LPRAM: A Low Power DRAM with Testability

Subhasis Bhattacharjee
University of Bristol
United Kingdom
Email: subhasis@cs.bris.ac.uk

Dhiraj K. Pradhan, *Fellow, IEEE*
University of Bristol
United Kingdom
Email: pradhan@cs.bris.ac.uk

[1] **Abstract:** To date all the proposal for low power designs of RAMs essentially focus on circuit level solutions. What we propose here is a novel architecture level solution. Our methodology provides a systematic trade off between power and area. Also, it allows tradeoff between test time and power consumed in test mode. Significantly, too, the proposed design has the potential to achieve performance improvements while reducing power. In this respect it stands apart from other approaches where the conventional wisdom of reducing power reduces speed.

I. Introduction

Progress in low-power VLSI technology including of low-power RAM designs is crucial for the future progress. Additionally, the success of future SOC depends heavily on innovations in low power embedded RAM design. All the previous works on RAM focus on circuit level solutions. There are mainly two directions in which research have been devoted to design of low power RAM [2], [3], [4], [5]. Specifically these are reduction in (a) charging capacitance [4], [7], (b) operating voltage [10], [7], [5].

The charging capacitance can be reduced by partial activation of Multi-divided Data Line (DDL) and/or Multi-Divided Word Line (DWL) [4]. [8] describes a method to reduce the charging capacitance during the data retention period along with the lowering of refresh frequency. However, lowering the refresh frequency shoots up the refresh busy rate unless devices are scaled properly to have higher *maximum refresh time*. Such modifications in devices are not always possible and is limited by the cell leakage current. Scheme to use of long word line for refresh operation and DDL/DWL for normal operation for reducing power is presented in [2].

The popular technique of lowering the supply voltage for low power design requires corresponding reduction of the threshold voltage (V_T). Apart from speed, the noise and DC current considerations also limit reducing the supply voltage arbitrarily.

Proposed methodology, here, departs radically from all these and provides an architectural high level solution. With our proposed design, we have achieved reduction in power consumptions for normal operation, refresh operation and during testing. This does not preclude application of lower circuit level techniques for low power design, in additions. Therefore, any existing circuit level techniques can also be applied to our proposed methodology to achieve further power savings. However, a unique feature of our design that cannot be accomplished through circuit approach that power reduction in our design is achieved with potential performance and test improvements. The speed of testing can also be varied allowing varying levels of power dissipations.

Further, our approach differs from earlier approaches [4] in that we don't share decoders between cell array partitions - therein providing additional power savings.

[1] This research was supported by EPSRC(UK) and is based on a patent filed in [11]

II. Proposed Architecture

The proposed architecture partitions the RAM into a number of modules, where each is a smaller RAM module with decoder and refresh circuitry. The modules are then interconnected by a H-tree which provides for planner layout and incorporation of a particular built-in-self-test technique. Major power/performance tradeoffs is achieved by allowing the modules to have arbitrary aspect ratios. Our design allows switching off of portions of the RAM during both normal operation and testing. Such a dynamic reconfiguration capability allows for smooth trade off of test application time and power dissipation during test.

III. Design Overview

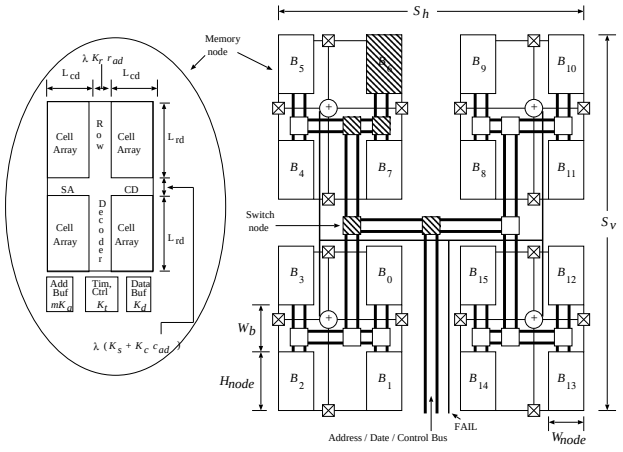

Fig. 1. LPRAM architecture including details of memory node, test structures

Our design for low-power RAM, assumes the $M = 2^m$ cells divided into equal-sized modules, M representing the size of the RAM in bits and m, the number of address lines (assuming an $M \times 1$ bit organization). These modules appear as leaf nodes

in a complete binary tree. The depth of the tree, l, and the number of modules or leaf nodes, Q, are related by $Q = 2^{l-1}$. The size of each node is $N = 2^n$, where $m = n + l - 1$. Note that the root node is at level 1. The parameters n and l define the properties of this architecture. A large l means a higher granularity, a higher degree of power saving, speed-up and testability, with increased chip size.

A Simplified Model of RAM for Comparisons: In this paper we assume a simplified model of RAM as shown in encircled portion Fig. 1. This model is used for both conventional RAM and the modules used in proposed architecture. The reason we use the simplified model is that it admits developing simple and accurate expressions for comparing power estimates as shown later.

Based on our Simplified Model of RAM, we observe that a conventional RAM can be thought as a special case of LPRAM with $l = 1$. In conventional RAM, $M = 2^m$ cells are arranged in 4 quadrants, each holding $\frac{M}{4}$ cells arranged in a two dimensional matrix of ρ_c rows and η_c columns. The address bus is divided into two equal (near equal, when m is odd) parts, one half is used to decode the row and the other to select the column. For the sake of comparison we assume a 4 quadrant architecture but the architecture allows each module to be built out of more number of cell array partitions.

Basically, two types of nodes are used in our design: memory nodes and switch nodes. Memory nodes have the cell array based on the traditional multi subarray (eg. four quadrant) organization with independent control units, refresh circuitry and certain built-in test circuitry. Each module itself can also be designed with larger number of sub arrays as done in current designs. For the sake of modeling, we propose that each module containing N cells is arranged in 4 quadrants, each quadrant holding $\frac{N}{4}$ cells. Each quadrant is a two-diamentional array of memory cells arranged in ρ_{mn} rows, each row containing η_{mn} cells. But, unlike conventional RAM, we divide the address bus (n address lines) into two parts r_{ad} and c_{ad} respectively, to give preferably a non-unit aspect ratio. These r_{ad} and c_{ad} address lines are separately decoded in the row and column decoders, respectively, to give $\rho_{mn} = 2^{r_{ad}}$ rows and $\eta_{mn} = 2^{c_{ad}}$ columns. We define α_{node}, the aspect ratio of the memory node in LPRAM, equals to η_{mn}/ρ_{mn}. Additionally, each memory node contains some tri-state switches on the runs of power line(s) to cut it off from the power source when require. The number of such switches will depend on the maximum number of elements active at any time and on the power line layout.

The switch nodes are simple 1-out-of-2 decoders with buffers. As Fig. 1 shows, the memory nodes are connected hierarchically, using the switch nodes, and laid out in a H-tree layout. Let each memory node be identified by B_j , where $0 \leq j < Q$. Therefore, as shown in the Fig. 1 (for $Q = 16$), the nodes are numbered $B_0, B_1, \cdots, B_{Q-1}$, consecutively numbered nodes are adjacent to each other in the layout.

The address/data/control bus is connected to the root, a switch node. The most significant bit is decoded, generating a left subtree or a right subtree select. The other signals are buffered and propagated down the tree. This action occurs repeatedly at each level until a single memory node is selected. At this point, the remaining address bits (n) are latched into the address buffers of the selected memory node only, and are then used to select a cell within the node. The address buffers of all other non-selected nodes remain completely unchanged, thereby nullifying any possibility of activity within them (other than normal refresh activity). Each cell be identified by the address $A_{i,j}$, where $0 \leq i < Q$ (node address) and $0 \leq j < (M/Q)$ (address within a node).

Test Structure: Fig. 1 shows the test structure to be used for the proposed low-power LPRAM during testing. All these Q modules have been divided into $\pi = Q/q$ quadrants (shown as dotted boundary in Fig. 1), each quadrant holding q (a power of 2, assumed 4 in the figure) modules. So, we have $Q = \pi \times q$. In each quadrant, comparators are placed between adjacent modules, shown in Fig. 1, $B_i, B_{(i+1) \bmod q}$, $0 \leq i < q$. The output of all those comparators is fed to a q input *OR* gate, centrally located in that quadrant. The output of all these π *OR* gates is tagged and sent as a single *FAIL* line, to generate error during testing. So, in each quadrant, all the *modulo* q adjacent nodes are compared simultaneously, eventually leading to a speed-up of q fold during testing.

Test Mode Operation: The LPRAM can be put into test mode by activating the *TEST* pin. Test data is fed into LPRAM, as usual, through the external tester by addressing as $A_{j,i,k}$, $0 \leq j < \pi$, $0 \leq i < q$. These i bits are ignored during testing, and data is written parallel into all q nodes simultaneously in the i th quadrant. Testing proceeds by activating each one of these π quadrants, one at a time. Identical data is simply written into all the modules in the quadrant, the data is then read back and compared against each other internally for test. Thus, all modules in a quadrant can be tested simultaneously.

IV. Power Estimation Model and Comparisons

The following is the active power equation for CMOS RAM of size $4 \times \rho \times \eta$ cells (i.e. $\rho \times \eta$ cells arranged in ρ rows and each containing η cells in each quadrant of a four-quadrant memory module), given by [2]

$P = [\eta \cdot i_{act} + \eta \cdot (\rho - 1) \cdot i_{hld} + (\rho + \eta) \cdot C_{DE} \cdot V_{INT} \cdot f_{OP} + C_{PT} \cdot V_{INT} \cdot f_{OP} + I_{DCP}] \cdot V_{DD}$

where V_{DD} is an external supply voltage, i_{act} is the actual current drawn by the selected cells, and i_{hld} is the data retention current required by any inactive or non-selected cell. C_{DE} is the output node capacitance of each decoder, V_{INT} is the internal supply voltage, C_{PT} is the total capacitance of the CMOS logic and driver circuits in the periphery. Let I_{DCP} represent the total static (DC) current of the periphery, and f_{OP} is the operating frequency.

The above equation can be simplified for high frequency DRAM operation (Fig. 2), and by the use of CMOS NAND Decoder, as well as by elimination of very low dc components, yielding the following reasonable approximation.

Data Reading Power of Conventional DRAM: The destructive readout of a DRAM cell requires successive operations of amplification and restoration for the selected cell on every data read. Here, each cell is, basically, a trench capacitor, requiring charging and discharging during each reading. This is accomplished by a latch-type CMOS sense amplifier on each data line. So, during the reading of a data line, the associated trench capacitor is charged and discharged with a large voltage swing

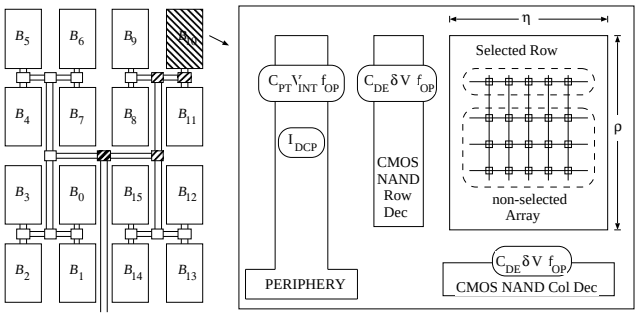

Fig. 2. Power dissipation model

of ΔV_D (usually $1.5 \sim 2.5$ V) and with charging current of $C_D \Delta V_D f$, where C_D is the data line capacitance. The active power consumption during read is given by :

$P_{read} = [\{\eta_c \cdot C_D \cdot \Delta V_D + C_{PT} \cdot V_{INT}\} f_{OP} + I_{DCP}] \cdot V_{DD}$

Data Retention Power of Conventional DRAM: In the data retention mode, internal data is retained and refreshed without any access from outside. The refresh operation is performed by reading data of all the cells on a single word line and restoring them to their original values. The refreshing circuitry selects each of the word lines in order, and during the whole time (called refresh busy time), the RAM is not accessible from the outside. For high performance RAMs, refresh busy time is expected to be as low as possible. The refresh cycle frequency equals ρ_c / t_{REF}, where t_{REF} is the refresh time interval of cells in the retention mode, and increases with reducing junction temperature. In general, t_{REF} is much smaller than the $t_{REF_{max}}$ which is provided in specification and depends on the cell technology for the trench capacitor. The power consumed for refreshing η_c cells can be derived as

$P_{reten} = [\{\eta_c \cdot C_D \cdot \Delta V_D + C_{PT} \cdot V_{INT}\}(\rho_c / t_{REF}) + I_{DCP}] \cdot V_{DD}$

It follows from the eqations of P_{read} & P_{reten} that the following factors are crucial to reduce the power during any read/write cycle: 1) reducing charging capacitance ($\eta_c \cdot C_D \cdot \Delta V_D$, C_{PT}), 2) lowering the external and internal voltages (V_{DD}, V_{INT}, ΔV_D), and 3) reducing static current (I_{DCP}), 4) reducing refresh cycle frequency (ρ_c / t_{REF}). As mentioned, several techniques have been offered to reduce circuit parameters. These techniques can be used in conjunction with our proposed architectural solution to low power design. It is to note that reducing design parameters like η and ρ can also reduce power consumption. Therefore, for instance, if previous researchers have proposed segmenting the word line, the proposed low power architecture allows a systematic way to reduce η. It further allows reduction of the ρ to reduce power, not possible previously by the circuit level techniques.

In the LPRAM, data is read out or written into by first choosing a selected module by the tree decoder, power is being dissipated only by the decoder (switch nodes) on its path (Fig. 2). The address is then decoded in selected modules to locate the exact cell containing data. For example, in Fig. 2, only those switch nodes that are hatched consume power while reading a data from module B_6. No other switch node is activated at all. This observation is used for modelling power for switching nodes.

Consider the example of a 16 M of DRAM, with $\eta_c = 8192$ and $\rho_c = 2048$. The same size of RAM implemented using LPRAM architecture will have $\eta_{mn} = 1024$ and $\rho_{mn} = 1024$, with 16 nodes of 1M each. In addition, if DWL is used for 16 divisions of the word line, then $\eta_{DWL} = 512$ for traditional RAM, and the corresponding value for LPRAM is 64. The power reduction in the proposed RAM is achieved primarily by reducing these parameters. In the following we develop various equations for power estimates.

Data Reading Power of LPRAM: The data read out power for LPRAM can be formulated as

$P_{LPRAM_{read}} = [\{\eta_{mn} \cdot C_D \cdot \Delta V_D + C_{PT} \cdot V_{INT} + C_{tree} \cdot V_{INT}\} f_{OP} + I_{DCP}] \cdot V_{DD}$

where η_{mn} could be as small as η_c / Q and C_{tree} (derived below) is the effective capacitance seen in the tree decoder of LPRAM.

Estimation of C_{tree}: In the tree decoder, each switch node consists of a simple 1-out-of-2 decoder and buffers. The decoder is a one bit decoder consisting of one level of logic. Additionally, each decoded signal is controlled by the preceding subtree select (chip enable for the first level), and this introduces another level of logic. In each switch node, one bit of the address is decoded and the rest of the address bits are simply transmitted. At each node, the signal has to drive a load of $2C_g$ (two gates each, offering a load of C_g), and the output gate has a drive capability of $2C_g$. The bus width is assumed to be 3λ.

All bus lengths of the tree are computed with respect to S_v, the length of the vertical side of the LPRAM Fig. 1. The input buffer drives the bus up to the root node-length $(S_v)/2$. Let l, the number of levels in the tree, be assumed odd. The length of the bus connecting level 1 to level 2 is $S_v/2$. Thus, if c_{mf} is the capacitance of metal over field oxide, then the load offered by the bus, between levels 1 and 2, is $(S_v/2)3\lambda c_{mf}$. Each node is connected to two nodes at the next lower level. Therefore, a buffer at level i has to drive *two* buffers at level $i+1$, each offering a load of $2C_g$. Thus, this load can be modeled as $4C_g$. The total load that has to be driven at level 1, by the second gate, is $((S_v/2)3\lambda c_{mf} + 4C_g)$ and is in parallel to $2C_g$. The total capacitance seen at level 1 can therefore be represented as $(S_v/2)3\lambda c_{mf} + 6C_g)$.

The capacitance at level 2 is the same as level 1 because the bus lengths are the same. Further, after every two successive levels, the length of the bus to be driven decreases by half of the level before. For example, level 3 and level 4 have to drive buses of lengths of $S_v/4$; and subsequent levels 5 and 6 have to drive buses of length $(S_v/8)$, and so on. Let $S_M = S_v 3\lambda c_{mf}$. In general, the bus length to be driven by the node at level i can be expressed as $(S_M / 2^{\lceil \frac{i}{2} \rceil})$. A tree of depth l will have $(l-1)$ decoding stages. Therefore, the total capacitance over the entire tree, from level 1 to the leaf nodes, can be modeled by their parallel operation, given by $\sum_{i=1}^{l-1} (S_M / 2^{\lceil \frac{i}{2} \rceil}) + 6C_g$, which evaluates to $S_M 2(1 - 2^{-(l-1)/2}) + 6(l-1)C_g$. So, the capacitance value seen from the root of the tree to the accessed node is given by

$C_{tree} = 2 \cdot S_v 3\lambda c_{mf} \cdot (1 - 2^{-(l-1)/2}) + 6 \cdot (l-1) C_g$

Data Retention Power of LPRAM: The LPRAM achieves a corresponding reduction in retention power, as well, because of the reduction in both η and ρ architectural parameters. The equation for data retention power is given by,

$P_{LPRAM_{retention}} = [(\rho_{mn} / t_{REF})\{\eta_{mn} \cdot C_D \cdot \Delta V_D + C_{PT} \cdot$

$V_{INT}\} + I_{DCP}] \cdot V_{DD}$

Refreshing is done independently within each module. Also, we have $\rho_c \cdot \eta_c = Q \cdot \rho_{mn} \cdot \eta_{mn}$. If we assume γ is of the form $2^x \leq Q$ as the ratio between η_c and η_{mn}; i.e., $\gamma = \frac{\eta_c}{\eta_{mn}}$, then $\rho_{mn} = \frac{\rho_c \cdot \gamma}{Q}$. So, by appropriate choice of γ both the data read out power and the data retention power can be reduced!

We have calculated the power dissipation of the proposed LPRAM for a large range of module sizes, and for four (4) different RAM sizes, 4 M, 16 M, 64 M and 256 M. The reduction in power dissipation over the traditional RAM is illustrated in Fig. 3 with a range of aspect ratios of individual memory node. These savings are shown as percentage of reduction in power dissipation over the same size of conventional RAM. From Fig. 3 we see for the same size of RAM, we achieve greater access power savings when aspect ratio of individual memory node (α_{node}) is more. However, when aspect ratio become too large the retention power increases. But, from Fig. 4 we also see that there is a saving of the retention power as well.

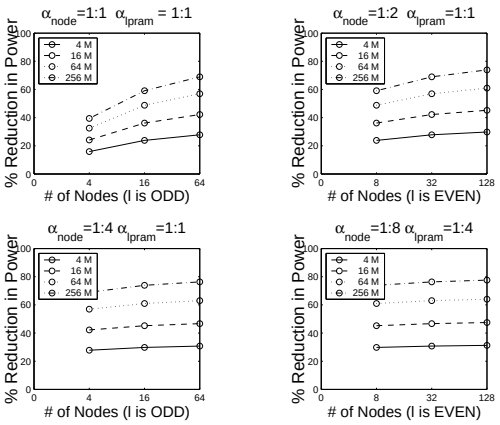

Fig. 3. Access Power Reduction in LPRAM

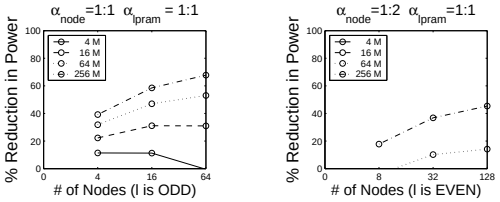

Fig. 4. Retention Power Reduction in LPRAM

V. Testing

The testability technique used here enables the q of Q nodes to be tested in parallel, as mentioned earlier. Depending upon the size of the RAM and the number of modules in LPRAM, we set the value of q. Thus, we get a test time saving of q fold, without dissipating much power, as well. A test algorithm with $C(M)^k$ steps now definitely requires $\pi * c(N)^k$ steps only. Testing the RAM involves three sets of tests: 1) testing the tree decoder, 2) testing the built-in test structure (BITS) and 3) testing the memory nodes. will discuss the testing of the memory nodes only, the test procedure of the other parts being the same as given in [1].

Comments About Power During Test: As all q modules are tested simultaneously, the instantaneous power consumption during testing also grows closer to q fold. But as LPRAM consumes very low power for accessing, and the test data is written and read locally within the quadrant, with q up to 4, we still get a reduction of about 20% power in 256 M RAM, depicted in the Fig. 5, compared to the traditional RAM when we apply MATS procedure, presented in [1]. At the same time we get a 4 times reduction in test time.

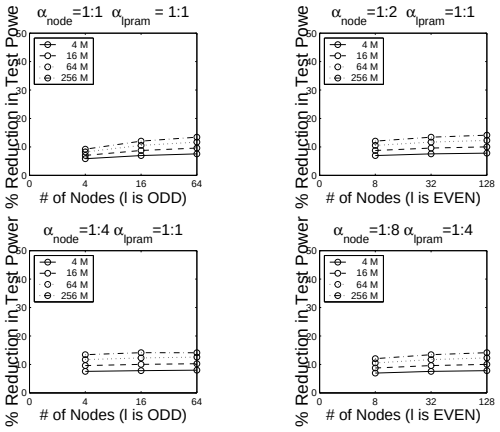

Fig. 5. Reduction in power during testing in LPRAM ($q = 4$)

VI. Conclusion

An architecture for low-power high-performance RAM is proposed. The LPRAM architecture saves about 35% power during normal operation for a 256M RAM compared to the traditional RAM. Also, for a 256M RAM, LPRAM provides about 20% reduction in power during testing, with 75% saving in test time due to the presence of BITS. Thus, it reduces power consumption both during normal operation and testing.

References

[1] Najmi T. Jarwala, D. K. Pradhan: "TRAM: A Design Methodology for High-Performance, Easily Testable, Multimegabit RAM's", *IEEE Transactions on Computers*, vol. 37, no. 10, pp. 1235-1250, October 1988.

[2] K. Itoh, K. Sasaki and Y. Nakagome: "Trends in low-power RAM circuit technologies", *Proceedings of the IEEE*, vol. 83, no. 4, pp. 524-543, April 1995.

[3] K. Itoh: "Trends in megabit DRAM circuit design", *IEEE J. Solid-State Circuits*, vol. 25, pp. 778-789, June 1990.

[4] K. Kimura *et al.* : "Power reduction in megabit DRAM's", *IEEE J. Solid-State Circuits*, vol. 21, pp. 381-389, June 1986.

[5] M. Margala and N. G. Durdle : "Noncomplementary BiCMOS logic and CMOS logic styles for low-voltage operation - A Comprehensive Study", *IEEE J. Solid-State Circuits*, vol. 33, no. 10, pp. 1580-1585, October 1998.

[6] A. Bellaouar and M. I. Elmasry : "Low-Power Digital VLSI Design, Circuits and Systems", *Kluwer Academic Publishers*, 1996.

[7] J. S. Caravella : "A Low-voltage SRAM for Embedded Applications", *IEEE J. Solid-State Circuits*, vol. 32, no. 3, pp. 428-432, October 1998.

[8] D. C. Choi *et al.* : "Battery operated 16 M DRAM with post package programmable and variable self refresh", *Symp. VLSI Circuit Digital Technical Papers*, pp. 83-84, May 1994.

[9] N. H. E. Weste, K. Eshraghian : "Principles of CMOS VLSI Design : a systems perspective", *Addison-Wesley* 1992.

[10] N. C. C. Lu and H. Chao : "Half-V_{DD} bit-line sensing scheme in CMOS DRAM", *IEEE J. Solid-State Circuits*, vol. 19, no. 8, pp. 451-454, August 1984.

[11] D.K.Pradhan : "Low power RAM", *British Patent application number:0315292.3*