Energy Effective 3D Stacked Hybrid NEMFET-CMOS Caches

Mihai Lefter, Marius Enachescu, George Razvan Voicu, and Sorin Dan Cotofana
Faculty of EEMCS, Delft University of Technology, The Netherlands
{m.lefter, m.enachescu, g.r.voicu, s.d.cotofana}@tudelft.nl

***Abstract*—In this paper we propose to utilise 3D-stacked hybrid memories as alternative to traditional CMOS SRAMs in L1 and L2 cache implementations and analyse the potential implications of this approach on the processor performance, measured in terms of Instructions-per-Cycle (IPC) and energy consumption. The 3D hybrid memory cell relies on: (i) a Short Circuit Current Free Nano-Electro-Mechanical Field Effect Transistor (SCCF NEMFET) based inverter for data storage; and (ii) adjacent CMOS-based logic for read/write operations and data preservation. We compare 3D Stacked Hybrid NEMFET-CMOS Caches (3DS-HNCC) of various capacities against state of the art 45 nm low power CMOS SRAM counterparts (2D-CC). All the proposed implementations provide two orders of magnitude static energy reduction (due to NEMFET's extremely low OFF current), a slightly increased dynamic energy consumption, while requiring an approximately 55% larger footprint. The read access time is equivalent, while for write operations it is with about 3 ns higher, as it is dominated by the mechanical movement of the NEMFET's suspended gate. In order to determine if the write latency overhead inflicts any performance penalty, we consider as evaluation vehicle a state of the art mobile out-of-order processor core equipped with 32-kB instruction and data L1 caches, and a unified 2-MB L2 cache. We evaluate different scenarios, utilizing both 3DS-HNCC and 2D-CC at different hierarchy levels, on a set of SPEC 2000 benchmarks. Our simulations indicate that for the considered applications, despite of their increased write access time, 3DS-HNCC L2 caches inflict insignificant IPC penalty while providing, on average, 38% energy savings, when compared with 2D-CC. For L1 instruction caches the IPC penalty is also almost insignificant, while for L1 data caches IPC decreases between 1% to 12% were measured.**

***Index Terms*—Computer Architecture, Memory Hierarchy, Caches, NEMS, Emerging Memories, Low Power, 3DS-IC**

I. Introduction

With the number of transistors on a single silicon die crossing the one billion threshold, power dissipation became of major concern in processor design, as it directly impacts operating costs and reliability. On the one hand, technology scaling can only marginally reduce power consumption, since the MOSFET threshold voltage (V_T) scalability frontier limits the power supply voltage reduction. On the other hand, leakage power, once insignificant for the micro-technology generation of ICs, increases abruptly and becomes the dominant fraction of the total power dissipation for sub-100 nm technologies [1].

With the advent of emerging nano-technologies, alternative memory arrays have been proposed, which make use of Nano-Electro-Mechanical (NEM) devices, e.g., NEM Field Effect Transistors (NEMFETs) [2],[3], NEM Relays (NEMRs) [4], in conjunction with CMOS devices to substantially reduce their energy consumption. In [5] we proposed a low power NEMFET-based dual-port (one write port and one read port) dual-tier 3D stacked hybrid NEMFET-CMOS memory cell (3D-HdpMC) that combines the appealing ultra-low leakage SCCF NEMFET inverter with the versatility of CMOS technology. We demonstrated that the proposed 3D-HdpMC outperforms the "best of breed", in terms of energy consumption, low-power dual-port SRAM cell (10T-DPMC) [6].

In this paper we propose the utilization of 3D Stacked Hybrid NEMFET-CMOS Caches (3DS-HNCC) as an alternative to traditional CMOS SRAM caches (2D-CC). We compare 3DS-HNCC of various capacities against state of the art 45 nm low power CMOS SRAM counterparts in terms of relevant metrics for battery operated SoCs, i.e., footprint, delay, and energy consumption. The footprint of 3DS-HNCC is about 55% larger than the one of CMOS only based caches, owing this to NEMFET's mechanical nature and TSV's size. Since an increased footprint impacts the access circuitry energy contribution to the total cache energy, 3DS-HNCC have a dynamic energy increase of 10% and 25% for cache capacities of 32-kB and 2-MB, respectively. However, 3DS-HNCC implementations provide two orders of magnitude static energy reduction, due to NEMFETs extremely low OFF current. 3DS-HNCC's read access time is on average 7% smaller, while its write access time is with about 3 ns higher, as it is dominated by the mechanical movement of the NEMFET's suspended gate.

In order to identify if the write latency overhead inflicts any performance penalty, we consider as evaluation vehicle a state of the art mobile out-of-order processor core equipped with 32-kB instruction and data L1 caches, and a unified 2-MB L2 cache. We evaluate different scenarios, utilizing 3DS-HNCC caches at different hierarchy levels, on sets of SPEC 2000 [7] benchmarks. We measure an Instructions-per-Cycle (IPC) decrease between 1% to 12% for 3DS-HNCC L1 data caches. However, L1 instruction and L2 caches inflict almost insignificant performance penalty (less than 1% IPC reduction on average). For the considered applications our simulations indicate that, despite of their increased write access time and dynamic energy, L2 3DS-HNCC provide substantial energy savings,

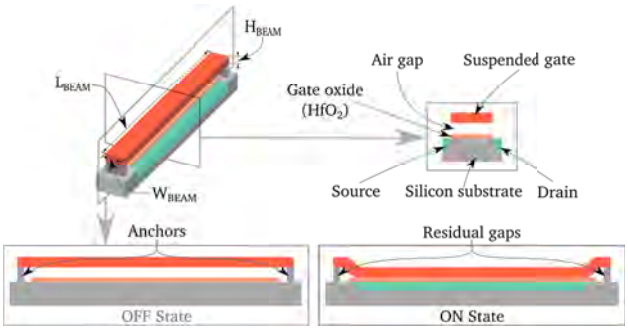

Fig. 1. NEMFET's Geometry [8].

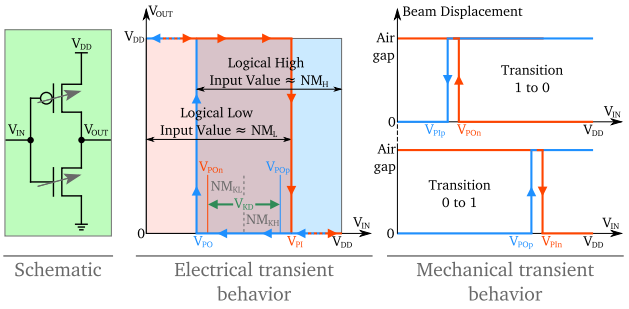

Fig. 2. NEMFET's Inverter Schematic and Transient Behavior.

i.e., 38% caches energy reduction on the average.

The remainder of the paper is organized as follows. In Section II we give a brief presentation of the SCCF NEMFET-based inverter. In Section III 3DS-HNCC are detailed. Next, the section continues with a comparison between 3DS-HNCC and state of the art low power SRAM cache implementations. In Section IV we evaluate the system level implications of utilizing 3DS-HNCC. Finally, Section V presents our conclusions.

II. Low Power Short Circuit Current Free NEMFET Inverter

The Nano-Electro-Mechanical FET (NEMFET), firstly described in [8], is a rather complex device with a 3D geometry and cross-section as presented in Fig. 1, where H_{BEAM} is the thickness of the suspended gate, W_{BEAM} is the width of the beam, and L_{BEAM} is the length of the beam. In the stable position, there is an $air\ gap$ between the $suspended\ gate$ and the $gate\ oxide$. If we apply a difference of potential between NEMFET transistor's gate and source, the gate plate, at some point, pulls-in, and touches the oxide. This happens when the transistor is heading towards the inversion region. On the other hand, when the device is about to leave inversion towards depletion, the electrical force gets smaller due to the reduction of the potential difference that generated the force in the first place, and the gate pulls-out to its original in-air position due to the spring force that pulls the gate plate towards its anchors. NEMFET has an extremely low OFF current and exhibits hysteresis as the Pull-In (PI) and Pull-Out (PO) effects occur at different gate voltage values, further denoted as V_{PI} and V_{PO}, respectively [9], [10].

An important power component in any standard CMOS logic gate is the power consumption induced by the short circuit current, which cannot be eliminated. In NEMFET based logic, however, this issue can be alleviated by separately controlling, at design time, the nNEMFET and pNEMFET hysteresis occurrence, as presented in [11]. The decision on the n/p channel NEMFET sizing (by means of adjusting the suspended beam dimensions) is made such that the following constraints are satisfied: (i) the nNEMFET PI event takes place after the pNEMFET PO event is completed, and (ii) the pNEMFET PI event completes before the nNEMFET PO event starts. From the NEMFET inverter transfer characteristic depicted in Fig. 2 it can be noticed that when the nNEMFET's beam is pulled-in, the beam of the pNEMFET is pulled-out, and vice versa. In [5] we also addressed the following issues related to NEMFET-based logic: (i) NEMFET based inverter noise margin, (ii) scaling for improved performance, and (iii) energy efficiency against "classic" CMOS technology. We concluded that the NEMFET-inverter *noise margin* is larger than the one of a typical MOSFET-based inverter counterpart, when carefully selecting the geometry of the n/p NEMFET device, such that, the V_{POp} is closer to V_{DD}, and V_{POn} closer to *ground* (see Fig. 2). NEMFET's Verilog-A compact model [11] was utilized to design a NEMFET based inverter with high *noise margin*, i.e., high hysteresis width. Hence, a $147mV$ (V_{PO}=0.573V V_{PI}=0.720V) hysteresis width was achieved, for $W_{BEAM} = 45nm$, $L_{BEAMn} = 0.9\mu m$, and $L_{BEAMp} = 1\mu m$ which has a 2 orders of magnitude lower leakage, and dynamic energy reduced with 70%, when compared with the CMOS counterpart.

The hysteresis behaviour of the NEMFET inverter makes it suitable for data retention. The storage functionality is the following: the inverter retains the output value unchanged as long as its input is kept at a voltage within the interval $[V_{POp}; V_{POn}]$ - see also Fig. 2. We further denote this voltage as V_{KD}. Moreover, Fig. 2 suggests that for an optimal NEMFET-based memory cell design in terms of *noise margin*, the geometry of the *n* and *p* channel NEMFET should respect the $NM_{KH} = NM_{KL}$ constraint, where: (i) $NM_{KH} = V_{POp} - V_{KD}$, and (ii) $NM_{KL} = V_{KD} - V_{POn}$.

In the next section we detail the utilization of the NEMFET based inverter in the design of low power caches.

III. Low Leakage 3D-Stacked Hybrid NEMFET-CMOS Caches

In this section we detail processor cache implementations with 3D-stacked hybrid NEMFET-CMOS memories. First, we present the design and the functionality of such hybrid memories. Next, we focus on cache implementations and provide a comparison against standard CMOS in terms of relevant metrics, i.e., dynamic energy, leakage power, footprint, and access time.

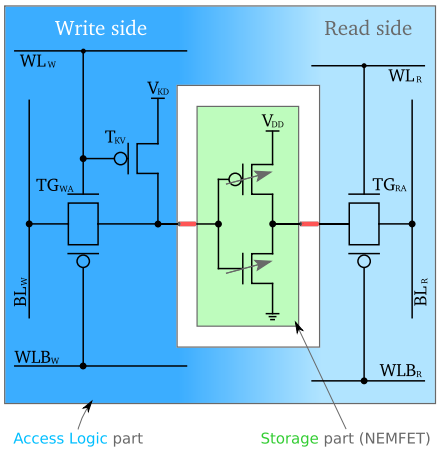

Fig. 3. Hybrid NEMFET-CMOS Memory Cell.

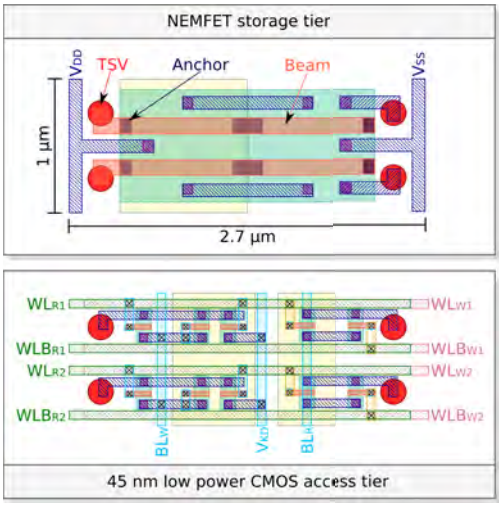

Fig. 4. Two Adjacent Hybrid NEMFET-CMOS Memory Cells Layout.

A. 3D-Stacked Hybrid NEMFET-CMOS Memory

The schematic diagram of the 3D-Stacked hybrid NEMS-CMOS memory cell that we proposed in [5] is presented in Fig. 3, and the storage functionality can be described as follows: (i) the to-be-stored value (0/1) is transmitted at the input of the NEMFET inverter, (ii) the inverter propagates the inverted value (1/0) at its output, and (iii) the inverter retains the output value unchanged as long as its input is kept at a voltage within the interval [$V_{POpNEMFET}$; $V_{POnNEMFET}$], denoted as V_{KD} - see also Fig. 2. We note that there is a clear and natural separation between the read and the write paths, as also reflected in Fig. 3. The CMOS logic consists of five transistors: four are forming the transmission gates TG_{WA} and TG_{RA}, which are utilised for write/read operations, and one is utilised for state retention, as explained further on.

In order to write to a memory cell the required value should be present on the write bitline (BL_W). When the write wordline (WL_W) is asserted the transmission gate TG_{WA} opens and the data item reaches the input of the NEMFET-based inverter, which further outputs its complementary value. During the write operation the pMOS transistor T_{KV} is kept closed by the asserted WL_W line. When WL_W is de-asserted the transistor T_{KV} keeps the inverter input at the stable voltage V_{KD}, such that the inverter output value is maintained unchanged. The write access time is mostly determined by the NEMFET switching time, which is rather slow, as due to mechanical considerations the NEMFET gate requires a certain time interval to stabilize when a state change occurs.

In order to read from a memory cell, the read wordline (WL_R) should be asserted. As a result, the transmission gate TG_{RA} opens and frees the stored data item on the read bitline (BL_R). Due to its large dimensions the NEMFET storage inverter behaves as a powerful driver during read.

We propose to employ TSV-based 3D stacking technology for the final memory structure as it smoothly facilitates the co-integration of NEM and conventional CMOS devices, which, for the time being, appears not to be feasible on the same tier. The proposed memory organization comprises two tiers: the NEMFET-based storage elements reside on the bottom tier, while the CMOS logic required to retrieve, maintain, and write the data is located on the top tier. Tier interconnection is realised through 2 TSVs per memory cell placed one at the input and the other one at the output of the NEMFET inverter, for write and read, respectively (see Fig. 3).

The 3D-stacked hybrid memory requires the same peripheral interfaces from the circuit designer's point of view. Its clear read and write paths separation does not allow common wordlines and bitlines to be employed for both operations. Thus, the memory can act either as a dual-port memory, and requires dedicated address decoders for read and write, or as a single port memory, with a single address decoder and an operation mode signal that selects which wordlines are driven by the decoder's output. Fig. 4 depicts the layout of two adjacent 3D-Stacked hybrid NEMS-CMOS memory cells that share the same bitline. Considering a TSV with a pitch of 0.4μm [12] and 45nm technology for the CMOS tier, a footprint of $2.7\mu\text{m}\times0.5\mu\text{m}$ is required for each memory cell.

B. 3D-Stacked Hybrid NEMFET-CMOS Caches

Given the low leakage benefit offered by the 3D-stacked hybrid NEMFET-CMOS memory arrays [5], in the following we address their utilization as caches in the memory hierarchy of a general purpose processor core.

In Table I we compare caches implemented with 3D-Stacked Hybrid NEMFET-CMOS memories against the most effective, to the best of our knowledge, existing low-power CMOS caches, that employ the 10T non-precharge dual-port SRAM memory cell design introduced in [6]. We consider relevant metrics for battery operated SoCs, i.e., energy, footprint, and delay. The CACTI 6.5 cache memory simulator [13] was utilized to derive the optimal memory partitioning as well as area, access latency, and energy information of peripheral circuitry. A 45 nm low power

TABLE I
CMOS VS. HYBRID NEMFET-CMOS CACHES COMPARISON.

	8-kB 2-way set associative 1 bank		32-kB 4-way set associative 1 bank		512-kB 8-way set associative 2 banks		2-MB 16-way set associative 4 banks	
	10T CMOS	Hybrid	10T CMOS	Hybrid	10T CMOS	Hybrid	10T CMOS	Hybrid
Read energy per access (pJ)	7.29	7.89	11.31	12.57	66.58	76.89	127.34	159.34
Write Energy per access (pJ)	8.08	7.64	11.76	12.05	62.64	70.63	123.99	153.78
Leakage (uW)	26.84	0.47	106.69	1.91	1715.48	33.39	6822.38	92.31
Footprint (mm^2)	0.10	0.17	0.30	0.46	5.15	7.84	17.08	27.12
Read access (ns)	1.78	1.55	1.89	1.67	2.66	2.47	3.47	3.30
Write access (ns)	1.41	4.28	1.52	4.36	2.13	5.15	2.88	5.81

CMOS technology node was considered, with additional changes being performed to the simulator in order to accommodate the different area required by 10T CMOS and 3D stacked hybrid NEMFET-CMOS cells.

For an accurate access latency, active energy, and leakage characterization the circuits of 10T CMOS and NEMFET-CMOS hybrid memory cell arrays are implemented in a commercial 45 nm low power multi-threshold CMOS technology utilising Cadence Virtuoso [14]. *DC* and *transient* simulations are performed using the Cadence Spectre [14] electric simulator, and Agilent ADS [15] in conjunction with Cadence Spectre, for the CMOS and for the hybrid NEMFET-CMOS memory cell, respectively. For the hybrid cell, we considered the write bitline driver output signal (having as load the bitline and the NEMFET inverter equivalent capacitances) generated by Spectre simulator, as input for the NEMFET inverter simulated with Agilent ADS, by means of the NEMFET's Verilog-A compact model from [11]. The TSV contribution was also considered, by means of an RLC model from [16], tailored to the TSV diameter from [12], [17]. Moreover, the bitline and wordline RC parasitics, including the bitline coupling capacitance, for each memory array aspect ratio, are extracted from layout, and included in the simulated circuit schematic. Our simulations were performed in typical case conditions, i.e., typical device models, 1.1V supply voltage, and $27°C$.

The footprint of 3D stacked hybrid NEMFET-CMOS caches is with about 55% larger than the one of CMOS only based caches, as it can be observed in Table I. This affects the length, and thus the wire capacitance, with more powerful drivers being necessary, especially for large caches. Therefore, for large capacities, the dynamic energy of the 3D stacked hybrid NEMFET-CMOS caches access logic is greater than for the CMOS counterpart, even though the cache capacity and the technology of the CMOS tier are the same. We can observe in Table I that for small caches, i.e., 8-kB and 32-kB, the dynamic read/write energies are similar for 3D stacked hybrid NEMFET-CMOS and CMOS only implementations, while for large caches, i.e., 512-kB and 2-MB, there is an increase for the 3D stacked hybrid NEMFET-CMOS implementations of approximately 13% and 25%, respectively.

But NEMFET's larger size is also of advantage, as it makes the inverter-based memory cell a more powerful driver. This has a direct impact on the read access time of the cache, which is on average 7% smaller when compared to CMOS implementations. The write access time of hybrid caches is with about 3 ns higher, as it is dominated by the mechanical movement of the NEMFETs suspended gate. The most important advantage of hybrid caches is given by their almost insignificant leakage, which is two orders of magnitude lower when compared to CMOS only caches.

In the next section we demonstrate that even with a write access time penalty and larger dynamic energy, 3D stacked hybrid NEMFET-CMOS caches provide significant power benefits when compared to CMOS only caches, due to their low leakage advantage.

IV. SYSTEM LEVEL EVALUATION

In this section we analyse the tradeoffs of utilizing 3D stacked hybrid NEMFET-CMOS caches as replacement for CMOS-only based caches at different levels of the memory hierarchy in a single-core environment. First we describe the simulation environment, i.e., assumed processor configuration model, simulator, benchmarks, and considered evaluation metrics. Provided that our proposed hybrid caches have a longer write time when compared to traditional CMOS, we investigate if this penalty propagates at the system level. Next, we focus on energy consumption and determine the benefits that can be achieved by utilizing low power 3D stacked hybrid NEMFET-CMOS caches as replacements for CMOS-only implementations.

A. Evaluation Methodology and Metrics

We consider a system which mirrors a mobile terminal, based on an ARMv7-A out-of-order processor core detailed in Table II, with a two level cache hierarchy: (i) first level consists of separate instruction and data caches, each 32-kB 4-way set associative, and (ii) second level consists of an unified 2-MB 16-way set associative cache. Cache access cycles were derived from Table I. We simulated the system with a modified gem5 cycle accurate simulator [18], able to accept different cache write and read latencies. In our simulations we employed a set of SPEC CPU2000 [7] benchmarks. To have the exact same workload executed we did not consider benchmarks with non-deterministic behavior. Moreover, some SPEC CPU2000 benchmarks could not be compiled and run in gem5 syscall emulation

TABLE II
PROCESSOR CONFIGURATION.

Frequency (GHz)	1.7
L1 size (kB)/associativity	32/4
L1 hit latency CMOS/Hybrid (cycles)	3/3
L1 extra write latency CMOS/Hybrid (cycles)	0/5
L2 size (MB)/associativity	2/16
L2 hit latency CMOS/Hybrid (cycles)	9/9
L2 extra write latency CMOS/Hybrid (cycles)	0/5
Cache line size (bytes)	64
Issue & commit width / Int & FP instruction queues	3/16
Reorder buffer size	32
Functional units	2 Int, 2 LS, 1 FP

mode. Nevertheless, the benchmarks set is representative for our comparison.

We considered four different scenarios: (i) **L1I NEMFET**, i.e., the L1 instruction cache is implemented with hybrid NEMFET-CMOS memory, and all other caches with CMOS, (ii) **L1D NEMFET**, i.e., the L1 data cache is implemented with hybrid NEMFET-CMOS memory, and all other caches with CMOS, (iii) **L2 NEMFET**, i.e., L2 cache is implemented with hybrid NEMFET-CMOS memory, and all other caches with CMOS, and, (iv) **All NEMFET**, i.e., both instruction/data L1 and L2 caches are implemented with hybrid NEMFET-CMOS memory. These four scenarios were compared against a **Baseline** design, i.e., all caches implemented with CMOS memories.

To accurately measure the impact on performance of a longer cache write latency, we utilize in our evaluation the cache average miss latency and Instructions-per-Cycle (IPC) metrics. The number of read/write cache accesses is computed based on the cache hits and misses statistics extracted from the gem5 simulator, and further utilized to determine the energy for each application.

B. Analysis

We start the analysis by observing the average miss latencies for all private L1 instruction and data caches, depicted in Fig. 5 and Fig. 6, respectively. It is clear that the extra write latency of NEMFET-CMOS based caches affects the miss latencies, which increase with 24% and 47% for **L1I NEMFET** and **L1D NEMFET**, respectively.

Fig. 7 presents the IPC reduction of the four scenarios relative to the **Baseline**, for the considered set of benchmarks (the higher the value, the greater the negative performance impact). It is apparent from the figure that the **L1I NEMFET** scenario exhibits insignificant IPC reduction, i.e., less than 0.5%, for almost all benchmarks. There are however two exceptions, i.e., $176.gcc$ and $252.eon$. Even though for these two benchmarks, a similar behavior in terms of average miss latencies is noticed (see Fig. 5), greater IPCs losses were obtained, i.e., 0.85% and 2.87%, respectively. Still, also these IPC reduction values are relatively low, translating in a low performance degradation. Similarly, almost insignificant IPC reduction can be noticed in Fig. 7 for the **L2 NEMFET** scenario.

For the **L1D NEMFET** scenario, IPC reductions between 1.08-12.21% are observed, when compared to the **Baseline**. This can be explained as follows: (i) L1 data cache is more often accessed in comparison with L1 instruction (data are read/written by he processor, while instructions are only read), (ii) L1 data cache is more often accessed in comparison with L2, being closer to the processor, (iii) miss rates for L1 data are greater than in other caches, and, (iv) larger L1 data average miss penalties when compared to the other caches are observed (see Fig. 6) due to the extra write latency.

Finally, the largest IPC reductions are obtained for the **All NEMFET** scenarios. This is natural, as all the caches are affected by the extra write latencies, which impact the overall performance. From the performance perspective **L1I NEMFET** and **L2 NEMFET** scenarios values exhibit almost insignificant IPC reductions.

Fig. 8 depicts the relative energy differences from the **Baseline** of the four scenarios. Positive values denote the energy reduction percentages, while negative values stand for energy increases. Due to the low leakage advantage of NEMFET-CMOS memories, an important energy reduction, i.e., 38% on the average, is obtained for the **L2 NEMFET** scenario. This happens because: (i) L2 is larger, thus its energy is dominant in the total cache hierarchy energy, and, (ii) much fewer accesses in L2 than in L1 caches generate an increase in the contribution of L2 static energy to the total energy. For **L1I NEMFET** and **L1D NEMFET** scenarios an energy increase is observed in all cases, with maximums of 3% and 5.58%, respectively. This is related to the fact that both L1 caches are very often accessed, thus, they have a high activity factor, leading to a higher dynamic energy contribution. Moreover, NEMFET-CMOS caches exhibit a higher dynamic energy when compared to their CMOS counterparts (see Table I). An important energy reduction, i.e., 35% on the average, is obtained for the **All NEMFET** scenario, since it cumulates the effects of all scenarios, thus the energy benefits in L2 caches compensate the energy losses in L1 caches.

V. CONCLUSIONS

In this paper we proposed to replace the traditional CMOS based processor caches with energy effective 3D-Stacked Hybrid NEMFET-CMOS ones. This solution provides two orders of magnitude cache static energy reduction, due to the NEMFET's extremely low OFF current, albeit with a slightly increased dynamic energy consumption, an approximately 55% larger footprint, and a longer write access time. We compared the performance of 3D-Stacked Hybrid NEMFET-CMOS against 45 nm low power CMOS SRAM based cache embodiments at different levels in the memory hierarchy. In order to find out if the write latency overhead inflicts any performance penalty, we considered as evaluation vehicle a state of the art mobile out-of-order processor core. Our simulation on a set of SPEC 2000 benchmarks proved that the extra write latency impact on the overall system performance is rather low or even negligible. Next we evaluated the energy benefits of utilizing 3D-Stacked Hybrid NEMFET-CMOS. Our simulations

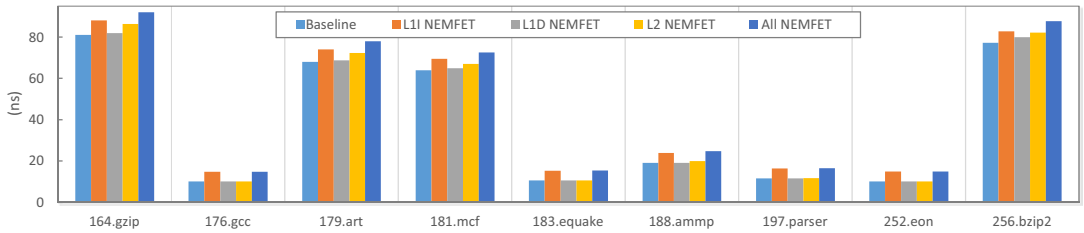

Fig. 5. L1 Instruction Cache Average Miss Latencies.

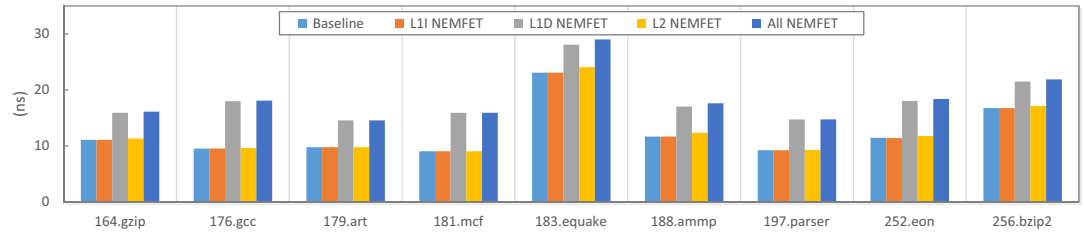

Fig. 6. L1 Data Cache Average Miss Latencies.

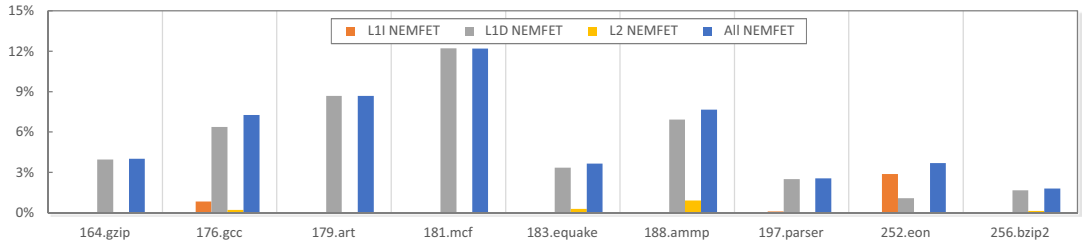

Fig. 7. IPC Reduction Relative to **Baseline** (lower is better).

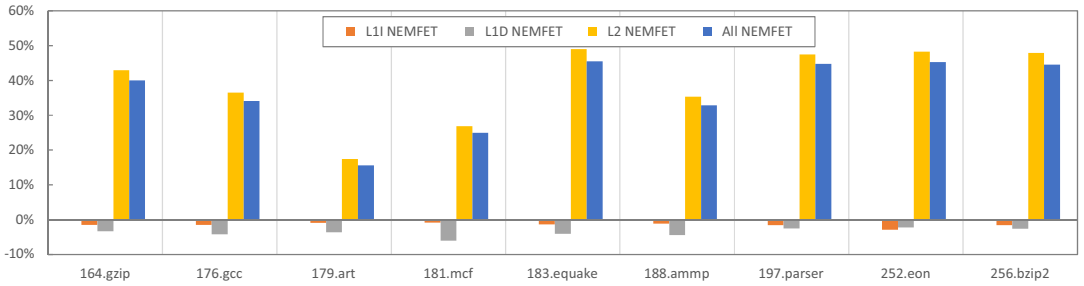

Fig. 8. Energy Difference Relative to **Baseline** (higher is better).

indicate that, in spite of their increased write access time and dynamic energy, 3D Stacked Hybrid NEMFET-CMOS L2 caches provide substantial energy savings, i.e., 38% total caches energy reduction on the average.

REFERENCES

[1] S. Borkar, "Exponential challenges, exponential rewards - the future of Moore's law," *VLSI-SOC*, 2003.

[2] N. Abele, A. Villaret, A. Gangadharaiah, C. Gabioud, P. Ancey, and A. M. Ionescu, "1T MEMS memory based on suspended gate MOSFET," in *IEEE IEDM 2006*.

[3] H. Dadgour and K. Banerjee, "Hybrid NEMS-CMOS integrated circuits," *IET 2009*.

[4] R. Venkatasubramanian, S. K. Manohar, and P. T. Balsara, "NEM relay based memory architectures for low power design," in *IEEE NANO 2012*.

[5] M. Enachescu, M. Lefter, and S. Cotofana, "Low-Leakage 3D Stacked Hybrid NEMFET-CMOS Dual Port Memory," submitted to *JETCAS*.

[6] H. Noguchi *et al.*, "Which is the best dual-port SRAM in 45-nm process technology?" in *IEEE ICICDT*, 2008.

[7] J. L. Henning, "SPEC CPU2000: measuring CPU performance in the new millennium," *Computer*, vol. 33, no. 7, pp. 28–35, Jul 2000.

[8] A. M. Ionescu *et al.*, "Modeling and design of a low-voltage SOI SG-MOSFET," in *ISQED*, 2002.

[9] K. Akarvardar *et al.*, "Analytical modeling of the suspended-gate FET and design insights for low-power logic," *IEEE Tran. on Electron Devices*, vol. 55, no. 1, pp. 48–59, 2008.

[10] M. Enachescu *et al.*, "Can SG-FET replace FET in sleep mode circuits?" in *International ICST Conference on Nano-Networks*, Luzern, Switzerland, October 2009, pp. 99–104.

[11] M. Enachescu, M. Lefter, A. Bazigos, A. M. Ionescu, and S. Cotofana, "Ultra low power NEMFET based logic," in *IEEE ISCAS*, 2013.

[12] A. Topol *et al.*, "Enabling SOI-based assembly technology for three-dimensional (3D) integrated circuits (ICs)," in *IEEE IEDM*, 2005.

[13] N. Muralimanohar *et al.*, "CACTI 6.0," Tech. Rep., 2007.

[14] "Cadence Design Systems," 2011.

[15] "ADS Agilent," 2012.

[16] H. Chaabouni *et al.*, "Investigation on TSV impact on 65nm CMOS devices and circuits," in *IEEE IEDM*, 2010.

[17] A. Topol *et al.*, *3D Fabrication Options for High-Performance CMOS Technology*, C. S. Tan, R. J. Gutmann, and L. R. Reif, Eds. Springer, 2008, vol. Wafer Level 3-D ICs Process Technology.

[18] N. Binkert *et al.*, "The gem5 simulator," *SIGARCH Comput. Archit. News*, vol. 39, no. 2, pp. 1–7, Aug. 2011.