Exploiting Heterogeneity for Energy Efficiency in Chip Multiprocessors

Vinay Saripalli, Guangyu Sun, Asit Mishra, Yuan Xie, Suman Datta and Vijaykrishnan Narayanan

***Abstract*—Heterogeneous multicores are envisioned to be a promising design paradigm to combat today's challenges of power, memory, and reliability walls that are impeding chip design using deep submicron technology. Future multicores are expected to integrate multiple different cores including GPGPUs, custom accelerators and configurable cores. In this paper, we introduce an important dimension - *technology* - using which heterogeneity can be introduced in multicores to improve their energy-performance envelope. Specifically, we analyze the benefits of heterogenous technologies for processor cores and cache subsystems. We discuss two promising device candidates (Tunnel-FET and Magnetic-RAM) for introducing technological diversity in the multicores and analyze their integration in the processor and cache hierarchy in detail. Our analysis shows that introducing such a kind of heterogeneity can significantly enhance the performance and energy behavior of future multicore systems.**

I. Introduction

The traditional frequency-centric processor design philosophy has now yielded to the power-aware multi-core processor technology. All mainstream processor vendors have embraced increasing number of cores in their road maps. In addition to increasing number of cores, many multicore chips integrate various types of compute engines yielding heterogenous multicore systems. Few examples of such single die integrations include Intel's Sandy Bridge architecture [1] and AMD's Fusion architecture [2] where processor cores, GPUs and memory controllers are all integrated on a single die. Going forward, it is widely believed that such heterogeneous integration is key solution to combat the challenges of power, memory and reliability that can impede chip design at nanoscale regimes [3]. Many early results lend credibility to the advantages of heterogeneous systems [4]–[6]. For instance, average execution time for an equal area heterogeneous CMP reduces by 41% compared to a 32-core symmetric CMP in work shown in [7], [8]. Further, using specialized cores can reduce energy consumption for applications by up to 2X compared to traditional cores as shown for few SPEC 2006 application [9]. In addition, heterogeneity can extend to other parts of the multicore systems including the on-chip networks. For example, use of routers of different sizes in a multicore GPU enhances performance by 24.5% [10].

Future architectures can embrace heterogeneity at multi-levels and across multiple subsystems. This paper adds a new dimension to heterogeneous architectures by introducing new forms of technology heterogeneity. Different technologies have intrinsically different performance delay and reliability trade-offs. For example, Tunnel-FET (TFET) [11] [12] based logic is preferable over CMOS based logic from both leakage energy and performance perspective at *sub*-0.5 V operation. However, TFETs have been unable to match the performance of CMOS devices operating at higher voltages. Consequently, applications that either have low throughput requirement (e.g. embedded sensors [13]) or those that exhibit immense parallelism (such as SPEC-OMP [14] and SPLASH [15]) are suited for energy efficient execution on TFET based logic. In contrast, applications that demand high single-threaded performance (such as SPEC 2006 [16] benchmarks) require CMOS cores.

As another example, SRAM memory cells provide very fast read and write accesses but do not achieve densities similar to that of magnetic-RAM (MRAM) cells [17]. Consequently, replacing SRAM caches with MRAM caches can yield larger and less leaky memory sub-system. However, the higher write latencies and energy of MRAM when compared to SRAM results in a design trade-off when considering such replacements. A heterogeneous architecture that can steer writes to SRAMs while retaining most other accesses in MRAMs can capture the best characteristics of the different technologies. The primary goal of our heterogeneous cache design exploration is to match the application characteristic with the best technology feature.

In this work, we show that the use of heterogenous technologies for processor cores and cache subsystems can significantly enhance the performance and energy behavior of the multicore systems. Our results show that (i) CMOS-only cores cannot match the energy efficiency of hybrid TFET-CMOS cores due to the fundamental limitation of supply voltage scaling in CMOS, while TFET-only cores cannot match the performance of hybrid TFET-CMOS cores; (ii) A hybrid cache architecture with MRAM and SRAM can provide 66 % power reduction and 2.7% better performance than the SRAM-only design. Thus, employing the hybrid architecture, the applications' performance and dynamic power consumption are as good as SRAM but the leakage power consumption is similar to that of MRAM.

The rest of the paper is organized as follows. In section II, of types of heterogeneity that can be deployed in processors, memory and on-chip networks. In section III, we demonstrate the utility of a heterogenous chip multiprocessor that contains cores that are optimized for both high and low supply voltage operation. Next, we evaluate the benefits of hybrid caches that combine the best features of multiple memory technologies while masking their drawbacks in section IV. Finally, we conclude in section V.

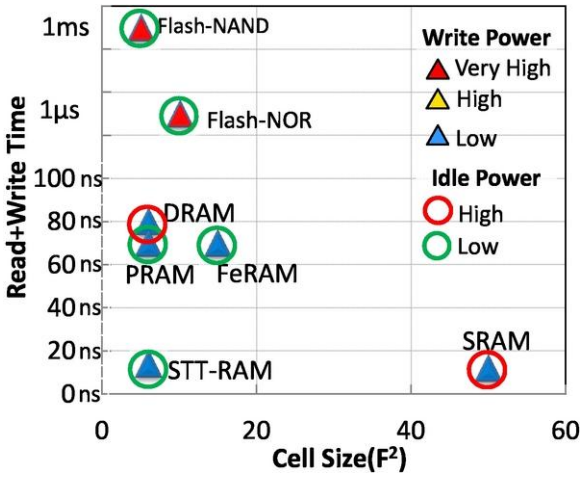

Fig. 1. Comparison of Memory Technologies (Data from [27])

II. Heterogeneous Architectures: An Overview

1) Heterogeneous Processor Cores: It is well-known that different applications/threads and application/thread phases have different characteristics [18]. As a result, a one-size-fits-all approach to designing multicore systems using the same type of cores (in terms of performance, power, and functionality characteristics) is known to be suboptimal in terms of performance and energy-efficiency [4]–[8]. Heterogeneity in processor cores can stem from heterogeneity in architecture design, ISA, frequency of operation or underlying technology used to implement the cores in the multi-processor. For example, small and big cores can be integrated on the same chip with the small cores supporting a subset of the big core's ISA. Li et al. [19] show the benefits of such an approach and demonstrate Operating System support for such a heterogenous system. The cores could also be asymmetric with respect to each other based on the operating frequency or underlying micro-architectural components such as register size, issue queue size, in-order/out-of-order issue and number of floating point units. Aide-De-Camp (ADC) [20] is one such example of processor cores exhibiting structural differences. Other works that demonstrate the utility of asymmetric cores include [21]–[24]. Researchers have also recently examined the special case of putting the CPU and the GPU on the same die [25], [26], as well as having reconfigurable or special purpose logic to augment general-purpose cores [9], [26]. However, none of the previous works on on-chip heterogeneous computing considered implications of heterogeneity in technology and examine how to combine heterogeneous cores with heterogeneous shared resources, or systematically examined what kind of heterogeneity to incorporate in core design. In this work, our first focus is to analyze how heterogeneous technology integrations can help architects to improve the energy envelope of multicores.

2) Heterogeneous Memory: Technology scaling of SRAM and DRAM (conventional memory technologies used in traditional memory hierarchy) is increasingly constrained by fundamental technology limits [28], [29] to mitigate the power and memory walls. Emerging non-volatile memory (NVM) technologies, such as Magnetic RAM (MRAM), Phase-Change RAM (PCRAM), and Resistive RAM (RRAM), taken together, have the features of combining the speed of SRAM, the density of DRAM, and the non-volatility of Flash memory. Emerging memory technologies, shown in Figure 1, enable a large set of options to build a heterogeneous memory hierarchy since individually they present different tradeoffs between power, performance, endurance and density. Hence, a memory hierarchy that employs a heterogeneous mix of both emerging and conventional technologies can not only enable scaling beyond the scaling limits of DRAM, but also provide large improvements in performance and power-efficiency. Unfortunately, some of the technologies have inherent challenges such as limited endurance, high write latency, and low write bandwidth. Thus, hybrid memory hierarchy using different technologies are emerging [30]–[32]

The introduction of the three-dimensional (3D) integration technology [33], [34] provides an opportunity to integrate these heterogeneous technologies and stack them on top of logic cores. Consequently these innovations help alleviate the critical memory bottleneck in CMPs. In this work, we illustrate the benefits of using a hybrid memory stacked on top of a multi-core system using 3D technology.

3) Heterogeneous On-Chip Networks: Network-on-chip (NoC) aid in interconnecting the multiple cores and cache banks in a scalable fashion on the chip and has become a critical shared resource in the emerging Chip Multiprocessor (CMP) era. Heterogeneity can occur at multiple levels in such networks. For instance, researchers have proposed multiple topologies [35] for tiled multicore architecture where each individual network is customized for a particular message type. Heterogeneity is possible in the resources allocated for the on-chip routers and links of this scalable communication backbone. Heterogeneous routers [36]–[38] have been proposed for application-specific architectures where individual resource in the network (e.g. buffer, link, crossbar, etc) is customized to the application hosted in the system). Heterogeneity in router operating frequencies was proposed in [39] for not only managing power but also dynamically managing congestion. Finally, heterogenous technologies have been used for enhancing the performance on on-chip networks. A combined RF-electric interconnect fabric is one of the early demonstrators in this domain [40], [41].

III. Technology Oriented Core Heterogeneity

As transistor scaling continues, some of the circuit-level implications of using MOSFETs with a 60 mV/decade sub-threshold slope become increasingly clear. The equations for the I_{On} and the I_{Off} of a nanoscale MOSFET are shown in eqs (1)- (2) [42].

$$I_{On}(V_{ds}) = C_{ox}.v_{sat}.(V_G - V_T) \quad (1)$$

$$I_{Off}(V_{ds}) = \mu_{eff} C_{ox} \frac{W}{L}(m-1).(\frac{kT}{q})^2. e^{-qV_{TH}/mkT}.(1 - e^{qV_{ds}/kT}) \quad (2)$$

The threshold voltage (V_T) is not scalable (by factor $\sqrt{2}$) due to the exponential dependence of I_{Off} on the threshold voltage, and has been kept nearly constant for the past few

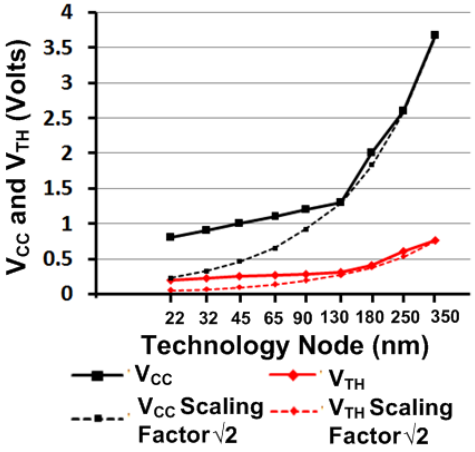

Fig. 2. Scaling of V_{CC} and V_{TH} with technology node.

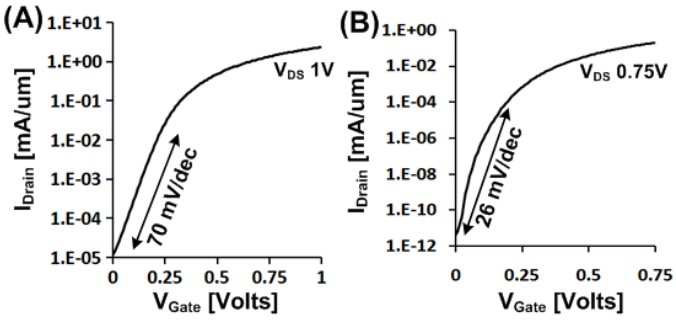

Fig. 3. Sub-60 mV/dec threshold slope of a Tunnel-FET.

technology generations as shown in Figure 2. In order to deliver a high I_{On}, a reasonable overdrive $(V_G - V_T)$ is required, and consequently, V_{CC} scaling has also slowed down, while only being scaled nominally due to reliability and power concerns.

Recently, novel Inter-band Tunneling Field Effect Transistors (TFETs) have been experimentally demonstrated with the potential to show sub-60 mV/decade sub-threshold slope [11], [43] (see Figure 3 for the steeper sub-threshold slope of TFETs). In the remainder of this section, we use TFET devices to achieve energy efficient operation at low V_{CC}, where the energy-delay trade-off diminishes for CMOS based circuits. At low supply voltages it is possible to take advantage of the steep sub-threshold slope to deliver higher Ion, while maintaining a good Ion/Ioff ratio. We take advantage of this characteristic of TFETs to propose a heterogeneous embedded processor architecture composed of CMOS and TFETs based processors whose Voltage-Frequency characteristics are shown in Figure 4. It can be observed that the TFET can achieve better performance than CMOS in the *sub*-0.5 V region. Consequently, it is possible to achieve better energy-efficiencies for a desired performance in this region using TFETs.

A. TFET Device Operation

We present a brief overview of TFET based device technology and refer the reader to [44] for a comprehensive review.

Off Condition: In a n-channel MOSFET, the population of electrons near the (N++) source - (P+) channel junction of the NMOS (Figure 5A) is given by the Fermi-Dirac distribution,

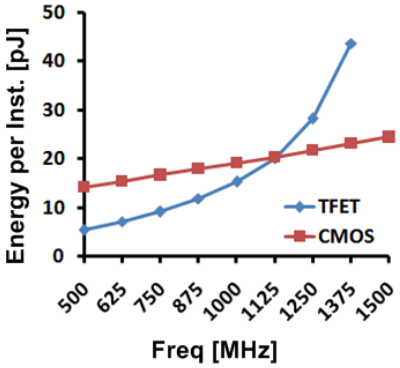

Fig. 4. Voltage-Frequency Characteristics of CMOS and TFET-based processor.

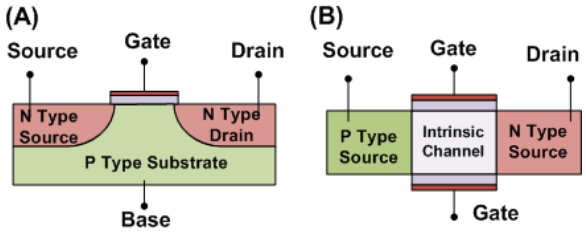

Fig. 5. Comparison of NMOSFET and NTFET.

$f(E) = (e^{\frac{E-E_F}{kT}} + 1)^{-1}$. The off-state conduction in a NMOS is caused by the diffusion of thermionically excited electrons across the P-N junction, which results in a sub-threshold conduction slope $>$ 60 mV/decade. In contrast, the source region of a n-channel TFET is P++ doped (Figure 5B), and the Fermi level in the source is a few kT *below* the Valence band-edge. As a result, the population of thermionically excited electrons near the (P++) source - (Intrinsic) channel junction of the NTFET, and *above* the Fermi level are filtered away by the Valence band-edge. Thus, TETs are able to demonstrate a $<$ 60 mV/Dec sub-threshold slope through band-edge filtering of the tail of the Fermi-Dirac distribution.

On Condition: When a sufficiently large positive gate voltage is applied to a NMOS, an inversion layer forms in the channel near the the oxide-semiconductor interface, leading to the conduction of electrons from source to drain, through electron drift. Applying a positive gate voltage to the NTFET also results in the formation of a n-type inversion layer in the channel. However, since the source region of the NTFET is degenerately P++ doped, the source-channel junction becomes a strongly reverse biased P-N junction thereby causing inter-band tunneling of electrons in the Valence-Band of the source, across the P-N depletion barrier, into the Conduction-Band of the channel.

B. Experimental Benchmarking of TFETs

The TFET characteristics are simulated using the device simulator TCAD Sentaurus [45], by using the non-local tunneling model for modeling interband tunneling. Figure 6 shows a good match between the experimental characteristics of an $In_{0.53}Ga_{0.47}As$ homojunction TFET from [11], and simulations using TCAD Sentaurus. The parameters used for simulating the single-gate Homojunction TFET are given in

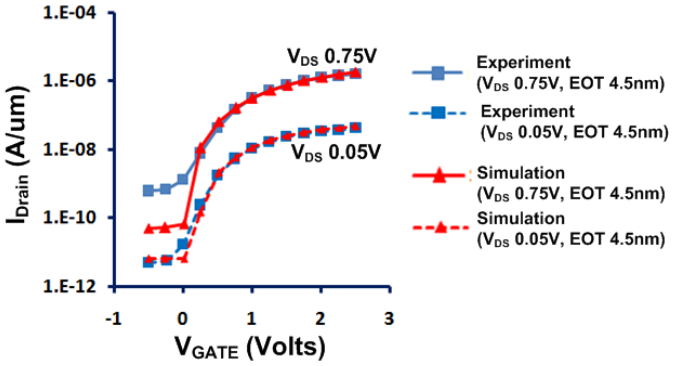

Fig. 6. Comparison of experimental and simulated characteristics of single-gate $In_{0.53}Ga_{0.47}As$ Homojunction TFET (EOT 4.5nm) [11].

Table I.Thus, the TCAD model of a TFET can be used to generate the transfer characteristics of a TFET over a wide range of voltages, which can then be used for simulating TFET-based circuits.

Gate Length,L_G	100 nm
Dielectric constant, ϵ_{ox}	9
Oxide Thickness, t_{ox} (EOT)	9nm (4.5 nm)
$In_{0.53}Ga_{0.47}As$ Bandgap, E_G	0.74 eV
Effective tunneling masses, m_c,m_v	0.05, 0.07
Tunneling Prefactors,m_c,m_v	0.07, 0.07

TABLE I
$In_{0.53}Ga_{0.47}As$ SINGLE-GATE HOMOJUNCTION TFET SIMULATION PARAMETERS

C. Heterojunction Tunnel FETs

We consider a GaSb/InAs Heterojunction Tunnel-FET (HTFET), and use the modeling technique described in Section III-B in order 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 7. By using the Heterojunction Tunnel-FET, a higher Ion can be obtained because of the higher critical-field strength provided by the staggered P-N heterojunction.

In order to understand the circuit level implications of using HTFETs, we compare the Ion versus Ion/Ioff 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 8. Fig 8A shows that at V_{CC} 0.8V, the highest Ion and Ion/Ioff ratio are provided by 22nm NMOS, making it the preferred device for operation at High V_{CC}. However, at V_{CC} 0.3V, the NMOS device cannot give both a good Ion as well as a good Ion/Ioff because of the 60 mV/Dec limit on the sub-threshold slope. The $In_{0.53}Ga_{0.47}As$ Homojunction TFET can

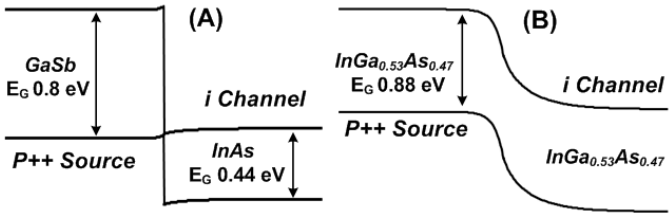

Fig. 7. Comparison of Heterojunction and Homojunction TFET (Band-Gap includes quantization effect due to Double-Gate structure with 7nm T_{Body})

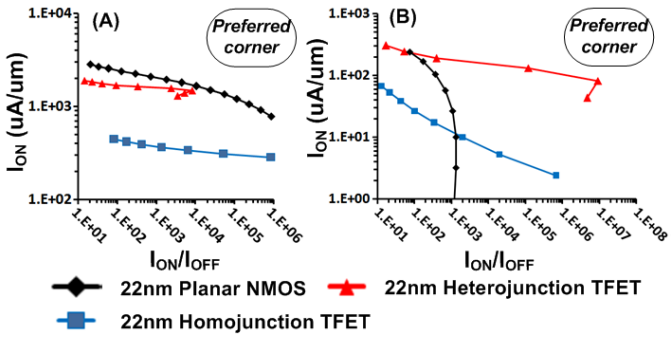

Fig. 8. Comparison of Ion versus Ion/Ioff ratio for different operating points on the $I_D - V_G$ for (A) a V_{CC} window of 0.8V and (B) a V_{CC} window of 0.3V.

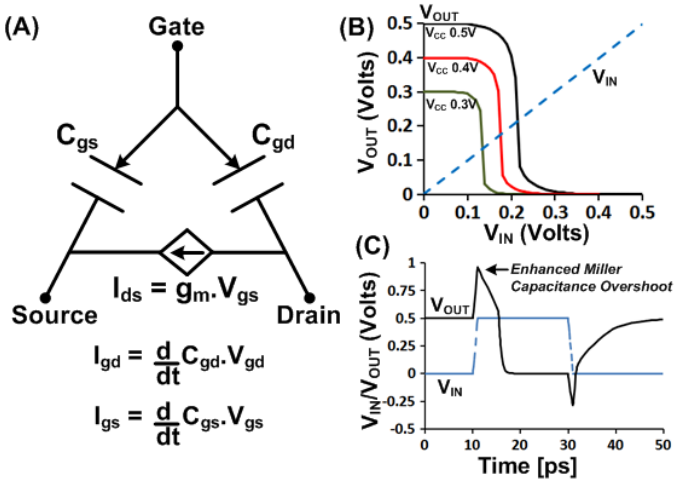

Fig. 9. Verilog-A small signal model used for Tunnel FET simulation.

provide a good Ion/Ioff but cannot provide a high Ion because the homojunction does not allow a strong tunneling current. In contrast, the Heterojunction TFET can provide a good Ion, as well as a good Ion/Ioff, due to the sub-60 mV/Dec sub-threshold slope, making it the preferred device for operation at low V_{CC}.

D. Tunnel FET Logic and Memory

1) Verilog-A Model: 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 simulation. 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 9A shows the Verilog-A small-signal model for Tunnel FETs, which uses the look-up tables in order to do circuit simulation. Fig 9A shows the Voltage Transfer Characteristics (VTC) and Fig 9B shows the transient output characteristic of a Heterojunction TFET inverter (V_{CC} 0.5V), which shows the validity of the Verilog-A lookup table based method.

2) TFET Logic: The energy-delay performance curve of a HTFET AND gate, shown in comparison to that of a CMOS AND gate in Fig 10A, shows a cross-over in the energy-delay characteristics. The CMOS AND gate has a better energy-delay tradeoff compared to the HTFET AND gate at

$V_{CC} > 0.5V$ and the HTFET AND gate has a better energy-delay trade-off at $V_{CC} < 0.5V$. 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 III-C 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 10B shows the energy-delay performance of a 32-bit prefix-tree based Hans-Carlson Adder. The Energy-Delay of this Adder was computed analytically using the Energy-Delay estimates for the gates, due to the excessive computational cost of simulating a 32-bit Adder using a look-up table. The delay of the Adder was computed using critical-path analysis of the Hans-Carlson Adder as described in [46]. A similar crossover is observed for the 32-bit Adder implemented using CMOS and HTFETs, where the CMOS-based 32-bit Adder has a favorable Energy-Delay trade-off at $V_{CC} > 0.5V$ and the HTFET-based 32-bit Adder has a better Energy-Delay trade-off at $V_{CC} < 0.5V$. Moreover, the energy consumption of a 32-bit Adder at low-activity and low V_{CC} is dominated by leakage energy, and does not show an energy-delay tradeoff below 0.4V. In contrast, the HTFET-based 32-bit Adder shows continued energy reduction with supply-voltage scaling upto 0.2V because of its sub-60mV/Dec sub-threshold slope.

3) TFET Cache: We modified CACTI to implement the 6T TFET SRAM cell design proposed in [47] 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 11 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 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.

E. CMOS-HTFET Heterogeneous Multi-Core Processors

The detailed processor and cache parameters for simulating single-core processors using Simics [48] are shown in Table II. 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% [49] and

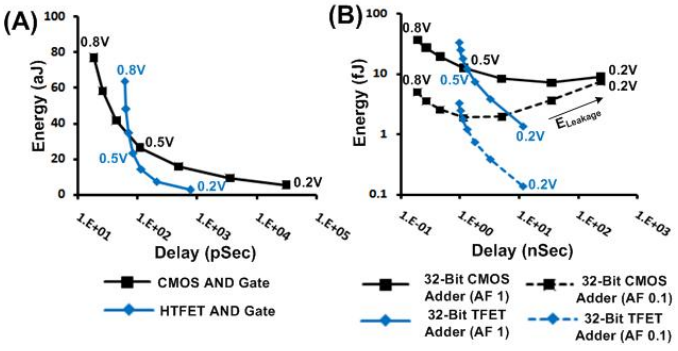

Fig. 10. Energy-Delay performance comparision for (A) a CMOS And-Gate and a HTFET And-Gate and (B) a CMOS 32-bit Adder and a TFET 32-bit Adder

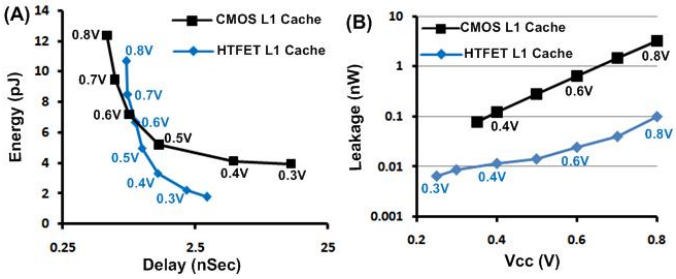

Fig. 11. (A) Energy-Delay performance comparison and (B) Leakage Power comparison for CMOS and H-TFET based L1 Cache.

the variations across instructions in commercial low power cores are minimal [50]). The delay and power numbers for each voltage/frequency pair for both TFET and CMOS gates are obtained using circuit simulations and incorporated into our simulator. We evaluate multi-threaded SPEC-OMP [14] and SPLASH [15] benchmark applications. Leakage power is consumed during both execution and stall cycles and no power-gating is assumed. The cache power models are based on modifications to CACTI as indicated earlier.

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 (4-Core@2GHz) as shown in Figure 12. Figure 13 shows the energy consumption compared to baseline CMOS for parallel program execution on 8-Core CMOS and 8-Core TFET, for *iso-performance* to baseline CMOS. When moving from 4 to 8 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 4-core CMOS and energy savings of 45% against 8-core CMOS.

While multiple TFET cores can provide energy efficiencies that are not possible using CMOS cores, some applications may still need to be executed on CMOS cores if single-threaded performance is critical. Further, in applications that do not meet their performance requirements when operating at low-voltages will also require CMOS cores. Consequently we believe a hybrid multi-core processor with both TFET and CMOS cores is desirable. Efficient scheduling techniques to bind the tasks to the appropriate cores and the degree of threading for each task is an open research direction that we are pursuing. Our hybrid architecture can provide additional energy efficiencies for multi-threaded applications by scheduling performance critical threads [24] on high performance CMOS cores and non-critical threads on energy efficient TFET cores. They can exploit imbalance across threads due to application behavior [51].

IV. Heterogeneous Memory

Magnetic-RAM (MRAM) combines the speed of SRAM, the density of DRAM, and the non-volatility of Flash memory, with excellent scalability. Furthermore, it has been shown that

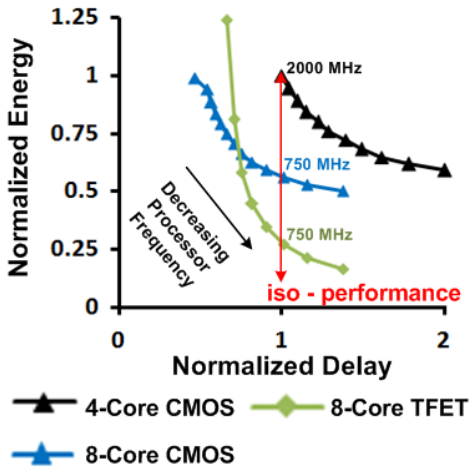

Fig. 12. Illustration of normalized energy-delay for of iso-performance for LU Benchmark application.

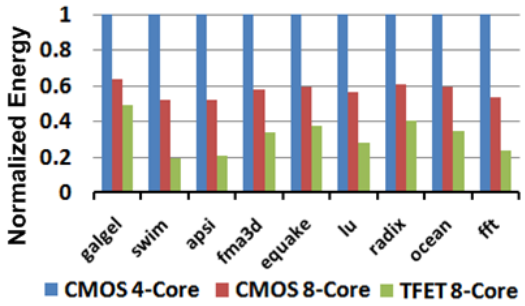

Fig. 13. Normalized multi-core execution energy for iso-peformance to CMOS @ 2 GHz.

with 3D stacking MRAM can be integrated with conventional CMOS logic [52], [53]. Thus, MRAM is potentially attractive to replace the traditional on-chip SRAM [31], [52], with benefits such as higher density and lower leakage compared to traditional SRAM-based cache architecture. Even though MRAM based cache architecture has many advantages, it suffers from a longer write latency and higher write energy consumption compared to SRAM. In this section, we show a hybrid memory architecture that combines the benefits of SRAM and MRAM technologies while masking the deficiencies of each of these technologies.

A. Mageneto-resistive RAM (MRAM) Overview

The basic difference between the MRAM and the conventional RAM technologies (such as SRAM/DRAM) is that the information carrier of MRAM is Magnetic Tunnel Junctions (MTJs) instead of electric charges. As shown in Figure 14, each MTJ contains a pinned layer and a free layer. The pinned layer has fixed magnetic direction while the free layer can change its magnetic direction by spin torque transfers [17]. If the free layer has the same direction as the pinned layer, the MTJ resistance is low and indicates state 0; otherwise, the MTJ resistance is high and indicates state 1.

The latest MRAM technology (spin-torque transfer RAM (STT-RAM)) changes the magnetic direction of the free layer by directly passing spin-polarized currents through MTJs. Comparing to the previous generation of MRAM using external magnetic fields to reverse the MTJ status, STT-RAM has the advantage of scalability, which means the threshold current to make the status reversal will decrease as the size of the MTJ becomes smaller. In this paper, we use the terms MRAM and STT-RAM equivalently.

Processors:	
# of cores	4 and 8
Frequency	2000MHz - 500MHz (125 MHz Steps)
Issue Width	1 (in order)
Memory:	
L1 cache	private, 32+32KB, 2-way, 64B line, write-through, 1 read/write port
SRAM L2	shared 2MB, 8-way, 64B line, write-back, 1 read/write port 5ns access delay
Main Memory	4 GB, 100-cycle latency @ 2GHz

TABLE II
CONFIGURATION PARAMETERS FOR HETEROGENEOUS MULTI-CORE STUDY

	Read Energy (nJ)	Write Energy (nJ)	Leakage Power (mW)	Read @ 3GHz (cycles)	Write @ 3 GHz (cycles)
SRAM	0.62	0.62	1.65	10	10
MRAM	0.76	5	0.23	10	33

TABLE III
COMPARISON OF THE MRAM AND SRAM TECHNOLOGIES.

The most popular structure of MRAM cells is composed of one NMOS transistor as the access device and one MTJ as the storage element (1T-1MTJ structure). As illustrated in Figure 14, the storage element, MTJ, is connected in series with the NMOS transistor. The NMOS transistor is controlled by the the word line (WL) signal. The detailed read and write operations for each MRAM cell is described as follows:

- Read Operation: When a read operation happens, the NMOS is turned on and a small voltage difference (-0.1V) is applied between the bit line (BL) and the source line (SL). This voltage difference causes a current through the MTJ whose value is determined by the status of MTJs. A sense amplifier compares this current to a reference current and then decides whether a 0 or a 1 is stored in the selected MRAM cell.
- Write Operation: When a write operation happens, a large positive voltage difference is established between SLs and BLs for writing for 0 or a large negative one for writing 1. The current amplitude required to ensure a successful status reversal is called threshold current. The current is related to the material of the tunnel barrier layer, the writing pulse duration, and the MTJ geometry [52], [54].

In this work, we use the writing pulse duration of 10ns [55], below which the writing threshold current will increase exponential. In addition, we scale the MRAM size of previous work [17] down to 65 nm technology node. Assuming the size of MTJs is 65nm X 90nm, the derived threshold current for magnetic reversal is about 195 μA.

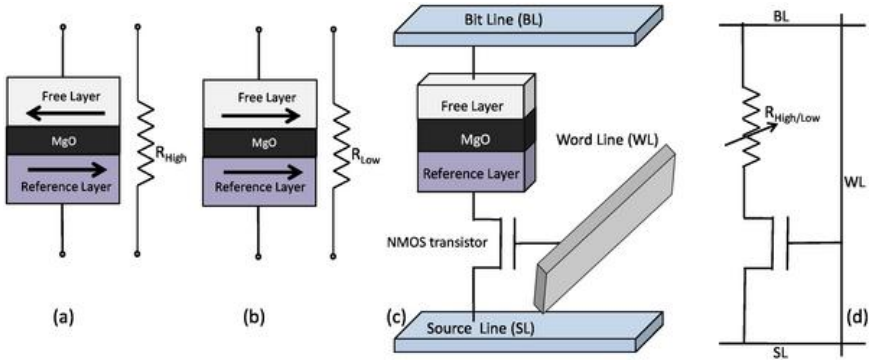

Fig. 14. MTJ and MRAM cell - (a) Anti-parallel (high resistance) indicating '1' state (b) Parallel (low resistance) indicating '0' state (c) MRAM structural view (d) MRAM schematic.

B. Hybrid SRAM-MRAM Architecture

Our baseline configuration for this study is an 8-core in-order processor using the Ultra SparcIII ISA. In order to predict the chip area, we investigate some die photos, such as Cell Processor [56], Sun UltraSPARC T1 [57], etc. and estimate the area of an 8-core CMP without caches to be 60 mm^2. By using a modified version of CACTI (details are described in [52], we further learn that one cache layer fits to either a 2MB SRAM or an 8MB MRAM L2 cache assuming each cache layer has the similar area to that of core layer (60mm2). The configurations are detailed in Table IV. Note that the power of processors is estimated based on the data sheet of real designs [56], [57]. We use the Simics toolset [48] for performance simulations. Our 3D NUCA architecture is implemented as an extended module in Simics. We use a few multi-threaded benchmarks from SPEC-OMP [14] and SPLASH [15].

Processors:	
# of cores	8
Frequency	3GHz
Power	6W/core
Issue Width	1 (in order)
Memory:	
L1 cache	private, 32+32KB, 2-way, 64B line, 2-cycle, write-through, 1 read/write port
SRAM L2	shared 2MB, 16-way, 64B line, read/write per bank: 10-cycle, write-back, 1 read/write port
MRAM L2	shared, 8MB, 16-way, 64B line, read penalty per bank : 10-cycle, write penalty per bank : 33-cycle, write-back, 1 read/write port
Main Memory	4 GB, 300-cycle latency

TABLE IV
CONFIGURATION PARAMETERS FOR HYBRID MRAM-SRAM STUDY

Since the performance and power of MRAM caches are closely related to transaction intensity, we select some simulation workloads as listed in Table V so that we have a wide range of transaction intensities to L2 caches. The average numbers of total transactions per kilo-instructions (TPKI) and write transactions per kilo-instructions (WPKI) of L2 caches are listed in Table V. The L2 caches utilize dynamic-NUCA (DNUCA) that dynamically migrates frequently accessed blocks to the closest banks [58]. For each simulation, we fast forward to warm up the caches and then run 3 billion cycles. We use the total IPC of all the cores as the performance metric.

Name	TPKI	WPKI	Name	TPKI	WPKI
galgel	1.01	0.31	lu	54	30
apsi	4.15	1.85	fft	78	64
equake	7.94	3.84	ocean	80	58
fma3d	8.43	4.00	radix	98	90
swim	19.29	9.76			

TABLE V
L2 TRANSACTION INTENSITIES

1) SRAM-MRAM Hybrid L2 Cache: In the hybrid cache implementation each cache set has a majority of MRAM cache ways and a minority of SRAM ways. The primary motivation is to keep as many write intensive data in the SRAM ways as possible and hence reduce the number of write operations to the MRAM. Therefore, we design an SRAM-MRAM hybrid L2 cache with 12 ways of MRAM and 1 way of SRAM (*12M1S*), in order to ensure area equivalence. After having these hybrid cache lines, the second step is to distribute MRAM cache lines and SRAM ones into separate cache banks. Considering the SRAM part is the minority in the proposed 12M1S cache, one partitioning alternative is to distribute these SRAM cache lines into different banks so that there are several SRAM cache lines close to each processing core. However, this method requires each cache bank to be a heterogenous memory array with SRAM and MRAM cells and increases the complexity of the cache design. In addition, this distributed partitioning of SRAM cells implies that the SRAM and MRAM cells have to be fabricated together. Considering the specialization of the MRAM fabrication process, this method also eliminates the cost advantages of stacking MRAMs on top of processing cores. Therefore, we use another alternative that, we reduce the number of cache lines in some MRAM cache banks compared to the pure MRAM cache structure (as shown in Figure 15 (a) that the MRAM banks at

four corners are smaller than other MRAM banks), compensate this cache line loss with SRAM ones, and collect all the SRAM cache lines together to build several entire SRAM banks on the core layer. As shown in Figure 15 (a), SRAM cache banks are placed in the center of the core layer instead of being distributed. In this method, SRAM and MRAM cache banks have no difference from the architectural point of view. Note that after placing one way of SRAM cache lines in the core layer, the area of the core layer will increase and the area of the cache layer will decrease. In this work, the total size of all the SRAM cache lines is 128KB, the derived area overhead is about 6.25%.

Hybrid Cache Management Policy: Another important issue is how to manage the hybrid L2 cache to improve the performance and reduce the power. Because the key point is to reduce the number of write operations to MRAM cache cells, we need to move as many write intensive data in SRAM cache banks as possible. The management policy of the hybrid cache can be described as follows:

- The cache controller is aware of the locations of SRAM cache ways and MRAM cache ways. When there is a write miss, the cache controller first try to place the data in the SRAM cache ways.
- Considering the high probability that a core write data to a specific group of cache lines repeatedly, data in MRAM caches should be migrated to SRAM caches if the some cache lines are frequently written to. In this work, data in MRAM caches will be migrated to SRAM caches when they are accessed by two successive write operations. This kind of data migration is named intra-migration to differentiate inter-migration policy introduced in Section 3. Due to the existence of this intra-migration policy, the number of write accesses from cores to MRAM caches can be reduced.
- Note that read operations from cores are also possible to cause data migrations, the number of which could be even larger than that of direct write accesses from cores. Therefore, a new type inter-migration policy is introduced. Figure 15 (b) and (c) compare the banks from which data can be migrated toward the core in upper left corner. Figure 15 (b) shows that, in original intermigration policy, the cache layer is divided into 4 uniform groups and there is only one core associative with each part. In this work, banks in each group are named as the host banks of their corresponding core. Data can only be migrated from non-host banks. For the traditional management policy, the data will be migrated to host bank. For the management policy proposed for the hybrid cache, the data can only be migrated to SRAM banks.

Two data migrations are illustrated in Figure 15 (b) for the traditional inter-migration. When using the hybrid SRAM-MRAM cache, the host banks for a core is redefined as shown in Figure 15 (c). Two corresponding data migrations are also shown in Figure 15 (c). Using this policy, there is no data migration between two MRAM cache lines, which reduces the number of write operations greatly. The drawback is that

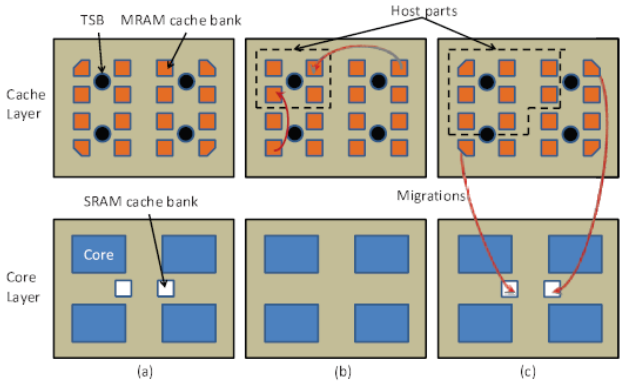

Fig. 15. SRAM-MRAM hybrid cache implementation (a) one placement method of SRAM and MRAM cache banks,(b) data migrations in original MRAM caches, (c)data migrations in hybrid SRAM-MRAM caches

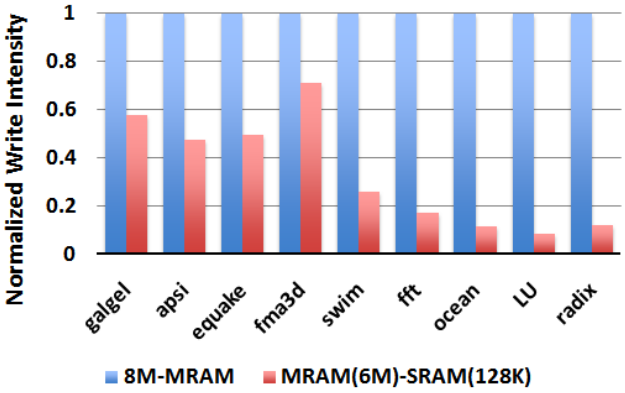

Fig. 16. The MRAM write intensity to MRAM before and after using hybrid SRAM- MRAM caches.

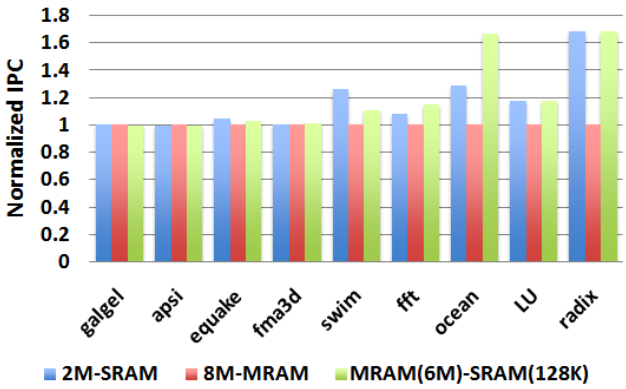

Fig. 17. The comparison of IPC among 2M SRAM cache, 8M MRAM pure cache, and 8M SRAM-MRAM hybrid cache (Normalized by the IPC of 8M MRAM pure cache)

SRAM banks are shared by all cores so that their limited sizes may increase L2 miss rates. Considering we have 8M of total cache size, which is considerably large for most applications, our simulation results show that the increase of L2 miss rates is very small.

Figure 16 shows the number of MRAM write operations per 1K instructions is reduced dramatically by using our hybrid SRAM-MRAM approach. As a result, the dynamic power associated with write operations to MRAM cells is also reduced and the performance is improved. Figure 17

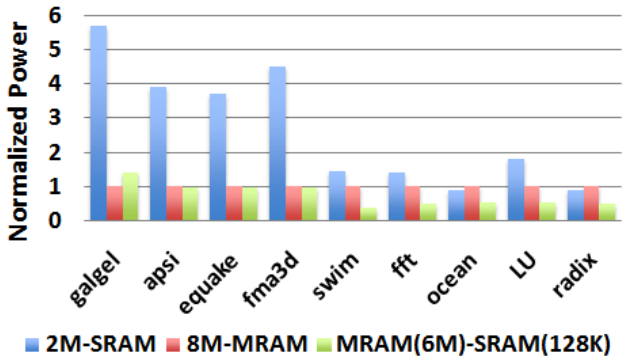

Fig. 18. The comparison of total power consumption among 2M SRAM cache, 8M MRAM pure cache, and 8M SRAM-MRAM hybrid cache (Normalized by the IPC of 8M MRAM pure cache)

shows the performance comparison. On average, the hybrid cache structure improves the performance by 2.7% compared to their SRAM counterparts, with a maximum performance improvment of 29% for *ocean* benchmark. The performance improvement is maximized for the *ocean* benchmark because of significant reduction in write intensity and reduced L2 cache misses due to the increased L2 cache capacity of the hybrid MRAM cache.

Fig. 18 shows the power comparison. We observe that the total power of the hybrid scheme is reduced compared to the MRAM-only cache, except for *galgel*. It is because both read and write intensities in *galgel* are so small that the dynamic power is very low. Consequently, the introduction of SRAM cache lines in the hybrid cache brings the leakage power back and eliminates the dynamic power reduction achieved by the hybrid structure. However, as the write intensity increases, the MRAM-SRAM hybrid cache can lower the total power consumption. For example, the total power consumption is cut by more than half compared to the MRAM-only cache, for the applications such as *fft, ocean, lu, radix* and *swim*. On average, after the transition from SRAM caches to the newly-proposed MRAM-hybrid cache, the total power consumption is reduced by 66% compared to the SRAM-only cache, and by 25% compared to the MRAM-only cache.

V. Conclusion

Heterogeneous integration of various functional/compute engines and multiple cores is expected to shape the computer architecture landscape going forward in the nanometer regime. Many recent works in this domain have looked into the design and architecture of such heterogeneous multicore systems to enhance their power-performance envelope. In this paper, we investigate an important dimension - *technology* - using which heterogeneity can be introduced in multicores to further improve this envelope. We discuss the benefits of integrating two new device candidates (Tunnel-FET and Magnetic-RAM) with traditional CMOS devices. Specifically, we observe that the TFETs can achieve better performance than CMOS in the *sub*-0.5 V region and MRAM has significantly lower standby power consumption than CMOS. However, TFETs are not competitive at higher voltages and MRAMs suffer from longer write-latencies as compared to CMOS. To combat these demerits our analysis shows that architectures that introduce technology heterogeneity can accentuate the desired features of a technology while utilizing another technology to mask its drawbacks. We believe, there is significant room for further research on studying interactions of heterogeneity of different types at multiple scales.

References

[1] "Intel sandy bridge information," http://software.intel.com/en-us/articles/sandy-bridge/.
[2] "Amd fusion apu," http://sites.amd.com/us/fusion/apu/Pages/fusion.aspx.
[3] S. Borkar, N. P. Jouppi, and P. Stenstrom, "Microprocessors in the era of terascale integration," in *Proceedings of the conference on Design, automation and test in Europe*, ser. DATE '07, 2007.
[4] M. Annavaram, E. Grochowski, and J. Shen, "Mitigating amdahl's law through epi throttling," *SIGARCH Comput. Archit. News*, vol. 33, pp. 298–309, May 2005.
[5] S. Balakrishnan, R. Rajwar, M. Upton, and K. Lai, "The impact of performance asymmetry in emerging multicore architectures," in *Proceedings of the 32nd annual international symposium on Computer Architecture*, ser. ISCA '05, 2005, pp. 506–517.
[6] M. D. Hill and M. R. Marty, "Amdahls law in the multicore era," *IEEE COMPUTER*, 2008.
[7] M. A. Suleman, O. Mutlu, M. K. Qureshi, and Y. N. Patt, "Accelerating critical section execution with asymmetric multi-core architectures," in *Proceeding of the 14th international conference on Architectural support for programming languages and operating systems*, ser. ASPLOS '09, 2009, pp. 253–264.
[8] M. A. Suleman, O. Mutlu, J. A. Joao, Khubaib, and Y. N. Patt, "Data marshaling for multi-core architectures," in *Proceedings of the 37th annual international symposium on Computer architecture*, ser. ISCA '10, 2010, pp. 441–450.
[9] G. Venkatesh, J. Sampson, N. Goulding, S. Garcia, V. Bryksin, J. Lugo-Martinez, S. Swanson, and M. B. Taylor, "Conservation cores: reducing the energy of mature computations," in *Proceedings of the fifteenth edition of ASPLOS on Architectural support for programming languages and operating systems*, ser. ASPLOS '10, 2010, pp. 205–218.
[10] A. Bakhoda, J. Kim, and T. M. Aamodt, "Throughput-effective on-chip networks for manycore accelerators," in *International Symposim on Microarchitecture (MICRO)*, 2010.
[11] 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.
[12] N. N. Mojumder and K. Roy, "Band-to-band tunneling ballistic nanowire fet: Circuit-compatible device modeling and design of ultra-low-power digital circuits and memories," *IEEE Transactions On Electron Devices*, vol. 56, pp. 2193–2201, 2009.
[13] "Open mote platform (http://www.cs.berkeley.edu/ prabal/projects/epic/)."
[14] "Spec omp benchmark suite (http://www.spec.org/omp/)."
[15] "Splash-2 benchmark suite (http://www.capsl.udel.edu/splash/)."
[16] "Spec cpu2006 (http://www.spec.org/cpu2006/)."
[17] M. Hosomi, H. Yamagishi, T. Yamamoto, K. Bessho, Y. Higo, K. Yamane, H. Yamada, M. Shoji, H. Hachino, C. Fukumoto, H. Nagao, and H. Kano, "A novel nonvolatile memory with spin torque transfer magnetization switching: spin-ram," in *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*, Dec. 2005, pp. 459 –462.
[18] P. J. Denning, "Working sets past and present," *IEEE Trans. Softw. Eng.*, vol. 6, pp. 64–84, January 1980.
[19] T. Li, P. Brett, R. Knauerhase, D. Koufaty, D. Reddy, and S. Hahn, "Operating system support for overlapping-isa heterogeneous multi-core architectures," in *High Performance Computer Architecture (HPCA), 2010 IEEE 16th International Symposium on*, Jan. 2010, pp. 1–12.
[20] S. Ghiasi, "Aide de camp: Asymmetric multi-core design for dynamic thermal management," Ph.D. dissertation, Boulder, CO, USA, 2004.
[21] R. Kumar, D. M. Tullsen, N. P. Jouppi, and P. Ranganathan, "Heterogeneous chip multiprocessors," *Computer*, vol. 38, no. 11, 2005.
[22] R. Kumar, D. M. Tullsen, P. Ranganathan, N. P. Jouppi, and K. I. Farkas, "Single-isa heterogeneous multi-core architectures for multithreaded workload performance," *SIGARCH Comput. Archit. News*, vol. 32, no. 2, 2004.

[23] T. Y. Morad, U. C. Weiser, A. Kolodny, M. Valero, and E. Ayguade, "Performance, power efficiency and scalability of asymmetric cluster chip multiprocessors," *IEEE Comput. Archit. Lett.*, 2006.

[24] 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.

[25] E. S. Chung, P. A. Milder, J. C. Hoe, and K. Mai, "Single-chip heterogeneous computing: Does the future include custom logic, fpgas, and gpus?" in *International Symposim on Microarchitecture (MICRO)*, 2010.

[26] M. A. Watkins and D. H. Albonesi, "Remap: A reconfigurable heterogeneous multicore architecture," in *International Symposim on Microarchitecture (MICRO)*, 2010.

[27] S. A. Wolf, J. Lu, M. R. Stan, E. Chen, and D. M. Treger, "The promise of nanomagnetics and spintronics for future logic and universal memory," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2256 –2320, dec. 2010.

[28] R. S., G. W. Burr, M. J. Breitwisch, C. T. Rettner, Y.-C. Chen, R. M. Shelby, M. Salinga, D. Krebs, S.-H. Chen, H.-L. Lung, and C. H. Lam, "Phase-change random access memory: a scalable technology," *IBM J. Res. Dev.*, vol. 52, pp. 465–479, July 2008.

[29] International Technology Roadmap for Semiconductors, "Process Integration, Devices, and Structures 2007 Edition," http://www.itrs.net/.

[30] Y. Park, S.-H. Lim, C. Lee, and K. H. Park, "Pffs: a scalable flash memory file system for the hybrid architecture of phase-change ram and nand flash," in *Proceedings of the 2008 ACM symposium on Applied computing*, ser. SAC '08, 2008, pp. 1498–1503.

[31] X. Wu, J. Li, L. Zhang, E. Speight, R. Rajamony, and Y. Xie, "Hybrid cache architecture with disparate memory technologies," in *Proceedings of the 36th annual international symposium on Computer architecture*, ser. ISCA '09, 2009, pp. 34–45.

[32] P. Mangalagiri, K. Sarpatwari, A. Yanamandra, V. Narayanan, Y. Xie, M. J. Irwin, and O. A. Karim, "A low-power phase change memory based hybrid cache architecture," in *Proceedings of the 18th ACM Great Lakes symposium on VLSI*, ser. GLSVLSI '08, 2008, pp. 395–398.

[33] B. Black, M. Annavaram, N. Brekelbaum, J. DeVale, L. Jiang, G. H. Loh, D. McCaule, P. Morrow, D. W. Nelson, D. Pantuso, P. Reed, J. Rupley, S. Shankar, J. Shen, and C. Webb, "Die stacking (3d) microarchitecture," in *Proceedings of the 39th Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO 39, 2006, pp. 469–479.

[34] Y. Xie, G. H. Loh, B. Black, and K. Bernstein, "Design space exploration for 3d architectures," *J. Emerg. Technol. Comput. Syst.*, vol. 2, pp. 65–103, 2006.

[35] J. Balfour and W. J. Dally, "Design tradeoffs for tiled cmp on-chip networks," in *Proceedings of the 20th annual international conference on Supercomputing*, ser. ICS '06, 2006, pp. 187–198.

[36] J. Hu and R. Marculescu, "Application-specific buffer space allocation for networks-on-chip router design," in *Proceedings of the 2004 IEEE/ACM International conference on Computer-aided design*, ser. ICCAD '04, 2004, pp. 354–361.

[37] S. Murali, M. Coenen, A. Radulescu, K. Goossens, and G. De Micheli, "Mapping and configuration methods for multi-use-case networks on chips," in *Proceedings of the 2006 Asia and South Pacific Design Automation Conference*, ser. ASP-DAC '06, 2006, pp. 146–151.

[38] Z. Guz, I. Walter, E. Bolotin, I. Cidon, R. Ginosar, and A. Kolodny, "Network delays and link capacities in application-specific wormhole nocs," 2007.

[39] A. K. Mishra, R. Das, S. Eachempati, R. Iyer, N. Vijaykrishnan, and C. R. Das, "A case for dynamic frequency tuning in on-chip networks," in *Proceedings of the 42nd Annual IEEE/ACM International Symposium on Microarchitecture*, ser. MICRO 42, 2009, pp. 292–303.

[40] M. F. Chang, J. Cong, A. Kaplan, M. Naik, G. Reinman, E. Socher, and S.-W. Tam, "Cmp network-on-chip overlaid with multi-band rf-interconnect," in *HPCA*, 2008, pp. 191–202.

[41] M.-C. F. Chang, J. Cong, A. Kaplan, C. Liu, M. Naik, J. Premkumar, G. Reinman, E. Socher, and S.-W. Tam, "Power reduction of cmp communication networks via rf-interconnects," in *MICRO*, 2008.

[42] Y. Taur and T. H. Ning, *Fundamentals of Modern VLSI Design*. Cambridge University Press, 1998.

[43] W. Y. Choi, B.-G. Park, J. D. Lee, and T.-J. K. Liu, "Tunneling field-effect transistors (tfets) with subthreshold swing (ss) less than 60 mv/dec," *IEEE Electron Device Letters*, vol. 28, no. 8, pp. 743–745, 2007.

[44] A. Seabaugh and Q. Zhang, "Low-voltage tunnel transistors for beyond cmos logic," *Proceedings of the IEEE*, vol. 98, no. 12, pp. 2095 –2110, Dec. 2010.

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

[46] V. G. Oklobdzija, B. R. Zeydel, H. Dao, S. Mathew, and R. Krishnamurthy, "Energy-delay estimation technique for high-performance microprocessor vlsi adders," in *Proc. 16th IEEE Symp. Computer Arithmetic*, 2003, pp. 272–279.

[47] 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.

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

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

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

[51] 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.

[52] G. Sun, X. Dong, Y. Xie, J. Li, and Y. Chen, "A novel architecture of the 3d stacked mram l2 cache for cmps," in *HPCA*, 2009, pp. 239–249.

[53] X. Guo, E. Ipek, and T. Soyata, "Resistive computation: avoiding the power wall with low-leakage, stt-mram based computing," in *ISCA*, 2010, pp. 371–382.

[54] Z. Diao, Z. Li, S. Wang, Y. Ding, A. Panchula, E. Chen, L.-C. Wang, and Y. Huai, "Spin-transfer torque switching in magnetic tunnel junctions and spin-transfer torque random access memory," *Journal of Physics: Condensed Matter*, vol. 19, no. 16, p. 165209, 2007.

[55] W. Zhao, E. Belhaire, Q. Mistral, C. Chapped, V. Javerliac, B. Dieny, and E. Nicolle, "Macro-model of spin-transfer torque based magnetic tunnel junction device for hybrid magnetic-cmos design," in *Behavioral Modeling and Simulation Workshop, Proceedings of the 2006 IEEE International*, Sept. 2006, pp. 40 –43.

[56] J. A. Kahle, M. N. Day, H. P. Hofstee, C. R. Johns, T. R. Maeurer, and D. Shippy, "Introduction to the cell multiprocessor," *IBM J. Res. Dev.*, vol. 49, pp. 589–604, 2005.

[57] P. Kongetira, K. Aingaran, and K. Olukotun, "Niagara: A 32-way multithreaded sparc processor," *IEEE Micro*, vol. 25, pp. 21–29, 2005.

[58] C. Kim, D. Burger, and S. W. Keckler, "An adaptive, non-uniform cache structure for wire-delay dominated on-chip caches," in *Proceedings of the 10th international conference on Architectural support for programming languages and operating systems*, ser. ASPLOS-X, 2002, pp. 211–222.