Dynamic Fine-Grain Body Biasing of Caches with Latency and Leakage 3T1D-Based Monitors

Shrikanth Ganapathy† Ramon Canal† Antonio Gonzalez‡ † Antonio Rubio§

†Department d'Arquitectura de Computadors
Universitat Politècnica de Catalunya
Barcelona, Spain
{sg,rcanal}@ac.upc.edu

‡Intel Barcelona Research Center
Intel Labs-UPC
Barcelona, Spain
antonio.gonzalez@intel.com

§Department d'Enginyeria Electrònica
Universitat Politècnica de Catalunya
Barcelona, Spain
antonio.rubio@upc.edu

***Abstract*—In this paper, we propose a dynamically tunable fine-grain body biasing mechanism to reduce standby leakage power in first level data-caches under process variations. Accessed physical arrays are forward body biased (FBB) to improve latency while idle (unaccessed) arrays are reverse body biased (RBB) for reducing standby leakage power. The bias voltage to be applied is computed at design time and updated at run-time to counter the negative effects of process variations. This ensures that under all scenarios, the cache will consume the lowest leakage power for the target access latency computed at design-time. A sensor-like hardware mechanism measures the variation in latency and leakage at run-time and this measurement is used to update the bias voltage. The backbone of the hardware used for measurement is a three-transistor one-diode(3T1D)DRAM cell embedded into a regular cache array. By measuring the access and retention time of the 3T1D cell, we show that it is possible to classify cache arrays based on run-time latency/leakage profiles. Our technique reduces leakage energy consumption and access latency of the cache on an average by 20% & 18% respectively. Finally we show that our technique will improve parametric yield by a maximum of 38% for worst-case scenario.**

I. Introduction

Tremendous advancements in chip design have made possible billion-transistor integration over the last decade. This can be largely attributed to the improving capabilities of the manufacturing processes. However, manufacturing has improved only to such extent that the problem of spatial variability of transistor parameters is inhibiting the power/performance gains achieved by scaling devices. Increasing power densities is another major cause of concern for high performance/low power designs. Power is dissipated in the form of heat leading to increased heat densities. With increase in temperature, the leakage power increases exponentially. Recent study has shown that operating temperature of chips can be as high as $90\,^{\circ}\mathrm{C}$ and in some cases as high as $120\,^{\circ}\mathrm{C}$ [1]. Frequent temperature shootups can result in functionality problems and cause permanent damage in the form of faults due to electromigration, thermal cycling & stress migration [2]. These faults can have a long lasting effect on the processor performance over the lifetime of the chip. While design-level techniques can be used to meet certain constraints by improving tolerance to process variations, it is impossible to design considering worst-case temperature or power (in-turn leakage & delay) conditions owing to reduced yield and revenues.

As caches are a very important component from an area point of view, it becomes extremely challenging to optimize chip yield keeping in mind the effects of spatial variations of process parameters and temporal variations of temperature & power. Recent proposals [3], [4], [5] have suggested that post-silicon adaptivity can be used effectively to improve parametric SRAM yield and also reduce power consumption significantly. Post-silicon adaptivity involves detecting changes in low-level circuit parameters (delay & leakage currents) post-manufacturing using on-chip canary structures and providing recovery circuits for effective repair. Body biasing (BB) is one such technique. The threshold voltage of the transistor dependent on body-source potential is modulated to improve performance or reduce power (leakage). In forward body biasing (FBB), application of positive bias voltage reduces threshold voltage making transistors faster but at the cost of increasing leakage. In reverse body biasing (RBB), a negative voltage increases the threshold making transistors slower and also less leakier. Tolerance to process variations can be improved by utilizing both RBB and FBB and this is called adaptive body biasing (ABB). Based on critical path delay measurements obtained at manufacturing time, bias voltages (either FB or RB) are set permanently for the lifetime of the chip. While this would greatly reduce the impact of spatial variations (process), susceptibility to temporal variations increases. Further, chip-wide body biasing does not take into account the effects of within-die variations. In order to reap maximum benefits would require fine-grain control by measuring both latency and leakage local to a particular block at run-time and applying an optimal body bias that trades off power for performance. This is called dynamic fine-grain body biasing (DFGBB) [6]. In essence, this is a 2 step mechanism that requires a sensor like unit (to measure the latency/leakage) to be interfaced with a body bias control unit for generating an optimal bias voltage based on the measurements.

The paper makes the following contributions,

1.) As a first step, we present a novel three-transistor one-diode (3T1D) DRAM-based latency/leakage measurement hardware specially targeted towards memory structures such as register files & caches. By embedding a 3T1D into a regular sram array, we show that each read (or write) to the 3T1D cell will suffer almost the same variation on access power and latency when compared to any sram cell in that array (since it will use the same periphery circuits and the physical variations will be almost identical between the cells due to their proximity). The retention and access time of the embedded 3T1D-DRAM cell are measured to determine the effects of process variation on leakage and latency respectively. Because of the transient nature of both latency and leakage, the mechanism behaves well in tracking temporal variations.

2.) The measurement hardware is then interfaced with a modified version of the look-up table based adaptive FBB generator [7]. In addition, a hybrid charge pumping circuit is used for generating the negative voltage required for RBB [8]. By exploiting the unique access patterns that caches exhibit, active arrays are forward biased while inactive/unused arrays are reverse biased in a very speed effective manner. Not only does this offer enhanced access speeds (forward biasing), tremendous leakage power reduction is made possible by reverse biasing multiple unused arrays.

This paper is organized as follows. Section 2 makes a review of existing work in the literature. In section 3, we discuss about the 3T1D cell and its performance in the presence of variations. In section 4, we propose a new hardware for classifying cache arrays based on latency/leakage. Section 5 presents the dynamic fine-grain body bias generator that is interfaced with the hardware presented in section 4. In section 6, leakage/latency improvements of the proposed scheme are analysed. Yield estimates are also presented. Section 7 presents the concluding remarks.

II. RELATED WORK

Chris *et al.* [9] propose a 2-level technique for reducing leakage in 6T-SRAM arrays. The cells are optimized for High-V_{th} using work function modification and at run-time forward body biasing is used for reducing the access latency. Results indicate a 64% leakage reduction when compared to conventional techniques without significant performance loss. However, a constant 500mV forward bias can have adverse effects on the source-body junction currents. In [5], latency of slow (parametric failure) critical paths is improved by boosting wordline voltage. Failing wordlines are tested during manufacturing and using an EEPROM, failure information is stored. For the lowest area overhead & boosting by 1V this technique enhances yield by 12% under worst-case process variations. The resulting dynamic power overhead is 37%. Though the technique is effective in enforcing optimization on a per-line basis, it assumes variation in latency as the only constraint while estimating parametric yield. This could be particularly misleading in sub-65nm designs where leakage is an important design parameter. In [10], Das *et al.* have proposed a new cache redundancy scheme called substitute cache that replicates data from cache lines affected by process variation. The only priority for replicating cache words in the redundant cache is to replace lines with very high latency. This technique also does not assume leakage to have equal priority as this can change by orders of magnitude under the effects of process variations. Singh *et al.* have partitioned the SRAM array into blocks of different voltage groups to account for intra-array variations [3]. While partitions that are very slow are connected to high V_{dd} lines, the remaining have lower voltage levels for power saving. This technique characterizes the spread of spatial variability using empirical results that may/may not correlate with on-chip measurements.

III. 3T1D CELL

Alternatives to 6T/8T SRAM based memories have been researched diligently for want of increased memory density and lower vulnerability to variations. One such proposal is the 3T1D cell proposed by Luk *et al.* [11]. The capacitorless DRAM cell stores the data using a gated diode that is tied to the read-wordline as shown in Figure 1. The 3T1D unlike 1T DRAM memory provides non-destructive reads and access speeds comparable to that of standard SRAM cells. When compared to regular SRAM cells, the transistors of the 3T1D can be asymmetrical in strength. This has 2 fold advantages: Primarily, process variations causing device mismatch are likely to cause less failures to the cell [12]. Secondly, it improves the overall stability making it radiation hardened. Data is written into

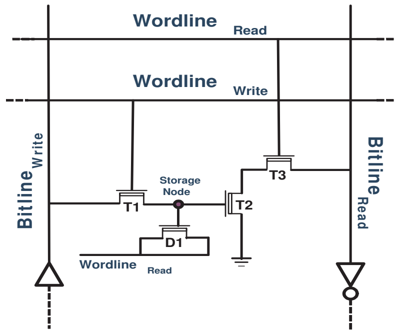


Fig. 1: Schematic of the 3T1D cell.

the cell by raising the write-wordline high and charging the bitline. The voltage level at the storage node is degraded and roughly about $0.6V_{dd}$. A strong T1 would further degrade the voltage resulting in lower retention time. This can be avoided by increasing the threshold of the write driver [11]. The read operation is initiated by precharging (to V_{dd}) the read-bitline and strobing the read-wordline. The retention or storage time of the cell can further be increased by holding the read-wordline at a negative voltage during idle state. This has shown to increase the retention rate by as much as 40X [11]. Liang *et al.* have proposed a 3T1D-only cache architecture that offers SRAM-like performance but better robustness to process variations [13]. In this paper, we interleave the 3T1D cells in the memory arrays to use them as latency and leakage sensors as we will explain in short. The arrays keep the original SRAM cells for program execution.

A. Retention & Access Time

For the sake of comparative study, a 3T1D and a 8T cell are embedded into the same array sharing the same wordline as shown in Figure 2. The 8T cell is a single ended 1R/1W ported cell found in register files. With minor modification to the column periphery, the 3T1D can be embedded into a conventional 6T-SRAM (dual-ended) based array. The access latency of both cells is measured

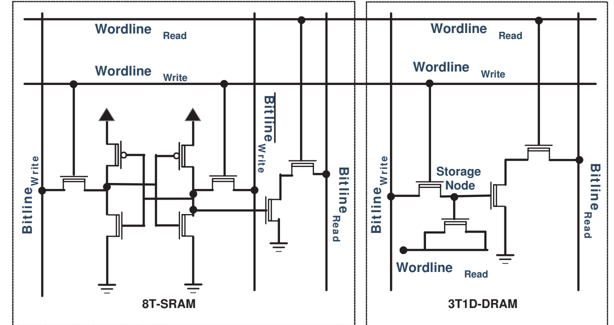


Fig. 2: 3T1D-DRAM embedded with a 8T-SRAM.

independently and normalised to value at 30 °C. Looking at Figure 3, only at high temperatures, the 8T cell is more prone to performance loss when compared to the 3T1D. As opposed to regular 8T cells, the 3T1D(s) are designed for single ended sensing. This combined with T2's (ref Figure 1) boosting action provides very high read speeds even at high temperatures. In addition to providing access speeds comparable to that of regular SRAM cell, the 3T1D with careful sizing can be made to outperform a regular SRAM cell for nominal-case process variations. It is the functional equivalence between the 2 cells that can be used to garner meaningful data about their structural disparities (process variations). While a regular

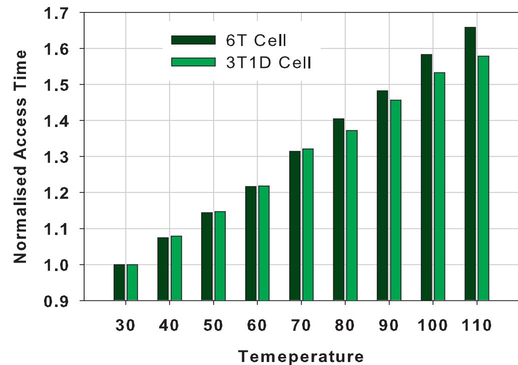


Fig. 3: Access time variation with temperature.

SRAM cell could already be used for measuring access latency with a suitable interface hardware, the 3T1D provides an extra measurable parameter called retention time. The retention time of the 3T1D is defined as the time taken for the voltage at the storage node to decay past $V_{dd}/4$. In [14], it is shown that the leakage through the cell is directly proportional to the retention (decay) time of the cell. In other words as leakage increases, the retention time decreases and vice-versa. Figure 4 shows the measured retention time for 500 cache samples simulated for spatio-temporal variability. The retention time

is normalised to the lowest retention time at 110 °C. The samples are organised in order of reducing magnitude of retention time. Retention time with zero-variability at 30 °C is found to be 9.3μs. Under the presence of process variations, operating at 30 °C, the retention can be as high as 34.2μs or as low as 5μs. Because of the exponential relationship between leakage and temperature, the retention time can be as low 980ns at 110 °C under worst-case process variations. It should be clear from the above argument that both access & retention time of the 3T1D are an important figure of merit that when measured properly can be used to reflect the memory array's latency and leakage run-time profiles under the effects of spatio-temporal variability.

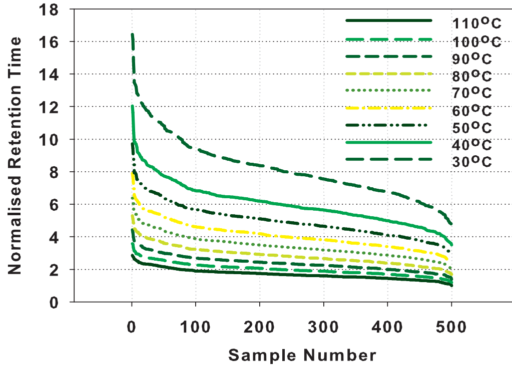


Fig. 4: Retention time variation with temperature.

B. Simulation Setup

The simulated cache is 32KB in size with multiple 1KB arrays. Each array is organized into 128 columns by 64 rows with a 32 bit read-out. Due to area constraints, the decoders are designed with dynamic CMOS and column multiplexer is tree-like design. 500 samples of the cache are simulated on HSPICE with 45nm PTM [15]. We modify only the periphery of one column to accommodate the single-ended 3T1D-DRAM cell in a regular dual-ended 6T-SRAM array. The area associated with a single 3T1D cell for 45nm technology is approximately 0.45μm^2 [13]. Assuming there is one 3T1D cell per row of SRAM, the associated area & energy overhead is estimated as 0.31% and 0.78% respectively. Because random variations are known to affect SRAM cells more than systematic variations and there is no definite way of tracking random variations, one 3T1D per whole array is sufficient to monitor run-time variations. For modelling process variations, we adopt the quadtree based multi-level partition scheme [16]. The σ for systematic and random variation of V_{th} is 6.4%. It is assumed that variances of intra-die systematic and random variation to be equal. Due to the strong correlation between parameter values in deep sub-micron technologies, it is assumed that variance of L_{eff} to be half of that of V_{th} [17]. The systematic and random variations of L_{eff} is derived as 3.2%. Inter-die variations of both parameters is set to an offset value of 3%.

IV. Latency/Leakage Measurement

A. Run-Time Classification of Cache Arrays

The purpose of classifying cache arrays based on latency/leakage profiles at run-time is very much alike calibration. Calibration enables to rectify from any deviations that arise out of manufacturing or during the lifetime of the chip (i.e. degradation). Most often run-time circuit-level optimization like body & source biasing, supply voltage minimization that have been proposed for leakage minimization, are enforced without this available information. Such optimizations resulting from homogeneous heuristics have been enforced across chips having different levels of process variations yielding non-uniform benefits. For any optimization that needs to extract maximum benefits using the available on-chip leakage/latency measurements, the granularity of the measurements should enable discretization of circuit macros which have different levels of process variations that manifest as a variation in latency and leakage as shown in Figure 5. A very high access time and low retention translates directly to high access latency and significantly high leakage power. Design choices resulting in such worst-case corners should be avoided at any cost. For the sake of simplicity, we would like to call each discrete combination of measured leakage and latency as a *bin*. The nomenclature used (min,low,high,max) is specific to our scheme and is not representative of the actual degree of separation. It is well established that both latency and leakage are transient and by generating this table-based data, on-die registers can be frequently updated with this information to be made available for cross-layer optimizations. By making the latency/leakage bounds more tight during classification, circuit optimizations can have more fine-grain control by having better cognizance of the power/performance profile of each memory array.

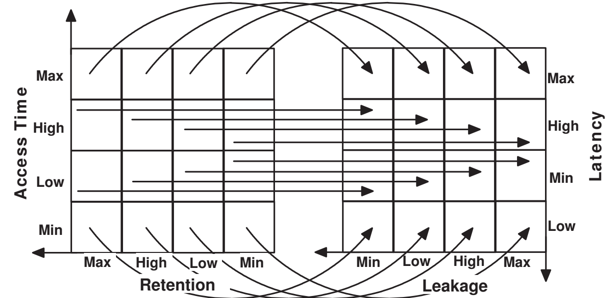


Fig. 5: Discrete classification based on measured latency/leakage profiles.

B. Discretization Architecture

As temperature has a more observable effect on the retention time when compared to access time, we begin with the classification based on leakage. A simple circuitry to measure the retention time involves the use a delay-to-pulse circuitry to count the number of cycles the voltage corresponding to a '1' at the storage node of the 3T1D takes to decay by sending a continuous stream of read requests and waiting till the voltage degrades completely [14]. This has 2 fold disadvantages. It was earlier shown that retention time is of the order of tens of μs. This means that it would require hundreds of thousands of cycles for a counter with a low pulse-width clock to complete the operation. As temporal variations can occur at a frequency of few nanoseconds, recovery mechanisms are designed to have very fast response in the order of thousands of cycles. Further, when the threshold of diode D1 (ref. Figure 1) is lower, the decaying rate reduces thereby increasing the time to drop below the reference voltage. This is exacerbated by the decaying of the voltage at the storage node with subsequent read accesses rendering the pulse circuitry inefficient for high speed operation. During decaying action of the diode in the forward biased mode, initial period of decay is very fast. With further reduction in the voltage at the storage node, the rate reduces as a result of increasing diode resistance. Thus, it is sufficient to measure the drop in voltage for the first few hundred nano seconds rather than wait for complete decay. This decaying behaviour of the storage node is replicated at the output of the sense amplifier. The proposed hardware will have to consider the difference in decay times between adjacent leakage bins and capable of amplifying even small differences into a clock pulse of one cycle. Any scheme that involves a delay-to-pulse circuitry can generate a clock cycle for every period that the output of the sense amplifier is held high [18]. Our proposed leakage-bin classification architecture is shown in Figure 6. The output of the cache array (sense amplifier) is linked to an adder which has the feedback of a

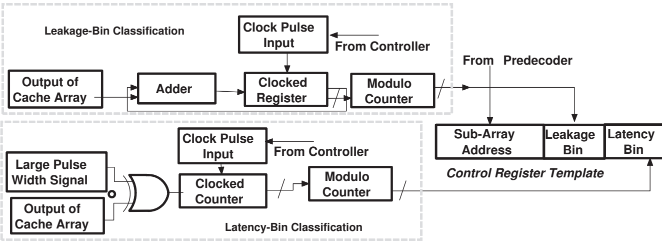


Fig. 6: Hardware based latency/leakage bin classification based on measured retention/access time.

clocked register. The register is clocked at a frequency bounded by the pulse width of the minimum difference between any 2 adjacent bins. This difference along with total number of bins can be computed at design-time for different process corners. With reducing number of bins, the bounds of leakage within which an array is placed into a bin is very loose. In other words, with minimum number of bins for classification, those arrays that have very different leakage profiles (reflected by the retention times of the embedded 3T1D) have a very high probability of being placed in the same bin. The bin selection procedure is initiated by writing a 1 to the 3T1D cell and signalling a read access and constantly strobing the read-wordline high. The output of the sense-amplifier after a given period begins to decay. As long as the output of the sense-amplifier is high enough to signal a 1, the adder increments the value of register by a 1 at the clock rate of the input pulse. The register is incremented at a predetermined frequency whose clock period is low enough to make sure adjacent bins exhibit a difference of at least 1 cycle as shown in Figure 7. It is

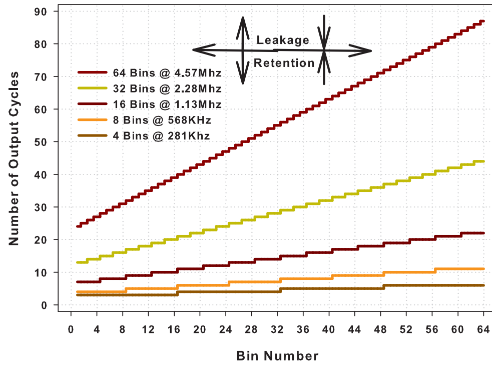
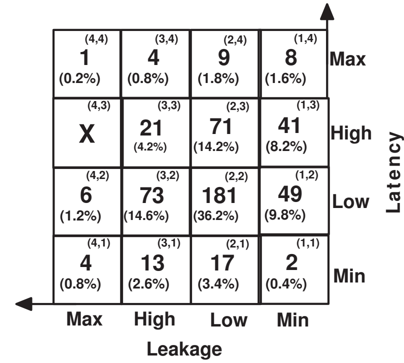


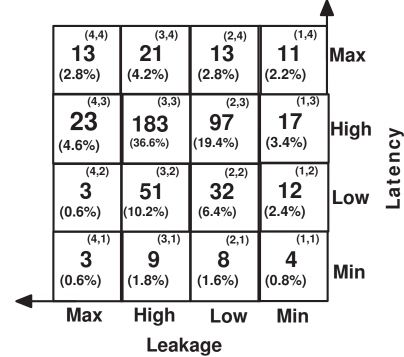
Fig. 7: Number of output cycles corresponding to leakage-bin field. (The number of output cycles (modulo) target number of bins is the value that is written into the control register.)

clearly observable from Figure 7 that the cycle count increases with the bin number. This is in direct relation to the fact that reducing leakage along bin number corresponds to increasing retention times which is reflected by the increase in cycle count along the x-axis. It can be seen that the clock frequency used for 64-bin classification is 16X higher than the frequency used for 4-bin classification. Thus, a linear relationship is exhibited between the number of bins to be used and the frequency of the clock. Some researchers may argue that it is not a viable option to have a frequency divider for bin-classification purposes. The simplest solution would be to design for a frequency that would cater to the maximum number of bins, for instance 64. For classification based on lesser number of bins, say 8, grouping is performed by placement of arrays into bins which are multiples of 8. This is made possible by using a modulo counter. Thus, in a 8-bin classification using 64-bins, bin-8 (in 64-bin) would represent bin-1 (in 8-bin) and bin-16 represents bin-2 and so on. This can be seen in Figure 7 where for a 4-bin classification using 64-bins, all the arrays that have their bin-number lower than 16 produce a 1 cycle output (lowest retention/highest leakage), and all those between 16-32 have 2 cycles output and so on.

In order to classify based on latency, we determine the time to read access the 3T1D cell, corresponding to the critical path delay. Under the impact of spatial variability, for a set of 500 samples, the access times have been found to vary between 14-18% when maintained at a fixed ambient-temperature. This translates to a difference of about 400ps between the slowest and fastest arrays. In effect, the separation between adjacent bins can be as low as 6ps. As a result, for 64-bin classification, even multi-GHz frequencies cannot produce a clock whose period is 6ps. Thus for multi-MHz frequencies, the maximum target-number of bins is 4. The procedure to measure delay in terms of cycle count is similar to the mechanism proposed in [18]. A signal with a very large pulse width is XOR'ed with the output of the cache array. The clocked counter starts incrementing on enabling the control signal to initiate reading a '1' from the 3T1D. As long as the output of the sense-amplifier is 0 and large-pulse width signal high, the counter is incremented for every cycle of the input clock. As soon as the output of the sense-amplifier reaches a high, the counting stops.



(a)



(b)

Fig. 8: (a) Runtime binning of whole cache at ambient temperature(30 °C). Numbers in each box indicate total & percentage number of caches in each bin. (Inset) (x,y) represents x output cycles obtained using leakage-bin classification hardware and y output cycles using latency-bin classification hardware. (b) Runtime latency/leakage binning of whole cache at 110 °C.

As within-die and random variations can have an observable effect on the proper functioning of the discretization hardware, it is imperative that on-chip calibration techniques be enforced. By sampling the output of both leakage and latency classification hardware under ambient conditions, the values can be stored in a master control register that can be used for reference purposes. This stored value can then be compared to the values obtained at design time for different process corners to gauge the impact of process variability. However, this is a cumbersome process when considering the effects

of degradation where the obtained delay is skewed from the reference value and needs self-calibration. This can be avoided by having multiple low area overhead 3T1D cells sharing the same interface hardware distributed across the array and disabling those cells that deviate from the mean latency and leakage bin number of the whole array [14]. It is safe to assume that all 3T1D cells inside a given array will have only different levels of random variations and systematic correlated variations will almost be the same. The effects of random variations manifest as a minimal difference in retention times or access latency across different 3T1D cells sharing the same interface hardware. The resulting shift in sampled output, in very rare cases, cause an erroneous binning.

For a fixed supply voltage of 1V and 500 cache samples, the classification was performed for 4 binning levels of latency and leakage. A cache is placed into a respective bin after measuring the retention and access time of the 3T1D embedded in each array and assuming the leakiest and slowest cell across all arrays is representative of the power and performance profile of the whole cache. It is strange that no cache has been placed in the high latency, maximum leakage bin as shown in Figure 8a. This phenomenon is characteristic to our single frequency grouped-levels binning methodology. As the 4-bins have been approximated by scaling the 64-bin classification, the bounds of each bin are loose resulting in misplacement of high latency, maximum leakage caches across the 3 immediate neighbouring bins along its Cartesian co-ordinates. By re-running the simulations adjusting the input-pulse frequency specific to 4-bin classification, a considerable number of caches were categorized into the high latency, maximum leakage bin. Assuming we consider all caches that have latency and leakage greater than *high* as yield loss, hardly 50% of caches are accepted. Further the presented yield estimates in Figure 8a hold true only when the cache is operating at nominal temperatures. Common phenomena such as sudden temperature shootups can result in the performance going from high to low and in some cases to minimum. The problem is compounded by increasing leakage with temperature. It is clearly observable from Figure 8a that number of caches placed in the low-latency low-leakage bin at $30\,^{\circ}\mathrm{C}$ shifts diametrically to the high-latency high-leakage bin at $110\,^{\circ}\mathrm{C}$ as shown in Figure 8b. Future DVS/DTM techniques can then use such available information at a more finer granularity for triggering mechanisms to lower temperature and simultaneously monitor energy-delay trade-off associated with the enforced optimization. In the next section, we will discuss as to how we can exploit these available measurements for standby leakage reduction and parametric yield improvement using dynamic fine-grain body biasing.

V. Applying Fine Grain Body Biasing

It was shown in [9] that reduction in leakage power is possible by optimizing the 6T-SRAM cell at design-time for high V_{th} and applying a large forward body bias at run-time to compensate for the increased latency. In other words, the array is purposefully designed for high latency (and low leakage) and is made to run faster during operation. This would mean that a large forward bias is applied irrespective of whether the array meets the required timing or not. From a statistical standpoint, both latency/leakage can be on either side of target design value as a result of process variations. Hence no forward biasing is required for those arrays that already meet both leakage and latency targets. It is this very non-determinism that we would like to exploit in order to generate optimal bias voltages dependent latency/leakage measured at run-time.

We propose to use a modified version of the look-up table-based adaptive forward body biasing mechanism designed in [7]. A global decoder inside the cache receives the address of the block to be accessed from the address buffer. Without loss of generality, we assume that the DFGBB generator receives this address at the same time. The index bits corresponding to the array address are decoded and obtained prior to the access. As shown in Figure 9, this decoded address is then referenced inside a look-up table to obtain a respective codeword. This codeword corresponds to the lowest forward bias voltage for which the desired cache access latency is met. As the LUT holds only the codewords and not whole index addresses, a comparator checks the decoded address with all control registers to obtain the correct latency bin value to be referenced. This LUT is defined at design time and can be stored in a small on-chip EEPROM. The codewords in the figure are only indicative and do not represent the actual bias voltages.

The FBB generator consists of four components - decoder, level shifter, demux & resistor tree. The resistor tree is used for generating the forward bias voltages. The resistor tree consists of series connected transistors acting as a potential divider. The number of transistors divide the range (VDDH-VDDL) into intermediate voltages that can be generated at nodal points of connection. The increase in leakage and forward source-body junction current limits the maximum source-body potential to 400mV in forward bias mode [19]. We assume the 20 series connected transistors will generate intermediate voltages of 100-400mV by connecting access switches to 4th,8th,12th and 16th transistors with a resolution of 100mV. The resolution (least 20mV) can be lowered by increasing the number of access switches that result in increased decoder area. The decoders are used to select the correct combination of switches to generate a *vbs* corresponding to the value obtained from the LUT. The generated bias voltages are then routed to the correct array using the demultiplexer.

A hybrid charge pump is used to generate negative bias for RB biasing inactive (unaccessed) arrays. Each of the N arrays require 3 amplifiers - one each for FBB & RBB to boost the body voltage to a level sufficient to bias the entire array and one (optional) for enabling/disabling sleep mode. Only one of Log_2N output lines is high (corresponding to the array being accessed) during any given access. This line is used as an enable signal for the forward bias amplifier to speed-up access while the remaining Log_2N-1 lines are low. Inverting these remaining lines act as an enable signal for the RBB amplifiers which route the RB signal only to inactive arrays. This ensures that the FBB and RBB amplifiers corresponding to a given array operate in a mutually exclusive manner reducing the transition latency (from RBB to FBB and vice-versa) tremendously. It was shown in [9] that if an array is accessed in a given cycle then it is likely to be accessed in the immediate next cycle and those that are idle are expected to remain idle for a considerable amount of time. This phenomena called temporal locality of reference, can be exploited to forward bias those arrays that are currently being accessed and reverse bias those that are idle. It also eliminates the need to regenerate the same FBB voltage on a per-cycle basis by constantly referencing the LUT. As a result, RBB generator needs to be aware of the idle arrays for a large number of cycles. Because it receives the address of the to-be accessed array only once, the state of idle arrays needs to be stored. An extra latch stores the state of all inactive arrays for enabling/disabling RBB mode. In addition to hiding the transition latency in a very time-effective manner, the transition energy involved in switching between RBB and FBB is also reduced significantly.

A. Associated Overhead

The only downside of body biasing is that it requires separate n-wells of whole arrays to be isolated from each other to improve immunity to substrate biasing. Modern day triple well processes offer this option at an increased area overhead. Techniques to improve immunity to substrate biasing include - providing low overhead control circuits to bias wells individually [20] or routing bias lines through upper layer metals [9]. It was shown in [19] that, when considering a multiple bias voltage design, the total area associated

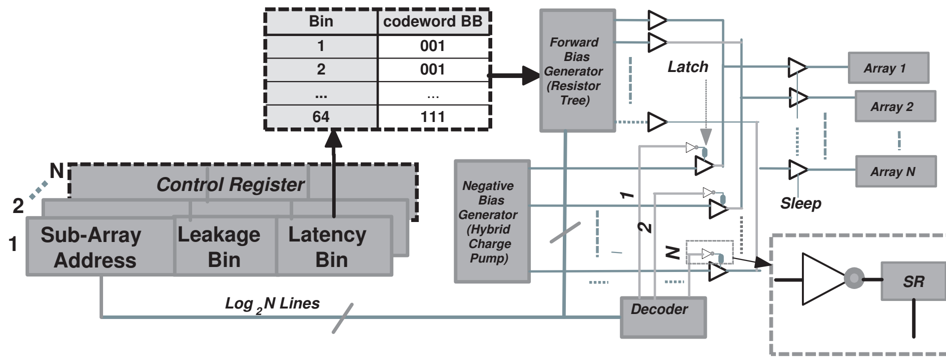


Fig. 9: Dynamic fine-grain body bias generator for caches.

with bias generation, routing and buffers is less than 2-3% of the die-area. Scaling the size of the bias generation unit reported (and used in our proposal) in [7] to 45nm technology, and adjusting for non-linear scaling of cell dimensions, we incur an additional overhead of 1.2%.

As explained before, the transition time is very minimal considering that separate generators for RBB'ing and FBB'ing are used and the pair of amplifiers corresponding to individual physical arrays function in a mutually exclusive manner. The only overhead that is expected is the extra delay in referencing the LUT for the first time an array is being accessed. This can be avoided by using effective way/set prediction techniques. This analysis though, goes beyond the scope of this work.

VI. Experimental Results

In accordance with the analysis presented in [6], BB voltages range from a minimum RBB of -500mV to a maximum FBB of 400mV. On a per-access basis, only one array is active and the remaining are inactive. This is the closest representation of the actual architectural state of the entire cache.

A. Leakage & Latency Reduction

Looking at figures 10(a) & (b), both leakage & latency are a very strong function of the bias voltages. The percentage savings are derived after taking into account energy consumption of bias generator and increase in leakage due to forward biasing of active arrays. The minimum and maximum values correspond to the lowest & highest savings observed for one array among all arrays (WID) of a cache across all samples (D2D). The average is the lowest of the arithmetic mean obtained for all arrays of a cache (WID) across all samples (D2D). It can be seen that the minimum average savings in energy is 12% (-0.1V) and the maximum is 24% (-0.5V). For -0.3V it is 20% and the improvement in energy savings is minimal for voltages above. Given the saturation in savings, it is not advisable to increase the RB voltage further as it known to worsen short channel effects. Process variations are known affect multiple transistor parameters (threshold, oxide thickness, effective channel length) which in-turn affect leakage and threshold voltage is the only parameter that can be dynamically altered with body biasing. Further our mechanism is designed only to ensure that leakage is within the bounds of maximum allowed leakage specified by heuristics used for estimating parametric yield. By providing a LUT based RBB generator, the leakage-bin field can be used to determine appropriate reverse bias voltages for further improvement of operational margin. Looking at the results of latency improvements in figure 10(b), it can be seen that there is a large discrepancy between minimum and maximum values. This is because, only the SRAM array is BB'ed and the latency is calculated for the entire access path constituting the periphery. The effectiveness of forward biasing in improving latency reduces with scaling of supply voltage. Since our mechanism can alter the source-body voltage based on measured latency, we can expect maximum latency reduction even under worst-case process variations.

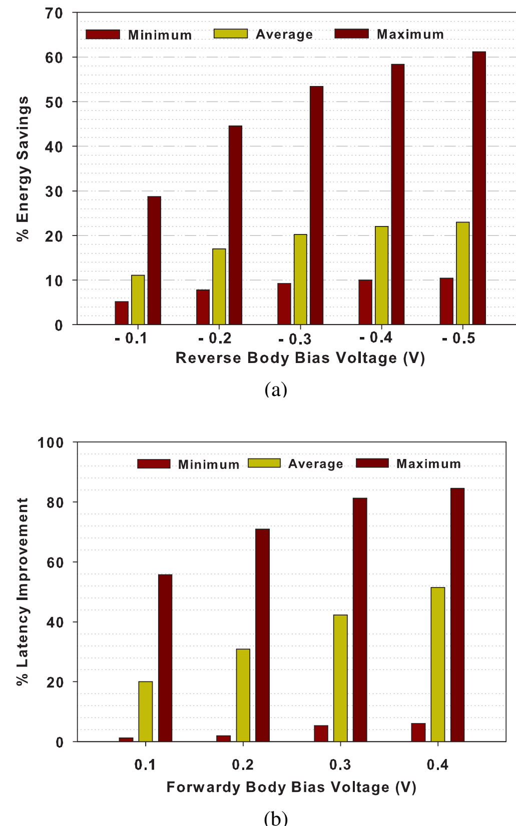


Fig. 10: (a) Percentage Energy Savings as a function of the reverse body voltage & (b) Percentage Latency Improvement as a function of forward body voltage. The bars represent savings when compared to Zero BB.

B. Evaluating Yield

Heuristics for estimating parametric yield suggest that individual cache lines which fail to meet the latency constraint (maximum allowed access latency under process variations) should be discarded. In sub-65nm designs, as leakage can play a very important role, it was shown that in addition to considering a latency cut-off, cache arrays whose leakage power is greater than 3μ should also be discarded [21].

Adopting the above heuristics, we determine the parametric yield for three different cases - No body biasing, forward body biasing active arrays with one voltage [9] & our proposal - dynamic fine-grain body biasing each array. We assume that all idle arrays are reverse biased at -0.3V for our case. The yield is determined for

multiple latency constraints and for one leakage constraint of 3μ. A cache is considered yield loss if more than 3 arrays fail to meet the constraints. It can be seen from Figure 11 that under zero body biasing (ZBB), yield reduces from 82% to 60% for tighter latency constraints. This reduction in yield results from arrays failing to meet only the latency constraints and not because of leakage constraints. The yield for forward biasing with one voltage is constant at 60% in all cases. While all arrays clear the latency cut-off because of lowering the threshold, some arrays fail to meet the leakage cut-off resulting in further yield loss. For all cases of latency constraint, the

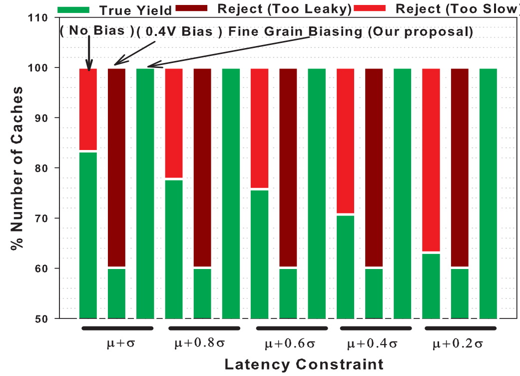


Fig. 11: Yield estimated for ZBB, Constant FBB and DFGBB as a function of Latency Constraints.

parametric yield is constant for our proposal. This is mainly because of 2 factors. Unlike adaptive body biasing where we decide to use either RBB or FBB, we employ both effectively in a time-shared manner. Secondly, the selected forward bias voltage is the minimum voltage for which the latency cut-off is met. This is to ensure that extra leakage energy because of forward biasing is not high enough to cause an yield loss. For the case when latency constraint is μ+σ, 82% of caches do not need any bias as shown in Figure 12. By providing a bias generator with just one programmable voltage of 0.1V, the yield can be significantly improved to 95%. With further increase in the number of available bias voltages, the increase in yield is minimal. The resulting parametric yield is actually not a function of the number of bias voltages but of the resolution (difference between minimum and maximum divided by the total number of available voltages) that ensure all arrays meet both latency & leakage targets. By increasing the number of available voltages (by reducing the intermediate steps), more fine-grain control can be achieved. For high performance designs, the latency constraints are between μ+0.2σ & μ+0.4σ and it is clearly evident that such caches require *vbs* of both 0.3V and 0.4V.

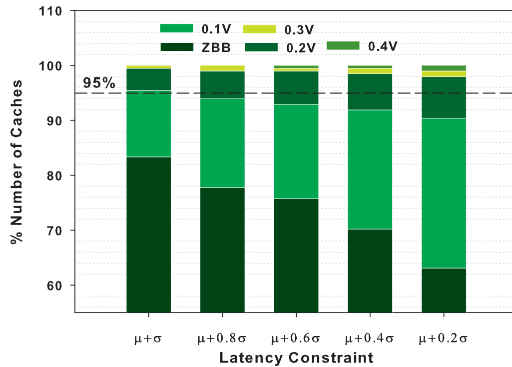


Fig. 12: Impact of FBB voltage granularity on parametric yield.

VII. Conclusion

In this paper, we propose a combination of latency/leakage monitoring and dynamically tunable fine-grain body-biasing techniques to maximize parametric yield and standby leakage reduction in caches. By measuring the time to access & retention time of the embedded 3T1D cell, it was shown that the SRAM arrays can be classified into discrete bins based on run-time leakage/latency measurements. Then, a look-up table based adaptive fine-grain body biasing mechanism uses this measurement to generate an optimal bias. While active arrays are forward biased to improve performance, inactive arrays are reverse biased to reduce leakage. The experimental results show that our technique on an average improves access latency & reduces leakage energy by 18% & 20% respectively. The adaptability to temporal changes ensures that the cache performance & power consumption over the lifetime of the chip is constant.

VIII. Acknowledgements

This work has been partially supported by the Spanish Ministry of Education and Science under grant TIN2010-18368 and TEC2008-01856, the TRAMS project of the FP7 program of the European Commission under agreement 248789, the Generalitat of Catalunya under grant 2009SGR1250 and Intel Corporation.

References

[1] W. Liao *et al.*, "Microarchitecture level power and thermal simulation considering temperature dependent leakage model," in *ISLPED'03*.
[2] D. Brooks *et al.*, "Power, thermal, and reliability modeling in nanometer-scale microprocessors," in *IEEE MICRO'07*.
[3] A. K. Singh *et al.*, "Mitigation of intra-array sram variability using adaptive voltage architecture," in *ICCAD'09*.
[4] M. Cho *et al.*, "Postsilicon adaptation for low-power sram under process variation," in *IEEE Design & Test'10*.
[5] Y. Pan *et al.*, "Selective wordline voltage boosting for caches to manage yield under process variations," in *DAC'09*.
[6] R. Teodorescu *et al.*, "Mitigating parameter variation with dynamic fine-grain body biasing," in *MICRO'07*.
[7] B. Choi *et al.*, "Lookup table-based adaptive body biasing of multiple macros," in *ISQED'07*.
[8] J. Jeong *et al.*, "Body bias generator for leakage power reduction of low-voltage digital logic circuits," in *Integrated Circuit and System Design. Power and Timing Modeling, Optimization and Simulation '06*.
[9] C. H. Kim *et al.*, "A forward body-biased low-leakage sram cache: device and architecture considerations," in *ISLPED'03*.
[10] A. Das *et al.*, "Evaluating the effects of cache redundancy on profit," in *MICRO'08*.
[11] W. Luk *et al.*, "A 3-transistor dram cell with gated diode for enhanced speed and retention time," in *VLSI'06*.
[12] K. Lovin *et al.*, "Empirical performance models for 3t1d memories," in *ICCD'09*.
[13] L. Xiaoyao *et al.*, "Process variation tolerant 3t1d-based cache architectures," in *MICRO'07*.
[14] S. Kaxiras *et al.*, "4t-decay sensors: a new class of small, fast, robust, and low-power, temperature/leakage sensors," in *ISLPED'04*.
[15] "Predictive Technology Models, http://www.eas.asu.edu/ ptm."
[16] Agarwal *et al.*, "Statistical timing analysis for intra-die process variations with spatial correlations." in *ICCAD'03*.
[17] Sarangi *et al.*, "Varius: A model of process variation and resulting timing errors for microarchitects." in *IEEE TSM'08*.
[18] C. Poki *et al.*, "A time-to-digital-converter-based cmos smart temperature sensor," in *IEEE JSSC'05*.
[19] A. Sathanur *et al.*, "Physically clustered forward body biasing for variability compensation in nanometer cmos design," in *DATE'09*.
[20] J. Gregg *et al.*, "Post silicon power/performance optimization in the presence of process variations using individual well-adaptive body biasing," in *IEEE TVLSI'07*.
[21] S. Ozdemir *et al.*, "Yield-aware cache architectures," in *MICRO'06*.