An Energy-Efficient Heterogeneous CMP based on Hybrid TFET-CMOS Cores

***Abstract*—The steep sub-threshold characteristics of inter-band tunneling FETs (TFETs) make an attractive choice for low voltage operations. In this work, we propose a hybrid TFET-CMOS chip multiprocessor (CMP) that uses CMOS cores for higher voltages and TFETs for lower voltages by exploiting differences in application characteristics. Building from the device characterization to design and simulation of TFET based circuits, our work culminates with a workload evaluation of various single/multi-threaded applications. Our evaluation shows the promise of a new dimension to heterogeneous CMPs to achieve significant energy efficiencies (upto 50% energy benefit and 25% ED benefit with single-threaded applications, and 55% ED benefit with multi-threaded applications).**

I. Introduction

Power consumption is a critical constraint hampering progress towards more sophisticated and powerful processors. A key challenge to reducing power consumption has been in reducing the supply voltage due to concerns of either reducing performance (due to reduced drive currents) or increasing leakage (when reducing threshold voltage simultaneously). The sub-threshold slope of the transistor is a key factor in influencing the leakage power consumption. With a steep sub-threshold device it is possible to obtain high drive currents (I_{ON}) at lower voltages without increasing the off state current (I_{OFF}). In this work, we propose the use of Inter-band Tunneling Field Effect Transistors (TFETs) [1] that exhibit sub-threshold slopes steeper than the theoretical limit of 60 mV/Dec. Consequently, TFETs can provide higher performance than CMOS based designs at lower voltages. However, at higher voltages the I_{ON} of MOSFETs are much larger than can be accomplished by the tunneling mechanism employed in existing TFET devices. This trade-off enables architectural innovations through use of heterogeneous systems that employ both TFET and CMOS based circuit elements.

Heterogeneous chip-multiprocessors that incorporate cores with different frequencies, micro-architectural resources, instruction-set architectures [2] are already emerging. In all these works, the energy-performance optimizations are performed by appropriately mapping the application to a preferred core. In this work, we add a new technology dimensionality to this heterogeneity by using a mix of TFET and CMOS based cores. The feasibility of TFET cores is analyzed by showing design and circuit simulations of logic and memory components that utilize TFET based device structure characterizations.

Dynamic voltage and frequency scaling (DVFS) is widely used to reduce power consumption. Our heterogeneous architecture enables to extend the range of operating voltages possible by supporting TFET cores that are efficient at low voltages and CMOS cores that are efficient at high voltages. For an application that is constrained by factors such as I/O or memory latencies, low voltage operations is possible, sacrificing little performance. In such cases a TFET core may be preferable. However, for compute intensive performance critical applications, MOSFETs operating at higher voltages are necessary. Our study using two DVFS schemes show that the choice of TFET or CMOS for executing an application varies based on the intrinsic characteristics of the applications. In a multi-programmed environment which is common on platforms ranging from cell-phones to high-performance processors, our heterogeneous architectures can improve energy efficiencies by matching the varied characteristics of different applications.

The emerging multi-threaded workloads provide an additional dimension to this TFET-CMOS choice. Multi-threaded applications with good performance scalability can achieve much better energy efficiencies utilizing multiple cores operating at lower voltages. While energy efficiencies through parallelism is in itself not new, our choice of TFET vs. CMOS for the application will change based on the actual voltage at which the cores operate and the degree of parallelism (number of cores). Our explorations shows TFETs based cores to become more preferred in emerging multi-threaded applications from both energy and performance perspective.

The rest of this paper is organized as follows: In section II, we introduce Tunnel FET device operation and modeling, and discuss III-V semiconductor-based TFETs. By comparing the transistor level characteristics of TFETs with state of the art MOSFETs, we identify the potential impact of III-V semiconductor-based TFETs at the architecture level. In section III, we demonstrate circuit modeling using TFETs, and compare the energy-delay performance of logic and memory elements for MOSFETs and HTFETs. In section IV we show the benefits of our heterogeneous multi-core. Finally we conclude in section V.

II. Tunnel FET Device Characteristics

A. Device Modelling of Tunnel FETs

Since compact models for the transfer characteristics of Tunnel FETs have not been fully developed, we use the device simulator TCAD sentaurus [3] in order to model the $I_D - V_G$ characteristics of TFETs. Fig 1(A) compares the experimental and simulated characteristics for a single-gate homojunction $In_{0.53}Ga_{0.47}As$ TFET from [1], and shows a good match between experimental and simulated curves. The

parameters used for simulating the single-gate homojunction TFET are from [1]. By reducing the gate oxide to Hi-K (ϵ_{ox} 21, t_{ox} 2.5nm (EOT 0.5nm), and by using a double-gated structure (T_{Body} 7nm), we obtain projected characteristics of a homojunction $In_{0.53}Ga_{0.47}As$ TFET as shown in Fig 1B.

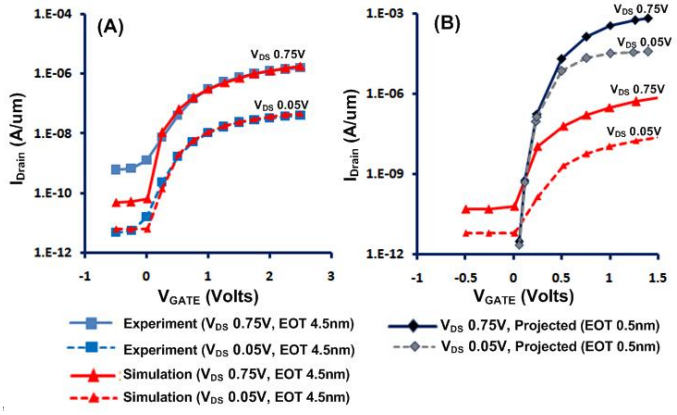

Fig. 1. (A) Comparison of experimental and simulated characteristics of single-gate $In_{0.53}Ga_{0.47}As$ homojunction TFET (EOT 4.5nm) [1] (B) Comparison of simulated characteristics of single-gate $In_{0.53}Ga_{0.47}As$ homojunction TFET (EOT 4.5nm) and projected double-gate $In_{0.53}Ga_{0.47}As$ homojunction TFET (EOT 0.5nm).

We capture the transfer characteristics of the tunnel FET obtained through device simulation across a range of voltages in a Verilog-A lookup table, in order to perform circuit simulations. The $I_{ds}(V_{gs}, V_{ds})$, $C_{gd}(V_{gs}, V_{ds})$ and the $C_{gs}(V_{gs}, V_{ds})$ characteristics are captured in two-dimensional look-up tables for modeling tunnel FETs. Fig 2(A) shows the Verilog-A small-signal model for Tunnel FETs, which uses the look-up tables for circuit simulation. Fig 2(A) and 2(B) show the Voltage Transfer Characteristics (VTC) and the transient output characteristic of a $In_{0.53}Ga_{0.47}As$ homojunction TFET inverter (V_{CC} 0.5V), which shows the validity of the Verilog-A lookup table based method.

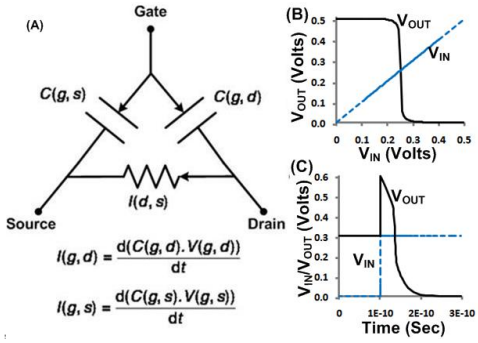

Fig. 2. Verilog-A small signal model used for Tunnel FET simulation.

B. Heterojunction Tunnel FETs

We consider a $GaAs_{0.1}Sb_{0.9}/InAs$ HTFET, and use the modeling technique described in Section II-A to obtain the transfer characteristics of the HTFET. A comparison of the $In_{0.53}Ga_{0.47}As$ homojunction TFET and the heterojunction TFET is shown in Fig 3. By using the HTFET, a higher I_{On} can be obtained because (1) $InAs$ is a smaller band-gap

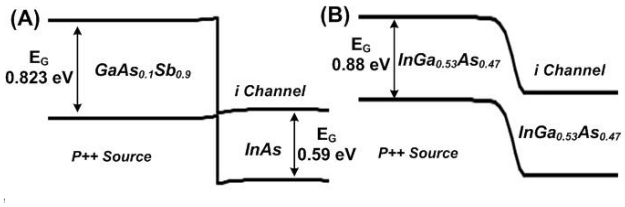

Fig. 3. Comparison of heterojunction and homojunction TFET (Band-Gap includes quantization effect due to Double-Gate structure with 7nm T_{Body})

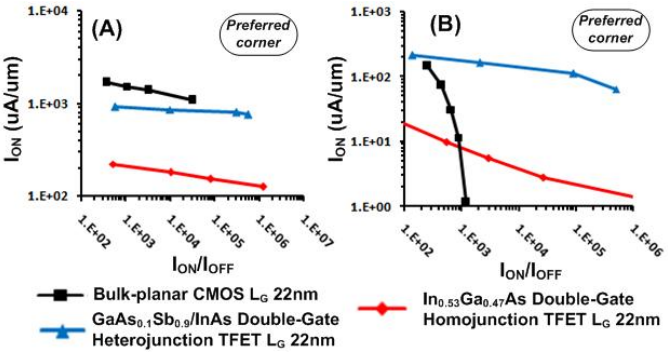

Fig. 4. Comparison of Ion versus Ion/Ioff ratio for different operating points on the $I_D - V_G$ for (A) a V_{CC} window of 0.8V and (B) a V_{CC} window of 0.3V.

material, and (2) the staggered P-N heterojunction provides a higher critical-field strength for efficient inter-band tunneling. In order to understand the circuit level implications of using HTFETs, we compare the I_{On} versus I_{On}/I_{Off} characteristics for the transistor candidates by considering different operating points along the $I_D - V_G$ curve for a given V_{CC} window, as shown in Fig 4. Fig 4A shows that at V_{CC} 0.8V, the highest I_{On} and I_{On}/I_{Off} ratio are provided by 22nm CMOS, making it the preferred device for operation at high V_{CC}. However, at V_{CC} 0.3V, the CMOS device cannot provide both a good I_{On} as well as a good I_{On}/I_{Off} ratio because of the 60 mV/Dec limit on the sub-threshold slope. The $In_{0.53}Ga_{0.47}As$ homojunction TFET can provide a good I_{On}/I_{Off} but cannot provide a high Ion since the homojunction does not allow a strong tunneling current. In contrast, the heterojunction TFET can provide a good Ion (due to the staggered P-N junction and the lower E_G material), as well as a good I_{On}/I_{Off}, due to the sub-60 mV/Dec sub-threshold slope, making it the preferred device for operation at low V_{CC}.

III. Characterization of HTFET based Logic and Memory

A. Tunnel FET Logic

In this section, we illustrate the energy-performance characteristics of logic gates constructed using CMOS transistors and HTFETs. We use a predictive BSIM model [4] for 22nm CMOS (V_T 0.2V) which provides an Ion of 1.4 uA/um and an Ion/Ioff of 3×10^3 when operating at its nominal V_{CC} of 0.8V. We also use a $GaAs_{0.1}Sb_{0.9}/InAs$ HTFET which provides an Ion of 100 uA/um and an Ion/Ioff of 2×10^5 at V_{CC} 0.3V. In order to build logic gates, a pull-up device is also

required. A PTFET can be constructed using a heterojunction with $InAs$ as the source, as shown in Fig 5(B). When a positive gate and drain voltage are applied to the H-NTFET (Fig 5(A)), electrons tunnel from the $GaAs_{0.1}Sb_{0.9}$ source into the $InAs$ channel (Fig 5(C)). In contrast, when a negative gate and drain voltage is applied to the H-PTFET (Fig 5(D)), holes tunnel from the $InAs$ source into the $GaAs_{0.1}Sb_{0.9}$ channel. By using the modelling techniques described in Section II-A, we obtain the energy-delay characteristics of HTFET logic gates. The energy-delay performance curve of a HTFET 40-stage ring-oscillator, when compared to that of a CMOS ring-oscillator in Fig 6(A), shows a cross-over in the energy-delay characteristics. The CMOS ring-oscillator has a better energy-delay compared to the HTFET ring-oscillator at $V_{CC} > 0.65V$ and the HTFET ring-oscillator has a better energy-delay trade-off at $V_{CC} < 0.55V$. Other logic gates, such as Or, Not and Xor (which are not shown here) also show a similar cross-over. This trend is consistent with the discussion in Section II-B where it has been illustrated that CMOS devices provide better operation at high V_{CC} and HTFETs provide preferred operation at low V_{CC}. Fig 6(B) shows the energy-delay performance of a 32-bit prefix-tree based Han-Carlson Adder has a similar crossover behavior for CMOS and HTFETs.

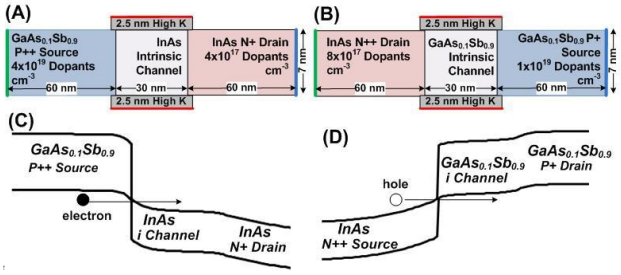

Fig. 5. (A-C) Double-Gate H-NTFET device structure and operation (D-F) Double-Gate H-PTFET device structure and operation.

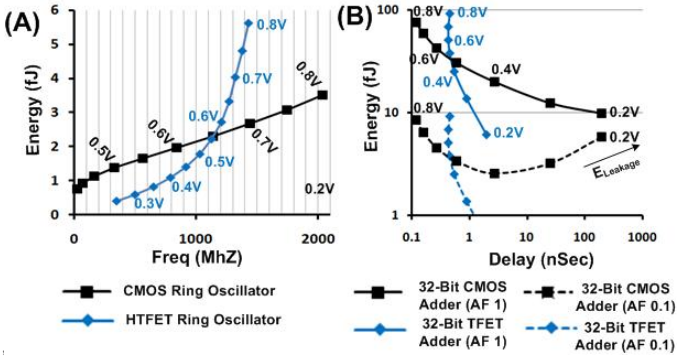

Fig. 6. Energy-Delay performance comparision for (A) a CMOS and a HTFET Ring-Oscillator and (B) a CMOS 32-bit Adder and a TFET 32-bit Adder

B. Tunnel FET Pass-Transistor Logic

As shown in Fig 7, due to their asymmetric source-drain architecture, HTFETs cannot function as bi-directional pass transistors. Though this may seem to limit the utility of TFETs in SRAM-cell design, several SRAM designs have been proposed to overcome this limitation [5], [6]. It is also important to consider a solution for logic, because of the ubiquitous usage of pass-transistors in logic design. We propose using a pass-transistor stack composed of N-HTFETs, with a P-HTFET for precharging the output. All the N-HTFET transistors in the pass-transistor stack will be oriented toward the ouput which allows them to drive the On current when the input signals are enabled. During the pre-charge phase, the P-HTFET precharges the output to V_{CC}, and during the evaluate phase, the N-HTFET stack evaluates the output based on the inputs to the pass-transistor stack.

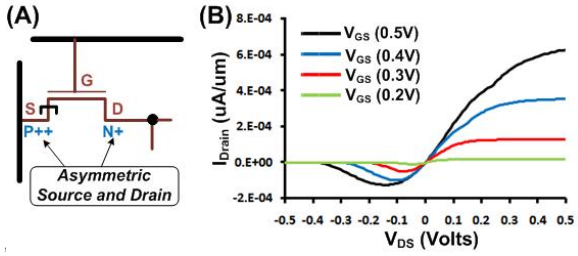

Fig. 7. (A) Asymmetric source-drain architecture for a heterojunction NTFET and (B) Asymmetric $I_D - V_D$ characteristics resulting from source-drain asymmetry

C. Tunnel FET SRAM Cache

$$\tau_{Delay} = \tau_f \sqrt{(log(V_s))^2 + 2.(\tau_{in}/\tau_f).b.(1 - V_s)} \quad (1)$$

$\tau_f = R_f \times (C_{Load} + C_{Eff})$ and τ_{in} is the input ramp

In order to model TFET-based processor architectures, it is important to consider the characteristics of the L1 cache, which is an integral on-chip component of a processor. We use the analytical method implemented in the cache analysis tool CACTI [7], in order to evaluate the energy-delay performance of a TFET-based cache. As discussed in Section III-B, in order to overcome the problem of asymmetric conduction in TFETs, we use the precharge-based pass-transistor mux which is implemented in CACTI, and we also assume a 6-T SRAM Cell with virtual-ground from [5]. CACTI uses the Horowitz approximation [8] given by eq (1) to compute the gate delay. R_{Eff} and C_{Eff} are estimated using simulation delay values as described in [9] which takes into account the effect of enhanced Miller capacitance effect in TFET resulting from the presence of a tunnel junction between the source and the channel. In order to validate the Horowitz model for TFETs, we compare the delay from the Horowitz analytical expression with the delay estimated using the Verilog-A table-lookup model for different input ramp times (τ_{in}), and obtain a good match as shown in Fig 8.

We modified CACTI to implement the 6T TFET SRAM cell design proposed in [5] and evaluated the energy-delay performance of a 32KB L1 cache with a 32Byte block-size, associativity 2 and consisting of 4 identical sub-arrays. Fig 9 shows that a cross-over point similar to that in logic exists for Low-V_T CMOS and TFET-based SRAM L1 caches. Due to

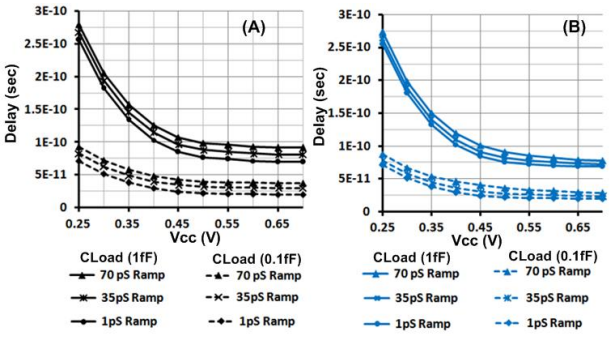

Fig. 8. Validation of Horowitz approximation for TFET gates.

the higher I_{On}/I_{Off} ratio of TFETs, the TFET L1 cache has lower leakage power than the CMOS Low-V_T L1 cache.

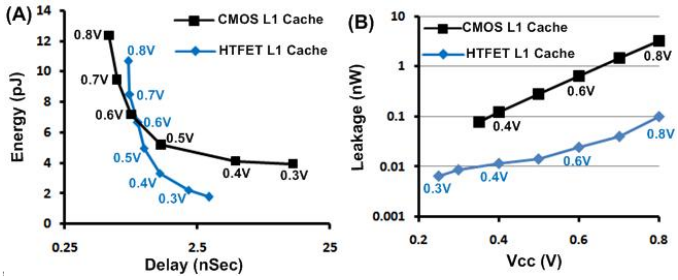

Fig. 9. (A) Energy-Delay performance comparision and (B) Leakage Power comparison for CMOS and H-TFET based L1 Cache.

IV. Architectural Analysis of CMOS and H-TFET Cores

TABLE I
Simulation Parameters.

Baseline Parameters

Parameter	Value
Processor Pipeline	Suns SPARC based core
Issue width	1
Fetch Queue	32
L1 cache	32KB, 2-way 32B block
L2 cache	2MB, 8-way 64B block
Mem. Lat / Baseline Freq.	70 cycles/ 2 GHz
Technology / Voltages	22 nm / V_{CC} = $0.7V - 0.3V$
DVFS Interval Period	200,000 Instructions

The detailed processor and cache parameters for simulating single-core processors using Simics [10] are shown in Table I. The delay and power numbers for each voltage/frequency pair obtained using circuit simulations are incorporated into our simulator. We evaluate both single-threaded (SPEC 2006) and multi-threaded (SPLASH) applications. For power analysis, we use a utilization based approach. The utilization is monitored by tracking the execution and stall cycles of the processor using Simics. For the execution cycles, the dynamic energy is modeled assuming 10% of the overall 20M gates in our core switch (typical switching activity in logic based data paths ranges from 10% - 15% [11] and the variations across instructions in commercial low power cores are minimal [12]). Leakage power is consumed during both execution and stall cycles and no power-gating is assumed. The cache power models are based on CACTI [7] that incorporates our modifications mentioned in Section III-C. For clarity, we highlight the results from 9 SPEC 2006 and 4 SPLASH benchmarks that capture the major trends observed across the suite.

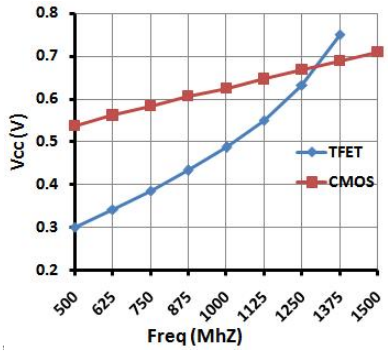

Fig. 10. Voltage Frequency Operating Points from H-TFET and CMOS processors.

Figure 10 shows the different voltage-frequency coordinates that can be achieved for a H-TFET and a CMOS based processor respectively (with a minimum frequency of 500 MHz and frequency increasing in steps of 125 MHz). It is clear from this figure that H-TFETs are the preferred device when operating below 1250 MHz. We consider a heterogeneous-technology asymmetric multi-core processor, with a TFET processor operating in the 1250-500 MHz frequency range, and a CMOS processor operating in the 0.7V to 0.5V range (frequency 1375-500 MHz). We then execute various benchmark applications (SPLASH benchmarks are executed using a single thread) using (1) an Energy-Aware DVFS policy which seeks to minimize the ED^2 [13], and (2) a purely IPC-aware DVFS algorithm [14]. The energy-aware DVFS policy monitors if the ED^2 in a DVFS interval (using the energy and delay incurred in executing 200,000 instructions) is better than the previous interval, and if so, it continues the voltage-frequency (VF) change (either continuing to increase or decrease) - otherwise, the direction of the VF change is reversed. We find that, when using the energy-aware DVFS policy on TFETs, most of the applications spend a significant amount of time (close to 60%) in 1000 MHz to 750MHz range, whereas when using CMOS most of the applications execute in 1375 MHz to 1250 MHz range (Figure 11(A)). As Figure 10 shows, the relationship between E and D^2 for TFET processors is non-linear, and the energy-aware DVFS algorithm sees a significant energy benefit when operating in these frequency ranges (1000 - 750 MHz) with TFETs. Consequently, there is a significant energy-delay benefit (average 50%) when using TFETs over the baseline CMOS based design (Figure 11(A)), but with a 40% cost in performance (Figure 11(B)).

The IPC-aware DVFS algorithm, on the other hand monitors the change in the IPC of the processor and ramps the frequency up or down by 125 MHz when it detects a 5% change in IPC. Figure 13(A) shows that the degradation in performance is less 12% than compared to baseline CMOS when using IPC-aware DVFS on TFET. Figure 13(B) shows that the energy reduction is significant when using DVFS on TFETs due to the lower energy of lower frequency modes in TFETs (Energy reduction

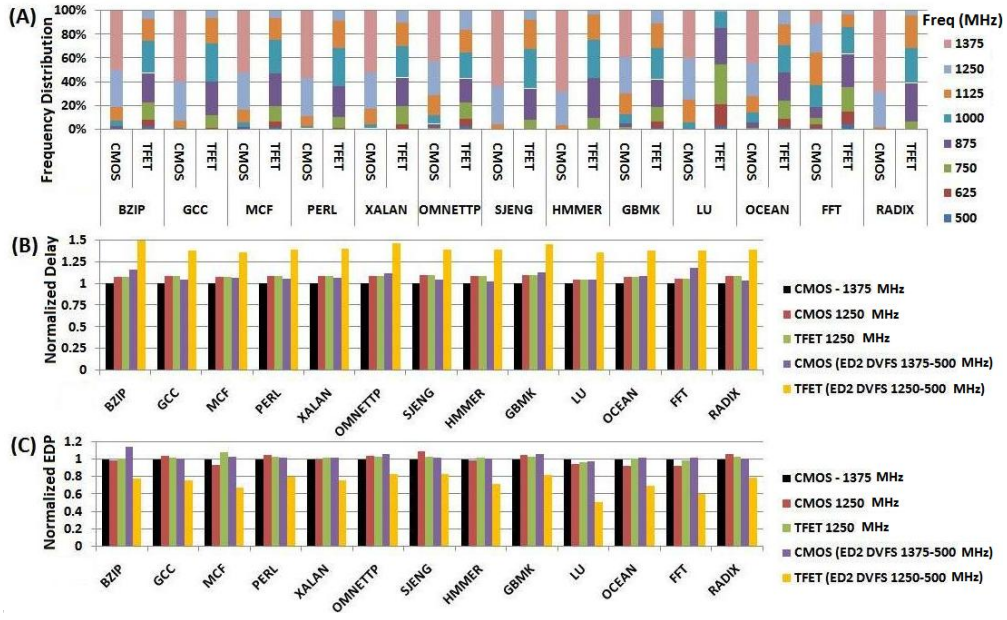

Fig. 11. (A) Frequency distribution (B) Normalized Delay and (C) Normalized EDP for Energy-Aware DVFS on benchmark applications.

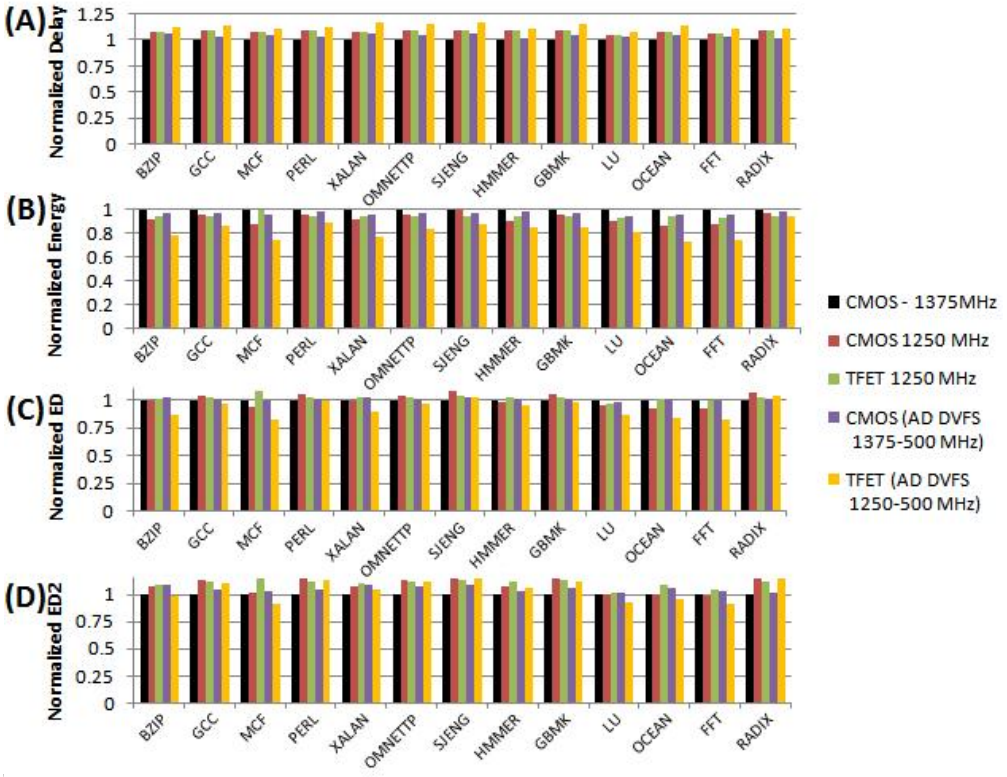

Fig. 13. (A) Normalized Delay (B) Normalized Energy (C) Normalized ED and (D) Normalized ED^2 for IPC-Aware DVFS on benchmark applications.

26% and ED reduction 18% over baseline CMOS). Further, Figure 13(D) shows that there is significant ED^2 reduction over baseline CMOS (upto 9% ED^2 benefit over baseline CMOS) for applications such as `bzip`, `mcf` and `ocean`. These applications have significant L2 miss-rates (shown in Figure 12) and consequently, the processors spend a lot of time stalling. Thus, by using energy oriented DVFS scaling on TFETs during these stall cycles gives us significant energy advantage when compared to a CMOS based design. Thus, we conclude that in heterogeneous-technology asymmetric-performance multi-core processor, single-threaded applications with higher miss-rates are more suited for execution on

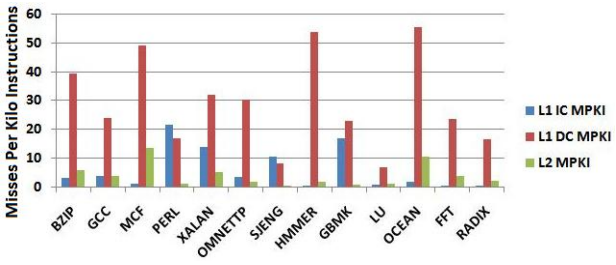

Fig. 12. Miss Rates for various benchmark applications.

TFETs with IPC-aware DVFS, since it results in significant ED^2 advantage. Pure energy conservation is best achieved by executing the applications on the TFET processor with energy-aware DVFS. Applications such as `sjeng`, `perl` and `radix` with low cache miss-rates are best executed on the CMOS processor with higher performance.

Multi-core processors can be used to minimize energy consumption by scaling down the operating frequency and increasing thread-level parallelism in order to regain *iso-performance* to baseline CMOS (1-Core @ 1375 MHz) as shown in Figure 14(A). Figure 14(B) shows the energy consumption compared to baseline CMOS for parallel program execution on 2-Core CMOS and 2-Core TFET, for *iso-performance* to baseline CMOS. When moving from 1 to 2 cores, we observe almost linear performance scaling with the number of cores, that drops the required operating frequency for *iso-performance* below the CMOS-TFET cross-over point. Due to the energy advantage of TFET processors at lower frequencies, TFET processors have a distinct energy advantage in *iso-performance* multi-core execution, giving an energy savings of 70% against single-core CMOS and energy savings of 55% against 2-core CMOS.

Our hybrid architecture provides additional energy efficiencies for multi-threaded applications by scheduling performance critical threads [15] on high performance CMOS cores and non-critical threads on energy efficient TFET cores. They can exploit imbalance across threads due to application behavior [16].

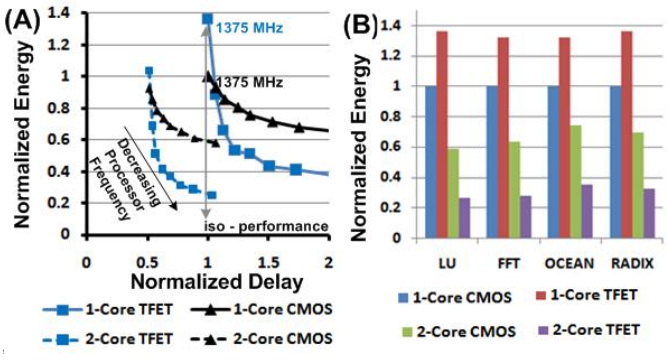

Fig. 14. (A) Illustration of normalized energy-delay for of iso-performance for `LU` Benchmark application and (B) Normalized multi-core execution energy for iso-peformance to CMOS @ 1375 MHz.

V. Conclusion

In this work we show the effectiveness of a hybrid TFET-CMOS core for exploiting inter-application characteristics in multi-programmed workloads. Our proposal can also be used to exploit intra-application characteristic. This can be done by detecting phases in applications that would benefit by being scheduled on a CMOS core and phases that would benefit by being scheduled on a TFET core (through OS support). We also show TFET cores become preferable in multi-threaded applications. Our future work will explore iso-performance scenarios to achieve the performance of multiple CMOS cores. Our initial results indicate promise in all these directions.

References

[1] S. Mookerjea, D. Mohata, R. Krishnan, J. Singh, A. Vallett, A. Ali, T. Mayer, V. Narayanan, D. Schlom, A. Liu, and S. Datta, "Experimental demonstration of 100nm channel length in0.53ga0.47as-based vertical inter-band tunnel field effect transistors (tfets) for ultra low-power logic and sram applications," in *Proc. IEEE Int. Electron Devices Meeting (IEDM)*, 2009, pp. 1–3.

[2] R. Kumar, D. M. Tullsen, N. P. Jouppi, and P. Ranganathan, "Heterogeneous chip multiprocessors," *Computer*, vol. 38, pp. 32–38, 2005.

[3] *TCAD Sentaurus Device Manual, Release: C-2009.06*, Synopsys, 2009.

[4] W. Zhao and Y. Cao, "New generation of predictive technology model for sub-45nm design exploration (http://ptm.asu.edu/)," in *Proc. 7th Int. Symp. Quality Electronic Design ISQED '06*, 2006. [Online]. Available: http://ptm.asu.edu/

[5] J. Singh, K. Ramakrishnan, S. Mookerjea, S. Datta, N. Vijaykrishnan, and D. Pradhan, "A novel si-tunnel fet based sram design for ultra low-power 0.3v vdd applications," in *Proc. 15th Asia and South Pacific Design Automation Conf. (ASP-DAC)*, 2010, pp. 181–186.

[6] D. Kim, Y. Lee, J. Cai, I. Lauer, L. Chang, S. J. Koester, D. Sylvester, and D. Blaauw, "Low power circuit design based on heterojunction tunneling transistors (hetts)," in *ISLPED '09: Proceedings of the 14th ACM/IEEE international symposium on Low power electronics and design*. New York, NY, USA: ACM, 2009, pp. 219–224.

[7] S. Thoziyoor, N. Muralimanohar, J. H. Ahn, and N. P. Jouppi, "Cacti 5.1," HP Labs, Tech. Rep., 2008.

[8] M. A. Horowitz, "Timing models for mos circuits," US Army Research Office, Tech. Rep., 1994.

[9] S. Mookerjea, R. Krishnan, S. Datta, and V. Narayanan, "Effective capacitance and drive current for tunnel fet (tfet) cv/i estimation," *IEEE Transactions on Electron Devices*, vol. 56, no. 9, pp. 2092–2098, 2009.

[10] "Simics product information (http://www.windriver.com/products/simics/)."

[11] "Xilinx power tutorials." [Online]. Available: http://www.xilinx.com/support/documentation/sw_manuals/xilinx12_2/ug733.pdf

[12] A. Sinha and A. P. Chandrakasan, "Jouletrack-a web based tool for software energy profiling," in *Proc. Design Automation Conf*, 2001, pp. 220–225.

[13] G. Magklis, P. Chaparro, J. Gonzalez, and A. Gonzalez, "Independent front-end and back-end dynamic voltage scaling for a gals microarchitecture," in *Proc. Int. Symp. ISLPED'06 Low Power Electronics and Design*, 2006, pp. 49–54.

[14] G. Semeraro, G. Magklis, R. Balasubramonian, D. H. Albonesi, S. Dwarkadas, and M. L. Scott, "Energy-efficient processor design using multiple clock domains with dynamic voltage and frequency scaling," in *Proc. Eighth Int High-Performance Computer Architecture Symp*, 2002, pp. 29–40.

[15] M. Aater Suleman, O. Mutlu, M. K. Qureshi, and Y. N. Patt, "Accelerating critical section execution with asymmetric multicore architectures," *IEEE Micro*, vol. 30, no. 1, pp. 60–70, 2010.

[16] I. Kadayif, M. Kandemir, and I. Kolcu, "Exploiting processor workload heterogeneity for reducing energy consumption in chip multiprocessors," in *Proc. Design, Automation and Test in Europe Conf. and Exhibition*, vol. 2, 2004, pp. 1158–1163.