2009 IEEE International Workshop on Memory Technology, Design, and Testing

Cost-Aware Lifetime Yield Analysis of Heterogeneous 3D On-Chip Cache

Balaji Vaidyanathan†, Yu Wang‡, Yuan Xie†
†*Department of Computer Science and Engineering, Pennsylvania State University, PA 16801, USA*
‡*E.E. Dept, TNList, Tsinghua University, Beijing, China*
E-mail: {bvaidyan@cse.psu.edu,yu-wang@tsinghua.edu.cn,yuanxie@cse.psu.edu}

***Abstract*—Technology scaling is increasingly yielding diminishing returns in terms of product performance, power, and its yield. Recent development in through-silicon via (TSV) technology has made multi-layer stacking (or 3D integration) a viable solution, opening possibility for coping with the issues related to poor interconnect scaling trend. In this direction there have been research works looking separately at performance, power, and area (or cost) benefits associated with the shift from 2D to 3D manufacturing process for SRAM. However, the poor scaling trend associated with devices still remains as a challenge in realizing large on-chip memories. Heterogeneous 3D integration has been widely adopted for bringing analog, RF, MEMS, DRAM, SRAM, among other wide application on a single chip. In this work, we propose to use heterogeneous 3D integration as an alternative means to manufacture SRAM with multiple technologies. This choice expands the design space that a SRAM designer has thus allowing graceful management of issues related to technology scaling. The main roadblock in realizing 3D integration is the manufacturing cost associated with the TSV process and its yield. Additionally, increased thermal congestion between 3D layers can potentially accelerate many of the reliability mechanisms (gate oxide degradation like Negative Bias Temperature Instability (NBTI)) bringing down the SRAM lifetime yield. Hence to help the system designer understand the overall benefit an integrated on-chip cache analysis flow is implemented to assess the shift from planar to 3D SRAM design under one platform. Our study shows performance, power, cost, and lifetime yield benefit in the move towards heterogeneous 3D cache compared with 2D caches and homogeneous 3D caches.**

I. Introduction

On-chip memory size at few Mega Bytes (MB), occupies 50% and more of total processor footprint [1]. On-chip cache capacity (L2 and L3) has shown an increasing trend to match with the increased core count, and also to alleviate the widening memory-core performance gap. Intel incorporates on-chip L3 cache of 24 Mega Bytes (MB) into its Montecito dual-core Itanium processors [2]. However the increasing on-chip memory size has hindered the performance, power, and yield of caches due to poor technology scaling trend and manufacturing induced process variation.

Recently, there has been a migration in the implementation of on-chip memory using multi-layer stacking process demonstrating performance, bandwidth, power, and heterogeneous integration benefits [3], [4]. Current 3D IC manufacturing process supports wafer-to-wafer, die-to-wafer, die-to-die chip stacking technology using face-to-face (f2f) or face-to-back (f2b) bonding (Figure 1) that enables denser and shorter inter stack communication leading to increased bandwidth and performance [5], [6]. Nho et al. [3] demonstrated 3D Static Random Access Memory (SRAM) architecture that could reduce the SRAM access delay by 1.8x and the active power consumption by 3.4x due to the decreased bit-line capacitance and decoder delay compared with SRAM implemented in 2D.

Previous work [3] on 3D SRAM concentrated on utilizing TSV for harnessing performance and power benefits within a single technology. In this work we identify the diminishing returns for on-chip memories as technology scales. As a result we propose to use heterogeneous 3D on-chip SRAM as an alternative way to obtain the performance benefits from the move towards newer technology while keeping the yield and power issues to minimum by using an older technology with minimal memory capacity impact.

This work was supported in part by grants from NSF 0903432, 0702617 and 0643902. Yu Wang's work is partially supported by grants from 863 program of China (No. 2009AA01Z130), and NSFC (No. 60870001, 90207002), and TNList Cross-discipline Foundation.

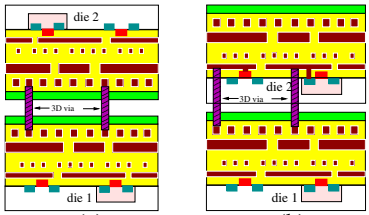

Figure 1. (a) Face-to-face bonding and (b) Face-to-back bonding showing two layer silicon stacks with 3D via or TSV

Poor thermal conduction between layers is an important challenge in the realization of multi-layer stacking [6]. Such increased temperature affects performance, leakage power as well as exponentially accelerates many of the progressive chip failure mechanisms like Negative Bias Temperature Instability (NBTI), gate-oxide Soft Breakdown (SBD), and Electron Migration (EM). More importantly, for the shift towards 3D IC one has to also factor the manufacturing yield associated with TSV, and the layer bonding process along with the known-good-die (KGD) test cost to assess the overall system cost benefit [5]. With the co-existence of both positives and negatives in 3D IC realization, one has to perform an overall assessment considering the system level benefits, as well as the manufacturing cost, and temperature induced yield loss at an early stage of design cycle. Consequently, we build an integrated assessment framework for heterogeneous 3D SRAM analysis, and project its benefits in the performance, power, yield, and manufacturing cost design space.

II. Motivation

In this section we briefly highlight the diminishing performance, power, lifetime yield, and manufacturing cost of on-chip SRAM with technology scaling to motivate the move towards heterogeneous implementation of 3D on-chip cache.

Given the size of on-chip SRAM, one can expect a 6σ variation in the process parameters, which may drive the SRAM cell to operate under worst-case conditions, leading to its reduced operational stability [7]. Further, increased dominance of Drain Induced Barrier Lowering (DIBL) in short channel devices is also considered to be another major source of SRAM stability issue with technology scaling [8]. Figure 2(a) and 2(b) shows the SRAM transfer characteristics and the corresponding read SNM in 45nm and 32nm. A relative decrease in SNM in 32nm is shown compared with the SNM in 45nm due to increased DIBL and also process variability due to reduced cell dimensions.

The pull-up PMOS transistors (PUx) in SRAM experiences NBTI stress during their operational lifetime due to the constant stressing under high temperature and voltage. With technology scaling NBTI worsens due to the poor voltage scaling

978-0-7695-3797-9/09 $26.00 © 2009 IEEE
DOI 10.1109/MTDT.2009.21

and hence increased electric field across the gate oxide. Further, the NBTI induced threshold voltage (Vt) shift in PMOS is shown to have intrinsic random variability [14,17] that is expected to worsen with technology due to decreased transistor dimensions. Additionally with increased SRAM density and decreasing feature size, the margin available for the stability of the SRAM cell is reduced and hence its yield.

The pull-up PMOS transistors (PUx) in SRAM experiences NBTI stress during their operational lifetime due to the constant stressing under high temperature and voltage. With technology scaling NBTI worsens due to the poor voltage scaling and increased electric field across the gate oxide. Further, the NBTI induced threshold voltage shift (ΔV_t) in PMOS is shown to have intrinsic random variability [9], [10] that is expected to worsen with technology due to decreased transistor dimensions. As a result of increased SRAM density and decreasing feature size, the margin available for the stability of the SRAM cell is reduced and hence its lifetime yield.

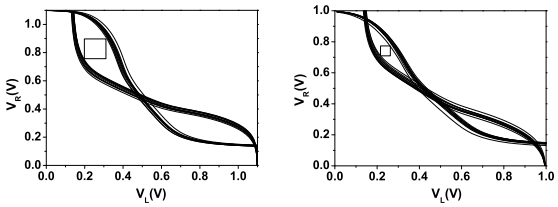

Figure 2. SRAM voltage-transfer characteristics (VTC) showing read SNM (10 HSPICE based Monte Carlo runs) in (a) 45nm, and (b) 32nm (Note: Read SNM denotes the side-length of maximum square nested inside the VTC as shown in the above figure)

SRAM performance improvement with technology has been diminishing with technology due to poor wire delay scaling. The technology road-map for interconnects predicts exponential increase in the wire resistance and poor scaling of the capacitance [11]. In such a scenario, the SRAM access critical path consisting of the word-line, 6T-cell, the bit line and the sense-amplifiers will be increasingly dominated by wire delay compared with the logic delay. Hence we can expect a diminishing trend for SRAM performance as technology scales. With on-chip SRAM occupying 50% of the processor area, one can expect its power mainly the contribution from leakage to be sizable. Additionally, the leakage dominance of SRAM will worsen with technology scaling due to increasing sub-threshold, and gate tunneling leakage. Finally and most importantly, the exponential increase in integration capacity with technology scaling confining to Moore's law has led to an increased challenge in manufacturing and the associated cost.

Advent of 3D IC technology provides a cost effective alternative path for continuing the Moore's Law in the third dimension. However careful selection of the 3D-stack count, die stacking process, and effective KGD testing strategy to overcome stacking yield loss needs attention. Hence architecting a design in 3D with cost-awareness at an early stage is inevitable. The overall technology-scaling trend leads to diminishing performance, power, yield and cost for SRAM. Hence we explore the possible shift towards heterogeneous on-chip 3D SRAM implementation, to obtain a cost-effective ideal SRAM configuration by balancing all the above-mentioned metrics.

III. Evaluation Framework for 3D Cache Analysis

We implemented an integrated analysis flow (Figure 3) to evaluate the performance, power, manufacturing cost, and NBTI induced lifetime yield loss for heterogeneous 3D caches. Detailed explanation for the assessment of each metric is given in the following subsections (III-A to III-D).

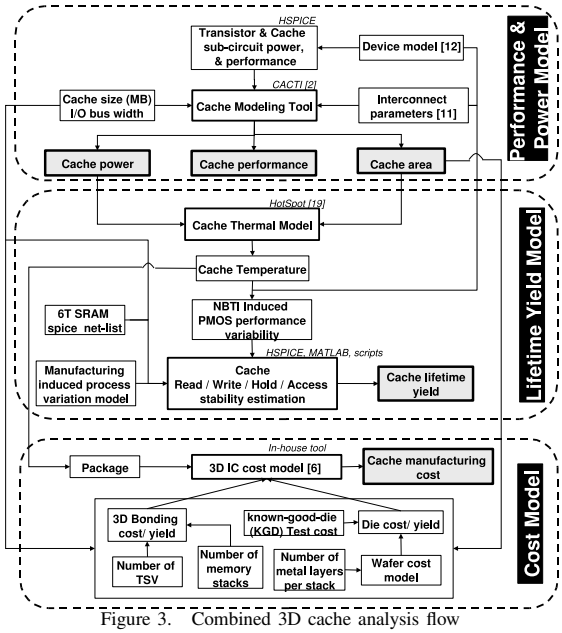

Figure 3. Combined 3D cache analysis flow

A. Cache Performance and Power Model

To assess the on-chip cache area, performance and power, we use a cache-modeling tool CACTI 6.0 [2] at the macro level with transistor models from BPTM modelcard [12] and interconnect parameters referred from ITRS [11]. At the circuit level, we assumed a conventional 6-Transistor (6T) SRAM cell (Figure 4(a)). The timing and power characteristics are derived based on HSPICE simulation using public domain BPTM modelcard [12]. We assumed a conventional layout of the 6T SRAM cell and the associated area model [13]. The SRAM cell transistors (PUx, PDx and PGx) are sized relatively to have a stable cell read, access and write properties [14]. SRAM transistor dimensions are scaled by 0.7x with technology scaling. A conventional SRAM macro (Figure 4(b)) is modeled consisting of decoder, word-line drivers, SRAM array, I/O block, sense-amp, bit-line pre-charge and output circuitry [15].

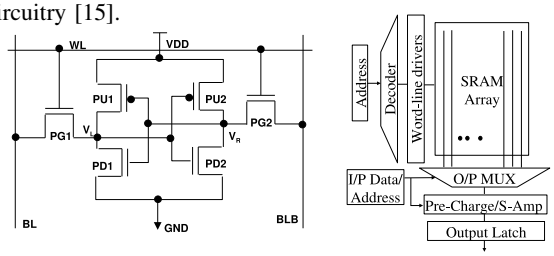

Figure 4. (a) 6T SRAM cell schematic and (b) SRAM macro

B. Manufacturing Induced Process Variation Model

Process variation can be subdivided into global and local variation. Global variation encompasses inter-die, inter-wafer, and inter-lot variation, while the local variation covers the within-die (WID) variations. We consider the effect of only the WID variation that is modeled through V_t (assuming Gaussian distribution) to track statistical SRAM stability. WID variation across technology is tracked using the device dimension dependent time0 V_t distribution [16], [7] shown in (1).

$$\sigma(V_t) = \frac{K2 * T_{Ox}}{\sqrt{A_{GOx}}} \quad (1)$$

,where, T_{Ox} is effective gate oxide thickness and A_{GOx} is its area, $K2$ is a process dependent constant.

C. Intrinsic NBTI Induced Variability Model

NBTI induces PMOS ΔV_t shift when it is negatively biased under high temperature conditions thus potentially risking the SRAM stability [9]. The NBTI induced PMOS ΔV_t shift is considered to be a combination of slow interface trapped charges and fast-hole-trapped charges in advanced technology (2). The slow NBTI induced PMOS V_t degradation (ΔV_{t_it}) is modeled as shown in (3) in accordance with the reaction diffusion theory, while the fast stress behavior that is attributed to the hole-trapping/de-trapping (ΔV_{t_h}) mechanism saturates at low voltages within few milliseconds [17]. At long stress periods slow interface traps dominates aging and hence we consider only ΔV_{t_it} to track NBTI induced PMOS aging in our work.

$$\Delta V_t = \Delta V_{t\_it} + \Delta V_{t\_h} \tag{2}$$

$$\Delta V_{t\_it} = \Delta V_{t0} * e^{A*E_{Ox}} * e^{-E_a/K_B T} * t^n \tag{3}$$

,where fitting parameters (ΔV_{t0}, and A), activation energy (E_a), Boltzmann constant (K_B), stress electric field across the gate oxide (E_{Ox}), operational temperature (T), stress time (t), and time exponent ($n = 1/6$ in accordance with reaction-diffusion theory) are used in modeling the NBTI behavior due to slow interface trapped charges [17]. One has to incorporate also the recovery model to understand the AC (or dynamic activity) behavior of the NBTI induced V_t stress in PMOS transistors. Universal recovery model is used in our analysis as proposed by Kaczer et al. [18] that follows from (5).

$$\Delta V_{t\_AC} = R * r(\xi) + P \tag{4}$$

$$r(\xi) = 1/(1 + B\xi^{\beta}) \tag{5}$$

$$\xi = \frac{1}{DF} - 1 \tag{6}$$

Where DF is the duty factor, B is the scaling parameter and ξ is the dispersion parameter [18]. The total NBTI induced ΔV_{t_AC} shift is considered to be a summation of permanent (P, permanent interface traps) and recoverable (R, recoverable interface traps) component (4), while the $r(\xi)$ describes the duty factor and technology dependence of the recovery as shown in (5) and (6). NBTI AC/DC factor derived from the above stress/recovery models is fed into the spice simulator for circuit lifetime extraction (Figure 5).

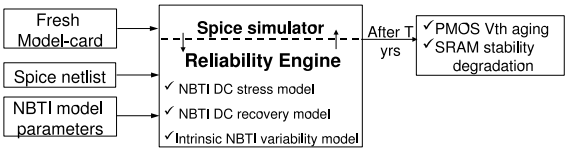

Figure 5. SRAM NBTI simulation setup

To predict SRAM circuit lifetime we statically calculate the V_t degradation at each transistor in SRAM assuming 0.5 static signal probabilities at their inputs. The NBTI induced V_t degradation is incorporated by adjusting the DELVTO parameter in HSPICE (using public domain BSIM4 model-card [12]).

Above-mentioned NBTI model, calculates the mean shift ($\mu(\Delta V_t)$) in device aging. Additionally, in the current and future technology, the intrinsic variability in NBTI (NBTI-induced statistical variation) leads to ΔV_t mismatch in SRAMs that needs to be considered as an additional source of random variation [9], [10]. Thus the intrinsic NBTI induced ΔV_t variability ($\sigma(\Delta V_t)$ modeled using (7)) is incorporated into the aging extraction framework to track the transistor ΔV_t) spread with stress time.

$$\sigma(\Delta V_t) = \sqrt{\frac{K1 * T_{Ox} * \mu(\Delta V_t)}{A_{GOx}}} \tag{7}$$

,where T_{Ox} is effective gate oxide thickness and A_{GOx} is its area, and $K1$ is an empirical constant equal to 1 [10].

D. 3D IC Cost and Thermal Model

Manufacturing cost is sensitive to the early design decision. For example the choice of technology, die area, metal layer count; thermal design package selection based on chip thermal limits can be critical in deciding the manufacturing cost. Manufacturing cost is all the more important to consider in analysis of 3D IC as it affects all the afore-mentioned cost sensitive parameters [5]. 3D IC manufacturing cost is modeled using our in-house 3D IC cost analysis tool that takes into account technology node, die area, metal layer count, die yield, bonding cost (we assume f2b bonding in this work), and KGD test cost among others [5]. Thermal profiling of the 3D Cache is done using a finite-element analysis based 3D Hotspot tool [19] that models both per-layer and interlayer heat flow.

IV. Manufacturing and NBTI Induced Variation Impact on 3D SRAM Yield

In this work, we calculate SRAM stability considering read, write, hold, and access stabilities.

A. Hold Stability Estimation

Hold stability is calculated at minimal supply voltage that ensures SRAM cell data retention at 6-sigma (σ) process corner. The minimal data retention or the hold voltage (V_{hmin}) decides the sleep mode voltage of the SRAM for leakage reduction. Hold failures are tracked by extracting the voltage dependent hold failure probability distribution (using Monte Carlo based HSPICE simulation) and fitted to non-central χ^2-distribution. Subsequently, the voltage (V_{hmin}) corresponding to the hold failure probability that leads to negligible failure at 6-sigma corner (Figure 6) [13].

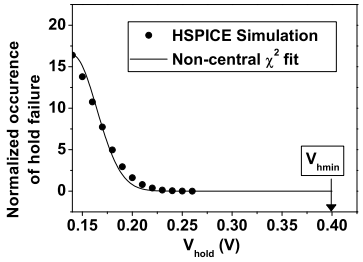

Figure 6. SRAM V_{hmin} estimation (using 32nm BPTM modelcard [12])

B. Access Stability Estimation

The access stability is calculated as the capability of the SRAM cell to drive the bit-line within a specified delay. However due to the variation in SRAM transistor parameters, the worst case (6σ) cell current (that is dependent only on the strength of PDx, and PGx) might be insufficient to provide enough margin for the sense amplifier detection thus creating access failure [13]. Access stabilities are calculated using a quasi-analytical Monte-Carlo based methodology [13] and the access timings are given enough margins to prevent any access failures at time0. SRAM lifetime yield is mainly affected by

NBTI in PMOS transistors. Hence we do not concentrate on access failures, as they are dependent only on NMOS transistors (PGx and PDx) in the 6T cell.

C. Read/Write Stability Estimation

The SRAM cells read (read SRAM cell data without its destruction) and write (successful write to cell within a specified time) stabilities are calculated using a quasi-analytical Monte-Carlo based methodology [14]. Read, and write stabilities are affected both at time0 (due to process variation) and during its operational lifetime (due to NBTI induced ΔV_t perturbation in the SRAM cell transistors). With SRAM operational time, NBTI induced ΔV_t shift and its spread increases, decreasing read stabilities of the 6T cell while enhancing write stability [13].

Read stability is assessed under process and intrinsic NBTI variation using a critical point sampling (CPS) technique introduced by Kang et al. [14] (Figure 7(a) and 7(b)). In CPS technique, three sampling positions (V1, V2, and V3) in the read VTC are chosen and their values are tracked by performing 1000 Monte Carlo based HSPICE simulation of SRAM cell incorporated with process variation and NBTI variation models. Finally, the read failure probability ($P_{c\_read}(t)$) of a SRAM cell is obtained using (8) [14].

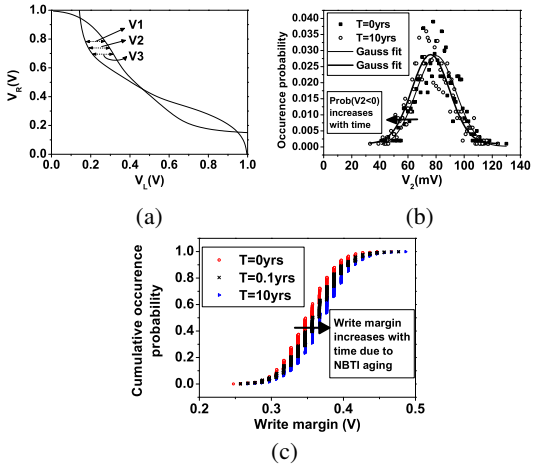

Figure 7. (a) Critical point (V1, V2, and V3) sampling technique based (b) read failure estimation using Gaussian approximation; (c) Write margin increase with time due to NBTI aging (all simulation results in 32nm)

$$P_{c\_read}(t) = P(\{V_1 < 0\} \cap \{V_2 < 0\} \cap \{V_3 < 0\}) \quad (8)$$

Write stability is modeled by assessing the bit-line write margin (BLWM) of the 6T cell under varying process and NBTI parameters using a quasi-analytical Monte Carlo based methodology [14] and model is extended for larger SRAM array size. BLWM is measured as the highest BL voltage needed to flip the cell while keeping WL high and BLB precharged high and BL ramped down from high [20]. Hence, when the cell becomes weak due to NBTI aging with time, we can expect the BLWM to increase as the cell flip is achieved at a much higher BL voltage in its ramp down path (Figure 7(a)). Statistical distribution of BLWM is obtained using 1000 Monte Carlo based HSPICE simulation and fitted to a normal distribution. Thus, the write failure probability ($P_{c\_write}(t)$) is obtained by fixing a specified lower limit for write margin at time0 (BLWM0), and tracking the statistical failure probability of a SRAM cell to meet this margin as shown in (9).

$$P_{c\_write}(t) = P(BLWM < BLWM_0) \quad (9)$$

D. 3D SRAM Yield Estimation

Hold failure probability is kept low by setting a lowest possible hold voltage (V_{hmin}). On the other hand, the access failures (that are not affected by NBTI induced PMOS device degradation) are taken care by setting memory access time to cover time0 6σ variations in cell current. The SRAM cell failure probability (P_c) (10) is thus calculated as a union of time-dependent read ($P_{c\_read}(t)$) and write-failure ($P_{c\_write}(t)$) probabilities [14].

$$P_c(t) = P_{c\_read}(t) \cap P_{c\_write}(t) \quad (10)$$

The calculated cell failure probability ($P_c(t)$) is used in calculating the 2D SRAM array failure probability ($P_{mem\_2D}(t)$) as shown in (14) [14]. The total number of SRAM cells in the memory array is calculated as $(N_{col} + N_{col\_red}) * (N_{row})$, where N_{col}, and N_{row} are number of SRAM cells along the column and row, while N_{col_red} is the number of redundant column cells incorporated into the memory array for yield improvement. The calculated per-layer 2D SRAM array failure probability is extended to calculate the overall 3D memory failure probability ($P_{mem\_3D}(t)$) as shown in (15).

$$P_{col}(t) = 1 - (1 - P_c(t))^{N_{row}} \quad (11)$$

$$P(F(0)) = \sum_{i=N_{col\_red}+1}^{N_{col}} \binom{N_{col}}{i} * P_{col}(0)^i * (1-P_{col}(0))^{N_{col}-i} \quad (12)$$

$$P(F(t)|S(0)) = 1 - (1 - P_{col}(t))^{N_{col}} \quad (13)$$

$$P_{mem\_2D}(t) = P(F(0)) + P(F(t)|S(0)) * (1 - P(F(0))) \quad (14)$$

$$P_{mem\_3D}(t) = 1 - \prod_{i=1}^{N_{layer}} (P_{mem\_2D\_i}(t)) \quad (15)$$

,where $P_{col}(t)$ (11) is time dependent SRAM column failure probability, $F(t)$ and $S(t)$ are time dependent per-layer SRAM failure and success probabilities, and $P(F(t)|S(0))$ (13) is the per-layer memory failure probability given its initial time0 success.

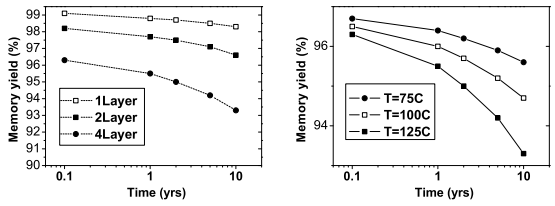

Figure 8. (a) 3D SRAM yield versus layer count with 4MB of memory per layer under 125C of operational temperature (in 32nm technology). (b) 3D SRAM with 4 layer (4x4MB) yield with time under different temperature conditions (in 32nm technology) (Note: We use N_{col}=8192, N_{row}=4096 (for 4MB SRAM), and $N_{col\_red} = 5\% * N_{col}$ based on the memory configuration assumed by Kang et al. [14])

Using the above-derived 3D SRAM failure probability, three main parameters affecting the 3D SRAM array yield namely, the layer count, temperature of operation, and its operational time is captured. Increase of layer count in the 3D SRAM stack comes with higher yield loss (Figure 8(a)) due to the increased probability of at-least one of the layer failing. Increase in temperature exponentially accelerates NBTI mechanism (3) and hence increases SRAM yield loss (Figure 8(b)). And finally the operational time increases yield loss, due to the time dependence (3) of NBTI mechanism (Figure 8(a) and 8(b)). In the 3D memory configuration, though inter-die variation exists

between layers, we do not consider them in our analysis. This is because inter-die or global variation leads to systematic V_t shift in each die that can be effectively reduced by post-silicon body bias tuning [21].

V. 3D SRAM Cost Versus Lifetime Yield Analysis

Overall benefits from different parameters like manufacturing cost, lifetime yield, performance, and power metrics drive the move towards multiple-layer stacking for on-chip caches. In this regard the move towards 3D integration comes with additional cost related to implementation of TSV technology, area increase due to TSV inclusion, and die bonding yield loss. Hence making an early analysis of the manufacturing cost along with other metrics (lifetime yield, performance, and power) is inevitable. Additionally adopting heterogeneous integration of multiple technologies (combining older technology with the newer one) can offset the concerns with the maturity (manufacturing cost, lifetime yield, and power metrics) of the newer technology. Thus gaining performance benefits from the move towards a newer technology while keeping cost, power, and lifetime yield at minimum.

A. Planar versus Homogeneous 3D Caches

To assess the benefits of the combined analysis, we initially take two cache configurations C1 and C2 with planar and homogeneous 3D cache implementation respectively (Figure 9). Cache configurations C1 and C2 implements 32MB of on-chip L2 cache. The cache density doubles with the same die footprint when moving from 32nm to 22nm technology. The area is modeled based on our transistor dimension scaling assumption (refer subsection III-A) plugged into the CACTI 6.0 cache-modeling tool [2] in our combined analysis flow (Figure 3) that includes both peripheral and SRAM array area models. TSV dimension is assumed to be 0.2μm [22] and the associated area overhead is factored in to the area model.

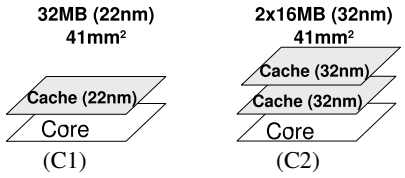

Figure 9. Planar and Homogeneous 3D cache configuration (C1 and C2 respectively) taken for yield versus cost analysis.

The access delay of the 3D cache is set to the layer with the maximum delay in the 3D cache configuration. Total cache power is calculated as a combination of dynamic and temperature dependent active leakage power dissipated in the cache bank that is activated, while rest of the power is calculated from the sleep mode leakage of other banks that are held at low voltage (V_{hmin}). Thermal profiling is done using HOTSPOT tool [19] in which we assume a microprocessor core to exist at the bottom layer in the multi-layered configuration, connected to the heat spreader and heat sink (the core is assumed to dissipate 100 Watts of power in accordance with the current generation processor power ratings excluding the power dissipated by on-chip memory). The power density for the given example is higher in comparison with the current generation micro-processor due to absence of mature power optimization assumptions. However, the conclusions derived based on these power density values will hold valid in a relative scale due to monotonous relation of the cost and yield on the power density. The lifetime yield is obtained using our 3D memory yield calculation setup (Section IV), and finally the manufacturing cost is obtained using our in-house 3D IC cost analysis tool [5].

Figure 10 compares the trend in manufacturing cost, lifetime yield, performance, and power across the chosen cache configurations C1 and C2. Though the cost of manufacturing a 32nm die is lower compared to 22nm, the move from C1 to C2 increases the total cost by 40%. The main reason being the increase in the number of layers in the 3D-stack (increasing the cost related to die bonding) and also the number of dies needed per product. Along with the cost, the yield for C2 also increased compared with C1. This is due to the decrease in power where the SRAM leakage power in 32nm is 2x lesser compared with 22nm for the same die footprint and hence lesser temperature effect on the NBTI induced lifetime yield.

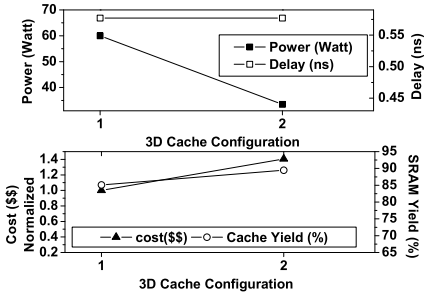

Figure 10. Manufacturing cost, lifetime yield (at 10yrs), performance, and power of 3D Cache configuration.

B. Homogeneous versus Heterogeneous 3D Caches

In this section, we compare the homogeneous with heterogeneous 3D caches and present the benefits of the latter. We analyze a homogeneous 3D cache configuration C3 with 32MB capacity in 22nm, and two alternative heterogeneous 3D cache configurations C4 and C5, implementing 24MB of heterogeneous on-chip L2 cache using both 22nm and 32nm technology (Figure 11). Homogeneous 3D cache configuration C3 is chosen to reduce the chip footprint (that can potentially save on manufacturing cost) while maintaining 32MB of memory to make a fair comparison with C1 and C2. C4 is mainly obtained to gain on performance, power, cost, and yield while taking a 25% hit on the memory capacity (moving from 32MB to 24MB), that is possible with the heterogeneous integration capability offered by 3D IC manufacturing process. Finally, C5 is obtained by swapping the layers of C4, the benefits of which will be discussed later in the section.

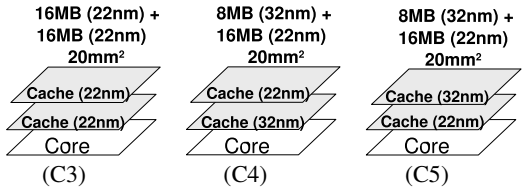

Figure 11. Homogeneous 3D cache configuration (C3) and Heterogeneous 3D cache configuration (C4 and C5) taken for yield versus cost analysis.

The access delay of 3D cache configuration C3 improved by 20% compared with C1 and C2, due to reduction in cache size by half and the associated delay with wire parasitic and decoder delay (Figure 10 and 12). The manufacturing cost of C3 goes down by 20% compared with C1 and 50% compared with C2. This is due to the die size reduction by half, leading to higher number of good-die yield per wafer in C3 compared with C1 and C2. However, the power consumption of C3 goes up by

25% compared with C1, and almost doubled in comparison with C2. This is due to increased power density and hence higher thermal congestion between layers in C3 compared with C1. More importantly, due to higher thermal congestion, the lifetime yield of C3 goes down to 61% from 85-90% for C1 and C2.

The performance and manufacturing cost benefits of C3 comes with high power and low lifetime yield. Subsequently, we analyze the heterogeneous 3D cache configuration C4 and C5 to alleviate the power and lifetime yield issue with C3. Manufacturing cost reduction in C4 (30% compared with C1 and 50% compared with C2) comes mainly from the die size reduction by half, leading to higher number of good-die yield in a wafer compared with C1 and C2. Additionally, manufacturing cost savings of 8% in C4 compared to C3 is obtained by averaging the manufacturing cost in 22nm and 32nm. The access delay of the C4 (same as C3) decreased by 20% going from configuration C1 and C2 to C4 due to the reduction in cache size by half and the associated delay with the wire parasitic and decoder delay. Configuration C4 has power that is 40% lesser than C3, as C4 is a heterogeneous combination of caches in 22nm and 32nm node. As a result of reduced power and thermal congestion, the lifetime yield in configuration C4 goes up to 82.3% from 61% in C3.

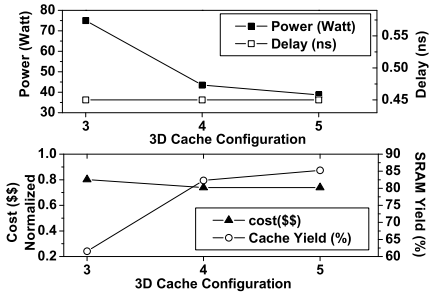

Figure 12. Manufacturing cost, lifetime yield (at 10yrs), performance, and power of 3D Cache configuration (C3, C4, and C5). (Note: Cost values are normalized to the manufacturing cost of configuration C1).

Configuration C5 is obtained by swapping the two layers bringing high power dissipating 16MB cache in 22nm closer to the heat sink while placing the cooler cache (8MB in 32nm) away from heat sink, thus reducing thermal congestion to an extent. Thus configuration C5 achieves 10% lesser power and 3% higher absolute lifetime yield compared with C3. Heterogeneous 3D cache stacking (C4 and C5) provides better performance (compared with C1 and C2); power savings (compared with C1 and C3), manufacturing cost reduction (compared with C1, C2, and C3), and lifetime yield reduction (much better than C3 and comparable to C1 and C2).

VI. Conclusion

A combined platform for 3D SRAM analysis provides an overall insight into the design space that spans lifetime yield, manufacturing cost, performance, and power, to help designer make a cost effective system decision at an early stage of the design. Heterogeneous 3D on-chip caches (manufacturing caches using multiple technology node) provide performance benefits of the newer technology while lowering the power, lifetime, and manufacturing cost issues with it. Using our integrated evaluation framework, we show that heterogeneous 3D caches can be manufactured with 20% lesser cost, while maintaining almost an equivalent lifetime yield, 50% lesser area, 30% lesser power, and 20% higher performance in comparison with the 2D cache. Additionally, homogeneous 3D caches are found to be either highly cost ineffective (2x higher) or lifetime yield limited (25% lower) in comparison with heterogeneous 3D caches. Important conclusion from the study indicates manufacturing cost, performance, power and lifetime yield benefits from the move towards heterogeneous 3D cache compared with 2D caches and homogeneous 3D caches while taking a minimal hit on the memory capacity.

References

[1] C. Molina, C. Aliagas, M. Garcia, A. Gonzalez, and J. Tubella, "Non redundant data cache," in *ACM/IEEE ISLPED*, 2003.

[2] N. Muralimanohar, R. Balasubramonian, and N. Jouppi, "Optimizing NUCA Organizations and Wiring Alternatives for Large Caches with CACTI 6.0," in *ACM/IEEE MICRO*, 2007.

[3] H. H. Nho, M. Horowitz, and S. S. Wong, "A High-speed, Low-power 3D-SRAM Architecture," in *IEEE CICC*, 2008.

[4] X. Dong, X. Wu, G. Sun, Y. Xie, H. Li, and Y. Chen, "Circuit and Microarchitecture Evaluation of 3D Stacking Magnetic RAM (MRAM) as a Universal Memory Replacement," in *ACM/IEEE DAC*, 2008.

[5] X. Dong and Y. Xie, "System-level cost analysis and design exploration for three-dimensional integrated circuits (3D ICs)," in *ACM/IEEE ASPDAC*, 2009.

[6] K. Bernstein, P. Andry, J. Cann, P. Emma, D. Greenberg, W. Haensch, W, M. Ignatowski, S. Koester, J. Magerlein, R. Puri, R, and A. Young, "Interconnects in the Third Dimension: Design Challenges for 3D ICs," in *ACM/IEEE DAC*, 2007.

[7] A. J. Bhavnagarwala et al., "The impact of intrinsic device fluctuations on CMOS SRAM cell stability," *IEEE JSSC*, Apr 2001.

[8] A. Carlson, "Mechanism of Increase in SRAM Vmin Due to Negative-Bias Temperature Instability," *IEEE TDMR*, Sept 2007.

[9] S. E. Rauch, "Review and Reexamination of Reliability Effects Related to NBTI-Induced Statistical Variations," *IEEE TDMR*, Dec 2007.

[10] S. Pae, J. Maiz, C. Prasad, and B. Woolery, "Effect of BTI Degradation On Transistor Variability in Advanced Semiconductor Technologies," *IEEE TDMR*, Sept 2008.

[11] http://www.itrs.net/Links/2008ITRS.

[12] http://www.eas.asu.edu/˜ptm.

[13] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in Nanoscaled CMOS," *IEEE Tran. On CAD of ICAS*, Oct 2005.

[14] K. Kang, H. Kufluoglu, K. Roy, and M. A. Alam, "Impact of Negative-Bias Temperature Instability in Nanoscale SRAM Array: Modeling and Analysis," *IEEE Tran. On CAD of ICAS*, Oct 2007.

[15] M. Margala, "Low-Power SRAM Circuit Design," in *IEEE MTDT*, 1999.

[16] J. A. G. Jess, K. Kalafala, S. R. Naidu, R.H.J.M, Otten, and C. Visweswariah, "Statistical Timing for Parametric Yield Prediction of Digital Integrated Circuits," *IEEE Tran. On CAD of ICAS*, Nov 2006.

[17] A. E. Islam, H. Kufluoglu, D. Varghese, S. Mahapatra, and M. A. Alam, "Recent Issues in Negative-Bias Temperature Instability: Initial Degradation, Field Dependence of Interface Trap Generation, Hole Trapping Effects, and Relaxation," *IEEE TED*, Sept 2007.

[18] B. Kaczer, T. Grasser, P. Roussel, J. Martin-Martinez, R. O'Connor, B. O'Sullivan, and G. Groeseneken, "Ubiquitous Relaxation in BTI Stressing-New Evaluation and Insights," in *IEEE IRPS*, 2008.

[19] W. Huang, K. Sankaranarayanan, R. J. Ribando, M. R. Stan, and K. Skadron, "Accurate, Pre-RTL Temperature-Aware Processor Design Using a Parameterized, Geometric Thermal Model Considerations," *IEEE Tran. On Comp.*, Sept 2008.

[20] Z. Guo, A. Carlson, L. T. Pang, K. Duong, T. J. K. Liu, and B. Nikolic, "Large-Scale Read/Write Margin Measurement in 45nm CMOS SRAM Arrays," in *IEEE Symp. On VLSI*, 2008.

[21] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "Reduction of Parametric Failures in Sub-100-nm SRAM Array Using Body Bias," *IEEE Tran. On CAD of ICAS*, Jan 2008.

[22] Y. Xie, G. H. Loh, B. Black, and K. Bernstein, "Design Space Exploration for 3D Architectures," *ACM JETC*, Apr 2006.