vPROBE:Variation Aware Post-Silicon Power-Performance Binning Using Embedded 3T1D Cells

***Abstract*—Variation-aware design involves designing circuits tolerant to process and temperature variations by estimating corner-case scenarios and providing minimum reliability guard bands. In order to extract maximum performance while consuming minimum power, systems need to adapt themselves by making them aware of their composition post-manufacturing. In this paper, we present an on-die post-silicon binning methodology that takes into account the effect of static and dynamic variations and categorizes every processor based on power/performance. The proposed scheme is composed of a discretization hardware that exploits the delay/leakage dependence on variability sources characteristic for categorization. We start by analyzing memory structures. The backbone of the binning scheme is the variation-tolerant 3T1D cell which is embedded into existing SRAM circuitry. By measuring the access and retention time of the 3T1D, we show that it is possible to gather sufficient information about the spread of spatial variability. This information can then be used for enforcing circuit-optimizations that are specific to every processor by making them aware of cross-die and intra-die variations. The results of our binning scheme indicate that post-manufacturing nearly 40% of chips fall into the high-performance low-power bin. However, this can change over the lifetime of the chip owing to constant activity resulting in degradation and performance reduction. To counter the effects of such high and low frequency variations, we also demonstrate the extended capabilities of this scheme in sensing high-frequency temperature variations and adapting to low-frequency changes. Plus, the information gathered by the binning can then provide a platform for cross-layer optimizations.**

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 variability in physical parameters is inhibiting the power/performance gains achieved by scaling devices. Reliable operation is further hindered by dynamic variations of temperature, supply voltage, transistor aging etc.

Increasing power densities is another cause of concern for designing high performance/low power circuits. With increase in temperature, the leakage component of power increases exponentially. Recent study has shown the 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 designers have cognizance of corner-case behaviors at design time, chips are seldom designed for worst-case owing to reduced performance. These dynamic variations to a great extent are dependent on the operating conditions and environment. As their frequency of occurrence and nature of existence is very random, it becomes virtually impossible to monitor and measure them. Designing systems with reduced guard bands would greatly improve performance but at the cost of reliable operation. Reliable operation can then be restored at the circuit level by making the system aware of the static variations and also sensing dynamic variations at regular intervals.

Sensing these variations, temperature in particular, has been mostly performed off-chip. Such a scheme lacks the ability to monitor high-frequency temperature variations due to the time-delay in sensing on-chip temperatures and off-chip regulation [3]. Further, fixed-point one time calibration does not account for the effect of degradation and aging. Almost most of the sensors implemented on-chip are very big and require special processes for components like thermistors or platinum resistors [4]. Such special requirement often make their usage near-to-impossible for regular designs. Dynamic Thermal Management (DTM) plays a vital role in offsetting the negative impact of temperature shootups and high power densities. Nevertheless, response mechanisms based on DTM principles are designed to mitigate emergency scenarios assuming homogeneous conditions across different processors. This calls for the need to develop a platform that discretizes every processor post-manufacturing enabling the possibility to provide ad hoc optimizations based on run-time power/performance.

In this paper, we present a novel 3T1D-based delay/leakage approximation scheme specially targeted towards memory structures such as register files & caches. The 3T1D cell (from now on shadow cell) is embedded along with existing SRAM circuitry. The sensing scheme has been incorporated into the read/write cycles thereby hiding all the synchronization and control overhead for a separate test cycle. Post-Silicon calibration through a single read and write has also been envisaged. The scheme is based on the fact that each read (or write) to the 3T1D cell will suffer almost the same variation on access power and delay when compared to the 6T cell it is paired with (since it will use the same periphery circuits and the physical variations will be almost identical between the cells due to their proximity). While keeping 6T cells for program execution, 3T1D cells are hidden (shadowed) from program execution and they keep track of the power/delay relation of each paired 6T cell. This information is later used to classify the different cells into power-delay bins.

The 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 spatio-temporal variations. In section 4, we propose the new binning methodology by introducing the composition inference scheme and then discussing about the discretization architecture. Section 5 discusses the extended use of our methodology towards temperature sensing and adapting to slow dynamic variations. Section 6 concludes the paper.

II. Related Work

Several works have highlighted the importance of post-manufacturing binning through the use of elaborate ATG tools aided by rigorous burn-in tests. Others have provided mechanisms on-chip which are manufacturer-controllable on an one-time basis [5]. For the sake of brevity, we will be discussing only on-die realizable self-adaptable sensing schemes.

In [6], tunable replica circuits resilient to dynamic variations have been proposed. On sensing high-temperatures of over $60\,^{\circ}\mathrm{C}$, the clock frequency is lowered to reduce dynamic power dependent temperature and also body bias lowered to reduce the leakage. Ituero *et al.* [7] have exploited the temperature dependent sub-leakage characteristics for gradient sensing of temperature. The sensor consists of

a chain of NOT gates fed by a input pulse. The pulse-width of the output is dependent on temperature and converted into a measurable form for temperature display. In [8], dynamic variation tolerance is achieved by means of using path-delay that amplify difference due temperature and aging. The change in delay is reported for corrective measures. In [9], the sensitivity of p-n junction diode to temperature is exploited for determining the temperature within a $3\,^{\circ}\mathrm{C}$ range. In [10], the dependence of threshold voltage on temperature is amplified through a chain of gated inverters. The scheme is extended using a voltage regulator to scale supply voltage reducing dynamic power consumption.

III. 3T1D Cell

Alternatives to 6T based SRAM 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 like 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

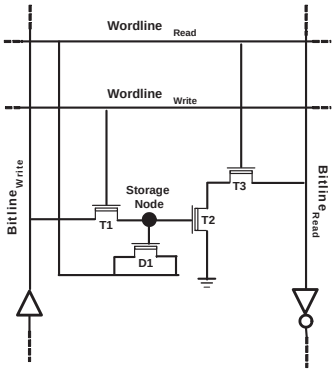

Fig. 1. Schematic of the 3T1D Cell

and access speeds comparable to that of standard 6T SRAM(s). While the cell can operate in the absence of transistor T2 in a 2T1D mode, T2 enhances the retention time and provides the gated diode the required boosting to improve the read speeds. The biggest advantage of the 3T1D is the non-requirement of similar strength transistors that constitute the cell. This has 2 fold advantages over 6T: 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.

Liang *et al.* [13] have proposed a 3T1D based cache architecture that relies on the fact that data stored in first level caches is transient and the time to the last cache reference to a given block (before it is erased) is well within the retention time of a 3T1D cell.

A. Write and Read Operation

Data is written into the cell by raising the write-wordline high and charging the write-bitline to the required value. The voltage level corresponding to a value '1' at the storage node is largely dependent on the strength of the transistor T1. A strong T1 would mean that the voltage level would be degraded when storing a 1. A lower voltage level would greatly reduce the retention time of the cell. 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. Due to boosting by T2, the value at the storage node increases temporarily close to the value of write-voltage. As the only path for sub-threshold leakage is through read-wordline tied to the gated diode, it can be reduced by holding the read-wordline at negative voltage (say -0.2V) when the cell is in its idle state (no access). This has shown to increase the retention rate by as much as 40X [11].

B. Retention & Access Time Measurement

As variation in device parameters are known to affect the performance of 3T1D to a lesser extent when compared to SRAM [12], for the sake of a comparative study, we place each of them next to each other on the same wordline and measure the retention and access time under similar conditions. The retention time of the 3T1D is defined

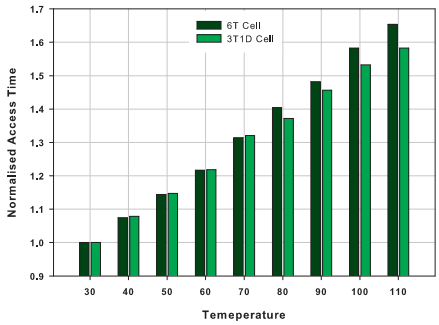

Fig. 2. Access Time Vs Temperature

as the time taken for the voltage at the storage node to decay past $V_{dd}/4$. It can be observed from Figure 2 that both 3T1D and 6T cells mimic their behavior. Although, Figure 2 for simplicity assumes just temperature variation, in the following sections, we will show how this behavior repeats when considering the other sources of variation. At high temperatures, 6T cell is more prone to performance loss when compared to the 3T1D. As opposed to regular 6T cells, the 3T1D are designed for single ended sensing. This combined with the boosting action provides very high read speeds even at high temperatures.

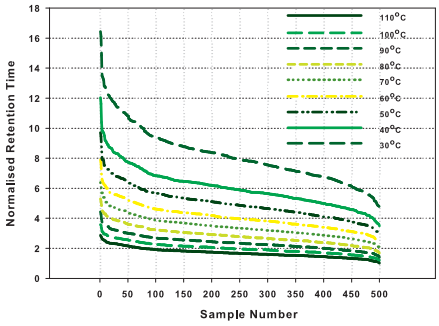

Fig. 3. Retention Time Vs Temperature

Figure 3 shows the measured retention time for 500 samples simulated for spatial variability. The simulated circuit is a 32KB cache composed of multiple 1KB array sub-blocks. Each array is organized into 64 rows by 128 columns with each column holding 32 active 6T and 3T1D cells. The retention time is normalized to the smallest retention time at $110\,^{\circ}\mathrm{C}$. This is to provide a better perspective on worst and best performing chips for estimation of power/performance guard bands. Retention time with zero-variability at $30\,^{\circ}\mathrm{C}$ is found to be 9.3μs. Under the presence of process variations, operating at $30\,^{\circ}\mathrm{C}$, the retention can be as high as 34.2μs or as low as 5μs. This is due to the high dependence of leakage

on physical parameters such as threshold voltage and channel length established by the relation

$$I_{Leakage} = ke^{-qV_{ht}}/(K_BT) \quad (1)$$

where q & K_B are physical constants, a and k are device parameters and T is absolute temperature [14]. Thus it is clear from the above argument that leakage is an important figure of merit that can be exploited to reflect static variations of physical parameters and simultaneously sense fine grain dynamic variations.

IV. Power/Performance Binning Scheme

A. Composition Inference Scheme

Conventional sensors that employ analog components require substantial calibration and often perform a one-time calibration not accounting for the effects of low-frequency variations such as degradation and aging which alter the behavior over the lifetime of the chip. Further, due to their large area requirements, it is not possible to integrate them into thermally unstable regions owing to existing device density [9]. It is well known that memory structures like register files and caches are sources of hotspots as a result of constant high activity. Also their importance from an area, energy & performance point make them top priority for constant monitoring.

In order to take advantage of the potential benefits offered by approximating leakage, in this work as a first step, we propose to measure the time taken by a 3T1D to discharge completely. One of the major bottlenecks of approximating leakage for inference of physical parameter distribution is the impact spatial variability across and within chips. In other words, variation of a parameter cannot be assumed as a single lumped value but a function of inter-die and intra-die(systematic & random) variations [15]. Inter-die variations are known to affect all the devices in a given die uniformly. Intra-die systematic variations are dependent on layout geometry and also correlated across distance. Intra-die random variations as the name suggests affect every device differently and are caused by issues such as random dopant effects and line edge roughness [16]. While regular structures like memories have lower levels of within-die systematic variations, they are easily affected by intrinsic variations because of minimum geometry transistors. These intrinsic random variations affect the critical charge of the SRAM cells making them extremely susceptible to bit-flips on accumulation of charge from alpha particles [15] .

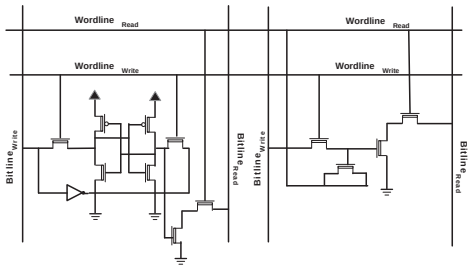

Fig. 4. 3T1D embedded with a 6T SRAM

The idea of placing a 3T1D next to a 6T SRAM is to gather sufficient information about the functioning of the SRAM as shown in Figure 4. We would like to leverage their functional equivalence to garner meaningful data about their structural disparities. The most important factor to be taken into account while sizing the 3T1D is the read/write speeds of the adjacent 6T cell. This is primarily to make sure that the system is unaware of the existence of the 3T1D by hiding all the synchronization and control information within the read/write cycle of the 6T. Secondly, the cells could double up as alternative storage space. By making the write driver stronger and the domino-sense amplifier weaker, read/write speeds can be made to match those of the 6T cell. Additionally by making the write driver stronger, the retention time is significantly improved. The fundamental assumption of the binning scheme is that leakage power (dominant source of power in sub 32nm designs) is a strong function of retention time and and the time to access a cell corresponds to the critical path delay. Looking at Figure 2 & Figure 3, the fact that access time and retention indeed reflect delay and leakage can be validated.

The purpose of binning chips individually post-silicon is very much alike calibration. Calibration enables the chip to rectify itself from deviations that arise out of manufacturing. From a statistical

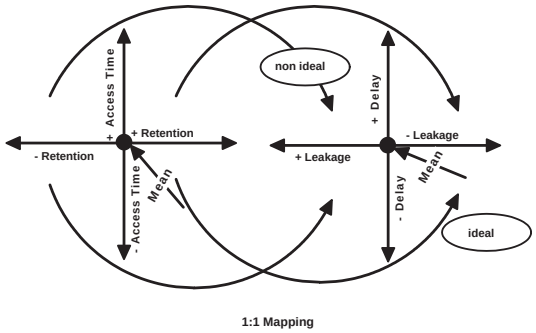

Fig. 5. 1:1 Mapping of Access Time/Retention to Delay/Leakage

standpoint, these deviations can be on either side of the mean corresponding to the value obtained at design time. Due to this difference in values of physical parameters across dies, a single-point fixed calibration will not address issues of inter-die variations. Thus it becomes extremely important to preclassify chips based on their static variations post-manufacturing. While it is extremely difficult and expensive to measure absolute values of parameter deviations, as shown in Figure 5 chips can be categorized into different bins based on predefined delay and leakage values. A very high access time and low retention translates directly to very slow and significantly high leakage power. Several circuit level optimization like dual-v_{th}, body biasing, supply voltage minimization have been proposed to maximize performance while maintaining power well within allowable budgets [17], [15]. However such optimizations resulting from holistic procedures have been enforced across varying chips yielding non-uniform benefits.

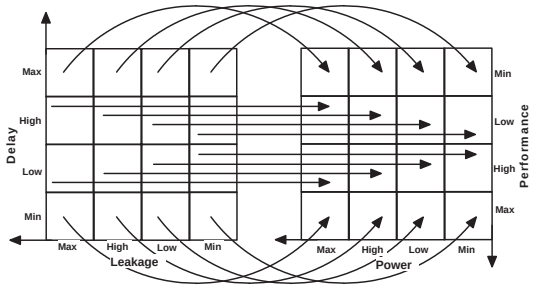

Fig. 6. Power/Performance Binning

An approach of finer-granularity binning based on power/performance is shown in Figure 6. In contrast to the coarse grain approach depicted in Figure 5, this mode of classification makes it possible to reduce bounds within which a chip is placed

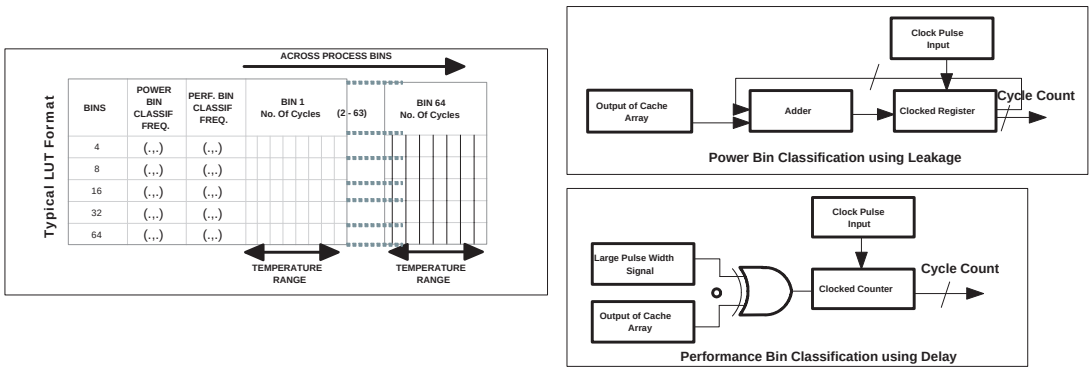

Fig. 7. Power/Performance Bin Classification based on Retention/Access Time

into a given bin. Further, it makes it attractive to extend such a mapping scheme to different regions on-die to account for intra-die variations. It is well established that both power and performance 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. In [18], it was shown that under the effect of v_{th} variation, leakage can be as high as 4.7X in comparison to the leakage at design time. Nevertheless, these chips can be restored to their original power configuration for a minimum performance overhead by iteratively assigning high-v_{th} to delay non-critical paths and reducing the standby supply voltage of idle blocks. However such optimizations also assume a standard percentage increase in v_{th} irrespective of the composition specific to each chip. By making the power/performance bounds more tight during binning, circuit optimizations can take advantage of the better cognizance of the composition. Assuming that under the effects of static variability it is possible to discretize systems based on their difference in retention and access times of the 3T1D, we have discussed a new binning approach that enables classification based on power and performance. It was also demonstrated that it can adapt over time dynamically and expose the difference in composition not only across chips but within them as well. In the next section, we discuss one possible methodology for discretization based on retention and access time.

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 power. The retention time is in the order of μs only when there are no accesses to a given cell. In other words when the read-wordline is not strobed. By holding the read-wordline high continuously, the retention is reduced by nearly 20X. It is this window of few hundred *ns* that is deeply impacted by temperature. The measurement scheme just has to convert this nanoseconds into something measurable on-chip. In this case, the simplest solution is to count the elapsed number of cycles. Any scheme that involves a delay-to-pulse circuitry generates a clock cycle for every period that the output of the cache array is held high [19]. In other words, the counter with a low-pulse width clock measuring a signal lasting typically few micro seconds will result in an output of a 100,000 thousand cycles. The lower the leakage, the higher the retention and so higher the number of output cycles. In addition to very fast response, the only other advantage of using high frequency pulses is the large difference in output cycles between adjacent bins. This pulse and the clock rate of the processor may be different, although for the sake of simplicity we suggest the use of multiples of the clock rate of the processor.

Our proposed power-bin classification architecture is shown in Figure 7. The output of the cache array is linked to an adder which has the feedback of a clocked register. The register is clocked at a frequency bounded by the pulse width of the minimum difference between any 2 adjacent bins. The total number of bins to be used for classification is user selectable. With increase in number of bins, those that are adjacent to each other will exhibit retention times of very small differences. In order to amplify such small differences for discretization, higher frequencies can be employed. The bin selection procedure is initiated by writing a 1 to a 3T1D cell and signaling a read access and constantly holding the read-wordline high. The output of the sense-amplifier after a given period begins to decay. As long as the output is high enough to signal a 1, the adder increments the value of register by a 1. The register is incremented at a predetermined frequency whose clock period is low enough to make sure adjacent bins exhibit a difference of atleast 1 cycle as shown in Figure 8.

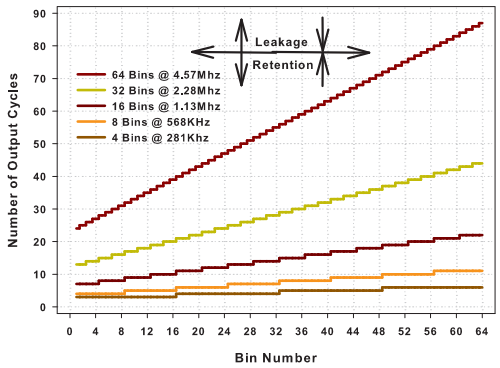

Fig. 8. Number of Output Cycles Vs Power Bin Number

It is clearly observable from Figure 8 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. Some researchers may argue that it is not a viable option to have a frequency divider for bin-classification purposes. The simplest solution to the problem would be to design for a frequency that would cater to the maximum number of bins, for instance 64. If the user intends to classify based on lesser number of bins, say 8, classification is performed by grouping into bins which are multiples of 8. This way in an 8-bin classification using 64 bins, bin 8 would represent 1 and bin 16 represents 2 and so on.

Look-up-table (LUT) as shown in Figure 7 is created statically at design time. This information is composition independent and can be loaded into control registers at run time. They are basically used for comparison of observable cycle count to obtained cycle count for classification purposes. The basic format of the table has an entry listing the cycle count for the first bin. For every next bin, the cycle count is incremented by 1. The extended version of the LUT is also shown which will be explained in the subsequent section. When area is not a constraint, and bounds of each bin have to be very tight, frequency dividers can be used for sharper resolution. While the classification frequency for a given-number-of-bins is a factor of the frequency for another set of given bins, the same does not hold true for cycle count. This is primarily due to variation in leakage dependent on random variations. As a result, the cycle count of the first bin for every target-number-of-bins is loaded and the rest obtained by a simple add.

In order to classify based on performance dependent on critical path delay, we determine the read access of the 3T1D cell. The access time is computed as the time from control-word initiation till the output of the sense amplifier going high. 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 chips. In effect, the separation between adjacent bins can be as low as 6ps. As a result even multi-Ghz frequencies for a 64 bin classification will not suffice. 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 one proposed in [19]. 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. 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. Orshansky *et al.* [20] have characterized the impact of of threshold variability on delay(D) using the following relation,

$$D \quad \sim \quad L_{eff}.V_{dd}/(V_{dd} - V_{th})^{\alpha} \tag{2}$$

where α is the velocity saturation assumed to be 1.3 for 45nm technology node. It is evident that delay is highly dependent on the inverse of the difference in supply voltage and threshold. In other words, for a fixed threshold by reducing the supply, the access time will increase considerably so as to make possible the amplification of adjacent performance bins. Under zero-spatial variability at ambient-temperature, the difference in access time of the shadow cell between 1.0V and 0.6V is 4.7X which is magnified even higher when considering process variations. Thus the target number of bins can be significantly increased by lowering the supply voltage resulting in the increase in delay.

On the outset, the entire scheme may look very slow incapable of reacting to high frequency variations. This frame of thought can be eliminated by considering the example below. We simulated the entire procedure on a shadow-cell based cache design along with the discretization module implemented with 45nm PTM library [21] on the HSPICE simulator. Multi-level quadtree based variability scheme was incorporated to account for cross-die and intra die variations of v_{th} and l_{eff}. The access and retention times of 500 cache samples generated from monte-carlo simulations were measured using the discretization architecture. For a fixed supply voltage of 1V, the classification was performed for 4 binning levels of performance and power. The results of the final classification is shown in Figure 9.

It is strange that no chips have been placed in the maximum power, low performance bin. 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 maximum power, low performance chips across the 3 immediate neighboring 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 samples were categorized into the maximum power, low performance bin. Majority of the chips have been placed in the high performance, low power bin indicating the goodness of the yield. During circuit operation, dynamic variations happen at such fine granularity that it is not possible for the chip to move along its polar co-ordinates. Thus there is a fixed probability the chips belonging to the low-high(high-low) bins move only to one of the 4 bins along the Cartesian direction. Chips placed in extreme bins can move in 1 of 2 directions and the remaining 8 in 1 of 3 directions. By loading the cycle-count of every neighboring bin at run time, within 4 comparisons dynamic variations of any chip placed in any bin can be tracked and appropriate response mechanisms initiated. This binning also can detect stuck-at faults. All measurements that go below the bottom bin indicate a stuck-at-zero fault and all measurements that go beyond the maximum bin can be considered as stuck-at-one faults. Since bin classification can be done periodically a cell can move in (and out) of the fault at every period. It then depends on operating system or any control system to decide whether to use this cell or not.

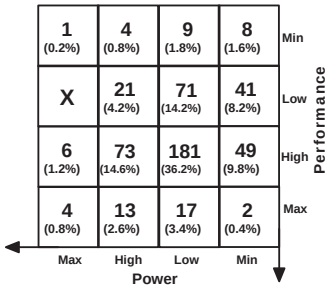

Fig. 9. Power/Performance Binning using Shadow Cells

V. Application

A. Temperature approximation under Spatio-Temporal Variability

Many smart temperature sensors (STS) exploit the delay variability across temperature characteristic for sensing the variations. As transistors are easily affected by process variations, elaborate calibration mechanisms are often used for compensation. This would greatly increase the complexity, area and testing costs [22]. However for non-critical applications such as everyday computing, measurement accuracy can be traded-off for simplicity in design. This prompts the need for an on-chip based temperature approximation scheme. Without any addition to the existing hardware, we create new entries in the LUT for temperature approximation as shown in Figure 7. The cycle count is dependent on the lowest measurable temperature. As we do not require a complete military range specification ($-55\,^{\circ}\mathrm{C}$ to $125\,^{\circ}\mathrm{C}$), the system can be designed to sense temperatures between $30\,^{\circ}\mathrm{C}$ & $110\,^{\circ}\mathrm{C}$. The exact same procedure enforced for power binning is employed for temperature approximation with a $10\,^{\circ}\mathrm{C}$ granularity. From Figure 10, it is evident that there is overlapping in the number of output cycles across different temperatures. This is primarily due to the effect of cross-die variations. The extended version of the LUT has one entry for every bin number corresponding

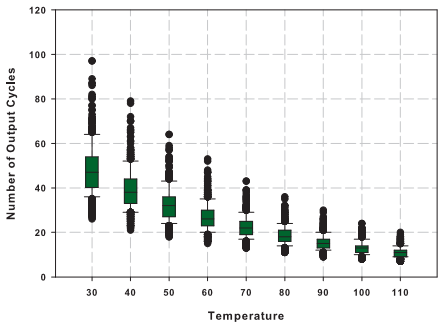

Fig. 10. Distribution of Number of Output Cycles Across Temperature Range

to the cycle count of the lowest measurable temperature. For subsequent temperature levels, a difference of 1-3 cycles from the lower temperature is accommodated. This approximation can be calibrated by taking advantage of the delay dependence on temperature and for extra area overhead, store entries for cycle count obtained from the delay test. The 2 different values obtained for the same temperature are compared with the entries in the LUT for better approximation. If both entries point to the same temperature but from different bins, a complete re-binning is enforced.

B. Tracking and Tolerating Low-Frequency Dynamic Variation

As high performance is traded off for reliability by significant reduction of guard bands, it becomes imperative that for correct operation reliability ought to be restored by monitoring the system functionality across its lifetime. The effect of process variations worsens the scenario by altering the composition of microprocessor that Mean Time to Failure (MTTF) is not the same across different processors. Further, runtime temperature variation stresses the circuit causing fatigue deformation [2]. A figurative representation of such a phenomenon is shown in Figure 11. Our proposed power/performance

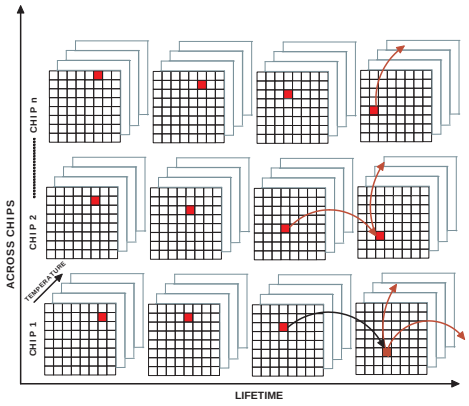

Fig. 11. Monitoring Low Frequency Dynamic Variations

binning scheme can be used for classification early in the lifetime of the chip. It is possible to define reliability guard bands by identifying emergency scenarios in terms of power/performance. Therefore, the binning process is a continuous one requiring movement of bins over the lifetime of the chip. For example, optimizations need to be enforced for certain chips running at more than $40\,^{\circ}\mathrm{C}$ while for others safe operation can be assured for temperatures even as high $70\,^{\circ}\mathrm{C}$. This is depicted by the flow of arrow indicating the need to restore reliable operation at the cost of performance. Due to high activity, certain chips degrade faster than others as in the case of Chip 2, thereby prompting the need for corrective measures. By identifying the direction of flow of bins over a given period, a definitive measurement of degradation can be enabled. The change in trend can be reported to the operating system for remedial optimizations.

VI. Conclusion

With reducing feature sizes, process variations are becoming a tough-challenge to cope with. This is worsened by high power densities making reliable operation tougher. In order to provide a platform for power/performance optimizations, it becomes important to make systems aware of their own composition. In this paper, we have presented a novel binning architecture that can be realized in processors for self-classification based on power/performance. The scheme proposed uses variation-tolerant 3T1D cells embedded into a conventional 6T-SRAM based cache for monitoring of processor structures, memory in particular. It was also shown that, the same design can be extended for temperature approximation. Further, by monitoring low-frequency dynamic variations such as aging and degradation, system reliability can be assured across the lifetime of the chip.

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] H. Sanchez *et al.*, "A cmos temperature sensor for powerpc risc microprocessors," in *VLSI Circuits'97*.
[4] M. Pertijs *et al.*, "A cmos smart temperature sensor with a 3σ inaccuracy of 0.5c from -50c to 120c," in *IEEE JSSC'05*.
[5] A. Das *et al.*, "Evaluating the effects of cache redundancy on profit," in *MICRO'08*.
[6] K. o. Bowman, "Circuit techniques for dynamic variation tolerance," in *DAC '09*.
[7] P. Ituero *et al.*, "Leakage-based on-chip thermal sensor for cmos technology," in *ISCAS'07*.
[8] M. Agarwal *et al.*, "Circuit failure prediction and its application to transistor aging," in *VTS '07*.
[9] S. Remarsu *et al.*, "On process variation tolerant low cost thermal sensor design in 32nm cmos technology," in *GLSVLSI '09*.
[10] Q. Chen *et al.*, "A cmos thermal sensor and its applications in temperature adaptive design," in *ISQED '06*.
[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] S. Borkar *et al.*, "Parameter variations and impact on circuits and microarchitecture," in *DAC'03*.
[16] K. Bernstein *et al.*, "Design and CAD Challenges in sub-90nm CMOS Technologies," in *ICCAD'03*.
[17] Liu *et al.*, "Leakage power reduction by dual-vth designs under probabilistic analysis of vth variation." *ISPLED*, 2004.
[18] S. Ganapathy *et al.*, "Modest: a model for energy estimation under spatio-temporal variability," in *ISLPED '10*.
[19] C. Poki *et al.*, "A time-to-digital-converter-based cmos smart temperature sensor."
[20] M. Orshansky *et al.*, "Characterization of spatial intrafield gate CD variability, its impact on circuit performance, and spatial mask-level correction," in *IEEE TSM'04*.
[21] "Predictive Technology Models, http://www.eas.asu.edu/ ptm."
[22] K. jae Lee *et al.*, "Analytical model for sensor placement on microprocessors," in *ICCD05*.