Titre: Title:	Performance modeling and analysis of CMOS, BiCMOS mega bit static ramdom access memories
Auteur: Author:	Venkatapathi Naidu Rayapati
Date:	1995
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Rayapati, V. N. (1995). Performance modeling and analysis of CMOS, BiCMOS mega bit static ramdom access memories [Thèse de doctorat, École Polytechnique de Montréal]. PolyPublie. https://publications.polymtl.ca/31798/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/31798/
Directeurs de recherche: Advisors:	Bozena Kaminska
Programme: Program:	Non spécifié

UNIVERSITÉ DE MONTRÉAL

PERFORMANCE MODELING AND ANALYSIS OF CMOS, BiCMOS MEGA BIT STATIC RANDOM ACCESS MEMORIES

Venkatapathi Naidu RAYAPATI

DÉPARTEMENT DE GÉNIE ÉLECTRIQUE ET DE GÉNIE INFORMATIQUE

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

THÈSE PRÉSENTÉE EN VUE DE L'OBTENTION

DU DIPLÔME DE PHILOSOPHIAE DOCTOR (Ph.D.)

GÉNIE ÉLECTRIQUE

December 1995

ISBN 0-612-11472-4

UNIVERSITÉ DE MONTRÉAL

ÉCOLE POLYTECHNIQUE DE MONTRÉAL

Cette thèse intitulée:

PERFORMANCE MODELING AND ANALYSIS OF CMOS, BiCMOS MEGA BIT STATIC RANDOM ACCESS MEMORIES

présentée par: RAYAPATI Venkatapathi Naidu

en vue de l'obtention du diplôme de: Philosophiae Doctor (Ph.D.)
a été dûment acceptée par le jury d'examen constitué de:

M. GHANNOUCHI Fadhel , Ph.D. président

Mme KAMINSKA Bozena , Ph.D. directeur de recherche

M. SAWAN Mohamed , Ph.D. membre

M. DE PAYREBRUNE Mark , Ph.D. membre

Dedicated to my beloved mother
(Kamalamma Rayapati)

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my professor Bozena Kaminska for her dedication, continuous encouragement, advise and her valuable assistance in the completion of this research work.

I would like to thank late professor Dinkar Mukhedkar for his encouragement and support.

I wish to thank Mark de Payrebrune for his continuous encouragement and suggestions.

I am thankful to my brother Subramanyam Rayapati for his encouragement, suggestions and support during the preparation of this thesis.

SOMMAIRE

Une évaluation de la performance est nécessaire à chaque étape du cycle de vie d'un circuit intégré SRAM offrant une capacité de l'ordre du mégabit, soit durant sa conception, sa fabrication, son acquisition, son utilisation, etc. Cette évaluation de la performance s'impose lorsqu'un concepteur de circuits intégrés SRAM veut comparer plusieurs nouveaux types de circuits afin de découvrir le meilleur. Même s'il n'en trouve pas, l'évaluation de la performance des circuits SRAM classiques aide à déterminer son niveau de performance, ce qui indique s'il convient de modifier les futurs circuits en raison des progrès technologiques. Ce que veulent les utilisateurs de mémoires SRAM, c'est obtenir le plus haut niveau de performance et de fiabilité à un prix donné.

Afin de concrétiser les attentes à l'égard de la performance des mémoires SRAM de l'ordre du mégabit, il faut déterminer de nombreux paramètres : la vitesse, la dissipation d'énergie, le temps de propagation, le bruit, la superficie du circuit, la densité d'intégration, la fiabilité, la détection des erreurs, la localisation des erreurs, la reconfiguration et le rendement de fabrication. À cette fin, cette recherche vise plusieurs objectifs : proposer des modèles de performance des circuits SRAM, analyser leurs paramètres de performance, valider chaque modèle de performance par des études expérimentales et, enfin, comparer l'amélioration de la performance par rapport aux conceptions classiques.

Un modèle de capacité d'interconnexion est proposé; il est basé sur la capacité des couches d'oxyde des cellules du SRAM, leur capacité marginale, leur capacité de couplage et leur capacité parasite. Ce concept pratique est ensuite traduit en termes mathématiques afin de calculer la capacité d'interconnexion et d'évaluer l'effet de la performance sur les paramètres des dispositifs SRAM de l'ordre du mégabit. La performance du dispositif MT5C1008 est simulée sur le simulateur haute vitesse HP 9000. La dépendance du temps d'accès par rapport à la température est mesurée par essai au thermocouple. On a analysé par microscopie électronique à balayage et par microscopie optique les défaillances du SRAM afin d'en identifier les mécanismes. Les résultats de la théorie, des simulations et des expériences sont en accord avec le modèle de capacité d'interconnexion.

En appliquant l'interconnexion à triple niveau à un circuit SRAM de 1 mégabit, le temps d'accès aux adresses est réduit de 35 ns à 25 ns. Lorsque la température augmente à 100°C, le temps maximal d'accès aux adresses est réduit à 30,8 ns, avec une dissipation de puissance de 1 W. La longueur nécessaire des fils électriques et la grosseur du circuit intégré ont été réduits de 69 % et 58 %, respectivement, par rapport aux interconnexions double métal. On a obtenu, pour la plaquette, un rendement de fabrication de 10 % supérieur grâce à l'implantation du modèle à triple niveau d'interconnexion dans la conception du circuit SRAM. Les résultats expérimentaux confirment que ce modèle est très utile pour la

conception des futurs circuits intégrés SRAM et DRAM.

On a développé une expression analytique pour le temps de propagation sur les circuits SRAM CMOS. On propose un modèle du temps de propagation, basé sur le couplage capacitif des cellules SRAM et sur les interconnexions en charge. La nature non linéaire de la porte d'attaque a été prise en considération, car on a utilisé un modèle de transistor du premier ordre pour les cellules SRAM. Le temps de propagation théorique pour le SRAM a été calculé en utilisant le programme MATLAB. On a simulé le temps de propagation d'un circuit SRAM de 16 Mb au moyen du programme HSPICE roulant sur un poste de travail HP A1097. On a utilisé un analyseur logique HP 8002 afin de mesurer le temps de propagation sur un circuit SRAM. L'expression analytique du temps de propagation donne un taux d'erreur maximal absolu inférieur à 4,8 % par rapport aux données mesurées. La valeur théorique des temps de propagation pour les circuits SRAM donne un taux d'erreur inférieur à 2 % par rapport aux résultats simulés sur HSPICE. L'expression analytique du temps de propagation est validée par les résultats expérimentaux.

Des techniques de conception des circuits haute vitesse SRAM BiCMOS sont présentement étudiées afin de réduire le temps d'interconnexion, d'accroître la vitesse et de réduire le bruit et la consommation d'énergie. Le système de simulation HP 9000 haute vitesse est utilisé afin de simuler les paramètres de performance des circuits intégrés SRAM BiCMOS. La vitesse des circuits SRAM

BiCMOS a été améliorée de 21 % par rapport au modèle classique. De plus, la consommation d'énergie a été réduite d'environ 10 % par rapport au modèle classique.

Un nouveau système de reconfiguration dynamique est proposé pour les circuits SRAM de l'ordre du mégabit, basé sur un système intégré de détection et de reconfiguration des mots défectueux afin d'économiser des cellules de mots en utilisant des éléments logiques à valeur multiple. On a implanté une architecture de reconfiguration dynamique pour les circuits SRAM de l'ordre du mégabit. Afin d'évaluer la fiabilité de ce type de circuit intégré, on a développé un modèle. Le système de simulation HP 9000 haute vitesse est utilisé pour simuler la fiabilité du circuit SRAM de 1 Mb. La fiabilité d'un tel circuit a été améliorée d'environ 30 % par rapport aux autres méthodes.

On a développé deux systèmes de reconfiguration dynamique des circuits SRAM BiCMOS de l'ordre du mégabit, ce qui permet de détecter les défectuosités au niveau du circuit et de reconfigurer automatiquement les cellules de mémoire sur celui-ci afin d'effacer les erreurs. Un modèle est proposé pour les circuits intégrés SRAM l'ordre du mégabit. L'implantation d'un mécanisme de reconfiguration dynamique est présenté. La conception de base est réalisée au moyen du progiciel de CAD de Mentor Graphics. La performance a été simulée sur le simulateur haute vitesse HP 9000. Le temps d'accès du SRAM BiCMOS a augmenté de 35 %, la surface du circuit intégré a été réduite d'environ 25 % et le

rendement de fabrication a augmenté de 10 % par rapport aux méthodes classiques. La fiabilité des circuits SRAM BiCMOS de 1 Mb a augmenté de 78 % par rapport aux méthodes classiques.

On propose une technique d'essai rapide des paramètres source afin d'améliorer la robustesse de la conception, de réduire les défaillances de fabrication et d'améliorer la fiabilité de fonctionnement des circuits SRAM de l'ordre du mégabit. La technique proposée est intégrée et mise en oeuvre durant les phases de conception et de prototypage des circuits SRAM. Les résultats de l'essai des paramètres source sont utilisés comme intrants pour la conception et la fabrication des circuits SRAM de 1 Mb afin d'en améliorer la fiabilité. Environ 20 % des défaillances de conception et environ 30 % des défaillances de fabrication peuvent être identifiées et évitées en vue de la production massive de circuits intégrés SRAM de l'ordre du mégabit. Leur fiabilité est améliorée d'environ 50 % au moyen de cette technique.

Les résultats présentés dans cette thèse sont très utiles pour la conception future des circuits SRAM, DRAM, EPROM, FPGA et ASIC de l'ordre du mégabit, et ils permettront d'en améliorer la performance.

ABSTRACT

Performance evaluation is required at every stage in the life cycle of mega bit Static Random Access Memory (SRAM) chip design, manufacturing, procurement, application and so-on. A performance evaluation is required when a SRAM chip designer wants to compare a number of alternative designs and find the best design. Even if there are no alternatives, performance evaluation of the current SRAM design helps in determining how well it is performing whether any improvements need to be made for future designs due to technology changes. The goal of mega bit SRAM users is to get the highest performance and reliability for a given cost.

In order to meet the performance expectations of mega bit SRAMs: speed, power dissipation, propagation delay, noise, chip area, chip density, reliability, fault detection, fault isolation, reconfiguration and yield issues need to be addressed. The objective of this research is to propose SRAM performance models, analyze mega bit SRAM performance parameters, validate each performance model with experimental studies, and compare performance improvement with conventional designs.

Interconnect capacitance model is proposed based on SRAM cell oxide capacitance, fringing capacitance, coupling capacitance, and parasitic capacitance.

A practical concept is translated into mathematical domain to solve for interconnect capacitance and evaluate performance impact on mega bit SRAM device parameters. MT5C1008 device performance simulated using the HP 9000 high speed simulation system. The temperature dependency of access time is measured through thermocouple test system. SRAM failure analysis has been performed to identify interconnect failure mechanisms using scanning Electron Microscope (SEM) and optical microscope. Theoretical results, simulations and experimental results are in close agreement with proposed interconnect model.

By applying the triple -level interconnection to 1-Mb SRAM chip, the address access time is reduced from 35 ns to 25 ns. Maximum address access time is decreased to 30.8 ns, when increasing the temperature to 100°C, with power dissipation of 1 W. The wiring length and chip size were reduced to 69% and 58% of those the double metal interconnection. A wafer yield of 10% high has been achieved by implementing the proposed triple-level interconnect in the SRAM design. It is confirmed from the experimental results that model is very useful for future SRAM and DRAM chip design process.

A closed-form expression for CMOS SRAM chip propagation delay is developed. A propagation delay model is proposed, based on the SRAM cell capacitive coupling and loaded interconnects. The non-linear nature of the driving gate has been taken into account by applying a first order transistor model for the SRAM cell. Theoretical propagation delay for the SRAM has been computed using

MATLAB program. A 16-Mb SRAM chip propagation delay simulated using HSPICE running on HP A1097 work station. A HP8002 Logic Analyzer is used to measure SRAM chip propagation delay. The closed-from propagation delay expression results in absolute maximum error smaller than 4.8% in comparison with the measured data. Theoretically computed propagation delay values for the SRAM chip result in an error smaller than 2% in comparison with HSPICE simulated results The closed form propagation delay expression is validated with experimental results.

High speed circuit design techniques for BiCMOS SRAM are investigated to reduce interconnect delay time, improve speed, reduce noise, and power consumption. HP 9000 high speed simulation system is used to simulate the BiCMOS SRAM chip performance parameters. The BiCMOS SRAM chip speed performance has been improved 21% higher than the conventional BiCMOS SRAM design. The power consumption has been reduced about 10% than the conventional BiCMOS SRAM design. A novel dynamic reconfiguration scheme is proposed for mega bit SRAMs, based on on-chip word failure detection and reconfiguration to spare word cell using multi-valued logic elements. Dynamic reconfiguration architecture for mega bit SRAM chip implemented. In order to evaluate reliability performance of the mega bit SRAM chip a model is developed. HP 9000 high speed simulation system is used for mega bit SRAM chip reliability simulations. A 1-mb SRAM chip reliability performance has been improved about 30% in comparison with other methods.

Two dynamic reconfiguration schemes are developed for mega bit BiCMOS SRAMs, which allow the failure detection at the chip level and automatic reconfiguration to fault free memory cell within the chip. A model is proposed for mega bit SRAM chip.Dynamic reconfiguration scheme implementation is presented. Basic design layout is produced using the Mentor Graphics CAD tool package. Performance simulations have been performed using the HP 9000 high speed simulation system. BiCMOS SRAM access time improvement of 35%, chip area reduced about 25%, and chip yield of 10% higher achieved in comparison with the conventional design methods. The 1-Mb BiCMOS SRAM chip reliability performance has been improved by 78% in comparison with conventional methods.

A rapid root cause test technique is proposed to improve design robustness, reduce manufacturing related failures, and improve mega bit SRAM reliability performance. The proposed technique is integrated and implemented during the SRAM chip design and prototype phases. The root cause test results are used as design and manufacturing feed back for improving mega bit SRAM chip reliability. About 20% of design related failures and about 30% of the fabrication process related failures can be identified and prevented for volume production of mega bit SRAM chips. The 1-Mb SRAM chip reliability performance improved by about 50% using this technique.

The results presented in this thesis are very useful for future mega bit SRAMs, DRAMs, EPROMs, FPGAs and ASIC design process to improve performance.

TABLE OF CONTENTS

CHAPTER 1

CHAPTER 2

CHAPTER 3

CHAPTER 4

CHAPTER 5

CHAPTER 6

CHAPTER 7

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 General Introduction

The recent strong demand for high-speed and high-density Static Random Access Memories (SRAMs) has come mainly from large scale computing systems, super computers, satellite systems, telecommunication systems and automotive control systems. On the other hand, high-density SRAMs with a low standby power dissipation characteristics, suitable for battery-backup application, are also increasingly in demand for such uses as portable communications equipment. The design and evaluation of mega bit SRAMs must consider both performance and reliability.

The ever increasing demand for high performance and high density mega bit SRAMs has resulted in a significant scaling of device geometries. This scaling, however, has been predominantly two dimensional. The thickness of the interconnecting metallization and interlayer dielectric are generally not reduced, because this would result in an increase in the associated resistance-capacitance(RC) time delay constant. Therefore, the aspect ratio of the contacts increases, and the spacing between features decreases with the scaling. This scaling has significant impact on the photolithography, etching, and interconnecting metallization aspects of the SRAM device fabrication. Poor step coverage at the contact or via sites will have negative effects on the performance and reliability of SRAM devices having sub-micron geometries.

One of the problems in developing mega bit SRAMs having fast access time and low power dissipation is the large parasitic capacitance of long data bus lines, estimated to be several picofarads per line. The value of polysilicon load resistance was quadrupled at each step in quadrupling the memory bit density. The relation between the resistance of the memory cell and the leakage current of the memory cell node becomes a very important design criteria. The data holding current in the high resistive load memory cell is limited by the resistance of the polysilicon load. Increased SRAM chip complexity result in scaled-down design rules, as well as increased chip size and reduced voltage have created an increasing number of device and circuit problems, which becomes a major road block in achieving high performance and reliability. Scaled-down bit-line spaces increase the coupling noise of the SRAM chip. Interconnects limits the device density performance of SRAMs, and reliability.

Major performance concerns in high speed SRAM design are not limited to SRAMs themselves but also include multi-layer interconnect design, packaging, circuit design techniques, redundancy, reconfiguration, fault isolation, fault recovery methods, and device interfaces. Improvement in MOSFET characteristics alone can not provide sufficient performance to meet computer system requirements. Improved circuit design techniques, multi-layer interconnect design, parasitics, delay, noise, speed, redundancy, reconfiguration, reliability, yield and failure analysis feedback are necessary for future mega bit SRAMs in order to meet advanced computers and telecommunication systems requirements.

Performance modelling and analysis of mega bit SRAMs has evolved into a broad discipline, one that encompasses all practical performance aspects of mega bit SRAM design. Mega bit SRAM performance analysis can be carried out by modelling all device performance parameters and how they interact during device operation. Performance modelling is the most critical aspect in the SRAM chip high level design. The SRAM chip performance parameters such as propagation delay, noise, chip density, chip area, speed, and power dissipation are strongly dependent on the device technology, multi-layer interconnects, and package design. Allowing the technology to influence device design in the early stages improves the SRAM performance. Systems approach need to be applied for high density, and high-speed SRAMs in order to meet performance requirements. Performance modelling and analysis plays a vital role for all future mega bit SRAM chip designs.

Performance modelling and analysis is broadly divided into two parts:

1) Performance modelling of mega bit SRAM chip addressing all critical device performance parameters (eg., speed, propagation delay, noise, power dissipation etc).

2) Performance analysis includes evaluating device reliability, reconfiguration, yield, and root cause test etc.

The development of high performance mega bit SRAMs requires special consideration of many research issues. The most critical areas are multi-layer interconnects, packaging, high speed design, redundancy, reconfiguration, yield and failure analysis. In this research work mega bit SRAM chip performance modelling and analysis of various performance parameters are modeled and analyzed.

1.2 Literature Survey

Performance Modelling and analysis of mega bit SRAMs play a vital role in the design of large-scale computer systems, super computers, and the telecommunication systems. Performance modelling of mega bit SRAMs related research problems pertaining to multi-layer interconnects, dynamic reconfiguration methods and rapid root cause test techniques have been searched for the last fifteen years. Special emphasis is placed on very large scale integrated (VLSI) circuit design related performance problems, rather than the material properties and mechanical design considerations. There is very limited literature available in this research field.

McGrevity [1] presents interconnects/gates in VLSI technologies through the 80's and beyond focused on the device complexity evaluation. In large VLSI chip, major fraction of the area is occupied by the interconnections and not by the devices. In a complex chip, the total wire length can be considerable, and the wiring load capacitance per gate can be the most important speed limitation. High speed circuits are most often operated at their power dissipation limit. From the device point of view, optimal chip design for high speed operation hence must satisfy two conditions. The first condition requires the reduction of the transistor active area. The second condition requires reducing the area of the access zones, which concerns with the technology.

Ho [2] confirms that the VLSI interconnect metallization is a major problem in the SRAM fabrication process due to limited choice of interconnect technologies. The overall maximum integration density achievable is often fixed by the interconnects. This, in turn affects the device performance parameters. Gardner [3] discussed interconnection and electromigration scaling theory. Interconnects scaling impact on electromigration has been studied.

Small and Pearson [4] discussed on-chip wiring for VLSI: status and directions. Different on-chip wiring technologies have been investigated. As a result intermetallic compounds formation was observed, which tend to degrade the reliability of on-chip interconnects. Interconnections are classified into two types: short range interconnections and long range interconnections.

Lin and Mead [5] presented signal delay in general RC networks. The formula used to calculate the signal delay is based on the approximation theory. The signal delay calculations were not straight forward at the transistor level. It can be used for a very small device with up to ten gates successfully. The method has serious limitations, when applied for a complex device. Sakuri [6] presented an approximation of wiring delay in MOSFET LSI circuits. The delay error was estimated about 30%. Parasitic capacitance associated with the device has not been accounted in this model.

Kolias [7] presented packaging/ performance design trade-offs in high speed digital computers. The package performance impact on the system was studied. Package design is one of the key performance problem for future VLSI circuits. Wireability is treated in some detail and Rents rule is discussed. Uttecht and Geffken [8] presented a four-level-metal interconnect technology used to wire high-density, high performance logic and SRAM chip designs. Process details are described along with the results of standard electrical tests and reliability stresses. Rayapati and Mukhedkar [9] confirmed interconnection problems in VLSI Random Access Memory chip. Mega bit SRAM chip interconnect problems were identified. A multi-level interconnect approach is proposed to over-come on-chip interconnection problems.

Coulton et al [10] addressed some fundamental issues on metallization in VLSI. Metallization and conductors are critical part of the VLSI chip, and can set limits on future down-scaling of integrated circuits. Due to decreasing lateral and vertical dimensions, interconnections are rapidly becoming a major problem in terms of device yield, reliability, signal delay time, and inter-device interactions. The authors discussed how interconnect limitations will affect the scaling of advanced devices. A number of issues regarding the interconnection technologies that will be required for future ULSI circuits were discussed.

Grosspietsch et al [11] had proposed a special reconfiguration scheme for stand-by memory systems. It involves conventional reconfiguration to replace faulty chips. It fails to detect all transition faults in the main memory. The proposed scheme is not cost effective for VLSI implementation. Gray [12] has described about an architecture that is capable of performing self-reconfiguration at the module level. This capability has been exploited to implement a special purpose architecture.

Raghavendra [13] has presented details of computer systems with large number of processors which cause major reliability problems. One solution to this problem is to build redundant communication paths using reconfiguration in computer network design. Lowrie and Kent Fuches [14] have developed an approach to the design of reconfigurable tree architecture. The spare processors are allocated using the sub-tree oriented fault tolerance(SOFT) approach. This approach is more reliable than the previous approaches and is capable of tolerating link and switch failures for both single chip and multichip tree implementation while reducing redundancy in terms of both spare processors and links. The author [15] proposed a new modular fault tolerant VLSI parallel processor architectures with dynamic redundancy. The proposed architectures significantly increase reliability and yield. The limitation is the speed reduction in the processor due to inherent architectural deficiencies. The literature survey indicates that there is no relevant literature available on dynamic reconfiguration architectures.

From the literature review, it is evident that the following research problems pertaining to mega bit SRAMs have not been analyzed:

1) Multi-layer interconnect capacitance modelling including all parasitic capacitance, fringing capacitance and coupling capacitance effects.

2) Multi-layer interconnect delay modelling related to SRAM device and device load effects. There is no closed-form solution available in the literature.

3) Multi-layer interconnect related BiCMOS SRAM chip package capacitance and inductance models were not reported in the literature.

4) High speed circuit design techniques related to multi-layer interconnects and performance impact on mega bit SRAM design were not addressed in the literature.

5) Dynamic Reconfiguration techniques pertaining to SRAMs have not been addressed either at the chip level or at the device level in the literature.

6) Rapid root cause test technique for high density VLSI circuits have not been found in the literature. In particular, for SRAMs there is no relevant information available in the literature survey.

The literature survey indicates that not much work has been done pertaining to performance modelling and analysis of CMOS, BiCMOS, and GaAs mega bit SRAMs. The present work focuses on multi-layer interconnect modelling, interconnect delay modelling, multi-layer interconnect related chip package modelling, high speed circuit design techniques, dynamic reconfiguration methods, rapid root cause test technique and performance evaluation for mega bit CMOS and BiCMOS SRAMs.

1.3 Research Problem and Scope

The ultra-high speed SRAM is now the target of competing technologies such as GaAs, BiCMOS, CMOS, and Bipolar. The systems on silicon concept demand is increasing for high speed, high density SRAMs. As the device complexity increases, device dimensions are scaled-down proportionately. Mega bit SRAM requires very high density, which in turn, increases chip size and lowers the yield. Reduced operating voltage is one of the key driver for new high speed circuit design innovations. All the above mentioned issues have created an increasing number of device and circuit problems that make it more and more difficult to achieve high density, high-performance SRAMs.

Major concerns in high speed SRAM design are not limited to SRAMs themselves but also include multi-layer interconnect design, packaging technology, performance, reliability and cost. The motivation for this research is to develop CMOS and BiCMOS SRAM performance models and analyzing SRAM performance issues pertaining to multi-layer interconnects, multi-layer interconnect delay, high-speed circuit design methods, dynamic reconfiguration at the device and chip level, and rapid root cause test technique development.

Early research in the SRAM design field did not address many performance problems related to multi-layer interconnects, high-speed design approaches, dynamic reconfiguration and rapid root cause test method. There is very limited literature available in this area of research. However, in order to achieve high performance, high-speed, high density and high reliability for SRAMs there is a definite need to investigate performance modelling and analysis aspects to fulfill future application requirements.

The objectives of this research work are discussed and stated as follows:

1) To develop mega bit CMOS, BiCMOS SRAM chip performance models and analysis.

2) To investigate high speed circuit design techniques for BiCMOS SRAMs.

3) To develop dynamic reconfiguration schemes for mega bit SRAMs and provide device reliability performance comparisons with other existing methods.

4) To develop integrated root cause test technique for mega bit SRAMs. In order to improve device reliability and eliminate/prevent manufacturing related failure mechanisms and failure modes.

The above mentioned research objectives are supported through original research papers. Each research paper presented in this thesis fulfills one specific research problem in detail. These research papers provide a systematic approach for problem formulation, model development, experimental validation, performance analysis results and applications for future mega bit SRAMs.

1.4 Organization of Dissertation

The dissertation structure is based on the original research papers presented in this thesis. The contents of this thesis have been organized into even chapters.

Each research paper is presented as one chapter in the thesis. The format and contents of each chapter is based on the research paper structure.

In chapter 2, multi-layer interconnect capacitance model is presented for mega bit SRAMs. Interconnect problems and performance issues were addressed. Experimental studies of 1-Mb SRAM chip are discussed.

In chapter 3, multi-layer interconnect delay model related to SRAM device and load effects are presented. A closed-form solution for multi-layer interconnect propagation delay is developed. Experimental validation of the propagation delay model for a 16-Mb SRAM chip is presented.

In chapter 4, High speed circuit design techniques for BiCMOS SRAMs related to multi-layer interconnects and performance analysis details are presented.

In chapter 5, a novel dynamic reconfiguration scheme for mega bit SRAM is proposed. A reliability model is presented to validate the proposed reconfiguration scheme.

In chapter 6, two dynamic redundancy schemes are proposed and analyzed. Design considerations and performance evaluation of a fault-tolerant BiCMOS SRAM is presented. A case study of 1-Mb BiCMOS SRAM chip is investigated and results are provided.

In chapter 7, a rapid root cause test technique for mega bit SRAM chip is presented. A 1-Mb CMOS SRAM chip root cause test results are provided.

Research results and conclusions are presented. Future research recommendations in the area of CMOS, BiCMOS and GaAs mega bit SRAM performance modelling and analysis are addressed.

REFERENCES

[1] D.J. MCGREVITY, "Interconnections/gates in VLSI technologies," in VLSI Technologies through the 80's and beyond, IEEE Computer Society Press, 1982.

[2] P.S. HO, "VLSI interconnection metallization,"Semiconductor International, August 1985.

[3] D.S. GARDNER etall, "Interconnection and electromagnetic scaling theory," IEEE Transactions Electron Devices, Vol. ED-34, March 1987, No. 3.

[4] H.B. SMALL and D.J. PEARSON, "On-chip wiring for VLSI: status and directions," IBM Journal of Research and Development, vol. 34, November 1990, No. 6.

[5] T. LIN and C.A. MEAD, "Signal delay in general RC networks,"IEEE Trans. on CAD, vol. CAD-3, October 1984, No. 4.

[6] T. SAKURI, "Approximation of wiring delay in MOSFET LSI,"IEEE Journal of Solid State Circuits, vol. SC-20, December 1985, No. 6.

[7] JOHN T. KOLIAS, "Packaging/Performance Trade-Offs in high speed computer systems," VMIC conference, June 1989, pp. 49-58.

[8] R.R. UTTECHT etal, "A four -level-metal fully planarized interconnect technology for dense high performance logic and SRAM applications," VMIC conference, June 1991, pp. 20-26.

[9] V.N. RAYAPATI and D.MUKHEDKAR, "Interconnection problems in VLSI random access memory chip, " in SPIE, Proc. Int. Conf. on advances in Interconnection and Packaging, vol. 1389, 1990, pp. 98-109.

[10] D.E. COULTON, K.R. GLEASON, K. JAMES, E.W. STRID, "Accurate measurement of high speed packages and interconnect parasitics," IEEE/CHMT IEMT Symposium Japan, October 1989, pp. 276-279.

[11] K.E. GROSSPIETSCH, J. KAISER and E. NETT, "A dynamic standby system for random access memories", proceedings of IEEE international symposium on fault tolerant computing, 1981, pp. 268-269.

[12] F. GAIL GRAY, "General purpose reconfigurable architectures,"Proceeding of IEEE international conference on circuits and computers, 1982, pp. 122-123.

[13] C.S. RAGHAVENDRA, "Fault tolerance in regular network architectures," IEEE micro, 1984, pp. 44.

[14] M.B LOWRIE and W. KENT FUCHS, "Reconfigurable tree architectures using subtree oriented fault tolerance,"IEEE Trans. on computers, 1987, vol. C-36.

[15] V.N. RAYAPATI, "Modular fault tolerant VLSI parallel processor architecture with dynamic redundancy," International Journal of Microelectronics and Reliability, vol. 30, No. 2, 1990, pp. 213-236.

CHAPTER 2

PERFORMANCE ANALYSIS OF MULTI-LAYER INTERCONNECTIONS FOR MEGA BIT STATIC RANDOM ACCESS MEMORY CHIP

Article by: Venkatapathi N Rayapati and Bozena Kaminska,

Published in IEEE Transactions on Components, Hybrids, and

Manufacturing Technology, Vol.16, No.5, pp. 469-477,August 1993

ABSTRACT

The objective of this paper is to analyze interconnection problems in the mega bit Static Random Access Memory (SRAM) chip. A multi-layer interconnect capacitance model is developed for mega bit SRAM chip. Interconnection effects on SRAM device performance parameters such as propagation delay, speed, power consumption, and noise characteristics are analyzed. A case study of 1 Mb SRAM chip interconnection problems is discussed. A multi-layer interconnect approach is proposed for SRAMs to overcome on-chip interconnection difficulties. Implementing a triple layer interconnect approach the wire length and chip size were reduced to 69% and 58% respectively. Maximum access time of 30.8 ns with 1 W at 100 ^{0}C and wafer yield as high as 10% is achieved. Experimental results of multi-layer interconnections for 1 Mb SRAM are provided. The analysis results are found to be very useful for future mega bit SRAMs.

2.1 INTRODUCTION

Effective utilization of very high speed Static Random Access Memories (SRAMs) requires the development of a suitable advanced interconnection technology. The interconnection technology used in high performance SRAM system configured with these high speed integrated circuits should be compatible with the following parameters: propagation delay, speed, power consumption and noise characteristics. The integration of a large number of components on a single SRAM chip requires sophisticated interconnections to minimize signal delays, power consumption, and simultaneously optimize the packaging density. Interconnections play a vital role in the mega bit SRAM chip packaging design because of cross talk, electromagnetic interference/compatability, reflections, and noise characteristics. Packaging density depends on the interconnect wire dimensions, and spacing between the adjacent conductors. Interconnect design for SRAMs depends on the dimensional tolerance of the interconnect wires and selection of dielectric material. Low dielectric constant materials usage for interconnects improve signal propagation characteristics and provides compatability with standard SRAM device manufacturing process. Silicides and polymides are attractive candidates for high density interconnect technology [1], because of their low dielectric constant values and excellent thermal stability properties.

The need for high-speed, high-density static RAMs (SRAMs) is rapidly increasing. Mega bit SRAMs enhance the system performance of computers, micro-computers, work stations, and high speed test systems. In the mega bit SRAM chip design era, interconnections have major importance. Today, a mega bit SRAM chip contains not only a memory circuit, but also contains a good part of an entire memory system. A complex 1 mega bit SRAM chip offers no flexibility in SRAM cells and associated circuits. This implies that interconnects limits the density of active devices and causes propagation delay, speed, power consumption and noise problems [2]. The effect of interconnects on mega bit SRAM chip performance is investigated in this paper.

This paper provides performance analysis of multi-layer interconnections for mega bit SRAM chip. A multi-layer interconnect capacitance model is presented in section 2.2. The multi-layer interconnect capacitance model is based on the substrate capacitance and the coupling between adjacent metal lines in the mega bit SRAM chip. Interconnection effect on SRAM device performance parameters, such as propagation delay, speed, power consumption and noise parameters are analyzed in section 2.3. A case study of 1 Mb SRAM chip interconnection problems is discussed in section 2.4. Performance evaluation of 1 Mb SRAM chip with double layer interconnections and triple layer interconnections is presented in section 2.5. Experimental studies of multi-layer interconnections for 1 Mb SRAM is provided in section 2.6. Conclusions and results are provided in section VII. Advantages of multi-layer interconnections for mega bit SRAMs are highlighted.

2.2 MULTI-LAYER INTERCONNECT CAPACITANCE MODELLING

In the past mega bit SRAM modelling techniques have concentrated on device local effects. However, with the decrease in feature device size, global effects are becoming more prominent, with interconnect capacitance playing a major role in the SRAM chip design. Capacitance modelling has been reported by Ward and Dutton [3], using the parallel-plate formula for estimating capacitances, resulting in large error. Mayer [4] developed MOS models and circuit simulation techniques, that gives approximate results. Shere et al and Oh et al [5,6] provided methods to calculate these parasitic capacitances in which finite parallel conductors are considered. Balaban [7] calculated the capacitance coefficients of planar conductors on a dielectric surface using the point fitting method. Silvester et al and Benedek [8,9] presented capacitances of a planar multiconductor configuration on a dielectric substrate by a mixed order finite method. However, for SRAM chip layout verification, the problems with these methods are: second-order integrals have to be evaluated and the direction of conductors to be identified. Second order integrals evaluation will take more computing time, while identification of conducting direction is an additional difficulty for a practical layout extractor.

In this section, multi-layer interconnect modelling and validation is presented. The physical concept of multi-layer interconnects are shown in Figure 2.1. In the figure, for a single metallization level, the capacitance is composed of substrate capacitance and the coupling capacitance. The substrate capacitance is the oxide parallel-plate C_{ox} plus the fringing capacitance C_f. In order to evaluate the fringing capacitance of a multi-layer interconnect a mathematical model is developed.

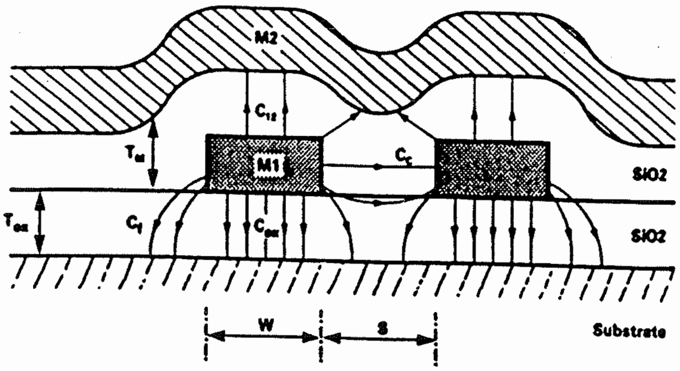


Figure 2. 1: Capacitance distribution in multilayer interconnections

Multi-layer interconnect is treated as a three-dimensional system as shown in Figure 2.2. The ground plane is represented as substrate (i.e., IC wafer) region 1, which is an equivalent representation of the first metallization layer of Figure 2.1. Region 2 corresponds to the second metallization layer of Figure 2.1. In order to evaluate the capacitance coefficients of the 3-dimensional multi-layer interconnect, an integral equation approach is used. In Figure 2.1 the concept of multi-layer interconnect is considered as a stratified medium with the conducting ground plane as the substrate and a sandwich of dielectric layers as oxide thickness SiO_2, in which the conductors 'float' as shown in Figure 2.2.

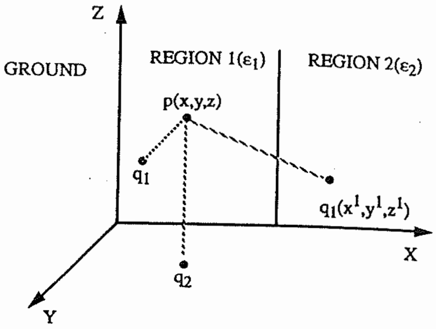


Figure 2.2: Capacitance 3-D representation as a stratified dielectric medium

The integral equation relating to the electrostatic potential $\Phi(P)$ to the charge density $\sigma(q)$ in a chargeless 3-dimensional dielectric medium can be represented by using Green's theorem,

$$\Phi(P)= \int_{all\ charge} G(p,q)\sigma(q)dq \qquad (1)$$

G (p,q) is Green's function, which is physically the potential at one point in space P (x,y,z) due to a unit point charge at another point q (x^1, y^1, z^1). For the multi-layer interconnect, the equation (1) can be rewritten as:

$$\Phi_j(P)=\sum_{i=1}^{M}\int_{S_i} G(p,q)\sigma_i(q)d_{Si}(q), \quad j=1,2,...,M \qquad (2)$$

where $\sigma i(q)$ denotes the charge density on the edges of the ith conductor surface S_i solving for Green's function and charge density. Now, the potential $\Phi_W(P)$ can be written as:

$$\Phi_W(P)=\Sigma_{j=1}^{M}\Sigma_{i=1}^{N_j}\alpha_{ji}F_{ji}(P), \quad W=1,2,\ldots,L \tag{3}$$

where $F_{ji}(P)$ = $_{ji\ th\ cell}$ G (p,q) $f_{ii}(q)$ $dl_{ji}(q)$, and L is the total number of SRAM cells within the SRAM chip

$$L=\sum_{j=1}^{M}N_j$$

Equation (3) can be represented in matrix notation as follows:

$$FA=J\phi \tag{4}$$

where F is an L X L matrix of the single integrals. The vector A^T denotes the set of unknown coefficients to be determined. The vector$\Phi^T=(\Phi_1, \Phi_2,....\Phi_M)$ is the set of the M distinct potentials of the multi-layer interconnects, while J is an L X M incidence matrix of nodes and conductors, in which $J_{ii}=1$, if conductor i contains node j. Now the capacitance matrix C is defined as:

$$Q = C\,\Phi, \tag{5}$$

where Q is the vector of total charge on the conductor.

From equations (3) and (5), we can have

$$Q = J^{T}A \tag{6}$$

and

$$A = F^{-1}J\phi. \tag{7}$$

Equation (7) is substituted into equation (6), and comparing with equation (5), finally we can get

$$C = J^{T}F^{-1}J. \tag{8}$$

The matrix equation (8) can be solved for the capacitance matrix by any standard technique.

Validation of the multi-layer interconnect capacitance modelling is explained as follows: the substrate capacitance is the oxide parallel plate capacitance C_{ox} plus the fringing capacitance C_f as shown in Figure 2.1. The fringing capacitance can be calculated by using the above derived multi-layer interconnect capacitance model. Therefore, the substrate capacitance is given as

$$S_c = C_{ox} + C_f \tag{9}$$

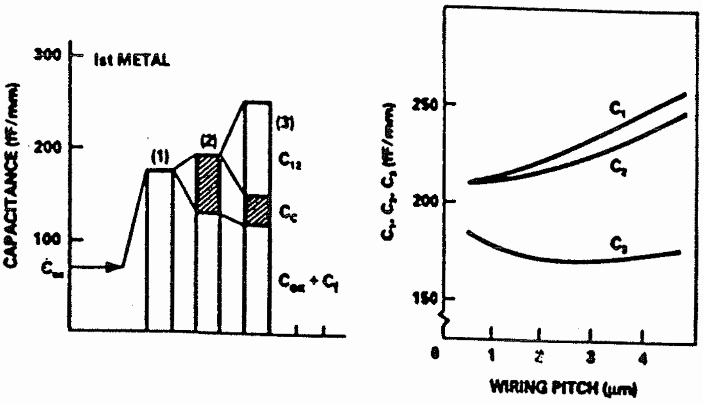
The coupling capacitance C_c between the adjacent interconnects of SRAM cells is shown in Figure 2.1. The parasitic capacitance C_p effect is considered between the interconnects of SRAM cells and other functional blocks of the SRAM chip. Therefore, the coupling capacitance is expressed as follows:

$$C_c = C_{ox} + C_p \tag{10}$$

For all practical purposes C_f and C_p are assumed to be of the same order of magnitude. Equations (9) and (10) can be expressed together as the interconnect capacitance C_I given as:

$$C_I = \epsilon_o L \frac{W + W_o}{T_{ox}} + \frac{T_m + T_{mo}}{S_c} \tag{11}$$

where L, W, S_c, T_{ox}, and T_{mo} are the dimensions of the interconnect as defined in Figure 2.1. ε_o is the dielectric constant of the oxide. W_o and T_{mo} are the correcting factors accounting for fringing and coupling capacitances. For a 2µm line width and T_{ox} = 0.575 µm, C_f has been found approximately of the same order of magnitude as C_{ox}. For multi-layer interconnect capacitance modelling results are shown in Figure 2.3(a) and 2.3(b). For metal single layer, two layers, and two level layers are shown in Figure 2.3(a). Figure 2.3(b) explains the variations of the total capacitance for 1st layer (C1), 2nd layer (C2), and 3rd layer as function of the pitch (W + S).



(a)

(b)

Figure 2.3: Capacitance in (a) for metal 1 layer : (1) single line, (2) multiple lines, (3) two levels; The dimensions are $T_m = T_{ox} = 1$ μm, W = S = 2 μm (b) Capacitance variations for first layer (C1), second layer (C2), and third layer (C3) as a function of thee pitch (W+S)

The fringing term of an isolated line in the first layer is of the same order of magnitude as the oxide capacitance. Conforming W_0 is about 2μm, however this term is strongly influenced by the presence of adjacent lines, so that the total capacitance of a line is affected by the proximity of another. The second layer has less significant impact on the first. When the thickness of both dielectric and metals are scaled as the square root of λ, the total capacitance per unit length of the first two layers decreases down to 1μm pitches, while the third is almost constant.

A practical concept is translated into mathematical domain and solved for capacitance matrix. By solving capacitance matrix, multi-layer interconnect capacitance values are calculated. The calculated multi-layer interconnect capacitance values are compared with the actual multi-layer design parameter. The results presented in Figure 2.3(a) and 2.3(b) are in agreement with theoretical calculations. Hence, the proposed multi-layer interconnect capacitance model is validated through practical SRAM design problem.

2.3 EFFECT OF INTERCONNECTS ON SRAM DEVICE ELECTRICAL PARAMETERS

(a) Propagation delay

The analysis of a SRAM cell shows that the propagation delay is the sum of the intrinsic delay associated with the device switching and the two transition times t_1 and t_2, involving the device access network and the loading elements. A simplified charge-control equivalent circuit concept is used to give a reasonable insight into the performance of a SRAM cell matrix.

The transition times t_1 and t_2 can be written as follows:

$$t_1 = (F_O\ (C\pi + 2\ C_U) + C_O + F_O\ .\ C_W)\ (V_{OH} - V_{TH})\,|\,I_{max}\text{-pull-down}$$

$$t_2 = (F_O\ (C\pi + 2\ C_U) + C_O + F_O\ .\ C_W)\ (V_{TL} - V_{OL})\,|\,I_{max}\text{-pull-up}$$

I_{max} pull-down is the maximum current delivered by the pull-down device, I_{max} pull-up device, F_O is the fan-out number, C_W is the interconnect capacitance per fan-out and C_O is the device output capacitance. The propagation delay of loaded Complementary Metal Oxide Semiconductor (CMOS) SRAM cell and can be derived from the above expressions. Assuming the logic threshold voltage is at half swing propagation delay of a SRAM cell and is expressed as:

$$t_{pd} = t_T + \frac{C_T \Delta V}{2 I_d (\Delta V/2)}$$

$C_T = (C_g + C_W)\ F_O + C_O$ means the total load capacitance and I_d is the current across the inverter biased at the logic threshold. An extra interconnection delay must be added to the total circuit delay, the propagation delay is given as

$$t_{pd} = t_T + \frac{1.7}{U_o}[F_o . L^2 + (\frac{F_o . C_w + C_o}{C_{ox}})\frac{L}{W}]$$

The L/W term is significantly influenced for small inverter load with long interconnecting lines in the SRAM chip. Then t_{pd} is highly dependent on C_{ox} and varies quite linearly with gate oxide thickness as shown in Figure 2.4. The figure shows propagation delay as a function of oxide thickness for 0.3μm channel transistors.

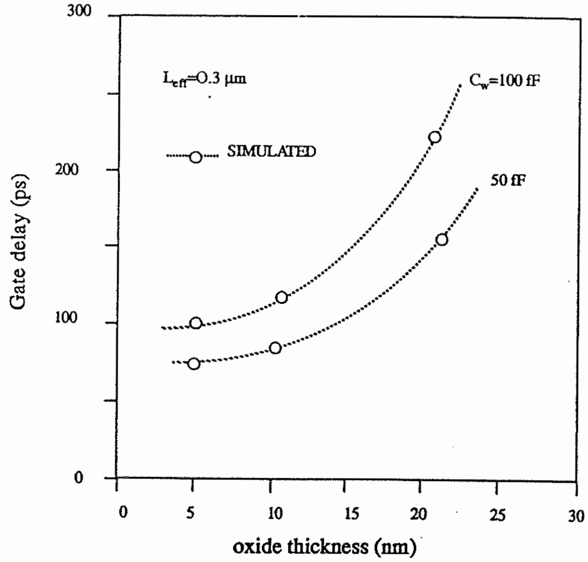


Figure 2.4: Influence of load capacitance on the device delay time as a function of oxide thickness for 0.3 μm channel transistors.

(b) Power dissipation

The static power dissipation P_d is extremely low in full CMOS based SRAM designs (leakage current < 1 nA per gate). The dynamic power dissipation however, has a significant impact on SRAM circuits, which operates continuously depending upon the application. The power dissipation is expressed as

$$P_d = N[C_g + C_W]V^2{}_{dd}FD_fF_o$$

The SRAM chip is assumed to have an N logic gates with an average fan-out F_o. When the circuit is operating at higher frequencies (for example >50 MHZ) then the dynamic power consumption of CMOS SRAM chip increases to high level unless the duty cycle can be kept low. Therefore, dynamic power consumption is dependent on operating frequency f_T and speed of SRAM device. The interconnects influences the dynamic power consumption in the SRAM chip.

(c) Speed

In addition to the intrinsic device characteristics, the power supply V_{dd} is of prime importance for the speed of the SRAM chip because it plays a major role in the device current. There is a strong pressure towards the use of lower voltages. This approach may drastically reduce the speed of CMOS SRAM's. The speed of SRAM is a function of intrinsic gate delays, interconnects induced delays and power supply voltage. The dependence of access time on supply voltage is shown in Figure 2.5. Therefore, interconnects play a vital role in the operational speed of the mega bit SRAM chip.

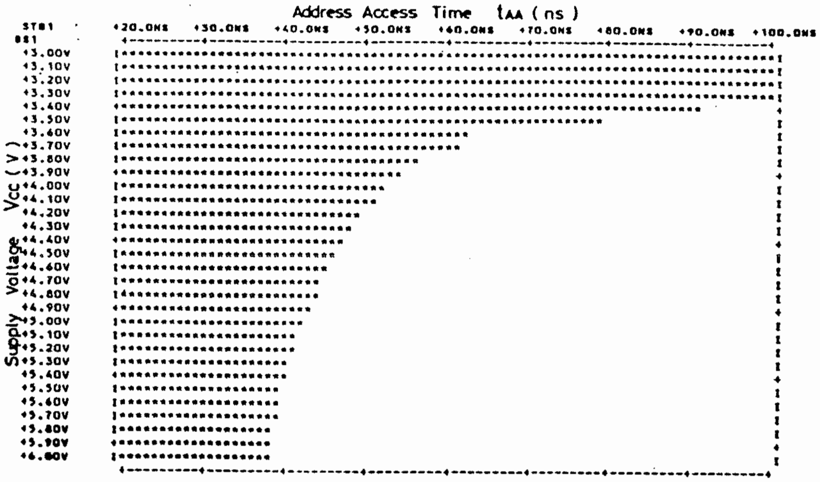


Figure 2.5: Access time versus supply voltage V_{CC}

(d) Noise margin

In a complex 1 mega bit SRAM chip, many sources of noise could cause the SRAM chip malfunction, and provide erroneous outputs. In general, noise could be classified into AC noise and DC noise. DC noise ranges from DC level offsets, voltage drop on power supplies or signal lines due to DC current loading which produces noise spikes whose pulse widths are longer than the propagation delay. AC noise spikes have pulse widths shorter than the logic gate delay. Noise spikes could be generated by capacitive coupling between signal lines caused by the transistor charging and discharging of large load capacitances. In the mega bit SRAM chip, input/output buffers are potential sources of this kind of noise. A typical noise spectrum for a multi-layer interconnect is shown in Figure 2.6. In the case of multi-layer interconnects, the coupling capacitance and the fringing capacitance causes noise problems in the mega bit SRAM chip.

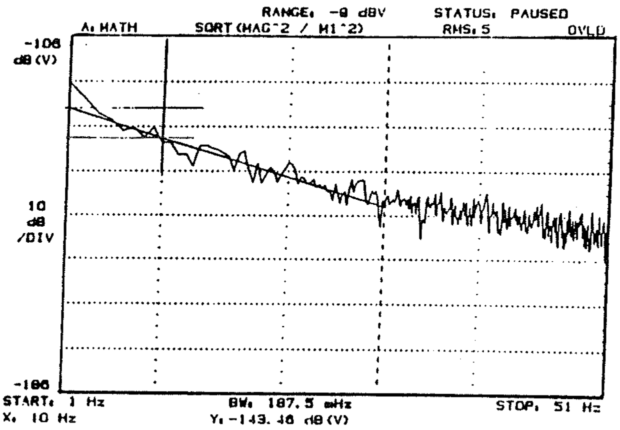


Figure 2.6: Noise spectrum for multilayer interconnections

2.4 INTERCONNECTION PROBLEMS IN 1 MEGA BIT SRAM CHIP

Interconnections in the mega bit SRAM chip will introduce propagation delays, noise transients, reduce speed, high power consumption, low packaging density problems. Mega bit SRAM chip integration density depends on the choice of interconnection technology. In the mega bit SRAM design, interconnects consume more power than the active devices. Interconnects induced AC and DC noise transients, which can affect the SRAM device functional performance.

Interconnects introduce significant propagation delays, which intern reduces mega bit SRAM device speed. Interconnects limit the density of SRAM cells placement within SRAM chip, complicates routing of SRAM cells, and reduced yield are also considered to be potential problems. The above-mentioned problems have significant impact on the mega bit SRAM device performance. Therefore, there is a definite need to investigate interconnection problems in the 1 Mb SRAM chip and improve performance. In order to analyze interconnection problems, a case study of 1 Mb SRAM chip interconnection problems and performance issues is evaluated in this section.

The functional block diagram of 1 mega bit SRAM chip architecture is shown in Figure 2.7(a). The memory cell array is divided into 16 sections. Each section consists of 512 word lines and 128 bit lines. Nine address buffers in the upperside to receive row address signals and send only bar signals to the row predecoders. Eight address buffers in the bottom side to receive column address signals and send only bar signals to the column predecoders. The row decoder is placed every two sections. The sense amplifier and the write data driver circuits are placed every 16 bit lines. The same lines are used for the sense amplifier output lines and write data driver input lines.

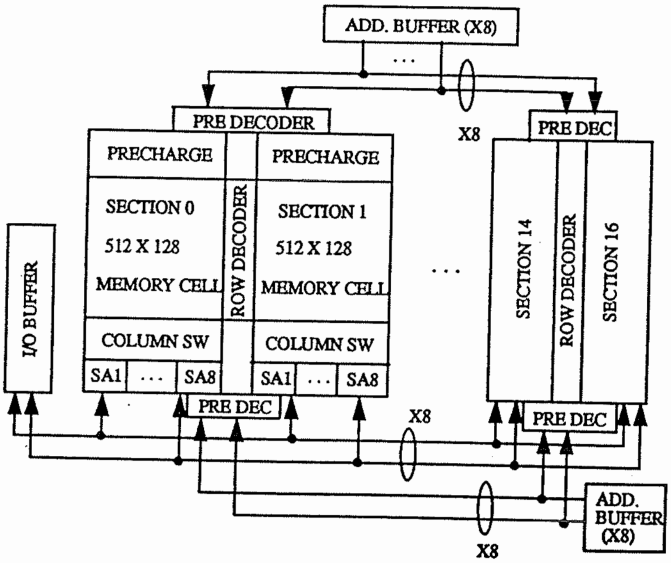


Figure 2.7 (a): Functional block diagram for 1-Mb SRAM

The SRAM cell layout is shown in Figure 2.7(b). The layout of the memory cell is based on PMOS transistor load. Multi-layer interconnect approach is used in this SRAM architecture. The SRAM has a single aluminum and triple-level polysilicon (including polycide) twin-well CMOS technology. The first polysilicon forms gate electrodes of the cell driver transistors and the word lines. The second polysilicon is used for the PMOS transistor loads and the V_{cc} lines. The third interconnect layer of silicide is used as the ground lines and upper word lines. The bit lines are formed by aluminum.

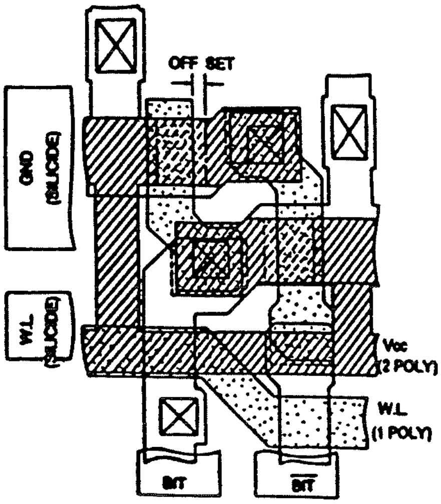


Figure 2.7 (b): SRAM cell layout

A cross sectional view of the memory cell is shown in Figure 2.7(c). The p $^{+}$ drain of the PMOS transistor is directly connected to the n^{+} polysilicon through a contact hole formed above the buried contact region. The current-voltage characteristics of a P-n junction between p^{+} polysilicon and n^{+} polysilicon is inferior to a diode formed between p^{+} diffusion and n-type substrate. The polysilicon diode characteristic is rather more ohmic than the P-n diode. Thus the effect of the polysilicon diode on the memory cell load device is negligibly small.

Both n-channel and p-channel devices have a lightly doped drain (LDD) structure. The gate oxide thickness is approximately above 200 A°. This assures reliability against hot-carrier induced device degradation.

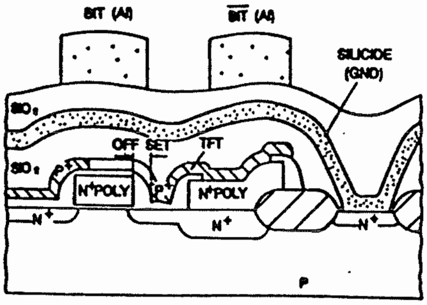


Figure 2.7 (c): Cross-sectional view of SRAM memory cell

The high-resistivity polysilicon resistor has been used as a SRAM cell load device for more than 10 years. But with increasing memory cell density, the required resistivity has become higher and higher. The result is a lower yield due to a smaller margin between load current and leakage current in the memory cell. A full-CMOS memory cell is ideal for this problem. However, it requires larger memory cell area and consequently leads to larger die size. This also lowers yield. Multi-layer interconnect approach is the solution to this problem. Multi-layer interconnect approach increases the complexity of SRAM chip. Multi-layer interconnects will have significant usage in the future mega bit SRAMs and DRAMS.

2.5 PERFORMANCE ANALYSIS OF 1 Mb SRAM WITH MULTI-LAYER INTERCONNECTION

The characteristics of the 1 Mb SRAM is evaluated through SPICE program using a high speed simulation system HP9000. Typical distributions of the address access time (t_{acc}) of the Mb SRAM with triple-level and the double-level interconnection technologies are shown in Figure 2.8 (a) and (b), respectively. In the test simulation, a galloping test pattern is used. By applying the triple-level interconnection, the maximum t_{acc} is reduced from 35 ns to 25 ns.

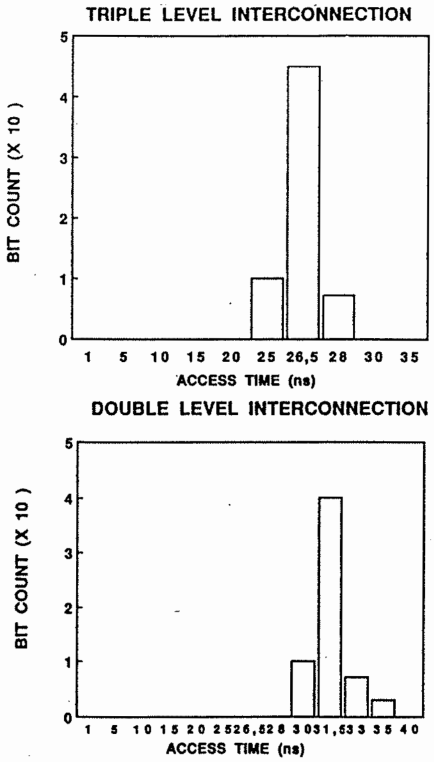


Figure 2.8: Access time comparison for mega bit SRAM's with multilayer interconnections

Based on a simple interconnection model, the improvement in the maximum t_{acc} is examined. We assumed that interconnect capacitance per unit length is the same for all the cases of (1) between the 1st and the 2nd level interconnections, (2) the 2nd and the 3rd, and (3) the 1st one and substrate, and it is denoted by C. We also assumed the same resistance per unit length for each level metal, which is denoted by R. The wiring length for the 1st and 2nd level are denoted by $_1$ and ℓ_2 respectively. The delay time t_1 through signal line ℓ_1 + ℓ_2 for double-level interconnection is written approximately as:

$$t_1 \approx (Rl_1 + Rl_2)(2cl_1 + cl_2) = RC(l_1 + l_2)(2l_1 + l_2) \tag{12}$$

The delay time, t_2, for triple-level interconnection is given by

$$t_2 \approx (R\alpha l_1 + R\alpha l_2)(2c\alpha l_1 + c\alpha l_2) = 2\alpha^2 RC(l_1 + l_2)^2 \tag{13}$$

where α is a given reduction ratio.

The above equations and the relation $\ell_1 \sim \ell_2$ result in the design of 1 Mb SRAM,

$$t_2 / t_1 \approx 4\alpha(2/3). \tag{14}$$

Here, the reduction ratio is 69%. So, we have t_2/t_1 = 64% from equation (3) and higher speed is expected. [From Figure 2.14, the maximum t_{acc} is roughly estimated to be reduced to 71% and it is close to the value obtained from equation (3). Therefore, it is confirmed that the improvement in the maximum t_{acc} is mainly achieved by the reduction of the wiring length.]

The temperature dependence of the maximum t_{acc} simulation are shown in Figure 2.9. It is less than 35 ns for the outside temperature range between 25°C and 100°C. Maximum t_{acc} decreases when increasing the temperature. At 100°C maximum t_{acc} is 30.8 ns with power dissipation of 1W.

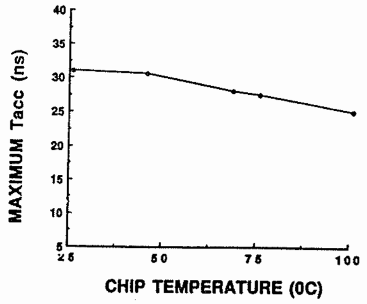


Figure 2.9: Temperature dependency of access time for 1-Mb SRAM

High speed and stable operation is accomplished for a wide temperature range between 25 and 100°C. From the analysis as expected, a wafer yield as high as 10% is achievable.

2.6. EXPERIMENTAL STUDIES OF 1 Mb SRAM CHIP

The purpose of experimental studies is to cover performance characteristics of SRAM device through functional verification, optical microscope and Scanning Electron Microscope (SEM) inspection observations pertaining to interconnect metal tracks, interconnect alignment, passivation protection to interconnects, and interconnect related failure mechanisms. SRAM Interconnects failure analysis has been performed to identify multi-layer interconnect failure mechanisms: step coverage, notching, electromigration, and voids using SEM experimental techniques.

The Micron MT5C1008 SRAM device is chosen for experimental studies. The SRAM device functional verification is discussed in this section. Micron MT5C1008 SRAM employs high-speed, low power CMOS design using a four transistor memory cell. MT5C1008 SRAM device is fabricated using double-layer metal, double-layer polysilicon technology. MT5C1008 SRAM device functional block diagram is shown in Figure 2.10. The flexibility of this design offers dual chip enable (CE1, CE2) lines as shown in Figure 2.10. This enhancement can provide high output impedance and allows additional flexibility in system design. Writing data into the SRAM device is accomplished, when write (WE) and CE1 inputs are both LOW and CE2 is HIGH as shown in Figure 10. Reading data from the SRAM device is accomplished, when WE and CE2 remain HIGH and CE1 goes LOW as shown in Figure 2.10. The device offers a reduced power standby mode when disabled. This allows system designers to achieve their low standby power requirements. These SRAM devices operate from a single +5 V power supply and all inputs /outputs are fully TTL compatible. The performance characteristics of 1 Mb SRAM device is evaluated using high speed test simulation system HP9000. Functional test verification has been performed using galloping test patterns. 1 Mb SRAM with multi-layer interconnects performance analysis results are provided in section 2.5.

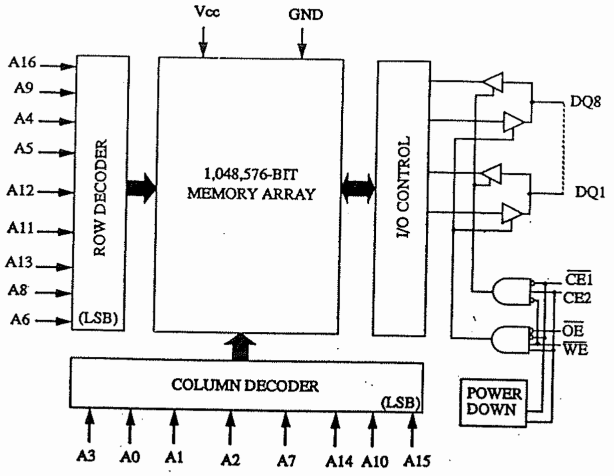


Figure 2.10: MT5C1008 1-Mb SRAM functional block diagram

Optical and Scanning Electron Microscope (SEM) instrumentation is used for multi-layer interconnect failure analysis examination. SEM having a resolution 250 A and variable magnification of 1000 to 5000 X is used to take photographs of SRAM multi-layer interconnects. The apparatus allows that the SRAM device specimen can be tilted to a viewing angle between 0 and 85, and can be rotated through 360 . This experimental method provides a means of judging the quality and acceptability of SRAM device multi-layer interconnect metallization on wafers. It addresses specific metallization defects that are batch process oriented and which can best be identified and evaluated SRAM performance utilizing this method.

Three Micron MT5C1008 SRAM devices are used for optical and SEM experimental investigations. The 1 Mb SRAM employs high-speed, low-power CMOS design using a four-transistor memory cell. The 1 Mb SRAM chip was fabricated using a double layer metal, double layer polysilicon technology. Optical microscope internal visual examination of 1 Mb SRAM shows two levels of metal tracks, are presented in Figure 2.11. As per the visual inspection, the 1 Mb SRAM chip wafer surface is very clean, and the two levels of metal tracks are fine. There are no voids observed between metal tracks of multi-layer interconnects.

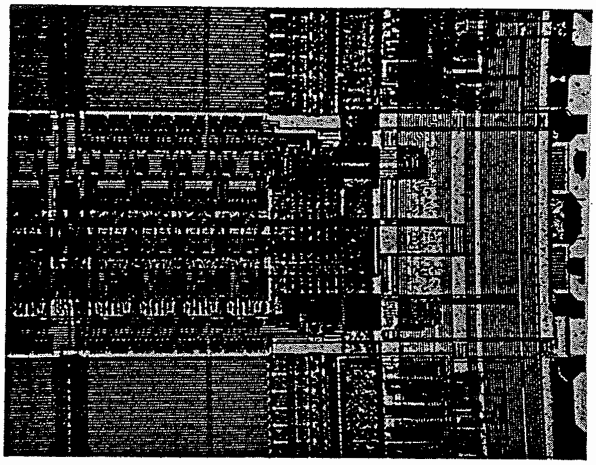


Figure 2.11: Optical microscope (100X) picture showing two levels of metal tracks in the 1-Mb SRAM (no voids observed in metal tracks).

The wafer passivation film for 1 Mb SRAM indicates that it has effective impurity barrier properties, physically hard, rugged and has perfect film integrity to protect multi-layer interconnect structure. Optical microscope internal visual examination shows 1 Mb SRAM passivation has smooth texture on the other end very well defined metal interconnect structure represented in Figure 2.12. SEM picture shows glassivation layer uniform aluminum step coverage and second level metal tracks, presented in Figure 2.13. From the SEM indicates a notching along lower edges of interconnect tracks. Notching has minor impact on 1 Mb SRAM, leading to degraded performance with time. If the current density is high in the metal lines, then 1/f noise characteristics will also affect the 1 Mb SRAM chip performance. The notching gets penetrated into interconnect metal tracks due to high current density, result in a open circuit condition in the multi-layer interconnects.

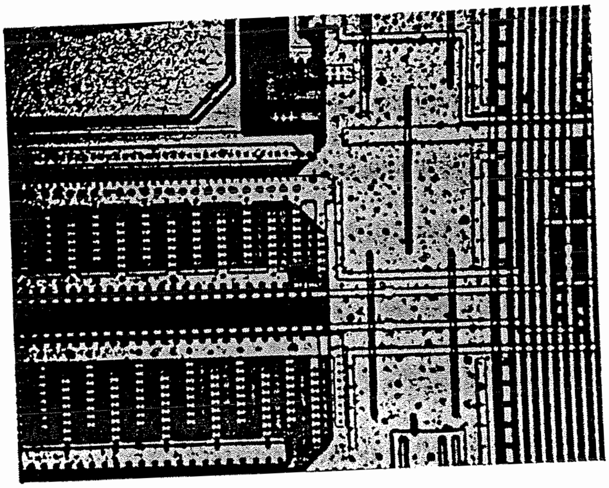


Figure 2.12: Optical microscope (400X) showing passivation texture and metal interconnect integrity within SRAM chip

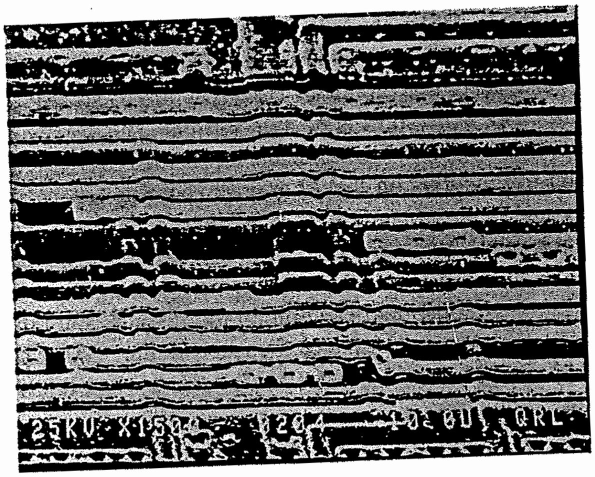


Figure 2.13: SEM showing passivation layer and second-level interconnect tracks. Notching observed along lower edges of metal tracks.

The 1 Mb SRAM multi-layer interconnects are examined for 1st and 2nd level metal tracks. Good alignment of polysilicon is shown in Figure 2.14. SEM picture was taken under DC current stress condition. The use of current signals that vary with time can lead to transients in the temperature of the metal interconnect line. These transients can be substantial for high current density stresses. As a result, they can affect the atomic flux primarily because of the temperature effect on the diffusion coefficient whose dependence is of the form (-E_a/KT). E_a refers to the activation energy of the diffusion process. K is the Boltzman's constant and T is the absolute temperature. The experimental conditions can cause electromigration in the 1 Mb SRAM chip. From the SEM picture Figure 2.14, it is clear that there is no electromigration phenomena observed in the multi-layer interconnects.

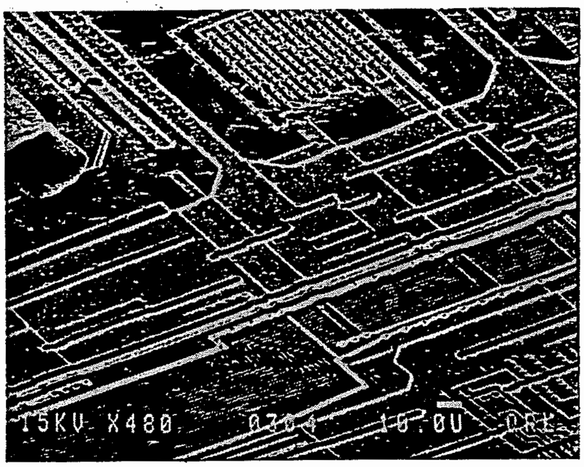


Figure 2.14: SEM showing well-defined first level metal connections, and good alignment of polysilicon visible tracks in the 1-Mb SRAM.

Optical microscope internal visual examination of interconnect alignment of 1 Mb SRAM is shown in Figure 2.15. The internal visual examination shows that there are no grain boundaries at the interconnect edges. Therefore, there are no voids or hillocks observed. SEM picture showing 1st and 2nd level metal tracks in the 1 Mb SRAM is presented in Figure 2.16. Under high current density conditions in the metallization tracks on the 1 Mb SRAM, there is no impact of electrons on the AL grains. Therefore, no voids are observed. Hence, open circuit conditions do not exist in the interconnects. Optical microscope and SEM photographs were taken for SRAM device multi-layer interconnects, to examine step coverage, notching, electromigration, voids, and hillocks, and evaluate performance impact on the 1 Mb SRAM device.

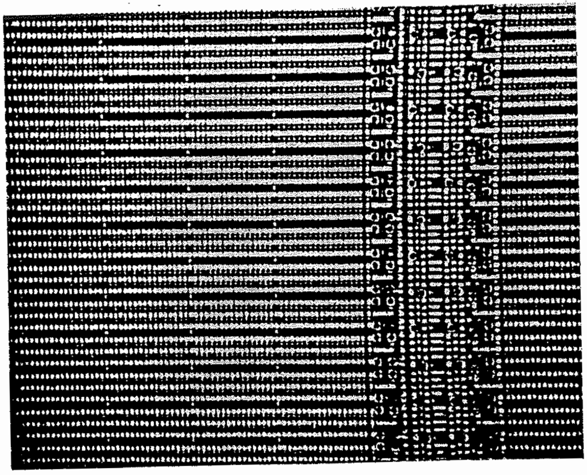


Figure 2.15: Optical microscope (400X) picture showing good multilayer interconnect alignment in the 1-Mb SRAM chip

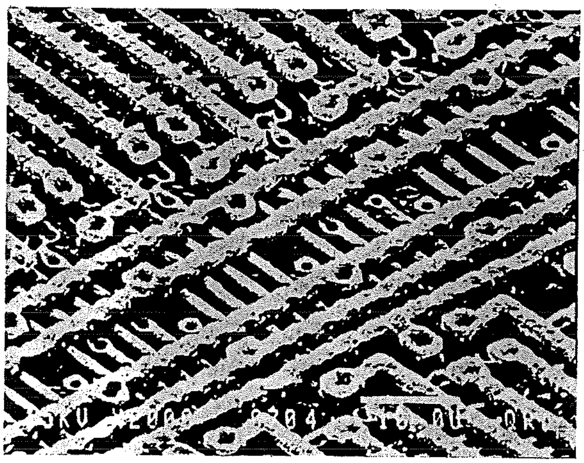


Figure 2.16: SEM showing first- and second-level metal interconnect tracks in the 1-Mb SRAM chip

Experimental observations of 1 Mb SRAM chip are as follows:

1) The 1 Mb SRAM chip interconnect length is well optimized to achieve optimum propagation delay, speed, and power dissipation performance characteristics.

2) The 1 Mb SRAM chip layout design minimizes crosstalk to reduce the level of inductance noise coming from random switching of individual memory cells and associated circuits.

3) All the interconnects are kept at minimum separation length. This reduces random fluctuations of the switching voltage at the contact.

4) The memory cell partitioning within the 1 Mb SRAM chip provides equally distributed power dissipation. This minimizes the device junction temperatures.

5) There is no electromigration problem observed. The multi-layer interconnects in the 1 Mb SRAM are very well designed, within the design limits.

6) The SRAM device multi-layer interconnect structure SEM examination shows that there are no potential interconnect failure mechanisms observed, which can lead to device catastrophic failures.

Experimental observations and data are used to support mega bit SRAM multi-layer interconnect performance analysis.

2.7 CONCLUSION

The following interconnect problems are anticipated in the mega bit SRAM chip design, by using conventional interconnect approaches.

1) Interconnections can themselves set some limits on the ultimate level of integration within SRAM chip.

2) The energy stored in the interconnects led to different lower limits on the logic delay times, depending upon whether the SRAM chip is device limited or interconnect limited in its performance.

3) The interconnect primarily sets limits on SRAM chip complexity, and performance.

In this paper a practical multi-layer interconnect capacitance model is developed and validated with the actual SRAM design. Interconnections effect on SRAM device parameters such as propagation delay, power dissipation, speed and noise parameters. A case study of 1 Mb SRAM interconnections is presented. A multi-layer interconnect approach is proposed to overcome on-chip interconnect problems in the mega bit SRAMs.

1) Multi-layer interconnects will provide higher level of integration within SRAM chip, for a given technology.

2) Multi-layer interconnects energy storage can be optimized by choosing different interconnect layers within the SRAM chip.

3) Routing and placement of SRAM cells within the SRAM chip becomes simpler by using multi-layer interconnects.

4) Mega bit SRAMs are interconnect limited designs, to overcome SRAM chip electrical performance problems, multi-layer interconnect approach is one of the best solutions to the problem.

5) High density of SRAM cells and associated circuits integration is practically feasible up to several mega bits, by using the multi-layer interconnects.

6) Multi-layer interconnects will provide high packaging density for feature SRAMs and DRAMs.

To reduce the interconnect capacitance of 1 Mb SRAM, a triple-level interconnection approach is proposed. The wiring length and chip size were reduced down to 69% and 58% of those of the double metal interconnection. Maximum t_{acc} of 30.8 ns with 1 W at 100°C and wafer yield as high as 10% is achieved. It is confirmed that the multi-level metal interconnection is very promising for realizing the high speed and high yield mega bit SRAMs.

REFERENCES

[1] K.K. CHAKRAVORTY et al. "High density interconnection using photosensitive polymide and electroplated copper conducted lines",IEEE Transactions on Components, Hybrids and Manufacturing Technologyvol. 13, March, 1990, pp. 200-206.

[2] V. N. RAYAPATI and D. MUKHEDKAR, "Interconnection problems in VLSI Random Access Memory Chip", SPIE, vol. 1389,International Conference on Advances in Interconnection and Packaging, 1990, pp. 98-109.

[3] D.E. WARD and R.W. DUTTON, "A charge-oriented model for MOS transistor capacitances,"IEEE J. Solid-state circuits, vol. SC-13, 1978, pp. 703-708.

[4] J.E. MEYER, "MOS models and circuit simulation,"RCA Rev., vol. 32, 1979, pp. 42-63.

[5] B.J. SHERE et al., "Measurement and modelling of short-channel MOS transistor gate capacitances,"IEEE J. Solid-state circuits, vol. SC-22, 1987, pp. 464-472.

[6] S.Y. OH et al., "Transient analysis of MOS transistors,"IEEE Trans. Electron Devices, vol. ED-27, 1980, pp. 1571-1578.

[7] P. BALABAN, "Calculation of the capacitance coefficients of planar conductors on a dielectric surface,"IEEE Trans. Circuit Theory, vol. CT-20, Nov. 1973, pp. 725-731.

[8] P. SILVESTER and R.L. FERRAI, "Finite element for electrical engineers," New York: Cambridge Univ. Press, 1983.

[9] P. BENEDEK, "Capacitances of a planar multiconductor configuration on a dielectric substrate by a mixed order finite-element method,"IEEE Trans. Circuit Systems, vol. CAS-23, May 1976, pp. 279-284.

CHAPTER 3

INTERCONNECT PROPAGATION DELAY MODELLING AND VALIDATION FOR 16-MEGA BIT CMOS SRAM CHIP

Article by: Venkatapathi N Rayapati and Bozena Kaminska,

Revised as per IEEE reviewers comments and sent for publication

in IEEE Transactions on Components, Hybrids, and Manufacturing

Technology.

ABSTRACT

In this paper, a closed-form expression for CMOS SRAM chip propagation delay is developed. This allows accurate calculation of the signal propagation delay of multilayer interconnects within the CMOS SRAM chip and also takes into account the delay of the CMOS SRAM cells driving the branched transmission line and the driving SRAM cell loading aspects of the interconnect line. Simulation results are presented to show the accuracy and efficiency of the propagation delay model. A case study of 16 MB CMOS SRAM chip performance evaluation is presented. The proposed closed-form delay expression results in an absolute maximum error smaller than 4.8% in comparison with the measured data. The proposed closed-form expression can be used for various high-speed, high-density multilayer interconnect SRAMs, DRAMs, FPGAs, and ASICs.

3.1 INTRODUCTION

CMOS technology focus on the very large-scale integration of high-speed SRAM cells within a single chip. To meet the expectations of SRAM speed and chip density, a number of special issues must be addressed in order to develop high-speed computer memory systems. Some important issues associated with the mega bit SRAM are : a) modelling, b) analyzing, and c) designing multilayer interconnects.

In order to achieve a correct design for mega-bit CMOS SRAMs, not only does the logical and geometrical correctness of the SRAM chip layout have to be verified, but the signal delays of the interconnect lines connecting SRAM cells, sense amplifiers, column/row decoders and buffer circuits must be considered as well to ensure SRAM chip design functionality. Typically, an aluminum layer is preferred for long interconnection lines, but the branching and crossing of interconnects force SRAM cells to use high resistance polysilicon layers. Also, the SRAM cells loading a branched line and the non-linear characteristics of the SRAM cell driving the interconnect line have to be taken into account in order to satisfy the accurate timing requirements. These signal propagation delays could be calculated by modelling the transmission lines appropriately and by applying circuit simulation techniques. But this approach has certain limitations when using semi-custom designed SRAMs. Whose timing must be quickly verified to enable an interactive improvement of cell placement and routing.

Propagation delay is a vital performance parameter for the SRAM chip design. Propagation delays introduced by multilayer interconnects within the SRAM chip are modeled using transmission line concepts. The transmission line is usually substituted by an RC line or an RC tree in the case of a branched line. Sakurai [1] showed that transmission-line modelling has many limitations. Depending on the SRAM cell's driving and load conditions, suitable T or Π ladder circuit models have to be chosen. The widely-used step L ladder circuit was found to be a very poor approximation, and better models with three to five step T or Π ladder circuit models have been chosen. Π ladder circuits require considerable computational effort. Another method, proposed by Penfield and Robinstein et al.[2], is the calculation of upper and lower bounds for the signal wave forms at the load terminals of an RC tree network driven by an independent voltage source. Application of this approach has shown that the bounds can differ by more than 30% from the exact step response, so that it is scarcely possible to deduce a measure for a multilayer interconnect delay that is sufficiently close to the exact results to be used in an interactive semicustom SRAM design.

In this paper, a closed-form expression for SRAM chip propagation delays due to multilayer interconnects is developed. The model is based on a delay approximation for an SRAM cell capacitively loaded with multilayer interconnects. Numerical results show that the linear delay estimation is valid for the entire range of SRAM chip capacitive loading. Furthermore, this approach is extended to branched transmission lines. The nonlinear nature of the driving gate is taken into account by applying a first-order transistor model for the SRAM cell. This results in a closed-form solution formula for the SRAM chip delay estimation, which allows accurate prediction using two to three orders of magnitude less computer time than SPICE simulations. A case study of 16-MB CMOS SRAM chip propagation delay is investigated.

3.2 DEFINITION OF PROPAGATION DELAY FOR THE SRAM CHIP

Most investigations of SRAM chip signal propagation delays use either Elmore's [3] delay definition or the definition of rise and fall time, i.e. the time a signal needs to change between 10% and 90% of the maximum signal value. Elmore [3] has defined a delay T_D as the first moment of the impulse response. This measure may be efficient for theoretical studies, but it does not provide a good prediction of the timing behavior of interconnected MOS SRAM cells. Another delay definition is the time required for a response to reach the threshold voltage of a MOS transistor used in SRAM cell design. This kind of definition causes some problems due to threshold voltage shifts in NMOS circuits with a back-gate bias. In most applications, acceptable delay definitions can be found by simulating typical driver-gate/line receiver-gate combinations. Using this approach, Carter and Guise [4] defined the propagation delay time the 62 % criterion of the average of pulling up and pulling down.

In order to over come the problems associated with the above mentioned delay time definitions, a practical delay time concept has been proposed in this paper. The objective of this research is to propose a delay time model, which represents the SRAM device operation in a practical manner and gives reasonable results. A closed-form expression for the SRAM cell delay time has been developed and validated with experimental studies.

The analysis of a SRAM cell [5] shows that the propagation delay is the sum of the intrinsic delay associated with the device switching and the two transition times t_1 and t_2, involving the device access network and the loading elements. A simplified charge-control equivalent circuit concept is used to give a reasonable insight into the performance of a SRAM cell matrix.

The transition times t_1 and t_2 can be written as follows:

$$t_1 = (F_O\,(C\pi + 2\,C_U) + C_O + F_O\,.\,C_W)\,(V_{OH} - V_{TH})\,|\,I_{max}\text{-pull-down}$$

$$t_2 = (F_O\,(C\pi + 2\,C_U) + C_O + F_O\,.\,C_W)\,(V_{TL} - V_{OL})\,|\,I_{max}\text{-pull-up}$$

I_{max} pull-down is the maximum current delivered by the pull-down device, I_{max} pull-up device, F_O is the fan-out number, C_W is the interconnect capacitance per fan-out and C_O is the device output capacitance. The propagation delay of loaded Complementary Metal Oxide Semiconductor (CMOS) SRAM cell and can be derived from the above expressions. Assuming the logic threshold voltage is at half swing propagation delay of a SRAM cell and is expressed as:

$$t_{pd} = t_T \quad + \quad \frac{C_T.\ \triangle V}{2I_d\,(\triangle V/2)}$$

$C_T = (C_g + C_W) F_O + C_O$ means the total load capacitance and I_d is the current across the inverter biased at the logic threshold. An extra interconnection delay must be added to the total circuit delay, the propagation delay is given as

$$t_{pd}=t_T+\frac{1.7}{U_o}[F_o.L^2+(\frac{F_o.C_w+C_o}{C_{ox}})\frac{L}{W}]$$

The L/W term is significantly influenced for small inverter load with long interconnecting lines in the SRAM chip. Then t_{pd} is highly dependent on C_{ox} and varies quite linearly with gate oxide thickness.

The parasitic capacitance C_P effect is considered between the interconnections of SRAM cells and other functional block of the SRAM chip. Therefore, the coupling capacitance is expressed as follows:

$$C_C = C_{OX} + C_P$$

Where C_C is the coupling capacitance, C_{OX} is the oxide parallel-plate capacitance and C_P is the parasitic capacitance. The total capacitance of the SRAM chip is expressed as:

$$C = C_1 + C_C$$

Where C_1 is the SRAM cell capacitance. A practical concept is translated into the mathematical domain and solved for the total capacitance C associated with the SRAM chip, in order to evaluate the chip propagation delay time.

Based on the interconnected CMOS SRAM cell simulations, a delay time concept has been proposed [6]. A delay time definition corresponds to a change in the output signal of 0.7 V_{DD} is presented. The SRAM cell logic levels can be made close to V_{DD} and ground. The SRAM cell logic swing can result in the same order of V_{DD} results. If the V_{DD} is lower than 0.7 value, then the SRAM cell demonstrates a hysteresis transfer characteristic and the SRAM cell can't be able to write or read. A delay time definition corresponding to a change in the output signal of 0.7 V_{DD} represents the practical device operating characteristics and has been shown to give reasonable results. In the case of different pull-up and pull-down conditions, the worst case delay time is considered. Based on this delay time concept, a delay time formula derived in this paper. The proposed delay time formula reduces significant computational effort and results in quite accurate delay time predictions for complex SRAM designs. If necessary, the delay time definition can be further improved, and only modification results in a change of the weighting factors in the delay time formula derived in this paper.

3.3 TRANSMISSION LINE MODEL FOR SRAM CELL MULTILAYER INTERCONNECTS

The split-word-line memory cell layout is shown in Figure 3.1. The split-word-line cell distributes the word systematically around the active area and first polysilicon layers. The split-word-line cell architecture keeps the effective transistor width stable and avoids the imbalance inherent in a multilayer interconnect layout. The symmetrical-cell layout in the SRAM chip allows low-voltage operation and provides immunity from SRAM cell operation noise. The SRAM cell cross section is shown in Figure 3.2. The cell uses 0.5μm CMOS technology with quintuple polysilicon and double metal layers. The first polysilicon layer is used for the gate electrode. The second and fourth layers are used for the PMOS TFT double gates. The third layer is used for the TFT channel and Vcc line. The fifth layer is used for the Vss line. Four polysilicon layers are connected in a contact hole through a polysilicon sidewall contact.

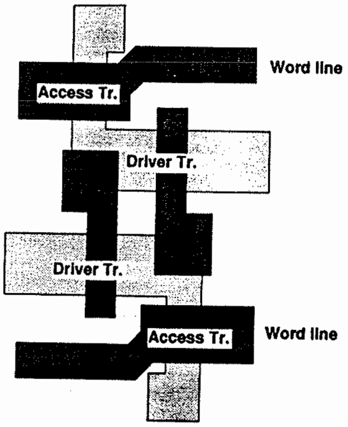


Figure 3.1: Split-word-line SRAM cell layout

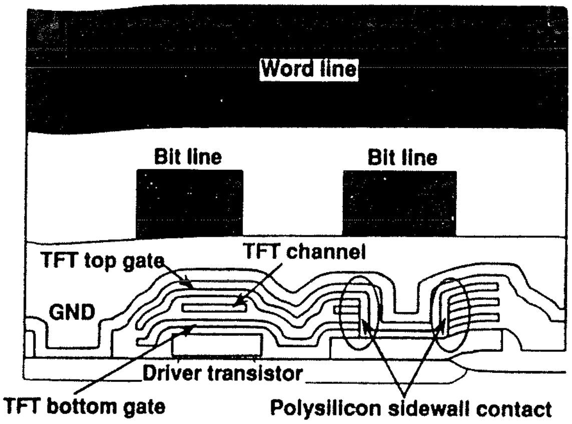


Figure 3.2: CMOS SRAM Cell Cross Section

A multilayer interconnect in the SRAM cell is modeled as a single memory-cell-driven transmission line by a step-voltage source with resistance R_T, a distributed R_C line with total resistance R and capacitance C (i.e., Total capacitance $C= C_1 + C_C$, where C_1 is the SRAM cell capacitance, and C_C is the coupling capacitance associated the interconnected SRAM cells) and a load capacitance C_T as shown in Figure 3.3. In order to simplify computational complexity the total capacitance C takes into account SRAM cell capacitance and coupling capacitance. For all practical purposes the fringing capacitance and parasitic capacitance are assumed to be the same order of magnitude. The fringing capacitance of an isolated interconnect layer is of the same order of magnitude as the oxide capacitance associated with the SRAM cell. The total capacitance of a line is affected by the proximity of another interconnect. The second layer has less significant impact on the first. When the thickness of both dielectric and metals are scaled as the square root of λ, the total capacitance per unit length of the first two layers decreases significantly. Where as the third or fourth interconnect layers capacitance remains constant.

A practical concept is translated into the mathematical domain and solved for delay time associated with the SRAM cell. Multilayer interconnects in the SRAM are considered to be distributed RC line. The SRAM cell capacitance and resistance distribution model is shown in Figure 3.3.

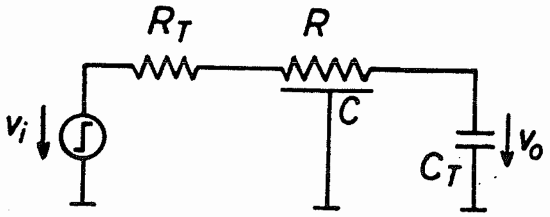


Figure 3.3: SRAM cell capacitance and resistance distribution model

The distributed RC line output step response in the Lapace domain can be written as:

$$V_o(s')/V_{DD} = H(s')/s' \tag{1}$$

where S' is the complex frequency.

The voltage transfer function for the RC distributed line can be expressed using [7] as:

$$H(s') = [\cos(jg)(1+s'\gamma_T C_T) - j\sin(jg)(s')^{\frac{1}{2}}(\gamma_T + C_T)]^{-1} \tag{2}$$

$$\text{where} \qquad g = (s')^{\frac{1}{2}} \tag{3}$$

is the normalized transfer constant and

$$\begin{aligned} \gamma_T &= R_T/R \\ C_T &= C_T/C \\ s' &= sRC = sT \\ t' &= t/T \end{aligned} \tag{4}$$

are the normalized driver resistance, load capacitance, frequency and time.

Applying Heaviside's second expansion theorem from [8] and using [1], the time response can be written as

$$V_o(t')/V_{DD} = 1 + \sum_{n=1}^{\infty} K_n e^{-\sigma_n t} \tag{5}$$

where

$$K_n = \frac{(-1)^n 2[(1+\gamma_T^2\sigma_T^2)(1+C_T^2\sigma_T^2)]^{\frac{1}{2}}}{(\sigma_n)^{\frac{1}{2}}[(1+\gamma_T^2\sigma_n^2)(1+C_T^2\sigma_n^2)+(\gamma_T+C_T)(1+\gamma_T C_T\sigma_n)]} \tag{6}$$

and the poles of (1): 0, σ_1, σ_2, ..., σ_n can be determined from

$$\tan[(\sigma_n)^{\frac{1}{2}}] = \frac{1-\sigma_n C_T \gamma_T}{(\gamma_T + C_T)\sigma_n^{\frac{1}{2}}}, \quad n - \frac{3}{2} < \frac{\sigma_n}{\pi} < n - \frac{1}{2} \tag{7}$$

Sakurai [1] has solved equation (7) numerically. From these numerical results, the higher-order terms of equation (5) can be neglected to get an excellent approximation for the step response in the rangeγ_T and $C_T < 1$:

Introducing the SRAM cell delay time concept $t_{0.7} = t(V_o = 0.7\ V_{DD})$ as stated

$$V_o(t')/V_{DD} = 1 + K_1 e^{-\sigma_1 t} \tag{8}$$

above, calculation of $t_{0.7}$ from equation (8) as a function ofγ_T or C_T shows that it depends nearly linearly on both γ_T and C_T in the γ_T, $C_T \leq 1$ range.

Consequently, delay can be written as

$$t'_{0.7} \approx \alpha_1 + \alpha_2 C_T + \alpha_3 \gamma_T + \alpha_4 \gamma_T C_T \tag{9}$$

where α_3 is equal to α_2 because of equation (8) which is symmetrical inγ_T and C_T. Taking into account that the influence of the driver SRAM cell is considered separately, equation (9) reduces to

$$t'_{0.7} = \alpha_1 + \alpha_2 C_T \tag{10}$$

weighting factor α_1 can be determined from equation (9) setting $C_T = 0$ to $\alpha_1 = 0.59$, and factor α_2 depends on C_T and increases by 7% between $C_T = 0.1$ and $C_T = 1$. To simplify the analysis, the maximum value$\alpha_2 = 1.21$ is taken, resulting in delay time error of $t'_{0.7}$ which is smaller than 1.5% in the $0 \leq C_T \leq 1$ range. Where $t'_{0.7}$ is the delay time error contribution resulting from the analysis. Thus we get the approximation

$$t_{0.7} = 0.59\,RC + 1.21\,RC_T \tag{11}$$

for the signal delay for a single interconnect line.

The multilayer interconnect is modeled as a transmission line with a tree structure. The branched line is now decomposed into single sections with no side branches within the SRAM cell. Each section is placed either with a polysilicon layer or an aluminum layer. Interconnections laid out in aluminum are modeled as capacitances. They can be considered as node capacitances at the nodes between sectors in poly-layers. For each of these interconnections, delay time can be calculated using equation (11).

However, for each single interconnect, the load capacitance C_T now consists of all capacitances between the output node of the actual interconnect and all succeeding nodes loading this interconnect. This follows from applying the theorem of Lin and Mead [8], and Corollay [10]. For the SRAM multilayer interconnect, the $C_T > 1$ range, where $| K_1 |$, decreases rapidly, as shown in Figure 3.4. Therefore, $| K_1 | \exp(-\sigma_1 t'_{0.7})$ deviates from the value 0.3, and so equation (8) is no longer a valid approximation of equation (5). On the other hand, the numerical results shown in Figure 3.5, which were obtained by solving the partial differential equations, show a linear dependence of $t_{0.7}$ from the C_T. Thus, the above approximation can be in the $C_T > 1$ range, but K_1 can no longer be understood as the variable defined in equation (6). It is interpreted as the constant value $\overline{K_1} = K_1 (C_T = 1) = -1.14$, which leads to the new approximation expressed as follows:

$$V_o(t')/V_{DD} = 1 - 1.14 e^{-\sigma_1 (C_T) t'} \qquad (12)$$

From equation (12) it is understood that α_2 is constant. Thus equation (11) is validated.

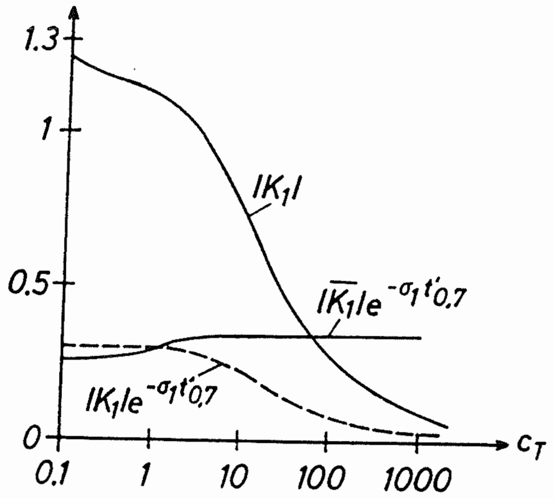


Figure 3.4: Effect of Poles on delay approximation

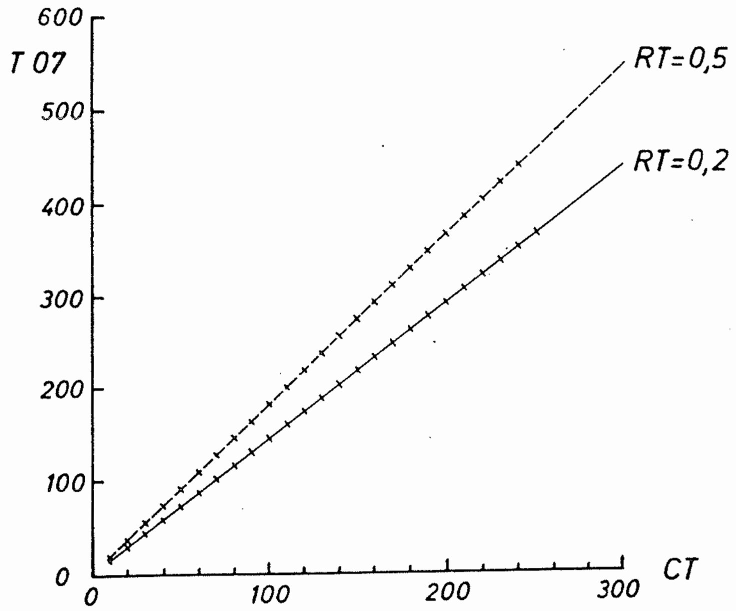


Figure 3.5: Capacitive load effect on SRAM cell delay numerical results

The resulting estimation for the delay $t'_{0.7}$ can be seen in Figure 3.4., showing that $|\overline{K_1}|$ exp (- σ_1 $t'_{0.7}$) is slightly smaller than 0.3 for $C_T > 1$ and increases up to 0.32 for the $C_T >> 1$.. Consequently, equation (12) can be expected to give reasonable results for all possible values of C_T.

Furthermore, equation (12) can be interpreted as the step response of a suitable RC circuit with an initially charged capacitor. This allows to apply the theorem of Lin and Mead [9], which states that the total delay of a path in a tree network is given by the sum of the delays of the single portion of the unique path between the input and the actual output. Therefore, the total delay of a path with K sections between the input node of the branched transmission line and an output node can be written as

$$t_{0.7} = \sum_{v=1}^{k} R_v (0.59 C_v + 1.21 \sum_{\mu} C_\mu)$$

where C_υ is the capacitance of transmission line interconnectυ and C_μ denotes the capacitances between the output node of sectionυ and all succeeding branches loading the interconnect.

3.4 SRAM CELL LOAD DRIVER PROPAGATION DELAY MODEL

The load of each SRAM cell branch of the transmission line is effected by the input capacitance of the driven SRAM cell. Since the non-linearity of the MOS SRAM cell capacitance depends on the SRAM cell voltage, the load can be modeled with sufficient accuracy by a constant capacitance

$$C_o = (W_p L_p + W_n L_n) C'_{ox} \tag{14}$$

where W, L are the width and length of SRAM cell load and driver transistor of the driven CMOS SRAM cell. C'_{ox} denotes the oxide capacitance per unit area. Load capacitance C_o can be included in equation (13), thus it is taken as term $C\mu$.

To model the propagation delay of the driving SRAM cell, the SRAM cell is interpreted as an additional RC circuit that represents the first portion of the transmission network. This implies that the effective load capacitance C_L need to be calculated from the SRAM cell delay includes not only the SRAM cell output capacitance C_D but also the total capacitance C_{tot}, which is much larger than C_D. The nonlinear portion of the resulting capacitance can be ignored without significant capacitance errors.

Hence

$$C_L = C_{tot} + A_{DP} C_{AP} + P_{DP} C_{PP} + A_{Dn} C_{An} + P_{Dn} C_{Pn} \tag{15}$$

where A_D and P_D denote the drain area and length of drain junction parameter, C_A / C_p are the drain junction capacitance per unit area/perimeter of the PMOS and NMOS transistors.

To calculate the propagation delay of the SRAM cell driver, a first order model [10] is used for the drain-source i_{DS} in the triode region and at saturation:

$$i_{DS} = \begin{cases} \pm\beta\,[(V_{GS} - V_T)\,V_{DS} - \dfrac{V_{DS}^2}{2}], & \text{triode region} \\ \pm\beta\,\dfrac{(V_{GS} - V_T^2)}{2}, & \text{saturation} \end{cases} \tag{16}$$

where the positive sign is valid for NMOS, and the negative sign for PMOS, with μ denoting the carrier mobilities

$$\beta = \mu\, C'_{ox} W/L \tag{17}$$

which gives different values for the NMOS and PMOS transistors.

A memory cell layout illustrating the driver transistor is shown in Figure 3.1. Driver-equivalent circuit models are shown in Figure 3.6. which represents the pull-up phase of a CMOS inverter with input $V_G = 0$, we get

$$C_L \frac{dV_o}{dt} = \frac{\beta_p (V_{DD} + V_{TP})^2}{2} \qquad (18)$$

in the saturation region, and

$$C_L \frac{dV_o}{dt} = \beta_p [(V_{DD} + V_{TP}) (V_{DD} - V_o) - (V_o - \frac{V_{DD}{}^2}{2}] \qquad (19)$$

in the triode region. Integrating equation (17) from $V_o = 0$ to $V_o = -V_{TP}$ (limit of saturation region) and equation (18) from $V_o = -V_{TP}$ to $V_o = 0.7\ V_{DD}$ (triode region) yields the driver propagation delay time

$$T_{IP} = \frac{C_L}{\beta_p} \left[\frac{|V_{TP}|}{(V_{DD} - |V_{TP}|^2)} + \frac{I_n \left(\frac{17}{3} - \frac{20}{3} |V_{TP}| / V_{DD} \right)}{2 \left(V_{DD} - |V_{TP}| \right)} \right] \qquad (20)$$

Generally, a SRAM cell-driver circuit is designed to generate a symmetric signal with $T_{IP} = T_{In}$. If not, the result is a worst case estimation. Thus the total delay of the driver/branched transmission line and SRAM cell-load combination is given by

$$T_{0.7} = t_{0.7} + \max\{T_{IP}, T_{In}\} \qquad (21)$$

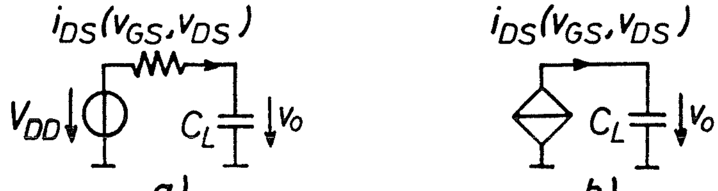


Figure 3.6: SRAM cell driver models for (a) Pull-up and (b) Pull-down

3.5 EXPERIMENTAL VALIDATION OF THE PROPAGATIONDELAY MODEL FOR A 16-MB CMOS SRAM CHIP

A. Chip Architecture for a 16-MB CMOS SRAM chip

The 16-MB CMOS SRAM chip functional block diagram is shown in Figure 3.7. The memory cell array is divided into four blocks with two global row decoders. Each block is divided into eight sections, and each section has 16 internal I/O bus pairs. The 16 I/O bus pairs are divided into two groups. Each group of data bus lines runs along the side of the memory cell arrays and is optimized by placing the I/O terminals on both sides of the array.

In high-density SRAM, the distance between the external voltage terminals and the output buffers gives rise to a parasitic impedance that causes noisy operation. This effect is distance-dependent and can be eliminated by choosing pin assignments that place the external voltage and output terminals together in the centre of the upper and lower sides of the cell array. The selected data-bus layout divides the bus into two groups to fit the new power supply and output terminal pairs. The divided data-bus layout is also facilitated by means of using interdigitated bit lines.

The interdigitated bit-line architecture for the 16-MB CMOS SRAM is shown in Figure 3.8. Four pairs of bit lines form one group. The first and second pairs are connected to the lower data bus lines and the third and fourth to the upper data bus lines. Each two pairs of bit lines are connected to the data bus on either side of the memory cell array. The I/O terminal for each data bus line is placed on the same side as the connection.

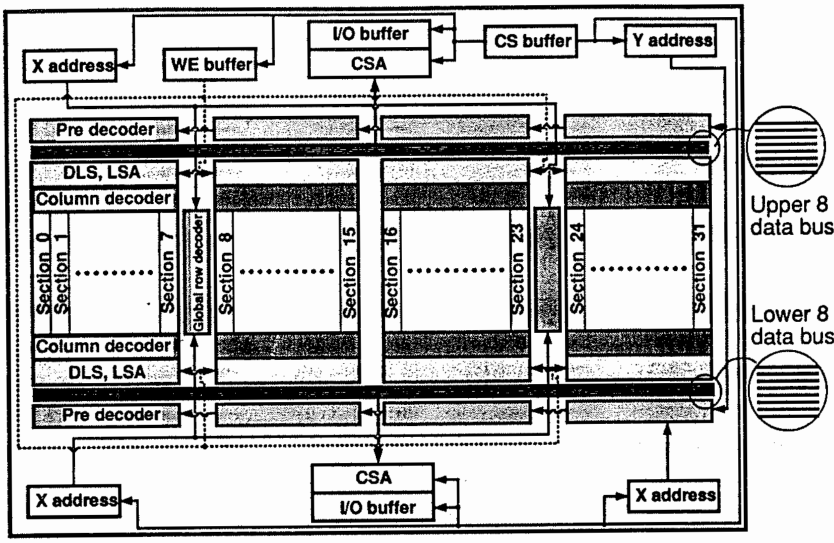


Figure 3.7: 16-Mb CMOS SRAM functional block diagram

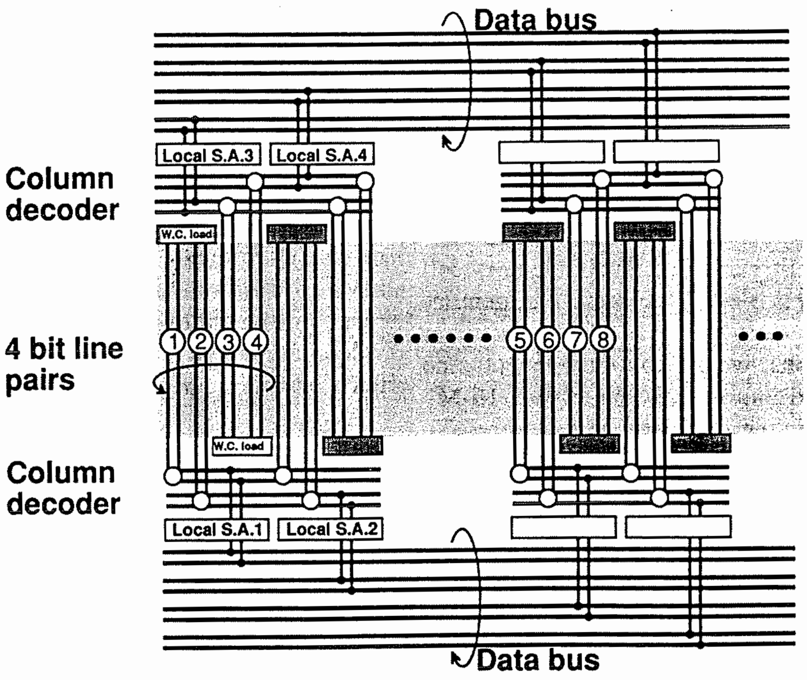


Figure 3.8: Interdigitated Bit Line Scheme and Divided Data Bus

The first and second pairs of bit lines have a column access gate on the lower side, while the third and fourth pairs have one on the upper side, which reduces the column access gate spacing by half. This leaves room for write-control load circuits on the opposite side to the column access gates. These write-control load circuits are intended to improve the read-access time after the cycle. In high-density SRAM, the bit line is too long to be charged up quickly by only one end-load- after-write cycle. The proposed write-control load is shown in Figure 3.9, and enables elimination of the write current while allowing a fast bit-line pullup because of the load on both ends of the bit line.

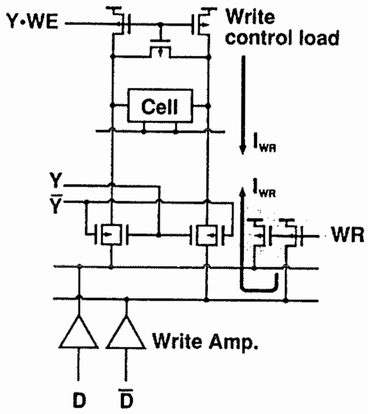


Figure 3.9: Write control load circuit for the 16-Mb SRAM

The hierarchical sense amplifier [11] shown in Figure 3.10, has been used to reduce the data bus voltage amplitude. The data bus is connected to the outputs of 32 local sense amplifiers and to the input of a Current Sense Amplifier (CSA). The local sense amplifier consists of a Data Level Shifter (DLS), and a latched cascaded sense amplifier (LSA). This structure enables a reduced parasitic capacitance from a column gate to the latched cascaded sense amplifier. The result is a short delay time in the latched cascaded sense amplifier with a fast reduced amplitude data bus.

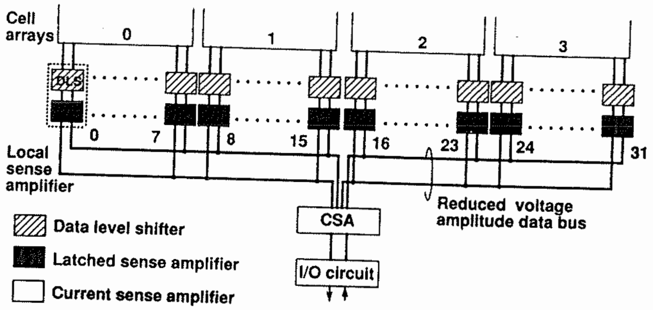


Figure 3.10: Hierarchical Sense Amplifier for the 16-Mb SRAM

The CMOS SRAM cell RC delay as a function of interconnect length and width simulations have been performed to optimize performance and interconnect density for the 16-Mb SRAM chip. Simulation results for the SRAM cell RC delay as a function of interconnect length and width are presented in Figure 3.11 (a) and 3. 11 (b) respectively.

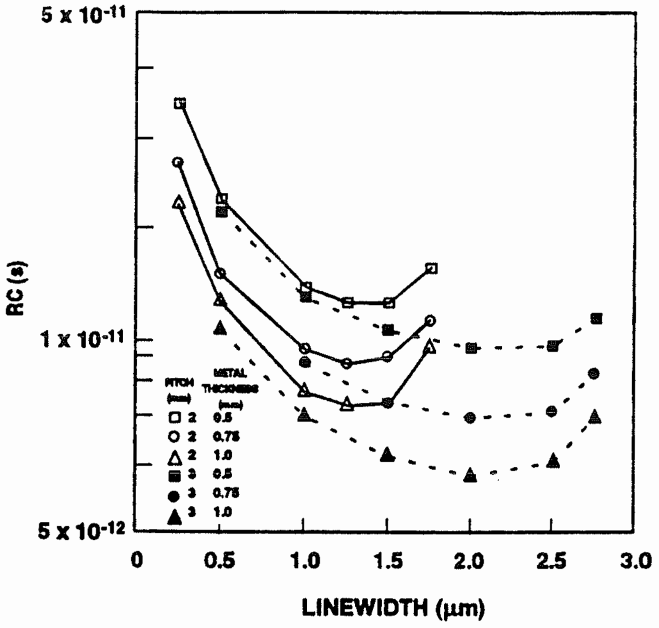


Figure 3.11(a): SRAM cell RC delay as a function of interconnect line width

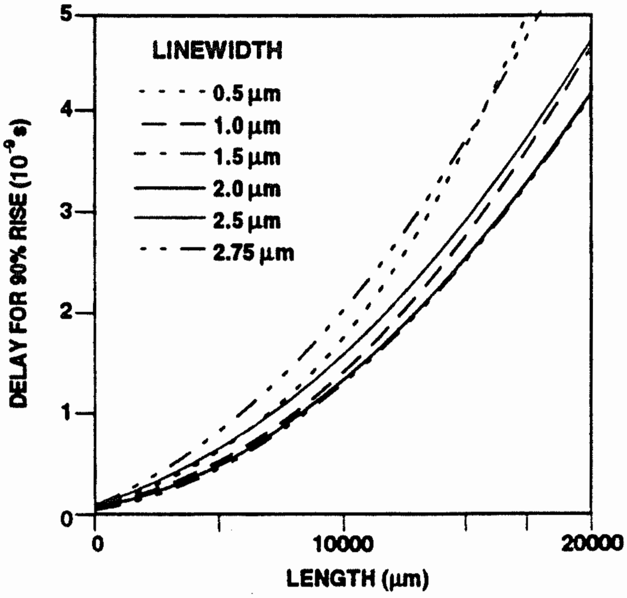


Figure 3.11(b): SRAM cell delay as a function of interconnect length simulation

B. Theoretical propagation delay computation for a 16-MB SRAM chip using the proposed model

The 16-MB SRAM chip shown in Figure 3.7, is used for validating the propagation delay model presented in this paper. The SRAM chip consists of the following circuits: address buffer, decoders, memory cell arrays, sense amplifiers, write/read-control circuitry, data buses and output buffers. In this SRAM chip, each polysilicon section of a line is modeled by a five-step $\prod$ ladder circuit. The theoretical calculation portions of the multilayer interconnect line laid out with aluminum layers were modeled as a capacitance. The equivalent model of the 16-Mb SRAM chip using the branched transmission line concept is shown in Figure 3.12. The SRAM chip consists of 11 sections of polysilicon layers and 8 sections of aluminum layers. Each section is again subdivided into corresponding interconnect structures within the device. In spite of the wide range in capacitance C_T, from 3.6 fF up to a value of 178.4 fF for the sections between nodes 3 and 5, the total error is smaller than 2% of the actual delay. The calculated delay between node 0 and the load is shown in Table 3.1. As can be expected from the approximation, the delay estimation has a negative error for short interconnect lines, which includes the fact that the C_T is in the lower range. For long transmission lines, the average C_T exceeds 1, which results in a positive error. MATLAB is used to compute these propagation delay calculations for the 16 MB SRAM chip. Other examples using the same model at the chip lowest circuit level has shown that the absolute delay time error for the 16-MB SRAM chip is smaller than 4.8%. Theoretically computed total propagation delay for the 16-MB SRAM chip is 15.72 ns.

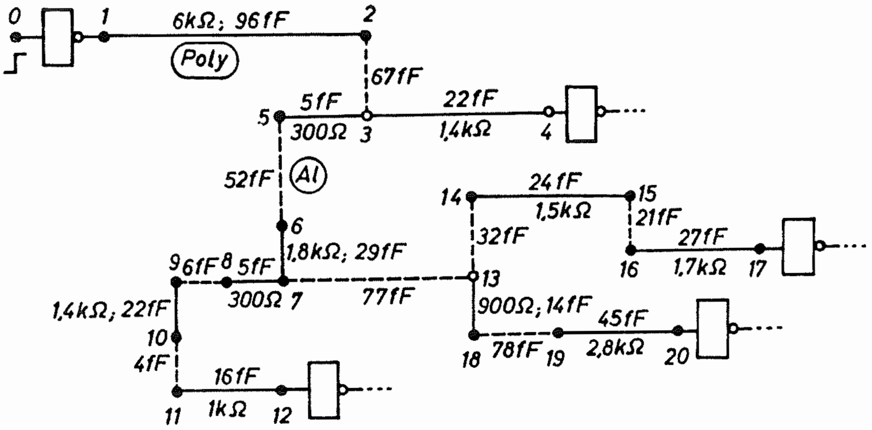


Figure 3.12: 16-Mb SRAM chip equivalent transmission line model for multi-layer interconnects

Table 3.1: Calculated propagation delay results

Node	Theoretically computed propagation delay	SPICE - Simulated propagation delay	Error
4	10.31 ns	10.50 ns	- 1.96%
12	12.94 ns	12.94 ns	0%
17	13.75 ns	13.73 ns	+ 0.17%
20	15.35 ns	15.30 ns	- 0.41%

C. SPICE simulations for the 16-MB CMOS SRAM chip propagation delays

In order to evaluate the overall propagation delay of the 16-MB SRAM chip, SPICE simulations were carried out using 0.5 µm CMOS technology. The CMOS Twin-well process-related parameters are presented in Table 3.2. The chip design is based on 0.5 µm rule and CMOS with quintuple polysilicon double metal layers. The n-channel transistor gate is 0.5µm long, the p-channel gate is 0.5 µm long, and the gate oxide is 11 nm thick. The TFT gate length, width and thickness are 0.5 µm, 0.3 µm, and 25 nm respectively. The 16-MB SRAM chip has been partitioned into subcircuits: SRAM cell matrix sense amplifiers, column/row decoders and buffer circuits. Each subcircuit simulation was performed separately. The summation of all subcircuit propagation delays results in an overall chip propagation delay. The total propagation delay of the 16-MB SRAM chip is about 15.3 ns.

Table 3.2: 16-Mb CMOS SRAM process parameters used for simulation

PROCESS TECHNOLOGY	0.5 μm design rule CMOS 5 polysilicon layers 2 Metal layers
Gate length	0.4 μm (NMOS) 0.5 μm (PMOS)
Gate oxide thickness	10 nm
TFT gate length / width	0.5 μm/0.3 μm
TFT gate oxide thickness	23 nm

D. Propagation delay measurements for the 16-MB SRAM chip

In order to evaluate the propagation delay for the 16 MB SRAM chip, practical measurements were taken using the simple test setup providing the input address to the chip and observed data output through the logic analyzer. Address access time is about 15 ns, when the reference voltage level is 1.5 V and the output is terminated by two resistor loads and a 30 pF capacitor.

The 16-MB SRAM used interdigital bit-line architecture, a reduced voltage-amplitude data bus, a latched sense amplifier, and a current sense amplifier, and provides a propagation delay for the overall chip of about 15 ns. The 16-Mb SRAM chip waveforms of address input and data output is shown in Figure 3.13.

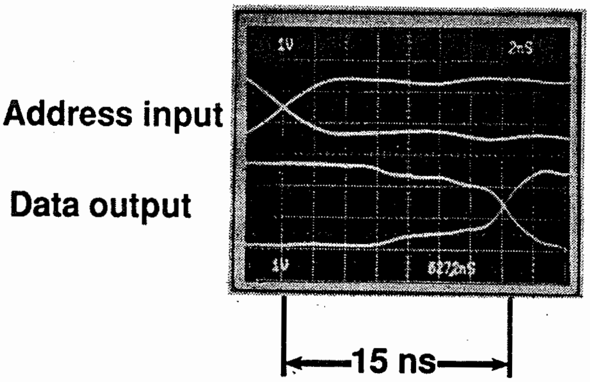


Figure 3.13: 16-Mb SRAM chip waveforms of address input and data output

3.6 CONCLUSION

A closed-form expression for the CMOS SRAM chip propagation delay has been developed. Numerical results shows that the linear delay estimation is valid for the entire range of SRAM chip capacitive loading. Furthermore, all the multilayer interconnections were modeled as transmission lines with a branched-tree architecture. The closed-form propagation delay expression results in an absolute maximum error smaller than 4.8% in comparison with the measured data. Theoretically computed propagation delay values for the SRAM chip results in an absolute maximum error smaller than 2% in comparison with the SPICE simulation results. The proposed closed-form expression for signal delay estimation provides accurate predictions for high-density, high-speed SRAMs. The proposed closed-form expression can be used for various high-speed, high-density multilayer integrated circuits, SRAMs, DRAMs, FPGAs and ASICs.

The closed-form expression reduces significant computational effort and results in quite accurate delay time predictions for complex SRAM designs.

ACKNOWLEDGEMENT

The authors thank K.Kamazawa from Fujitsu Laboratories, for his technical contribution and suggestions.

REFERENCES

[1] T. SAKURAI., "Approximation of wiring delay in MOSFET LSI",IEEE J.

Solid State Circuits, Vol.SC-18, no.4, pp. 418-426, Aug 1983.

[2] J. RUBINSTEIN, P. Penfiled, and M.Horowitz., "Signal delay in RC tree networks", IEEE Trans. Computer Aided Design, Vol.CAD-2, No.3, pp. 202-211, July 1983.

[3] W.C. ELMORE, "The transient response of damped linear networks with particular emphasis on wideband amplifiers",J. Applied Physics, Vol.19, no. 1, pp. 55-63, Jan 1987.

[4] D.L. CARTER, and D.F. GUIS., "Effects of interconnections on submicron chip performance", VLSI Design, pp. 63-68, Jan 1984.

[5] V.N. RAYAPATI, and B. KAMINSKA., "Performance Analysis of Multi-layer Interconnections for Mega bit Static Random Access Memory Chip",IEEE Transactions on Components, hybrids, and Manufacturing Technology, Vol.16, No.5, pp. 469-477, August 1993.

[6] V.N. RAYAPATI, and B. KAMINSKA., "Mega Bit BiCMOS SRAM chip Package Modelling and Performance Analysis", Proceedings of IEEE International Workshop on Memory Technology, Design and Testing, pp. 10-15, August 1994.

[7] K.SINGHAL and J.VLACH., "Approximation of non-uniform RC distributed networks for frequency and time domian computations",IEEE Trans. on Circuit Theory, Vol.CT-19, pp. 347-354, July 1972.

[8] B. VAN DER POL, and H. BREMMER., Operational calculus based on the two-sided Lapalace integral, chapter 7,Cambridge University Press.

[9] T.M. LIN and C.A. MEAD., Signal delay in general RC networks,IEEE Trans. CAD, Vol. 3, no. 4, pp. 331-349, Oct 1984.

[10] M.I.ELMASTRY., "Digital MOS Integrated Circuits: A Tutorial"Digital VLSI Systems edited by M.I Elmastry, pp. 10-37, 1985.

[11] K.SASAKI et al., A 7-ns 140-mW 1-Mb CMOS SRAM with current sense amplifier, IEEE J. Solid State Circuits, Vol. 27, no. 11, pp. 1511-1518, November 1992.

CHAPTER 4

HIGH SPEED CIRCUIT DESIGN TECHNIQUES FOR MEGA BIT BiCMOS STATIC RANDOM ACCESS MEMORIES

Article by: Venkatapathi N Rayapati and Bozen Kaminska, Sent for publication in International Journal of Microelectronics, Elsevier Advanced Technology publishers, Oxford, England.

ABSTRACT

In order to improve mega bit BiCMOS SRAMs performance, high speed circuit design techniques need to be investigated. The objective of this paper is to present high speed circuit design techniques for mega bit BiCMOS SRAMs and performance evaluation. Address transition detection, divided word line, dynamic double word line, pulsed word line, dynamic bit-line loads, and hierarchical decoding circuit design techniques for BiCMOS SRAMs are discussed. A 1-Mb BiCMOS SRAM chip architecture is proposed based on these high speed circuit design techniques. Electrical performance of 1-Mb BiCMOS SRAM has been evaluated. Performance analysis results indicate that speed has been improved by 21% using high speed circuit design techniques, in comparison with the conventional BiCMOS SRAM designs. Power reduced by about 10 %. High speed circuit design techniques are very useful for future mega bit BiCMOS SRAMs and DRAMs.

4.1 INTRODUCTION

High performance computer systems require large scale, high speed Static Random Access Memories (SRAMs). In recent years, sub-micrometer BiCMOS technologies have helped SRAMs to improve memory density and performance [1-4]. Because BiCMOS SRAMs contain both bipolar devices and CMOS devices, they are easily adjusted to both ECL and TTL interfaces. High speed ECL I/O SRAMs are indispensable in main memories of mainframe computers and super computers, and TTL I/O SRAMs operating with microprocessors at low supply voltage are required for high performance work stations. In order to meet both of these performance requirements, there is a definite need to investigate high speed circuit design techniques for BiCMOS SRAMs.

The objective of this research work is to investigate high speed circuit design techniques for BiCMOS SRAMs. In order to improve speed, density, and reduce power consumption of BiCMOS SRAMs the following design issues have been considered:

1) How to reduce signal delay time of the word line
2) How to improve sense amplifier speed
3) How to reduce noise caused by the output circuit
4) How to provide signal level conversion
5) How to reduce power consumption

In order to address the above mentioned design problems, there is a definite need to investigate high speed circuit design techniques. The following high speed circuit design techniques have been analyzed for BiCMOS SRAMs:

1) Address Transition Detection (ATD) Method
2) Divided Word Line (DWL) Method and Modified Divided Word Line
3) Dynamic Double Word Line (DDWL) Scheme
4) Pulsed Word Line (PWL) Technique
5) Dynamic Bit-line Loads (DBL) Scheme
6) Hierarchical Word Decoding (HWD) Architecture

The combination of the above mentioned circuit design techniques provides very high speed and low power performance characteristics for mega bit BiCMOS SRAMs.

This paper has been organized in the following way. In section 4.2, address transition detection circuit implementation is discussed. In section 4.3, divided word line circuit and modified divided word line circuit details are presented. In section 4.4, dynamic double word line scheme is presented. In section 4.5, pulsed word line technique is addressed. In section 4.6, dynamic bit-line loads scheme details are discussed. In section 4.7, hierarchical word decoding architecture is presented. In section 4.8, 1-Mb BiCMOS chip architecture and performance analysis results are presented. In section 4.9, conclusions are presented.

4.2 ADDRESS TRANSITION DETECTION (ATD) METHOD

The ATD pulse is the basic clock-of all the internal clocks. For the SRAM high speed operation particularly, ATD pulse width requires a severe control. If the pulse width of ATD circuit is high, it is difficult to keep the pulse width constant. The pulse width variations depend on the number of address changes or address skews. It is difficult to minimize the address skews, and the access time changes.

The ATD generator circuit diagram is shown in Figure 4.1 (a). The load impedance is adjusted synchronously with an address change to provide both the rising and falling edge [5-7]. The local ATD circuit detects a change of only one address signal. The positive pulse of local ATD (L0, L1,...,Ln) pulls down the node ATD. The impedance of the load transistor Q1 is kept high when the node ATD goes down, while the impedance becomes low before the node ATD goes high. As a result, both the rising and falling edge becomes sharp and the pulse width is easily controlled. The significance of the ATD circuit controls pulse width variations. The ATD circuit timing diagram is shown in Figure 4.1(b). Therefore, the time from the last address change is kept constant, and offers no access time increase in the presence of address skew or in any other condition.

An ATD scheme is employed to speed up access and cycle time. Transition pulses are triggered both by address and chip enable inputs and are summed by a wired-OR to generate the global pulses that precondition the data path. A monostable circuit ensures that a fixed minimum pulse width is obtained insensitive to small width variations of the input pulses and normal skew conditions. Figure 4.1(a) and 4.1(b) shows the ATD circuit and ATD timing diagram respectively.

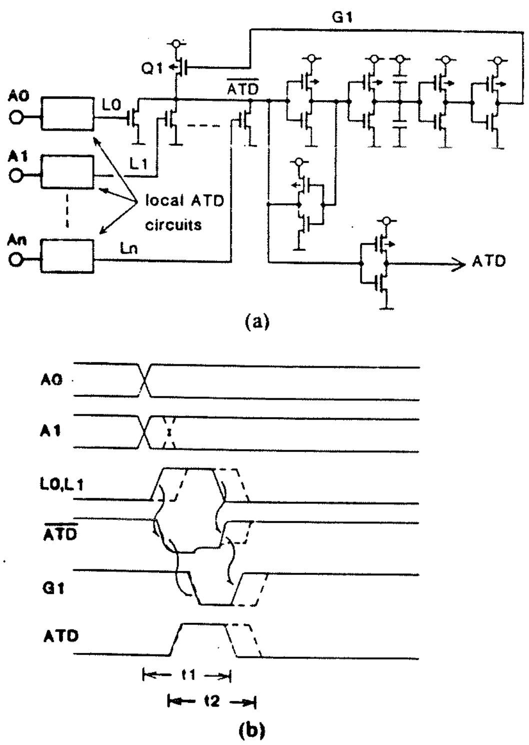


Figure 4.1(a): BiCMOS SRAM ATD Circuit, (b) BiCMOS SRAM ATD Timing Diagram

New ATD design techniques improve both speed and power performance characteristics [7]. This design technique utilizes an internally generated clock sequence to equalize bit lines and data lines after a row or column address change and then turns off the sense amplifier after the READ cycle is completed. By choosing this architecture such that only half the bit lines are precharged high during an address change and data for the four outputs of the Vss, thus power is reduced significantly. Sensing an address change immediately is the key design feature of the ATD, since this function has an impact on the access time. The circuit that detects an address change generates a time pulse to turn off all word lines and equalize all bit lines of the row decoder. Once the READ cycle is completed, then all sense amplifiers will be turned off.

For the purpose of address detection in the BiCMOS SRAM, an address buffer circuit need to be added. An address buffer senses an address change immediately, since this affects the access time. The first stage of the address buffer in the BiCMOS SRAM satisfies interface requirements with TTL inputs. Negative glitch protection circuit has been provided to prevent any over shoot occurring at the 5V level from starting a new address sequence through an address buffer. The address detection circuit is split into rows and columns. The column selection is synchronized and controlled by the row selection. During the address detection in the BiCMOS SRAM, this technique optimizes power consumption.

4.3 DIVIDED WORD LINE (DWL) STRUCTURE

The divided word-line structure [8] is illustrated in Figure 4.2. The cell array as well as a word-line is divided into N_B blocks. If the SRAM has N_C columns, N_C / N_B columns are dedicated to each block. The output of a row decoder is connected to a row select line which runs horizontally. The word-line of each block (divided word-line) is activated by each switching gate (AND gate), which has two inputs, the horizontal row select line and the vertical block select line. Consequently, only memory cells connected to one divided word-line within a selected block are accessed in a cycle.

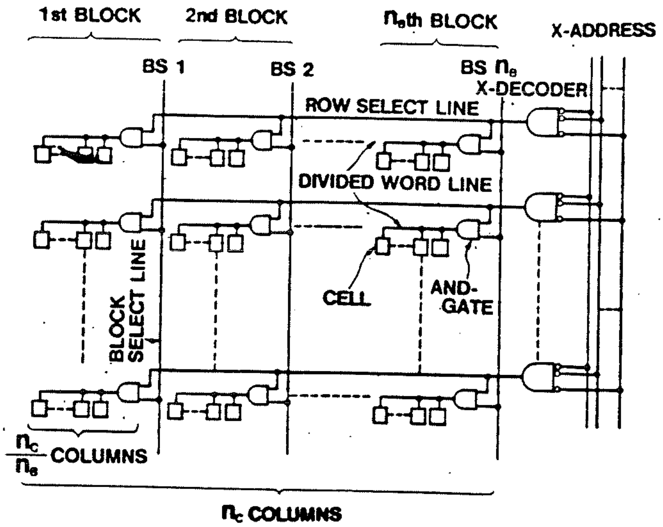


Figure 4.2: Concept of Divided Word Line Circuit Design for BiCMOS SRAM

The divided word-line scheme described above, provides the following advantages:

1) Elimination of the wasted column current. Therefore, the total power consumption in a high density static RAM with a number of columns is reduced drastically without the internal clocks [8-9]. Because it reduces the instantaneous precharge current, thus minimizing the peak current.

2) Reduction of the word selection delay. The total word-selection delay is expressed as the sum of the row select line delay and the divided word-line delay in the DWL structure.

3) Enhancement of the alpha particle-immunity in the BiCMOS devices with high resistive polysilicon loads.

However, the DWL structure suffers with a little area penalty. Decoding technique has been implemented using the multi-word decoder circuits in the

BiCMOS SRAM. A comparison of chip area penalty between DWL approach and multi-decoder approach is shown in Figure 4.3. The increase of chip size depends on the number of blocks used in the RAM chip, and the RAM cell arrangement within the chip. The chip architecture provides row decoder arrangement within the chip in different forms, which has significant impact on chip area.

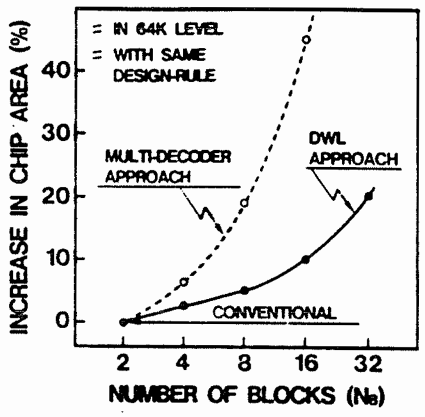


Figure 4.3: Comparison of area penalty between DWL approach and multi-decoder approach for BiCMOS SRAM

The modified DWL structure has been introduced for BiCMOS SRAM, which has significant performance improvement than the conventional DWL [9]. The word-selection circuit modification is shown in Figure4.4(a). X0-X8 and Z0-Z4 are address signals for row selection and block selection, respectively. The upper address group of X2-X8 is predecoded in the global row decoder which activates one of the row group select lines. The row group consists of four rows, each of the row select signals is input to eight NAND gates of each block row decoder. The lower address of X0 and X1 are selected by four rows of predecoded signals. The capacitance of the bit line, row select line and the width of the row decoders effect using the modified DWL structure is shown in Figure 4.4(b). The bit lines and the row select lines are fabricated by the first and second level aluminum, respectively. As the number of rows in group (N_R) increases, the bit-line capacitance decreases because of the crossover capacitance associated with the row select lines. However, the slope of the curve is very broad, especially beyond N_R=4. In contrast, the capacitance of the row select lines increases rapidly as N_R increases. Moreover, the capacitance on the predecoded signals of X0 and X1 decreases. On the other hand, the width of the global row decoder decreases and the width of the block row decoder increases, when N_R is larger. Therefore, the width of the row decoder is minimized at N_R=4 . Consequently, this modified DWL structure has been applied to the 1-Mbit BiCMOS SRAM, setting N_R=4. The modified DWL architecture with the small bit line capacitance contributes to the fast access time by a 1.5 % decrease of chip size compared to the conventional DWL.

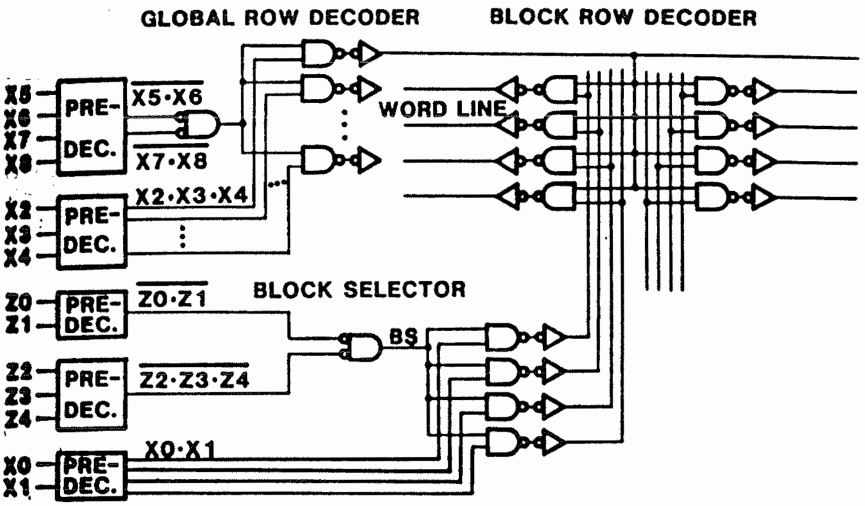
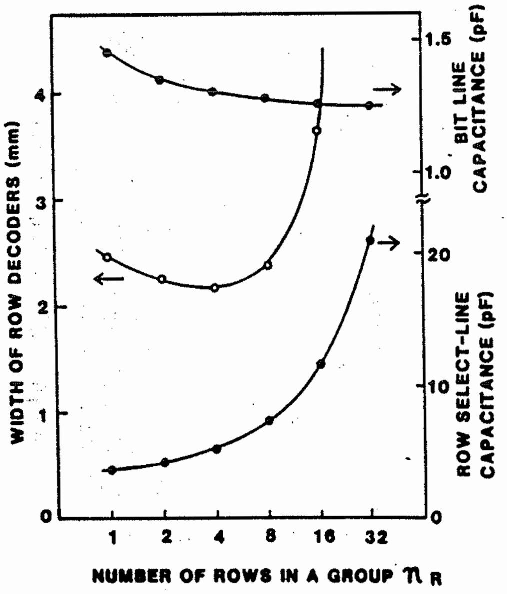


Figure 4.4 (a): Modified Divided Word Line Circuit Design for BiCMOS SRAM



Effect of modified DWL.

Figure 4.4(b): Effect of Modified Divided Word Line Structure for BiCMOS SRAM cell interconnect design

4.4 DYNAMIC DOUBLE WORD LINE (DDWL) SCHEME

The fast access time and low power operation are achieved by using a unique word line scheme, called Dynamic Double Word Line (DDWL) scheme. The DDWL is a combination of a Double Word Line (DWL) structure [10-11] and an Automatic Power Down (APD) scheme. In order to implement the APD function, a word line is activated dynamically by a pulse wider than the access time. The DDWL cuts down the operating power at 1MHZ to about 1/15th of a conventional device. Address Transition Detectors (ATD) [11] are incorporated in the RAM to generate the activation pulse for the APD function. This APD also provides a means for equilibration of bit lines (BLs) and sense lines (SLs), which turn out to be very effective in realizing a fast access time and power reduction. The equilibration circuitry together with DDWL scheme, reduces the access time to about one-half, compared with a conventional SRAM.

The basic configuration of DDWL is shown in Figure 4.5. Word lines are doubly placed, namely main word lines (MWLs) and selection word lines (SWLs) [12]. Since an MWL is not directly connected to memory cells, capacitance of the MWL is relatively small, reducing the WL delay. The MWL is made by the second aluminum layer, which also reduces the delay. Low power dissipation is expected because only one SWL is selected at a time, and consequently only a small number of memory cells are activated. From the delay time and power dissipation point of view, the greater the number of sections, the better the performance. However, the chip area is dependent on the SWL select NOR circuits, which introduce 2.5 % additional area of each cell. Therefore, the chip area increases linearly with the number of sections. Hence, the number of sections can't be set too large, although the area overhead is much smaller than a simple division of WLs, where many row decoders are required. This trade-off is illustrated in Figure4.6. In the present device, the number of sections is chosen to be 16, where the delay and power improvements are saturated. At this point, the delay and power reduction are about 30 ns and 31 %, respectively, with the chip area being the same, compared with the 4-divided conventional poly-WL structure.

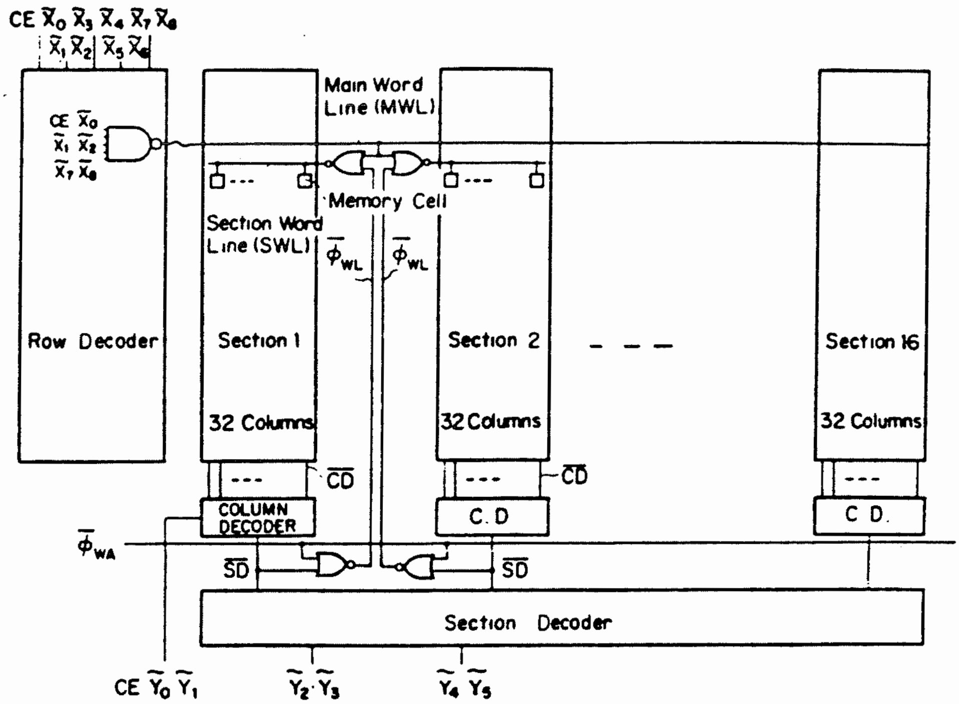


Figure 4.5: Basic Circuit Design Concept of Dynamic Double Word Line Scheme for BiCMOS SRAM

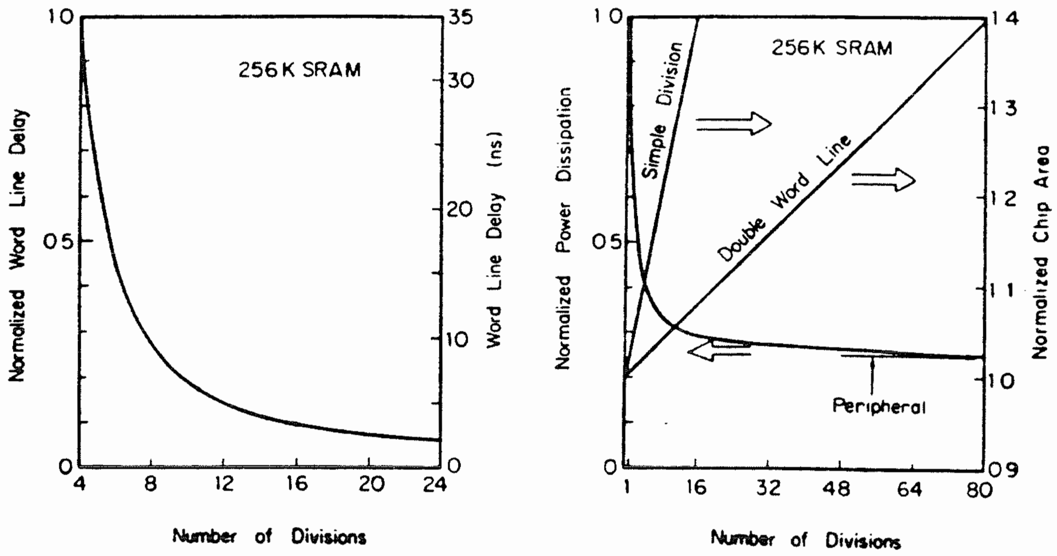


Figure 4.6: Performance trade-off results for the BiCMOS SRAM using Double Word Line Circuit Design Technique

The activation clocks have a pulse width wider than an access time, and control sense amplifiers and SWLs so as to cut all the dc path in the RAM after a read operation is over. Consequently no dc power is consumed. This is the principle of the Automatic Power Down (APD) function. Since the activation pulse width is set at about 100 ns, the supply current depends linearly on the operating frequency based on the simulation results as shown in Figure 4.7. The activation pulse should have a width of 1.5 to 2.0 times the access time in order to salvage the slowest bits and to improve yield. The operating power at 1MHz is reduced to about 1/10th of a conventional SRAM through the application of the APD function only.

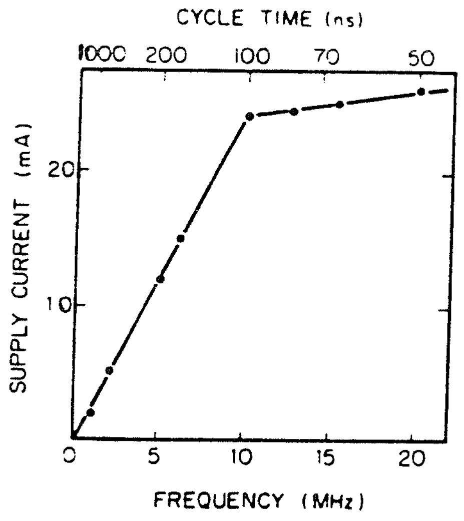


Figure 4.7: BiCMOS SRAM supply current versus operation frequency performance

4.5 PULSED WORD LINE (PWL) TECHNIQUE

Several ATD techniques have been proposed for SRAMs. The simplest usage of ATD is widely accepted for fast SRAMs, in order to equalize the data lines [13]. This is very useful to achieve fast access time but can not reduce power dissipation. The Pulsed Word Line (PWL) technique [14] is a new approach to the usage of ATD. The READ operation itself is asynchronous as in a conventional SRAMs. The internal clocks are used only to achieve the peripheral circuits for a short period, when they are really needed. This PWL technique has a high immunity to address signal noise.

The internal timing diagram of the RAM is shown in Figure 4.8. When an address signal is changed, the ATD circuit generates a base clock. The base clock enables word line and sense circuits. After the READ data are sent to output buffers and latched there, the word line and the sense amplifiers are disabled. Common data lines are equilibrated by the internal clocks, which provides fast access time. Power has been reduced only by 3.6 % through this technique.

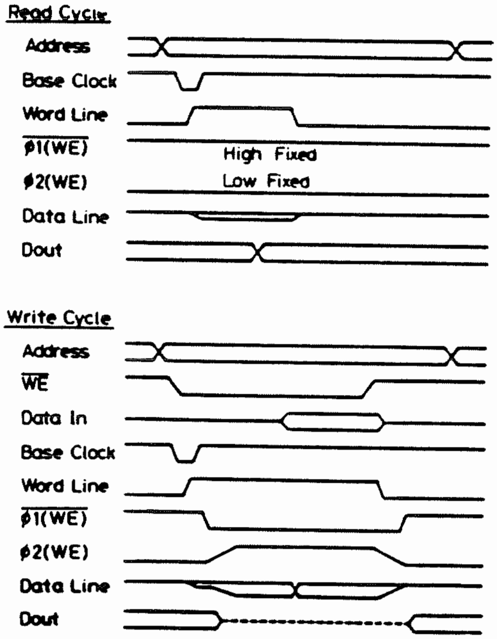


Figure 4.8: BiCMOS SRAM Timing diagram for the Read and Write cycle operation using PWL design technique

4.6 DYNAMIC BIT-LINE LOADS (DBL) SCHEME

The DBL circuit is developed to achieve high-speed bit-line precharge, equalization, and discharge during the initial stage of the read operation. A major asset of the DBL circuit is that no special transistors, such as two kinds of NMOS transistors [15], are necessary to achieve this function. The DBL circuit is shown in Figure 4.9, which consists of five PMOS transistors, five BiCMOS inverters, and four BiCMOS NOR gates. The control signal(CE.WE) is at low level in the read operation and at high level in the write operation. Figure4.10 shows a timing diagram for the read operation. The timing relation between ATR and OFQ is shown in Figure 4.10, and these pulses are generated by the internal clock generator. When an address transition occurs, the bit lines in a selected section are precharged and equalized to the Vcc level by the PMOS transistors Q1, Q2, and Q3 during the ATR pulse. Just before the ATR pulse goes low, the bit line load PMOS transistors Q4 and Q5 are cut off by the OFQ pulse, which is delayed 5 ns from the positive edge of the ATR pulse. As a result of this operation, the bit-line load consists of only the stray capacitance. Subsequently, a selected SWL goes to the high level, and the memory cells are connected to the bit lines. Due to small capacitance load on the bit line the memory cell can rapidly drive the bit-line load. Therefore, the transition speed of the voltage difference in the bit-line pair (BL-BL) is 8 ns faster than the conventional circuit in which bit lines are connected through NMOS transistors [16]. The concept of DBL is shown in Figure 4.11. The conventional circuit is placed in the left part of Figure 4.11, the bit-line loads consist of the NMOS transistor and the stray capacitance. On the other hand, the DBL during the OFQ pulse is high level, and the bit-line load is only the stray capacitance. Therefore, the memory cell can rapidly drive the bit-line load, resulting in a fast access time.

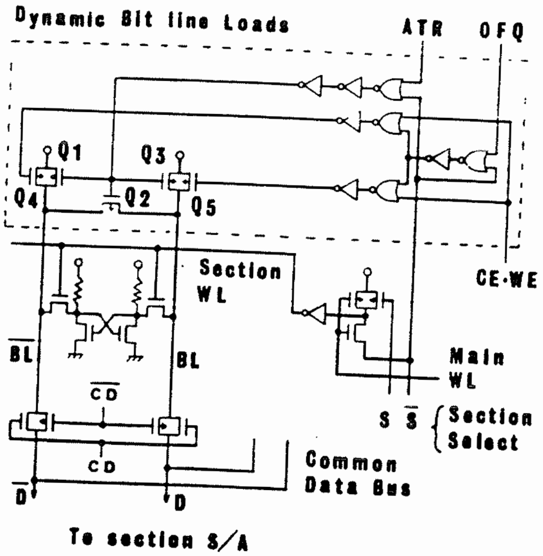


Figure 4.9: Dynamic Bit Line design concept for BiCMOS SRAM

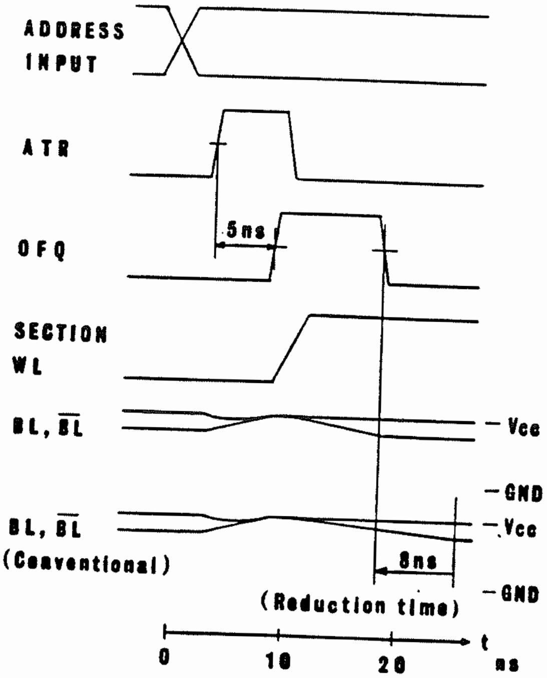


Figure 4.10: BiCMOS SRAM Timing diagram for the Read operation using Dynamic Bit Line Load concept

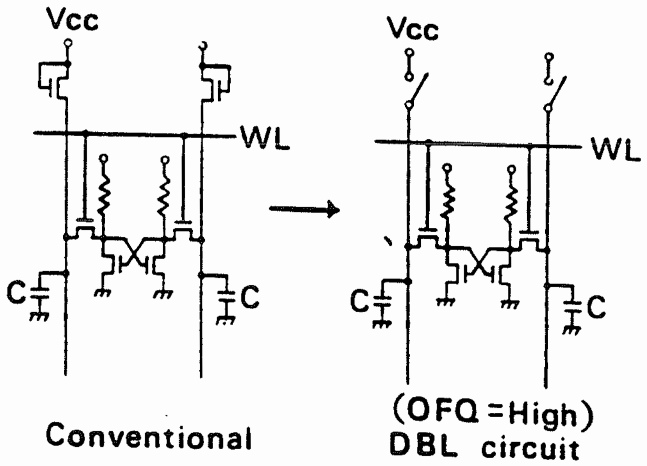


Figure 4.11: BiCMOS SRAM Dynamic Bit Line Load conventional design and Modified design concept

4.7 HIERARCHICAL WORD DECODING (HWD) ARCHITECTURE

Large part of the access time results from word-line selection time, which is the delay time from the address input to the word-line. In order to improve the speed and power performance characteristics of the word decoder circuit, the Divided Word Line (DWL) structure [17] has been widely used in high density SRAMs. Figure4.12 (a) shows the concept of the conventional DWL structure. It is adapted to the multi-divided memory cell array, whose one portion is called "block" or "sub-array". A local word-line, which is placed in each block, is activated by a global word-line and block is activated by a global word-line and a block select line. Since only one block is activated, the DWL structure reduces both word-line delay and power consumption.

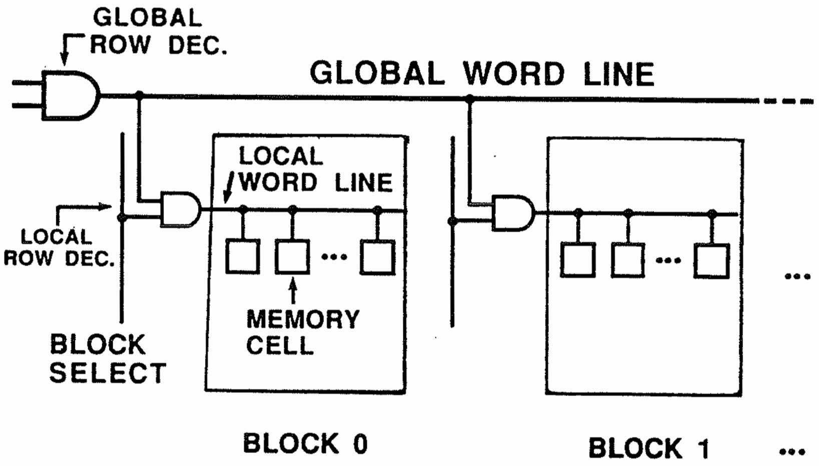


Figure 4.12 (a): BiCMOS SRAM conventional Divided Word Line Circuit Design

However, in high density SRAMs that have more than 4-Mb capacity, the number of multi-divided blocks will have to be increased. It is a technological requirement to replace the word-line RC delay and the array current. Therefore, even in the DWL, increasing load capacitance of the global word-line causes a significant increase of both delay time and power consumption. From this point of view, a novel word decoding architecture that replaces the conventional DWL will be indispensable to the future mega bit BiCMOS SRAMs.

Figure 4.12 (b) shows the concept of the hierarchical word decoding (HWD) architecture [18]. In this architecture, the word select line is divided into more than three levels. The number of hierarchies depends on the level of word-line division, which is determined by the total load capacitance of the word decoding path. In this example, the word select line has a hierarchical structure of three levels, that is, global word-line, sub-global word-line, and local word-line. In the HWD architecture , the load capacitance of the word decoding path is efficiently distributed. Therefore, the HWD architecture realizes a significant reduction in both delay time and power consumption. Figure 4.13 shows the simulation result of the minimum delay time and the total load capacitance of the word decoding time and the total load capacitance of the word decoding path. The conventional DWL and the proposed HWD are compared with each other at the optimum point of 256-Kb, 1-Mb, and 4-Mb BiCMOS SRAMs.

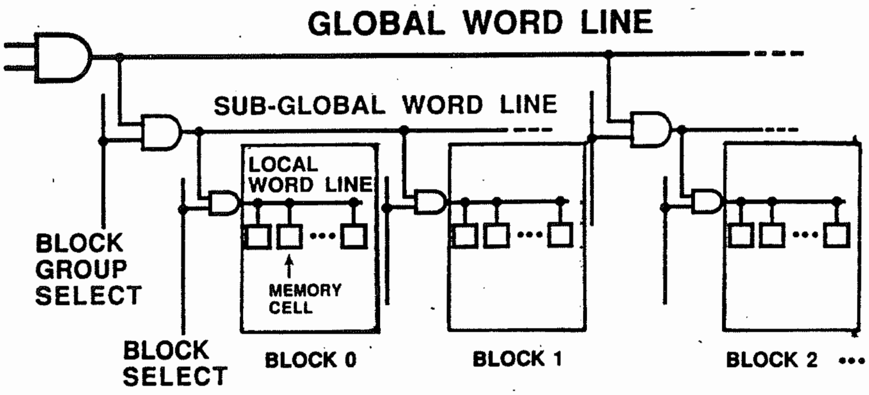


Figure 4.12 (b): BiCMOS SRAM Hierarchical Word Decoding Architecture

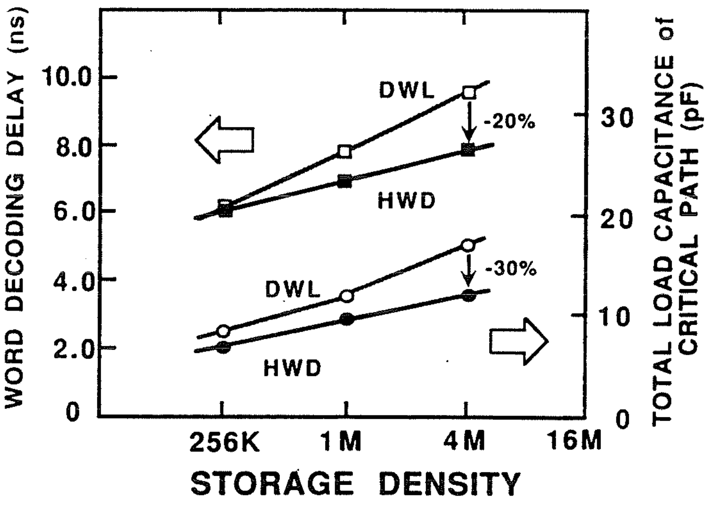


Figure 4.13: BiCMOS SRAM Performance comparison using Divided Word Line Design and Hierarchical Word Decoding Design

In a smaller density, like a 256-Kb BiCMOS SRAM, there is no significant difference. However, as the density increases the effect of the HWD is very significant. Although the HWD needs an extra decoding stage compared with the DWL, it shows better performance than that of the DWL. In 1-Mb BiCMOS SRAM, for example, the HWD architecture can reduce the delay time 20 percent and the total load capacitance 30 percent compared to the conventional DWL. This estimation predicts that the HWD architecture will be very effective for future high density BiCMOS SRAMs.

4.8 1-Mb BiCMOS SRAM CHIP PERFORMANCE COMPARISON

The BiCMOS 1-MB chip architecture block diagram is shown in Figure4.14

[19]. A high resistive poly silicon load with four transistor memory cell is used. The BiCMOS array structure has a resemblance to a CMOS RAM [19]. High packaging density is achieved through a four transistor (4T) memory cell architecture. Signal flow for the RAM is as follows: first, ECL-to-CMOS-level converters translate the input ECL signal level to CMOS signal level. Address signals are decoded by address buffers, partial decoders, row decoders, and column decoders in turn, which select a specific memory cell. The decoder circuit consists of seven stage logic gates: five are BiCMOS gates, and the other two are standard push-pull CMOS gates. The cell array consists of 512 rows by 2048 columns and is divided into 16 sections. Each section contains 128 columns and is further divided into four sub-matrices. Each sub-matrix of 512 rows X 32 columns contains local amplifiers connected internally to 4 bit-wide data bus. During in access, the differential bit-line signal is amplified by a two-stage pseudo-ECL bipolar differential sense amplifier. The first sense amplifier is placed at every local section, the second global sense amplifier is placed near the output buffer.

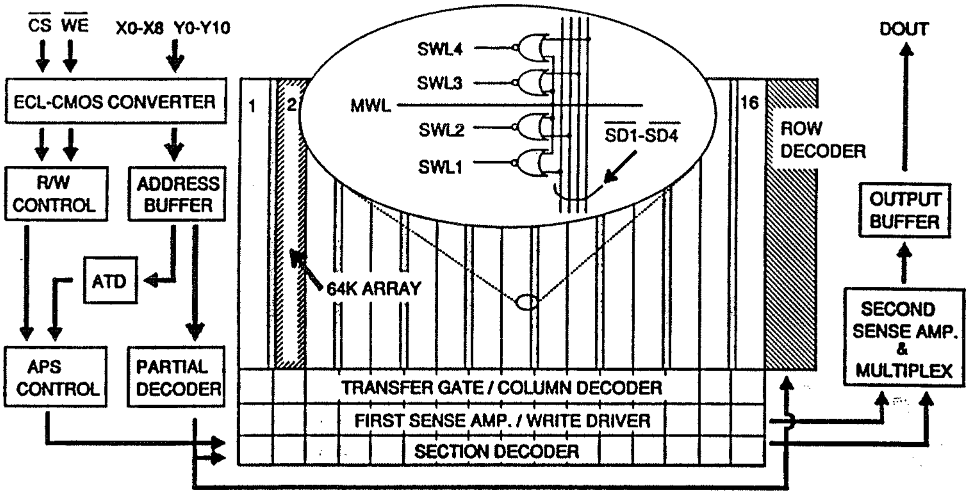


Figure 4.14: 1-Mb BiCMOS SRAM chip architecture functional block diagram

Differing from a standard CMOS RAM, an asynchronous internal architecture is used in the BiCMOS SRAM [20-21]. The address transition detector (ATD) is employed for power reduction purposes rather than being used for bit-line or data-line equilibrations. The following advanced circuit design techniques have been used in the 1-Mb BiCMOS SRAM chip:

1) Modified double word line structure

2) ECL-to-CMOS-level converter

3) Bit-line peripheral circuitry

4) Automatic power saving function

The 1 Mb BiCMOS SRAM chip architecture is presented in this paper, based on the above mentioned high speed circuit design techniques. The 1-Mb

BiCMOS SRAM design is based on 1.2 um process technology. The 5V supply voltage has been assumed for simulation. An 8-ns typical address access time has been achieved by developing a modified double-word-line structure, a bit line equalization circuit, and a double latch ECL-to-CMOS-level converter. The low power dissipation is due to the 16-divided cell array as well as the automatic power saving function. The BiCMOS SRAM chip architecture provides 21% higher speed performance than the conventional BiCMOS RAM chip as shown in Figure 4.15. The power dissipation of 1-Mb BiCMOS SRAM chip has been reduced about 10% than the conventional BiCMOS SRAM design as shown in Figure 16. Therefore significant speed and power performance improvement has been achieved using the high speed circuit design techniques.

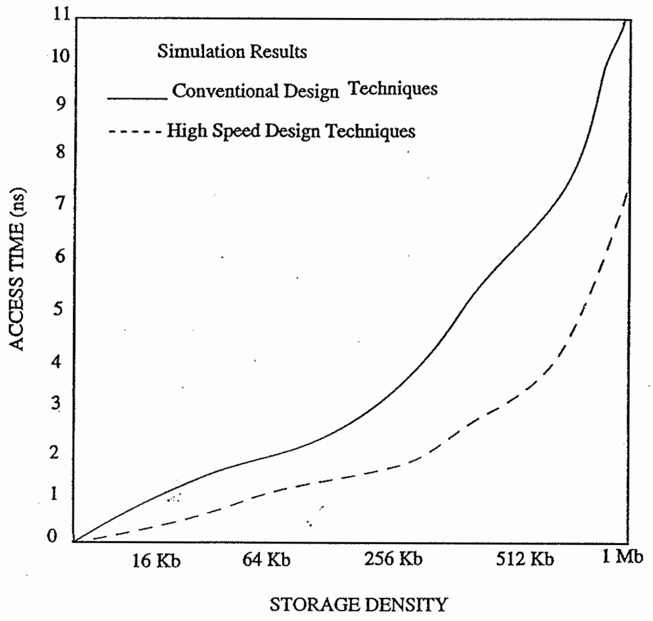


Figure 4.15: 1-Mb BiCMOS SRAM speed performance comparison

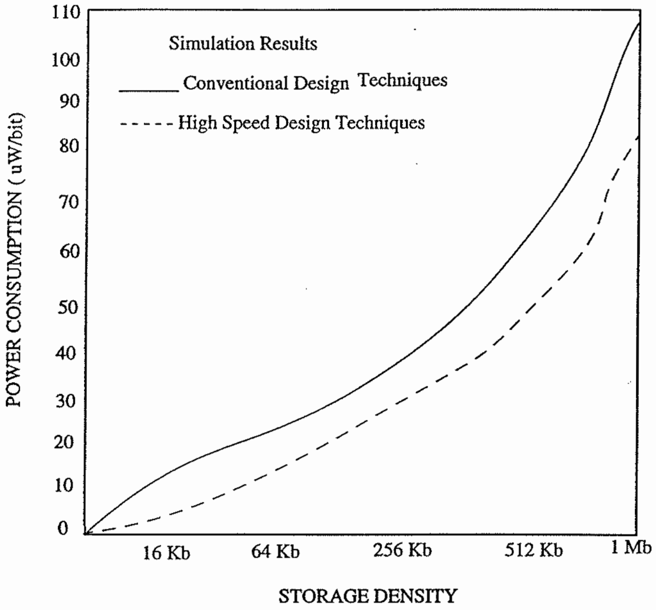


Figure 4.16: 1 Mb BiCMOS SRAM power consumption performance comparison

4.9 CONCLUSIONS

The following high speed circuit design techniques for mega bit BiCMOS SRAMs have been analyzed: address transition detection, divided wordline, dynamic double wordline, pulsed wordline, dynamic bit-line loads, and hierarchical word decoding circuit design technique.

1-Mb BiCMOS SRAM performance analysis results are presented as follows:

1) 1-Mb BiCMOS SRAM chip speed of 21% has been improved in comparison with the conventional BiCMOS SRAM design.

2) 1-Mb BiCMOS SRAM chip power consumption has been reduced about 10% compared to conventional BiCMOS SRAM design.

3) 1-Mb BiCMOS chip area has been reduced about 3% compared to the conventional BiCMOS SRAM design. Chip area calculations have been performed using analytical methods.

The above mentioned high speed circuit design techniques can be used for future mega bit BiCMOS SRAMs, DRAMs, PROMs, EEPROMs, and gate arrays.

REFERENCES

[1] T. DOUSEKI et al., "Fast access BiCMOS SRAM architecture with a VSS generator", VLSI Circuits Design Symposium Tech.Papers, June 1990, pp.45-46.

[2] M. TAKADA et al., "A 5 ns 1 MB ECL BiCMOS SRAM ", IEEE J. Solid -State Circuits, vol.25, Oct. 1990, pp. 1057-1062.

[3] Y. MAKI et al., A 6.5 ns 1 Mb BiCMOS ECL SRAM ",in ISSCC Dig. Tech. Papers, 1990, pp. 136-137.

[4] Y. KOBAYASHI et al., " Bipolar CMOS merged structure for high speed Mbit DRAM", IEDM Tech. Dig. Papers, Dec 1986, pp.802-804.

[5] M. SUZUKI et al., " A 3.5 ns 500 mW 16 Kb BiCMOS ECL SRAM " inISSCC Dig. Tech. Papers, Feb. 1989,pp. 32-33.

[6] S. FANNAGAN et.al., "Two 13-ns 64K CMOS SRAMs with very low active power and improved asynchronous circuit techniques",IEEE J.Solid State circuits, Vol.Sc-21, October 1986, pp.692-702.

[7] W.C.H.GUBBELS et.al., "A 40-ns/100-pf low power full CMOS 256K (32KX8) SRAM", IEEE Solid State Circuits, Vol.Sc-22, October 1987, pp. 741-756.

[8] K.OCHII, et al., " A 15-nW standby power 64Kb CMOS RAM",ISSCC Dig.Tech.Papers, Feb 1982, pp. 260-261.

[9] S.KONISHI et al., "A 64Kb CMOS RAM ",ISSCC Dig.Tech.Papers, Feb 1982, pp.258-259.

[10] T.SAKURAII et al., " A low power 46-ns 256Kbit CMOS static RAM with dynamic Double Word Line", IEEE J.Solid State Circuits, October 1984, pp. 578-584.

[11] M.MATSUI et al., "A 25 ns 1-M bit CMOS SRAM with loading free bit lines", IEEE J.Solid State circuits, October 1987, pp.733-738.

[12] K.C.HARDEE et al., " A fault tolerant 30 ns/375 mw 16KX1 NMOS static RAM", IEEE J. Solid State Circuits, October 1981, pp. 435-443.

[13] O.MINATO et al., " A 20-ns 64K CMOS SRAM",ISSCC Dig.Tech.Papers, Feb.1984, pp.222-223.

[14] S.T.CHU et.al., "A 25-ns low power full CMOS 1-Mbit (128KX8) SRAM",

IEEE J.Solid State Circuits, Vol.Sc-23, October 1988, pp.1078-1083.

[15] M.MATSUI et.al., "A 25-ns 1-Mbit CMOS SRAM with loading free bit lines", IEEE J.Solid State circuits, Vol.Sc-22, October 1987, pp.733-740.

[16] T.KOMASTSU et.al., "A 35-ns 128KX8 CMOS SRAM",IEEE J.Solid State Circuits, Vol.Sc-22, October 1987, pp.721-726.

[17] K.SASAKI et.al., " A 23-ns 4 Mb CMOS SRAM with 0.5 uA standby current", ISSCC Dig.Tech.Papers, Feb 1990, pp.130-131.

[18] K.OGIVE et.al., "13-ns 500 mw, 64Kbit ECL RAM using HI-BiCMOS technology", IEEE J.Solid State circuits, Vol.Sc-, October 1986, pp.681-685.

[19] T.HIROSE etal., A 20 ns 4Mb CMOS SRAM with hierarchical word decoding architecture, ISSCC Dig.Tech.Papers, Feb 1990, pp 132-133.

[20] M.MATSUI et al., " A 25-ns 1-Mbit CMOS SRAM with loading free bit lines", IEEE J.Solid State Circuits, Vol.Sc-22, October 1987, pp.733-740.

[21] M.MATSUI et al., " An 8-ns 1-Mbit ECL BiCMOS SRAM with Double-Latch ECL-to-CMOS-Level converters", IEEE J.Solid State Circuits, Vol.Sc-24, October 1989, pp. 1226-1231.

CHAPTER 5

A DYNAMIC RECONFIGURATION SCHEME FOR MEGA BIT STATIC RANDOM ACCESS MEMORIES

Article by: Venkatapathi N Rayapati and Bozena Kamiska,

Published in International Journal of Microelectronics and

Reliability, Vol.34, No.1, pp. 107-114, 1994.

ABSTRACT

The objective of this paper is to present a novel dynamic reconfiguration scheme for mega bit Static Random Access Memories (SRAMs). Most of the conventional reconfiguration methods are implemented using two-way switching elements. The proposed scheme is based on on-chip word failure detection and reconfiguration to spare word cell using multi-valued logic elements. The physical concept of the dynamic reconfiguration scheme and implementation details are discussed. Based on the SRAM dynamic reconfiguration implementation a reliability model is developed. Dynamic reconfiguration scheme reliability comparisons with other existing methods are presented. The advantages of the proposed dynamic reconfiguration scheme are highlighted.

5.1 INTRODUCTION

The era of Mega bit (Mb) Static Random Access Memories (SRAMs) has promised further advances in electronic data processing, such as super-computers, engineering work-stations, and microprocessor applications, etc. Device technology that ensures the device long term reliability and provides high performance characteristics, which is a key design issue for developing high performance Mb SRAMs.

Status of the research work done so for is provided through literature survey, Wada et al [1] developed a 34 ns 1 Mb CMOS SRAM using triple polysilicon gate architecture. Recently Kohno et al [2] designed a 14 ns 1 Mb CMOS RAM with variable organization. However, the variable organization does not consider failure tolerance and reliability aspects. A multiple word/bit line redundancy configuration for semiconductor memories was proposed by Schuster [3] and a static redundancy configuration for semiconductor memory was addressed. A general fault tolerant technique for memory components was proposed by Arzali [4]. More recently, the fault tolerance in NMOS RAMs with dynamic redundancy methods were proposed by Rayapati et al [5]. Two dynamic redundancy approaches were implemented to provide fault-tolerance capability at the chip level. Finally, a fault model for the multi-valued NMOS Dynamic RAM was introduced by Rayapati et al [6]. Multi-valued DRAM fault conditions and applications are discussed. Some fault tolerant hardware design considerations are presented by Lala [7]. So for in the literature chip level reconfiguration methods are addressed, the current research problem is to investigate device level fault-tolerance capability and device level reconfiguration. In this paper a dynamic reconfiguration scheme for Mb SRAMs is proposed. The proposed approach provides a parallel word redundancy with failure detection and on-line switching dynamic reconfiguration. The memory cell architecture is designed in such a way that the memory cell provides built-in fault-tolerance, single event upset tolerance and soft failure detection capability. Device reliability performance comparisons are provided using the proposed scheme with other existing methods. Applications of the proposed reconfiguration scheme for future mega bit SRAMs are presented.

5.2 DYNAMIC RECONFIGURATION SCHEME

Redundancy is indispensable in high speed, high density SRAMs for obtaining a satisfactory yield, fault-tolerance capability and reliability. However, in using conventional redundancy schemes in which a spare word replaces a defective word by switching between decoders, an access time delay of a few nanoseconds is inevitable. This access delay time is unacceptable for high speed SRAMs. Therefore, the objective of this research to investigate device level built-in fault tolerance with dynamic redundancy methods for SRAMs.

In this paper a dynamic reconfiguration scheme is proposed, which provides provision for a faulty word is replaced with a standby spare word through a dynamic reconfiguration network using multi-valued logic elements. A standby spare word, which replaces a faulty word, is placed in the cell array block adjacent to the cell array in which a defective word cell exists. The proposed dynamic reconfiguration architecture is shown in Figure 5.1, where the standby spare word 1 can be replaced with the word 1 in the memory cell array and is placed in cell array 2. Standby spare word 2 can be replaced with word 2, and placed in the memory cell array 1. Suppose the array block section of input signal x_1, when corresponding to the output signal y_1 is of the high level. In this case, both word 1 and standby spare word 1 are simultaneously activated. Therefore, the cell data in the standby spare word 1 reads the sense amplifier 2 no later than the cell data in word 1 reaches the sense amplifier 1. In the normal operation, the sense amplifier 1 is selected by y_1, and the cell data in word 1 appears on the data bus. If word 1 contains a faulty memory cell, the connections in the switch are changed by detecting the faults in the word address, and sense amplifier 2 is selected. Thus, the cell data in standby spare word 1 appears on the data bus. Failures in word 1 and standby spare word 1 are detected by error correcting codes. Once a failure is detected on the SRAM device, automatically the reconfiguration network switches to standby spare words. Hence, there is no data access interption in the SRAM device.

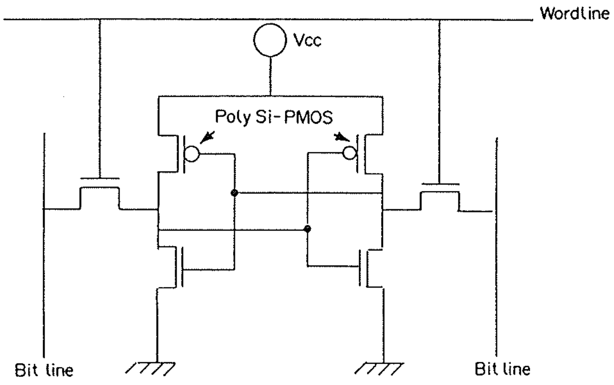


Figure 5.1: Dynamic reconfiguration scheme basic concept for mega bit SRAMs

The delay time is considered to be the time required for both failure detection in the defective word address and on-line reconfiguration to the standby spare word is less than 1ns, as compared to the time required for activating a word line and transmitting the cell data to a sense amplifier. Therefore, the parallel-word-access reconfiguration architecture provides both the normal cell array block and the redundant cell array current becomes twice as that of the conventional reconfiguration methods. One standby spare word line, which is based on the parallel-word-access reconfiguration scheme, will be incorporated in each of the 32 cell array blocks of the 1 Mb SRAM. The proposed SRAM dynamic reconfiguration scheme is very efficient for high speed computer applications, where the access time has very stringent requirements.

5.3 RECONFIGURATION SCHEME IMPLEMENTATION

In this section, the proposed dynamic reconfiguration scheme implementation details are discussed. Memory cell circuit diagram is shown in figure 5.2.

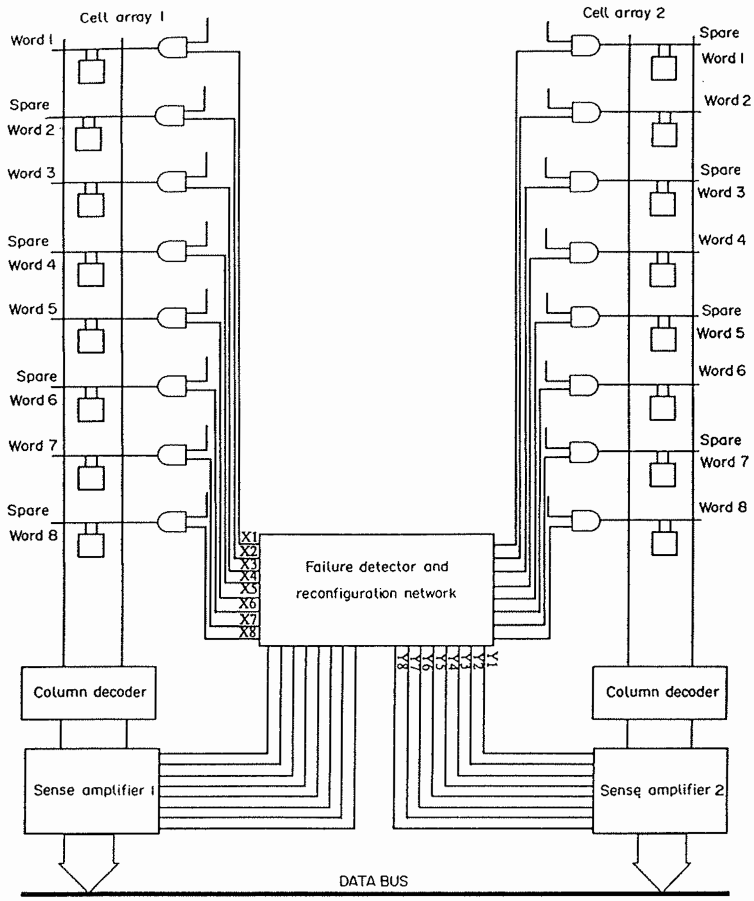


Figure 5.2: Dynamic reconfiguration architecture implementation for mega bit SRAM

Its implementation is based on the polysilicon PMOS transistors, which are used as memory cell loads instead of conventional high resistance polysilicon loads. The gate electrode of polysilicon PMOS transistor is formed by the second polysilicon layer. The source and the drain channels are formed by the third polysilicon layer. The second polysilicon layer is also used for a self-aligned bit-line contact, which reduces the memory cell area. After SRAM device read/write operations, the high level cell node voltage drops to the threshold voltage, which is due to backgate bias effect in the conventional SRAM circuits.

How fast the lower cell node voltage is charged up to V_{cc} is very important for redundancy and the soft failure detection problem. The proposed memory cell has high radiation resistance capability, which provides built-in fault tolerance and radiation induced soft failure tolerant capability to the SRAM device. Because of the large storage charge, the polysilicon PMOS load memory cells are very efficient, soft failure tolerant elements. This reconfiguration scheme is used as a radiation hardness design technique for mega bit SRAMs. As the supply voltage decreases, the soft failure rate will be decreased by two to three orders of magnitude, depending on the intensity of the ionizing radiation. In the 1 Mb SRAM functional block diagram, the cell array is divided into 32 blocks in the column direction. Each cell array block is controlled by a word decoder. Thirty-two local sense amplifiers, each consisting of a CMOS pre-amplifier, the first and second PMOS cross coupled amplifiers are located on the side of each cell array block. All peripheral circuits will be placed on the two long sides of the chip. High speed sense amplification of the small memory cell is the key design feature to achieve fast access time for the mega bit SRAMs.

5.4 RELIABILITY MODEL FOR A 1 Mb SRAM

The proposed dynamic configuration scheme implements two consecutive functions, such as failure detection in the memory words and on-line dynamic switching through multi-valued logic elements to the standby spare words within the same SRAM device.

The reliability model is developed based on the proposed dynamic reconfiguration scheme implementation. Reliability of the mega bit SRAM with the dynamic reconfiguration scheme having S standby spare words is given as follows:

$$R = 1 - (1 - R_m)^{s+1} \qquad (1)$$

where R_m denotes the reliability of the memory device active word or standby spare word in the 1 Mb SRAM, and s represents the number of standby spare words per cell array.

The arrangement of memory words in each cell array represents a hot-standby system. The hot-standby memory system is to operate two memory words in parallel with either memory word acting as a standby spare word. A failure detection circuit continuously checks each word in each cell array of the memory system. Failure detection circuit hardware is based on the choice of the error correcting codes. If a failure is detected in a memory word of any cell array within the chip, on-line reconfiguration switching mechanism isolates the failed word and automatically switches to the standby spare word. The reliability of 1 Mb chip can be expressed as follows:

$$R = [R_m^2 + 2F_cR_m (1-R_m)] R_s R_c \qquad (2)$$

where R_m denotes the reliability of each memory cell array within the chip; R_c stands for the reliability of failure detection circuit hardware; R_s is the reliability of the reconfiguration switching mechanism; and F_c represents the fault coverage correction factor.

Reliability improvement using the proposed dynamic reconfiguration scheme for 1 Mb SRAMs is compared with other reconfiguration methods and is shown in figure 5.3. From figure 3, it is evident that the proposed dynamic reconfiguration scheme significantly improves reliability.

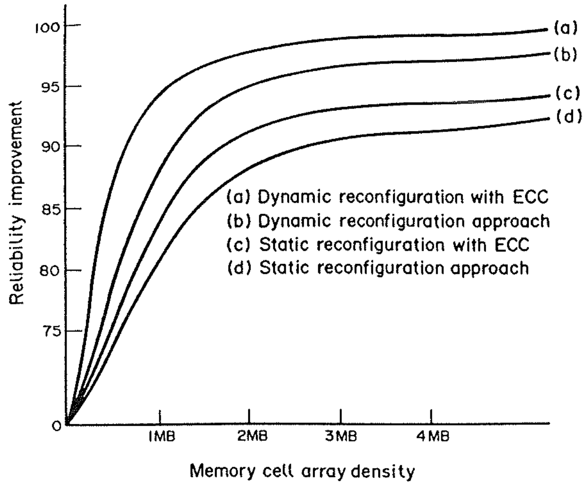


Figure 5.3: Dynamic reconfiguration scheme reliability improvement comparison with other methods

5.5 CONCLUSIONS

The proposed dynamic reconfiguration scheme for Mb SRAMs, has the following advantages over the conventional reconfiguration methods:

1) On-line failure detection and dynamic reconfiguration switching mechanism at the device level considerably reduces word access time penalty.

2) Significant reliability improvement in comparison with the conventional approaches.

3) High speed sense amplification design methodology provides improved speed.

4) Soft failure tolerant capability within the memory cell increases the fault tolerance at the chip level.

5) Parallel-word redundancy approach improves yield and reliability of Mb SRAMs.

6) Provides radiation tolerant capability to the SRAM device for single event upsets and ionizing total dose effects.

The proposed dynamic reconfiguration scheme can be used for future mega bit DRAMs, programmable ROMs and field programmable gate array devices.

REFERENCES

[1] T. WADA et al, "A 34 ns 1 Mb CMOS RAM using triple poly", ISSCC Digest of Technical Papers, Feb. 1987, pp. 262-263.

[2] Y. KOHNO et al, "A 14 ns 1-Mbit CMOS RAM with variable organization", IEEE J. Solid State Circuits, Vol. 23, Oct. 1988, pp. 1060-1066.

[3] S.E. SCHUSTER, "Multiple word/bit line redundancy for semiconductor memories", IEEE J.Solid State Circuits, Vol.SC-13, Oct. 1978, pp. 698-703.

[4] L.M. ARZALI, M. KUBO, and T. MANO, "Fault tolerant techniques for memory components", ISSCC Dig. Tech. papers Vol. 28, Feb. 1985, pp.231.

[5] V. RAYAPATI and S. MAHAPATRA, "Fault tolerance in NMOS Random Access Memories with Dynamic Redundancy Methods",Microelectronics and Reliability: An International Journal, Vol.28, No. 2, 1988, pp. 193-200.

[6] V. RAYAPATI and S. MAHAPATRA, "A Fault Model for Multi-Value NMOS Dynamic Random Access Memories", Microelectronics and Reliability: An International Journal, Vol. 29, No. 2, 1989, pp. 42-50.

[7] P.K. LALA, "Fault Tolerant and Fault Testable Hardware Design",Prentice Hall International, 1985.

CHAPTER 6

DYNAMIC RECONFIGURATION SCHEMES FOR MEGA BIT BiCMOS SRAMs AND PERFORMANCE EVALUATION

Article by: Venkatapathi N Rayapati and Bozena Kaminska,

Accepted for publication in International Journal of

Microelectronics and Reliability, Pergamon press Ltd, Oxford, U.K.

ABSTRACT

In this paper two dynamic reconfiguration schemes are discussed for mega bit BiCMOS SRAMs. Dynamic reconfiguration schemes allow the failure detection at the chip level and automatic reconfiguration to fault free memory cells within the chip. The first scheme is a standby system approach where the I/O lines of the memory can be dynamically switched to spare bit slices in the SRAM. This scheme is implemented through a switching network at the memory interface. Every memory access is controlled by a fault status table (FST) which memorizes the fault conditions of each memory block. This fault status table is implemented outside the memory system. Second dynamic reconfiguration scheme for BiCMOS SRAMs is addressed through a graceful degradation approach. Basic design considerations and performance evaluation of mega bit BiCMOS SRAMs using dynamic reconfiguration schemes are presented. BiCMOS SRAM access time improvement of about 35 %, chip area of 25%, and chip yield of 10% are achieved respectively, as compared to the conventional methods. Comparison of reliability improvement of 1-MB BiCMOS SRAMs using dynamic configuration schemes is addressed. These two dynamic reconfiguration schemes have considerable importance in reliability improvement when compared to the conventional methods. The major advantage is that the size of reconfiguration of the system can be considerably reduced.

6.1 INTRODUCTION

In any computer system the memory usually comprises the largest number of components, which constitute the most unreliable part, even though the mean failure rates of BiCMOS SRAMs is quite low. There is a growing demand for BiCMOS SRAMs due to their high speed, high density, and low power consumption characteristics. The complexity of computer system depends on the hard core memory requirements. The computer system potential failures result in due to BiCMOS mega bit SRAMs. Many computer applications require additional memory and high reliability. In order to increase the reliability it is necessary to provide some internal fault tolerance to BiCMOS mega bit SRAMs.

Fault tolerance memory techniques are error detecting and correcting codes. The major advantage of the codes is to identify errors and correct the fault on the fly (i.e. no interruption of the normal SRAM operation is necessary). Especially, codes for correcting single bit errors in the memory cells can be easily implemented. To detect multiple bit errors in memory cells requires additional hardware for checking and coding. Therefore, in fault tolerant BiCMOS SRAM implementation the use of codes is confined to single error correcting/double error detecting (SEC/DED) codes [1]. Using these codes, commonly occurring transient faults can be adequately handled as long as only such transient fault holds the SRAM, it is masked by the choice of codes. After a relatively short time transient fault errors may be assumed to end. As a result transient fault errors cannot be combine with static errors in the SRAM. Static bit errors in the memory cells are not removed and these errors can be neither detected nor corrected by the codes [2]. This effect limits the reliability of BiCMOS SRAMs that is achievable using coding techniques.

In order to achieve high reliability of mega bit BiCMOS SRAMs, the additional use of dynamic redundancy technique is of specific importance for computer systems. Stiffler[3] and Sarrizir and Malek[4] proposed redundancy techniques, which adds access time penalty, low yield, and increases chip area. In this paper a dynamic redundancy configuration scheme is proposed for BiCMOS SRAMs. In this approach, once a fault condition is detected then the failed memory cell arrays/columns are logically removed from the SRAM chip using the dynamic reconfiguration method. Two main strategies of dynamic redundancy can be achieved: (1) the standby approach, (2) the soft failure (graceful degradation) technique. The standby approach requires additional (spares) hardware components. If a memory array/column fails, its function is taken over by a spare. An alternative approach is the graceful degradation technique to improve mega bit SRAM reliability by using redundancy which is inherently built-in in the SRAM chip.

For large memories the standby approach usually requires extra hardware for storing fault conditions and carrying out necessary reconfiguration. In this paper new approaches are proposed in which the reconfiguration hardware components are considerably reduced as compared to existing standard techniques.

The most significant contribution is that this additional hardware module is interfaced with the main memory of the system.

In the following, first a BiCMOS Static Random Access Memory (SRAM) model is considered and then a corresponding fault tolerance model is introduced. A special standby reconfiguration scheme implementation is discussed. This approach is then modified for the case that no additional spare bit slices are available within the SRAM. The resulting reconfiguration scheme implements a graceful degradation of the BiCMOS SRAM for computer systems.

6.2 MODEL FOR THE BiCMOS MEGA BIT STATIC RANDOM ACCESS MEMORY (SRAM)

Logically the BiCMOS SRAM consists of a number of word cells each consisting of i bit cells to store a data word of L bits, where L is defined as the word length of the SRAM. The bit cells i (i = 0,1, ..., n-1) of all word cells together form the so called "bit slice i" of the memory, as shown in Figure 6.1.

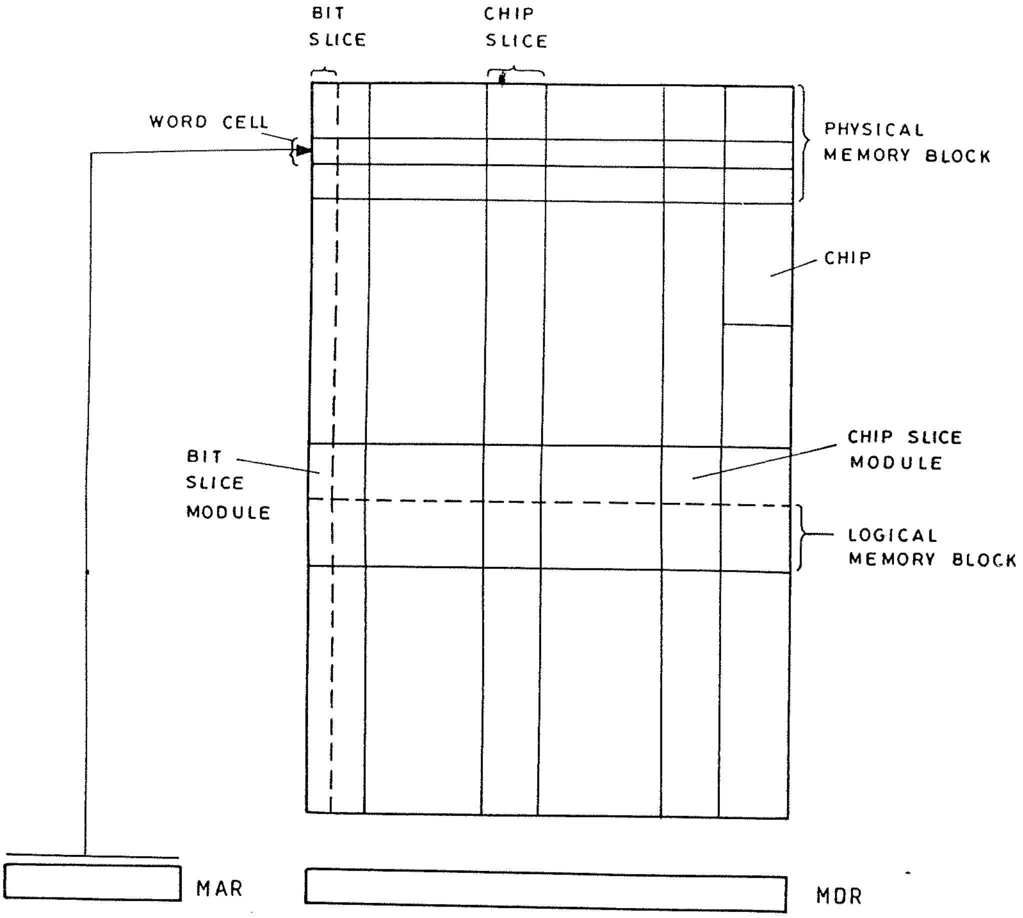


Figure 6.1: BiCMOS SRAM Memory Structure and Notation

Usually BiCMOS SRAM is constructed as a two-dimensional array of m rows and n columns. A row of the array is called a "memory block". A column of a shell is called a "chip slice". Inside each chip there are a number of d bit cells. So, a physical memory block has a capacity of m*d bits. Functionally, the d bit cells of a chip may be arranged in a two-dimensional array of d/L rows, each having L bit cells. Each row may be addressed and the contents of the addressed row can be concurrently accessed through read/write lines.

Based on the physical architecture, every memory block is used to constitute d/L word cells of n = L*m bits. The entire memory has C*d/L words. A word cell is addressed by decoding its memory address by an appropriate decoding logic. Addressing of a word cell is usually performed in the following way. First, the decoder determines the address block "lab" to which the word cell belongs. All m chips of that particular block are activated, while on each chip an intra - block decoder reads the `least significant bit' La (d/n) of the memory address which defines the position of the word cell inside the memory block. All the bit cells of the addressed word cell are activated, so that data can be read/write into the word cell.

In many cases, the chip architecture provides only for storage of L = 1 bit/word, i.e. each chip has only L = 1 read/write lines and we have d words of n = m bits on each memory block. Then a chip slice module contains a part of only one bit slice.

Independent of the physical size of the chip, here we introduce a new concept "bit slice modules". Each bit slice may be logically divided into binary number b' + f*b, where f = 1, 2, 4, 16, ... , etc. of identical bit slice modules, as shown in Figure 1. so each chip contains L*f such modules arranged in f rows and L columns. Those bit slice modules in the n bit slices which have the same column position together form a logical memory block, as shown in Figure 1. In this approach "bit slice modules" will constitute the basic units of "dynamic reconfiguration" scheme. Therefore, the size of these units can be reduced to the requirements of SRAM application in computer systems.

Example: The number f may also be chosen to be f = 1, i.e. the part of a bit slice residing on one chip contains exactly one bit slice module. For f = 1 and L =1 we have the standard case that the bit slice module just coincides with the chip size and with the size of the chip slice module.

6.3 DYNAMIC STAND BY CONFIGURATION METHOD FOR MEGA BIT BiCMOS SRAMs

The basic idea of standby system for SRAMs is usually to provide a number of S spare bit slices to tolerate faults within the SRAM array. In the BiCMOS architecture, for each word cell of n bits length there are S spare bit cells, so that it will function with upto S defective bit cells in each word cell. It would require a large overhead of storage if each word cell of memory were to be reconfigured individually. By introducing the new concept of "bit slice modules" the replacement of faulty memory cells will be more efficiently performed, i.e. if there is an arbitrary fault within a bit slice module, this module is totally replaced by a spare module. Therefore, the system user has a wide choice of reduction in the size of "bit slice modules" for that particular application. In the case of chips having very large storage capacity, units of reconfiguration which are smaller than chip size may be very advantageous.

In the case of defective bit slice module, the I/O lines to these modules must be switched by corresponding modules in the spare bit slices, as shown in

Figure 6.2. This is the basis of the reconfiguration scheme, in contrast to conventional standby systems. This reconfiguration scheme is carried out by a network at the memory interface. The logic switches for this network are set independently for every memory access; the system is therefore called a "dynamic standby system (DSS)".

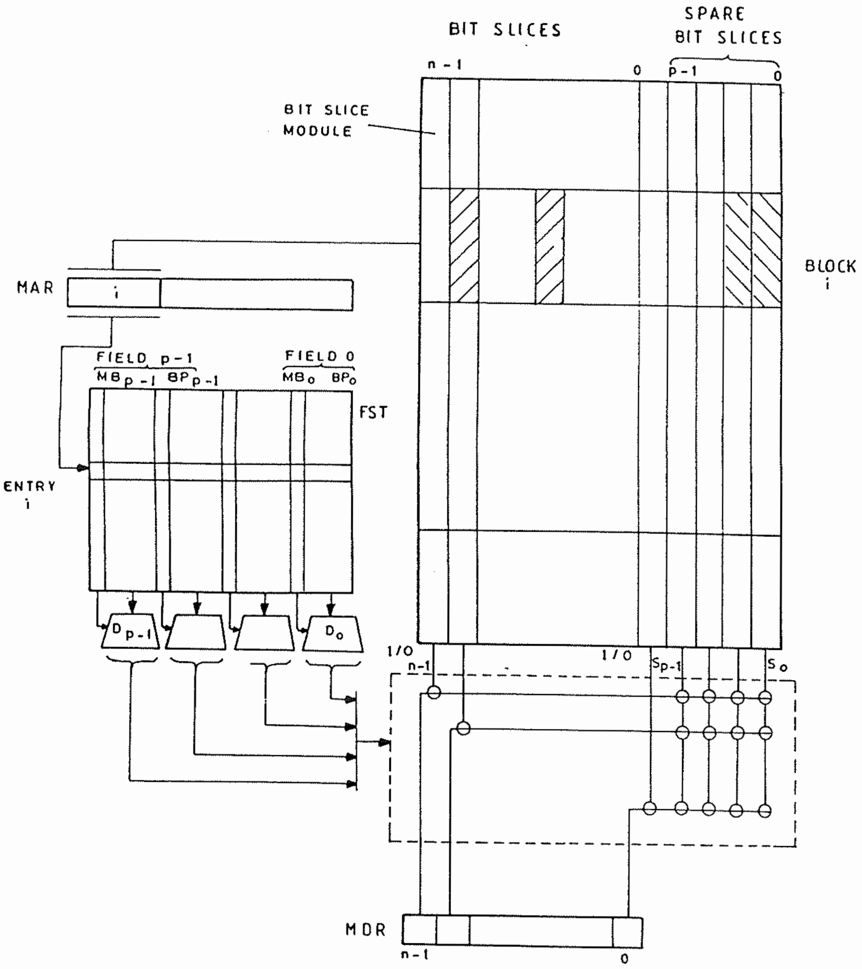


Figure 6.2: BiCMOS SRAM chip Architecture implementation for Dynamic Reconfiguration Scheme

The network is inserted between the n bits of the memory data register and the n+S I/O lines leading from the n+S bit slices of the SRAM array. If a word is read/write into the SRAM, it is transferred through n bit cells of the memory data register and n lines out of the n+S I/O lines. These lines are selected by logic switches of the reconfiguration network. Depending on the status of the logic switches, a bit cell i (i = 0, 1, 2, ..., n-1) of the memory data register, instead of being connected to a bit cell in a defective bit slice module i of a block j, can be logically switched to a corresponding bit cell in a spare module associated to block j, as shown in Figure 6.3.

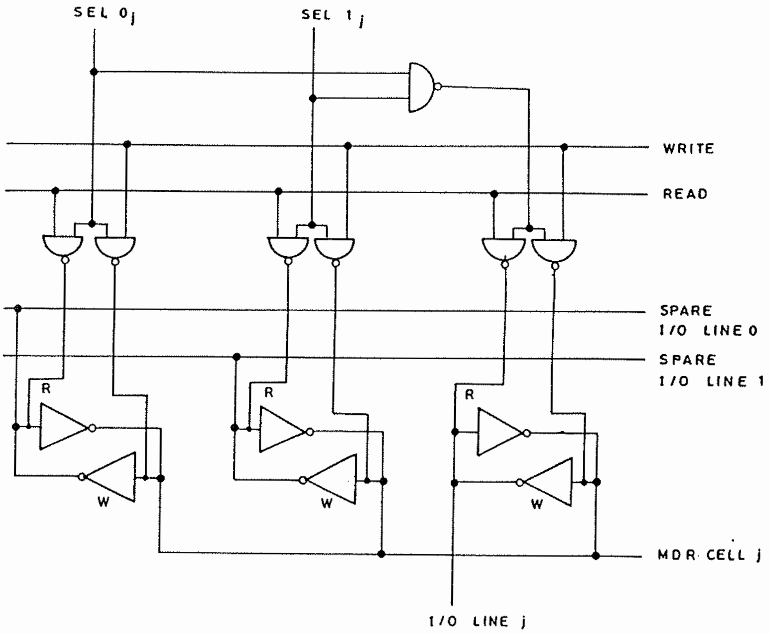


Figure 6.3: BiCMOS Switch Structure Functional Block Diagram

The status of the logic switches is controlled by the contents of a fault status table (FST) implemented outside the SRAM with an extra memory of small capacity and short access time. In this table, for each logical memory block j (j = 0, ..., b' - 1) an entry is provided to store information concerning faults. This entry comprises S fields, each assigned to one of the corresponding P spare bit slice modules of the block within each field K (K = 0, ..., S - 1). A mark bit MB denotes whether spare K is free (MB = 0) or used (MB = 1) to replace a defective bit slice module. The column position i (i = 0, 1, 2, ..., n - 1) of this replaced module inside its logic memory block is stored in the subfield BP of the entry.

If a word cell with address d in a memory block j is to be accessed, the most significant bits of d determining j are used to address the corresponding entry j in the FST as shown in Figure 6.4. As the FST is stored in its own memory, the entry can be read out of the FST concurrently with the main memory access. Outside FST memory its S fields are decoded by S : 1 - out - of -n decoders D_0, D_1, D_2, ..., D_{n-1}. If the mark bit MB of the corresponding field K (K = 0,1, ..., S -1) is 0, the decoder D_k remains passive and normal memory access is performed. Otherwise decoder D_k is activated and from the contents i (i = 0, 1, ..., n - 1) of the subfield BP a control signal is derived. This signal causes the logic switches of the reconfiguration network to connect the bit cell i of the memory data register to the required space bit slice module K. As the FST access is performed concurrently with the main memory access, the introduced scheme implies no relevant performance degradation.

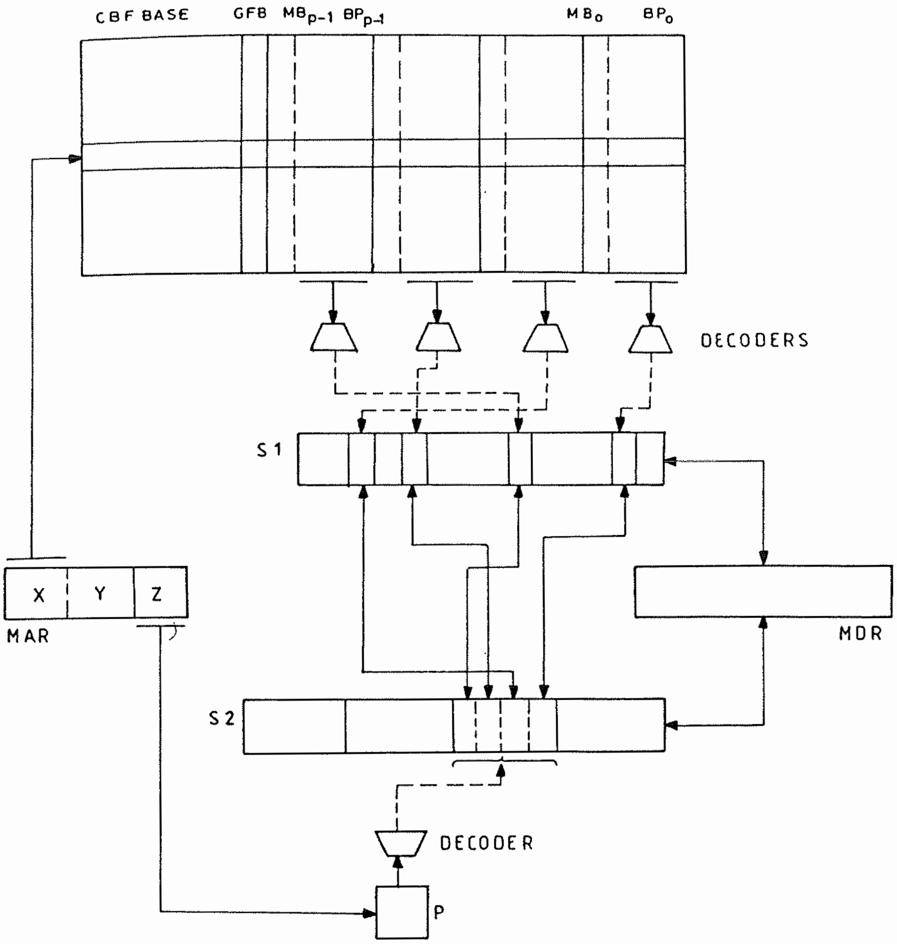


Figure 6.4: Dynamic Reconfiguration Scheme Bit Transfer Unit Architecture for BiCMOS SRAM

Example: Reduction of reconfiguration network to half of its former value. This can be performed without any change in its internal structure of the SRAM array. It only implies that the number of entries which have to be provided in the FST is doubled [5]. This can be generalized to arbitrary fractions 1/f = Z ** (-(df)) of the chip capacity. So, this method is more practical for mega bit BiCMOS SRAMs.

6.4 DESIGN CONSIDERATIONS AND PERFORMANCE EVALUATION OF A FAULT TOLERANT BiCMOS SRAM

BiCMOS technology promises a large number of switching elements available in a single SRAM chip. BiCMOS circuits lead to high speed of operation, smaller size, high density, and operate at low power levels. Memory band width is considerably enhanced. Here a standard 1.2μm BiCMOS process is used to analyze the fault tolerant SRAM [6-9]. The manufacturing cost of BiCMOS can be expressed as

$$\text{Manufacturing cost} = K\,(10^{0.0211A}),$$

where A is the area of the chip in square millimetres per 1000 and K is constant for a given process. Here an exponential relationship between manufacturing cost and chip area is valid. The main memory of the computer system provides faster access time, but lower density. The fault tolerant BiCMOS SRAM provides low access time, low density, low power consumption, which coincides with the CMOS, ECL, TTL interface capability within the system.

Due to the inherent regularities of the FST at the reconfiguration network, these BiCMOS circuits appear to be ideal one for SRAM realization. IBM, NEC, Micron, HoneyWell, and Siemens have developed BiCMOS SRAMs those can be moderately added to the interface of the proposed BiCMOS SRAM chip. This chip is called fault-tolerant BiCMOS SRAM interface chip. The hardware design plan of the BiCMOS SRAM is shown in Figure 6.5. The basic BiCMOS SRAM cell layout is shown in Figure 6.6. The SRAM cell shown in Figure 6.6 forms the main component of the BiCMOS mega bit SRAM chip. BiCMOS SRAM chip hardware layout is shown in Figure 6.7.

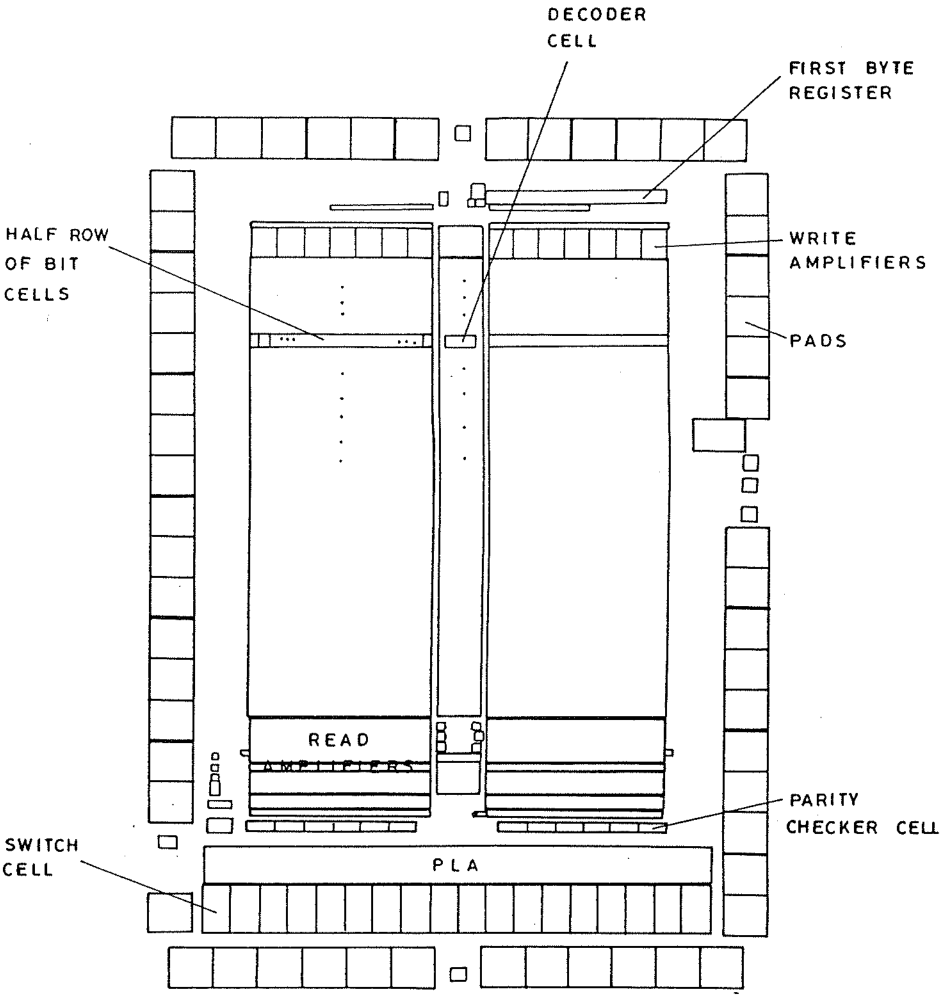


Figure 6.5: BiCMOS Fault-Tolerant Interface Chip Hardware Layout Plan

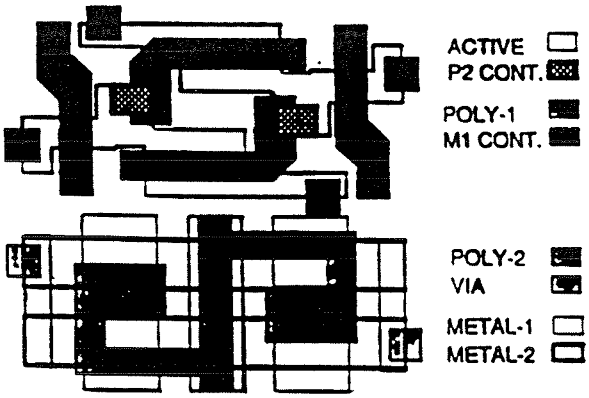


Figure 6.6: BiCMOS SRAM Cell Layout

The FST is realized as a SRAM based on the use of a static six transistor BiCMOS SRAM cell. The FST is laid out to store 256 different Memory blocks. It is reduced to main memory with a data word length of 16 bits and a additional parity bit for each of its two bytes. As a result, it is possible to encode the resulting 16 different possible bit positions of a bit cell with a word cell. It is assumed that p = 2 spare bit slices are added to the main memory. So, each entry has to comprise 2 Subfields, each Subfield consisting of a mark bit MB and a fault bit P section BP.

The 256 entries of 16 bits are physically implemented [10] as a cell array of 64 rows, where each row has 64 bit cells. For each subsequent group of 4 bit positions one I/O is provided. The I/O lines, as well as the power supply lines VDD and GND, are laid out vertically to the row of bit cells. In the middle a column of 64 decoder cells is arranged.

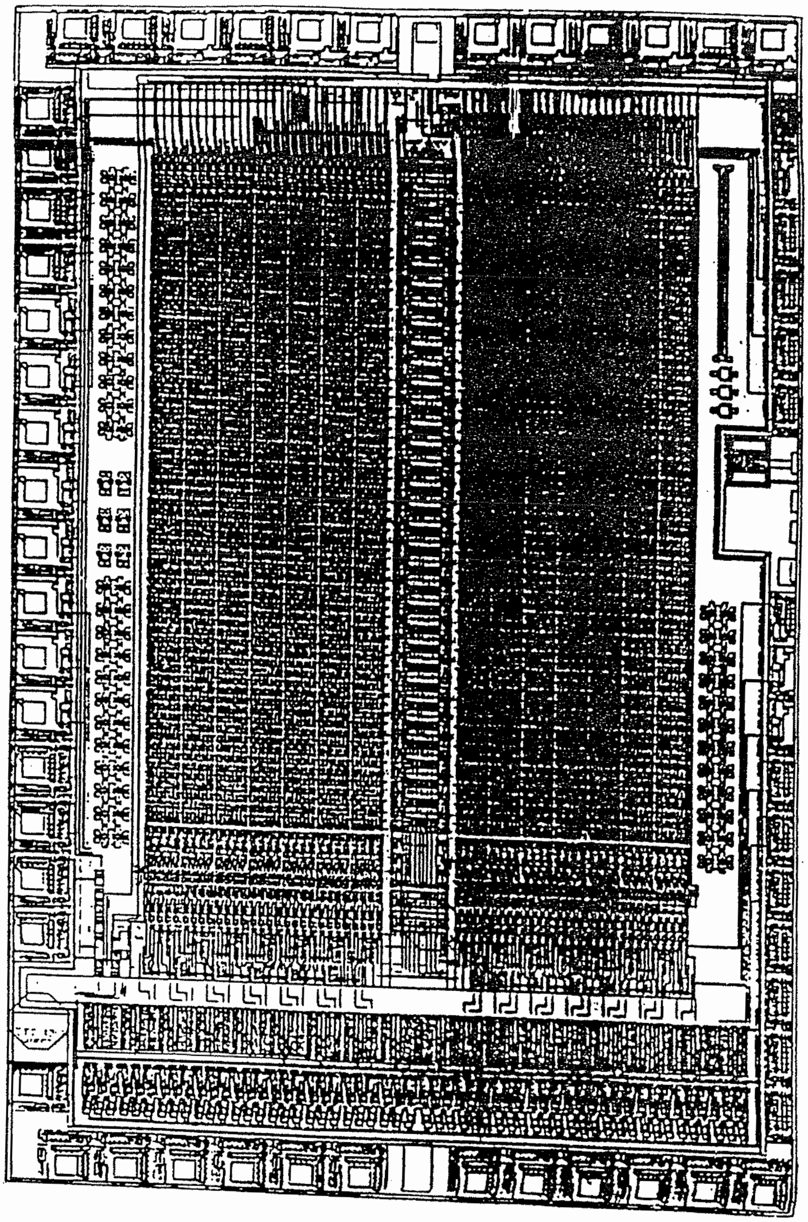


Figure 6.7: BiCMOS Fault-Tolerant SRAM Interface Chip Layout

The BiCMOS SRAM performance parameters: access time, noise margin, yield, and reliability have been analyzed to validate the proposed dynamic redundancy schemes.

A. Access time

One of the key issue for design of mega bit BiCMOS SRAM is how to achieve fast access time without increasing the chip size. Two design approaches are considered for this purpose. First a BiCMOS bit-line sense amplifier, and second is direct connection with the CMOS current mirror circuit and the BiCMOS differential amplifier for the bit-line sensing. The second design approach can not be employed because of slow bit-line sensing operation. Therefore the BiCMOS bit-line sense amplifier featured by the fast access time has been optimized for reducing the chip area penalty.

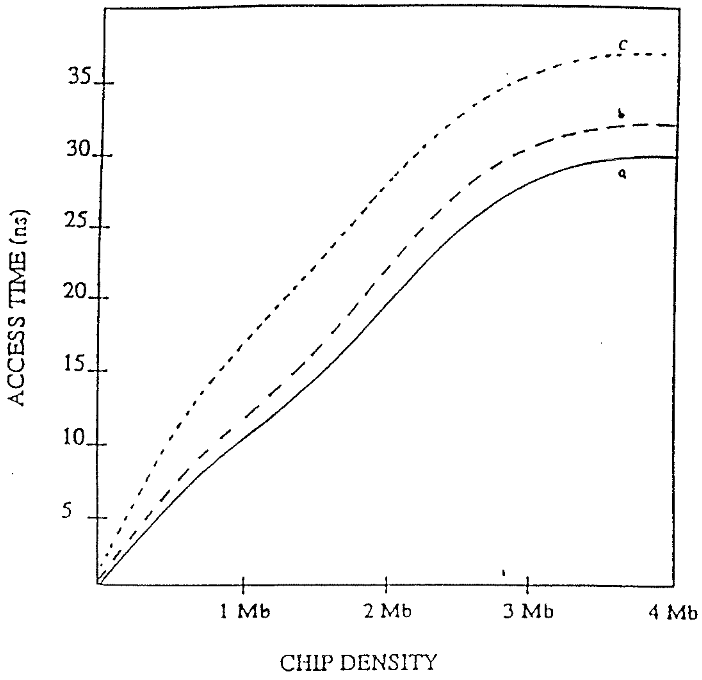


Figure 6.8 BiCMOS Mb SRAM Chip Access Time Comparison

Notations used is as follows:

(a) ----------- Non-redundant BiCMOS SRAM

(b) - - - - - - Dynamic reconfiguration schemes [Rayapati & Kaminska]

(c) .... Conventional redundancy methods [Stiffler, Sarrazin & Malek]

The write cycle time and read cycle time will be considerably reduced compared with the CMOS SRAM, because the BiCMOS SRAM peripheral circuit is used for reducing the delay time. By using the BiCMOS buffer with large current drivability, the large load capacitance of address and wordline wiring can be successfully driven in short time. Based on the BiCMOS SRAM chip simulation results, the access time improvement with respect to CMOS SRAMs[3,4] of about 35% is achieved. Access time improvement results comparison is presented in Figure 6.8. The reduction of wiring resistance and capacitance for wordlines and sense amplifiers driving lines is a special design feature, which improves access time. From this estimation, the BiCMOS technology is verified to be more appropriate for use in the mega bit SRAMs.

B. Noise margin

Noise margin is the difference between the worst case output level of driving SRAM device and the worst case input level at which the receiving device no longer recognizes the input as the intended high or low logic level. Both static and dynamic noise margins are of concern for high speed mega bit BiCMOS SRAMs. Static noise margins define the low-frequency safety margins built into the BiCMOS SRAM input-output level. Dynamic noise margin relates to the sensitivity of the BiCMOS SRAM to noise spikes.

Based on the actual loaded input-output levels of fault -tolerant BiCMOS interface chip design, the actual low level noise margin is about 300 mV and the high-level noise margin is 500 mV. BiCMOS SRAM device tend to react to pulses as narrow as 2 or 3 ns. Thus, narrow ground bounce and cross coupling spikes are of great concern for mega bit BiCMOS SRAMs.

C. Chip area

The BiCMOS bit-line sense amplifiers and buffer circuits optimization process result in significant silicon area reduction. All BiCMOS peripheral circuits employed in the fault-tolerant interface chip, reduced the overall chip area about 25%. The mega bit BiCMOS chip area comparison results are presented in Figure 6.9.

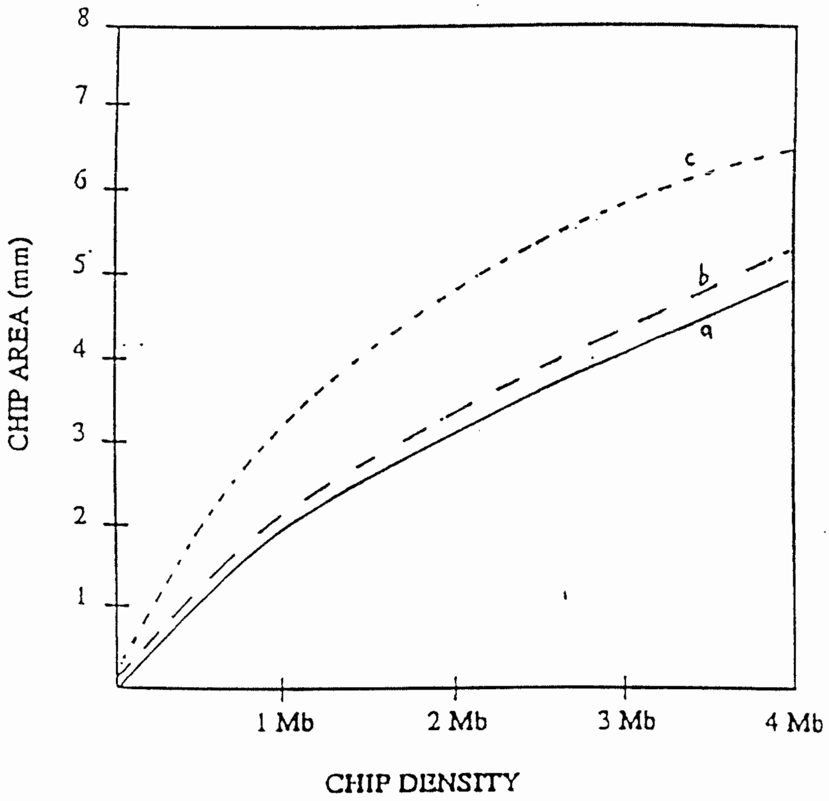


Figure 6.9: BiCMOS Mb SRAM Chip Area Comparison

Notations used is as follows:

(a) ----------- Non-redundant BiCMOS SRAM

(b) - - - - - -Dynamic reconfiguration schemes [Rayapati & Kaminska]

(c) .... Conventional redundancy methods [Stiffler, Sarrazin & Malek]

Note also that the proposed dynamic redundancy scheme not only reduced the area, but also provides significant cost savings at the system level.

D. Yield

Suppose the system consists of n 1-Mb BiCMOS SRAMs and requires the capacity of m 1-Mb SRAMs and that the yield of each 1-Mb BiCMOS SRAM chip is p. The yield of the memory system using a conventional repair scheme is calculated by summing the combinations of pass and fail chip possibilities and can be expressed as follows:

$$P = \sum_{i=0}^{n-m} nC_i p^{n-i} (1-p)^i$$

The yield of the BiCMOS memory system with dynamic redundancy is calculated by summing of the combinations of pass and fail rows/columns within the chip by the following formula:

$$P = \sum_{i=0}^{n-m} [\, nC_i p^{n-i} (1-p)^i \sum_{j=0}^{(n-i)(S-m)\times S} (n-i)\times SC_j p^{(n-i)\times(S-j)} (1-p)^j]$$

Where P is a memory system yield, p is the interface chip yield, S chip density factor, C complexity factor, n number of rows, m number of columns, i and j are integer variables. BiCMOS mega bit SRAM chip yield has been calculated using the above derived formula. Yield improvement results are presented in Figure 6.10 where the improvement about 10 % is achieved in comparison with the conventional redundancy methods.

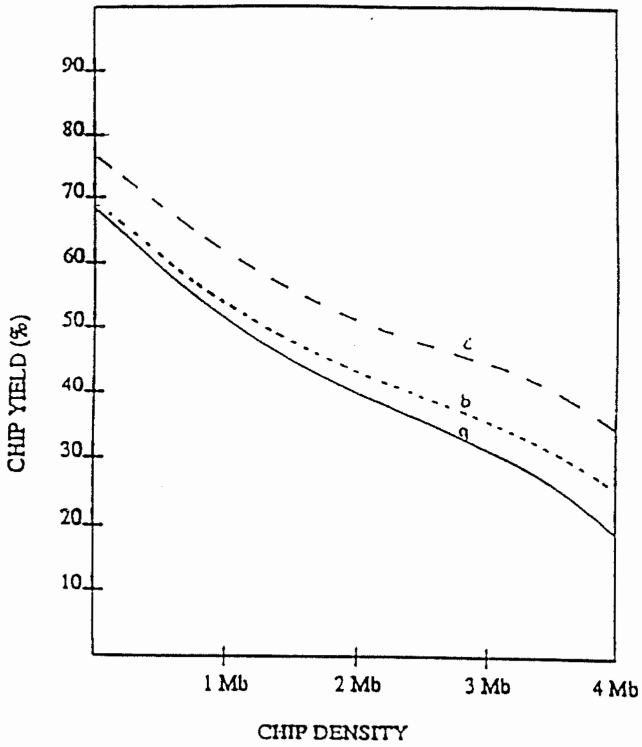


Figure 6.10: BiCMOS Mb SRAM Chip Yield Comparison

Notations used is as follows:

(a) ----------- Non-redundant BiCMOS SRAM

(b) - - - - - - Dynamic reconfiguration schemes [Rayapati & Kaminska]

(c) .... Conventional redundancy methods [Stiffler, Sarrazin & Malek]

E. Reliability

The BiCMOS mega bit SRAM device reliability is a major concern for all applications. In order to improve reliability two dynamic redundancy schemes for BiCMOS mega bit SRAM are presented in this paper. The proposed schemes provide the capability of built-in fault-tolerance, fault detection, and fault isolation within the SRAM chip. For all real life and life critical applications the fault-tolerant BiCMOS interface chip provides predicted reliability of 0.9995 for 15 year mission. BiCMOS mega bit SRAM chip reliability comparison results are presented in the Figure 6.11.

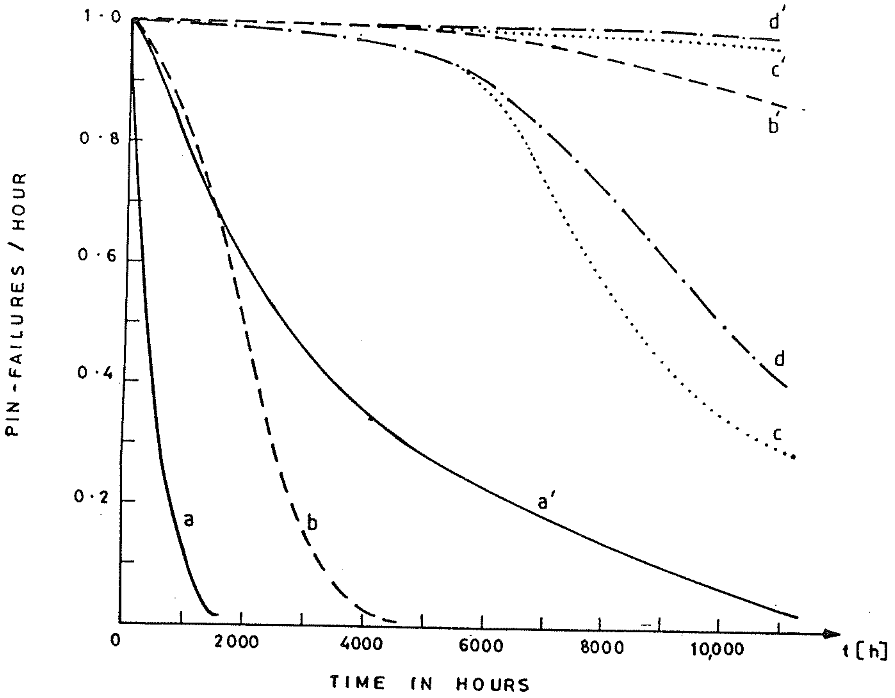


Figure 6.11: BiCMOS 1-Mb SRAM Chip Reliability Comparison

Case (1) TR = 10^{11} and PIN = 10^{7} failures per hour

Case(2) TR = 10^{12} and PIN = 10^{9} failures per hour

Notations used is as follows:

(a) ----------- Non-redundant BiCMOS SRAM

(b) - - - - - - One bit error correction scheme

(c) .... Dynamic reconfiguration scheme with p=4

(d) _._._._._ Graceful degradation scheme with p=4

6.5 BiCMOS SRAM GRACEFUL DEGRADATION SCHEME

One great advantage of dynamic standby system is that the hardware for reconfiguration and fault status storage can be modularly added to the interface of SRAM chip. This additional effort can be implemented independently of the internal structure of the main memory. Spare slices are provided to the BiCMOS SRAM. This reconfiguration method identifies fault conditions and provides fault masking and fault isolation capabilities within the chip. The BiCMOS SRAM has soft failure tolerance capability. If such a dynamic reconfiguration scheme is not implemented, then the SRAM won't be able to provide soft failure tolerance capability, then the storage capacity of BiCMOS SRAM has to be "gracefully degraded". Reliability of 1 Mb BiCMOS SRAM performance comparison is provided to validate the proposed dynamic reconfiguration schemes.

Replacement of defective memory cells can be done by adopting the strategy that the unit of reconfiguration is a single word cell. A different method would be adequate here is the use of entire memory blocks as replaceable with an alternative dynamic standby system. Logically it deactivates only defective bit slice modules, the basic idea is to store the fault coincident bits of a defective memory block. Using this dynamic redundancy approach, we can effectively manage very high density memory allocation within the computer system.

6.6 CONCLUSIONS

Two schemes of dynamic redundancy are introduced which allow the treatment of SRAM chip level faults at the interface of the main memory of the computer system. The proposed dynamic redundancy schemes are:

1) The standby system approach.

2) The graceful degradation (soft failure technique).

ADVANTAGES:

1) Flexibility is available in the SRAM chip design in order to obtain smaller units of reconfiguration, the size of the bit slice modules is reduced significantly.

2) This method can be effectively applied to BiCMOS PROMs, ROMs, DRAMs, and sequential index memories.

3) Significant access time improvement achievable compared to CMOS SRAMs using conventional methods.

4) The size of the reconfiguration can be reduced to the demands of the user applications.

5) High reliability of BiCMOS SRAM can be achieved. System level fault-tolerance can be achieved with low cost.

These results are very useful to improve system speed, reduce power consumption and improve access time of SRAMs. The BiCMOS SRAM access time improvement of about 35%, chip area of 25%, and chip yield of 10% are achieved respectively using the proposed dynamic reconfiguration schemes, as compared to the conventional methods.

REFERENCES

[1] J. LOSQ, "Influence of fault-detection and switching mechanisms on the reliability of stand-by systems".Proceedings of the International Symposium on Fault-tolerant Computing, 1975, pp. 81-86.

[2] W. W. PETERSON, "Error Correcting Codes",MIT Press, Cambridge, 1972.

[3] J. J. STIFFLER, "Error correcting Coding techniques for Random Access Memories", IEEE Transactions on Computers, C-27, 1978, 256-531.

[4] D.B. SARRAZIN and M. MALEK, "Fault-tolerant, Semiconductor memories", Computer Magazine, pp. 49-56, August 1984.

[5] K.E. GROSSPIETSCH, J. KAISER and E. NETT, "A dynamic reconfiguration scheme for random access memories",Digital Processes, 1980, pp. 257-270.

[6] N.G. EINSPRUCH, "VLSI Electronics Microstucture Science", Vol.3, Academic press, 1982, pp. 1-24.

[7] K. YOKOMIZO and K. NAITO, "Design Techniques for high-throughput BiCMOS Self-timed SRAMs", IEEE Journal of Solid State Circuits, Vol-28, No.4, April 1993, pp. 484-489.

[8] B.BASTANI et al, "1.0 um BiCMOS technology for high speed 256K SRAMs", VLSI Technology Symposium proceedings, 1987, P.41-45.

[9] K.NAKAMURA et al., " A 6-ns ECL 100K I/O and 8-ns 3.3V TTL I/O 4-MB BiCMOS SRAM", IEEE Jouranal of Solid State circuits, Vol 276, No 11, 1992, pp. 1504-1510.

[10] R.M. TANNER, "Fault-tolerant 256K memory designs,IEEE Transactions on Computers, C-33, 1984, pp.314-322.

[11] B.R. BORGERSON and R.F. FREITAS, " A reliability model for gracefully degrading and stand-by sparing systems",IEEE Transactions on Computers, C-24, 1975, pp. 517-525.

CHAPTER 7

ROOT CAUSE TEST TECHNIQUE FOR MEGA BIT CMOS SRAM AND RELIABILITY PERFORMANCE EVALUATION

Article by: Venkatapathi N Rayapati and Bozena Kaminska,

Accepted for publication in International Journal of Microelectronics and Reliability, Pergamon press Ltd, Oxford, U.K.

ABSTRACT

A rapid root cause test technique for mega bit CMOS SRAM chip proposed in this paper. The proposed test technique uses two complementary techniques for mega bit SRAM fault localization and fault analysis. This method is very useful in the CAD environment. The SRAM chip functional test allows the detection of permanent/intermittent faults during design phase, which could cause the SRAM chip functional failure. These faults are stuck-at-1/ stuck-at-0, coupling faults, and bridging faults based on physical failures like metallization shorts and capacitive coupling. Then a laser beam integrated in an automatic test equipment and provides an accurate localization of SRAM chip failures using memory chip layout. This paper demonstrates that the association of an electrical tester with integrated laser beam makes easier for the localization failures in the SRAM chip and consequently reduces the test cost. A reliability model is presented based on design and manufacturing faults accounted separately. Multi-layer interconnect design impact on the SRAM chip reliability has been addressed through this model. The overall SRAM reliability can be improved by about 50% by implementing root cause test feedback and fixing failures during design and prototype phases. The proposed techniquie can be applied to DRAM's, EPROM's and ASIC's.

7.1 INTRODUCTION

Root cause identification test is considered to be a critical step in mega bit CMOS Static Random Access Memories (SRAMs) Yield and Reliability improvement. Root cause test technique provides quick and accurate feedback on what is wrong in a mega bit SRAM device and leads to the right fixes and improvements. Root cause testing of mega bit SRAMs has become a bottleneck issue in the industry, due to SRAM cell density, multi-layer interconnects and device scaling problems. The importance of testability of mega bit CMOS SRAM chip is intimately connected with the fact that the test cost is continuously rising as a percentage of the SRAM cost. Today mega bit CMOS SRAM chip designers must conceive circuits which can be efficiently tested in order to reduce escalating test costs [1].

Failure detection becomes very hard with new types of failures owing to the increasing complexity of mega bit SRAMs. Analysis of intermittent failures is particularly challenging because the electrical failure must be reproduced in the laboratory for accurate diagnosis. One type of defect that can produce intermittent failure is open metal interconnect, for which the failure symptoms can be affected by temperature, voltage and frequency [2]. Testing of mega bit CMOS SRAM chip allows detection of permanent faults which could cause the SRAM chip function incorrectly. These faults are stuck at 1 or 0, based on the physical failures like metallization shorts and capacitive coupling. The root cause of these failure is poor step coverage of the metallization over an oxide step or oxide breakdown/interconnect defects. The mega bit SRAM functional testing alone would not be able to localize these faults. Therefore, there is a strong need to use integrated laser beam to identify fault sites in the SRAM chip.

Failed SRAM devices are brought to the Lab for root cause identification from many sources. They can be internal in process failures, reliability test failures, failures from ESD testing, customer returns or even good parts for analysis of certain variable parameters. From an understanding of the failure and the SRAM architecture of the chip we can deduce where the fault is occurring. Further isolation of the fault can be obtained by using photo emission microscopy, liquid crystal analysis [3], voltage contrast [4 - 5] or internal probe [6 - 7]. These techniques are time consuming and are not cost effective.

In this paper a rapid root cause test technique for mega bit SRAM chip presented. The proposed test technique uses two complementary techniques for mega bit SRAMs failure localization and analysis. A mega bit SRAM chip root cause test technique is presented in Section 7.2. Root cause test technique uses functional testing approach and integrated laser beam scanning to localize faults within the SRAM chip. Experimental validation of SRAM root cause test technique is presented in Section 7.3. SRAM data analysis performed using the SRAM chip layout and scanning for failures. A reliability model is developed based on SRAM dynamic reconfiguration scheme, design failures, and manufacturing failures. 1-Mb SRAM chip reliability performance analysis results are presented in section 7.4.

Conclusions and future applications are provided in Section 7.5. Advantages of the proposed root cause test is presented.

7.2 MEGA BIT SRAM CHIP ROOT CAUSE TEST TECHNIQUE

The proposed SRAM chip root cause test technique uses two complementary test methods: a) External Electrical Functional Test, and (b) Internal Contactless Laser Beam Testing.

A External Electrical Functional Testing

The SRAM chip functional testing has been performed using the electrical tests shown in Figure 7.1. Testing of SRAM chip allows detection of permanent faults or intermittent faults which cause the circuit to function incorrectly. These faults are stuck at 1 or 0, based on physical SRAM device failures due to metallization shorts and captive couplings. One type of defect in the SRAM chip that can produce intermittent failure is open metal interconnect. Functional testing of mega bit SRAMs usually means testing of each SRAM cell by writing with 1's or 0's using standard memory test algorithms and reading the same data contents based on the device specification.

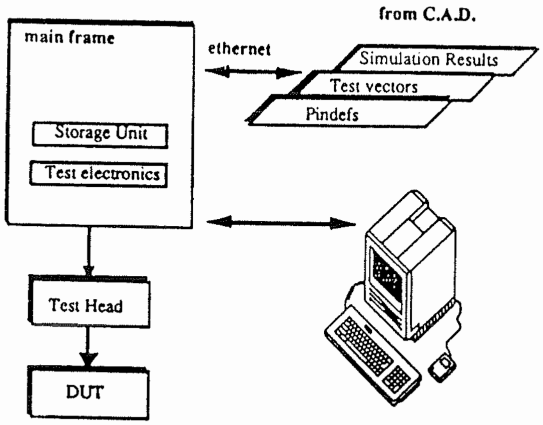


Figure 7.1: Mega bit CMOS SRAM Automated test set-up integrated with laser

SRAM chip functional testing can be performed without knowing the detailed implementation of chip architecture. Although manufacturers have detailed knowledge of the logical and parametric behaviour of the mega bit SRAM chips produced, they usually only apply specific testing of SRAMs due to economic reasons. To ensure reliable operation of mega bit SRAM chip, additional testing must be performed by the user. SRAM chip functional testing has been performed using the standard test vectors on which a failure is observed. An electrical tester is integrated into the CAD environment to generate test vectors appropriate to the SRAM chip. This approach reduces the test vector generation time and number of iterations used to modify the SRAM design. A concurrent design and test development methodology used for the mega bit SRAM chip functional testing.

Even for complex mega bit SRAM chips, the test vector generation and test sequence to cover all SRAM cells and associated circuitry within very short time period. The test vector generation time and testing time for an SRAM chip represents an important cost constraint. The functional test should cover all permanent faults and intermittent faults within the SRAM chip. Once functional faults are identified, then proceed for the second phase of testing for localizing fault locations in the SRAM chip using the architectural and layout information.

B Internal Contactless Laser Beam Testing

A non-destructive laser photoscanning technique has been used to analyze stress induced permanent failures or intermittent failures in the SRAM chip. The laser beam is integrated into the automatic test equipment and is shown in Figure 7.1, which provides an accurate localization of the SRAM chip failures. A laser photoscan system for the ATE is shown in Figure 7.2. A He-Ne laser spot is scanned in the roaster pattern on the device under test using a standard metallurgical microscope and commercially available scan mirrors. The laser scanning beam generates hot-electron pair in a volume about ten microns in radius and less than eight micron deep. Both dimensions depend on laser wavelength. Carriers generated within about a diffusion length of reverse biased wells or drain junction in device power supply current. The objective is to analyze stress related permanent or intermittent failure localization.

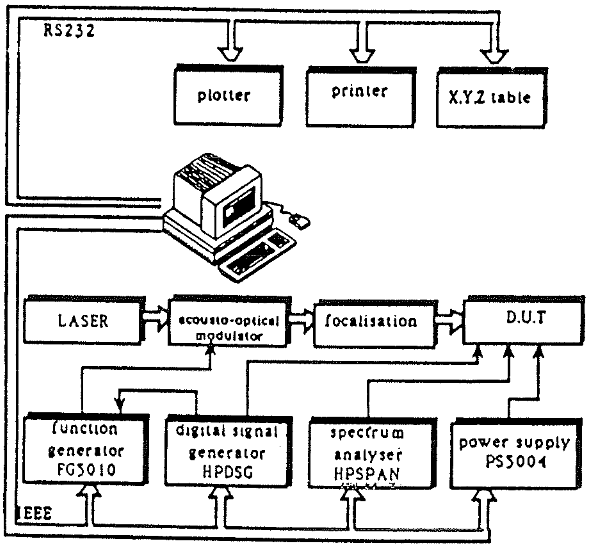


Figure 7.2: Mega Bit SRAM a automated test equipment integrated with photoscan system

A functional test has been performed using the electrical tester shown in Figure 7.1, which results in permanent or intermittent failures identification in the SRAM chip. The faults will indicate either pass or fail condition for all memory cells within the SRAM chip. On the schematic representation of the SRAM chip, different blocks can have stuck at 1 or 0 faults.

On the SRAM chip layout by relation to the source of the chip architecture and establish logic state errors within the device. The electrical tester, by means of a functional test, allows a briefly localization in the layout of sensible areas to be scanned with laser beam tester.

The use of a laser beam test integrated in automatic test equipment allows logic state analysis of the SRAM cells inside the chip. Analysis of photo-induced currants generated in the laser beam silicon interaction and detected in the power line of the device under test, permits to find stuck at faults or intermittent faults within the device.

7.3 EXPERIMENTAL VALIDATION OF THE SRAM CHIP ROOT CAUSE TEST TECHNIQUE

A functional test is designed to detect permanent or intermittent stuck-at faults that cause the SRAM chip to function incorrectly. These faults are stuck at 1 or 0, based on physical SRAM device failures due to metallization shorts, opens and capacitive couplings. A straight-forward functional test algorithm is presented in this paper for the SRAM chip [7]. This algorithm can be modified depending on the complexity of the SRAM chip.

SRAM MULTI-BIT TEST ALGORITHM:

1. Initialize the cell array to 1.
2. Call the test procedure x (0, 0, 0).
3. For each even cell (i,j) except (0,0) do cell the test procedure x (i,j,0).
4. Initialize the cell array to 0.
5. For each even cell (i,j) except (0,0) do cell the test procedure x (i,s,1).

Test procedure x(i,j,k): for cell (i,j), for states will cell value k.

```
Step 1.     calculate P and Q:
            check $P \geq Q$
```

```
              weight = parameter 2 - 1.
Step   2.     For soft error free cell 1 = parameter 2 - 1 to 0 $d_o$
              for soft error free cell 2 = parameter 2 -1 to 0 $d_o$
                          cell-address;
              if cell 2 $\leq$ weight then read (K); write (K)
              else read (K); (write (K).
Step   3.     If weight < 0 then stop.
Step   4.     Repeat step 2, with K in place of K, and K in place of
              K.
Step   5.     For soft error free cell 2 = weight;
                    for free cell 1 = parameters 1 - 1 to 0 $d_o$
                    cell-address;
                    read (K); write (K);
Step   6.     Weight -;
              go to step 2.
        Calculate  P  and  Q:
              $P = \min \{n, \max \{2^x\}\}$; $i \bmod 2^x = 0$;
        $Q = \min \{n, \max \{2^x\}\}$; $j \bmod 2^x = 0$;
        cell-address:
        Rou = $[i - P/2 + (P/2 + free - i) \bmod P] \bmod n$;
        Column = $[j - Q/2 + (Q/2 + free - j) \bmod Q] \bmod n$;
              check  $P \geq Q$
```

```
If  $P \geq Q$ then        else
parameter 1 = P;           parameter 1 = Q;
parameter 2 = Q;           parameter 2 = P;
soft error free cell 1 = free-i; soft error free cell 1 = free - j;
soft error free cell 2 = free-j; soft error free cell = free - i
```

This algorithm has the capability to detect multi-bit errors (fine cells) as well as neighbour-bit errors in 1-Mb SRAMs. Multi-bit test algorithm has more advantages than other standard test algorithms.

A 1 - Mb SRAM chip is tested using an automated test set-up shown in Figure 7.1. A software interface has been developed between the CAD environment and the electrical tester. SRAM chip pin connections, test vectors and simulation results (i.e. test patterns) for the circuit are transferred and adapted to the tester as shown in Figure 7.2. SRAM chip functional test has been performed using the test vectors from the CAD station as shown in Figure 7.1. An output fail flag is mapped out in the tester on the test vector number N for the Pin Wi of the SRAM chip.

	Input	Output
Vector	N/001100	000101/T01 FAIL Wi

A parametric test shows at this time that the output Wi on the N vector test, is at a high level "1" instead of low level "0". Now, we observed a bit failure in the SRAM chip. The different SRAM blocks can have stuck at faults 0's or 1's within the SRAM chip.

SRAM chip layout details are required to perform laser beam scanning, which will result in logic state analysis. SRAM chip logic state analysis has been performed using the integrated laser beam tester shown in Figure 7.2. In the 1 - Mb SRAM chip layout partitioning and interconnect routing information is used to establish test pattern's for the laser beam tester. Each pad Wi, we establish a logic sequence between the output panel and the laser beam scan path within the SRAM chip.

SRAM chip logic state analysis has been performed using the laser beam tester. The SRAM chip logic state analysis establishes the relationship between the address, control and data output. From the functional test data it is evident that the SRAM chip has few stuck at 1's or 0's type of faults identified within the SRAM chip. The test results are given below:

Test Vector	**Input Data**	**Output Data**
N	1010...	HLHH.../PASS
N + 1	1011...	HLHH*.../FAIL on Wi

N + 2	1011...	HLHH.../PA SS
N + 3	1011...	HH*HL.../FA IL on Wj
...	....	....
N + m	1111...	HHHH.../PA SS STOP

This test result incriminates logic operations in the laser beam scan path. Thus, the integrated laser beam tester by means of a functional test, allows a briefly localization in the SRAM chip layout sensible areas have been scanned for failure identification.

The use of a laser beam tester integrated in automatic test equipment allows logic state analysis of any operator inside the SRAM chip. Analysis of photo induced current generated in the laser-beam-silicon interactions and detected in the power line of the SRAM device under test, permits to find stuck at 0's or 1's type of faults.

The basic principle is illustrated using CMOS inverter circuit is shown in Figure 7.3. Figure 7.3 (a) and 7.3(c) illustrate the junctions which are expected to collect photocurrent for the two electrical states of a CMOS inverter, which is used

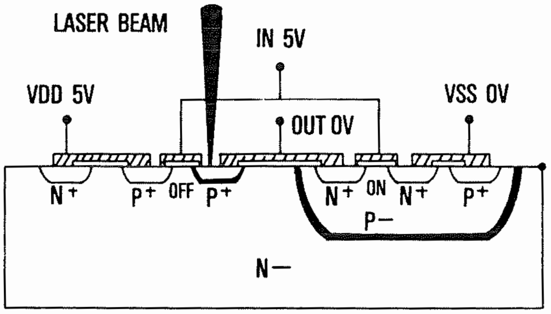


Figure 7.3(a): CMOS Invertor cross section. Input high. P-channel OFF, N-channel ON

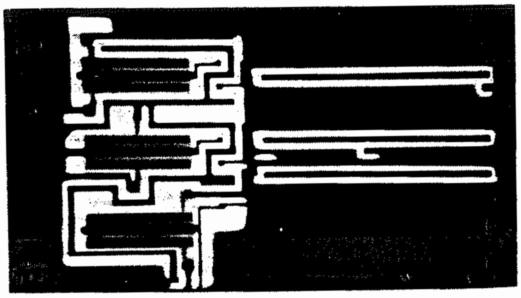


Figure 7.3 (b): CMOS Invertor photoscan image. Inputs high.

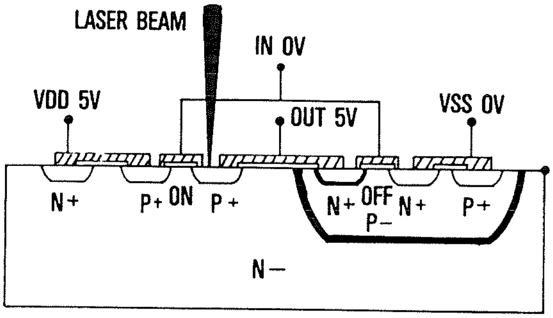


Figure 7.3 (c): CMOS Invertor cross section. Input Low. P-channel ON, N-channel OFF

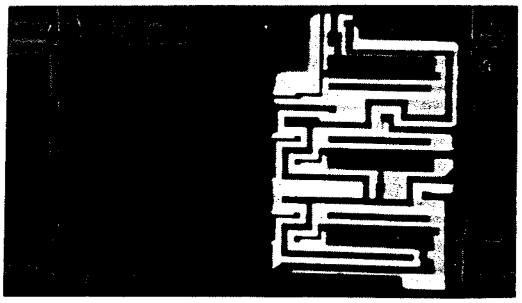


Figure 7.3 (d): CMOS Invertor photoscan image. Inputs Low.

Laser beam tester scanning and modulating the beam intensity in proportion to the device current. When the inverter is in a high impedance state, a photo-induced current flows out-side of the circuit which is recorded by the laser beam tester. When the inverter is in a low impedance state the photo-induced current flows only inside the circuit and can not be detected outside. Photocurrent and commutation current are split by a spectrum analyzer and detection of a photo-current corresponds with a high output level for the CMOS inverter circuit. For the CMOS inverter, similar results are obtained, a photo-current is noticed when the output is at a low logic level.

The 1 - Mb CMOS SRAM chip functional block diagram shown in Figure 7.4 is tested using the integrated laser beam tester. The test patterns are applied to the SRAM chip tester through the CAD system. A micro instruction code prepared to STOP on the vector number N, where a fail flag appeared on the Wi output. This specific vector, any logic operator from the pad Wi is tested with a laser beam scanning technique. With this technique we establish the SRAM cell columns/rows responsible for the failure observed by the electrical tester. A stuck at 1 fault is found at the output of the SRAM chip due to interconnect metallization failure. The CMOS SRAM laser beam reflected light is shown in Figure 7.5. The normal photoscan image for the 1 - Mb SRAM chip is shown in Figure 7.6.

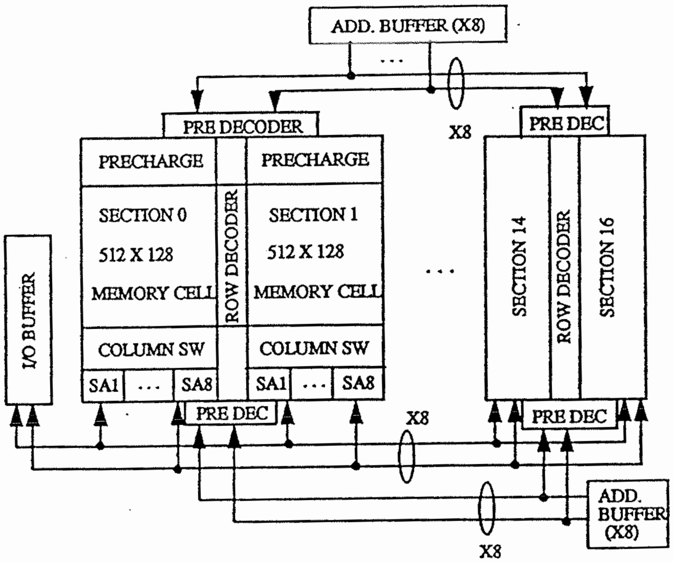


Figure 7.4: 1-Mb SRAM chip functional block diagram

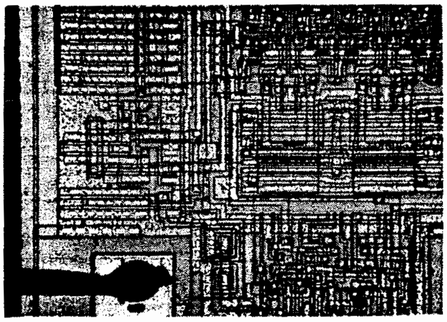


Figure 7.5: 1-Mb CMOS SRAM chip Laser beam reflected light showing interconnect failures

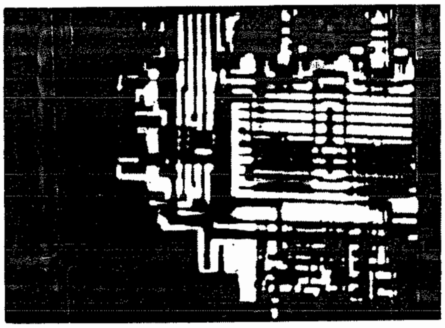


Figure 7.6: 1-Mb CMOS SRAM normal Laser beam photoscan image

A correlation has been established between the functional faults and the root-causes using the integrated laser beam testing method [8]. The SRAM chip test results are shown in Table 7.1.

Table 7.1: 1 - Mb SRAM Chip Results Summary

Electrical Test Data For **1 - Bit Failures**	**Electrical Test Data For** **2 - Bit Failures**
0 0 - x - 0 0 - x - 0	0 0 - x - 0 0 - x - 0
0 - x - 0 0 *x - 0 0 - x	0 - x - 0 0 x - 0 0 - x
0 0 - x - F 0 - x - 0	0 F - * - F 0 - x - 0

0 - x - 0 0 - x - 0 0 - x		0 - x - 0 0 - x - 0 0 - x	
0	good SRAM cell	0	good SRAM cell
-	good peripheral circuitry	-	good peripheral circuits
F	failed SRAM cell	F	failed SRAM cell
*	defective peripheral circuits	*	defective common bit line cell contacts
x	common bit line cell contact	x	common bit line cell contact

SRAM chip fault localization process is discussed based on the integrated laser beam testing results. The CMOS memory cell is shown in Figure 7.7. It is a six-transistor cell, consisting of two cross-coupled inverters and two access transistors. Drain to gate connections between the inverter pairs are through polysilicon feedback resistors and common drain connections are metal. The laser beam scanning image of the following memory cell written to the "0" logic state is shown in Figure 7.8. The capacitance coupling image of the failed memory cell while writing the "1" state is shown in Figure 7.9. Through this approach permanent/intermittent stuck at 0 or 1 faults are localized within the SRAM chip. The a root cause for each failure localization deduced based on SRAM chip architecture and layout information analyzed through integrated laser beam testing method. The proposed root cause test technique is very effective and CAD based tool to eliminate failures in the early design phase, manufacturing phase and improve reliability and yield of mega bit SRAMs.

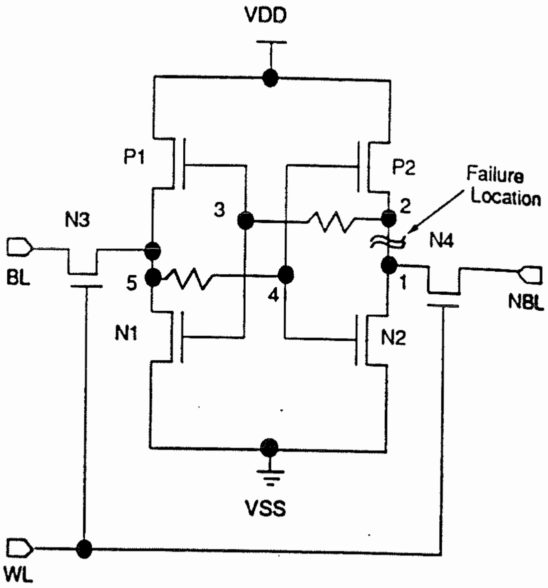


Figure 7.7: CMOS Memory cell circuit diagram

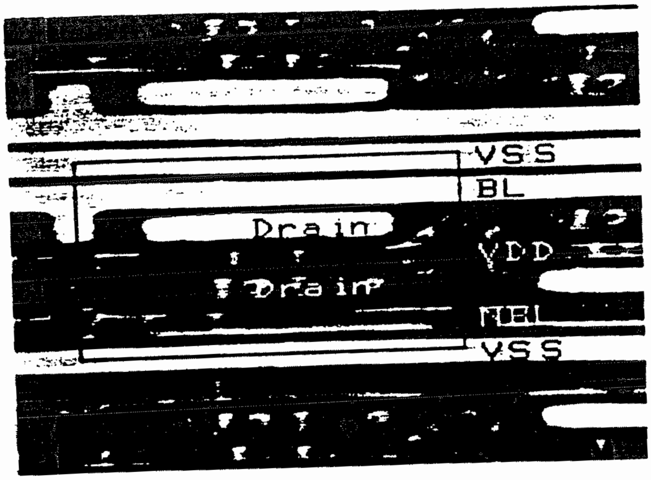


Figure 7.8: CMOS SRAM cell stuck-at-0 fault identification through Laser beam scanning image

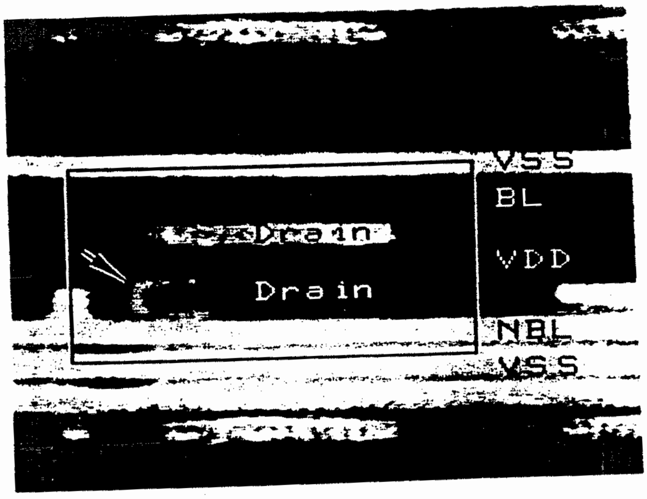


Figure 7.9: CMOS SRAM cell Stuck-at-1 fault identification through Laser beam scanning image

7.4 RELIABILITY MODEL FOR A 1 Mb SRAM

The 1-Mb CMOS SRAM functional block diagram is shown in Figure 7.4. The memory cell array of 512X2048 cells and is divided into 32 blocks, each of which consists of 512X64 cells. Each cell array block contains column and word decoders. The adjacent blocks have a set of local sense amplifiers and Write circuit for four input/output operation. Bonding pads are placed on four sides of the chip to make it easy to mount the chip in a small PLCC package. Therefore, the peripheral circuits such as address buffers, predecoders, control buffers, I/O buffers, and redundancy circuits are also placed on the four sides of the chip.

The 1-Mb CMOS SRAM chip is considered as a series system for reliability modelling. Every element in the 1-Mb SRAM chip is required to function correctly to fulfill the overall SRAM functionality. In this model the failures related to interconnects, design failures (i.e., Read/Write timing faults, redundancy switch over failures, logic failures), and manufacturing related failures (i.e., package failures, die assembly failures, assembly process failures etc.) are addressed. The proposed reliability model considers: dynamic reconfiguration, interconnects, design, and manufacturing related failures, which can affect the 1-Mb CMOS SRAM chip field reliability performance. A typical failure distribution for 1-Mb SRAM chip is shown in Figure 7.10.

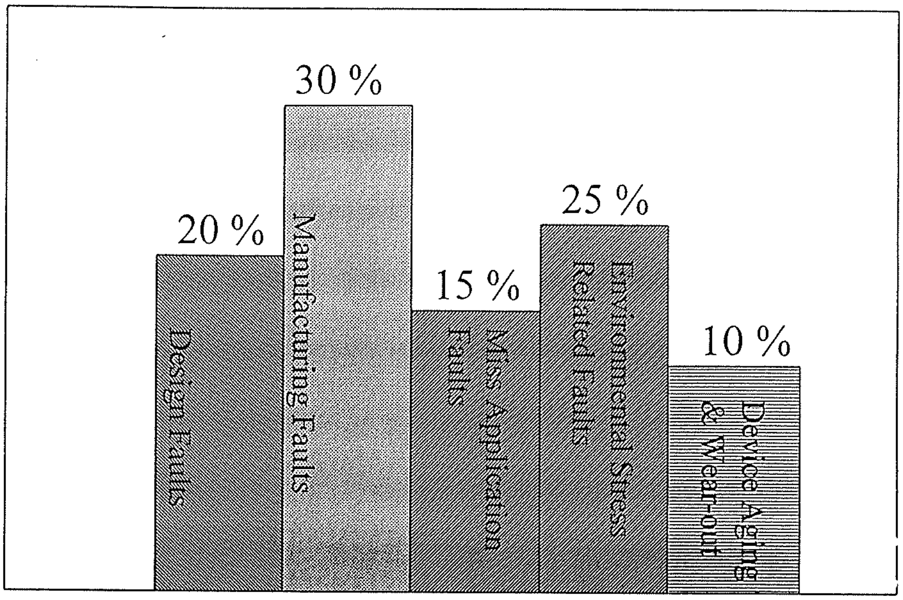


Figure 7.10: 1-Mb SRAM typical field failure distribution

The reliability model is developed based on the dynamic reconfiguration scheme implementation for 1-Mb CMOS SRAM chip. Reliability of the mega bit SRAM with the dynamic reconfiguration scheme having S standby spare words is given as follows:

$$R = 1 - (1 - R_M)^{S+1}R_i \tag{1}$$

where R_M denotes the reliability of the memory device active word or standby spare word in the 1 Mb SRAM, and S represents the number of standby spare words per cell array. 1-Mb SRAM chip design is based on the three level multilayer interconnect architecture. Interconnects limit the density of mega bit SRAM and speed performance. The interconnect reliability contribution factor R_i is taken into account in this model.

The arrangement of memory words in each cell array represents a hot-standby system. The hot-standby memory system is to operate two memory words in parallel with either memory word acting as a standby spare word. A failure detection circuit continuously checks each word in each cell array of the memory system. Failure detection circuit hardware is based on the choice of the error correcting codes. If a failure is detected in a memory word of any cell array within the chip, on-line reconfiguration switching mechanism isolates the failed word and automatically switches to the standby spare word. The reliability of 1 Mb CMOS SRAM chip can be expressed as follows:

$$R = [R_M^2 R_i^2 + 2F_C R_M R_d R_m (1-R_M R_i)] R_S R_C \qquad (2)$$

where R_M denotes the reliability of each memory cell array within the chip; R_C stands for the reliability of failure detection circuit hardware; R_S is the reliability of the reconfiguration switching mechanism; F_C represents the fault coverage correction factor. R_d is design reliability factor (i.e., design robustness assurance factor implemented through front end design process); R_m is the manufacturing process reliability factor (i.e., package failures, die assembly failures, assembly process failures etc). Memory chip failures, and interconnect failures still dominate the overall reliability the 1-Mb SRAM chip. This model gives a good over view of field reliability performance impact due to design and manufacturing of the SRAM.

Reliability improvement using the dynamic reconfiguration schemes for 1 Mb SRAMs, design failures identification through root cause test and implementing design fixes comparison is shown in Figure 7.11. The results demonstrate that the integrated design root cause test identifies about 20% of design failures and provides feedback for continuous improvement during the design cycle, which reduces design cycle time, improves reliability, and reduces cost over runs.

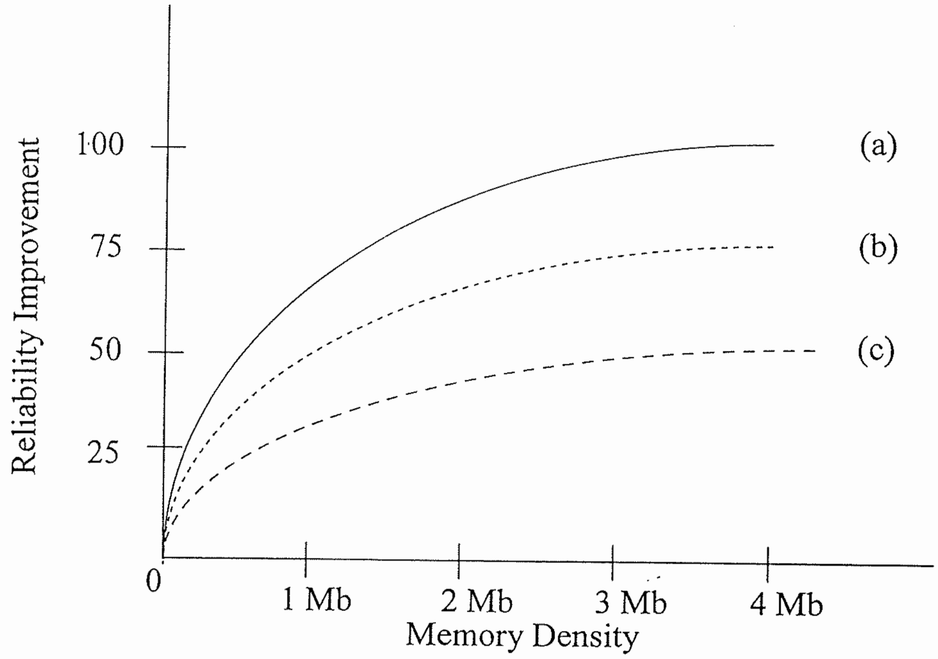


Figure 7.11: Mega Bit SRAM Design reliability improvement using the root case test feedback and fixing the design

Notations used: (a) --------- with root cause test feedback design fixes

(b) .... using dynamic reconfiguration scheme

(c) - - - - - without using root cause test feedback & reconfiguration scheme

Reliability improvement using dynamic reconfiguration schemes, and manufacturing process failures identification though the proposed root cause test technique is shown in Figure 7.12. During the prototype SRAM chip development the proposed root cause test detects about 30% of failures, which can affect SRAM chip field reliability performance.

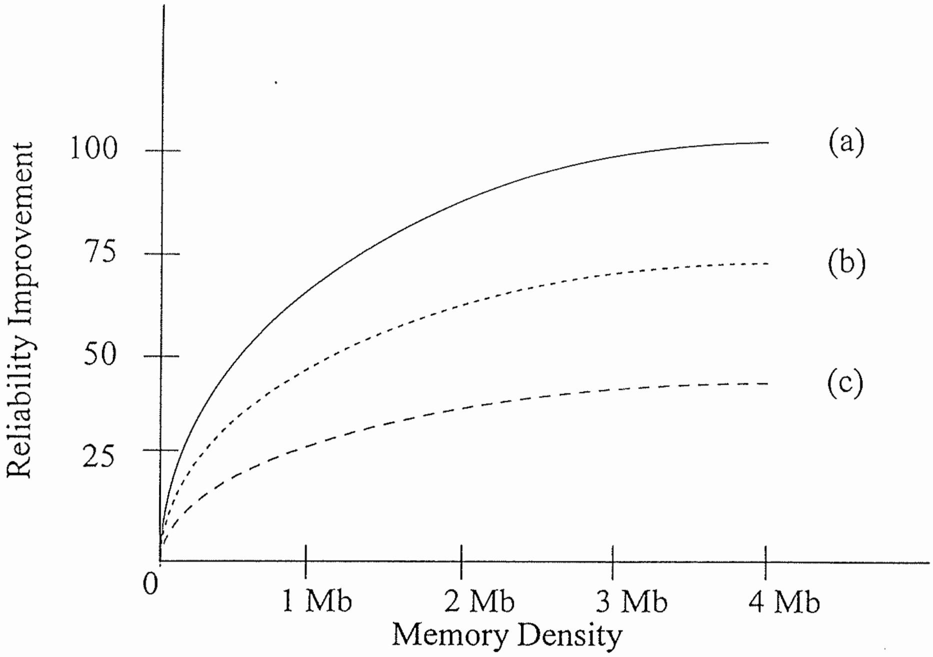


Figure 7.12: Mega Bit SRAM Manufacturing reliability improvement using the root case test feedback and fixing the design

Notations used: (a) --------- with root cause test feedback manufacturing fixes

(b) .... using dynamic reconfiguration scheme

(c) - - - - - without using root cause test feedback & reconfiguration scheme

Reliability improvement using dynamic reconfiguration schemes, root cause test feedback process for detecting design, manufacturing faults and fixing before releasing for volume production is shown in Figure 7.13. The proposed reliability model clearly demonstrates the interconnect failures, design failures, and manufacturing failure impact on 1-Mb CMOS SRAM chip filed reliability performance.

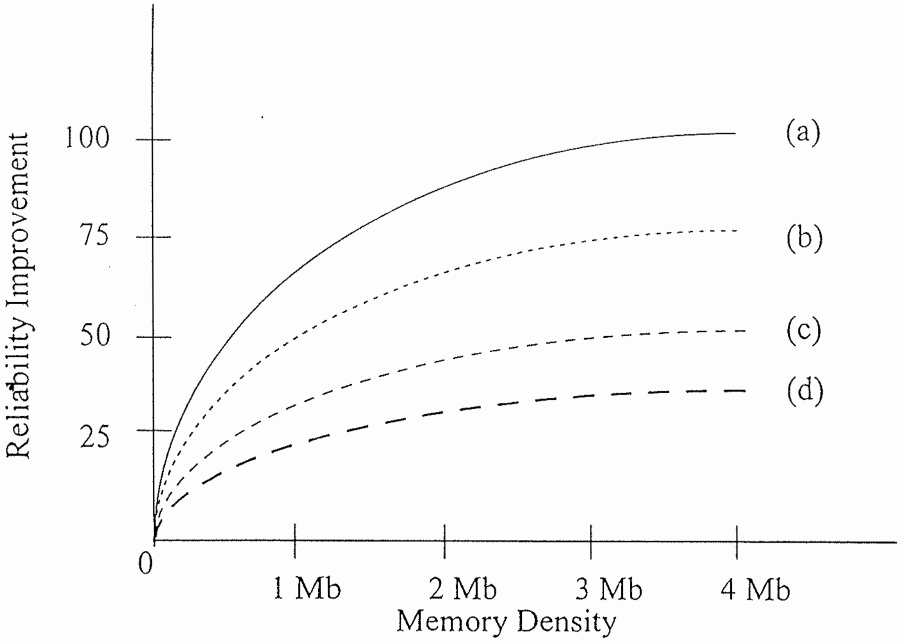


Figure 7.13: Mega Bit SRAM Reliability improvement using the root case test Technique

Notations used: (a) --------- with root cause test feedback design fixes

(b) .... using dynamic reconfiguration scheme

(c) - - - - - without using root cause test feedback & reconfiguration scheme

(d) - - - - - without redundacy and ECC

7.5 CONCLUSION

Mega bit CMOS SRAM's are growing more and more complex. Error detection becomes very hard with new types of failures owing to the increasing complexity of the SRAM devices. A rapid root cause test technique is presented for mega bit SRAM chip. The proposed root cause test technique uses two complementary methods for mega bit SRAM functional testing and failure localization process. Based on the failure location process a root cause defect identified within the mega bit SRAM chip. This method is very useful in the CAD environment. Where the test pattern is automatically created it is essential to find incriminate sensitive areas in a highly complex SRAM chip layout.

Even for a complex 1 - Mb SRAM chip, the test sequence requires a very short time to accomplish root case for an observed defect within the SRAM device. The time required to perform SRAM chip failure analysis can be significantly reduced using the proposed failure analysis method. So this failure analysis method must be performed in conjunction with the SRAM CAD environment. The integrated laser beam testing approach reduces test time about 25% in comparison with conventional methods. In the 1 - Mb SRAM chip the intermittent failure to write the "1" logic state to a specific memory location, due to the open metallization in a metal-to-silicon contact was identified as the root cause of the problem. The proposed root cause technique effectively detects stuck at 0's or 1's within the SRAM device.

The 1-Mb SRAM reliability model presented in this paper clearly demonstrates close to the practical field use. The model takes into account interconnect, design, and manufacturing related failures. The reliability improvement of 1-Mb SRAM chip os 20% achieved by using the root cause test feedback and fixing the design failures. The reliability improvement of about 30% is achieved by implementing root cause feedback and fixing the manufacturing process. The overall SRAM reliability can improved by about 50% by implementing root cause test feedback and fixing failures during design and prototype phases.

Advantages:

1) SRAM product development time can be reduced about 10 %.
2) It reduces significant SRAM testing time and cost.
3) The proposed root cause test can be integrated into the desing and manufacturing process to improve reliability and reduce cost.
4) It can be applied to DRAMs, EPROMs and ASICs.

REFERENCES

[1] C.L. HENDERSON, J.M. SODEN and C.F. HAWKINS, "The Behaviour and Testing Implications of CMOS IC Logic Gate Open Circuits",proc. Int. Test Conference, Nashville, TN, 1991.

[2] NORIO KUJI, K. MATSUMABO, "A Marginal Fault Diagnosis Based on E-

Beam Static Fault Imagery with CAD Interface", International Test Conference, 1990.

[3] M. MARZOUKI, J. LAURENT, B. COURTOIS, "Comply Electron Beam Probary with Knowledge Based Fault Localization", International Test Conference, 1991.

[4] J. TEURINO, "Design to Test",Logical Solutions Technology,1989.

[5] M.T. PRONOBI AND D.J. BURNS, "Laser Die Probary for Complex CMOS IC's", prog. of the ISTFA, Oct. 1982, pp. 178-181.

[6] C.F. HAWKINS, J.M. SODEN, E.I. ODE and E.S. SNYDER, "The Use of Light Emission in Failure Analysis of CMOS IC's", proc. Int Symp for Testing and Failure Analysis, 1990, pp. 55-67.

[7] VENKATAPATHI N. RAYAPATI, "VLSI Semiconductor Random Access Memory Functional Testing" International Journal of Microelectronics and Reliability, Vol. 30, No. 5, 1990, pp. 877-899.

[8] VENKATAPATHI N. RAYAPATI and BOZENA KAMINSKA, "Mega Bit SRAM Chip Failure Analysis Using External Electrical Testing and Internal Contactless Laser Beam Testing", Proceedings of IEEE International Workshop on Memory Technology, Design, and Testing, August 1994, pp.32-37.

CONCLUSIONS

General Conclusion

This research was started with the objective of developing performance modelling and analysis for mega bit CMOS, BiCMOS Static Random Access Memories (SRAMs), which is considered as the main bottleneck problem in the semiconductor industry. During the investigations, theoretical considerations were blended with practical mega bit performance problems in order to develop performance models. Systematic analytical modelling approach of the complex multi-layer interconnects, chip performance issues were analyzed. Mega bit SRAM performance issues such as: multi-layer interconnects, chip complexity, dynamic reconfiguration methods and failure analysis technique were investigated in this research. The complex performance issues tackled pertinent to the mega bit SRAMs and the associated results sufficiently demonstrate that the research results represent a significant contribution to fulfill future application requirements.

Some of the unique advantages offered by these mega bit SRAM performance models are given below:

1) Multi-layer interconnect effect on SRAM device performance issues such as propagation delay, speed, power consumption, noise characteristics are analyzed.

2) A closed-form expression for the CMOS SRAM chip propagation delay developed, which reduces significant computational effort, design time and cost for the SRAM designer.

3) High-speed circuit design techniques investigation clearly demonstrates mega bit

SRAM device performance trade-off issues.

4) SRAM Device level dynamic reconfiguration implementation improves reliability, fault-tolerance and yield of mega bit SRAMs.

5) Dynamic reconfiguration schemes at the chip level improves, access time, chip yield, fault-tolerance and reliability, which are having major applications in Aerospace, Real-time control systems, telecommunication systems, and medical applications.

6) Rapid root cause test technique provides localization of interconnects, design, and manufacturing related failures within the SRAM chip. This technique is very useful from an industry standpoint.

Research Contribution

This research is an unique attempt to develop comprehensive performance models and analysis for mega bit CMOS, BiCMOS SRAMs. Multi-layer interconnect, SRAM chip propagation delay performance models were developed, which enable us to assess the impact of mega bit SRAM performance issues in a systematic way during all phases of the design. Mega bit SRAM chip speed-performance issues investigated in detail to explore innovative circuit design techniques for future SRAMs. In order to improve reliability, fault-tolerance, yield, speed, and noise performance measures, dynamic reconfiguration schemes were implemented at the device level and chip level for mega bit SRAMs.

The approach adopted in identifying actual mega bit SRAM performance problems and providing definite solutions through performance models and analysis result in a major contribution. Mega bit CMOS, BiCMOS SRAM performance models and analysis results are more practical value for semiconductor industry and users as well. The results provide guidance for future SRAM developments.

The summary of the research results are provided below:

1) A multi-layer interconnect capacitance model is developed for mega bit SRAM chip. Multi-layer interconnect effect on SRAM device performance parameters such as propagation delay, speed, power consumption, and noise characteristics were analyzed. Implementing triple-layer interconnect approach, the wire length and chip size were reduced to 69% and 58% respectively. Maximum access time of 30.8 ns with 1 W at 100°C and wafer yield in addition 10% more is achieved. Experimental results of multi-layer interconnections for the 1-Mb SRAM chip are presented.

2) A closed-form expression for the mega bit CMOS SRAM chip propagation delay is proposed. This allows accurate calculation of the signal delay of multi-layer interconnects within the SRAM chip and also takes into account the delay of the CMOS SRAM cells driving the branched transmission line and the driving SRAM cell loading aspects of the interconnect line. The proposed closed-form expression results in an absolute maximum error smaller than 4.8% in comparison with the measured data.

3) High-speed circuit design techniques for mega bit BiCMOS SRAMs have been

investigated for sub-micron designs. A case study of 1-Mb BiCMOS SRAM chip performance analysis results were presented. The results demonstrate the speed, power consumption, and reliability trade-offs.

4) A novel dynamic reconfiguration scheme for mega bit SRAM chip is proposed. The proposed scheme is based on on-chip word failure detection and reconfiguration to spare word cell using multi-valued logic circuits. A reliability model for 1-Mb SRAM chip is presented to evaluate the performance of the mega bit SRAM chip. 1-Mb CMOS SRAM reliability improvement of 30% has been achieved in comparison with the conventional approaches.

5) Two dynamic reconfiguration schemes are proposed for mega bit BiCMOS SRAMs. Dynamic reconfiguration schemes allow reconfiguration to fault-free memory cells within the chip.BiCMOS SRAM chip access time improvement of about 35%, chip area of about 25%, and additional chip yield of 10% are achieved respectively in comparison with conventional methods. 1-Mb BiCMOS SRAM chip reliability performance improvement of 78% is achieved in comparison with the conventional methods.

6) Rapid root cause test technique for mega bit CMOS SRAM chip is presented. The proposed root cause test technique uses two complementary methods for mega bit SRAMs failure localization and analysis. The proposed root cause test technique demonstrates that the association of an electrical tester and an internal laser beam tester makes easier for the localization of failures in the SAM chip and consequently reduces the test time and cost. A reliability model is developed

based on interconnect, design, and manufacturing test failures. 1-Mb CMOS SRAM Reliability improvement of 50% is achieved by effectively applying the root cause test feedback during the SRAM design process..

This thesis has demonstrated the performance modelling and analysis for mega bit CMOS, BiCMOS SRAMs. Multi-layer interconnects, propagation delay, dynamic reconfiguration schemes, and root cause test performance models and analysis for mega bit SRAMs , successfully demonstrated the significance of this research work.

Future Research Recommendations

Multi-layer interconnects performance models and analysis for mega bit SRAMs, successfully demonstrate the significance of the research work. This research ends with a discussion of possible extensions of the work. Performance models developed in this thesis could be further developed for application specific integrated circuits (ASICs) and dynamic random access memories (DRAMs).

Dynamic redundancy and reconfiguration schemes at the device level, chip level, and board level has various applications. Dynamic reconfiguration schemes implementation at the wafer level has not been addressed. Dynamic reconfiguration is a potential research area using sub-micron technologies at the wafer level, and system level. Because of advances in semiconductor technology, fault detection, fault isolation at the lowest level provides significant cost benefit to the industry and the component users. Dynamic reconfiguration techniques are very useful for hardware and software combined applications. Similar concepts could be further extended for communication system architectures.

Rapid root cause test technique could be further extended for DRAMs, ASICs, and field programmable gate arrays (FPGAs). Rapid root cause test technique has various practical applications in semiconductor industry. The closed-form expression for propagation delay presented in this thesis, could be further extended to ASICs. The closed form expression could be further modified and extended for sub-micron devices.

BIBLIOGRAPHY

ARZALI, L.M.; KUBO, M.; MANO, T., "Fault tolerant techniques for memory components", *ISSCC Dig. Tech. papers*, Vol. 28, Feb. 1985, pp.:231.

BALABAN, P., "Calculation of the capacitance coefficients of planar conductors on a dielectric surface",*IEEE Trans. Circuit Theory*, Vol. CT-20, Nov. 1973, pp.: 725-731.

BASTANI, B., "1.0 um BiCMOS technology for high speed 256K SRAMs", *VLSI Technology Symposium proceedings*, 1987, pp.:41-45.

BENEDEK, P., "Capacitances of a planar multiconductor configuration on a dielectric substrate by a mixed order finite-element method",*IEEE Trans. Circuit Systems*, Vol. CAS-23, May 1976, pp.: 279-284.

BORGERSON, B.R.; FREITAS, R.F., "A reliability model for gracefully degrading and stand-by sparing systems",*IEEE Transactions on Computers*, C-24, 1975, pp.: 517-525.

CARTER, D.L.; GUISE, D.F., "Effects of interconnections on submicron chip performance", *VLSI Design*, Jan 1984, pp.: 63-68.

CHAKRAVORTY, K.K. etall, "High density interconnection using photosensitive polymide and electroplated copper conducted lines", *IEEE Transactions on Components*, Hybrids and Manufacturing Technology, Vol. 13, March, 1990, pp.: 200-206.

CHU, S.T., "A 25-ns low power full CMOS 1-Mbit (128KX8) SRAM",*IEEE J.Solid State*

Circuits, Vol.Sc-23, October 1988, pp.:1078-1083.

COULTON, D.E.; GLEASON, K.R.; JAMES, K.; STRID, E.W., "Accurate measurement of high speed packages and interconnect parasitics",*IEEE/ CHMT IEMT Symposium Japan*, October 1989, pp.: 276-279.

DOUSEKI, T., "Fast access BiCMOS SRAM architecture with a VSS generator",*VLSI Circuits Design Symposium*. Tech.Papers, June 1990, pp.:45-46.

EINSPRUCH, N.G., "VLSI Electronics Microstucture Science", Vol.3, pp.: 1-24*Academic press*, 1982.

ELMASTRY M.I., " Digital MOS Integrated Circuits: A Tutorial",*Digital VLSI Systems edited by M.I Elmastry*, 1985, pp.: 10-37.

ELMORE, W.C., "The transient response of damped linear networks with particular emphais on wideband amplifiers",*J. Applied Physics*, Vol.19, No. 1,Jan 1987, pp.: 55-63.

FANNAGAN, S., "Two 13-ns 64Kb CMOS SRAMs with very low active power and improved asynchronous circuit techniques",*IEEE J.Solid State circuits*, Vol.Sc-21, October 1986, pp.:692-702.

GARDNER, D.S. et al., "Interconnection and electromagnetic scaling theory",*IEEE Transactions Electron Devices*, Vol. ED-34, March 1987, No. 3.

GRAY, F. GAIL, "General purpose reconfigurable architectures",*Proceeding of IEEE international conference on circuits and computers*, 1982, pp.: 122-123.

GROSSPIETSCH, K.E.; KAISER, J.; NETT, E., "A dynamic reconfiguration scheme for random access memories", *Digital Processes*, 1980, pp.: 257-270.

GROSSPIETSCH, K.E.; KAISER, J.; NETT, E., "A dynamic standby system for random access memories", *Proceedings of IEEE international symposium on fault tolerant computing*, 1981, pp.: 268-269.

GUBBELS, W.C.H., "A 40-ns/100-pf low power full CMOS 256K (32KX8) SRAM"*IEEE Solid State Circuits*, Vol.Sc-22, October 1987,pp.: 741-756.

HARDEE, K.C., "A fault tolerant 30 ns/375 mw 16KX1 NMOS static RAM"*IEEE J. Solid State Circuits*, October 1981, pp.: 435-443.

HAWKINS, C.F.; SODEN, J.M.; ODE, E.I.; SNYDER, E.S., "The Use of Light Emission in Failure Analysis of CMOS IC's",*proc. Int Symp for Testing and Failure Analysis*, 1990, pp.: 55-67.

HENDERSON, C.L.; SODEN, J.M.; HAWKINS, C.F., "The Behaviour and Testing Implications of CMOS IC Logic Gate Open Circuits",*proc. Int. Test Conference*, Nashville, TN, 1991.

HIROSE, T., "A 20 ns 4Mb CMOS SRAM with hierarchical word decoding architecture", *ISSCC Dig.Tech.Papers*, Feb. 1990, pp.: 132-133.

HO, P.S., "VLSI interconnection metallization"*Semiconductor International*, August 1985.

KOBAYASHI, Y., "Bipolar CMOS merged structure for high speed Mbit DRAM"*IEDM Tech. Dig. Papers*, Dec. 1986, pp.: 802-804.

KOHNO, Y., "A 14 ns 1-Mbit CMOS RAM with variable organization"*IEEE J. Solid State Circuits*, Vol. 23, Oct. 1988, pp.: 1060-1066.

KOLIAS, JOHN T., "Packaging/Performance Trade-Offs in high speed computer systems", *VMIC conference*, June 1989, pp.: 49-58.

KOMASTSU, T., "A 35-ns 128KX8 CMOS SRAM", I*EEEJ.Solid State Circuits*, Vol.Sc-22, October 1987, pp.:721-726.

KONISHI, S., "A 64Kb CMOS RAM ",*ISSCC Dig.Tech.Papers*, Feb 1982, pp.:258-259.

KUJI, NORIO; MATSUMABO, K., "A Marginal Fault Diagnosis Based on E-Beam Static Fault Imagery with CAD Interface,"*Iteration Test Conference*, 1990.

LALA, P.K., "Fault Tolerant and Fault Testable Hardware Design", *Prentice Hall International*, 1985.

LIN, T.; MEAD, C.A., "Signal delay in general RC networks",*IEEE Trans. on CAD*, Vol. CAD-3, October 1984, No. 4.

LIN, T.M.; MEAD, C.A., "Signal delay in general RC networks",*IEEE Trans. CAD*, Vol. 3, No. 4, Oct 1984, pp.: 331-349.

LOSQ, J., "Influence of fault-detection and switching mechanisms on the reliability of stand-by systems".*Proceedings of the International Symposium on Fault-tolerant Computing* , 1975, pp.: 81-86.

LOWRIE, M.B.; FUCHS, W. KENT, "Reconfigurable tree architectures using subtree oriented fault tolerance", *IEEE Trans. on computers*, 1987, Vol. C-36.

MAKI, Y., "A 6.5 ns 1 Mb BiCMOS ECL SRAM ",*ISSCC Dig. Tech. Papers*, 1990, pp.: 136-137.

MARZOUKI, M.; LAURENT, J.; COURTOIS, B., "Comply Electron Beam Probary with Knowledge Based Fault Localization", *Interation Test Conference*, 1991.

MATSUI, M., "A 25 ns 1-M bit CMOS SRAM with loading free bit lines",*IEEE J.Solid State circuits*, October 1987, pp.:733-738.

MATSUI, M., "An 8-ns 1-Mbit ECL BiCMOS SRAM with Double-Latch ECL-to-CMOS-Level converters", IEEE J.Solid State Circuits, Vol.Sc-24, October 1989, pp.: 1226-1231.

MEYER, J.E., "MOS models and circuit simulation",RCA Rev., Vol. 32, 1979, pp.: 42-63.

MCGREVITY, D.J., "Interconnections/gates in VLSI technologies",VLSI Technologies through the 80's and beyond, IEEE Computer Society Press, 1982.

MINATO, O., " 20-ns 64K CMOS SRAM", ISSCC Dig.Tech.Papers Feb.1984, pp.:222-223.

NAKAMURA, K., "A 6-ns ECL 100K I/O and 8-ns 3.3V TTL I/O 4-MB BiCMOS SRAM", IEEE Jouranal of Solid State circuits, Vol 276, No 11, 1992, pp.: 1504-1510.

OCHII, K., "A 15-nW standby power 64Kb CMOS RAM",ISSCC Dig.Tech.Papers, Feb 1982, pp.: 260-261.

OGIVE, K., "13-ns 500 mw, 64Kbit ECL RAM using HI-BiCMOS technology"IEEE J.Solid State circuits, Vol.Sc-, October 1986, pp.:681-685.

OH, S.Y. et al., "Transient analysis of MOS transistors", IEEE Trans. Electron Devices Vol. ED-27, 1980, pp.: 1571-1578.

PETERSON, W. W. , "Error Correcting Codes",MIT Press, Cambridge, 1972.

PRONOBI, M.T.; BURNS, D.J., "Laser Die Probary for Complex CMOS IC's"prog. of the ISTFA, Oct. 1982, pp.: 178-181.

RAGHAVENDRA, C.S., "Fault tolerance in regular network architectures",IEEE micro, 1984, pp.: 44.

RAYAPATI, V.N.; MUKHEDKAR, D., "Interconnection problems in VLSI random access memory chip, " SPIE, Proc. Int. Conf. on advances in Interconnection and

Packaging, Vol. 1389, 1990, pp.: 98-109.

RAYAPATI, V.N., "Modular fault tolerant VLSI parallel processor architecture with dynamic redundancy", International Journal of Microelectronics and Reliability, Vol. 30, No. 2, 1990, pp.: 213-236.

RAYAPATI, V.; MAHAPATRA, S., "Fault tolerance in NMOS Random Access Memories with Dynamic Redundancy Methods", Microelectronics and Reliability: An International Journal, Vol.28, No. 2, 1988, pp.: 193-200.

RAYAPATI, V.; MAHAPATRA, S., "A Fault Model for Multi-Value NMOS Dynamic Random Access Memories", Microelectronics and Reliability: An International Journal, Vol. 29, No. 2, 1989, pp.: 42-50.

RAYAPATI, V.N.; KAMINSKA, B., "Performance Analysis of Multi-layer Interconnections for Mega bit Static Random Access Memory Chip", IEEE Transactions on Components, hybrids, and Manufacturing Technology, Vol.16, No.5, August 1993, pp.: 469-477.

RAYAPATI, V.N.; KAMINSKA, B., "Mega Bit BiCMOS SRAM chip Package Modelling and Performance Analysis", Proceedings of IEEE International Workshop on Memory Technology, Design and Testing, August 1994, pp.: 10-15.

RAYAPATI, V. N., "VLSI Semiconductor Random Access Memory Functional Testing" International Journal of Microelectronics and Reliability, Vol. 30, No. 5, 1990, pp.: 877-899.

RAYAPATI, V. N.; KAMINSKA, BOZENA, "Mega Bit SRAM Chip Failure Analysis Using External Electrical Testing and Internal Contactless Laser Beam Testing", Proceedings of IEEE International Workshop on Memory Technology, Design, and Testing, August 1994, pp.: 32-37.

RUBINSTEIN, J.; PENFILED, P.; HOROWITZ, M., "Signal delay in RC tree networks", IEEE Trans. Computer Aided Design, Vol.CAD-2, No.3, July 1983, pp.: 202-211.

SAKURI, T., "Approximation of wiring delay in MOSFET LSI",IEEE Journal of Solid State Circuits, Vol. SC-20, December 1985, No. 6.

SAKURAI, T., "Approximation of wiring delay in MOSFET LSI",IEEE J. Solid State Circuits, Vol.SC-18, No.4, Aug 1983, pp.: 418-426.

SARRAZIN, D.B.; MALEK, M., "Fault-tolerant, Semiconductor memories",Computer Magazine, August 1984, pp.: 49-56.

SASAKI, K., "A 7-ns 140-mW 1-Mb CMOS SRAM with current sense amplifier"IEEE J. Solid State Circuits, Vol. 27, No. 11, November 1992, pp.: 1511-1518.

SASAKI, K., "A 23-ns 4 Mb CMOS SRAM with 0.5 uA standby current",ISSCC Dig.Tech.Papers, Feb 1990, pp.:130-131.

SAKURAII, T., "A low power 46-ns 256Kbit CMOS static RAM with dynamic Double Word Line", IEEE J.Solid State Circuits, October 1984, pp.: 578-584.

SCHUSTER, S.E., "Multiple word/bit line redundancy for semiconductor memories"IEEE J.Solid State Circuits, Vol.SC-13, Oct. 1978, pp.: 698-703.

SHERE, B.J. et al., "Measurement and modelling of short-channel MOS transistor gate capacitances", IEEE J. Solid-state circuits, Vol. SC-22, 1987, pp.: 464-472.

SILVESTER, P.; FERRAI, R.L., "Finite element for electrical engineers", New York: Cambridge Univ. Press, 1983.

SINGHAL, K.; VLACH, J., " Approximation of non-uniform RC distributed networks for frequency and time domian computations", _IEEE Trans. on Circuit Theory_, Vol.CT-19, July 1972, pp.: 347-354.

SMALL, H.B.; Pearson, D.J., "on-chip wiring for VLSI: status and directions"_IBM Journal of Research and Development_, Vol. 34, November 1990, No. 6.

STIFFLER, J. J., "Error correcting Coding techniques for Random Access Memories"_IEEE Transactions on Computers_, C-27, 1978, 256-531.

SUZUKI, M., "A 3.5 ns 500 mW 16 Kb BiCMOS ECL SRAM",_SSCC Dig. Tech. Papers_, Feb. 1989, pp.: 32-33.

TAKADA, M., "A 5 ns 1 MB ECL BiCMOS SRAM ",_IEEE J. Solid -State Circuits_, Vol.25, Oct. 1990, pp.: 1057-1062.

TANNER, R.M., "Fault-tolerant 256K memory designs", IEEE Transactions on Computers, C-33, 1984, pp.:314-322.

TEURINO, J., "Design to Test",_Logical Solutions Technology_. 1989.

UTTECHT, R.R. et al., "A four-level-metal fully planarized interconnect technology for dense high performance logic and SRAM applications",_VMIC conference_, June 1991, pp.: 20-26.

VAN DER POL, B.; BREMMER, H., Operational calculus based on the two-sided Lapalace integral, chapter 7, _Cambridge University Press_.

WADA, T., "A 34 ns 1 Mb CMOS RAM using triple poly",_ISSCC Digest of Technical Papers_, Feb. 1987, pp.: 262-263.

WARD, D.E.; DUTTON, R.W., "A charge-oriented model for MOS transistor capacitances", IEEE J. Solid-state circuits, Vol. SC-13, 1978, pp.: 703-708.

YOKOMIZO, K.; NAITO, K., "Design Techniques for high-throughput BiCMOS Self-timed SRAMs", IEEE Journal of Solid State Circuits, Vol-28, No.4, April 1993, pp.: 484-489.