Process Variation Aware SRAM/Cache for Aggressive Voltage-Frequency Scaling

Avesta Sasan (Mohammad A Makhzan), Houman Homayoun, Ahmed Eltawil, Fadi Kurdahi
{mmakhzan,hhomayou,aeltawil,kurdahi}@uci.edu
University of California Irvine

Abstract-*this paper proposes a novel Process Variation Aware SRAM architecture designed to inherently support voltage scaling. The peripheral circuitry of the SRAM is modified to selectively allow overdriving a wordline which contains weak cell(s). This architecture allows reducing the power on the entire array; however it selectively trades power for correctness when rows containing weak cells are accessed. The cell sizing is designed to assure successful read operations. This avoids flipping the content of the cells when the wordline is overdriven. Our simulations report 23% to 30% improvement in cell access time and 31% to 51% improvement in cell write time in overdriven wordlines. Total area overhead is negligible (4%). Low voltage operation achieves more than 40% reduction in dynamic power consumption and approximately 50% reduction in leakage power consumption.*

I. INTRODUCTION

With fabricated device dimensions approaching the limits of process technology capabilities, a rapid increase of manufacturing process variation induced defects is observed [1-5]. This in turns makes the defect rate in the memory device sensitive to changes in operation parameters including temperature, voltage and frequency. Due to the random nature of local process variation, resulting defects have random and uniform distribution [1-5] that adversely affect the expected system yield. Furthermore, voltage scaling exponentially increases the impact of process variation on memory cell reliability, resulting in an exponential increase in the fault rate. In this paper, we propose Selective Charge Pumping Cache (SCPC) architecture with improved wordline driver architecture. The architecture allows aggressively scaling the supply voltage on the entire memory structure, and selectively charge pumping the wordlines with weak cells to higher voltages. The cell is appropriately sized to safe guard against read failures thus avoiding flipping the content of the cells when the wordline is in overdrive mode. The increase in leakage and area overhead as a result of this resizing is negligible (< 4%). The selective nature of the circuit allows for significant savings on both leakage and dynamic power consumption while only targeting failing cells with selective overdriving.

II. PRIOR WORK

There exists a multitude of techniques to handle failing cells in SRAM structures. Among these techniques row and column redundancy are widely used. However these techniques are limited to a small fixed percentage of the memory array cells[7][8], and thus are poorly suited to dynamic parametric errors. Another commonly used technique is Error Correcting Codes (ECC) [10][11] to deal with transient defects. ECC guarded memories can handle dynamic faults albeit at a heavy cost in power consumption, area and complexity [17]. Statistical sizing and optimization of the SRAM cell for yield enhancement is suggested in [12]. This pre-silicon technique could improve production yield, however, it is limited by conflict in sizing requirement for different types of failures [9]. In addition to design and circuit level techniques, architectural level techniques have also been proposed to address manufacturing induced process variation. Inquisitive Defect Cache (IDC) in [6] is a small direct or associative cache that works in parallel with L1 cache and provides a defect free view of the cache for the processor in the current window of execution. However, in this work the basic assumption is that the data, if lost, could be recovered from lower level cache or memory and could only work for hierarchical structures. Finally, the area overhead of this method (about 12%) is large. Resizable caches are suggested in [9]. In this technique, it is assumed that in a cache layout, two or more blocks are laid in one row, therefore the column decoders are altered to choose another block in the same row if the original block is defective. However, for tightly coupled loops this approach will always result in a miss [6]. In the following section, we discuss voltage scaling as the most viable approach for low power consumption. Section 4 then presents a discussion of the proposed architecture. Section 5 discusses the design considerations associated with overdriving the wordline drivers. Section 6 presents the performance improvement while section 7 discusses area and power impact. The paper is concluded in section 8.

III. PROCESS VARIATION AND MEMORY CELL OPERATION UNDER VOLTAGE SCALING

Process variation in SRAM under voltage scaling was studied in [6][18] in which variation in process parameters was lumped into an independent Gaussian distribution characterizing the Vth fluctuations of each transistor. In order to setup a baseline for comparison of our proposed solution a simulation was setup reproducing the observations in [1-5]. In this simulation, the circuit under test is a standard six transistor SRAM memory bit cell. The SPICE models used for the simulation were obtained from the Predictive Technology Model (PTM) [14] website in 32nm. In the simulation, Vdd is lowered from its nominal voltage (1V) to (0.6 V). In order to obtain a probability of failure we assumed a constant access time and used Monte Carlo simulations to calculate the probability of failure of a cell (accumulated read, write and destructive read probabilities). The SRAM cell under

978-3-9810801-5-5/DATE09 © 2009 EDAA

consideration was designed for a 48.1 ps access time and 39.8 ps write time at nominal voltage (1V). 120ps operation (access or write) time was used to access the cell regardless of the voltage level.

Another important issue to consider is the frequency scaling policy which will have a major effect on the probability of failure. Along with the voltage scaling the frequency will or will not scale. If the frequency is constant and the voltage is scaled the memory frequency management policy will be referred to as a Fixed Frequency Voltage Scalable (FFVS) policy. On the other hand, if the frequency is scaled along with voltage it will be referenced as a Frequency Scalable and Voltage Scalable (FSVS) policy. In the example previously given, the FFVS policy was used where the cycle time is kept constant and an increase in the mean delay of the memory cells shifts a larger portion of access time distribution out of range. For future reference in the paper we will define the "safe margin" for both these policies as the increase in the access time from mean access time that reduces the probability of failure below a controlled threshold and provides immunity from process variation.

IV. PROPOSED ARCHITECTURE

To counter the effect of process variations, typically designers overdrive the entire memory array. This leads to extra power consumption as both leakage and dynamic power increase. We are proposing to use a modified wordline driver peripheral circuit to allow selective wordline overdriving utilizing a small one step charge pump. The wordline peripheral circuit will drive the wordline in two phases. In the first phase, using the supplied Vdd the wordline is driven to Vdd. In the second phase the charge pump will overdrive the wordline voltage increasing the Vgs above the supply Vdd. Increase in Vgs improves both access and write time to the cell as will be described in the following sections.

A. *Improved wordline driver*

Figure 1 outlines the proposed wordline driver architecture. It consists of two consecutive NAND gates(possibly preceded by two or more inverters to increase fan-out). The second NAND gate's pull up PMOS (P1) as shown in figure 1 is connected to the supply voltage. Pull up PMOS (P2) is connected to the charge pump output. Using this driver P1 first drives the wordline to Vdd and conditioned upon activation of P2 the charge pump and wordline start charge sharing. Being at a higher voltage the charge pump's capacitor charge migrates to the wordline, effectively raising its voltage as shown in Figure 2.

The proposed two phase configuration for wordline overdriving scores two advantages over single phase charge pumping. First, charge sharing is a slow process, driving the wordline by solely charge sharing requires longer time for the wordline to reach its peak value. Secondly, the two phase configuration results in higher wordline peak value as compared to a wordline driver with a one phase charge pump configuration. The delay unit in Figure 1 controls the timing difference between the first and second phase of driving the wordline. The delay unit consists of small inverters (in increasing order of driving power –fan out-) and its delay is controlled by channel length and/or the number of inverters in the chain. The output of delay unit is used to switch from P1 to P2. Figure 2 illustrate the timing and operation of the wordline driver.

The selective behavior of this circuit is controlled by the input to the Nand-gate in the delay unit of the wordline driver. If the external-input to this Nand-gate is high, the wordline overdrive will be inactive whereas a low input initiates the overdriving behavior. Typically, the input to the Nand-gate will be from a defect map that stores the results of a BIST run that is performed for each new setting of voltage, frequency and temperature. This approach assures support for a large number of operation modes. Alternatively, if the system is known to have a small number of operating modes, the configuration information can be stored in fuses that are configured for each mode.

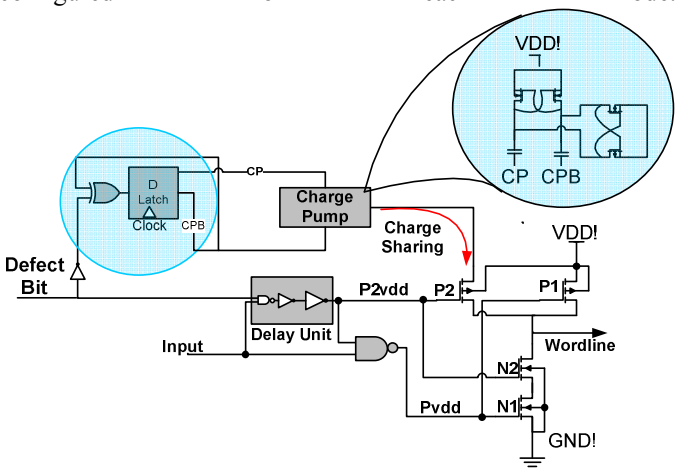

Figure 1: Proposed wordline driver

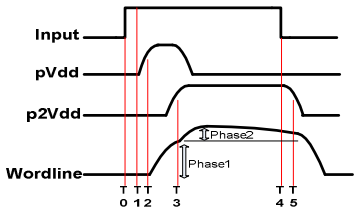

Figure 4: Signals timing order in the proposed charge pumped wordline driver

V. WORDLINE DRIVER DESIGN CONSIDERATIONS

In this paper we used the charge pump basic cell introduced in [15] as illustrated in Figure 1. For the purpose of our proposed circuit we only use one stage of this charge pump cell. CP and CPB will have opposite polarities (phases). Note that two capacitors are used in this charge pump and depending on the frequency of CP and the output capacitance load the range of output fluctuation could be controlled. The voltage of the wordline after the second phase of wordline driving by the charge pump is related to the applied supply voltage, size of the charge pump capacitances, CP frequency and related device sizes of NMOS and PMOS transistors used within the charge pump. If CP and CPB stay constant during cell access time and transistors sizes are chosen large enough, the wordline voltage after overdrive is determined by:

$$V_{OverDrive} = V_{dd} + \frac{C_{cp}(V_{dd} - V_{th})}{C_{wl} + C_{cp}} \quad (1)$$

$$C_{wl} = \sum_{i=1}^{N} [Cgl_i + Cgh_i] + C_{wire} \quad (2)$$

In which N is the number of cells in each word-line. Cgl_i and Cgh_i are the Gate Capacitances of the i^{th} cell's access transistor connected to the low (Cgl_i) and high (Cgh_i) side of the cell. C_{wire} is the word-line wire capacitance. Cgl_i and Cgh_i are obtained from PTM in 32 nm and the wire capacitance was obtained from CACTI 5.0 [16] wire models in 32 nm technology.

VI. Wordline Overdriving and Cell Stability

Overdriving the wordline driver may cause cell stability problems. The access and pull down transistors in a SRAM cell during a read as shown in figure 3 form a voltage divider. The size of pull down transistor is chosen large enough to assure that the rise in the intermediate node of the voltage divider smaller than threshold voltage of devices used in the cell. Increasing the voltage of the access transistor lowers the resistance of the access transistor effectively increasing the voltage at this intermediate node. This in turn increases the likelihood of a bit flip during a read operation. In order to prevent this effect one could trade cell area increasing the size of pull down transistor to counter the voltage overdrive impact on cell stability. Increasing the size of the pull down device also impacts the Static Noise Margin (SNM) and the leakage of the SRAM cell.

To establish a fair comparison, we linearly upsize the traditional cell and compare the failure probability of both architecture with equivalent cell areas. The following analysis is based on a charge pump architecture that effectively increases the wordline voltage by 40% over the supplied cell voltage. The charge pump capacitances to achieve this level of overdrive could be calculated from equation 1. The rise in the intermediate point of the voltage divider (during a read operation) can be expressed as:

$$\Delta V = \frac{V_{DSAT} + CR(V_{DD} - V_{Tn}) - \sqrt{\Gamma}}{CR} \quad (3)$$

$$\Gamma = V_{DSATA}^2.(1+CR) + CR^2.(V_{DD} - V_{Tn})^2 + 2.CR.V_{DSAT}.(V_{DD} - V_{WL}) \quad (4)$$

In which CR is the cell ratio of the pull down transistor to the access transistor and is obtained from:

$$CR = \frac{W_{PD}}{L_{PD}} . \frac{L_{AC}}{W_{AC}} \quad (5)$$

Figure 4 illustrates the voltage rise in node L (node storing a 0) in both charge pumped and traditional architecture varying based on Pull-down cell size. In this simulation the threshold voltage of devices is 0.31V. If the 0.27V voltage rise is chosen in the traditional design (to allow proper SNMread margin) the pull down to access transistor ratio in the traditional cell base on figure 4 should be 1.2. For the same voltage rise when overdriving the wordline the access transistor should be 1.4. This means increasing the size of the pull down device from (38.4nm to 44.8nm in a 32nm design) or (78nm to 91nm in 65nm). This in turns mean a larger layout and a larger leakage. It is important to note that the change in cell area is strongly dependent on the cell's layout. In the rest of this paper we build our discussion based on the cell layout illustrated in figure 3 however for any other SRAM layout a similar methodology could be developed.

By changing the ratio of the pull up to pull down (λ) the SNM is degraded however for the slight change in the pull down device (from 1.2 to 1.4) the change in the SNM is measured to be less than 4%.

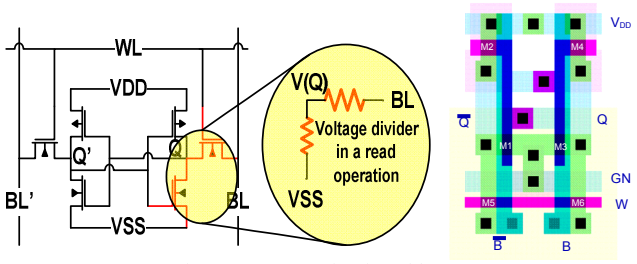

Figure 3: SRAM circuit and layout

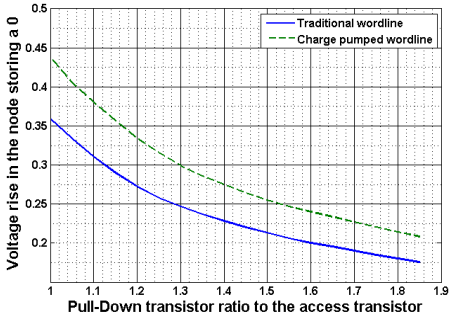

Figure 4: Voltage rise in node with store value of 0 during a read operation.

VII. Impact of Charge Pump on Access/Read Time

Overdriving the wordline using a charge pump improves both mean and standard deviation of the access/write time distribution. As previously mentioned, in order to guarantee read stability during a read operation via wordline overdriving, the pull down transistor is upsized to assure the same voltage rise as in traditional cell, while the area of the modified cell and the proposed cell is equivalent.

A. Mean access/write time improvement

We previously discussed that by means of charge pumping the wordlines we could improve the read/write time of the cell and make it less sensitive to process variation. In this section, the gains associated with this procedure are quantified. A simulation was setup using Berkley 32nm PTM where the mean access/write times of the cells in the proposed scheme was compared to that of the traditional architecture. Figure 5 illustrates the simulation results comparing the access time and write time of the proposed architecture in charge pumping and regular access mode to that of the traditional architecture.

It is interesting to note that decreasing the voltage increases the percentage improvement in the access time but up to a point (0.63v) after which a further decrease in the voltage does not yield an improvement in the charge pump architecture. This is because at lower voltage the charge pump requires a

longer time to charge up and therefore when the charge pump is activated it does not reach full charge.

B. Improvements in the standard deviation from mean.

In order to understand the effectiveness of the proposed wordline driver to mitigate process variation effects a Monte Carlo simulation was setup in which the threshold voltage values, which are based on a Gaussian distribution [1-5], were varied. The normalized result of this simulation is illustrated in Figure 6. As discussed in the previous section voltage scaling increases the mean access time. In addition, results obtained from this simulation show that not only does the mean shift but also the standard deviation from mean is modified by voltage scaling. This poses a limitation on defining a safe margin for a FSVS frequency management policy. In other words, the safety margin can no longer be specified as a fixed value from the mean since the standard deviation varies depending on the applied supply voltage. For example, by referring to Figure 6, it is clear that the same safety margin that is used for the nominal 1 v will result in a higher failure rate if used for the 0.8 v setting due to the larger standard deviation.

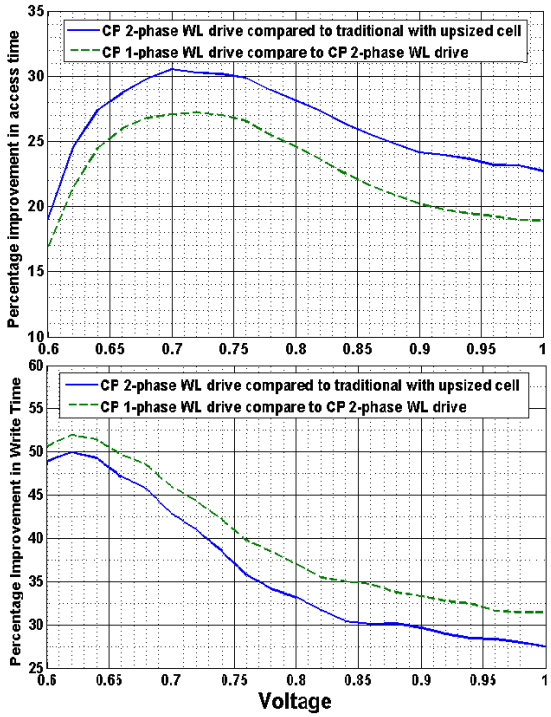

Figure 5: Percentage improvement in Access time (Top) and write time (Bottom) of the charge pumped architecture compared to traditional architecture.

As discussed in the previous section using the proposed wordline driver results in a smaller shift in the mean access time of the cell upon voltage scaling. Figure 6 also reveals another interesting characteristic of wordline overdriving: Although the charge pumped architecture is also affected by process variation, the standard deviation from mean access time is reduced. Therefore for a fixed safe margin policy the number of weak cells is decreased. Change in the mean and standard deviation of the access time for charge pumped and traditional architecture is summarized in table 1.

C. Improvements in the probability of failure at lower voltages

As explained in section 3, voltage scaling could be accomplished with either FSVS or FFVS frequency management policies. In the following section we investigate how the probability of failure changes with each of these frequency management policies.

- *Probability of failure in a system with FFVS frequency management policy*

Reducing the voltage increases the cell access and write time. Therefore in a system with an FFVS frequency, voltage scaling reduces the gap between maximum realizable and clocked frequency. Reducing this gap increases the sensitivity to process variation such that if in nominal voltage 6σ variation in V_{th} result in a faulty behavior in a cell, at lower voltages due to both mean shifting (as explained in 6.1) and increase in the deviation of the access time (as explained in section 6.2) a much smaller variation could result in a defective behavior.

TABLE 1: CHANGE IN THE MEAN AND STANDARD DEVIATION OF ACCESS TIME

VOLTAGE	μ SHIFT CP → TRAD	σ SHIFT CP → TRAD	μ SHIFT 1-PHASE → 2- PHASE	σ SHIFT 1-PHASE → 2- PHASE
1.0 V	11.53	1.72	8.11	1.52
0.9 V	15.3	2.37	11.4	2.28
0.8 V	19.9	6.97	16	5.98
0.7 V	31.4	10.46	27.5	17.14
0.6 V	34.7	15.27	28	26.13

To quantify how the probability of failure of a FFVS system changes due to voltage scaling, a simulation was setup where an SRAM with maximum realizable access time of 48.1ps and write time of 39.8ps experiences voltage scaling. The SRAM operates at a fixed 120ps and the access time is quantified from nominal voltage down to a lower voltage (close to Vth voltage). An error occurs if the access time of 120ps is not honored. The process variation is modeled as a variation in Vth of each transistor. Process variation from - 6σ to 6σ [1-5] for each transistor is considered when Monte-Carlo simulation are performed. The obtained results are used to produce a probability of failure for each voltage as shown in Figure 7 (left) for both the traditional and the charge pump based approach. The probability of failure of the charge pump architecture is a several orders of magnitude smaller that the traditional architecture.

- *Probability of failure in a system with FSVS frequency management policy*

In a Frequency Scaling and Voltage Scaling (FSVS) frequency management policy, memory access time (frequency) is scaled as the voltage is scaled. In this case a Monte Carlo simulation was setup with a fixed safety margin of 30 ps. Figure 7 (right) illustrate the result of this simulation. Again a similar trend is observed where the charge pump architecture outperforms the traditional architecture by several orders of magnitude.

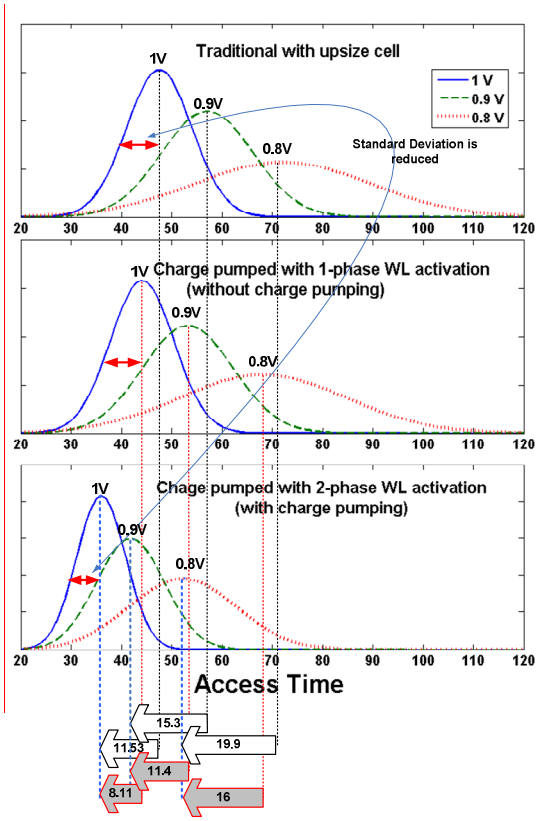

Figure 6: Comparison of the effect of process variation on access time in charge pumped and traditional architecture.

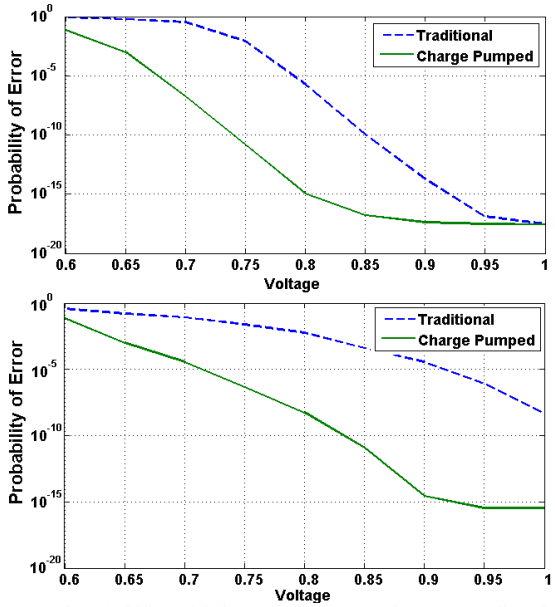

Figure 7: Total probability of failure (WF+AF+BF) for FFVS policy (left) and FSVS policy (right)

VIII. POWER CONSUMPTION SIMULATION RESULTS AND DISCUSSION

The charge pumped architecture consumes more power compared to traditional architecture when operated at the same supply voltage assuming that the charge pumping circuit is not power gated. However, the ability to tolerate process variation defects allows the charge pumped architecture to down scale the supplied voltage and operate at lower voltages or alternatively, achieve a higher production yield when operated at nominal voltage.

The extra power consumption is composed of two parts: leakage and dynamic power consumption. Leakage power consumption is increased because not only the number of transistors in each wordline driver is increased but also the capacitor that is used for charge pumping will add to the leakage when it is idle and not power gated. Furthermore, the charge pumped architecture requires a small defect map to identify the faulty locations. The extra dynamic power consumption on the other hand is a result of defect map query, overdriving the wordline driver and the extra switching activities by the charge pumping circuit. Over driving the wordline driver from "V_{dd}" to "$V_{dd} + \Delta V$" is made possible by charge sharing between the charge pump capacitor and the wordline capacitor. During the charge sharing process (Phase 2 of wordline driving) there is no path to voltage source but in the non-active part of the clock cycle the capacitors will be recharged and therefore it drains current from the supply voltage. As previously stated a policy to control the CP and CPB signals is needed since the frequency of CP and CPB signals will define the number of times the charge sharing is performed within one access. In this simulation we used a simple implementation where only after charge pumping the signals CP and CPB will switch. Using this method the charge sharing is only performed once and the capacitor that previously was discharged will have the off cycle of the current clock period and the entire next clock cycle to recharge. This behavior is controlled by the latch and XOR unit in the circuit illustrated in Figure3. In order to demonstrate the power savings, obtained as a result of voltage scaling, a spice simulation was setup. In this spice simulation the total power consumption of two memory banks identical in all aspects except the wordline drivers is compared.

At each voltage level the average power consumption over 10,000 read operations is obtained. The voltage scaling in each structure is possible for as long as the inequality in (6) holds:

$$E\left(P_F[WL]\right) = N_{Row}.P_F[WL] \le R \quad (6)$$

Where R is the redundancy budget (8 in this case), N_{Row} is the number of rows in the bank and $P_F(WL)$ is obtained from:

$$P_F(WL) = 1 - P_H(WL) \quad (7) \qquad P_H(WL) = \left[1 - P_F\right]^{N_{cells}} \quad (8)$$

In which P_F is the probability of failure (and could be replaced by P_{Fcp} or P_{Ftrad}) where P_{Fcp} is the probability of failure of the charge pumped architecture while P_{Ftrad} is the probability of failure of the traditional architecture. N_{cell} is the number of cells on each wordline. At each voltage the number of wordlines N_w that require charge pumping is expected to be:

$$N_w = E\left(P_{F_{Trad}}[WL]\right) = P_{F_{Trad}}[WL].N_{Row} \quad (9)$$

For the purpose of this simulation we used a FSVS policy for voltage scaling with P_{Fcp} and P_{Ftrad} drawn from Figure 7.

For the charge pumped architecture we need a defect map to store the faulty locations. In practice many configurations could be used for the defect map such as a very small SRAM always operating at full supply voltage, flash bits implemented at each wordline, or set of latches updated externally. In the following simulation we setup our defect map as a set of latches implemented on each wordline. To create a fair comparison we have included the power consumption and area of these latches as part of the analysis. Note that in very low voltages, the cache/SRAM architecture might have all the wordlines in the charge pumping mode. Although this means incremented dynamic power in almost all accesses, however still we significantly save on leakage since the leakage is managed using the lower supply voltage.

Figure 8 depicts the power savings of the two architectures. Based on the probabilities of the failure,
the expected number of weak rows (non operational at that voltage level) is obtained and compared between the traditional and the charge pumped architecture as depicted in Figure 8 (top). In all the simulations presented in Figure 8 a redundancy of 8 rows is assumed. Figure 8 (center) illustrates the dynamic power savings for the two architectures. It can be seen that a savings in dynamic power of 34% (on top of traditional voltage scaled approaches) is possible using the charge pumped architecture by running the array at a lower voltage.
In a 16KB SRAM, arranged in 256 rows, with each row containing 512 cells, and using a redundancy of 8 rows the traditional architecture, given the probability of failure in Figure 7 could only scale the voltage down to 0.87V. However, the charge pumped architecture can be downscaled to 0.67 volts while maintaining the same performance. The savings in leakage power are even more pronounced due to the exponential dependence on supply voltage as shown in Figure 8 (bottom), where a savings of more than 43% is achieved as compared to the traditional voltage scaling approaches. Finally, using the charge pump architecture results in a savings of 50% in dynamic power and 62% in leakage power when compared to a traditional architecture working at full voltage (1volt). These savings in power consumption could be improved with introduction of the power gating to the wordline drives and sharing a charge-pump between multiple wordlines.

IX. Conclusion

In this paper we presented a novel architecture for low power and high yielding memory arrays. Proposed approach utilizes a charge pump wordline driver and selectively overdrives the wordlines containing weak cells. This architecture achieves power savings of more that 50% in dynamic, and 60% in leakage power as compared to the traditional architecture running at nominal voltage. Alternatively, when operated at the same voltage as traditional memory it provides an improvement in memory array yield.

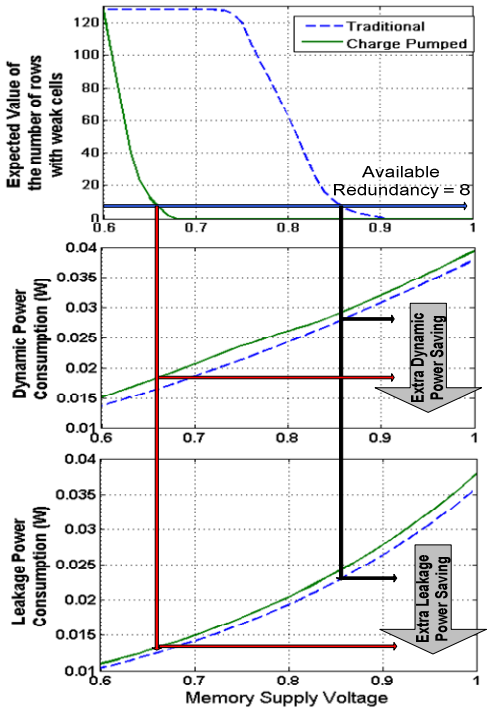

Figure 8: (Top): Expected number of rows containing defective weak cells (failing cells) for an FFV Voltage scaling policy with 120ps access time. (center): Dynamic power savings. (Bottom): Comparison of Leakage power savings.

References

[1] S. R. Nassif "Modeling and Analysis of manufacturing variation" in Proc. CICC, 2001 pp/ 223-228

[2] S. Borkar, T. Karnik, et al., "Process Variation and impact on circuits and micro architectures," in Proc DAC 2003 pp338-342

[3] S. Mukhopadhyay, H. Mahmoodi, K. Roy "Modeling of Failure Probability and Statistical Design of SRAM Array for Yield Enhancement in NanoScaled CMOS" CADICS Vol.24 NO. 12, DEC 2005

[4] A. Bhavnagarwala, X. et al. " The impact of intrinsic device fluctuation on CMOS SRAM cell stability," IEEE J. Solid-State Circuits vol.36, no.4 pp 658-665 Apr 2001

[5] H. Mahmoodi, at al.. "Modeling of failure probability and statistical design of SRAM array for yield enhancement in nano-scaled cmos," IEEE Trans CAD , 2003

[6] Avesta Sasan (Mohammad A Makhzan), Amin Khajeh, Ahmed Eltawil, Fadi Kurdahi, "Limits of Voltage Scaling for Caches Utilizing Fault Tolerant Techniques". ICCD 2007.

[7] S. E. Schuster, "Multiple word/bit line redundancy for semiconductor memories," IEEE J. Solid-State Circuits, vol. SC-13, no. 5, pp. 698–703, Oct. 1978.

[8] M. Horiguchi, "Redundancy techniques for high-density DRAMS," in Proc. 2nd Annu. IEEE Int. Conf. Innovative Systems in Silicon, Oct. 1997, pp. 22–29.

[9] A. Argawal, B. C. Paul, S Mukhopadhyay, K. Roy "Process Variation in Embedded Memories: Failure Analysis and Variation Aware Architecture." IEEE Journal of Solid State Cuircuits, VOL. 40, NO. 9, SEPTEMBER 2005

[10] H. L. Kalter et al., "A 50-ns 16-Mb DRAM with a 10 ns data rate and on chip ECC," IEEE J. Solid-State Circuits, vol. 25, no. 5, pp. 1118–1128, Oct. 1990.

[11] D. Weiss, J. J. Wuu, and V. Chin, "The on-chip 3-MB subarray-based third level cache on an itanium microprocessor," IEEE J. Solid-StateCircuits, vol. 37, no. 11, pp. 1523–1529, Oct. 1990.

[12] S. Mukhopadhyay, et al., "Statistical design and optimization of SRAM cell for yield enhancement," in Proc. Int. Conf. Computer Aided Design (ICACD), Nov. 2004, pp. 10–13.

[13] P. P. Shirvani and E. J. McCluskey, ''PADded Cache: A New Fault-Tolerance Technique for Cache Memories'', In Proc. Of 17th IEEE VLSI Test Symposium, pp.440-445, April 1999.

[14] http://www.eas.asu.edu/~ptm

[15] Bhalerao, et al., " A CMOS Low Voltage Charge Pump" VSLID 2007

[16] http://quid.hpl.hp.com:9082/cacti/

[17] G. Sohi, "Cache Memory Organization to Enhance the Yield of High Performance VLSI Processors", IEEE Trans. Comp., vol.38(4), , pp.484-492, April 1989

[18] Avesta Sasan (Mohammad A Makhzan), Houman Homayoun, Ahmed Eltawil, Fadi Kurdahi, "Architectural and Algorithm Level Fault Tolerant Techniques for Low Power High Yield Multimedia Devices". ICCD 2007.