Thermal-aware application scheduling on device-heterogeneous embedded architectures

Karthik Swaminathan, Jagadish Kotra, Huichu Liu, Jack Sampson, Mahmut Kandemir and Vijaykrishnan Narayanan
Pennsylvania State University, University Park, PA 16802

***Abstract*—**

The challenges of the *Power Wall* manifest in mobile and embedded processors due to their inherent thermal and form-factor constraints. The power dissipated over a fixed area, namely, the power density, directly affects acceptable core temperatures even for low-power devices. In this paper, we examine techniques to counter this power density increase with device and microarchitecture-level heterogeneity. We explore the design space in which various parameters such as frequency and micro-architectural complexity can be traded off against each other in order to achieve the optimal configuration for a fixed temperature limit. Since conventional CMOS technology based cores may not satisfy our performance and power requirements, especially under tight thermal constraints, we propose a heterogeneous CMOS-Tunnel FET multicore for obtaining the optimal operating points under power and thermal limitations. Using a profiling based static assignment scheme, we demonstrate the improvement obtained by coupling this device-level heterogeneity to architectural modifications. We also propose an instruction slack-based scheme to map applications on the heterogeneous multicore. Our schemes show an improvement of up to 47% performance and 30% energy above the best homogeneous configuration.

I. Introduction

Transistors have continued to scale down to smaller feature sizes, albeit without the constant field scaling that has marked their miniaturization for more than two decades. However, the coming CMOS technology generations are expected to be bound by an inherent physical limitation to the transistors' sub-threshold slope, which prevents them from switching quickly and efficiently at near- or sub- threshold supply voltages. This poses either an increased delay or unacceptable increase in leakage power. Consequently, supply voltage reduction has not kept pace with reduction in transistor size and has resulted in increased power density.

Recently, a new generation of steep slope devices enabling complete turn on/turn off with a limited voltage swing have emerged. The physics of these devices enables them to achieve sub-60 mV/decade sub-threshold slopes. This leads to higher $(I_{on})/(I_{off})$ ratios at low voltages, which translates into higher drive currents (better performance) at low voltages and lower off-state leakage currents than CMOS transistors. Different implementations of such steep slope devices include NEMS [1] and *Inter-Band Tunnel Field Effect Transistor (TFET) transistors* [2]. Efficient logic and memory structures have already been demonstrated using TFETs [3] and it is projected as one of the commercial front-runners of the new post CMOS technologies in the future technology nodes [4].

Thermal constraints assume additional importance in embedded domains since the thermal limit directly impacts the product and packaging costs. They also affect overall form factors due to additional cooling schemes and diminish the energy efficiency. To this end, we examine tuning architectural parameters like processor issue-width and frequency to operate under thermal constraints for various application domains and employing heterogeneous CMOS/TFET cores to maximize both the performance and energy efficiency. These various knobs present a multi-dimensional design space for various architectural domains ranging from the embedded space to the high-end server space. In an application domain such as mobile computing, these constraints on peak temperature are especially stringent as there is limited flexibility in terms of cooling techniques and longer wirelengths to reduce on-chip hotspots, as compared to higher end systems. For instance, cooling mechanisms are unaffordable due to the small form factor of the entire processor system, while increasing chip area significantly can increase die and manufacturing costs. The design space of exploring the appropriate micro-architectural trade-offs in hybrid architectures under these constraints remains heretofore unexplored.

In this paper we make the following novel contributions:

- We propose using steep-slope device-based processors as complementary cores in mobile processors that operate under tight thermal constraints. These processors would serve to expand the design space along with modifying architecture and system parameters, enabling us to achieve improvements to both performance and energy efficiency under these constraints.
- We demonstrate techniques to optimally map single and multiple application workloads onto this heterogeneous mobile processor. We also propose techniques to dynamically swap application threads from one core-type to another, depending on the dynamic behavior of the application.
- We re-examine the design of existing heterogeneous architectures such as the ARM big.LITTLE [5] processor when allied with device heterogeneity and conclude that a heterogeneous CMOS-TFET design can effectively run applications that prefer either a high operating frequency or wide-issue configurations that exploit high instruction level parallelism.

The rest of the paper is arranged as follows. Section II explains the problem addressed by this paper and the motivations for our proposed scheme. Section III provides details on the device characteristics and our modeling techniques. Section IV explains the dynamic algorithms that we adopt to obtain the optimal state. Section V describes the simulation infrastructure with our system configuration while Section VI discusses our experiments and the results we obtained. Section VII presents the related work while distinguishing our work. We conclude with Section VIII.

II. Motivation

One of the major limitations observed in TFET processors at the 22nm node is their low peak performance, which is due to their higher device switching delay. However, this disparity between CMOS and TFET devices reduces when extrapolated to future technology nodes [6]. By the 10nm node, TFET cores can attain 75% of the peak performance of 10nm CMOS, as compared to 50% for the current technology node. In addition, TFETs become more and more power efficient w.r.t CMOS with each subsequent generation, with the relative power consumption between CMOS and TFET processors increasing from 3.2× at 22 nm to 4.6× at 10 nm.

A. *Variations in application behavior with microarchitectural complexity*

Given fixed power and thermal constraints for a multicore architecture, there are several points in the design space that can be explored. From a microarchitectural perspective, this includes varying core complexity in terms of number of instructions fetched per cycle, issue width, size of register file and issue queue and the number of execution units. Depending on the microarchitecture configuration, the relative contributions of dynamic and leakage energy with respect to performance vary significantly. Further, workload characteristics also impact the efficiency of the various microarchitecture components. Based on the nature of the application, the impact on performance and energy due to the intrinsic datapath frequency or external resources such as the memory subsystem would also vary in different proportions.

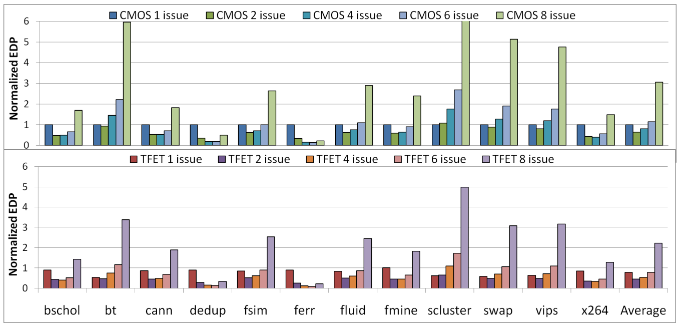

Fig. 1. Variation of energy delay product (normalized to single issue CMOS) for CMOS and TFET cores with varying issue widths

We first examine the diversity in application behavior across different device and microarchitecture configurations. Details of the methodology used in obtaining a device-to-processor abstraction for TFETs are provided in Section III. Figure 1 demonstrates the variation in energy-delay product (EDP) for different core configurations for both Si FinFET and TFET based core designs. There is a wide variation across applications for the best core configuration that minimizes the EDP In general, applications with high throughput (*dedup*, *ferret*) are better able to exploit the higher core complexity on account of their higher instruction-level parallelism (ILP). On the other hand, applications like *streamcluster* show hardly any improvement with increase in issue width and would prefer operating more energy-efficiently on lower issue width cores.

B. *Thermal constraints based microarchitecture design*

Depending on domain, the peak permissible temperatures vary. For instance, a mobile processor can tolerate a far lower peak temperature than a server. For instance, Samsung Galaxy phones containing ARMv6 processors are rated to operate at a maximum of less than 57°C (330K) [7], while most servers can attain upto 100°C (373K) temperatures. Since the work in this paper is primarily restricted to the embedded domain, we examine different architectural configurations with thermal limits in the 330-350K temperature range.

We use the Hotspot-5.02 [8] thermal estimation tool for obtaining the peak core temperatures and generating thermal profiles. Since TFET devices share the same substrate and material characteristics as Si FinFETs apart from the few atoms used in doping, the thermal characteristics of TFETs are similar to that of CMOS devices. The CMOS and TFET power profiles used as input to the tool are obtained from periodic power traces using McPAT.

In this work the compatibility of TFET technology with CMOS, leading to possible heterogeneous integration [9] has led us to focus on a joint evaluation of CMOS and TFET cores at the microarchitecture level for expanding this thermally bound design space.

C. *Frequency-complexity Tradeoffs*

Since it is also possible to achieve higher processor complexities for a given thermal limit by simply reducing the processor frequency, we jointly examine the design space of both core frequency and processor complexity.

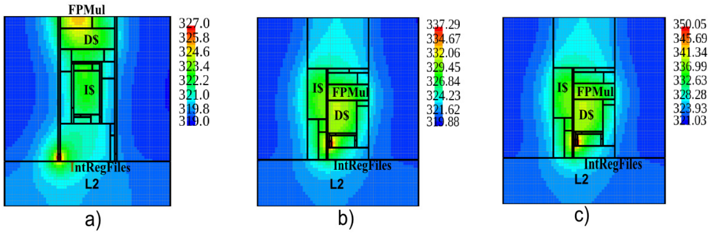

Fig. 2. Comparison of thermal profiles of cores corresponding to the best performing CMOS configuration for temperature limits of a) 330K (1 issue, 1750MHz), b) 340K (2 issue, 2 GHz) and c) 350K(4 issue, 1.75GHz). CMOS frequencies below the crossover are not shown as TFETs are inherently more energy efficient at those points.

Figure 2 shows the variation in temperature across different micrarchitectural components in each processor configuration. The peak temperature of these cores determines which configuration is permissible under that thermal constraint. We observe that, at lower thermal limits, CMOS cores have a very limited set of permissible microarchitectural configurations. In fact, even a single issue CMOS processor cannot operate at frequencies above 1.75 GHz for the 330K limit. On the other hand, TFET cores are able to operate at much higher issue widths. This compensates for the lower operating frequency of TFET processors as compared to CMOS. As the temperature limit exceeds 350K, CMOS cores are also able to operate at higher frequencies with sufficiently high issue width configurations. Since we observed the benefits of increasing issue width beyond 4 in mobile applications to be negligible, our simulations restrict the peak issue width to 4. Consequently, the higher frequency of CMOS cores becomes the dominant factor and they outperform TFET cores.

In addition to performance, battery life is also a concern, especially for embedded devices. Hence we also look to minimize the energy under the previously described temperature constraints. Based on these permissible configurations, it is possible to determine the best configuration in terms of both performance and energy.

III. TFET Device Characteristics and Simulation Details

In this section we provide a brief overview of the TFET device characteristics and describe a methodology for modeling entire processors realized from these devices.

A. Modeling of low leakage TFET processor

To fulfill both operating frequency and leakage power requirements, we tuned the device characteristics to realize a low static power HTFET (*LSTP HTFET*). Compared to the previously reported TFET designs optimized for dynamic operation power (LOP HTFET [10], the low leakage HTFET employs a relaxed channel length (10% relaxation) to achieve reduced static leakage power, since a reduced short channel effect has been observed in 20nm gate TFETs [11]. This design of LSTP HTFET still offers a desired drive-current at low supply voltage with an optimized leakage power, since the field across the tunnel junction is the primary determinant of the performance. We developed circuit models of the low leakage TFET for extracting parameters for our architectural simulation. This model has been calibrated with atomistic simulations and is consistent with fabricated devices [12]. As shown in 3(a), LSTP HTFET shows comparable on-state current with LOP HTFET, which outperforms Si FinFET technology used for our CMOS cores at 0.5V operation. For the off-state leakage current, the LOP HTFET shows a cross-over with Si FinFET at 0.5V operation, while the Low Leakage HTFET shows 2× leakage reduction compared to LOP HTFET at the same supply voltage and with a cross-over at 0.3V compared to Si FinFET (Figure 3(b).

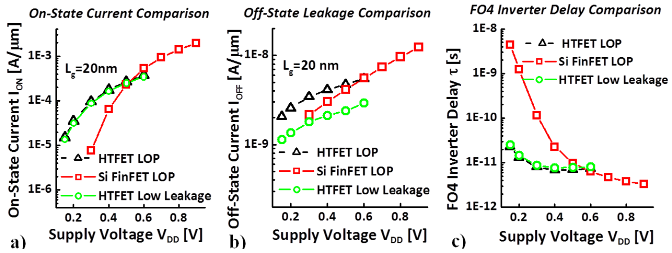

Fig. 3. (a) On-state current (b) off-state leakage (c) FO4 inverter delay comparison with V_{dd} scaling for HTFET LOP model, Si FinFET model and HTFET low leakage model at 20nm gate-length.

B. Extrapolation to Processor Model

Circuit simulations of FO4 inverters have been conventionally used to determine the relationship between the voltages and circuit frequencies. In prior works such as [10], the processor critical path delay have been modeled using multi-pipeline stage ring oscillators. However, due to the diversity in critical paths in more complex out-of-order cores, and the corresponding impact of non-logic components such as interconnects, a this model may not be sufficiently accurate to model the performance and power characteristics of these cores. Hence, a more detailed abstraction model is required, especially at the architecture level.

The device models used for simulating TFET and FinFET characteristics are similar to those used in [6]. We used the GEMS full system simulator [13] for running performance simulations for different processor configurations, while core power estimates were obtained using McPAT-1.0 [14] for a 20 nm Si FinFET technology running at different core frequencies. In addition, changes were made to the McPAT source code to incorporate wire delay and power overheads as well. In order to obtain TFET core power numbers, the FinFET logic power was scaled in the ratio of the TFET to CMOS transistor switching power. The wire power remained almost constant, with the exception of repeaters and buffers which also underwent device scaling. Validation of these models were done with the help of Fabscalar [15], which generates synthesizable HDL code for different micro-architectural configurations. Synthesizing the Fabscalar cores of different issue widths enabled us to determine critical path delay and power of the cores and match it to those obtained from our models. The modeling details are shown in Figure 4.

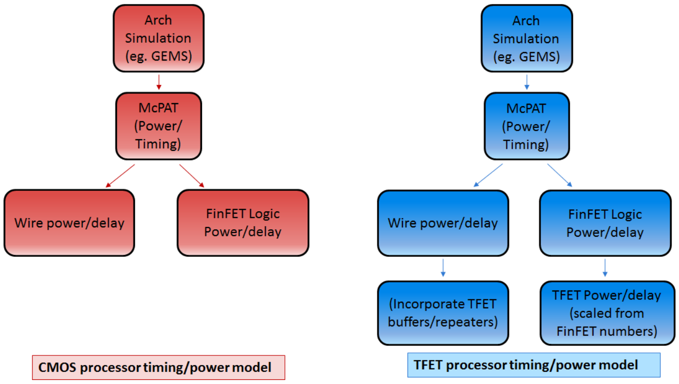
Arch Simulation (eg. GEMS)
McPAT (Power/ Timing)
Wire power/delay
FinFET Logic Power/delay
CMOS processor timing/power model
Arch Simulation (eg. GEMS)
McPAT (Power/ Timing)
Wire power/delay
FinFET Logic Power/delay
(Incorporate TFET buffers/repeaters)
TFET Power/delay (scaled from FinFET numbers)
TFET processor timing/power model

Fig. 4. Design flow for modeling a)CMOS and b)TFET cores

Figure 5 shows the variation in total core power with frequency for the Si FinFET and both the LOP and Low Leakage TFET Models. The crossover frequency F_c is defined as the frequency below which TFET processor operation is more energy efficient than that the CMOS FinFET based processor. The lower leakage energy of the LSTP HTFET results in more efficient operation as compared to the LOP TFET device. This increases the desired operating frequency for LSTP HTFET compared to CMOS to almost 1.5 GHz (compared to 1.2 GHz for LOP TFET). Further, incorporating wire components to the existing processor model results in an increase in the crossover frequency in comparison to the crossover frequency in the absence of wire effects. This is because, the non-scaling of wire-delays causes the frequency gap between CMOS and TFET processors to shrink, increasing the feasible design space for TFETs.

Works such as [16] and [17] have demonstrated heterogeneous integration of Si-FinFET and III-V devices, which can make it possible to manufacture CMOS and TFET cores on a single layer. Hence, our design comprises of a heterogeneous dual-core system with both CMOS and TFET cores.

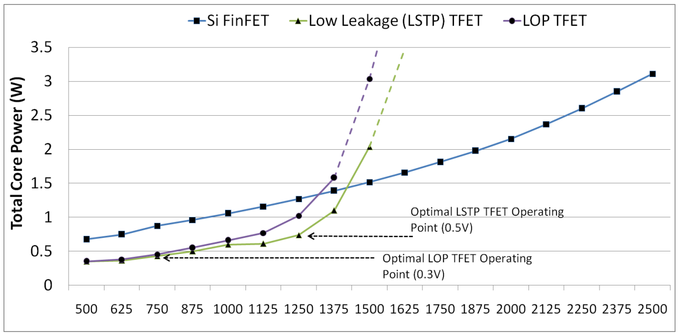

Fig. 5. Comparison of Power-frequency characteristics of Si FinFET, LOP TFET and Low Leakage TFET based processors

IV. Architecture Design Details

In this section we describe the techniques used to map applications on the embedded processor, under thermal constraints. In addition to a simple static mapping scheme, we also explain a dynamic mapping scheme based on the runtime instruction slack of the application.

A. Static mapping of applications

In these evaluations, we determine the best possible operating point in the frequency-issue-width design space, for each application for different thermal limits. We then arrive upon the configuration preferred by a majority of the applications for each temperature domain. Each application was run at its optimal frequency for that configuration. Depending on the static profiling results, it is possible to determine whether the application has a higher affinity for a CMOS or a TFET core.

Statically mapping applications to CMOS/TFET cores may not always achieve the desired results on account of periodic changes in program phase and characteristics. Hence we attempt to exploit the application phases with high ILP by running them on TFET cores, which are capable of attaining more complex configurations than CMOS within the same thermal budget. However, their limited performance at high operating voltages precludes them from optimally running low ILP applications which prefer higher frequencies. Hence we require a metric to determine the degree of ILP of an application. For this purpose, we employ a runtime slack estimation technique derived from [18].

B. Slack-based dynamic mapping

Figure 6 shows our slack estimation method. Assuming no dependent instructions in the ROB, instruction I_i can be delayed at most by C_k number of cycles before instruction I_k is executed in its designated cycle. To estimate C_k (in cycles), we divide the number of instructions between I_i and I_k by the IPC of the current epoch (IPC_{epoch}). This is defined as $\Delta(i,k) = I_i - I_k$. This implies that I_i has enough slack until the exact cycle where I_k gets committed. However, since only a finite number of instructions can be committed every cycle, I_i should be ready latest by $\Delta(i,k)/max\_commit$ cycles before I_k is ready to commit. Thus the total slack in this case would be given by C_k, as shown. However, if I_j has a true dependency on I_i, then I_i has to be executed before I_j, and the slack would be denoted by C_j.

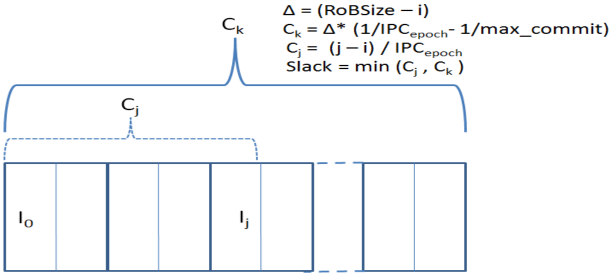
Δ = (RoBSize – i)
C_k = Δ* (1/IPC_epoch- 1/max_commit)
C_j = (j – i) / IPC_epoch
Slack = min (C_j , C_k)
C_k
C_j
I_O
I_j

Fig. 6. Runtime Slack Estimation

Most of the information required to compute slack is already in the processor. The dependency information for each instruction is stored in the ROB. The IPC can be estimated from the hardware performance counters.

Based on the slack estimated at runtime for each epoch, we migrate the application to run on a best configuration core based on the applications sensitivity to slack. If the slack determined in the epoch is higher than a prescribed threshold, it means that the application is less sensitive to slack. Hence we migrate the application to run on a low frequency high issue-width TFET core. Similarly, if the application is currently running on a TFET core and is very sensitive to slack, we migrate it to a high frequency CMOS core. We assumed the low overhead thread migration techniques from [19] to migrate threads at runtime.

In a real system, these techniques are implemented by affixing temperature sensors near the thermal hotspots corresponding to each core. These sensors trigger a frequency scaling or core migration operation, whenever the temperature approaches the prescribed limit.

V. Simulation Infrastructure

TABLE I. Configuration of the evaluation platform.

CMOS Processor Technology	upto 4 issue SPARC ISA 20nm Si FinFET
TFET Processor Configuration	upto 4 issue SPARC ISA 20nm Interband Tunnel FET
L1 Cache	32 KB D/I 64B Cache Line, 4 way S.A
L2 Cache	512KB shared LLC
DRAM	2GB, DDR2-1600, 1 memory channel

A. Thermal-aware Core configurations

Table I shows the processor parameters used during our architectural simulations. The experiments run on this configuration consist of applications from the *mibench* suite [20], which are typically used as commercial representatives for embedded system architectures. Our baseline comprises a single CMOS core with the optimal configurations as shown in Table II. The heterogeneous configuration comprises 2 cores, 1 CMOS and 1 TFET. For the dynamic case, which we term as *DynMap*, our architecture consists of a single issue CMOS core and a multiple (4-) issue TFET core, both running at the highest possible frequency within the thermal limit.

TABLE II. Configurations at different thermal limits

Processor	330K	f_{max}	340K	f_{max}	350K	f_{max}
	Issue	(GHz)	Issue	(GHz)	Issue	(GHz)
CMOS	1	1.75	2	2.0	4	1.75
TFET	4	1.25	4	1.25	4	1.5

B. Simulation tools

For device simulations, we employed the look-up table based Verilog-A 20nm technology models developed from TCAD Sentaurus [21]. We used a calibrated Si FinFET Verilog-A model also obtained from TCAD Sentaurus simulation for baseline comparison. We used the GEMS simulator [13] for performance estimation of each workload. We used McPAT [14] to estimate the power consumption of the entire core as well as individual microarchitecture components. For power estimation of TFET processors, we modified the other technology parameters in McPAT to correspond to our TFET device models. The power numbers obtained periodically from McPAT were then used by Hotspot-5.02 [8] to create a power trace and consequently a thermal profile of the core during the execution of the workloads.

VI. Results

A. Static mapping of applications

Figures 7 and 8 show the speedup and energy of a static scheduling scheme, where the best core configuration is selected for each application. All results are normalized to a homogeneous system comprising of the best CMOS architectural configuration for that application.

We can observe that the overall speedup and energy savings increases as the thermal limit is raised. This is because as the thermal budget increases, the number of attainable configurations in terms of issue width for the baseline CMOS core also increases. Hence, at the 350K limit, it is possible to operate a 4 issue CMOS core at a higher frequency than its TFET counterpart, negating any improvements due to heterogeneity. The maximum harmonic mean speedup due to static mapping is observed to be 43% at 330K with an energy savings of 27%.

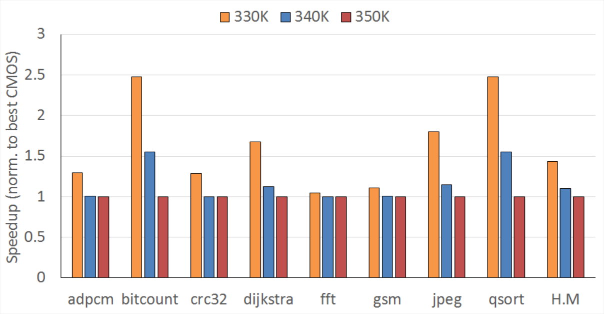

Fig. 7. Speedup on heterogeneous multicore with static mapping on best homogeneous CMOS configuration for thermal limits of 330K, 340K, 350K

B. Dynamic migration of applications

When the dynamic migration scheme described in section IV is implemented, it is possible to account for intra-application phases, thus further boosting the speedup and energy savings, Figure 9 and 10 show the speedup and energy of the dynamic scheduling scheme, *DynMap*. All results are normalized to a homogeneous system comprising of the best CMOS architectural configuration for that application.

DynMap outperforms the static scheme across most workloads. The largest improvement in performance and energy savings is seen in *FFT*. *DynMap* causes slight degradation in the performance of *adpcm* and *gsm* at 330K. Both these applications show high ILP as well as sensitivity to frequency. Consequently, migration to TFET, even for a few epochs, degrades performance significantly. *DynMap* obtains improvements of 4%, 22% and 14% over the static scheme at 330K, 340K and 350K respectively. While *DynMap* is more energy efficient than static mapping at lower temperatures, consuming up to 10% lower energy at 340K, the energy for the *DynMap* increases at 350K. This is because the TFET core operates above F_c, at around 1500 MHz . Hence, the energy penalty for migrating to TFET cores is also high.

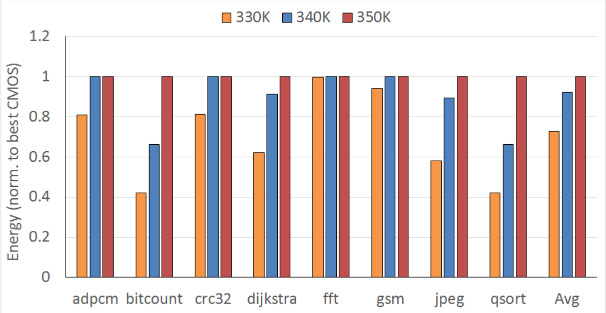

Fig. 8. Normalized energy in heterogeneous multicore with static mapping on best homogeneous CMOS configuration for thermal limits of 330K, 340K, 350K

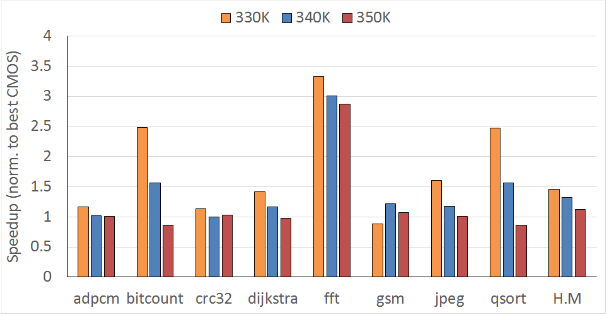

Fig. 9. Speedup on heterogeneous multicore with *DynMap* on best homogeneous CMOS configuration for thermal limits of 330K, 340K, 350K

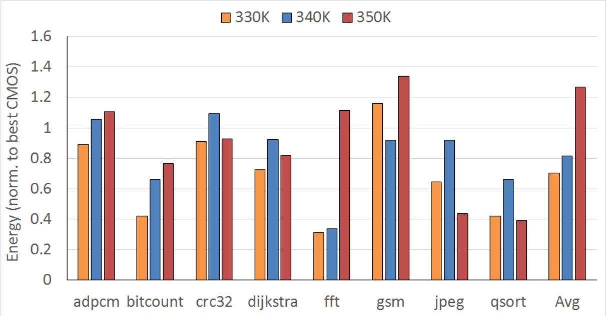

Fig. 10. Normalized energy in heterogeneous multicore with *DynMap* on best homogeneous CMOS configuration for thermal limits of 330K, 340K, 350K

VII. Related Work

Heterogeneous Asymmetric CMP cores have been proposed in [22] which deals with the varying demands from the applications in terms of ILP and TLP. Our work, however, also considers the prominent thermal concerns arising from subsequent generations of process scaling and the move toward increasingly mobile platforms.

Allying TFET cores with traditional CMOS cores has been investigated in [23] and [10]. However, these papers do not look at jointly addressing device and microarchitectural heterogeneity. While works such as [6] examine thermally constrained mapping on 3D stacked device-heterogeneous many-core architectures, the techniques used in that paper are static and do not change state during runtime. Further, the device heterogeneity is restricted to a layer-level granularity, unlike our paper, which considers more fine-grained intra-die CMOS-TFET heterogeneous cores.

Instead of migrating to a whole new process technology to combat the power wall, [24] proposed the concept of *Near Threshold Computing* (NTC), where the supply voltage is held to a near-threshold voltage level. However, in addition to the operational inefficiencies of this method, the reliability in terms of soft error vulnerability is also affected adversely, making TFETs a better choice for operating in this design space [25].

[26] proposes an iterative heuristic static task mapping algorithm formulated as a binary quadratic programming problem to optimize energy consumption. However, this work does not include studies on mapping of energy-aware algorithms on heterogeneous platforms. In [27], the authors propose PROMETHEUS, a framework for thermal-aware scheduling of workloads on a heterogeneous MP-SoCs, to predict the future core temperature given current state. Similarly, [28] proposes a proactive thermal management technique in MP-SoCs. In contrast to our work, these techniques do not consider device or micro-architecture level heterogeneity while scheduling tasks.

VIII. Conclusion

With increased technology scaling, the problem of limited on-chip power and power density has led to inefficient utilization of available hardware. Temperature is a critical constraint that affects packaging costs and lifetimes. These thermal constraints restrict the degree of complexity that can be incorporated into the core at the microarchitectural level, especially in mobile form-factor platforms.

In this context, steep slope devices like Tunnel FETs open up new opportunities for embedded mobile-type applications with strict thermal budgets. We explored the microarchitecture design space to determine the processor configuration with maximum performance and introduced an additional knob, i.e a new transistor technology. We proposed device-heterogeneous multicores with different issue widths and operating frequencies. Using static and dynamic mapping techniques, we demonstrated a peak performance improvement of 47% and energy improvement of 30% on the heterogeneous multicore. Adoption of this heterogeneous technology, especially in the embedded application space, can thus enable us to achieve more thermal and energy efficient designs.

Acknowledgments

This work was supported in part by the Center for Low Energy Systems Technology (LEAST), one of the six SRC STARnet Centers, sponsored by MARCO and DARPA. It was also supported by NSF awards 1205618 and 1213052.

References

[1] T. Skotnicki *et al.*, “The end of CMOS scaling: toward the introduction of new materials and structural changes to improve MOSFET performance,” *IEEE Circuits and Devices Magazine*, 2005.

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

[3] 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.

[4] P. Clarke, “Intel’s Gargini sees tunnel FET as transistor option,” 2011.

[5] P. Greenhalgh, “Big.LITTLE Processing with ARM Cortex-A15 and Cortex-A7 ,” 2011.

[6] K. Swaminathan, H. Liu, J. Sampson, and V. Narayanan, “An examination of the architecture and system-level tradeoffs of employing steep slope devices in 3d cmps,” in *ISCA*, 2014.

[7] V. Alzieu, “Samsung Galaxy S II: Clock Rate and Temperature Differences,” 2012.

[8] W. Huang *et al.*, “Accurate pre-RTL temperature-aware design using a parameterized, geometric thermal model,” in *ISSCC*, 2008.

[9] D. Mohata *et al.*, “Demonstration of MOSFET-like on-current performance in arsenide/antimonide tunnel FETs with staggered hetero-junctions for 300mv logic applications,” in *IEDM*, 2011.

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

[11] L. Liu *et al.*, “Scaling length theory of double-gate interband tunnel field-effect transistors,” *Electron Devices, IEEE Trans.*, 2012.

[12] U. Avci *et al.*, “Understanding the feasibility of scaled III-V tfet for logic by bridging atomistic simulations and experimental results,” in *VLSIT*, 2012.

[13] M. Martin *et al.*, “Multifacet’s general execution-driven multiprocessor simulator (GEMS) toolset,” *SIGARCH Comput. Archit. News*, 2005.

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

[15] N. Choudhary *et al.*, “Fabscalar: composing synthesizable RTL designs of arbitrary cores within a canonical superscalar template.” in *ISCA*. ACM, 2011.

[16] R. Royackers *et al.*, “A new complementary hetero-junction vertical tunnel-fet integration scheme,” in *IEDM*, 2013.

[17] K. Tomioka *et al.*, “Integration of iii-v nanowires on si: From high-performance vertical fet to steep- slope switch,” in *IEDM*, 2013.

[18] B. Fields, R. Bodík, and M. Hill, “Slack: maximizing performance under technological constraints,” in *ISCA*, 2002.

[19] J. A. Brown, L. Porter, and D. M. Tullsen, “Fast thread migration via cache working set prediction.” in *HPCA*, 2011, pp. 193–204.

[20] M. Guthaus *et al.*, “Mibench: A free, commercially representative embedded benchmark suite,” in *WWC*, 2001.

[21] “TCAD Sentaurus Device Manual,” 2010.

[22] R. Kumar *et al.*, “Single-isa heterogeneous multi-core architectures: The potential for processor power reduction,” in *MICRO*, 2003.

[23] K. Swaminathan *et al.*, “Improving energy efficiency of multi-threaded applications using heterogeneous CMOS-TFET multicores,” in *ISLPED*, 2011.

[24] R. Dreslinski *et al.*, “Near-threshold computing: Reclaiming moore’s law through energy efficient integrated circuits,” *Proceedings of the IEEE*, 2010.

[25] H. Liu *et al.*, “Technology assessment of Si and III-V FinFETs and III-V tunnel FETs from soft error rate perspective,” in *IEDM*, 2012.

[26] A. Hussien, A. Eltawil, R. Amin, and J. Martin, “Energy aware task mapping algorithm for heterogeneous mpsoc based architectures,” in *ICCD*, 2011.

[27] S. Sharifi *et al.*, “Prometheus: A proactive method for thermal management of heterogeneous MPSoCs,” *TCAD*, 2013.

[28] A. Coskun *et al.*, “Proactive temperature management in mpsocs,” in *ISLPED*, 2011.