Modeling Steep Slope Devices: From Circuits to Architectures

Karthik Swaminathan*, Moon Seok Kim*, Nandhini Chandramoorthy*,
Behnam Sedighi†, Robert Perricone†, Jack Sampson* and Vijaykrishnan Narayanan*
*Pennsylvania State University, University Park, PA 16802
†University of Notre Dame, Notre Dame, IN 46556

***Abstract*— Steep Slope devices, with Heterojunction Tunnel FETs (TFETs) in particular, have been proposed as a viable solution to overcome the subthreshold slope limitation in existing CMOS technology and achieve ultra-low voltage operation with acceptable performance. However, state-of-the-art FinFET technologies continue to demonstrate superior performance than steep slope devices in application domains demanding peak single threaded performance. In this context, we examine different computing paradigms where TFET technologies can be used, not just as a 'drop in' replacement, but as an additional parameter to augment the architectural design space. This greatly widens the scope of optimizations for performance and power. We investigate the tradeoffs between device and architectures in general purpose processors when performance, power and temperature are individually constrained. We also synthesize examples of domain-specific accelerators used in computer vision using in-house TFET standard cell libraries to demonstrate the energy benefits of designing TFET-based accelerators. We demonstrate that synthesizing these accelerators using TFETs reduces energy by over 6X in comparison to an equivalent iso-voltage CMOS-based design and by over 30% in comparison to an iso-performance CMOS design.**

I. Introduction and Motivation

As it has been for many years, CMOS remains the dominant transistor model in currently fabricated devices. However, as chip designers attempt to move tasks with large computational demands into increasingly power-sensitive and energy-sensitive domains, there is increasing interest in alternative devices more directly suited to low-power operation. This interest arises due to the relationship among supply voltages, performance, and power in low-voltage domains; For CMOS devices in a low-voltage domain, further lowering the supply voltage would significantly reduce dynamic power, but at the cost of profound reductions in performance that can lead to increases in total energy. A concise summary of the tradeoffs between power and performance at subthreshold voltages is provided by a device's subthreshold slope. So-called *steep-slope* devices, such as *Heterojunction Tunnel FETs*(HTFETs) [1, 2], offer a superior exchange rate between reductions in performance and supply voltage compared to CMOS.

So far, none of the proposed steep-slope devices are silver bullets, that is, drop-in replacements for CMOS that offer universally superior properties. While steep-slope devices may offer superior performance at sufficiently low voltages, state-of-the-art CMOS FinFET technologies continue to offer superior peak performance on highly sequential, latency-bound computations. However, we shall show that, for tasks with thread or data parallelism, with fixed, real-time performance demands, or with tight thermal constraints, adding steep-slope devices to the chip designer's tool chest extends the range of realizable designs in fundamental ways. For much of this paper, we focus on a particular class of steep-slope devices, HTFETs, show both their potential and their limitations with respect to CMOS, and how CMOS/TFET tradeoffs change in both general and specialized computing domains.

In this paper, we make the following contributions:

- **We compare Heterojunction TFETs with CMOS and beyond-CMOS devices**: We evaluate HTFETs and other emerging technologies using standard (NRI) benchmarking methodologies. These, and additional evaluations carried out on simple circuits, such as inverters and adders, show HTFETs to offer promising tradeoffs between circuit delay and computation energy.
- **We explore the relationship between device selection and optimal microarchitecture**: Due to the energy-efficiency properties of HTFETs in low-voltage operating conditions, processors built with HTFETs can more aggressively pursue wide issue designs than their CMOS counterparts before encountering thermal limitations. This makes microarchitectural techniques for scaling ILP that trade thermal density and/or area for performance attractive for HTFET based designs, even though these techniques are no longer viable for CMOS designs approaching their thermal limits.
- **We highlight the potential benefits of TFETs in general purpose processor designs**: We show that TFET-based designs can produce performance-viable processors beyond the set of thermal envelopes that are available with traditional CMOS approaches.
- **We show the promise of TFETs for domain-specific computing**: Domain-specific tasks, such as computer vision, often have two properties that make them attractive for employing steep-slope devices: abundant data parallelism and deadline-based computation. Using vision accelerators synthesized with a TFET-based standard cell library, we show two key findings. First, for an iso-voltage point, TFETs offer vastly superior performance. Second, for an iso-performance point, the TFET design consumes lower power and energy.

978-3-9815370-2-4/DATE14/©2014 EDAA

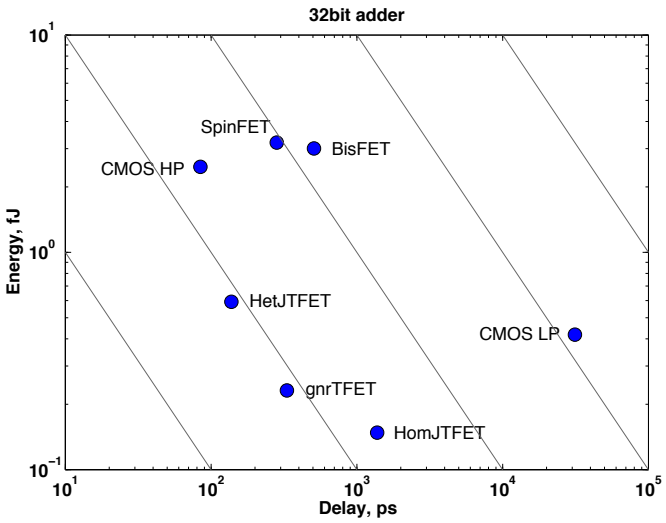

Fig. 1. Switching time and energy of CMOS and several beyond-CMOS devices for a 32-bit full adder circuit in 15 nm node. Lower left corner is more preferable (HetJ: Heterojunction, HomJ: Homojunction, gnr: Graphene Nano-Ribbon, Bis: Bilayer pseudospin).

II. Benchmarking Steep Slope Devices with other Emerging Technologies

The goal of benchmarking beyond-CMOS devices is to identify the devices that are the most promising as potential alternatives to CMOS technology. In this work, we use the NRI method [3] which aims to provide a consistent benchmarking methodology for beyond-CMOS devices. To this end, the proposed methodology is based on simplicity, uniformity, and transparency; for simplicity, analytical expressions are chosen over simulations for benchmark calculations. Uniformity is achieved by applying similar methodologies across all relevant devices. By making equations and parameters transparent, others can easily recreate or utilize the metrics for their own research.

Two key metrics used to compare and contrast devices are switching time and energy. To estimate the switching time and energy of electronic devices, the benchmarking model utilizes several analytical expressions. Equations (1) and (2) show how the intrinsic switching time (t_{int}) and energy (E_{int}) are computed for an electronic device.

$$t_{int} = C_{dev}V_{dd}/I_{dev} \tag{1}$$

$$E_{int} = C_{dev}V_{dd}^2 \tag{2}$$

C_{dev} represents the device capacitance, and I_{dev} is the on-current of the device. C_{dev} and I_{dev} vary in proportion with the transistor width. The performance of a simple circuit is then calculated (see [3] for details) . The switching time and energy are computed for a 32-bit full adder circuit as:

$$t_{add} = 32M_{tadd}(5M_{tinv}t_{int} + 5t_{ic}) \tag{3}$$

$$E_{add} = 32M_{Eadd}(10M_{Einv}E_{int} + 5E_{ic}), \tag{4}$$

where M_{tadd}, M_{tinv}, M_{Eadd}, and M_{Einv} are circuit performance parameters chosen to make the results agree with PETE simulation [4]. The time and energy of an interconnect (t_{ic} and E_{ic}, respectively) are estimated with the interconnect

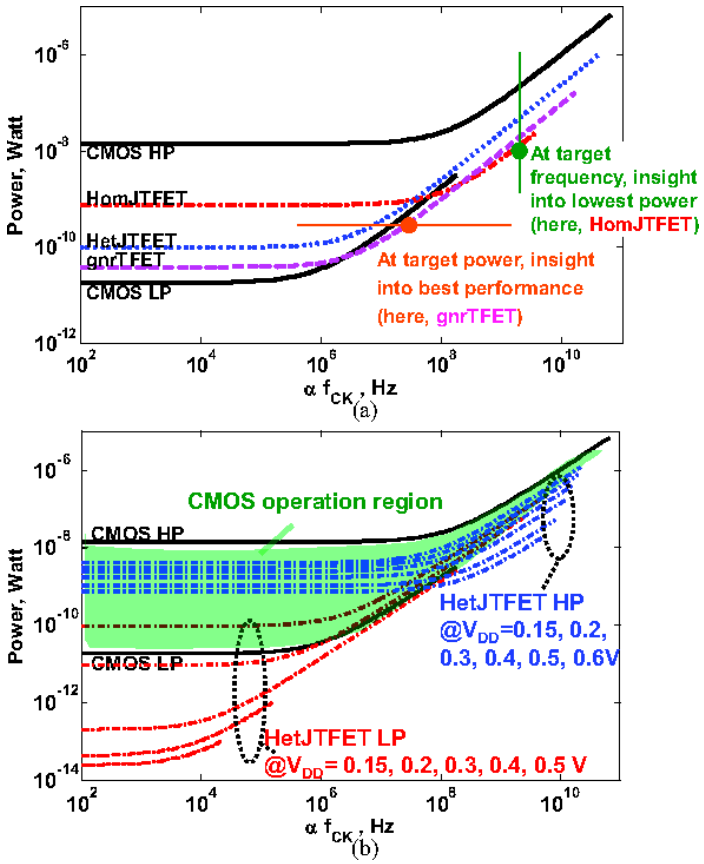

Fig. 2. (a) Comparison of power dissipation of an inverter in several technologies. (b) Comparing HP and LP versions of HetJTFET with CMOS.

capacitance replacing the device capacitance in (1) and (2). The results, given in Fig. 1, show how CMOS devices (i.e. CMOS HP and LP) compare with these beyond-CMOS devices. The uniform approach used to derive the characteristics of each device allows for better one-to-one comparisons to be made. For example, HTFETs can be observed to have a lower switching energy, but their switching speed is between a high performance and low-power CMOS.

One shortcoming of Fig. 1 is that it does not provide any information about the standby power dissipation. It is preferred to compare the total power dissipation instead of the dynamic power dissipation. The total power dissipation of a gate is given by [5] as:

$$P_D = I_{OFF}V_{DD} + \alpha C_L V_{DD}^2 f_{CK}, \tag{5}$$

where I_{OFF} is the leakage current, f_{CK} is the clock frequency, and α is the activity factor. C_L is the capacitive load including both the interconnect and device capacitances which can be found for a given fanout (e.g. fanout of four or FO4). The speed of the technology is reflected in the maximum clock frequency ($f_{CK,max}$), which is determined by the delay of the critical path in a system. Fig. 2(a) compares the power dissipation of a FO4 inverter in several technologies. It is assumed that the ratio of the critical delay path to the delay of an inverter is 10 in all technologies and the curve corresponding to each technology extends up to the resulting $f_{CK,max}$. As it can be seen, each technology is more suitable for a certain application domain in

terms of a lower power dissipation or a higher performance. A more detailed comparison amongst different technologies can be performed when (5) is evaluated for devices optimized for high performance (HP) or low-power (LP) at different power supplies. For instance, power dissipation of HP and LP HTFET-based inverters is depicted in Fig. 2(b); the shaded area between CMOS HP and CMOS LP curves indicates the application ranges that can be covered by the CMOS technology. The curves outside the shaded area indicate that HTFET is a promising option in two applications domains; the LP version is attractive for ultra-low-power systems operating at low clock frequencies (e.g. biomedical applications). The HP version is attractive for low-voltage low-power GHz-range operation (e.g. high performance mobile devices).

III. Device to Architecture Abstraction

In this section, we extend the comparison between the CMOS and HTFET devices evaluated in Section II to a processor architecture level of abstraction. Processing paradigms differ, depending on the application characteristics. These can be broadly classified into general purpose and domain-specific computing paradigms. The modeling methodologies in each case are very different. For instance, use of architecture simulators like GEMS [6] and power estimation tools like McPAT [7] are sufficient to model the former with a significant degree of accuracy, as different general purpose processors share many common or parametrized structures. On the other hand, the customized datapaths of application-specific processors require detailed hardware synthesis. We describe the methodology for each of these cases in detail.

A. Device to Architecture Extrapolation of a General Purpose Processor Model

In [8], the authors modeled a simple in-order processor as a ring oscillator with matching critical path delay. In contrast, modeling an out-of-order processor in a similar manner is more difficult, due to its inherent non-sequentiality of operation. For this purpose, core power numbers for a Niagara-3 like processor were obtained using McPAT-0.8 integrated with GEMS-2.1, for a 20 nm Si FinFET technology. The corresponding HTFET core power values were obtained by scaling the per-transistor power of the FinFET and HTFET at different supply voltages.

Figure 3 shows the variation in total core power with frequency for the Si FinFET and HTFET Models at the 20 nm technology node. As explained in Section II, devices with different characteristics and configurations are attractive for different application domains. Here, the HTFET model that is used for a general purpose core is akin to a low leakage, high performance device with an operating voltage of 0.5V. The crossover frequency F_{cross} is defined as the frequency below which HTFET processor operation is more energy efficient than that the CMOS FinFET based processor. From the figure, this crossover frequency can be observed to be around 1.4 GHz.

B. Domain-Specific HTFET processors: Design of HTFET Standard Cell Library

Figure 4 shows the design flow of a 20 nm III-V $In_{0.9}Ga_{0.1}As/GaAs_{0.18}Sb_{0.82}$ HTFET based standard cell

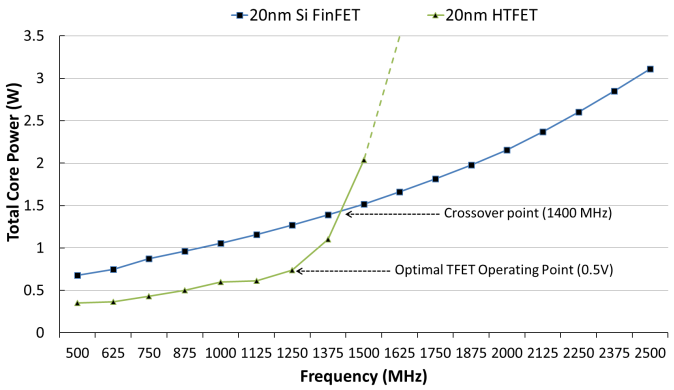

Fig. 3. Comparison of Power-frequency characteristics of 20nm Si FinFET and HTFET-based processors

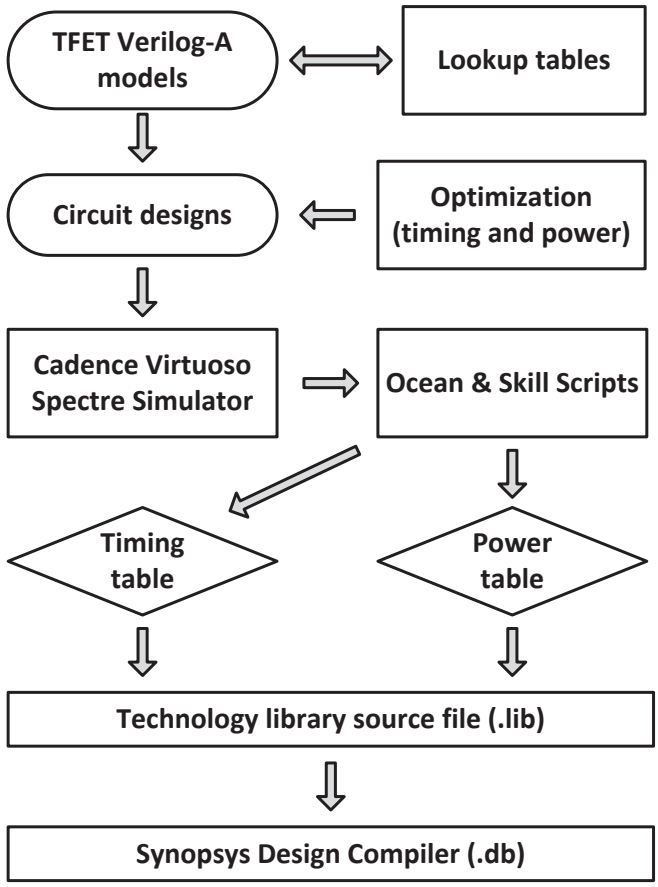

Fig. 4. The design flow of 20nm III-V InAs/GaSb HTFET based standard cell library

library. The performance of HTFET at high frequencies and its energy efficiency at low supply voltage ($< 0.5V$) has been demonstrated in [9]. In order to characterize each cell library, Verilog-A models with lookup tables from TCAD Sentaurus [10] device simulation have been employed for HTFET based combinational and sequential cell designs. This standard cell library provides a maximum fanout of 16 for each of these cell designs. Each cell in the standard cell library refers to characterized 7×7 lookup matrices with input

transition time and total net capacitances in order to determine specific value for the timing and power consumption during the processing of synthesis.

For device-logic-cell simulations, the noise characterized HTFET Verilog-A models with intrinsic noise components (e.g., shot noise, white noise and flicker noise) are implemented, and each cell design is optimized for timing and power consumption. For timing and energy calculations, a bottom-up methodology with device-circuit level simulations is utilized to generate the HTFET based standard cell library at 0.3V supply voltage using the Cadence Virtuoso Spectre simulator [11]. Further, this library file is compiled and converted into the database format by Synopsys Design Compiler [12] in order to make it compatible with the architecture level description languages with the HTFET based standard cell library. A This makes it possible to synthesize various accelerator designs and carry out accurate performance and power comparisons with corresponding CMOS cell-based designs.

IV. Examination of General Purpose Computing Paradigms

General purpose processors are meant to cater to a large diversity of applications. These applications may vary in their scalability with respect to number of cores (thread-level parallelism), in their utilization of available processor or memory resources or in their sensitivity to peak single-threaded performance. In addition, the evaluation metrics and constraints may vary depending on the application domain. For instance optimizing energy and consequently battery life or minimizing power for a given performance baseline may be more important than peak performance in a power constrained domain such as mobile processors. On the other hand, high end server domain processors aim to maximize performance amidst controlling temperature in order to minimize cooling costs. These hybrid power-performance-temperature metrics opens up a huge design space in the micro-architecture and architecture domains. As described in Section I, CMOS and HTFET processors are capable of optimal operation in different regions of this design space. Although the high frequency operation of HTFET processors is limited in comparison to CMOS which prevents HTFET devices from being a direct replacement for CMOS technology, intelligent device-architecture co-design can enable us to bridge this gap in performance.

A. Power and Thermally constrained application scheduling

Power constrained performance optimization in the context of a CMOS-HTFET heterogeneous multicore has been explored in [8, 13]. These works assume a uniform simple microarchitecture across all cores. This reduces the likelihood of thermal hotspots developing due to asymmetric microarchitectures and makes power capping feasible.

On the other hand, when microarchitectures vary across cores, merely capping the overall power and limiting the entire power consumption to a small fraction of the chip in itself does not adequately address the thermal concerns resulting from the increasing power density problem. The Thermal Design Power (TDP) of a processor chip, defined as the power which the processor can dissipate without exceeding the maximum allowable chip temperature, is used as a metric to determine the power budget of processors. Dissipating all the power in a smaller area causes a significant increase in peak temperature due to higher power density. Hence, one should also take into account the wide range of application domains in which the processor can be utilized. These domains can be effectively characterized by the thermal limit that they entail. For instance, a mobile-based ARM-like embedded core operates under a much more stringent temperature limit than a Xeon-based server architecture. In the former case, CMOS cores are forced to operate at sub-optimal frequencies with limited microarchitecture flexibility. This provides opportunities for HTFETs, which, being more power and consequently thermal-efficient at these temperatures can operate over a much wider range of microarchitecture complexities. Thus HTFETs can attain more optimal states in the frequency-issue-width design space.

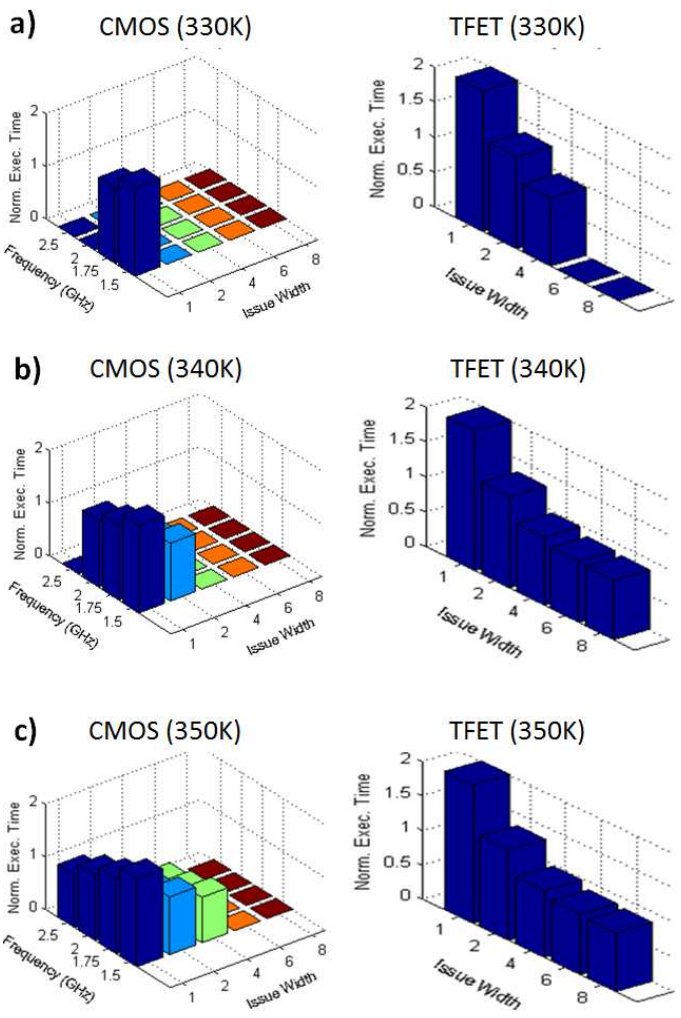

Fig. 5. Permissible states in the frequency-issue-width design space for CMOS and HTFET processors at a) 330K, b) 340K and c) 350K temperature limits

Figures 5 shows the possible configurations that can be attained under different thermal budgets by CMOS and HTFET processors. This figure clearly demonstrates that HTFETs can operate at higher issue widths at lower temperatures, while CMOS cores can reach higher operating frequency as the thermal limit is increased.

The results from all our tradeoff comparisons at the architecture and microarchitecture level are summarized in Tables I, II and III. Table I shows the performance improvement obtained by replacing a homogeneous multicore system with a heterogeneous CMOS-HTFET configuration, under a 1 W per core power budget. The advantages of task scheduling and power partitioning techniques over a naive technology substitution on the multicore are also highlighted.

Table II shows the speedup obtained using different microarchitectures in CMOS and HTFET processors under different thermal constraints. It can be observed that HTFET cores actually perform worse than CMOS cores at 350K. Similarly Table III shows the Energy-Delay Product (EDP) of the best performing HTFET and CMOS core configurations at the above temperature limits. The disparity in the EDP between CMOS and HTFET cores reduces as the thermal limit increases, until CMOS is more energy efficient at thermal limits in excess of 350K. These results clearly illustrate the diminishing returns that HTFETs demonstrate as we transition to higher temperature domains.

Processor Configuration	Application Type /Power Mapping	Normalized Speedup
Homogeneous multicore (CMOS or HTFET)	Static	1.0
Heterogeneous multicore (CMOS-HTFET)	Static	1.05
Heterogeneous multicore (CMOS-HTFET)	Dynamic	1.22

TABLE I. POWER CONSTRAINED APPLICATION MAPPING ON A CMOS-HTFET HETEROGENEOUS MULTICORE, WITH A 1 W/CORE POWER BUDGET

Processor Configuration	Type of Benchmark	Normalized Speedup		
		330K	340K	350K
Heterogeneous CMOS-HTFET Multicore	Single threaded /programmed	1.46	1.19	1.01
Heterogeneous CMOS-HTFET Multicore	Multi threaded /programmed	1.26	1.12	0.84

TABLE II. SPEEDUP W.R.T A HOMOGENEOUS CMOS (OR HTFET) BASELINE FOR SINGLE AND MULTITHREADED WORKLOADS FOR DIFFERENT TEMPERATURE LIMITS

Processor Configuration	Type of Benchmark	Normalized Speedup		
		330K	340K	350K
Heterogeneous CMOS-HTFET Multicore	Single threaded /programmed	0.41	0.65	0.80
Heterogeneous CMOS-HTFET Multicore	Multi threaded /programmed	0.45	0.68	1.11

TABLE III. NORMALIZED ENERGY-DELAY PRODUCT W.R.T A HOMOGENEOUS CMOS (OR HTFET) BASELINE FOR SINGLE AND MULTITHREADED WORKLOADS FOR DIFFERENT TEMPERATURE LIMITS

Although HTFET cores for general-purpose processing may have an advantage over CMOS only in domains with tighter thermal constraints, they can still play a valuable role in augmenting high-end devices by providing very efficient specialized coprocessors. Over the last few processor generations, customizing architectures with logic optimized for domain-specific applications, such as graphics, multimedia, or cryptography kernels, has gained importance alongside traditional process-shrinking based improvements in processor performance. In the following section, we examine the viability of using HTFET-based accelerators as an energy efficient alternative to conventional technology without compromising performance.

V. EXAMINATION OF DOMAIN SPECIFIC COMPUTING PARADIGMS

A. Synthesis of domain-specific accelerators

As part of our experiments, we synthesized a sample accelerator that computes the Euclidean distance between two vectors. This accelerator can be employed in feature-matching algorithms, where the Euclidean distance between a test feature descriptor vector (such as those obtained from algorithms like SIFT or SURF) and every feature vector from the training database is computed for recognition tasks. The input consists of 64-dimensional vectors, each element being 16 bits in width. The accelerator consists of of an 8-stage pipelined execution unit consisting of an array of multipliers and adders that compute the sum of squares of element-wise difference between a pair of vectors. The accelerator HDL code was simulated and verified using Synopsys VCS [14]. The designs were synthesized using the in-house HTFET standard cell library, using Synopsys Design Compiler [12].

Figure 6 shows the block diagram corresponding to an accelerator for computing Euclidean distance between two vectors.

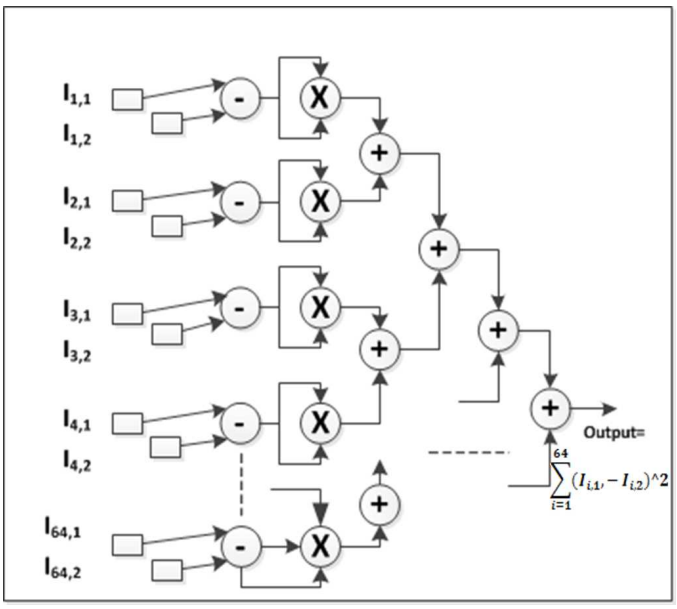

Fig. 6. Block diagram for computation of Euclidean distance

The overall execution time includes the data transfer time for streaming the input to the accelerator using DMA transfer from external memory and the accelerator computation time

for 200 feature vectors in the training database and 1 test vector. For an input stream of images, the throughput of the HTFET accelerator was computed to be 19392 frames/second.

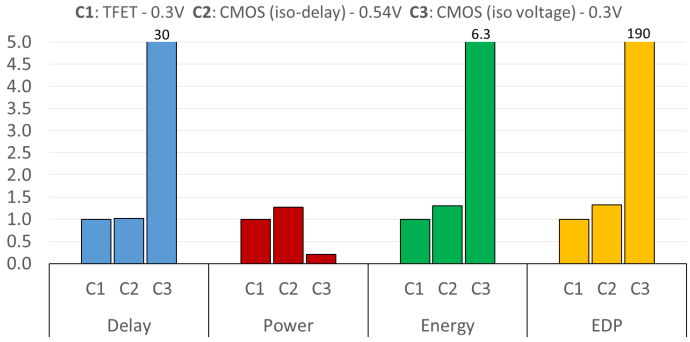

Fig. 7. Normalized delay, power, energy and EDP of HTFET, iso-performance CMOS and iso-voltage CMOS accelerator designs. All results are normalized w.r.t the HTFET design.

B. Comparison with existing CMOS designs

The salient feature of HTFET based accelerators is their capability to achieve high throughput with energy efficiency even while operating at low supply voltages. This is evident when the HTFET-based Euclidean distance accelerator was compared with equivalent CMOS designs. An iso-voltage CMOS design operating at 0.3V is severely limited by its inherent sub-threshold operation and consequently has a transistor delay that is over 30X higher than HTFET. Hence a TFEET accelerator design is more feasible, even though its total power consumption is higher than the iso-voltage CMOS design due to its far superior switching speed. Consequently, the accelerator is much slower, and can be clocked at less than 100 MHz in order to meet timing constraints. In order to match the performance of the HTFET accelerator, the CMOS design will have to operate at 0.54V. This increases the power and hence, the energy overheads. The performance, power, energy and energy-delay product results are summarized in Table 7.

VI. Conclusion

In this paper we examined the feasibility of adopting steep slope devices as viable alternatives to the FinFETs used in existing CMOS-based processors. We first evaluated Heterojunction Tunnel FET devices in simple circuits such as inverters and then carried out benchmarking comparisons with different types of CMOS technologies. We explored the use of HTFET processors in various application domains, determined by the power and thermal constraints placed on the processor, and demonstrated that, while HTFETs are not a universal replacement for CMOS, they are dominant in some very relevant domains. Following from these findings, we extended our studies to encompass domain-specific accelerator architectures and used our in-house TFET standard cell library to synthesize the accelerator designs. Our evaluation demonstrates the significant energy and performance benefits of adopting TFET based technology for accelerator designs.

Acknowledgments

This work was supported by the Center for Low Energy Systems Technology (LEAST), one of six centers supported by the STARnet phase of the Focus Center Research Program (FCRP), a SRC program sponsored by MARCO and DARPA. It was also supported by the NSF NERC award 1160483. The authors also wish to thank Huichu Liu for her valuable inputs.

References

[1] S. Mookerjea *et al.*, "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 *IEDM*, 2009.

[2] A. C. Seabaugh and Q. Zhang, "Low-voltage tunnel transistors for beyond CMOS logic," *Proceedings of the IEEE*, Dec.

[3] D. Nikonov and I. Young, "Overview of beyond-CMOS devices and a uniform methodology for their benchmarking," *Proceedings of the IEEE*, vol. 101, no. 12, pp. 2498–2533, Dec. 2013.

[4] C. Augustine, A. Raychowdhury, Y. Gao, M. Lundstrom, and K. Roy, "PETE: A device/circuit analysis framework for evaluation and comparison of charge based emerging devices," in *Quality of Electronic Design, 2009. ISQED 2009. Quality Electronic Design*, 2009, pp. 80–85.

[5] E. Morifuji, T. Yoshida, M. Kanda, S. Matsuda, S. Yamada, and F. Matsuoka, "Supply and threshold-voltage trends for scaled logic and SRAM MOSFETs," *IEEE Trans. Electron Devices*, vol. 53, no. 6, pp. 1427–1432, Jun. 2006.

[6] M. Martin *et al.*, "Multifacet's general execution-driven multiprocessor simulator (GEMS) toolset," *SIGARCH Comput. Archit. News*, 2005.

[7] S. Li *et al.*, "McPAT: An integrated power, area, and timing modeling framework for multicore and manycore architectures," in *MICRO*, 2009.

[8] E. Kultursay *et al.*, "Performance enhancement under power constraints using heterogeneous CMOS-TFET multicores," in *CODES*, 2012.

[9] R. Bijesh *et al.*, "Demonstration of In0.9Ga0.1As/GaAs0.18Sb0.82 near broken-gap tunnel fet with Ion=740uA/um, GM=700us/um and gigahertz switching performance at Vds=0.5V," 2013, accepted.

[10] "TCAD Sentaurus Device Manual," 2010.

[11] Cadence, "Cadence virtuoso spectre circuit simulator," 2009.

[12] Synopsys, "Synopsys design compiler," 2010.

[13] K. Swaminathan, E. Kultursay, V. Saripalli, V. Narayanan, M. Kandemir, and S. Datta, "Steep-slope devices: From dark to dim silicon," *Micro. IEEE*, vol. 33, no. 5, pp. 50–59, 2013.

[14] Synopsys, "Synopsis VCS (verilog compiled code simulator)," 2009.