET transistors. The impact of logic depth, switching activity and minimum supply voltage has been also included in our analysis. Provided results suggest that benefits in terms of a certain metric, in which a higher weight is placed on power or delay, are strongly determined by the selected device. Particularly, the suitability of two of the explored TFET technologies to improve CMOS performance for different metrics is pointed out. A circuit level benchmark is evaluated to validate our analysis.

Keywords— Tunnel transistors, Steep subthreshold slope, Energy efficiency, Low supply voltage, Optimal design points.

I. INTRODUCTION

Tunnel transistors (TFETs) are currently receiving a lot of attention as potential candidates to substitute or complement CMOS devices [1]-[4]. They are one of the most attractive steep subthreshold slope devices. Steep subthreshold slope (SS) enables low voltage operation with acceptable speed leading to power and energy savings. Thus, they are being explored to overcome the power density and energy inefficiency problems exhibited by CMOS due to its 60mV/decade minimum subthreshold slope [5], [6].

Many works have addressed benchmarking of TFETs at the circuit level with different aims, including the extraction of information useful to guide device design [7], the identification of circuit design challenges or opportunities due to the distinguishing features of these transistors [8]-[10], or the comparison to CMOS to evaluate obtained gains and assessing which applications are candidate for replacing or complementing CMOS with TFETs. In particular, there are a lot of works comparing TFET transistors versus CMOS ones for logic applications [11]-[18].

The latter comparative analyses also follow very different approaches and methodologies. On one hand, some of them rely on analytical expressions using a reduced set of technological parameters such as on current, off current, input capacitance and supply voltage ($V_{DD}$). Others use simulations in order to take into account the great impact of specific features of the TFET transistors including super-linear onset, unidirectional conductance, enhanced Miller Capacitance or dominant gate to drain capacitance on their performance [19], [20].

There are also differences in terms of the circuits considered, from the typical FO4 inverter to system level benchmarks. Finally, there are many distinct criteria on which basis tunnel and CMOS technologies are evaluated. Each allows illustrating some aspects of the differences among both types of transistors. In several cases, realizations of a given circuit implemented with TFETs are compared with their CMOS counterpart at iso-performance or iso-power points. This is interesting for the practical scenery of designing circuits under operating frequency targets or under power budgets. In other works [18], [11], [21], the minimum energy point is used as a figure of merit to summarize the energy advantages. That is, energy optimized designs are compared. However, energy is just one metric of the generalized family of metrics of the form $P^*D^*$ [22] which represents different trade-offs between power and delay. There is a fundamental relationship between the optimal operating points of a design and the generalized design metrics.

In this paper, a comparative analysis in terms of optimization of a set of widely accepted power-delay design metrics is carried out. Main outstanding key point in our simulation based study is the broad set of technologies which are evaluated: five projected TFET technologies and four CMOS, including MOSFETs and FinFETs, designed for both high performance and low power applications. In addition, switching activity, logic depth and minimum supply voltage are taken into account in our analysis. Using $V_{DD}$ as a design parameter, the average energy per operation ($E$), energy-delay ($EDP$) and power-energy product ($PEP$) have been obtained and their minimum values evaluated and compared. The rest of the paper is structured as follows. Section II describes the experiments we have carried out. Results obtained are shown and discussed in Section III. In Section IV, 8-bit adders are evaluated and compared. Finally some conclusions are given in Section V.

Manuscript received July 27, 2016, revised October 20, 2016, accepted November 10, 2016. This work has been funded by Ministerio de Economía y Competitividad del Gobierno de España with support from FEDER (Project TEC2013-40670-P).

Juan Núñez and María J. Avedillo are with the Instituto de Microelecánica de Sevilla, IMSE-CNM (CSIC/Universidad of Sevilla), Av. Américo Vespucio s/n, 41092, Sevilla, Spain. e-mail: (juanen/avedillo)@imse-cnmc.csic.es.

Copyright (c) 2016 IEEE. Personal use of this material is permitted.
However, permission to use this material for any other use must be obtained from the IEEE by sending a request to pubs-permissions@ieee.org.
II. EXPERIMENT DESCRIPTION

A. Transistors

Five different tunnel transistor models have been used in this work. All of them are available from the nanoHUB website [23]. Two of them have been derived by Pennsylvania State University and the other three by Notre Dame University. They are briefly described below.

TFET models from Pennsylvania State University [24]

These are look-up table based Verilog-A models for III-V interband TFETs based on calibrated Synopsys TCAD device simulations. The calibrated TCAD TFET models serve as an approximation of full-band atomistic calculation of TFET band diagram and band-to-band tunneling current to generate the DC characteristics. The gate-source and gate-drain capacitance characteristics obtained from the TCAD small-signal simulation are validated with measured transient characteristics of TFETs. For p-channel transistors drive-currents identical to those of the n-channel are assumed. Its gate-capacitance characteristic is obtained from a TCAD simulation to take into account that density of states of electrons and holes can be quite different in III-V materials. Models with gate lengths of 20nm are available for both a double gate InAs Homojunction TFET (PSUHOMO) and a double gate GaSb-InAs Heterojunction TFET (PSUHETE).

TFET models from Notre Dame University [25], [26]

The current model, based on the Kane-Sze formula for tunneling, is valid in all four operating quadrants of the TFET. It uses a simple analytic model of the gate drain capacitance. The model parameters derived for different TFET structures showed good agreement with atomistic or TCAD device simulations. p-channel transistors assume identical drive-on currents and capacitances. Gate length for both transistors is 20nm. In this work, we use a model for a planar InAs double-gate TFET (NDHOMO), an AlGaSb/InAs double gate TFET (NDHETE,1) and a GaN/InN single gate TFET (NDHETE,2).

CMOS transistors

Four different CMOS transistors have been also evaluated for comparison purposes. All of them are predictive models obtained from the PTM web page [27]. The ones selected were those with channel lengths similar to the available TFETs, namely: 22nm MOSFET devices for both high performance (MOSFETHP, nominal \(V_{DD}=0.8V\)) and low power (MOSFETLP, nominal \(V_{DD}=0.95V\)) applications, and 20nm FinFET transistors for high performance (FinFETHP, nominal \(V_{DD}=0.9V\)) and for low stand by power (FinFETLP, nominal \(V_{DD}=0.9V\)).

B. Circuits and measurements

Fig. 1 shows the circuit used to evaluate and compare logic gates with different fan-in. An inverter, a two-input NAND gate (NAND2) and a three-input NAND (NAND3) gate have been evaluated (in blue). Note that gates under test have been loaded with the parallel connection of four minimum inverters and their inputs were not ideal but generated with chains of inverters.

Transistor sizing

In all benchmarking circuits transistors have been sized using minimum gate length. n-type transistors width is also the minimum allowable in each case (one finger for the FinFETs). MOSFET p-type transistors have been widened (to twice the minimum value) to compensate for mobility differences. Minimum p-type TFET transistors have been used since the models already assumed identical drive-on currents. Applying typical scaling rule, n-type transistors have been doubled (multiplied by three) in width in the NAND2 (NAND3) gates to keep similar rise and fall characteristics in all technologies except for FinFETs circuits for which a single finger has been used for all gates.

Measurements

The benchmarking circuits have been characterized by simulation at different supply voltages in order to take into account the effect of distinctive characteristic of these transistors that impact performance, as mentioned above. Specifically, \(V_{DD}\) has been varied from 0.05V to 1V (with voltage step equal to 0.05V) for CMOS transistors and from 0.05V to 0.7V for TFET devices. For each circuit and technology, minimum allowable \(V_{DD}\) has been determined as the minimum supply voltage at which correct functionality is observed with maximum logic swing degradation of 10%. Worst case high-to-low and low-to-high propagation delays have been measured (at \(V_{DD}/2\)) and the average of these delays, \(\Delta_{FOM}(V_{DD})\), has been used to calculate \(f_{MAX}(LD,V_{DD})\), the maximum achievable frequency at a given \(V_{DD}\), \(f_{MAX}(LD,V_{DD})=1/(LD \cdot \Delta_{FOM}(V_{DD}))\).

Evaluation and comparison in terms of optimization of a set of widely accepted power-delay design metrics, which are members of the generalized family of metrics of the form \(P^mD^n\) [22], have been carried out. Thus, different weights are assigned to power and delay depending on which specification is the most relevant concern.

The average energy per operation (E) has been calculated as \(E=E(V_{DD},f_{MAX})/f_{MAX}\), which corresponds to \(m=1\) and \(n=1\) in the generalized form of the power-delay metrics. \(P(V_{DD})\) is the average power consumption for certain values of switching activity (\(a\)) and \(f_{MAX}\) (and, thus, of LD and \(V_{DD}\)).

We have also evaluated the energy-delay product (EDP) and the power-energy product (PEP) in order to estimate trade-offs between power and speed performances.
have been calculated as follows: \( EDP = P(\alpha, f_{\text{MAX}})/f_{\text{MAX}}^2 (m=1, n=2) \) and \( PEP = P(\alpha, f_{\text{MAX}})^2/f_{\text{MAX}} (m=2, n=1) \).

III. EVALUATION OF POWER-DELAY METRICS

In this Section, the performances in terms of energy, \( EDP \) and \( PEP \) of the circuits shown in Fig. 1 are evaluated and compared. For each figure of merit (described in subsections III.A-III.C), results are firstly discussed in detail for the \( \text{FO}_4 \) inverter and subsequently provided for NAND2 and NAND3 gates.

A. Energy

Average energy per operation \( (E) \) of the \( \text{FO}_4 \) inverter versus \( V_{\text{DD}} \) curves are shown in Fig. 2 for \( LD=50 \) and \( \alpha=0.1 \). The minimum energy for each technological node does not necessarily correspond to the lowest \( V_{\text{DD}} \). Critical \( V_{\text{DD}} \) value for minimum energy \( (V_{\text{DD,opt}}) \) is much lower in TFET than in MOSFET/FinFET in agreement with previous works [18], [11], [21], because the \textit{on}/\textit{off} current ratio in tunnel technologies is larger (steeper \( SS \)) for low values of \( V_{\text{DD}} \). In this context, it should be noted that \( \text{NDHETE}_1 \) and \( \text{NDHETE}_2 \) curves have no elbows. It can be also observed that TFET inverters exhibit smaller values of minimum energy than CMOS ones.

The impact of \( LD \) and \( \alpha \) on the energy performance has been illustrated in Fig. 3. In Fig. 3a, \( LD \) has been reduced to 25, without significantly modifying the energy curves but for slight differences in \( V_{\text{DD,opt}} \). On the other side, a downward shift of the energy is observed when the switching activity factor is decreased by 10, as shown in Fig. 3b for \( \alpha=0.01 \). Note that in Fig. 3a \( LD \) was only divided by 2 (doubling frequency for a given \( V_{\text{DD}} \)), which explains the more remarkable impact of \( \alpha \) variation with respect to the experiment in which \( LD \) is varied.

To complete the analysis, minimum average energy per operation has been evaluated for six \( LD \) and \( \alpha \) combinations and each technology. Results are reported in Table I. Minimum energies among all the technologies are observed in \( \text{NDHETE}_1 \) inverters (marked in red in the Table) for \( \alpha=0.01 \) and \( \alpha=0.1 \). However they are obtained for low frequencies: \( 1.09\text{MHz} \) \( (LD=50) \) and \( 2.18\text{MHz} \) \( (LD=25) \) and at the minimum \( V_{\text{DD}} \). For \( \alpha=0.5 \), \( \text{PSUHETE} \) is the most efficient design. The minimum energy is achieved at \( 14.07\text{MHz} \) \( (LD=50) \) and \( 28.14\text{MHz} \) \( (LD=25) \) with \( V_{\text{DD}} = 0.05\text{V} \).

Fig. 4 depicts minimum energy versus frequency of the previous experiments for all CMOS technologies and \( \text{NDHETE}_1 \) and \( \text{PSUHETE} \) devices. For each technology, we have highlighted the area described by the six solutions. As expected, minimum energy at low frequencies is obtained for \( \text{NDHETE}_1 \). On the other hand, \( \text{PSUHETE} \) would be the best option to achieve high-frequency operation with the lowest energies. Note that MOSFETHP node exhibits the largest energy values for a wide frequency range and, thus, being not competitive regarding tunnel technologies. Technologies that have not been depicted in Fig. 4 (\( \text{NDHOMO} \), \( \text{NDHETE}_2 \) and \( \text{PSUHOMO} \)) cover the medium energy/frequency range.
<table>
<thead>
<tr>
<th>Technology</th>
<th>(LD,α)</th>
<th>(25,0.01)</th>
<th>(25,0.1)</th>
<th>(25,0.5)</th>
<th>(50,0.01)</th>
<th>(50,0.1)</th>
<th>(50,0.5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MOSFET_{LP}</td>
<td>0.0008</td>
<td>0.3737</td>
<td>2.4606</td>
<td>8.8567</td>
<td>0.3980</td>
<td>3.0362</td>
<td>10.3631</td>
</tr>
<tr>
<td>MOSFET_{HP}</td>
<td>0.0002</td>
<td>0.4055</td>
<td>1.8506</td>
<td>2.4745</td>
<td>0.5561</td>
<td>2.2638</td>
<td>6.803</td>
</tr>
<tr>
<td>FinFET_{LP}</td>
<td>0.0005</td>
<td>0.2993</td>
<td>1.6818</td>
<td>4.5602</td>
<td>0.3590</td>
<td>2.0569</td>
<td>6.3455</td>
</tr>
<tr>
<td>FinFET_{HP}</td>
<td>0.0003</td>
<td>0.6313</td>
<td>2.1621</td>
<td>4.5993</td>
<td>0.9795</td>
<td>2.9028</td>
<td>7.7446</td>
</tr>
<tr>
<td>ND_{HOMO}</td>
<td>0.0002</td>
<td>0.0994</td>
<td>0.3317</td>
<td>0.5844</td>
<td>0.1246</td>
<td>0.45</td>
<td>1.171</td>
</tr>
<tr>
<td>ND_{HETE,1}</td>
<td>0.0003</td>
<td>0.0003</td>
<td>0.0023</td>
<td>0.0011</td>
<td>0.0008</td>
<td>0.0048</td>
<td>0.0224</td>
</tr>
<tr>
<td>ND_{HETE,2}</td>
<td>0.0008</td>
<td>0.0002</td>
<td>0.0022</td>
<td>0.0110</td>
<td>0.0004</td>
<td>0.0044</td>
<td>0.0213</td>
</tr>
<tr>
<td>PSU_{HOMO}</td>
<td>0.0003</td>
<td>0.0972</td>
<td>0.1785</td>
<td>0.5397</td>
<td>0.1851</td>
<td>0.2668</td>
<td>0.6279</td>
</tr>
<tr>
<td>PSU_{HETE}</td>
<td>0.0003</td>
<td>0.0027</td>
<td>0.0928</td>
<td>0.1571</td>
<td>0.0483</td>
<td>0.1133</td>
<td>0.170</td>
</tr>
</tbody>
</table>

Key

- $E$ [aJ]
- $EDP$ [aJ·µs]
- $PEP$ [aJ²/µs]

Fig. 4 Minimum energy versus frequency regions for $LD=\{25,50\}$ and $α=\{0.01,0.1,0.5\}$
On the other hand, they are from 75 (NAND3) to 33 (NAND2), both compared again to CMOS technology, FinFET LP. As expected, ND HETE,1 is the performance than the best CMOS. Note that, in terms of values correspond to TFET technologies exhibiting better have been represented in logarithmic scale, so that negative α\(\leq 0.1\) ratios are between 13 and, thus, only for the NAND3 gate. This result can be explained on the basis of the sizing strategy (described in Section III.B), in which wider pull-down transistors are used in MOSFET and TFET gates when the fan-in is increased (stacked transistors) but identical in the FinFET gates. Unlike the inverter and the NAND2, in which normalized EDP results are quite similar, FinFET HP is the most competitive CMOS technology for the NAND3, which explains the apparent degradation of this figure of merit.

None of the other TFET devices in any of the simulated gates are able to improve best CMOS technology but for NDHOMO in the NAND3. Finally, it has been verified PSU HETE is the most efficient for the explored (LD, α) design space, where no significant variations on the EDP ratios are observed.

### Power-energy product (PEP)

As shown in Table I for the FO4 inverter, NDHETE,1 is the best technology in terms of minimum PEP (highlighted in blue) for all (LD, α) pairs. In fact, minimum energy designs are also optimum in terms of PEP for α\(\leq\)0.1. Normalized PEP ratios are depicted in Fig. 5c for the three gates with LD\(\geq\)25 and α\(\geq\)0.1. Again NDHETE,1 is the most efficient technology (compared to FinFET LP) for the inverter (88 ratio), NAND2 (129) and NAND3 (8040). Again, benefits due to the stack factor compared to FinFET LP are observed, as described above for energy results.

The impact of LD and α variations is very similar to that exhibited for energy results. That is, PEP ratio of NDHETE,1 (compared to FinFET LP) increases (decreases) when LD (α) does.

### Effect of the variation of VDD,MIN

The impact of VDD,MIN on these figures of merits has been analyzed for NDHETE,1 (in Fig. 6a) and PSU HETE (Fig. 6b) technologies. Similar behavior are observed for both devices: energy savings regarding MOSFET/FinFET are reduced for larger values of VDD,MIN because, as shown in Fig. 3a, advantages of TFET technologies for low VDD are exploited to a lesser extent. A similar trend is obtained for PEP when VDD,MIN is increased since, for those technologies, also power advantages due to their operation at ultra-low power are weakened. However EDP does not exhibit significant variations because, for both TFET technologies, VDD for minimum EDP is 0.2V and, thus, only for the VDD,MIN =0.25V scenario the performance is slightly reduced.
that, since random inputs are applied, large values of
values.
NAND2 and NAND3) and with different switching activity
the outputs are now a combination of different gates (inverter,
those performed at gate level. Logic paths from the inputs to
these experiments exhibit significant differences regarding
device.
Results have been normalized with respect to the best CMOS
comparisons carried out at gate level in the previous sections.
Again, all TFET technologies exhibit better energy
performance than the CMOS device, being NDHETE,1 and PSUHETE the most efficient technologies. Note that the
significant differences between NDHETE,1 and PSUHETE in Fig.
5a are not observed now. This is due to the distinct switching
activity in both experiments. Increasing $\alpha$ has a larger impact
in the minimum energy value for NDHETE,1 than for PSUHETE.
This can be observed in Fig. 4 from the shapes of the drawn
regions for both technologies. In fact, comparing points 3 (the
largest $\alpha$ value represented) for both technologies, it is clear
the similarity of the minimum energy values. In terms of $EDP$,
PSUHETE exhibit the best performance, keeping the relative
differences with respect to the other TFET devices. Finally,
for $PEP$, NDHETE,1 is still the best TFET device, whereas
similar trends for the other tunnel devices are observed.

V. CONCLUSIONS
In this paper, we evaluate a set of metrics which provides a
measure of the performance of distinct circuits implemented
using CMOS and TFET technologies. Thus, limitations of a
simple estimation of TFET energy savings, achieved by
reducing $V_{DD}$ with respect to CMOS technologies, can be overcome. Those metrics include power and delay (frequency)
since they are the two most important design specifications.
Moreover, we have discussed the impact of switching activity,
logic depth or minimum supply voltage on the performance
and advantages of TFETs devices.
As in conventional CMOS technologies, a single TFET device
is not competitive to optimize a design to achieve a minimum
value for any metric. Among the five explored TFET technologies in this work, two of them have been identified as candidates to improve CMOS performance. Specifically, $ND_{HETE}$ transistors exhibit significant advantages for both energy and $PEP$ with respect to CMOS (even regarding LP devices) since power is weighted equal or higher than frequency. On the other hand, the PSU $HETE$ has been shown to be much advantageous in terms of $EDP$, for which frequency is the primary concern. It is also competitive in terms of energy with respect to $ND_{HETE}$ for high switching activity applications.

In the circuit level example, only PSU $HETE$ exhibits advantages in the three analyzed figures of merit compared to CMOS technologies.

REFERENCES


