On Microarchitectural Modeling for CNFET-based Circuits

Tianjian Li, Hao Chen, Weikang Qian, Xiaoyao Liang and Li Jiang
Shanghai Jiao Tong University, Shanghai, China
Email: {ltj2013, ljiang_cs}@sjtu.edu.cn

***Abstract*—Carbon Nanotube Field-Effect-Transistors (CNFETs) show great promise to be an alternative to traditional CMOS technology, due to their extremely high energy efficiency. Unfortunately, the lack of control over the Carbon NanoTube (CNT) growth process causes CNFET circuits to suffer from the CNT count variation, which degrades the CNFET circuit performance. Compared to the CMOS process variation, the CNT count variation exhibits asymmetric spatial correlation. In this work, we propose an analytic model that integrates the impact of the asymmetric spatial correlation into the key microarchitectural blocks. We use this model to evaluate the variations in circuit performance for different layout styles and microarchitectural parameters. We further explore the opportunity of leveraging the asymmetric spatial correlation for performance enhancement. Experimental results based on SPICE simulation and architectural simulations showed the accuracy and effectiveness of the proposed model.**

I. Introduction

Carbon nanotube field-effect transistors(CNFETs) have gained significant momentum recently as one of the most promising alternatives for CMOS technology. A CNFET uses several doped CNTs as the transistor channels, as shown in Fig. 1. Except transistor channel, the remaining components of CNFET and their fabrication process are compatible to the CMOS technology. Due to the perfect electrical characteristics of CNTs, i.e., the extremely low dynamic power and near-zero leakage power, CNFET circuits can improve the energy-delay product, a measure of energy efficiency, by more than an order of magnitude, compared to the CMOS technology [12].

Despite the promising benefits of CNFETs, some imperfections are observed in the CNFET fabrication process, including the metallic CNTs (m-CNTs), CNT density variations and mis-positioned CNTs [16]. Layout techniques [12] are proposed to guarantee functional correctness in CNFET-based circuits with mis-positioned CNTs. M-CNTs can be efficiently removed with various chemical and electrical methods [13]. Due to the lack of effective control on the CNTs growth, the number of CNTs in the CNFETs exhibits large variability, resulting in huge variations in the driving capability of CNFETs and delay of the CNFET-based circuit. This in turn degrades the circuit performance. Only methods such as naive CNFET upsizing [18] and CIDER [14] are known to alleviate the impact. The hardware cost, energy, and the complexity in fabrication process, however, limit the use of these methods in complex modules of microprocessors. Studies have shown the effective variation toleration in mircoarchitectural level for CMOS circuits [9], indicating their superiority compared to device level solutions. Thus, a high level timing model is essential to evaluate the impact of the CNT variation.

The basic device-level CNFET delay models have been developed in [5], based on which the device and circuit delay in the presence of CNT density variations can be evaluated [11]. Ghavami et al. [7] proposed a statistical model to predict the timing yield of CNT-based circuits. This model gives the distribution of the paths with respect to CNT density variation and the probability of residual m-CNTs. However, existing timing models omit one underlying characteristic that is unique for CNFET-based circuits: strong asymmetric spatial correlation in the CNT density variation [18]. That is, the CNFETs aligned along the CNT growth direction share the same set of CNTs, and thereby exhibit the same driving capability. By exploring this asymmetric spatial correlation, study [18] suggested a dramatic yield enhancement by enforcing all CNFETs to be perfectly aligned along the CNT growth direction.

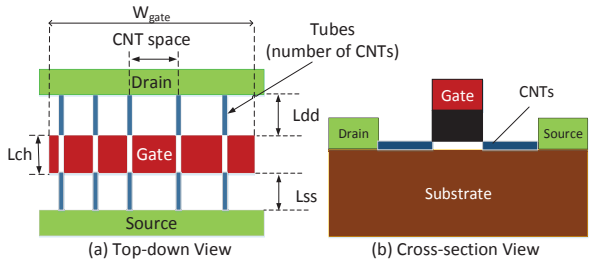

Fig. 1. 2-D view of CNFET

The asymmetric spatial correlation may induce two impacts. First, the existing variation-aware timing models may become too optimistic without considering this asymmetric spatial correlation. For example, if a path lies along the CNT growth direction, then the delay of gates in the path are highly correlated. The resulting path becomes either extremely fast or slow. Second, the asymmetric spatial correlation may be amplified in those microarchitectural blocks with regular structures(e.g., register files, caches etc.). The layout decisions (e.g., the direction of critical paths versus the CNT growth direction) can significantly affect the performances of these blocks. We use the CNFET-based SRAM as our compelling starting point for case study because i) it composes the major portion of the modern billion-transistor scale microprocessor; and ii) it outperforms the MOSFET SRAM in terms of power, access latency, and static noise margins [8]. Therefore, it is of highest priority to build an analytical timing model for the CNFET-based SRAM, in the presences of the CNT related variations and the asymmetric spatial correlation in these variations. With this model, we may explore the architectural design of CNFET-based cache. We propose a novel technique, by disabling the cache line suffering extreme performance degradation, under above variations. The experimental results show significant performance enhancement.

978-1-4673-9094-1/15/$31.00 ©2015 IEEE

The rest of the paper is organized as follows. Section II reviews the background of CNFET variation and modeling. Section III presents the microarchitectural timing model for general logic unit and array-based unit. We evaluate the model with SPICE simulations in section IV, based on which we explore the CNFET-SRAM based cache design in section V. Section VI concludes this paper.

II. Background

A. CNFET-based circuit model

The first step to model the CNFET circuit is to obtain the CNFET driving capability, which depends on the CNT count, i.e., the number of parallel CNTs that connect source and drain contacts in that CNFET. Its value, on the one hand, depends on the density variation. By assuming that the CNT spacing follows the Gaussian distribution, Zhang et al. [6] showed that the CNT count within a FET follows an asymptotic Gaussian distribution. On the other hand, the semiconducting CNTs (s-CNTs) may be mistakenly removed when the m-CNTs are removed. Taking above issues into account, the number of "survived" s-CNTs in a CNFET N_{surv} can be derived [13]. With the s-CNT count, the delay of a CNFET-based gate is estimated based on a compact SPICE CNFET device model [5]. The delay of a CNFET device with n parallel s-CNTs is estimated from the driving current of the device and the total gate capacitance of the device with n parallel CNTs:

$$T_{CNFET} = \frac{C_{CNFET,n} \times V_{supply}}{I_{CNFET,n}} \quad (1)$$

where $C_{CNFET,n}$ and $I_{CNFET,n}$ are the total capacitance and driving current of the CNFET device with n parallel CNTs, respectively, and V_{supply} is the supply voltage.

Using the above device model, Ghavami et al. estimate the CNFET-based gate delay[7]. As the number of survived s-CNTs in equation **??** relates to the number of nanotubes and s-CNTs, by assuming the latter two parameters as random variables, they add up the delay of each CNFET device on the critical path to derive the path delay:

$$T_{path}(n_{cp}) = \sum_{k=1}^{n_{cp}} T_{gate}(k) \quad (2)$$

where $T_{path}(n_{cp})$ is the delay of a path with n_{cp} length, and $T_{gate}(k)$ is the delay of the k-th gate, in which all the CNFET devices in this column contain n survived s-CNTs. It should be noted that other sources of variations, including the wire delays, are ignored.

B. Motivation

Many imperfections-tolerant CNFET-based SRAM designs have been proposed [17], which promote the idea of applying CNFET-based caches in microprocessors. Among various factors, the performance is the most important concern in designing the CNFET-based SRAMs. Previous works modeled the cells in CNFET-based SRAMs [8], from which the CNFET-based SRAM model was derived by considering other logics, e.g., the address decoder [4]. With these timing models, we can accurately evaluate the cache access time, and examine the effect of cache sizes and aspect ratios on the cache access time for on-chip cache, as in [15]. However, the asymmetric spatial correlation associated with CNT density variations was not included in the SRAM timing model developed by previous works, which may cause a large inaccuracy in the model and mislead the CNFET-based cache design in the future.

The asymmetric spatial correlation is an inevitable result of two facts: i) the CNTs served as transistor channel may grow very long and be covered by many CNFETs; ii) the active regions of CNFETs are enforced to be aligned to each other along the CNT growth direction to improve yield [18]. The aligned CNFETs along the CNT growth direction share the same set of CNTs and thereby have highly correlated delay. While the CNFETs aligned perpendicular to the CNT growth direction exhibit little correlation as they utilize different sets of CNTs. For example, all the gates in path 1 in Fig. 2 have almost the same delay, while the delays of all the gates in path 2 are almost independent. Consequently, path 1 has much higher likelihood to become extremely fast or extremely slow, showing higher variance in path delay. On the contrary, since the delays of all the gates in path 2 are independent, their variances can mutually cancel out each other and hence, the path delay does not have a large variance.

With this asymmetric spatial correlation, the timing models for CNFET-based SRAMs proposed by previous works are no longer accurate. For instance, if CNTs are parallel with the wordline, then all the cells driven by the same wordline have almost the same delay and all the sense amplifiers (SA) will also have the similar characteristics. If CNTs are parallel with the bitline, all the cells connected to the same bitline will show strong correlation. In this paper, we take this into account and revisit the timing model for CNFET-based SRAM. Moreover, the above asymmetric spatial correlation may affect some microarchitectural blocks with regular structures, such as array-based register files and caches. This motivates us to map the timing model from circuit level to microarchitectural level. We use the CNFET-based SRAM as a case study. Leveraging our microarchitecural model, we explore several designs of CNFET-based SRAM and evaluate their performances.

III. Microarchitectural Timing Model

We classify the microarchitectural blocks into two types: i) logic blocks and ii) array-based blocks, because the asymmetric spatial correlation affects their timing in different ways.

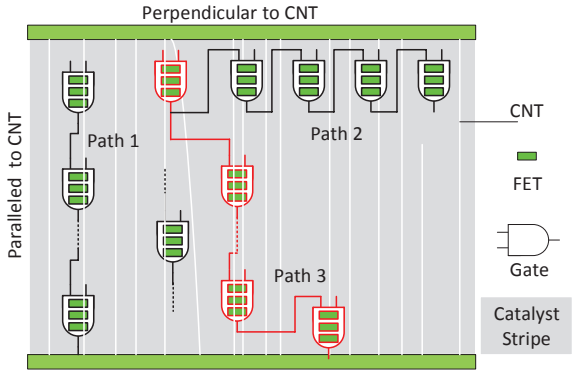

Fig. 2. The correlation in different critical paths

A. General Logic Model

To obtain a general timing model for logic blocks, we first apply Equation (2) to derive the path delay T_{path}. Afterwards, we can use the FMAX model [3] to develop the microarchitectural timing model, in which the delay is decided by the important parameters such as n_{cp} (the number of gates on a critical path) and N_{cp} (the number of critical paths in a unit block). And the delay of a logic unit is expressed as

$$T_{unit} = \max_{1 \le k \le N_{cp}} T_{path,k}(n_{cp}) \tag{3}$$

where $T_{path,k}$ refers the delay of the kth path. The remaining questions are i) how to integrate the asymmetrically correlated CNT variation to the above model, and ii) how to derive an accurate coefficient of the asymmetric spatial correlation (CNFETs aligned along CNT growth direction are not 100% correlated).

1) *Considering of the asymmetric spatial correlation:* To handle the spatial correlation in the CMOS technology, the quart-tree method was used to get the delay mean and the delay variance of critical path in a circuit [1]. Unlike the CMOS technology, we decompose the entire circuit into columns which are parallel with the CNT growth direction. The width of the column equals the gate width and therefore, two CNFET gates which are not aligned with the CNT growth direction belong to two different columns. Here, we assume all the CNFETs within a logic gate have the same CNT count. The length of the column equals to the width of catalytic strip. Note that, the CNTs in different catalytic strips are fabricated independently and therefore no correlation on CNT count exists across the catalytic strips. Consequently, we divide the path into different columns as follows:

$$T_{path} = \sum_{k=1}^{n_c} T_{column}(k) \tag{4}$$

where n_c is the number of columns over which the gates on the critical path spans, and $T_{column}(k)$ is the total delay of the gates in the k-th column. For example, the gates in path 1 are correlated in 1 column, while gates in path 2 are divided into 5 columns and 3 columns about path 3. The total delay of the gates in the k-th column is calculated as follows:

$$T_{column}(k) = \sum_{k=1}^{n_g} T_{gate}(k) \tag{5}$$

where n_g is the number of gates in the column.

2) *Derivation of the accurate correlation coefficient:* The coefficient of the asymmetric spatial correlation is based on the following observation: the correlation coefficient is as high as 100% [6] when two gates are close to each other in the same column, and decreases as the distance between the two gates increases. However, as long as the two gates are within the same catalytic strip, their correlation coefficient can still maintain as high as 90% [6]. The non-100% correlation, i.e., the CNT count divergence in the same column, is mainly because of the misaligned CNTs and fractured CNTs. Suppose the CNT count divergence between gate i and j is $\Delta CNT_{i,j}$. Intuitively, $\Delta CNT_{i,j}$ increases as the distance of gate i and j increasing. Suppose the probability of X CNTs are fractured or misaligned is $P(X)$ in a column, the probability of CNT count divergence can be estimated as

$$P(\Delta CNT_{i,j}) = \frac{L_{i,j}}{L_{strip}} \times X \times P(X) \tag{6}$$

wherein $L_{i,j}$ is the distance between gate i and j, and L_{strip} is the length of the catalytic strip. To simplify the calculation, we assume the CNT count of the middle gate (located in the middle of that catalytic strip) in each column equals to the CNT count randomly assigned to that column, CNT_{mid}. Then, the CNT count in other gates in the same column can be calibrated by counting the distance of that gate to the middle gate using the following equation:

$$Cor(i, mid) = CNT_{mid} + \Delta CNT_{i,mid} \tag{7}$$

where $CNT_{i,mid}$ can be derived using equation (6) by randomly generating $P(X)$ (relates to the probability of CNT length variation and misaligned probability). Combine equation (2) and (7), we get the delay of the column

$$T_{column} = \sum_{k=1}^{n_g} T_{gate,Cor(k,mid)}(k) \tag{8}$$

where, $T_{gate,Cor(k,mid)}(k)$ is the k-th gate delay, which can be calculated by taking $Cor(k, mid)$ into the equation . Now, we can use equation (4), (3) and (8) to calculate the delay of the logic unit.

B. Model of Array-based Block

The array-based blocks occupy a large portion in modern microprocessor. Their regular structures and layouts are strongly correlated to the CNT layout. Thus, this type of blocks are sensitive to the asymmetric spatial correlation of CNT count. We use the SRAM as a case study to illustrate our modeling methodology. According to Wada et al [15], the access time for reading an SRAM can be divided into four different parts: decoder delay, wordline delay, bitline/sense amplifier delay and data-out delay:

$$T_{access} = T_{decoder} + T_{wordline} + T_{bit/sense} + T_{dataout} \tag{9}$$

In this paper, we model the access time for SRAM array following the method in [15], but we emphasize the impact of CNT density variation and the asymmetric spatial correlation.

1) *Decoder delay*: Each sub-bank has its own row decoder circuit. We use the proposed logic model to calculate the decoder delay. The delay of decoding is the time that signals pass from input to the output. The decoder can be considered as a circuit with a large fanout. The critical paths of accessing the two adjacent wordlines share most of the gates in decoder.

2) *Wordline delay*: The wordline driver is a CNFET inverter, which pulls up the voltage of selected wordline to reach the threshold voltage of the access transistors in all memory cells. The wordline circuit is shown in Fig.3. Conventionally, each memory cell is assumed to have the same capacitance and each wordline driver has the same size. But it is not valid for CNFET-based SRAM, due to the asymmetric spatial correlation in CNT variation. The wordline access delay, hence is based on the layout: if the wordline is parallel with the CNT growth direction, the pass transistors driven by the same wordline have strong correlations, and thereby these pass

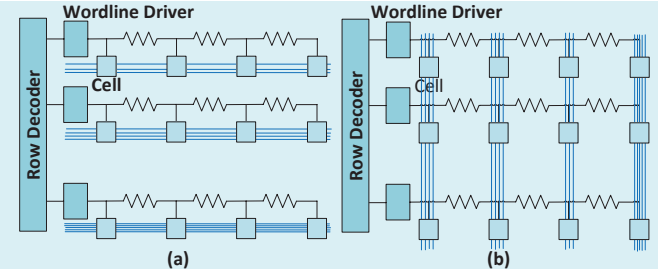

Fig. 3. Circuit of Wordline:(a) wordline parallel with CNT growth (b)wordline perpendicular to CNT growth

transistors have the similar gate capacitance and access delay, as shown in Fig. 3(a). Otherwise, each pass transistor has dramatically different gate capacitance, and the weakest pass transistor dominates the access delay of the whole wordline, denoted as "worst case". In addition, the load capacitance on the interconnect is the sum of unit capacitance associated with each pass transistor. Therefore, the latter layout cancels out the variation among the unit capacitances.

3) *Bitline/Sense Amp delay*: The timing model for bitline and sense amplifier is the most complicated one, involving the bitline pre-charge, charge/discharge of the bitline load from the memory cell, and the sense amplifying. To read a cell, we first pre-charge the two bitlines. After all the pass transistors driven by the active wordline are switched on, they begin to charge/discharge the bitline load. If the value in the cell is 1, the BL will be logic 1 and the BLB will be discharged to produce the voltage difference. If the value is 0, the opposite would happen. When the voltage difference between BL and BLB reaches the minimum value that can be sensed, the sense amplifier amplified this difference to produce the stable value: VDD for 1 and GND for 0.

In the bitline/sense amplifier delay, the memory cell, bitline load and sense amplifier play the important roles, and they are affected by the asymmetric correlated variation differently. The bitline load is the sum of the unit capacitance (mainly the drain junction capacitance of the pass transistors) of the 1-bit cell along the bitline, whose total number is the number of wordlines in one bank. Thus, in the layout like Fig. 3(a), the variation of bitline load is canceled out. As the memory cell only charges/discharges the bitline in a very short period we focus on the delay variation for the sense amplifier. In the critical path of accessing the bitline, there is only one sense amplifier, therefore, there are no cancelling out effect for the sense amplifiers, no matter which layout is chosen. On the contrary, they exhibit the "worst case": if the wordline is parallel with the CNT growth direction, the driving capability of all the sense amplifiers are highly correlated, as shown in Fig. 3(a). Otherwise, each sense amplifier has dramatically different driving capability, and the overall sensing delay is dominated by the weakest sense amplifier.

4) *Data out delay*: The content stored in the selected cell should be output through output driver to the data buffer. The output capacitance of the data-bus driver and the capacitance of the data buffer have the similar effect with the sense amplifier under the asymmetric correlated variation.

5) *Summary:* As mentioned above, four different parts

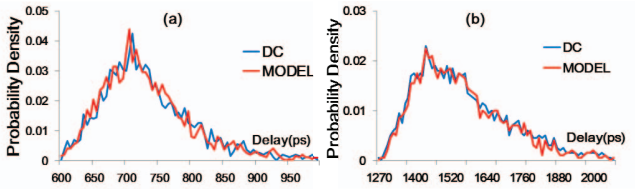

Fig. 4. Validating the logic model: (a) 16-bit multiplier and (b) M1 CPU

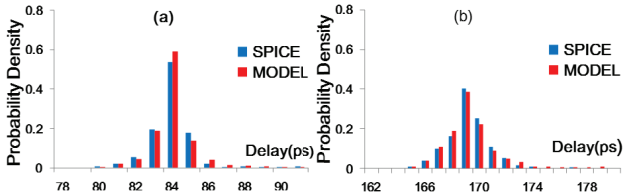

Fig. 5. Result of the wordline delay: (a) 64-bit wordline; (b) 128-bit wordline

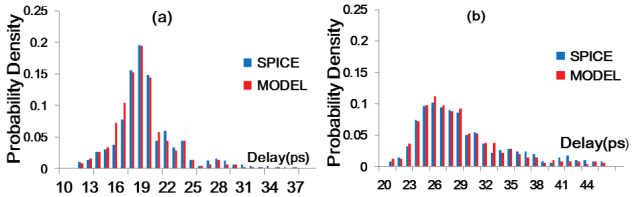

Fig. 6. Result of the bitline delay: (a) 64-bit bitline; (b) 128-bit bitline

are combined to get the total access time. According to the access timing model, we can imply that the access time of a cache memory is determined by cache size, bank numbers, block size, associativity, and more importantly, whether the cell arrangement direction is parallel with or perpendicular to the CNT growth direction. Changing the organization of the cache may optimize the access time.

IV. Model Validation

Based on the Stanford CNFET SPICE model [5], we apply Monte Carlo simulations to validate our analytical timing model. The gate width and CNT spacing are set as 32nm and 4nm, respectively. The probability of m-CNT is set between 10% to 30% randomly. The probability of misaligned CNTs and fracture CNTs is set between 1% to 5% randomly. It should be noted that, the latter two parameters is user-defined and can be calibrated using sample data.

A. Validating the Logic Model

First, we randomly generate 2000 CNFET chip samples to validate the logic model. Fig. 4 presents the delay distribution of two microarchitectural units, i.e., a 16-bit multiplier and M1 CPU from OpenCore benchmark. The result derived from our model, denoted as "MODEL", is compared to the results using static timing analysis on the design layout with detailed place&route, denoted as "DC". We apply our model on the layout, i.e., by dividing critical paths into columns. The width of the column is fit to the width of gates. We calculate the maximum delay among all the critical paths. Noted that we calibrate the gate capacitance in our model with an estimated interconnect load capacitance, which is derived by assuming an constant and average distance between gates. The results for both designs show significant delay variations, which indicate

that the CNT variation is a critical challenge for very large scale circuit design. Generally, the proposed logic model have a good match with the STA results.

B. Validating the Array Model

Next, we validate the array-based model using 500 sample CNFET-based SRAM arrays. Here, only read operation is considered because read operation consumes longer time than that consumed by write operation. Fig. 5 and Fig. 6 present the results of the worst wordline and bitline delay with the parameters simulated by the analytical model and SPICE respectively. The results show that the proposed model has a good match with the SPICE simulation with the wordline length of 64 and 128 bits. Same observation can be seen for 64 bits and 128 bits length bitline delay. In these results, the CNT growth direction is parallel to the wordline. We also validate the model by setting CNT growth direction perpendicular to wordline. Due to the paper limit, we omit the results here.

Compare Fig. 5(a) and (b), the wordline delay increases significantly as the wordline contains more cells. For CNFET-based circuits, the CNT-metal capacitance dominates the gate capacitance. The increment of the pass transistors mainly contributes to the wordline delay. Moreover, Fig. 5(b) exhibits larger variation (see the scale of X-axis) than (a). Because the pass transistors in the same wordline exhibit the highly correlated gate capacitance, the load capacitance may be extremely large or extremely small in this layout due to the asymmetric spatial correlation. The longer wordline amplifies the variation. In Fig. 6, we also observe the above phenomenon, because the sense amplifiers, which are aligned parallel to wordline (along the CNT growth direction), dominate the bitline charging (discharging) duration.

Note that, in this work, we ignore the delay variations in metal wires, because the circuits, though with different CNT growth direction, have the same layout of the MOS-like structures and the routing of metal wires.

V. Case Study: Cache Design Exploration

With the derived model, we choose the on-chip cache as our case for studying the impact of the CNT layout. Cache is implemented with SRAM arrays. We assume they are designed to allow variable latency—-a widely used technique for variation tolerance [10]. Therefore, the access latency for the cache can be variable, depending on which wordline is accessed (not dominated by the worst bit).

A. Stand Alone Characterization

Fig. 7 shows the access time of different cache sizes with the fixed wordline length. The access time of each bit may show variations due to the CNT variations, but the worst case delay of all bits determines the cache access time. The access time increases as the cache size increases, due to the increasing length of the bitline. In addition, the variation of access time also increases more obviously as the cache size increases. That is because, larger cache has more wordlines, each of which exhibits large variation. Comparing Fig. 7(a) with Fig. 7(b), the cache with wordline parallel to the CNT

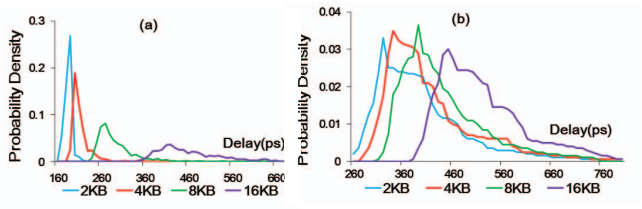

Fig. 7. Results of the access time varying the cache size when wordline is (a) parallel and (b) perpendicular to CNT growth direction

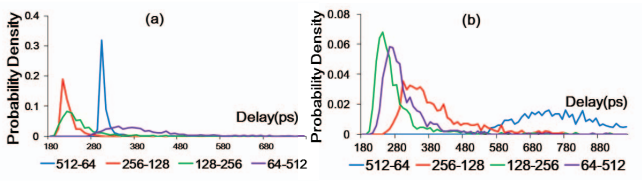

Fig. 8. Results of the access time varying word-bit ratio when wordline is (a) parallel and (b) perpendicular to CNT growth direction

growth direction performs better than the opposite: the mean delay and the variance are smaller, and the 99% time yield is higher. The underlying reason can be explained as follows: when wordline is perpendicular to the CNT growth direction, each pass transistor in a wordline show significant variation (as well as the sense amplifiers), among which the worst case pass transistor dominates the wordline delay. Similarly, the bitline charging/discharging time is dominated by the sense amplifier with the weakest driving capability. Moreover, due to the asymmetric spatial correlation of CNT count variation, the slowest transistor gate and the weakest sense amplifier must be in the same bitline, i.e., the worst case bitline. Therefore, the read operation is dominated by this worst "bit". On the contrary, when wordline is parallel to the CNT growth direction, each pass transistor and each the sense amplifier have almost the same delay. Therefore, the access time is dominated by the worst "line" of pass transistors and sense amplifiers, which is much smaller in statistics.

Fig. 8 shows the access time of a direct mapped 8KB cache with different wordline-bitline length ratio for a fixed cache size. In Fig. 8(a), the larger ratio contributes to less variation, because the cache access has less chance to hit the worst case wordline access delay. We find a sweet-point ratio, i.e., 256-128, which outperforms others in terms of both mean delay and variance. In Fig. 8(b), each wordline has chances to hit the worst case access delay. However, this chance reduces as the wordline becomes shorter. Interestingly, the sweet-point ratio is different, i.e., 128-256, in this layout.

B. Architectural level characterization

We integrate our analytical CNFET-based SRAM model into the cycle accurate architectural simulator Gem5 [2] for the performance evaluation using SPEC2006. The nominal core frequency is set at 2GHz while the supply voltage is set at 0.8V, with 32KB L1 data cache and 16KB L1 instruction cache. In the default simulator, it takes 2 cycles to access the L1-cache. We revise the simulator to allow variable latency for the cache using the results generated by the proposed model.

Fig. 9 shows the results of a L1-cache (32Byte block

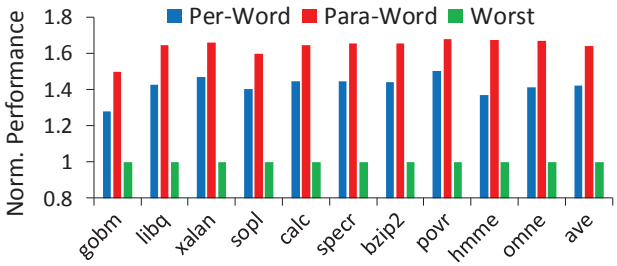

Fig. 9. Simulation results on cache performance under asymmetric correlation of CNT count variation

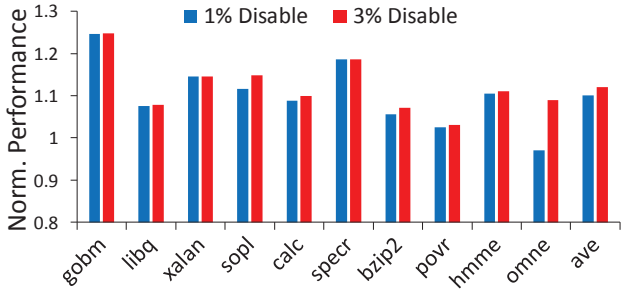

Fig. 10. Simulation results on cache performance after applying the cache line disabling method

and 2-way associativity), which are normalized to the worst-case cache design with CNT variation. The worst case sets the cache access time as the bit cell with the worst delay in the entire cache, while the other two adopt the variable latency design. The Per-Word/Para-Word denotes the layout style where the CNT growth direction is perpendicular/parallel to the wordline. We observe that Per-Word is worse than Para-Word in performance by about 20%, which meets the stand alone characterization. Both cases require a similar level of testing and hardware overhead to figure out and store the latency settings and this again shows the importance of the CNT layout on the architecture decision. In the Para-Word scenario, significant access time variation exists in different cache lines. We advocate to disable the cache line with long access time using the power gating. Fig. 10 shows the simulation results of the above disabling method, wherein the results are normalized to Para-Word in Fig. 9. It shows that disabling 1% cache lines can improve the performance by around 10%. The extreme long access latency only takes a limited part. Therefore, disabling 3% cache lines can't get much improvement.

VI. Conclusion

In this paper, we describe an analytical timing model for CNFET-based microarchitectural blocks, emphasized in characterizing the impact of CNT asymmetric spatial correlation, which can quickly and precisely get the delay of blocks. We also propose the design technology in the case study to mitigate its impact. The experimental results indicate that the unique CNT variation plays an important role in microarchitectural design.

VII. Acknowledgement

This work is partly supported by Shanghai Science and Technology Committee (Grant No. 15YF1406000), the National Natural Science Foundation of China (Grant No. 61202026 and No. 61332001) and Program of China National 1000 Young Talent Plan.

References

[1] A. Agarwal, D. Blaauw, and V. Zolotov, "Statistical timing analysis for intra-die process variations with spatial correlations," in *International Conference on Computer Aided Design*, Nov 2003, pp. 900–907.

[2] N. Binkert et al., "The gem5 simulator," *ACM SIGARCH Computer Architecture News*, vol. 39, no. 2, pp. 1–7, 2011.

[3] K. A. Bowman et al., "Impact of die-to-die and within-die parameter fluctuations on the maximum clock frequency distribution for gigascale integration," *J. Solid-State Circuits*, vol. 37, no. 2, pp. 183–190, 2002.

[4] D. Cheng et al., "Optimal Redundancy Designs for CNFET-Based Circuits," *Asian Test Symp.*, pp. 25–32, 2014.

[5] J. Deng and H.-S. Wong, "A compact spice model for carbon-nanotube field-effect transistors including nonidealities and its applicationpart ii: Full device model and circuit performance benchmarking," *IEEE Transactions on Electron Devices*, vol. 54, no. 12, pp. 3195–3205, 2007.

[6] J. Z. et al., "Carbon nanotube circuits in the presence of carbon nanotube density variations," in *Design Automation Conference*, 2009, pp. 71–76.

[7] B. Ghavami, M. Raji, and H. Pedram, "Timing yield estimation of carbon nanotube-based digital circuits in the presence of nanotube density variation and metallic-nanotubes," in *Intl. Symp. on Quality Electronic Design*. IEEE, 2011, pp. 1–8.

[8] Y. B. e. a. Kim, "A low power 8t sram cell design technique for cnfet," in *Intl. SoC Design Conf.*, 2008, pp. I–176–I–179.

[9] X. Liang and D. Brooks, "Mitigating the impact of process variations on processor register files and execution units," in *IEEE/ACM International Symposium on Microarchitecture*. IEEE, 2006, pp. 504–514.

[10] S. Ozdemir et al., "Yield-aware cache architectures," in *IEEE/ACM International Symposium on Microarchitecture*, Dec 2006, pp. 15–25.

[11] N. Patil and et al., "Circuit-level performance benchmarking and scalability analysis of carbon nanotube transistor circuits," *Nanotechnology, IEEE Transactions on*, vol. 8, no. 1, pp. 37–45, 2009.

[12] N. Patil et al., "Design methods for misaligned and mispositioned carbon-nanotube immune circuits," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 27, no. 10, pp. 1725–1736, Oct 2008.

[13] N. P. Patil et al., "Vmr: Vlsi-compatible metallic carbon nanotube removal for imperfection-immune cascaded multi-stage digital logic circuits using carbon nanotube fets," in *IEEE International Electron Devices Meeting*. IEEE, 2009, pp. 1–4.

[14] M. M. Shulaker et al., "High-performance carbon nanotube field-effect transistors," in *Electron Devices Meeting (IEDM), 2014 IEEE International*. IEEE, 2014, pp. 33–6.

[15] T. Wada, S. Rajan, and S. A. Przybylski, "An analtical access time model for on-chip cahce memories," *Journal of Solid-State Circuits*, vol. 27, pp. 1147–1156, Aug 1992.

[16] H.-S. Wong et al., "Carbon nanotube field effect transistors-fabrication, device physics, and circuit implications," in *Solid-State Circuits Conference, 2003. Digest of Technical Papers. ISSCC. 2003 IEEE International*. IEEE, 2003, pp. 370–500.

[17] Z. Zhang and J. G. Delgado-Frias, "Carbon nanotube sram design with metallic cnt or removed metallic cnt tolerant approaches," *Nanotechnology, IEEE Transactions on*, vol. 11, no. 4, pp. 788–798, 2012.

[18] J. Zhang et al., "Carbon nanotube correlation: promising opportunity for cnfet circuit yield enhancement," in *Proceedings of the 47th Design Automation Conference*. ACM, 2010, pp. 889–892.