Impact of the RT-Level Architecture on the Power Performance of Tunnel Transistor Circuits

María J. Avedillo and Juan Núñez
Instituto de Microelectrónica de Sevilla, IMSE-CNMTCSIC/Universidad de Sevilla)

Abstract— Tunnel field-effect transistors (TFETs) 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, we analyze the relationship between devices and RT-Level architecture choices. We claim that architectural issues should be considered when evaluating this type of transistors because of the differences in delay versus supply voltage behavior exhibited by TFET logic gates with respect to CMOS gates. More specifically, the potential of pipelining and parallelism, both of which rely on lowering supply voltage, as power reduction techniques is evaluated and compared for CMOS and TFET technologies. The results obtained show significantly larger savings in power and energy per clock cycle for the TFET designs than for their CMOS counterparts, especially at low voltages. Pipelining and parallelism make it possibly to fully exploit the distinguishing characteristics of TFETs, and their relevance as competitive TFET circuit design solutions should be explored in greater depth.

Keywords— Tunnel transistors, Steep subthreshold slope, Low power, Concurrency, Pipelining, Parallelism.

1. INTRODUCTION

Intensive research is currently being conducted into devices with steeper subthreshold slopes (SS) below the physical limit of 60mV/dec of CMOS technologies. A smaller SS makes it possible to lower threshold voltage while keeping leakage current under control, facilitating low voltage operation with acceptable speed and thus overcoming the power density problems and energy inefficiency of scaled CMOS.

Tunnel transistors (TFETs) are one of the most attractive steep subthreshold slope devices [1]-[5]. Experimental TFETs with SS under 60mV/dec have already been obtained in different material systems, including Silicon TFETs and III-V TFETs. Band-to-band tunnel field-effect transistors based on two-dimensional transition metal dichalcogenide semiconductors are being explored as a potential means of improving their on currents ($I_{ON}$) [6].

Many benchmarking efforts have been made to evaluate gains over CMOS and, thereby, identify those
devices which are the most promising candidates for replacing or complementing CMOS under different metrics or in different application areas. Several works have shown power benefits for iso-performance or higher performance at iso-power up to moderate operating frequencies [7]-[12]. This is due to the fact that current TFETs do not reach the on current values obtained by CMOS transistors at their nominal supply voltages. Nevertheless, opportunities to enhance higher performance domains also exist in applications with stringent power budgets [13].

More recently, different studies have addressed analyses in which architecture choices at different levels of abstraction are taken into account. That is, they focus on the relationship between devices and architectures, showing that TFETs are more attractive than CMOS for specific design techniques or computing paradigms. This suggests that the conclusions drawn from comparisons between TFETs and CMOS could be different if these aspects are taken into account and that it may be possible to extend the application domains of these transistors in different directions. In [14]-[16], for example, processor level issues were addressed, showing speedup over CMOS due to the fact that a large number of TFET cores can function within the thermal limit and that TFETs can operate over a much wider range of microarchitecture complexity. In [17], [18], studies into lower levels of abstraction addressed issues related to circuit logic depth selection. Our work analyzes the relationship between devices and register transfer (RT) level architecture options. The conclusions drawn will be useful in the design of application specific circuits.

In this paper, we explore and compare the impact of pipelining and parallelism - two well-known register transfer level optimization techniques for increasing operating frequency and throughput - in TFET and CMOS technologies. The two techniques can also be applied to reduce supply voltage while maintaining speed, leading to power reductions or, in general, different power-speed trade-offs. However, their impact on power may vary due to the differences in the $I_{on}$ versus $V_{dd}$ behavior of the two types of transistors. Therefore, a justifiable need to explore these architecture options when comparing the two technologies.

The rest of the paper is structured as follows. Section 2 compares the delay versus supply voltage of a tunnel transistor technology and a CMOS, and explains the rationale behind our work. Simulation results from benchmark circuits are described and discussed in Section 3. Finally, some conclusions are presented in Section 4.
2. IMPACT OF CONCURRENCY ON POWER

Two well-known techniques for increasing concurrency at circuit level are pipelining and parallelism. They can be used to improve frequency performance or for power optimization [19]. Both techniques are shown in Fig. 1. In Fig. 1b, pipeline registers are added in order to cut down signal propagation paths in the combinational block C shown in Fig. 1a. Thus, assuming equal delay for C1 and C2 and ideal registers, the operating frequency and the throughput (data produced per time unit) can be doubled with respect to the implementation in Fig. 1a. Equivalently, the frequency and throughput of the implementation in Fig. 1a can be maintained at lower $V_{DD}$ by that in Fig. 1b thanks to the shorter signal paths, thus producing power benefits. Different frequency-power trade-offs are possible. In Fig. 1c, a copy of the processing circuit and some extra circuitry are added in order to have two clock cycles available to propagate signals through

Fig. 1 Block diagrams illustrating: (a) original implementation, (b) pipelining and (c) parallelism.
combinational blocks. Assuming an ideal multiplexer, throughput/frequency can thus also be doubled with respect to the original implementation. Equivalently, the throughput and the frequency of the circuit in Fig. 1a can be maintained at lower $V_{DD}$. As in the pipelined implementation, different throughput-power trade-offs could be possible.

In both cases, frequency and throughput can be maintained at a reduced $V_{DD}$ because timing constraints are relaxed. In the ideal concurrent architectures (Fig. 1b and Fig. 1c) described, constraints are fulfilled even if logic gate delays are doubled. The amount by which supply voltage (and therefore power) can be lowered depends on the extent to which speed degrades with supply voltage reduction. That is to say, the delay versus $V_{DD}$ behavior of the logic gates determines the power savings (or the speed-power tradeoffs in general) that can be achieved by applying these techniques. Because of the reduced $SS$ of TFET devices, steep reduction of $I_{ON}$ occurs at reduced supply voltage values with respect to FinFETs. That is, TFETs’ current barely varies in a range of $V_{DD}$ values in which CMOS transistors present rapid current decrement. Since gate delay depends on $I_{ON}$, it is expected that supply voltage can be lowered by a larger amount in the TFET circuit than in the CMOS one for equal delay degradation.

To illustrate these differences in behavior regarding delay and supply voltage, we simulated fan-out 4 (FO4) inverters at electrical level in CMOS and TFET technologies respectively. The TFET was a projected 20nm GaSb-InAs Interband Heterojunction device developed at Pennsylvania State University, a model obtained from nanoHUB [20]. The model is a look-up table based Verilog-A 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. Parasitic resistances are also incorporated. This model has been widely used in previously published works on the design and evaluation of TFET based circuits. The CMOS device was a 22nm predictive FinFET transistor, suitable for HP applications, obtained from the PTM web page [21]. Nominal supply voltage for this transistor is $V_{DD}=0.9V$. For each supply voltage value ($V_{DD}$), we measured FO4 delay ($D$) and evaluated how much supply voltage could be reduced before the gate delay more than doubled its value ($2D$). This reduced $V_{DD}$ value, as measured in the experiment, is denoted $V_{DD,RED}$.

Fig. 2a depicts $V_{DD,RED}$ versus $V_{DD}$ for both technologies. Note that $V_{DD}$ values up to 0.5V are shown for TFET since these transistors are suitable for low voltage operation. It can clearly be observed that smaller $V_{DD,RED}$ values were obtained for the TFET transistor, as we anticipated in our previous analysis. This translates into a greater potential for power savings associated with the use of optimization techniques which relax timing constraints, like pipelining or parallelism, as described below. As explained, frequency
and throughput in ideal implementations of pipelining (Fig. 1b) and parallelism (Fig. 1c) are identical to those in the implementation shown in Fig 1a when logic gate delays are doubled. The average dynamic power dissipation of a logic gate can be expressed as:

\[ P = C \alpha f V_{DD}^2 \]  

(1)

where \( C \) is the total capacitive load, \( \alpha \) is the activity factor, \( f \) is the operating frequency and \( V_{DD} \) the supply voltage.

Fig. 2 Impact of concurrency on \( V_{DD} \) and PR for FinFET and TFET technologies. (a) \( V_{DD} \) reduction, (b) \( PR \) vs. \( V_{DD} \) and (c) \( V_{DD} \) vs. \( FO4 \) delay.
In an ideal implementation of the pipelined circuit (ideal pipeline registers), exactly the same gates as those in the original circuit are operated at the same frequency. The power ratio ($PR$) between the ideal pipelined design and the original circuit can be approximated as:

$$PR \simeq (V_{DD,RED})^2/(V_{DD})^2$$  \hspace{1cm} (2)

Note that (2) is independent of $C$, $f$, and $\alpha$, so the selected transistor sizing criteria and interconnection capacitances for the benchmark circuits do not impact this approximation of $PR$. (2) is also valid for the ideal parallel circuit, where each gate is now replicated (two copies of the original circuit are required) but switching activity is halved.

Thus, power savings are simply $100 \cdot (1 - PR) \cdot (V_{DD,RED})^2/(V_{DD})^2$, shown in Fig. 2b, can be used to estimate of the power savings achieved by the ideal implementation of the circuit with the constraints relaxed with respect to the original circuit when dynamic power dominates.

In both technologies it can be observed that the estimated power ratio increases when supply voltage is reduced. That is, the relative power savings achieved through pipelining or parallelism decrease with a reduction in $V_{DD}$. For a given $V_{DD}$, the power ratio achieved by TFET is smaller than that obtained by CMOS. TFET therefore offers larger power savings (as a percentage). For a low supply voltage (under 0.5V), CMOS gives power savings of between 40% (at $V_{DD} = 0.4V$) and 15% (at $V_{DD} = 0.2V$), while the TFET savings are between 80% at $V_{DD} = 0.5V$ and 45% for $V_{DD} = 0.2V$. These results show that the impact of pipelining and parallelism as power optimization techniques at low voltages is more significant in TFET than in CMOS.

The previous analysis does not consider speed. However, it is interesting to evaluate the same power impact taking this factor into account. To complete our study, Fig. 2c shows supply voltage versus FO4 delay. Of the four curves depicted, CMOS-ORIG and TFET-ORIG correspond to the $V_{DD}$ required for the CMOS and the TFET FO4 inverters, respectively, to produce a given delay. CMOS-RED (TFET-RED) correspond to the $V_{DD}$ required for the CMOS (TFET) FO4 inverter to double a given delay, i.e., the $V_{DD,RED}$ for a given FO4 delay. FO4 delays of up to 20ps are shown. As expected, for a given speed, the $V_{DD}$ for CMOS-RED (TFET-RED) is smaller than that of CMOS-ORIG (TFET-ORIG). It can also be observed that the TFET inverter (TFET-ORIG curve) is not able to operate as fast as the CMOS inverter, as is also well known. Furthermore, as found in many previous works, for equal delay the $V_{DD}$ for TFET-ORIG (TFET-RED) is smaller than the $V_{DD}$ for CMOS-ORIG (CMOS-RED), thus producing power advantages for iso-performance. To achieve a delay of 10ps, for instance, $V_{DD}$ is 0.7V for CMOS (the CMOS-ORIG curve) and 0.42V for TFET (the TFET-ORIG curve).
It is even more interesting to analyze the results in Fig. 2c from the perspective of applying concurrency techniques. In this scenario, the ORIG versions of the curves correspond to the $V_{DD}$ required to operate the designs at a given frequency. The RED versions of the curves correspond to the $V_{DD}$ required by the ideal pipeline or parallel implementation operating at the same frequency. That is, minimum supply voltages needed to operate at a given speed, with (RED versions) and without (ORIG versions) applying concurrency can be compared.

Here, a number of important points should be mentioned. Firstly, note that for $FO4>10.5$ps the supply voltage of CMOS-RED is larger than that required by TFET-ORIG. Secondly, the smallest delays (highest frequency) obtained by CMOS-ORIG could be achieved in TFET through the application of ideal concurrency (TFET-RED) at supply voltages of around 0.25V, much smaller than in CMOS (0.9V or 0.47V, when concurrency is applied). This suggests that the application domain in which TFET can offer power advantages could be extended to larger frequencies. In other words, supply voltage reduction is significant enough to anticipate advantages even taking into account possible differences in interconnection capacitances.

Finally, it is also interesting to consider applications with stringent power budgets requiring low operating voltages. Assume a maximum allowed supply voltage of 0.5V. Note here that the TFET speed could be further increased. TFET-RED, which achieves the smallest delay with 0.26V, can be operated at a larger supply voltage to increase speed. In contrast, CMOS-RED, which achieves the smallest delay with 0.47V, leaves little room for improvement.

These results support our claim that introducing concurrency could be more efficient for optimizing power in TFET technologies than in CMOS, especially at low voltages. However, this model for estimating power advantages is very simple. It takes into account only dynamic power, and not the impact of supply voltage into transistor capacitances. Neither does this simple experiment consider the power and delay overheads associated with the extra circuitry that needs to be added to support pipeline or parallel operations. Moreover, only the simplest logic gate, the inverter, is evaluated, whereas the behavior of more complex gates is not considered, and other sources of power consumption, like glitches, are ignored. To overcome these limitations, electrical simulations were carried out at circuit level. These are described in the next section.
3. CIRCUIT LEVEL EVALUATION

A. Applying parallelism

Two versions of an 8-bit adder were evaluated and compared. They were implemented using the architecture shown in Fig. 1a (ORIG), in which the combinational block (C) corresponds to a single Ripple Carry Adder (RCA). The parallel circuit (PARA), based on that shown in Fig. 1c, had two adders, I/O registers with enable signals (EN) and multiplexers to combine the results produced by both processing channels. An RCA was built by interconnecting full-adders (FA). Each FA comprised four inverters and ten NAND gates (two and three inputs). Registers were implemented with typical 8-gate D flip-flops. This circuit is representative of a generic logic circuit built from logic gates.

For each supply voltage value ($V_{DD,ORIG}$), we evaluated the maximum operating frequency of the ORIG design ($f_{max,ORIG}$) and the minimum $V_{DD}$ value required by the PARA design to be operated at that frequency ($V_{DD,PARA}$). This was done through electrical simulations. Average power for each design was measured by applying long random input combinations at $f_{max,ORIG}$. $V_{DD,ORIG}$ was used for the ORIG design and $V_{DD,PARA}$ for the PARA design. In each case, complete circuits were electrically simulated. This experiment was carried out for both CMOS and TFET technologies. $V_{DD}$ values up to 0.9V (0.5V) were simulated for the CMOS (TFET) designs. $V_{DD}$ steps of 0.1V were used in the experiment. The insignificant variation of power ratio with frequency (suggested by expression (2)) was validated in the frequency range of interest.

Fig. 3 shows the results obtained. Both $V_{DD,PARA}$ and the measured power ratio ($PR=\text{Power[PARA]}/\text{Power[ORIG]}$) are depicted. It can be observed that the $V_{DD,PARA}$ values obtained concur with the results of the experiment carried out with the FO4 inverter (Fig. 2a). For TFETs, at $V_{DD,ORIG}$ 0.5V (0.2V), $V_{DD,PARA}$ was 0.22V (0.15V) for both the inverter and the adder. In CMOS, the reduced $V_{DD}$ values obtained for the adder were slightly higher than those obtained for the inverter.

![Fig. 3 PR(%) and $V_{DD,PARA}$ vs. $V_{DD,ORIG}$, including the PR(%) vs. $V_{DD,ORIG}$ estimation from expression (2).](image-url)
Power ratio values of between 25% and 60% were obtained for TFET (slightly higher than those estimated for the inverter). This small increment was partially due to the multiplexer we added to support the parallelism. Note that this extra circuitry was not taken into account in the preliminary inverter experiment. For CMOS, larger differences with respect to the inverter experiment were observed. Note that, unlike the inverter, there were no power savings ($PR>100$) under $V_{DD, ORIG}=0.4V$. For purposes of comparison, the simple estimation of $PR$ given by expression (2) using the measured $V_{DD, PARA}$ values is superposed (in black) for each technology. As expected, the measured power ratios were higher than the estimated values. In particular, note that the small differences in $V_{DD, PARA}$ do not explain the larger power ratios obtained for CMOS. Note also that the differences were greater at lower supply voltages. To interpret the results, we analyzed the simulations in depth. It can be concluded that power was wasted in glitches occurring in the parallel designs operated at reduced supply voltages (slower circuits).

The power savings achieved in the TFET technology were significantly larger than those achieved in CMOS at low voltages. As pointed out in Section II, similar relative power savings can be achieved with CMOS at

![Parallel Designs Power Ratio](image)

Fig. 4 (a) $PR(\%)$ vs. $V_{DD, ORIG}$ and $V_{DD, PIPE}$ vs. $V_{DD, ORIG}$ plots (b) $PR(\%)$ vs. $V_{DD, ORIG}$ without considering the impact of pipeline registers.
higher supply voltages. However, these parallel CMOS implementations cannot compete with TFET in terms of power consumption. To operate at the maximum frequency of CMOS-ORIG at its nominal $V_{DD}$, for example, the CMOS-PARA design requires 0.5V, while the TFET-PARA design requires only around 0.3V. Significant power advantages can therefore be expected from the TFET-PARA design. In our experiment, we found that this design consumes 10% (3%) of the power of the PARA (ORIG) circuits in CMOS. Although these results could vary, depending on the selected transistor sizing criteria or due to parasites associated with interconnections the differences obtained are large enough to confirm the power savings advantages of TFET designs.

B. Applying pipelining

Two versions of a two-level adder tree were evaluated and compared. Each adder was an RCA, implemented as described in the previous section. A two stage pipeline design with registers between the first and second level adders (PIPE) was compared with the original adder tree (ORIG), and an experiment similar to the one carried out for the parallel benchmark was conducted.

Fig. 4 shows the results obtained. Again, the reduced supply voltage $V_{DD,PIPE}$ and both the simulated and estimated $PR$ achieved by applying pipelining are shown in Fig. 4a. No power savings were produced in CMOS for $V_{DD,ORIG}$ under 0.5V ($PR > 100\%$), whereas a saving of around 30% ($PR \sim 70\%$) was achieved by the pipeline TFET at the smallest $V_{DD,ORIG}$ analyzed (0.2V). Power savings of up to 75% were obtained for TFETs ($PR \sim 25\%$ at 0.5V).

To further analyze the results, the power of both the adders and the pipeline registers was measured separately. Fig. 4b shows the power ratios produced without considering the impact of the pipeline registers. The power ratios obtained were better than those predicted by the simple estimation model (the dashed line), indicating larger power savings. This is partially attributable to the reduction in glitches in the second-level adder associated with the incorporation of the registers.

Again, significant power advantages were achieved in TFET-PARA in comparison with both CMOS designs. At the maximum frequency of CMOS-ORIG at its nominal voltage, we found that TFET-PIPE required a supply voltage of 0.3V and that its power consumption was 12% and 5% of the values obtained for the respective CMOS-ORIG and CMOS-PIPE (with $V_{DD,PIPE}=0.55V$) designs.

Power savings obtained by both concurrent designs (PIPE and PARA) with respect to the ORIG one translate into energy savings. Note that comparison between concurrent and original designs is carried out at the same operating frequency. So, energy ratios and power ratios are identical. In the same way, power comparisons between TFET designs and FinFET ones, at the end of each sub-section, can be as well extended to energy.
4. CONCLUSIONS
The experiments carried out confirm that, as expected from the delay versus $V_{dd}$ behavior of TFET logic gates, the reductions in supply voltages achieved by applying pipelining or parallelism while maintaining speed are larger in TFET circuits than in CMOS circuits. It has been shown that applying pipelining or parallelism as a means of relaxing timing constraints on signal propagation paths is a more efficient power and energy reduction technique for tunnel transistor circuits than for CMOS. These results suggest that such architectural issues should be considered when evaluating this type of transistors. That is to say, the benchmarking of TFETs versus CMOS should not be limited to comparing logic gates or identical circuit structures, since the impact of RT-level optimization techniques can vary greatly between the two technologies. From a complementary point of view, it is preferable to use RT/logic architecture when designing with TFETs, in order to fully exploit their specific features, including their $I_{ON}$ advantages for low supply voltages. Techniques which make it possible to reduce supply voltage while maintaining operating frequency and throughput should therefore be considered as a means of designing more competitive TFET logic circuits.
As in previous studies at higher abstraction levels, our results also demonstrated the potential of TFETs to achieve power and energy advantages even at frequencies above those of CMOS designs operated at nominal supply voltages in severely power-limited applications.

ACKNOWLEDGMENTS
This work was funded by Ministerio de Economía y Competitividad del Gobierno de España with support from FEDER (Project TEC2013-40670-P).
Maria J. Acedillo and Juan Nuñez are with the Instituto de Microelectrónica de Sevilla, IMSE-CNMM (CSIC/Universidad de Sevilla). Av. Américo Vespucio s/n, 41092, Sevilla, Spain. e-mail: {avedillo/jnunez}@imse-cnm.csic.es.

REFERENCES