Assessing Merged DRAM/Logic Technology

Submitted to Integration, the VLSI Journal

Yong-Bin Kim*(Contact person) Tom W. Chen**

*Department of Electrical Engineering

University of Utah

50 South Central Campus Drive

Salt Lake City, UT 84112

Tel (801) 581-4445 Fax (801) 581-5281

Email: ybk@ee.utah.edu

**Department of Electrical and Computer Engineering

Colorado State University

Fort Collins, CO 80523

Tel: 970-491-6574 Fax: 970-491-2249

Email: chen@engr.colostate.edu.

November 11, 1998

Abstract

This paper describes the impact of DRAM process on the logic circuit performance of Memory/Logic Merged Integrated Circuit and the alternative circuit design technology to offset the performance penalty. Extensive circuit and routing simulations have been performed to study the logic circuit performance degradation when the merged chip is implemented on DRAM process. Three logic processes(0.5 μm, 0.6μm, and 0.8μm) and two corresponding contemporary DRAM(64Mb and 256Mb) processes have been selected for the study knowing that the performance difference between the logic and DRAM processes can be extrapolated for the advanced processes. The simulation results show that the logic circuit performance is degraded about 22% on DRAM process including the increased interconnect delay due to less interconnect layers available in DRAM process. The silicon area is increased up to 80% depending on the number of net and components when implementing a logic circuit in a DRAM process. Simulation results show that the performance penalty can be well offset if the same circuit used in the simulation is implemented using dynamic circuit techniques.

1 Introduction

Since the actual memory bandwidth of a system is limited by off-chip interconnects, the memory access time is improved dramatically by utilizing the bandwidth available from internal arrays if a core processor and DRAM are integrated into a single chip. Typically, there are less than two dozen off-chip interconnects available on a DRAM part, and for the densest DRAM technologies the interconnects provide typically 50MB/sec of bandwidth per part even though the actual bandwidth is much higher. This is in contrast to the 1 to 2 GB/sec bandwidth needed to support the core of most modern microprocessor. The bandwidth requirement on graphics chip is even higher. In fact, the actual bandwidth inside the memory part is much higher. Only a small percentage of the actual number of memory bits read from these internal memory arrays are actually made available off-chip. Most of the bandwidth actually present inside the memory chip is discarded on the system level because of the limited I/Os, and then painfully regained through replication [1]. This limited bandwidth then causes designers of even higher performance CPU chips to spend more of their CPU silicon chip area and external glue logic on bandwidth acceleration and memory subsystem support circuits [2]. DRAM/Logic merged technology now permits very significant amount of logic to be placed on DRAM chips, meaning that the bandwidth available from internal memory arrays can be utilized directly by one or more CPUs placed directly on the chip. The second advantage is heat removal. If memory bus can be eliminated by integrating DRAM and core processor into a single chip, almost of the memory bus related power dissipation which is major portion of total power dissipation of the system can be removed.

With demand for better graphics driving frame refresh rates above 75Hz, increasing pixel resolution to 1280×1024, added color bits like 24 bits/pixel, and new graphics software, the bandwidth requirement on DRAM is increasing significantly, from several tens of Mb/sec in 1990 to over 500Mb/sec [3]. The number of DRAM bits per chip is so large that one to two chips have the capacity for most graphics applications. Therefore, how to get high bandwidth from one or two chips with limited I/O becomes an issue, driving thoughts like integrating both DRAM and logic on the same chip. This concept is being more emphasized in graphics controller chip design, because a total integration with a graphics controller and macro based DRAM frame buffers is the promising approach with respect to cost, performance, and power. DRAM macros provide bandwidth by wide data bus and on-chip speed. Independent address per macro also provides flexibilities in memory-to-screen mapping [4].

As silicon fabrication technology develops, it is a matter of time until DRAM/Logic merged chip is a

commonplace. It is likely that embedded DRAM will become as common as embedded SRAM at some point. Since the first paper was published on the DRAM/Logic merged chip design in 1990[5], several papers have been published on DRAM/Logic merged chip using DRAM process technology. A product was announced on revolutionary graphics controller chip that embedded 1 Mb frame buffer on the die alongside the controller logic using an available 16Mb DRAM process [6]. Another chip design that merged 100K custom circuits and 4.5Mb DRAM onto a single die has been announced [2]. Even at higher level integration using more advanced process technology, the potential exists to integrate a considerable amount of logic circuits and a memory, especially DRAM on one chip. For example, it is possible to integrate a microprocessor, 256Kb SRAM and 16Mb DRAM all at one chip using 256Mb DRAM process technology modified to add the necessary metal levels. Just as there are reasons the embedded DRAM will run fast, there are reasons the logic next to it will run slowly. Those include the limitations of DRAM processes.

However, any quantitative investigation has not been carried out on many open issues which deals with the slower speed of the DRAM transistors, routing area penalty due to two less metal layers of DRAM process, circuit technology to offset those device performance penalty. In this paper the logic device performance penalty on DRAM process has been investigated, penalty model has been extracted, the gate delay penalty model has been verified with realistic critical path circuit simulations of the state of the art 64-bit microprocessors, and routing area overhead has been predicted based on experiments.

2 Differences of Logic and DRAM Processes

Logic and memory processes are being developed in different directions. Logic process has been developed for speed performance and DRAM process has been developed for density and reliability. Therefore, there are many aspects of both processes that are very different.

2.1 Leakage Current of Logic Process

For transistors in DRAM, it is extremely important to reduce the leakage current of storage capacitor through pass transistor, and threshold voltage is increased by applying substrate bias to reduce the leakage current. On the other hand, the substrate is biased at 0V in logic process for low threshold voltage and high speed. Therefore, leakage current is the big obstacle in implementing DRAM using logic process, and it is worthwhile to investigate the leakage current of logic process and its impact on DRAM refresh cycle.

Leakage current components are tunneling currents through gate oxide, drain junction in the sub-micron device, and sub-threshold currents.

Gate tunneling current density can be determined making the worst-case assumption that the entire supply voltage appears across the entire gate:[7]

$$J_{ox} = J_0 E_{ox}^2 e^{-kt_{ox}} \tag{1}$$

where ($J_0 = 6.5 \text{ x } 10^{10}$ (A/cm^2)) was adjusted to match experimental data. The imaginary part of the wave vector k is given by

$$k = \frac{2k_0}{3}\frac{\phi}{V}(1 - (1 - min(1, \frac{V}{\phi}))^{1.5}) \tag{2}$$

These expressions are valid for voltages both above and below the barrier potential ϕ, which was taken to be 3.2V. The pre-exponential constant $k_0 = 1.2/\AA$ was used. Using those Equations (1) and (2), tunneling current of a transistor assuming W=8.25μm, L=0.4μm for gate oxide of 95Åand 3V power supply is 0.0018fA, which is very negligible. That specific transistor geometry was chosen because the minimum storage capacitance for DRAM is 30fF considering α particle effect.

The other leakage current component is junction leakage current which becomes more significant as doping concentration increases. In order to find out the junction leakage current at a given doping concentration, the maximum electric field across the drain junction has to be found out first. And the maximum electric field in the drain junction can be determined from the junction voltage, which in the worst case will be the supply voltage plus the built-in junction potential.

$$E_j = \sqrt{\frac{2qn(V + V_b)}{\epsilon_{si}}} \tag{3}$$

For simplicity step junction approach has been taken and it gives us the junction built-in voltage $V_b = 1.1V$. Given the maximum electric field the junction tunneling current can be determined using the following equation[7]:

$$J_j = G_0 V \frac{E_i}{E_0} e^{-\frac{E_0}{E_j}} \tag{4}$$

The constant E_0 = 2.9 x 10^7 V/cm was taken from Fair and Wivell [8], and the pre-exponential factor G_0 = 3 x 10^9A/Vcm^2 was chosen as an empirical constant. Using Equations (3) and (4) the junction tunneling current can be determined as 0.003fA for 1.73286 x $10^{17}/cm^3$ doping concentration, which is again very negligible amount.

So far we have found out that the tunneling current components of the leakage current in logic process are not considerable. Now let us investigate the sub-threshold current of the same device used for tunneling currents in logic process. There was a discontinuity problem in the classical SPICE3 model, which fails to provide continuous I-V curve. That problem has been solved in the BSIM model by considering the continuity of drain current and the first order equation in the sub-threshold voltage region. And the sub-threshold current in the BSIM model is given by:[9]

$$I_{sub} = \frac{I_{exp} \times I_{lim}}{I_{lim} + L_{lim}} \tag{5}$$

where I_{exp} is the term to express the diffusion due to bias voltage and given by:

$$I_{exp} = \beta_o (V_{tm})^2 e^{1.8} e^{\frac{V_{GS}-V_{TH}}{nV_{tm}}} (1 - e^{-\frac{VDS}{V_{tm}}}) \tag{6}$$

where V_{TH} is BSIM model threshold voltage and given by:

$$V_{TH} = V_{FB} + PHI + K_1(PHI - V_{BS})^{0.5} - K_2(PHI - V_{BS}) \tag{7}$$

where V_{FB} is flat band voltage of the device, PHI is two times the Fermi potential, K_1 represents body effect, and K_2 explains the non-uniform doping profile in the channel. β_0 is intrinsic transconductance and given by:

$$\beta_o = \mu_o C_{ox}(W/L) \tag{8}$$

where μ_o is gate field mobility reduction factor.

An empirical expression for I_{lim} can be derived such that sub-threshold is saturated as V_{GS} increases

toward strong inversion area:

$$I_{lim} = \frac{\beta 3(V_{tm})^2}{2} \tag{9}$$

where V_{tm} is thermal voltage which equals to 0.026V at room temperature and n is a sub-threshold coefficient and is given by:

$$n = n_0 + nb_0 V_{BS} + nd_0 V_{DS} \tag{10}$$

where nb_0 and nd_0 are the proportional constants which represent the dependency of n on V_{BS} and V_{DS}, respectively.

The size of pass transistor in DRAM is usually minimum and assumed to be 0.5μm/0.5μm for 0.5μm logic process. Using 0.5μm logic process parameters of t_{ox} =95Å, μ = 326.682cm^2/(V-sec), W/L=0.5/0.5, V_{GS}=0.1V, V_{TH}=0.6V, and V_{DS} = 3.0V, I_{exp} and I_{lim} can be obtained to be 2.53 x 10^{-11}A and 1.08 x 10^{-5}A, respectively. Consequently, the sub-threshold current for the given device is 2.53 x 10^{-11}A. It turns out that sub-threshold leakage current is dominant over tunneling leakage current, and the total amount of leakage current in the logic process requires 3.19ms refresh cycle time because 10% leakage of the total stored charge is regarded as a failure. 3.19ms refresh cycle is faster than commercially available 64Mb DRAM refresh cycle by factor of twenty. This faster refresh cycle will increase power consumption and heat dissipation quite a bit because power consumption is directly proportional to operating frequency. If the word line low voltage is worse than the assumption(0.1V), the refreshment rate has to be even higher. Typically DRAM process development is one generation ahead of logic process development in terms of minimum feature size. 0.5μm logic process and 64Mb DRAM process are contemporary processes. For the advanced contemporary technologies such as 0.25μm logic process and 256Mb DRAM process, the refresh rate differences stay about same, because leakage current in the logic process decreases and refresh rate of 256Mb DRAM increases to refresh more memory cells.

2.2 Threshold Voltage

It is extremely important to keep the leakage current as low as possible in DRAM process, and the transistors are designed for low leakage current. In order to reduce leakage current high threshold voltages

are required, and that, in turn, slow switching speed down. A typical way to increase threshold voltage in DRAM is to apply substrate bias. The threshold voltage shift due to substrate bias, ΔV_t, is proportional to the voltage between source and substrate and γ which is given by

$$\gamma = (1/C_{ox}) \times \sqrt{2q\epsilon_{si}N_A} \tag{11}$$

Furthermore, the substrate doping concentration(N_A) of DRAM process is higher than that of logic process for the device scaling purpose. The threshold voltage of DRAM process, therefore, is increased more than 40% comparing to that of logic process. This threshold voltage shift causes performance degradation of logic circuits if they are implemented on DRAM process. However, this problem may be solved by doing different implant levels in logic transistors than memory array or biasing different wells at different voltages. Consequently, the complexity of DRAM process will be further increased, which will have a negative impact on yield and cost. Performance degradation of logic gates due to higher threshold voltage in DRAM process can be mitigated by circuit techniques, which will be discussed in later section.

2.3 On-chip Interconnect

Another problem in using DRAM process for logic component than slower switching might be lack of metal layers. Even the state-of-the-art DRAM process has only four metal layers, whereas the state-of-the-art logic process has seven metal layers. The number of metal layer difference between contemporary DRAM and logic process is two or three. If chip is routed with less metal layers, the average metal line will be longer and the longer average metal line will carry higher impedance, stressing drivers that were, after all, designed for more metal layers environment. In order to minimize the routing penalty of using DRAM process extra efforts have to be taken for a cell library development that meets synthesis needs. Despite that multi-metal layer DRAM process that provides more than four metal layer is being developed to reduce the average metal length and to lower the resistance of the metal line, it is extremely difficult to develop DRAM process that has more than four metal layers because of DRAM's rough and hilly silicon surface unless costly trench DRAM cell is used. Therefore, due to stringent yield and reliability requirement in DRAM process, the number of metal interconnect layers in DRAM process is not likely to match that in standard logic process in the near future. That leaves the performance penalty due to on-chip interconnect in merged DRAM/Logic chips as an issue.

2.4 Other Issues Related to Logic Process

In the current DRAM process, the DRAM cell leakage current should be less than 1fA/μm^2 in order to pass reliability specification. This is one of the major reasons that DRAM can not be in implemented using logic process.

In order to increase DRAM density, capacitance per unit area for storage capacitor has to be increased. That increase requires three dimensional capacitor structure instead of planar technology, and dielectric layer thickness reduction causes lower breakdown voltage. DRAM processes uses a special dielectric structure(Oxide/Nitride/Oxide) to increase the breakdown voltage of the storage capacitor, but this structure is not available on logic process. The last problem of using logic process for DRAM is the capacitance of bit lines. In order to secure enough sensing margin, the ratio of bit-line capacitance to storage capacitance should not be more than 15 to 1. As die size increases by using logic process the routing capacitance will be increased and it will be another challenging issue to keep that ratio.

Considering those issues described above, it would be reasonable to conclude that DRAM/Logic merged integrated circuit would be built on a DRAM process.

3 Performance Impact of Merged Logic/DRAM on Static Logic Gates

As discussed in the previous section, there are many reasons that DRAM cannot be built using logic process. Even in DRAM process problem is not free in implementing logic components on the same die as DRAM. The major problem of using DRAM process to build logic devices is slower speed of logic circuits because of substrate bias. Despite the fact that DRAM transistors are slower than logic transistors, high performance systems can still be implemented using DRAM transistors. A DRAM/Logic merged chip that runs on chip RAMDAC at 100MHz was published in 1995 [6].

Since logic and memory processes are on diverging paths, there are so many differences between DRAM and logic process parameters. It is very difficult to study device performance differences analytically using those device parameters. Therefore, three logic processes and two corresponding contemporary DRAM processes have been used for comparison, and very extensive circuit simulations have been performed using the device parameters to find out gate performance penalty of using DRAM process for logic components. The three logic processes are 0.5μm, 0.6μm, and 0.8μ CMOS processes. The two DRAM processes are 0.5μm 64Mb and 0.4μm 256Mb DRAM processes. Even though minimum channel lengths of those processes

are different, the processes are contemporary processes, and the performance gap between the logic and DRAM processes can be extrapolated for advanced process technologies. The process characteristics are shown in Table I. The simulation was done for the worst case process corner at room temperature with fanout of four which is average fanout in real microprocessor designs. The device geometry of those gates were chosen so that the effective β_p/β_n ratio becomes two and effective driving strength becomes comparable. For example, the device sizes of inverter is that $(W/L)_p = 20/0.6$ and $(W/L)_n = 10/0.6$. The device sizes of two input nand gate is that $(W/L)_p = 20/0.6$ and $(W/L)_n = 20/0.6$. Table II shows the detailed comparison of the average logic gate delay of five different gates for both DRAM and logic process. In order to find out the minimum average logic performance penalty of using DRAM process for logic circuit implementation, SPICE simulations have been performed changing the device sizes of five different primitive gates. Fig.1 shows the average penalty as a function of (MD:ML). MD:ML is the ratio of $(W/L)_{DRAM}$ to $(W/L)_{LOGIC}$. As shown in Fig.1 the average logic gate performance degradation is about 21% as minimum when (W/L) of the transistors on DRAM process is increased 29% comparing to the (W/L) of the transistors on logic process for the same gate. The dip in the curve of Fig.1 represents the optimum gate size on DRAM process. If the device size is smaller than the optimum, the driving strength of the gate is not enough to drive the average fanout(four). On the other hand, if the device size is larger than the optimum, the benefit of increased driving strength is undermined by the increased gate loading. The 29% device geometry increase will not become a dominant factor for chip area increase, because chip area increase will be dominated by routing capability of DRAM process which has less level metal layers. This issue will be addressed in Section 4 in more detail.

Fig.2 shows the average driving capability of the DRAM and logic transistors with different fanout. Each fanout gates have the same sizes of device and the average gate delay was taken out of the five gates used for the Table II simulation results. DRAM process device tends to have more capacitance loading because the devices are optimized not for performance but for device density and yield. The input capacitance C_{IN} of a MOS transistor consists of the following components as shown in Fig.3[10].

1. C_{OS}, C_{OD} – the source and the drain overlap capacitances resulting from the overlap of the gate on the source and drain diffusions.

2. C_{GS}, C_{GD} – the gate-to-channel capacitances lumped at the source and drain regions of the channel respectively.

3. C_{GB} – the gate-substrate capacitance.

4. C_{OX} – Gate Oxide capacitance

$f_B(v)$, $f_D(v)$, $f_S(v)$, $F_O(v)$ are voltage dependent functions. Combining those individual terms, C_{IN} is given by

$$C_{IN} = C_{GB} \times L \times f_B(v) + C_{GD} \times W + C_{GS} \times W + C_{OX} \times W \times L \times f_O(v) + C_{OX} W(l_s + l_d) \qquad (12)$$

where l_s and l_d are the overlap length. In DRAM process, C_{GB}, C_{GD}, C_{GS}, C_{OX}, C_{OS}, C_{OD} are larger than those of contemporary logic process, because DRAM process is not optimized for performance but yield. Consequently, those factors increase the input gate capacitance up to 60% for the same fanout gates comparing to logic devices. And also as mentioned earlier device geometry is increased 29% from logic process transistor to achieve minimum gate delay penalty. The 21.3% gate delay performance degradation in Fig.2 matches with the data in Table II.

4 Performance Penalty Model on Real CMOS Circuit

In real circuit design, especially very complex systems design like microprocessors, a lot of custom circuits which have a variety of driving strengths of gates and latches are used. In most of critical path circuits, device sizes are optimized and ratioed so that a particular edge is faster than the other. If the critical path circuit has to drive a long distance, the circuit may have to be designed so that it is less sensitive to any noise and ground bouncing. Therefore, most of the critical path circuits are non-standard type of gates. In order to investigate the realistic circuit performance penalty on DRAM process, these non-standard type circuit's performance has to be studied to find out if the circuit performance penalty model obtained from standard gates can be applied to the non-standard type circuits. The realistic standard gate circuits have to be also tested to verify the circuit performance penalty model , because the circuit performance penalty model was extracted from individual gates. The non-standard type circuits have been selected from both data path and control logic critical path circuits, and the standard type circuits have been selected from synthesized blocks.

More than 50 circuits have been reviewed out of a 64 bit microprocessor design to test the circuit performance penalty model, and total twenty one real critical path circuits have been selected to verify

the gate delay penalty model from a state of the art 64-bit microprocessor. The PMOS transistors had to be resized and optimized to minimize the logic performance penalty on DRAM process, since the PMOS performance on DRAM is much worse than PMOS on logic process comparing to NMOS. The circuit1 in Table III is a 32 bit adder which is used for an integer execution unit. The rest of the circuits are custom logic circuits used for both data path and control logic circuits. The last ten circuits(circuit12 $\sim$ circuit21) are selected from synthesized circuits. The simulation results and gate delay penalty percentages are shown in Table III. As shown in Table III the logic gate delay penalty is well matched with the model, and it can be concluded that the penalty model can be applied to both standard gate circuits and non-standard type circuits.

5 Routing Area Penalty

Memory processes have many poly interconnects but less metal layers than logic process. Logic processes have only one or two poly layers, but it usually has more layers of metal interconnect. With less metal layers, a penalty has to be paid in routability, size and performance. Longer average metal line will add more capacitive and resistive load to the driver.

In order to investigate area penalty of CMOS circuits implemented in DRAM process due to less routability and wiring capacitance increase per net, nine circuits have been selected out of a state-of-the-art 64-bit microprocessor design, and routing simulations have been performed with three metal layers and five metal layers using the same automatic routing and placement tool because metal layers of DRAM process are at least two layer less than contemporary logic process. The average wiring length per net can be found by dividing the total metal interconnect length of all the metal layers by the total number of nets for both three and five metal layer cases. The wiring capacitance per net of three level metal layer routing can be compared to that of five metal layer routing. Table IV shows the area and wiring capacitance per net of these nine test circuits for both three and five metal layer processes.

The results show that the wiring capacitance per net is increased by 20.85% if the same circuit is routed using three level metal layer process. But the absolute capacitance increase per net is only 7.6ff per net. When the wiring capacitance increase is reflected on the total gate delay performance penalty, it is additional 1.5%. But this data is valid for intra-block routing. Considering global clock and power distribution, the performance penalty due to routability will be much worse than that at intra-block routing

level chip level depending on chip size and block sizes. Therefore, block partition is critical in order to minimize the routing penalty, and extra efforts should be taken for cell library and macro cell development so that they meet synthesis requirements by leaving more routing channels inside the cell. It is obvious that the routing channel area will be increased as total number of nets and components in a circuit increase. Using the routing simulation results in this research, the area penalty model is derived by curve-fitting the simulation results. Fig.4 shows the trend of area penalty as a function of total component used and total number of net.

As shown in Fig.4, the area penalty is more sensitive to the number of net than to the number of components. The fitting equation representing the area penalty trend is given by

$$Z = -82.71 + 14.97 \times ln(X) + 0.015 \times Y \tag{13}$$

where Z is the area penalty and X and Y are number of component and number of net, respectively.

6 Alternative Circuit Technology to Offset the Performance Penalty

It is important for circuit designers to select appropriate logic family to implement logic function and to meet the design specification. Currently low power logic design uses static, dynamic and pass transistor logic, with current mode logic and special logic families playing a minimal role in the future low power designs.

Pass transistor logic offers reduced transistor count by eliminating the PMOS transistors and it provides improved speed. However, complementary pass networks are desirable to achieve good logic levels [11]. The number of drain-source connections at the output is effectively doubled, hence doubling the output capacitance and making CMOS pass transistor networks inherently slow. As a result, this pass logic offers only moderate area and speed improvements over CMOS static gate logic [12]. Furthermore, pass transistor logic is difficult to implement and NMOS pass gate logic may not fully turn off PMOS transistor in buffer stages leading to the static power dissipation problem.

Special logic circuits include adiabetic circuit, self timed circuits, and current mode logic. Current mode logic is generally not included in low power circuits because of its inherent static current. And the primary disadvantage of adiabetic logic is the requirement of the ramp clock. The power supply for this circuit requires a power resonant circuit to provide the ramp clock signal.

The primary disadvantage of static CMOS is the requirement of complementary P and N MOS device networks that consumes area and increases gate loading capacitance by a factor of three comparing to an equivalent dynamic logic gate. On the other hand dynamic logic gates use only one of the two networks which form a static gate, taking into account the simplicity of interconnection, for complex blocks of gate with several inputs a saving of more than 50% in the layout area should be expected. Of course, gate loading capacitance is a lot smaller than static gates. The CMOS dynamic gates offer much faster gate delay than CMOS static gates not only because it has less gate loading but also its gate threshold is actual device threshold(Static gate's threshold is generally (VDD-VSS)/2). Of course, there are several circuit issues like coupling noise and charge sharing problems, which can be taken care of.

The dynamic gate performance benefit was further examined using the same methodology for static gate performance penalty in Section 4. Table V. shows the comparison of static and dynamic individual gates' when they are implemented using DRAM process. The data has been extracted from the comparable driving strength gate with same fanout(three). The worse edge delay has been picked among rising and falling edge edge for static gates delay, and the falling edge has been picked for dynamic gates because dynamic gates usually have enough precharging time. As shown in Table V, NOR gates are benefited more than nand and inverter gates by using dynamic circuits because PMOS transistor sizes in static NOR gates are connected in series and gives slower rising time. The performance gain is even more outstanding for complex gates.

The same critical path circuits used for the logic gates performance penalty analysis in Section 4 have been designed using dynamic gates and simulated with DRAM process parameters to test the performance gain by using dynamic gates for the logic implemented using DRAM process. The results are shown in Table VI. As shown in Table VI, in most of the cases the penalty of using DRAM process for logic circuits is overcome by using dynamic gates. Furthermore, since dynamic gates are positive logic gates(AND, OR, etc) the number of logic level could be reduced by taking advantage of the logic polarity. If the circuit is composed of a lot of complex gates and XOR gates, the performance gain will be more outstanding.

7 Conclusion

It has been found that DRAM/Logic merged chip is to be implemented on DRAM process instead of logic process based on leakage and sub-threshold current, DRAM refresh rate, threshold voltage, and

reliability issue. Even in DRAM process problem is not free in implementing logic components on the same die as DRAM. The major problem of implementing logic circuits on DRAM process is slower speed of logic circuits because of substrate bias.

This paper presented the circuit performance and routing penalty of using DRAM process for logic components of DRAM/Logic merged integrated circuits, and alternative circuit technology to offset the penalty. The simulation results show that there is 22% logic circuit performance degradation including routing capacitance penalty when they are implemented using DRAM process. That amount of penalty can be offset by using dynamic circuit technologies. Even though this paper has demonstrated the penalty of using DRAM process in a quantitative manner and will be a good reference for DRAM embedded circuit design in the future, a lot of device and process issues have to be investigated to develop and improve the performance of DRAM/Logic merged chips. As some enhancements of DRAM process are being done to help the performance of logic circuits, more metal layers will be available in 256Mb to 1Gb DRAM processes. However, the routing penalty on DRAM process are still likely to exist when the contemporary logic and DRAM processes are compared, because the advanced logic process provide better performance and more metal layers. Of course, there are some alternatives to minimize the penalties at higher costs. The first is to provide different substrates to the logic and DRAM parts by using triple well process. The second is to use two type of gate oxide thickness. The last alternative is to use trench DRAM cell instead of stack cell to put more metal layers by providing easy planarization. All of the alternatives can be accomplished at higher costs. Therefore, the performance and cost analysis is yet to be analyzed to make DRAM/Logic merged chip more promising.

References

[1] Peter M. Kogge, "EXECUBE-A New Architecture for Scaleable MPPs", *Proceedings of International Conference on Parallel Processing*, 1994, 77-84.

[2] Peter M. Kogge, Toshio Sunaga, Hisatada Miyataka, Koji Kitamura, "Combined DRAM and Logic Chip for Massively Parallel Systems", *Proceedings of Sixteenth Conference on Advanced Research in VLSI*, March 1995, 4-16.

[3] Akira Kanuma, "The Best DRAM Approach for Graphics Application", *ISSCC95/Evening Discussion Session/WE2*, February 1995, 96-97.

[4] Toshio Sunaga, "The Best DRAM Approach for Graphics Application", *ISSCC95/Evening Discussion Session/WE2*, February 1995, 96-97.

[5] Mark D. Weir, Jay F. Kita, Thomas J. Cockerill, Paul F. Hynek, Thomas M. Lepsic, Trang K. Ta. Bobby K. Wong, Steven R. Woodham, Richard R. Young, "A 250K-Circuit ASIC Family Using a DRAM Technology", *Proceedings of IEEE 1990 CICC*, 1990, 4.6.1-4.6.5.

[6] EE TIMES, "Embedded-DRAM arrival stirs debate" ,*Electronic Engineering Times*, April 10, 1995, 43.

[7] Carver A. Mead, "Scaling of MOS Technology to Submicrometer Feature Sizes", *Journal of VLSI Signal Processing* vol 8, July 1994, 9-25.

[8] R.B. Fair and H.W. Wivell, "Zener and avalanche breakdown in as-implanted and low-voltage Si n-p junctions", *IEEE Trans. Electron Dev.* vol. ED-23, 1527-1534.

[9] Meta-software, "HSPICE User's manual",*Meta-software* vol.2, 1995, 7-71∼ 7-80.

[10] Mohamad I. Elmasry, "Digital MOS Integrated Circuits: A Tutorial", *Digital MOS Integrated Circuits 2.* , *IEEE Press* 3-33.

[11] Nei H.E. Weste, Kamran Eshraghian, "Principles of CMOS VLSI Design" *Addison-Wesley Publishing Company*, 1993, 307.

[12] John H. Pasternak, Alex S. Shubat, "CMOS Differential Pass-Transistor Logic Design", *Journal of Solid State Circuits* vol. sc-22 NO.2, April 1987, 216-222.

Table I. Characteristics of the DRAM and logic processes.

–	*DRAM Process*		*Logic Process*		
Parameter	64Mb	256Mb	0.5μm	0.6μm	0.8μm
Channel length	0.5μm	0.4μm	0.5μm	0.6μm	0.8μm
Gate Oxide thickness	100$\AA$	70$\AA$	95$\AA$	93$\AA$	125$\AA$
N-tr flatband vol.(V)	-.927792	-1.00806	-.3012973	-.4577923	-1 .00171
P-tr flatband vol.(V)	-.318455	-.674946	-.4225464	-.3786467	-.3 55492
N-tr 2 $\times$ (Fermi vol.)(V)	.820873	.859140	.500237	.654587	.850000
P-tr 2 $\times$ (Fermi vol.)(V)	.740000	.834164	.880000	.827692	.800000
N-tr mobility(cm^2/(V-sec))	405.245	422.953	326.682	385.900	499.832
P-tr mobility(cm^2/(V-sec))	159.703	98.8012	131.000	113.400	156.953

Table II. Logic gate performance comparison.

–	*Logic Process(nS)*	*DRAM Process(nS)*	*Performance Penalty(%)*
Inverter	0.186	0.231	24.2
2-Nand	0.279	0.344	23.3
3-Nand	0.443	0.514	16.0
2-Nor	0.389	0.484	24.4
3-Nor	0.670	0.803	19.8
AVERAGE	–	–	21.5

Table III. Performance differences of critical paths for DRAM and logic process.

–	*Logic Depth*	*No.of Gates*	*Logic Process(nS)*	*DRAM Process(nS)*	*Perfo rmance Penalty(%)*
circuit1	12	516	3.02	3.68	21.85
circuit2	14	308	3.43	4.05	18.07
circuit3	19	308	4.45	5.54	24.50
circuit4	11	308	3.35	4.02	20.00
circuit5	15	308	5.07	6.20	22.28
circuit6	17	186	5.76	7.45	29.34
circuit7	7	186	2.29	2.79	21.83
circuit8	8	186	2.65	3.28	23.77
circuit9	9	186	3.63	4.60	26.72
circuit10	16	141	5.36	6.53	21.82
circuit11	5	141	1.69	2.02	19.52
circuit12	9	101	1.59	2.00	25.78
circuit13	6	101	1.31	1.57	19.85
circuit14	7	101	1.61	1.92	19.25
circuit15	7	101	1.46	1.79	22.60
circuit16	6	101	1.32	1.65	25.00
circuit17	6	101	1.43	1.78	24.48
circuit18	9	240	2.49	3.13	25.70
circuit19	7	240	1.56	1.94	24.36
cricuit20	8	240	1.88	2.30	22.34
circuit21	5	16	1.00	1.22	22.00
AVERAGE	–	–	–	–	22.91

Table IV. Routing and wiring simulation results

–	*Block Size*		*Area*			*Cap. per net*		
Block	♯ of Net	♯ of Comp.	4 layer(sq.μm)	2 layer(sq.μm)	penalty(%)	4 lay er(ff)	2 layer(ff)	penalty(%)
block1	587	409	219,340.8	238,280	8.63	27	31	14.81
block2	2334	2566	1,033,881.6	1,736,582.4	67.90	50	64	28.00
block3	962	962	241,990.4	346,032	42.99	15	23	53.33
block4	580	476	323,251.2	364,540.8	12.77	40	59	47.50
block5	651	651	341,990.4	441,686.4	29.15	41	55	34.14
block6	468	468	215,424	225,836	4.83	48	49	2.08
block7	1321	1400	1,108,224	1,576,872	42.28	53	63	18.86
block8	241	231	147,993.6	152,716.8	3.19	21	14	-33.33
block9	3111	2519	811,008	1,455,648	79.48	27	33	22.22
Average	–	–	–	–	–	–	–	20.85

Table V. Static and dynamic gate performance comparison on DRAM Process.

–	*Static Gates(nS)*	*Dynamic Gates(nS)*	*Performance Gain(%)*
2-Nand	0.325	0.300	7.69
3-Nand	0.395	0.380	3.79
2-Nor	0.265	0.260	1.89
3-Nor	0.400	0.285	28.75
2-Xor	0.520	0.350	32.69
AVERAGE	–	–	14.96

Table VI. Logic gate performance comparison.

–	*Static Circuit(nS)*	*Dynamic Circuit(nS)*	*Performance Gain(%)*
circuit13	1.31	1.31	0.00
circuit14	1.61	1.46	9.32
circuit15	1.46	1.36	6.85
circuit17	1.43	1.43	0.00
circuit18	2.49	2.22	10.84
circuit19	1.56	1.48	5.13
circuit20	1.88	1.24	34.04
circuit21	1.00	0.98	2.00
AVERAGE	–	–	8.52

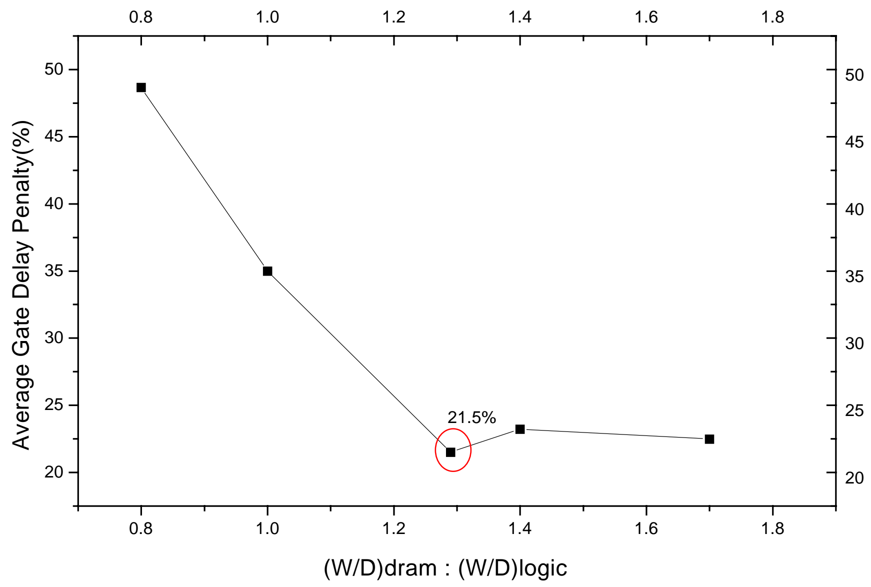

Fig.1. Average Gate Penalty VS. (MD:ML).

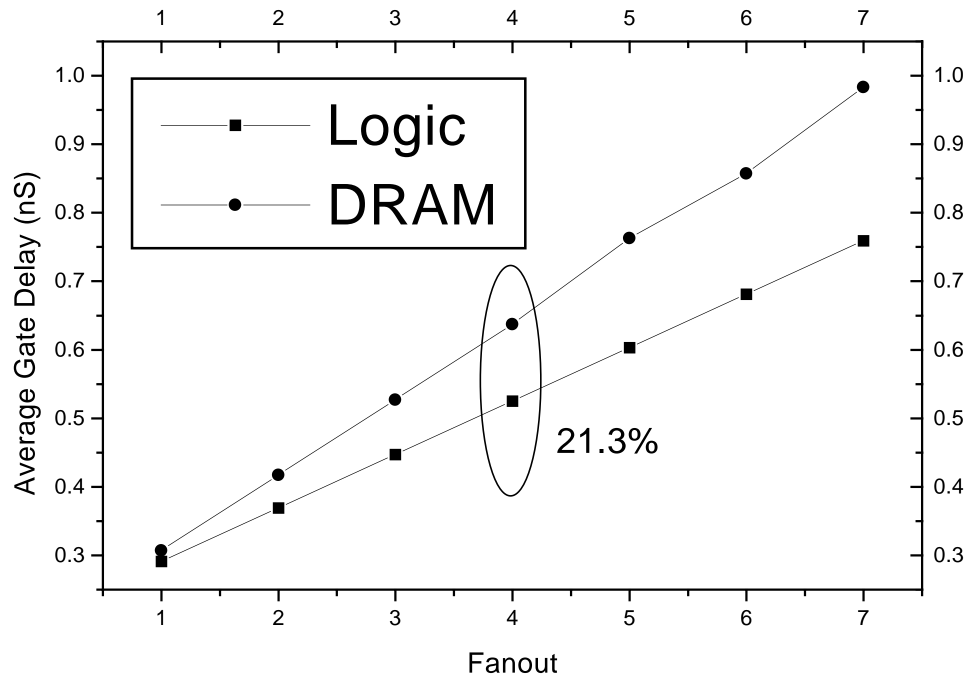

Fig.2. Simulated driving capabilities of DRAM and logic process gates with fixed load.

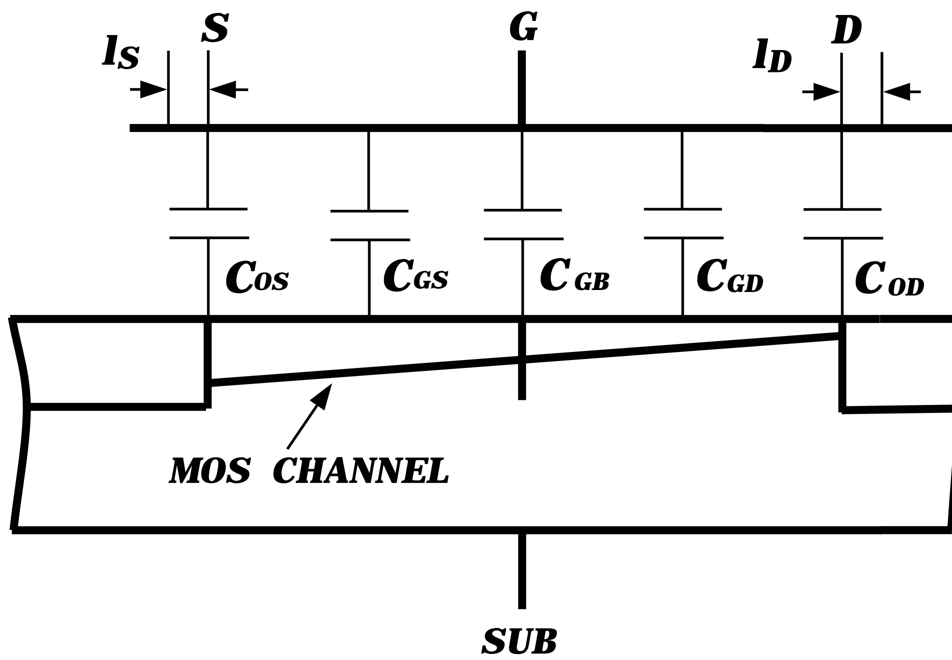

Fig.3. The different components of C_{IN}.

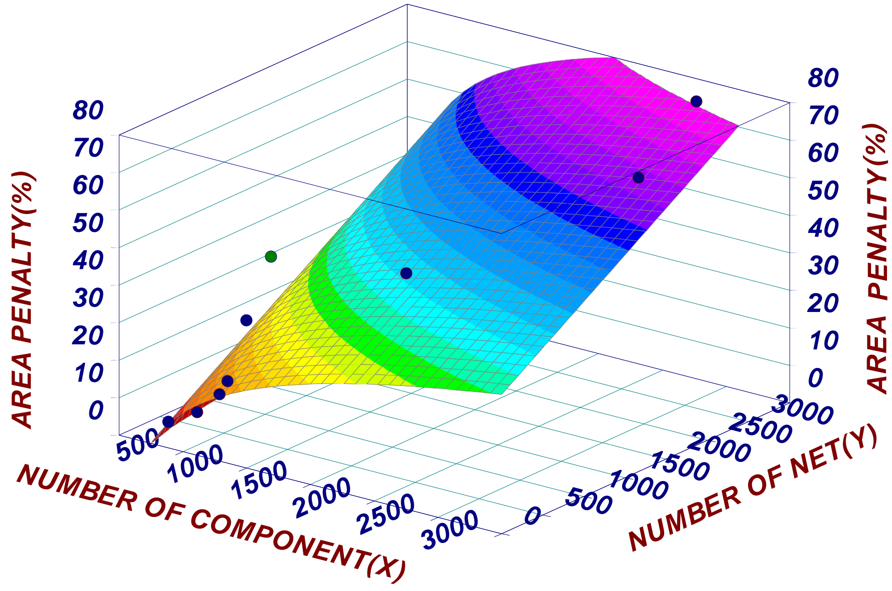

Fig.4. Routing area overhead.

Index terms

DRAM/Logic Merged Technology, Embedded DRAM, DRAM, memory, VLSI

LIST OF TABLES

Table I: Characteristics of the DRAM and logic processes . . . .

Table II: Logic gate performance comparison . . . .

Table III: Performance differences of critical paths for DRAM and logic process . . . .

Table IV: Routing and wiring simulation results . . . .

Table V: Static and dynamic gate performance comparison on DRAM process . . . .

Table VI: Performance difference of critical paths for DRAM and logic processes . . . .

LIST OF FIGURES

Fig.1: Average gate penalty vs. (MD:ML) . . . .

Fig.2: Simulated driving capabilities of DRAM and logic process gates with fixed load . . . .

Fig.3: The different components of C_{IN} . . . .

Fig.4: Routing area overhead . . . .

Yong-Bin Kim was born in Seoul, South Korea in 1960. He received the B.S. degree in electrical engineering from Sogang University in Seoul, South Korea,in 1982, and the M.S. and PH.D degree all in electrical engineering from New Jersey Institute of Technology in 1982 and Colorado State University 1996, respectively.

From 1982 to 1987 Dr. Kim was with Electronics and Telecommunications Research Instituture in South Korea as a Member of technical Staff. From 1990 to 1993 he was with Intel Corp. as a Senior Design Engineer, and involved in micro-controller chip design and Intel P6(Pentium Pro) microprocessor chip design. From 1993 to 1996 he was with Hewlett Packard Co., Fort Collins, Colorado as a Member of Technical Staff, and involved in HP PA-8000 RISC microprocessor chip design. From 1996 to 1998 he was with Sun Microsystems, Palo Alto, california as an individual contributor, and involved in 1.5GHz Ultra Sparc5 CPU chip design. Since 1998 August, he has been an assisstant professor in the Dept. of Electrical Engineering of University of Utah. His research focuses on circuit design, especially VLSI design.

Tom Chen received the B.Sc. degree in electronic engineering from Shanghai Jiao-Tong University, China, and the Ph.D. degree in electrical engineering from the University of Edinburgh, Scotland, U.K., in 1982 and 1987, respectively.

From 1987 to 1989 Dr. Chen was with Philips Semiconductors in Hamburg, Germany. From 1989 to 1990 he was with the department of Electrical and Computer Engineering, New Jersey Institute of Technology. He is currently an associate professor at the Department of Electrical Engineering at Colorado State University. Dr. Chen's research interests are in the areas of novel architectures in computation and signal processing, VLSI design and testing, and real-time computer vision systems.