Low Power SRAM Design using Hierarchical Divided Bit-Line Approach *

Ashish Karandikar †
Intel Corporation, Santa Clara, CA 95052, USA
E-mail: akarand1@td2cad.intel.com

Keshab K. Parhi
Dept. of Electrical and Computer Engineering
University of Minnesota, Minneapolis, MN 55455, USA
E-mail: parhi@ece.umn.edu

Abstract

This paper presents a novel hierarchical divided bit-line approach for reducing active power in SRAMs by reducing bit-line capacitance. Two or more 6T SRAM cells are combined together to divide the bit-line into several sub bit-lines. These sub bit-lines are again combined to form two or more levels of hierarchy. This division of bit-line into hierarchical sub bit-lines results in reduction of bit-line capacitance, which reduces active power and access time. Optimum values for number of levels of hierarchy and number of blocks combined at each level have been derived. Experimental results show that the observed parameters and estimated ones follow the same trend. It is shown that the reduction in bit-line capacitance reduces active power consumption by $50-60\%$ and the access time by about 20% at the expense of approximately 5% increase in the number of transistors. This approach is further extended by incorporating the controlled voltage swing on bit-lines. This extension reduces the power consumption by another $20-30\%$.

1 Introduction

Designing a low power system not only reduces weight and size of batteries for portable systems but also helps in reducing the ever-important packaging costs of integrated circuits. To this end, the design of low power digital systems is becoming increasingly important. With memories accounting for the largest share of power consumption in the processors, an emphasis has been placed on the design of low power memories [4] [3] [7].

As described in [4], active power is a major component of the total memory power. This paper first describes a novel *divided bit-line approach* for reducing the active power by reducing the bit-line capacitance and then extends it to a *hierarchical* divided bit-line approach. It is shown that by reducing this capacitance, not only power reduction is achieved but access time is reduced as well. While hierarchical word-decoding has been used to reduce power consumption by reducing the number of columns activated during a read operation [9][8], this paper is the first attempt to reduce SRAM power consumption by the divided bit-line concept. Another advantage of this approach is the stability of the SRAM cells. By dividing bit-line into sub bit-lines, SRAM cells become more stable, as they are guarded from the noise on bit-lines through pass transistors.

In section 2, the *divided bit-line* approach is described and an optimum value for the number of SRAM cells to be combined is derived. Section 3 describes the *hierarchical divided bit-line* approach. Section 4 presents some experimental results. Section 5 presents an extension to the proposed approach by limiting the voltage swing on bit-lines which leads to further savings in power consumption. Finally, section 6 concludes the paper.

*This work was supported by the Defense Advanced Research Project Agency under contract number DA/DABT63-96-C-0050.

†This work was performed when author was with the University of Minnesota.

2 Divided Bit-Line Approach

2.1 Concept

Power consumption in SRAMs, for a normal read cycle, is given by

$$P = V_{dd} \times I_{dd} \quad (1)$$

$$I_{dd} = (mI_{act}\Delta t + C_{PT}V_{INT})f + I_{DCP} \quad (2)$$

where, V_{dd} is an external supply voltage, I_{dd} is the total current, I_{act} is the effective active current, V_{INT} is an internal supply voltage, C_{PT} is the total capacitance of the peripheral circuits, I_{DCP} is the total static current, m is the number of columns and f is the operation frequency. This equation is based on the fact that in SRAMs, holding current is very small [4] and decoder charging current is negligible because of NAND decoders [4] [6]. To reduce the total power consumption, active current should be reduced as it dominates the total current. Active current is the current that flows during word line activation, *i.e.,* during charging or discharging of bit-line capacitance. This active current is directly proportional to bit-line capacitance. Divided word line and hierarchical word decoding approaches reduce the I_{dd} by reducing the value of m, *i.e.,* the number of columns activated during a read operation [9][8]. Approaches involving the pulse operation of word line and column circuitry reduce the I_{dd} by reducing the value of Δt in (2) [4]. In contrast, the approach presented in this paper reduces I_{dd} by reducing the active current, I_{act} in (2).

The total effective charging current flowing through a bit-line, during a read operation, can be expressed as

$$I_{eff} = C_{eff} \times \frac{\Delta V}{\Delta t} \quad (3)$$

where C_{eff} is the effective bit-line capacitance, ΔV is the voltage swing on the bit-line and Δt is the word line activation time.

Similarly, expression for power can be written as,

$$Power = (I_{eff} \times V \times \Delta t) \times f \quad (4)$$

$$= (C_{eff} \times \Delta V \times V) \times f. \quad (5)$$

If the same voltage swing is allowed, then power consumed during read or write is directly proportional to the capacitance of bit-line.

The bit-line capacitance is mainly composed of the drain capacitance of the pass transistors of the SRAM cell and metal capacitance of bit-line. To reduce this capacitance, drain capacitance and metal capacitance should be reduced. Bit-line capacitance can be reduced

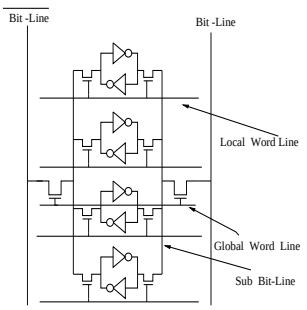

Figure 1. Divided Bit-Line Approach, M = 4

by the proposed *divided bit-line approach*, where the number of transistors connected to the bit-line is reduced by combining two or more SRAM cells. Fig. 1 shows 4 SRAM cells combined together and connected via one pass transistor to the bit-line. Thus, the number of pass transistors connected to the bit-line is reduced by 4.

2.2 Theoretical Basis

In this subsection, we provide a theoretical basis for our approach and derive an optimal value for the number of SRAM cells to be combined. The bounds for signal delay in RC tree networks have been derived in [5]. Similar bounds can be obtained for the access time of SRAM by modeling the pass transistors connecting pull-up and pull-down transistors of the SRAM cell with bit-lines as resistors. Though, pass-transistors are not always in linear region of operation, but modeling them as resistors gives us a suitable first order approximation. Fig. 2 shows the modeling of pass-transistors as RC chain. As bit-lines are always precharged before reading, analysis is performed for $\overline{bit-line}$ (or bit-line) which has to be pulled down to read a '1' (or a '0').

By using the equations and notations given in [5], we can write the access time T_{delay} in terms of parameters T_P, T_{D2} and T_{R2} as

$$T_{D2} - T_{R2} + T_{R2} \ln \frac{T_{R2}}{T_P[v(t)]} \leq T_{delay} \leq T_P - T_{R2} + T_P \ln \frac{T_{D2}}{T_P[v(t)]} \quad (6)$$

where $v(t)$ is a normalized voltage with respect to the supply and is given by

$$v(t) = 1 - \frac{\Delta V}{V}. \quad (7)$$

The parameters T_P, T_{D2}, T_{R2} are given by following

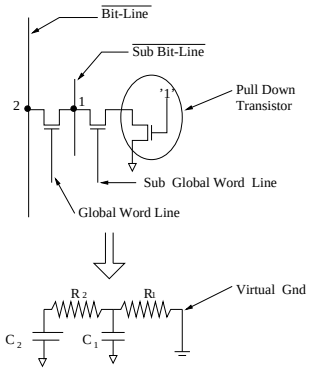

Figure 2. R-C model for Pass transistors, reading '1' stored in SRAM cell

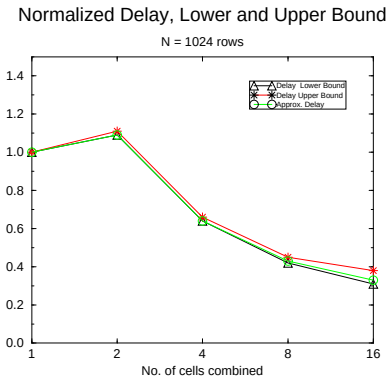

Figure 3. Normalized upper bound, lower bound and approximate delay

equations,

$$T_P = T_{D2} = R_1C_1 + (R_1 + R_2)C_2 \tag{8}$$

$$T_{R2} = \frac{R_1^2C_1}{R_1 + R_2} + (R_1 + R_2)C_2 \tag{9}$$

where R_1 is the resistance of the pass-transistor 1 and R_2 is the resistance of the pass-transistor 2. C_1 and C_2 are the capacitances at node 1 and at node 2, respectively, in Fig. 2.

Let us assume that the number of rows in the memory array is N and the number of cells combined in the *divided bit-line* is denoted by M. Then, the capacitances C_1 and C_2 can be obtained as

$$\begin{aligned} & & C_1 &= C\frac{M+1}{N} \\ for \quad N \gg 1, & & C_1 &= C\frac{M}{N} \qquad (10) \\ & & C_2 &= \frac{C}{M} + 0.1 \times C \qquad (11) \end{aligned}$$

where C denotes the original drain capacitance of N rows. Metal capacitance contribution to total bit-line capacitance is assumed to be 10% of the total drain capacitance. To make a first order approximation, we can assume the resistances of two pass-transistors to be equal.

With the approximation $R_1 = R_2 = R$, and from (9)(10), we can rewrite (7) as

$$\begin{aligned} & RC(\frac{M}{2N}) + RC(\frac{M}{2N} + \frac{2}{M} + 0.2) \\ & \times [\ln(\frac{\frac{M}{2N} + \frac{2}{M} + 0.2}{\frac{M}{N} + \frac{2}{M} + 0.2}) + \ln(\frac{1}{v(t)})] \\ & \leq T_{delay} \leq \\ & RC(\frac{M}{2N}) + RC(\frac{M}{N} + \frac{2}{M} + 0.2)\ln(\frac{1}{v(t)}). \end{aligned} \tag{12}$$

For standard SRAM cells, by using the same model as given in (7), we can write the delay as

$$T_{delay}(standard) = 1.1RC \times \ln(\frac{1}{v(t)}). \tag{13}$$

If we assume $M \ll N$, we can rewrite (11) as

$$T_{delay} \approx RC(\frac{M}{N} + \frac{2}{M} + 0.2) \times \ln(\frac{1}{v(t)}). \tag{14}$$

This equation is equivalent to the well known capacitance discharge through a series resistance. Using this approximated delay we can write normalized delay as

$$T_{delay}(normalized) = \frac{1}{1.1} \times [\frac{M}{N} + \frac{2}{M} + 0.2]. \tag{15}$$

If we assume a voltage swing of 10% on bit lines, we can plot lower and upper bounds for delay and approximate delay, using (11)(12)(13), as a function of M for a given value of N. Fig. 3 shows such plots.

Next, we find an optimal value of M for minimum power consumption. Power consumption, as a function of N and M, can be expressed as

$$\begin{aligned} & power = f(M) = (C_2 \times \Delta V \times V + 2 \times C_1 \times V^2) \times f \\ & = [(\frac{C}{M} + 0.1 \times C) \times \Delta V \times V + 2 \times C(\frac{M}{N}) \times V^2] \times f. \end{aligned} \tag{16}$$

The above equation is derived from the fact that sub bit-lines are not precharged and they can swing to full supply voltage during a read operation. Normalized power, with respect to standard SRAM, can be given as

$$Power(normalized) = (\frac{1}{M} + 0.1) + 2 \times \frac{M}{N \times \frac{\Delta V}{V}}. \tag{17}$$

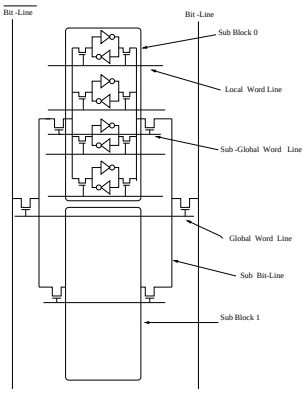

Figure 4. Hierarchical Divided Bit-Line Architecture, L = 3, M_1 = 4, M_2 = 2

By differentiating (15) with respect to M and solving this, we get an optimum value of M for minimum power as

$$M_{opt} = \sqrt{(\frac{N}{2} \times \frac{\Delta V}{V})}. \quad (18)$$

For example, with $N = 1024$ and $\frac{\Delta V}{V} = 0.1$, we get $M_{opt} \approx 8$.

3 Hierarchical Divided Bit-Line Approach

In high density SRAMs, the number of sub bit-lines will increase and even with *divided bit-line* architecture, bit-line capacitance can be significant. From this point of view, *hierarchical divided bit-line* architecture is developed. Fig. 4 shows the concept of *hierarchical divided bit-line* approach. In this architecture, the bit-line is divided into more than two levels. The number of hierarchies, as it will be shown further, is determined by the number of rows in the SRAM array.

Theoretical results of previous section on *divided bit-line* can be extended to the *hierarchical divided bit-line* architecture. Let us assume that the total number of levels in the hierarchy is L and at each level i, the number of blocks combined to form a new block is M_i. Then capacitance, C_i at each node is given by

$$\begin{aligned} & C_i = C \times \frac{M_i + 1}{N} \quad \forall \quad i \neq L \\ & for \quad N \gg 1, \\ & C_i = C \times \frac{M_i}{N} \\ & C_L = C \times \frac{1}{\prod_{i=1}^{L-1} M_i}. \end{aligned} \quad (19)$$

Using these capacitances and again making the assumption that resistance values of all pass transistors are equal, we can write the expression for active power as

$$\begin{aligned} & Power = f(M_1, ..., M_{L-1}, L) \\ & = \frac{2V^2C}{N} \times \sum_{i=1}^{L-1} M_i + \frac{C\Delta VV}{\prod_{i=1}^{L-1} M_i} + 0.1C\Delta VV. \end{aligned} \quad (20)$$

Similarly, we can write the expression for delay for hierarchical bit-line as

$$T_{Delay} = RC \times (\sum_{i=1}^{L-1} \frac{M_i i}{N} + \frac{L}{\prod_{i=1}^{L-1} M_i} + 0.1 \times L). \quad (21)$$

To obtain the optimum parameter for the number of blocks combined at each level, we differentiate (19) partially with respect to all M_i's and solve the following equation

$$\frac{\partial f}{\partial M_i} = 0 \quad \forall \quad i = 1 \; to \; L-1. \quad (22)$$

The solution to the above equation gives the optimum values for M_i's, which are given by

$$M_1 = M_2 = \cdots = M_{L-1} = (\frac{N}{2} \times \frac{\Delta V}{V})^{\frac{1}{L}}. \quad (23)$$

By substituting these values for M_i's in (19) and (20), expressions for Power and Delay in terms of N and L can be obtained as

$$Power = \frac{2V^2C}{N} L \times (\frac{N}{2}\frac{\Delta V}{V})^{\frac{1}{L}} + 0.1C\Delta VV \quad (24)$$

$$\begin{aligned} T_{Delay} &= RC \times (\frac{(L-1)L}{2} \times (\frac{N}{2} \times \frac{\Delta V}{V})^{\frac{1}{L}}) \\ &+ RC \times (\frac{L}{(\frac{N}{2} \times \frac{\Delta V}{V})^{\frac{L-1}{L}}} + 0.1L). \end{aligned} \quad (25)$$

The optimum value of L can be obtained by differentiating (23) and is given as

$$L = \ln(\frac{N}{2} \times \frac{\Delta V}{V}). \quad (26)$$

Fig. 5 shows the variation in power and delay with $N = 1024$ and with optimum values of M_i's for different values of L.

4 Experimental Results

To test our ideas, we have modeled a SRAM test chip. This modeling is based on the standard capacitance extraction and resistance estimation techniques

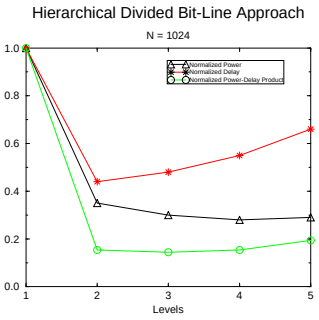

Figure 5. Normalized Power, Delay and Power-Delay product for different levels with N = 1024 and optimum M_is

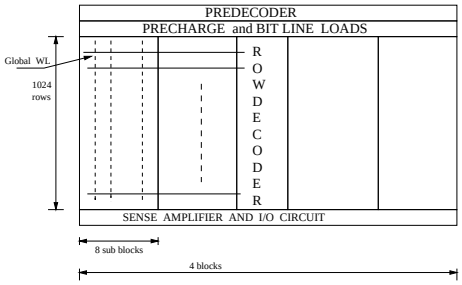

Figure 6. Model of Test SRAM

described in [10]. To model our test SRAM, we have assumed the configuration as shown in Fig. 6. We have assumed the hierarchical word decoding and standard precharging and sense amplifier techniques. Circuits for precharging and sense amplifier are the same as given in [2]. The experimental results for memories based on the *divided bit-line approach* are shown in Fig. 7. These results show that the experimental observations and theoretically obtained results follow the same trend. It is clear from the plots that the delay does not reduce as much as that predicted by the plots shown in Fig. 3. This is mainly because of our simplified assumption of same resistance for both pass-transistors. Fig. 7 also shows the results for hierarchical divided bit-line approach. As it is evident from the plots, the delay again follows the same trend. However it does not exactly follow the same trend as predicted for power. This is mainly because of our approximation of optimum value of M_i's to the nearest power of 2. This approximation is assumed as it makes the design of decoders simple.

These results show that the power consumption is reduced approximately by $50 - 60\%$ and access time is reduced by $20 - 30\%$. It is important to note that this reduction in power is valid only for active power

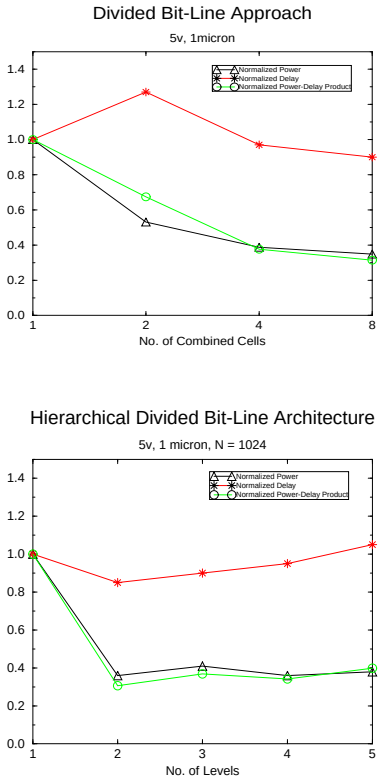

Figure 7. Experimental Results for Divided Bit-Line and Hierarchical Divided Bit-Line architecture.

and does not take into account the power consumed by sense amplifier and decoders.

To test our ideas in silicon, we have custom designed a $2K \times 8$ bits SRAM chip using MAGIC layout tool. The architecture of this chip is based on the model of Fig. 6 and 8 cells are combined at the sub bit-line level. The complete layout of the chip is shown in Fig. 9. This chip contains approximately $105K$ transistors and requires 6.81 mm^2 area in 0.5μ technology. The simulation results on this chip match very closely with that predicted by the model.

5 Extension of Divided Bit-Line Approach

Power consumption in memories can also be reduced by reducing the voltage swing on the bit-lines as active current also depends upon the voltage swing on the bit-lines (see (3) and (4)). Voltage swing can be controlled by using the word line and creating a desired pulse on the word line by using a replica feedback [1]. The technique described in [1] uses one extra reference column and one row for creating a pulse. Using the technique proposed in [1], the *divided bit-line approach* can be modified to provide limited voltage swing. In *divided bit-line approach*, bit-line is divided into sub bit-lines of SRAM cells. The basic idea behind this extension is to make use of these sub bit-lines, which have much smaller capacitance, to develop a limited voltage swing on the bit-line. The charge developed on the bit-lines due to precharging is shared with the much smaller capacitance of the sub bit-line to produce a limited swing. This charge sharing between a small capacitance of sub bit-line and the large capacitance of bit-line produces a limited voltage swing on the large capacitance.

This idea can be more suitably explained with the help of an example. Consider a SRAM with 1024 rows and 8 cells combined together to reduce the bit-line capacitance as shown in Fig. 8. The capacitance of the bit-line is approximately 16 times that of the capacitance of a sub bit-line. Since, bit-lines are always precharged before reading, initially both bit-line and $\overline{bit-line}$ will be at V_{dd}. While reading, depending upon whether a 1 or 0 is stored in the memory cell, sub bit-line (sub $\overline{bit-line}$) will be charged (discharged) or discharged (charged). We begin the reading operation by making the pass transistors of individual cells ON and those connecting bit-lines to sub bit-lines OFF. After charging or discharging of the sub bit-lines, the pass transistor for sub bit-line is switched off and for bit-line is switched on. This can be done by applying a small pulse to pass transistors. This pulse can be created by simple pulse generation techniques. As bit-lines are

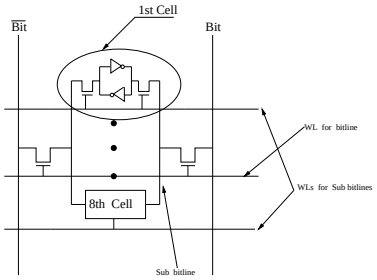

Figure 8. SRAM configuration with 8 cells combined

at V_{dd} because of precharging and either sub bit-line or sub $\overline{bit-line}$ will be at V_{dd} and other at Gnd, depending upon 1 or 0 stored in the cell, charge will be supplied by the bit-line to the sub bit-line which is at Gnd. This will lead to charge sharing between the big capacitance of the bit-line and the small capacitance of sub bit-line and the final voltage, V_{final}, on the bit-line is given by

$$V_{final} = V_{dd} \times \frac{C}{C + C/15} = V_{dd} \times \frac{15}{16}. \qquad (27)$$

Equation (27) shows that a voltage swing of $V_{dd}/16$ is obtained on the bit-line. Thus, we can obtain a controlled voltage swing on bit-line which will further reduce the power consumption.

In this approach, there is no need of an extra row and extra column and special circuitry as described in [1] to generate limited swing. This extension uses the charge sharing technique over the *divided bit-line* to develop a limited voltage swing with an extra pulse generation circuit, to reduce the power consumption. This approach reduces the power consumption by another $20 - 30\%$. Fig. 10 shows a bar-chart for the extended approach with 8 cells combined together.

6 Conclusions

In this paper, a novel approach for reducing bit-line capacitance has been presented. Both experimental and theoretical results and a theoretical explanation of the reduction in power and delay have been presented. This approach shows a considerable improvement over the existing standard technique of a single bit-line, both in terms of power and in terms of speed.

References

[1] B. S. Amrutur and M. Horowitz. Techniques To Reduce Power In Fast Wide Memories. *1994 IEEE*

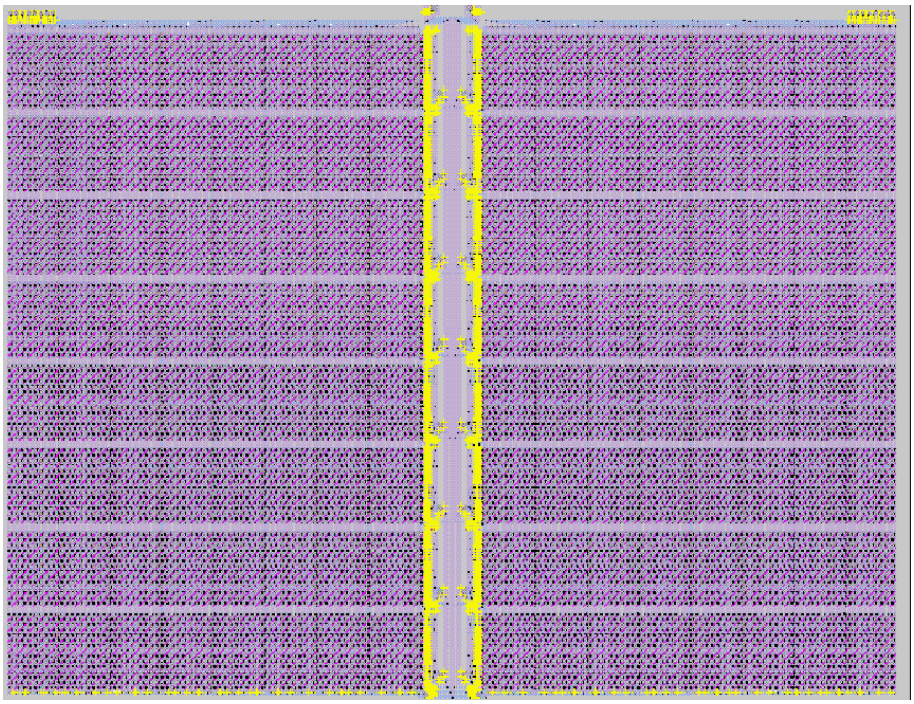

Figure 9. The complete layout of SRAM chip

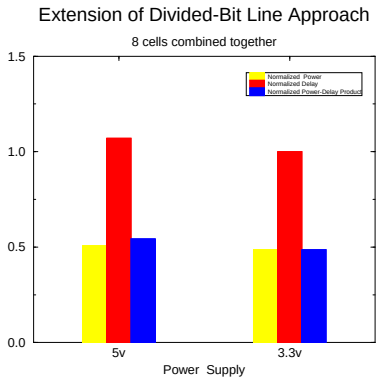

Figure 10. Experimental Results for Extension of Divided Bit-Line

Symposium on Low Power Electronics, pp.92-93, Oct. 1994.

[2] S. Dutta. VLSI Issues and architectural tradeoffs in advanced video signal processors. *Phd Thesis, Department of Electrical Enginnering, Princeton University, Nov. 96, Chapt.4, pp 100-145.*

[3] J. L. Hennessy and D. A. Patterson. *Computer Architecture, A Quantitative Approach.* Morgan Kaufmann, San Francisco, CA, 1996.

[4] K. Itoh, K. Sasaki, and Y. Nakagome. Trends in Low-Power RAM Circuit Technologies. *Proceedings of the IEEE, Vol. 83, No. 4, April 1995.*

[5] J. Rubinstein *et. al.* Signal Delay in RC Tree Networks. *IEEE Transactions on Computer-Aided Design, Vol. CAD-2, No. 3, pp. 202-211, July 1983.*

[6] K. Kimura *et. al.* Power reduction techniques in megabit DRAM's. *IEEE Journal of Solid State Circuits, vol. SC-21, pp. 381-389, June 1986.*

[7] M. Takada *et. al.* Reviews and prospects of SRAM technology. *IEICE Trans., vol. E74, no. 4, pp. 827-838, Apr. 1991.*

[8] M. Yoshimoto *et. al.* A divided word line structure in the static RAM and its application to a 64k full CMOS RAM . *IEEE Journal of Solid State Circuits, vol. SC-18, pp. 479-485, Oct. 1983.*

[9] T. Hirose *et. al.* A 20-ns 4-Mb CMOS SRAM with Hierarchical Word Decoding Architecture. *IEEE Journal of Solid-State Circuits, Vol. 25, No. 5, Oct. 1990.*

[10] N. H. E. Weste and K. Eshraghian. *Principles of CMOS VLSI Design, A Systems Perspective, 2nd Ed.* Addison Wesley, 1993.