ABSTRACT

PARK, JONG BEOM. 3D-DATE: A Circuit-Level Three-Dimensional DRAM Area, Timing, and Energy Model. (Under the direction of W. Rhett Davis and Paul D. Franzon.)

Three-dimensional stacked DRAM technology has emerged recently. Many studies have shown that 3D DRAM is most promising solutions for future memory architecture to fulfill high bandwidth and high-speed operation with low energy consumption. It is necessary to explore 3D DRAM design space and find the optimum DRAM architecture in different system needs. However, a few studies have offered models for power and access latency calculations of DRAM designs in limited ranges. This has led to a growing gap in knowledge of the area, timing, and energy modeling of 3D DRAMs for utilization in the design process of processor architectures that could benefit from 3D DRAMs. This paper presents a circuit level DRAM Area, Timing, and Energy model (DATE) which supports 3D DRAM design with TSV. DATE provides front-end and back-end DRAM process roadmap from 90 nm to 16 nm node and provides a broader range 3D DRAM design model along with emerging transistor device. DATE is successfully validated against several commodity planar and 3D DRAMs and published prototype DRAMs with emerging device. Energy verification has a mean error of about -5% to 1%, with a standard deviation of up to 9.8%. Speed verification has a mean error of about -13% to -27% and a standard deviation of up to 24%. In the case of the area, the bank has a mean error of -3% and the whole die has a mean error of -1%. The standard deviation for area is up to 4.2%. In the case study, we demonstrate that 1Gb DDR3 DRAM designs achieve up to about 0.7 Gb/sec data throughput and energy efficiency of 510 bit/nJ using 3D design options with 16 nm DRAM technology.

© Copyright 2018 by Jong Beom Park

All Rights Reserved

3D-DATE: A Circuit-Level Three-Dimensional DRAM Area, Timing, and Energy Model

by
Jong Beom Park

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Electrical Engineering

Raleigh, North Carolina

2018

APPROVED BY:

James Tuck

Hans Hallen

W. Rhett Davis
Co-chair of Advisory Committee

Paul D. Franzon
Co-chair of Advisory Committee

DEDICATION

My Lord, Jesus

My wife, Jina, and my family.

BIOGRAPHY

Jong Beom Park was born in Seoul, Korea in March 1978. He earned a Bachelors of Science at Hanyang University at Ansan in 2001. In 2003, he earned a Master of Science in Electronic, Electrical, Control and Instrumentation Engineering from Hanyang University at Seoul in 2003, with a thesis entitled "Implementation of the Multirate Viterbi Algorithm for IEEE 802.11a Wireless LAN System." After working in the industry for several years, Mr. Park entered the ECE graduate program at North Carolina State University in 2009, where he earned a Masters of Science in Computer Engineering from North Carolina State University in 2010. He initiated his Ph.D. studies in Electrical Engineering in 2011 working on the NSF's Underwater Optical Communication program with Dr. John Muth. In 2012, Mr. Park switched his research focus to the circuit design area rather than embedded system. Thus, he joined Dr. Paul D. Franzon's research group in 2012. He started working on the DARPA PERFECT program in 2012 and 2013, focusing on the design of a custom, low power memory. Mr. Park also maintains an active interest in computer architecture, digital VLSI design, and machine learning.

ACKNOWLEDGEMENTS

First, I would like to thank Dr. Paul D. Franzon, my advisor. I still remember the moment I first joined his research group. Dr. Franzon told me, "Welcome aboard" with a generous smile. It was a great fortune for me to be on his ship. Without his mentorship and guidance, this journey would not have been possible. I also would like to thank my co-advisor, Dr. W. Rhett Davis for being supportive on many occasions, his valuable comments on my research, and providing the research opportunity. In addition, I would like to thank the following faculty members: Dr. John Muth for giving me my first research opportunity at NC State; Dr. James Tuck for his mentoring on PERFECT projects and for being my committee; and Dr. Hans Hallen for his valuable comments on my research and for being my committee.

I would like to thank the following people for their contributions that have made my dissertation possible: Joshua C. Schabel for motivating me and helping me to write this thesis with creative discussions; Kirti Bhanushali and Wenxu Zhao for being great colleagues throughout the research with discussions that made ambiguities clear; Randy and Weiyi Qi for sharing insights on modeling algorithms into the program; and Lee B. Baker for sharing insights on machine learning.

Finally, I would like to thank my parents and parents-in-law who would be glad in Heaven with God, my wife Jina and two lovely daughters, Songee and Yuni. I appreciate their sacrifices, support, and patience.

TABLE OF CONTENTS

LIST OF TABLES	vii
LIST OF FIGURES	ix
Chapter 1 Introduction	1
1.1 Motivation	1
1.2 Original Contributions	2
1.3 Related Work	3
1.4 Organization of Dissertation	5
1.5 Abbreviations	5
Chapter 2 DRAM Process Roadmap	7
2.1 Transistor Model and Scaling	9
2.1.1 Gate Transistor Model and Scaling	11
2.1.2 High-Voltage and Peripheral transistor	17
2.2 Interconnect	18
2.2.1 Wire	18
2.2.2 Through Silicon Via	24
2.3 Roadmap and discussion	27
2.3.1 Gate Transistor	27
2.3.2 High Voltage and Peripheral Transistor	31
2.3.3 Wire	35
2.3.4 Through Silicon Via	40
Chapter 3 DRAM Circuit Level Modeling	43
3.1 Component Modeling	46
3.1.1 General Layout and Drain Capacitance	46
3.1.2 Digital Logic and Driving Buffer	51
3.1.3 Repeater for Wire	58
3.1.4 Address Decoder	60
3.1.5 Bitline and Bitline Sense Amplifier	63
3.2 Architecture Level Modeling	65
3.3 Validation	72
3.4 Comparison with Other Models	79
Chapter 4 Case Study: DRAM Design Space Exploration	82
4.1 Planar Design Space Exploration in 35 nm Node	83
4.1.1 Single Bank Design Space	83
4.1.2 Multi-bank Design Space in 35 nm Node	98

4.2	3D Design Space Exploration in 35 nm Node	106
4.2.1	Area Efficiency	107
4.2.2	Energy Efficiency	109
4.2.3	Throughput	111
4.2.4	Product of Design Metric	113
4.2.5	Design Metric Comparison in Different Technology	116
Chapter 5	Conclusion and Future Work	119
5.1	Summary of Contributions	119
5.2	Future Work	121
BIBLIOGRAPHY		122
APPENDICES		129
Appendix A	Derivation of the Leakage Current Equation	130
Appendix B	TCAD Simulation Code	133
B.1	Sentaurus Structure Editor Code	133
B.2	Sentaurus Device Code	140
B.3	Inspect Code	144
Appendix C	Definition and Derivation of the Path Effort	147
Appendix D	Reference of Commodity DRAM Part	150
Appendix E	How to run DATE	153
E.1	Read-me First	153
E.2	Executable file with comments	154

LIST OF TABLES

Table 2.1	Material and doping method of gate transistor	13
Table 2.2	Leakage Current Criterion (fA/cell)	15
Table 2.3	ITRS Saturation Current Roadmap of Supportive NMOSFET at 25 °C .	18
Table 2.4	Gate transistor roadmap	29
Table 2.5	High voltage transistor roadmap	33
Table 2.6	Peripheral transistor roadmap	33
Table 2.7	DATE wire roadmap	34
Table 2.8	Wire comparison with commodity logic design process	39
Table 2.9	ITRS TSV physical dimension roadmap	41
Table 2.10	CACTI-3DD TSV physical dimension roadmap	41
Table 2.11	DATE TSV area, capacitance, and resistance roadmap	41
Table 2.12	ITRS TSV area, capacitance, and resistance roadmap	42
Table 3.1	The logical effort of logic gates	52
Table 3.2	Validation of energy calculation	73
Table 3.3	Validation of key timing parameter calculation	74
Table 3.4	Validation of area calculation	76
Table 3.5	Validation of timing parameter of VCAT based and 3D DRAM	77
Table 3.6	Validation of area calculation of VCAT based and 3D DRAM	77
Table 3.7	Timing parameter change according to process change	78
Table 3.8	Energy change according to process change	79
Table 3.9	Circuit level model comparison	81
Table 4.1	Address bit and physical dimension of bank matched to each page size at in 1 Gb single bank, 6 F^2 layout	84
Table 4.2	Area efficiency of single bank DRAM	85
Table 4.3	Read energy efficiency of single bank DRAM, 6 F^2 layout	86
Table 4.4	Read operation energy change in each component as page size change	88
Table 4.5	Bank size of single bank DRAM as subarray size change, 6 F^2 layout .	89
Table 4.6	Read operation energy change as subarray row size change	90
Table 4.7	Read operation energy change as subarray column size change	91
Table 4.8	Read energy efficiency of single bank DRAM, 4 F^2 layout	92
Table 4.9	Throughput of read operation, single bank DRAM, 6 F^2 layout	94
Table 4.10	Speed of each component and read throughput as page size change .	94
Table 4.11	Speed of each component and read throughput as subarray row size change	97
Table 4.12	Speed of each component and throughput as subarray column size change	98
Table 4.13	Throughput of read operation, single bank DRAM, 4 F^2 layout	99

Table 4.14	Area efficiency of planar multibank DRAM	100
Table 4.15	Read Energy and Efficiency of 1 Gb 2D Multibank DRAM in 35 nm node, 6 F^2 layout	101
Table 4.16	Read Energy and Efficiency of 1 Gb 2D Multibank DRAM in 35 nm node, 4 F^2 layout	103
Table 4.17	Throughput of 1 Gb 2D Multibank DRAM, 6 F^2 layout	105
Table 4.18	Area efficiency of 1 Gb 3D multibank DRAM in 35 nm node	109
Table 4.19	Energy efficiency of 1 Gb 3D multibank DRAM in 35 nm node	111
Table 4.20	Throughput of 1 Gb 3D multibank DRAM, 6 F^2 layout in 35 nm node	112
Table 4.21	The best performance metric results in 16 nm node	118

LIST OF FIGURES

Figure 2.1	Cross-section of various gate transistor	10
Figure 2.2	3D Schematic diagram of VCAT-based DRAM cell	11
Figure 2.3	DRAM Cell Layout	12
Figure 2.4	Gate transistor structures	14
Figure 2.5	Recessed gate transistor threshold voltage trend	16
Figure 2.6	MASTAR graphical user interface	17
Figure 2.7	Wire and wire cross section for resistance calculation	19
Figure 2.8	Wire cross section for capacitance calculation	19
Figure 2.9	Typical cross-section of interconnect architectures	22
Figure 2.10	Cross section and top view of single TSV with capacitance	25
Figure 2.11	TSV bundles and coupled capacitance	25
Figure 2.12	TCAD simulation snapshot of recessed transistors	27
Figure 2.13	Three dimensional view and cross section of TCAD simulation of VCAT	28
Figure 2.14	Gate transistor roadmap	30
Figure 2.15	Detail view of source or drain junction of MOSFET	31
Figure 2.16	Wire resistance roadmap	35
Figure 2.17	Wire capacitance roadmap	36
Figure 2.18	Metal capacitance comparison	37
Figure 2.19	Metal resistance comparison	38
Figure 3.1	DATE program flow	45
Figure 3.2	Wide width transistor layout	46
Figure 3.3	Folded transistor layout	47
Figure 3.4	Drain region of the folded transistor	48
Figure 3.5	Internal layout height assumption	49
Figure 3.6	Two input NAND gate schematic and layout example	50
Figure 3.7	Drain region of series-connected transistor	50
Figure 3.8	DATE logic design assumption	51
Figure 3.9	Transistor size of inverter, NAND, and NOR gate	55
Figure 3.10	Horowitz gate model	56
Figure 3.11	Two input NAND gate	57
Figure 3.12	Interconnect line with repeater	59
Figure 3.13	Nine bit row address decoding path	61
Figure 3.14	Predecoder structure	62
Figure 3.15	Two input NOR gate	62
Figure 3.16	Bitline sense amplifier	63
Figure 3.17	Eight bank DDR DRAM floor plan	66
Figure 3.18	Subarray and related peripheral circuits	68

Figure 3.19	Schematic diagram of primitive core array for the conventional 6 F^2 DRAM.	70
Figure 3.20	Schematic diagram of primitive core array for the 4 F^2 DRAM	71
Figure 4.1	Energy efficiency as subarray size change with 6 F^2 layout, 16384-bit page size	88
Figure 4.2	Energy efficiency as subarray size change with 4 F^2 bitcell layout, 16384-bit page size	92
Figure 4.3	Read throughput as subarray size change at 6 F^2 layout, 16384-bit page size	96
Figure 4.4	Data Throughput as subarray size change at 4 F^2 layout with 16384-bit page size	99
Figure 4.5	Energy sum of wire component in multi-bank 2D DRAM, 6 F^2 Layout	102
Figure 4.6	Energy sum of wire component in multibank 2D DRAM, 4 F^2 Layout	104
Figure 4.7	Sum of each component delay in multibank 2D DRAM, 6 F^2 Layout	105
Figure 4.8	Rank level die-stacking	107
Figure 4.9	Chip micrograph of the fabricated DRAM die and cross-sectional view of TSVs	108
Figure 4.10	Bank level die-stacking	108
Figure 4.11	Energy sum of wire components and TSV in 35 nm node	110
Figure 4.12	Delay sum of each design component with TSV in 35 nm node	113
Figure 4.13	Multiple design metric trend in 35 nm node	115
Figure 4.14	Design metric comparison between 68 nm, 35 nm and 16 nm node in 6 F^2 cell layout	117
Figure A.1	Variation of bitline voltage with time	132

CHAPTER

1

INTRODUCTION

1.1 Motivation

Three-dimensional die stacking involves connecting multiple silicon dies with a vertical interconnect, such as through-silicon vias (TSVs) or micro-bumps. Three-dimensional die stacking reduces global wire routing inside of integrated circuits[1]. Implementing dynamic random access memory(DRAM) in three-dimensional stacks could minimize random access latencies, internal cycle time and power consumption. These benefits have motivated industry to implement 3D die stacked DRAM for off-chip, and on-chip stacked memory applications.

One example is Micron’s Hybrid Memory Cube (HMC) which utilizes an off-chip, 3D DRAM. A single HMC provides 160 GB/s to 320 GB/s peak transfer bandwidth while DDR3 DRAM module offers tens of GB/s [2, 3]. The Wide-I/O and Wide-I/O 2 standards also have been proposed for on-chip stacked DRAM by Joint Electron Device Engineering Council (JEDEC) [4, 5]. Samsung has shown that Wide-I/O has 330.6 mW read operating power in 50 nm process which is almost equal to LPDDR2 read power at the same process node. Samsung also has shown that Wide-I/O has 12.8 GB/s data bandwidth, which is four times of LPDDR2’s [6].

Many studies have shown that 3D DRAM provides higher bandwidth with lower power consumption, as well as methods to utilize 3-D DRAM in memory hierarchies [2, 7–10]. However, few studies have offered models for power and access latency calculations of custom designs. This has led to a growing gap in knowledge of the area, timing, and energy modeling of 3D DRAMs for utilization in the design process of processor architectures that could benefit from 3D DRAMs.

1.2 Original Contributions

The goal of this work is to provide a 3D DRAM Area, Timing and Energy (DATE) model. DATE not only can be used to model existing standard planar DRAM, but also for custom 3D DRAM designs or to find the optimal 3D DRAM design for architectures under exploration using traditional or emerging devices. To support the goal, this work includes the following original contributions:

- DATE provides transistor-level accuracy across various DRAM process nodes, from 90 nm to 16 nm process nodes.

- DATE presents four different transistor models for modeling DRAM. The recessed channel array transistor (RCAT) and the Sphere-shaped-RCAT (SRCAT) models are provided in DATE for modeling traditional commodity DRAMs. DATE also provides an emerging gate transistor device, the vertical channel access transistor (VCAT) to reflect the future DRAM layout trend and thus its effect on area, energy, and speed. To support modeling of general transistor models in DRAM peripheral circuits, a conventional metal-oxide-semiconductor field-effect transistor (MOSFET) model is provided in DATE.
- DATE demonstrates a new core design to support emerging VCAT based cell array layout as depicted in [11]. The new core design includes sense-amplifier (SA) rotation and hybridization, conjunction restructuring, word line (WL) strapping, etc.
- DATE is validated against 22 planar and 3D DRAMs from 80 nm to 30 nm technology. The details are shown in Section 3.3.

A more detailed comparison with other models are presented in Section 3.4.

1.3 Related Work

There are two approaches for analyzing DRAM: (a) circuit-level and (b) system-level power models. Briefly, the circuit-level model examines DRAM based on given front-end, back-end process, and architectural assumptions. Thus, this model can calculate energy, speed, and area of DRAM. The accuracy of the model depends on the accuracy of the DRAM process model and architectural assumptions.

The system-level power model utilizes the DRAM's JEDEC standard operating-scenario current numbers from vendor's data-sheet for calculating specific DRAM operation energy.

While this energy model gives precise energy numbers for standard DRAMs, the system-level power model is limited to only those DRAMs with data sheets provided. Thus, the system-level power model cannot be utilized to explore new DRAM architectures in which there is no datasheet available. The system-level power model is also limited in that it cannot address sub-logic level power numbers. Thus, this model cannot clarify specific parts of the architecture if the power optimization is required.

Many circuit-level power, area, and timing models have been introduced. CACTI [12] is the most widely known of these models. CACTI models caches, SRAMs and DRAMs. The architectural and circuit level model includes assumptions of optimizing cache and SRAM and is suitable for modeling embedded DRAM.

Rambus has also proposed a circuit-level commodity DRAM power model [13]. The Rambus model calculates power and area but does not calculate the speed nor provide a detailed circuit model of the DRAM sub-logic blocks. Although it allows the user to choose design assumptions, without providing detail design guidance, the user could encounter the pitfall of wrong assumptions on energy and area prediction by choosing false DRAM sub-logic blocks. Both the CACTI and Rambus models are derived from planar die models.

CACTI-3DD is published to model commodity 3D-DRAMs [14]. The model does not support DRAMs implemented in or below the 21 nm technology node nor DRAMs implemented with emerging gate transistor devices with related architectural changes. CACTI-3DD does not provide a gate-transistor model, but rather a model designed upon the planar transistor model with ideal assumptions.

For modeling power at system-level, Micron provides support for power analysis of their planar DRAM in their application notes [15]. Chandrasekar et al. have improved Micron's system-level energy model and released it online, applying it to the Wide-I/O standard for 3D DRAMs [16–18]. Weis et al. have shown 3D DRAM design space exploration and

optimization at the extension of system-level power model [19]. The Weis' study shows area, energy, and speed of 58 nm, 46 nm, and 45 nm process node DRAM. In the study, necessary information is extracted from real measurement or simulations of commodity DRAM device. The Weis' study is limited to specific technology node (58 nm, 46 nm and 45 nm) and not support emerging devices as well.

1.4 Organization of Dissertation

This dissertation is organized as follows. Chapter 2 presents DRAM process node characterization. Transistors, wires, and through silicon via (TSV) models, modeled from 90 nm to 16 nm technology nodes, are discussed. Chapter 3 presents circuit-level model and architectural-level model of 3D DRAM. Chapter 4 presents the first case study, which explores the benefits of 3D design space using a 1 Gb standard double-data-rate DRAM. Summary and future work are outlined in Chapter 5.

1.5 Abbreviations

ASC	Asymmetric Channel Doping
BL	BitLine
DATE	DRAM Area, Timing, and Energy model
DDR	Double Data Rate
DRAM	Dynamic Random Access Memory
F	minimum Feature size
FEOL	Front-End-Of-Line
FinFET	Fin Field Effect Transistor

ITRS	International Technology Roadmap for Semiconductors
JEDEC	Joint Electron Device Engineering Council
LPDDR	Low Power Double Data Rate
MASTAR	Model for Assessment of cmoS Technology And Roadmaps
MOSFET	Metal-Oxide-Semiconductor Field-Effect Transistor
MWL	Main WordLine
NMOS	n-channel MOSFET
PMOS	p-channel MOSFET
RCAT	Recessed Channel Array Transistor
SRAM	Static Random Access Memory
SRCAT	Sphere-shaped Recessed Channel Array Transistor
SWL	Sub-WordLine
TCAD	Synopsys Technology Computer-Aided Design
TSV	Through Silicon Via
VCAT	Vertical Channel Access Transistor
WL	WordLine

CHAPTER

2

DRAM PROCESS ROADMAP

As process technology advances, the needs of differing applications have led to different process roadmaps. The international technology roadmap for semiconductors (ITRS) provides application-specific roadmaps that reflect the industry needs [20]. For the DRAM, ITRS provides a roadmap of scaling for several key features. In 2001 and 2003, ITRS provided cell size, storage cell dielectric thickness and minimum retention time from 130 nm node down to 18 nm node. In 2005, ITRS added more features including storage cell capacitor dielectric thickness, gate transistor dielectric thickness, maximum wordline voltage level, electric field of capacitor dielectric, and electric field of gate transistor dielectric. ITRS2005 roadmap started from 80 nm node. In 2007, ITRS added prospective of gate transistor

structure, supportive transistor supply voltage, saturation current of NMOS and PMOS supportive transistor with gate materials, and oxide thickness of NMOS supportive transistor from 68 nm node. Since 2007, the ITRS has updated information on each feature.

The ITRS roadmap is not sufficient to reveal overall area, energy and performance information. ITRS2001 and ITRS2003 does not provide any DRAM transistor information. ITRS2005 roadmap provides partial information of gate transistor with wordline voltage from 80 nm node. From the ITRS2007 roadmap, ITRS provides information for supportive transistors from 68 nm node but no information about the gate transistors.

Rambus has built DRAM power model which provides DRAM process technology roadmap from 140 nm node [13]. The roadmap contains projection for the capacitance and voltage of transistors and interconnects. In detail, it provides length, width and oxide thickness for transistors on each technology node. However, the roadmap does not provide any resistance and current information to calculate speed. Instead, the Rambus model evaluates the power according to the clock speed from DDR specification.

CACTI-3DD [14] utilizes the ITRS low standby power (LSTP) process technology roadmap, also implemented in CACTI. The LSTP process technology roadmap is designed for modeling low power digital IC processes, but the bitcell transistors modeled in CACTI-3DD utilizes a constant turn-on current, which can lead to inaccurate performance estimations in future process technologies.

DATE presents DRAM roadmap from 90 nm technology node. As we discussed above, previous roadmaps do not provide sufficient information about area, speed, and energy from 90 nm technology node. Since there are discrepancies, and indeed even inaccuracies between DRAM roadmap, we deploy the DRAM process roadmap in this chapter. The various gate transistor will be discussed as well as interconnects and TSV.

2.1 Transistor Model and Scaling

In DRAM, a gate transistor is required to reduce the leakage current and to retain the stored data in the cell capacitor during the required data retention time. As feature size reduces, conventional planar transistors suffer from higher leakage current, mainly due to higher electric field across the channel since supply voltage does not linearly scale with the channel length. Increasing channel doping suppresses the subthreshold current with a counter effect of an increase in the electric field across the device junction to the storage capacitor. This increases the junction leakage in the storage node.

Researchers have proposed several different devices for a bitcell transistor to reduce leakages [21–28]. Samsung proposed the recessed channel array transistor (RCAT) with 88 nm DRAM technology [21] and scaling RCAT down to 50 nm process [22]. The recessed gate structure increases the effective channel length of gate which helps reduce the leakage current. The channel doping density can be reduced; therefore RCAT reduces junction leakage and overall leakage current [29]. Samsung also proposed a sphere-shaped recessed channel array transistor (SRCAT) with the 70 nm process and expected extendable scaling down to sub-50 nm process. SRCAT provides more recessed channel effect than RCAT [23]. Figure 2.1 shows the cross-section of various gate transistor. Arrows indicate channel length. As depicted in the Figure 2.1, SRCAT has longer channel length than RCAT or planar MOSFET.

FinFET or its hybrid are also studied as a bitcell transistor in DRAMs [25–27, 30]. FinFETs have a more extensive channel width compared to a planar transistor which helps to suppress short channel effect. Thus, FinFET can be used as a gate transistor in a smaller technology node rather than a planar transistor [26, 29]. However, FinFET needs negative word line scheme or work function engineering to satisfy the off-leakage current requirement

Figure 2.1 Cross-section of various gate transistor.

[31]. These limitations make FinFET less attractive than RCAT and SRCAT.

Vertical channel access transistor (VCAT) is another transistor that has been proposed as a bitcell transistor alternative for DRAMs [11, 24]. The major benefit from VCAT is area efficiency; the VCAT is a three-dimensional structure in which the channel exists vertically, surrounded by the gate as depicted in Figure 2.2. This allows for the bitcell transistor to be placed at the cross section of bitline and wordline and also allows VCAT dedicated denser cell layout such as $4 F^2$. Even though VCAT does not support RCAT or SRCAT based cell layouts such as $8 F^2$ or $6 F^2$ cell layout, VCAT drives cell area from $8 F^2$ or $6 F^2$ to $4 F^2$ as shown in Figure 2.3. The unit, F , is denoted in minimum feature size (half pitch of the first metal layer). Since $4 F^2$ cell array layout could increase the gross die about 1.35 times compared to $6 F^2$ cell array layout, the industry expected VCAT as the next gate device [20, 24].

Compared to all other supportive circuits, satisfying the speed margin of DRAM standard (i.e. DDR, DDR2, etc.) is the driving force that underlines most design and technology choices [22]. Studies show that commodity DRAM uses conventional planar transistors [34] or mixed devices (bitcell transistor for NMOS and a planar transistor for PMOS [22]) for these supportive circuits. High voltage driving transistors, required to drive poly wordline

Figure 2.2 3D Schematic diagram of VCAT-based DRAM cell [11].

in the bitcell array area are depicted in reference [13].

2.1.1 Gate Transistor Model and Scaling

In this work, our gate transistor roadmap is deployed with Synopsys Technology Computer-Aided Design (TCAD) device simulator technology. Unlike MOSFET or FinFET, RCAT, SRCAT, and VCAT do not have an appropriate numerical device model. DATE provides a gate transistor roadmap of recessed devices like RCAT, SRCAT, and of emerging devices such as VCAT since the industry has extensively used it and is expected to continue its use in the future. RCAT, SRCAT, and VCAT structures and simulation conditions are shown in Figure 2.4 and Table 2.1. As shown in Figure 2.4c, the top part of the VCAT pillar diameter would be assumed as $0.5 F$ due to the etching process. During the evaluation, general feature size and scaling is employed from Rambus DRAM process roadmap. Between technology nodes,

(a) $8F^2$ Cell Layout [32]

(b) $6F^2$ Cell Layout [33]

(c) $4F^2$ Cell Layout [24]

Figure 2.3 DRAM Cell Layout.

Table 2.1 Material and doping method of gate transistor

Structure	Tech. node (nm)	Depth or Height (nm)	Gate Material	Substrate Doping Method
RCAT	90,75	200	WSi	Uniform
SRCAT	65,55,45,40, 36,31,27,24, 21,18,16	190	WSi (65 nm), W	Uniform (65 nm), Asymmetric
VCAT	90,65,45,31, 21,16	250	Poly Silicon	Uniform

DATE adopts linearly interpolated values.

In RCAT and SRCAT, we assume RCAT would dominate in 90 nm and 75 nm process, and SRCAT would dominate from 65 nm to 16 nm [23, 35]. The trench depth of the recessed devices has dependency with the threshold voltage. A deeper trench would result in lower threshold voltage even while all the other conditions are unchanged [36]. The trench depth would follow the results published in reference [23], which examined 110 nm to 60 nm process. Below 60 nm process, we assume that the trench depth would remain as it is on the 60 nm process. For the gate material, we assume it is tungsten silicide from 75 nm and tungsten from 55 nm [34]. These work functions are 4.82 eV and 5.12 eV in each case [37, 38]. Asymmetric Channel Doping (ASC) is assumed from 55 nm to reduce junction leakage between capacitor and gate transistor [34, 39].

VCAT would be used as gate transistor from approximately 28 nm and below according to ITRS roadmap [20]. However, DATE provides roadmap from 90 nm for the comparison since vendors fabricate test chip in larger technology nodes [11]. The pillar height would remain as 250 nm, as shown in [11]. The gate material would be polysilicon [24], i.e., 4.15 eV for the work function.

Figure 2.4 Gate transistor structures.

Low leakage current (I_{off}) is the primary decision criterion for the gate-transistor design parameters. The JEDEC standard requires 64 ms data retention time at 85 °C for the storage node [40]. The relationship between storage node retention time (t_{REF}) and I_{off} is described by the equation [41, 42]:

$$I_{off} = \frac{C_S(V_{array}/2 - \Delta V_{BL}) - C_B \Delta V_{BL}}{t_{REF}}. \quad (2.1)$$

ΔV_{BL} is the bitline sensing voltage and is given by the equation,

$$\Delta V_{BL} = \frac{C_S}{C_B + C_S} \times \left(\frac{1}{2} V_{array} - \Delta V_{MAX} \right) \quad (2.2)$$

where C_B is the bitline capacitance, V_{array} is the bit array operating voltage, ΔV_{MAX} is maximum allowable data loss for reading data with a sense amplifier, and C_S is the storage

Table 2.2 Leakage Current Criterion (fA/cell)

t_{ref} (ms)	DRAM Process Node (nm)												
	90	75	65	55	44	40	36	31	27	24	21	18	16
64	94.2	96.1	97.7	71.7	63.9	63.6	58.6	60.0	61.3	56.3	57.5	59.1	60.3
500	12.1	12.3	12.5	9.2	8.2	8.1	7.5	7.7	7.8	7.2	7.4	7.6	7.7
1000	6.0	6.2	6.3	4.6	4.1	4.1	3.8	3.8	3.9	3.6	3.7	3.8	3.9

capacitance. Detailed derivation is provided in Appendix A.

DATE adopts internal and supply voltage projection from the Rambus roadmap and assumes storage capacitor has 30 fF as CACTI model. Bitline capacitance could change according to bank design. Rambus assumes bitline capacitance about 192 fF on 90 nm node [13]. DATE calculates bitline capacitance about 90 fF to 100 fF on 80 nm node while evaluates 1 Gb commodity DRAM [43]. For conservative prediction, we assume bitline capacitance as 300 fF on 90 nm node and linearly reduce according to the technology node. The ΔV_{MAX} is 10% of V_{array} for calculation.

Table 2.2 shows the leakage current calculation result that meets the required retention time with these assumptions. The DRAM vendors set the I_{off} criterion as less than 1 fA/cell [34, 44, 45]. However, based on Table 2.2, 5 fA/cell would be a good criterion to satisfy, even though t_{REF} is 500 ms. Thus, we assume 5 fA is the requirement for gate transistor leakage current as our TCAD device simulation result.

The leakage current is inversely proportional to the threshold voltage of the device. Thus, when the threshold voltage and the trend are known, the remaining device design parameters can be approximated.

RCAT and SRCAT threshold voltage projections have been provided in references [22, 23, 29, 31, 46]. Among these empirical data, RCAT threshold voltages are collected as shown

Figure 2.5 Recessed gate transistor threshold voltage trend.

in Figure 2.5. For DATE model, it is assumed that the trend for RCAT threshold will best fit the straight line of the mean value of threshold data. Thus, the trend would follow equation (2.3).

$$V_{th-trend} = -0.0056 \times \text{Process Node} + 1.672 \quad (2.3)$$

The straight line in Figure 2.5 represents the RCAT threshold trend shown in Equation 2.3. The standard deviation of data from the trend line is 0.0664. For the SRCAT, the threshold voltage is assumed to be 200 mV lower than RCAT when all other conditions are kept constant [23].

Overall, for the recessed gate transistors, DATE admits the result to the roadmap when the I_{off} is less than 5 fA/cell, when comparing the result with the threshold projection (within the standard deviation range). For the VCAT, the leakage current is the only criterion for this study.

Figure 2.6 MASTAR graphical user interface [47].

2.1.2 High-Voltage and Peripheral transistor

Peripheral and high voltage transistor roadmaps are deployed with a Model for Assessment of cmos Technology And Roadmaps (MASTAR) from ITRS [47]. Figure 2.6 shows the graphical user interface of MASTAR. MASTAR has high performance (HP), low stand-by power (LSTP) and low operating power (LOP) process roadmaps with physical models of planar bulk, double gate (DG) and silicon on insulator (SOI) transistor. MASTAR could calculate expected transistor characteristics (i.e., on/off current, threshold voltage, mobility,

Table 2.3 ITRS Saturation Current Roadmap of Supportive NMOSFET at 25 °C

	DRAM Process Node (nm)												
	90	75	65	55	44	40	36	31	27	24	21	18	16
I_{sat-n} ($\mu A/\mu m$)	500	500	500	465	450	410	410	400	400	400	450	450	450

etc.) with several transistor geometry values like gate length, oxide thickness and so on. We assume peripheral and high voltage transistors to be planar bulk, and the additional fabrication process for peripherals would be optimized for speed with low leakage current. From this assembly, we rely upon MASTAR process assumptions along with Rambus size projections.

ITRS provides saturation current roadmap of supportive transistors as shown in Table 2.3. DATE admits the ITRS projection for adjusting channel doping concentration. Since the ITRS roadmap was generated at 25 °C, we extended temperature from 300 K to 400 K using MASTAR.

2.2 Interconnect

2.2.1 Wire

For the wire resistance and wire capacitance calculation, DATE adopts Horowitz wire model [48]. From the model, the general metal wire resistance is given by Equation 2.4:

$$R = \rho \frac{\text{Length}}{\text{Conductor's Cross-sectional Area}} \quad (2.4)$$

where ρ is resistivity.

Figure 2.7 Wire and wire cross section for resistance calculation [48].

Figure 2.8 Wire cross section for capacitance calculation [48].

In the case of copper wire, a thin barrier layer is needed on three sides to prevent copper from diffusing into surrounding oxides, as shown in Figure 2.7. The copper wire resistance per unit length is given as,

$$R_{unit-length} = \rho \frac{1}{(\text{Thickness} - \text{BT}) \times (\text{Width} - 2 \times \text{BT})} \quad (2.5)$$

As shown in Figure 2.8, copper wire capacitance consists of the surrounding sheet capacitance with fringe capacitance. The capacitance is driven as,

$$C_{unit-length} = C_{horizontal} + C_{vertical} + C_{fringe} \quad (2.6)$$

$C_{horizontal}$ and $C_{vertical}$ are given by the equation,

$$C_{horizontal} = 2 \times \epsilon_{dielectric} \epsilon_o \frac{\text{wire thickness}}{\text{wire spacing}} \quad (2.7)$$

$$C_{vertical} = 2 \times \epsilon_{ILD} \epsilon_o \frac{\text{wire width}}{\text{ILD thickness}} \quad (2.8)$$

For the general metal wire material, Horowitz and ITRS expected the technology would migrate from aluminum to copper because aluminum wires have a resistivity of $282 \Omega \cdot \text{cm}$ while copper wires have a resistivity of $170 \Omega \cdot \text{cm}$ at 20°C [48–50]. DATE adopts copper as a wire material as ITRS and Horowitz. Rambus model [13] and the cross-section of specific commodity DRAM [51] has shown that aluminum is used wire material in DRAM. However, even though copper has smaller resistivity compare to aluminum because of the thin barrier layer, wire resistance does not quite decrease as much as the ratio of two materials [48].

For the wordline and bitline, polysilicon or tungsten silicide or the combination of both are used in DRAM [51]. DATE assume tungsten silicide is used as a poly-wire material.

Tungsten silicide could have different resistivity according to the different process recipes. Higher temperature and longer time on annealing process give lower resistivity [52]. For calculating resistance, DATE use $80 \mu\Omega\cdot\text{cm}$.

Figure 2.9 shows a cross-section of interconnect architecture. Figure 2.9a shows the interconnect architecture of the general microprocessor. Figure 2.9b shows the cross-section view of the DRAM interconnect architecture. In Figure 2.9b, cylindrical capacitors are connected to the drain region of the RCAT, not the polysilicon bit line.

As depicted in Figure 2.9, in general, commodity DRAM has capacitors between poly and metal layer one (M1) at the cell region and uses fewer metal layers (overall two to four layers [13, 49]) than the microprocessor technology. In the technical report [51], DRAM uses a metal size similar to the global wire size of a microprocessor process.

ITRS provides the detail size projection for general microprocessor interconnect with dielectric material properties and effective copper resistivity according to the metal size [49]. ITRS also offers M1 pitch and contact resistance and few more information as the indicative key feature for the DRAM wire projection but the provided information is not detailed for revealing entire wire projection: there is no DRAM wire size projection.

From the DRAM cross-sectional report [51], we can find detailed physical dimensions of the entire DRAM wire layer of a specific commodity DRAM, but little detail for the dielectric material properties. Thus, for deploying DRAM wire roadmap, ITRS roadmap alone or the technical report alone is insufficient.

As in ITRS roadmap, DATE assumes copper as a base wire material. This allows DATE to use ITRS wire material property roadmap. In addition, DATE adopts the physical dimension from the cross-sectional report to construct the DRAM wire roadmap. For the bitline and wordline, DATE assumes aspect ratio of 2.2 in all technology node as similar in the cross-sectional report [51]. DATE also assumes polysilicon wire pitch of two feature-size ($2F$) for

(a) Cross-section of microprocessor [49].

(b) Cross-section of 6 F^2 layout DRAM [51].

Figure 2.9 Typical cross-section of interconnect architectures.

a condensed bit-cell array layout. Silicon-oxide is assumed as a dielectric material of poly-wires since the wires are used as a gate material of the gate transistor. The oxide thickness follows Rambus projection of the gate transistor.

DATE assumes three copper metal layers with polysilicon wordline and polysilicon bitline. DATE limits the use of the first metal layer (M1) to the inter-cell routing within small peripherals. There is a significant difference in the choice of inter-cell routing materials assumed between DATE and the technical report [51]. In the technical report, polysilicon plays a role of inter-cell routing layer. During peripheral circuit speed and energy calculations, the M1 capacitance is only included for energy calculations, but the M1 resistance is ignored for the speed due to the short routing distance. The M1 capacitance has a relatively small portion of the peripheral circuit compared to the capacitance of the transistors, so with either copper or polysilicon, the impact of the inter-cell routing layer on the entirety of the calculations in DATE is limited. Since the lack of physical dimensions of M1 for the inter-cell routes in the technical report, the width, pitch, and aspect-ratio of the M1 layer and other properties would follow ITRS M1 layer projection.

For the other metal layers, DATE adopts similar width sizes and aspect ratios from the cross-sectional report [51]. The metal layer two (M2) and the metal layer three (M3) of DATE match M1 and M2 of the cross-sectional report, respectively. The M2 wire pitch is assumed 8.8 times the feature size and the M3 wire pitch is assumed 15 times the feature size. The width of each wire is assumed half of the wire pitch. The aspect ratios are 1.5 and 1.75 for the M2 and M3, respectively. For the M2, effective resistivity and dielectric properties are following ITRS semi-global wire roadmap. For the M3, material properties are following ITRS global wire projection.

Once unit resistance and capacitance are deployed, DATE use the value to calculate the size and the location of inserted buffers/repeaters to reach the optimal performance of the

wire bus.

2.2.2 Through Silicon Via

Through silicon via (TSV) is an essential component in configuring 3D DRAM. TSVs are classified into different categories according to the fabricated order compared to the metal layer. DATE uses front-end-of-line (FEOL) TSVs which are fabricated right before the first metal layer processing. FEOL TSV enables the interconnection between the top metal of bottom die and the first metal layer of the top die. Thus, DATE adopts the analytic model of the FEOL-TSV proposed in reference [14, 53]. The equations presented in this section are taken from the TSV references [14, 53].

Figure 2.10 shows the cross-sectional view and top view between two stacked dies using the FEOL TSV. In the figure, r_{tsv} , r_{ox} , and r_{dep} are the radius of TSV, oxide, and depletion region respectively. Figure 2.11 shows a top view of the FEOL TSV bundles along with coupled capacitance.

The TSV resistance model is given by Equation:

$$R_{tsv} = \rho \frac{l_{tsv}}{\pi r_{tsv}^2} \quad (2.9)$$

where ρ is resistivity, and l_{tsv} represents the length of the TSV.

As depicted in Figure 2.10 and Figure 2.11, TSV capacitance consists of intrinsic capacitance with coupling capacitance. The TSV capacitance model is given by Equation:

$$C_{tsv} = C_{intrinsic} + C_{coupling}. \quad (2.10)$$

Figure 2.10 Cross section and top view of single TSV with capacitance [53].

Figure 2.11 TSV bundles and coupled capacitance [14].

The TSV intrinsic capacitance, $C_{intrinsic}$, is modeled by Equation:

$$C_{intrinsic} = \frac{C_{ox} C_{dep}}{C_{ox} + C_{dep}} \quad (2.11)$$

where C_{ox} and C_{dep} are given as following Equations:

$$C_{ox} = \frac{2\pi\epsilon_{ox}l_{tsv}}{\ln(r_{ox}/r_{tsv})} \quad (2.12)$$

$$C_{dep} = \frac{2\pi\epsilon_{ox}l_{tsv}}{\ln(r_{dep}/r_{ox})}. \quad (2.13)$$

In the DRAM layout, the TSVs are arranged close together as in Figure 2.11. Thus, there is coupling capacitance between the TSVs. The $C_{coupling}$ is modeled by the equation,

$$C_{coupling} = \alpha \frac{\epsilon_{si}}{S} \pi d_{tsv} l_{tsv}. \quad (2.14)$$

where α is a fitting constant which is accounting for technology and nonlinearity of coupling capacitance. The d_{tsv} is a distance between TSVs. For the detailed calculation for each technology nodes, DATE follows CACT-3DD size roadmap for a conservative size scaling; ITRS provide size roadmap of TSVs. CACTI-3DD adds conservative industry perspective on top of the ITRS projection.

DATE includes TSVs in the driving circuits. The drivers are inserted instantly before and after a TSV to ensure the driving strength of the TSV. The logical effort method is utilized to calculate the size and number of buffer stages [54]. The area calculations for TSVs includes the buffer chain unless the buffer chain can fit into the pitch of the TSV.

2.3 Roadmap and discussion

2.3.1 Gate Transistor

RCAT, SRCAT and VCAT have been simulated with Synopsys TCAD under the condition proposed in Table 2.1 and Figure 2.4. The simulation calculates gate capacitance, device turn-on/off currents (I_{on} , I_{off}) and threshold voltage. The simulation sweeps the temperature variation from 300 K to 400 K in 10 K increments, and runs at 358 K to check I_{off} current.

(a) RCAT simulation with uniform channel doping. (b) SRCAT simulation with asymmetric channel doping.

Figure 2.12 TCAD simulation snapshot of recessed transistors.

Figure 2.12 shows TCAD simulation of the recessed transistors. RCAT is simulated with uniform channel doping, and SRCAT is simulated with asymmetric channel doping. Since the simulation is performed in a 2D cross-sectional environment, TCAD computes the results per unit length. Figure 2.13 shows three-dimensional view and cross-section of

Figure 2.13 Three dimensional view and cross section of TCAD simulation of VCAT.

VCAT TCAD simulation. For the VCAT, the simulation results are per device value since the simulation runs on a single device. The detail simulation commands of 44 nm SRCAT device structures are included as an example in Appendix B.

Table 2.4 Gate transistor roadmap¹

Technology (nm)		90	75	65	55	44	40	36	31	27	24	22	18	16
Gate Capacitance (aF/device)														
Rambus	MOSFET	55.7	39.2											
	Recessed			41.0	32.9	24.2	21.2	18.4	14.9	12.5	10.7	8.7	7.0	6.1
CACTI-3DD	Unknown	76.5		46.8		24.1			14.5			8.35		
DATE *	RCAT	210.4	188.3											
	SRCAT			174.9	128.7	103.8	98.0	106.2	96.7	85.6	75.6	70.4	61.2	52.8
	VCAT	194.4		164.4		125.4			101.5			91.8		79.0
Ion current (μ A/device)														
CACTI-3DD	Unknown	20		20		20			20			20		
DATE *	RCAT	24.1	18.8											
	SRCAT			19.0	17.6	14.0	12.3	12.8	10.7	9.4	4.2	4.9	3.8	3.4
	VCAT	36.9		39.6		53.1			36.1			26.3		22.3
Ioff current (fA/device) at 85°C														
CACTI-3DD	Unknown	1		1		1			1			1		
DATE *	RCAT	3.1	5.0											
	SRCAT			2.1	3.9	4.2	3.3	2.0	1.5	1.3	1.2	1.1	0.9	0.9
	VCAT	4.9		2.9		2.0			1.4			4.9		3.6
V_{th} (V)														
Expected V_{th} **		1.17	1.25	1.11	1.16	1.23	1.25	1.27	1.30	1.32	1.34	1.35	1.37	1.38
DATE *	Recessed **	1.17	1.25	1.10	1.16	1.20	1.20	1.27	1.31	1.32	1.34	1.36	1.38	1.38
	VCAT	0.31		0.84		0.72			0.92			0.89		0.68

* Simulation result at 300 K

** RCAT at 75 nm, SRCAT at 65 nm ~ 16 nm

Table 2.4 shows results from the overall device simulation with CACTI-3DD and Rambus projection. Rambus provides capacitance projection on each node. CACTI-3DD provides

¹The Rambus and CACTI-3DD projection was derived and calculated based on the source code or data provided by the author.

Figure 2.14 Gate transistor roadmap.

90 nm, 65 nm, 45 nm, 32 nm, 22 nm and 16 nm process node projections in the source code. The CACTI-3DD source code, provided by the author, does not work below 22 nm, so we exclude 16 nm. Between these nodes, CACTI-3DD assumes linearly interpolated values at each node. Both Rambus and CACTI-3DD assume similar capacitance scaling projection as shown in Figure 2.14. For DATE model, TCAD simulation results exhibit 4 to 13 times larger capacitance compared to the CACTI-3DD projection from 75 nm to 16 nm nodes. These differences are to be expected since Rambus and CACTI-3DD estimate shorter and scaling down channel length while DATE assumes constant trench depth and pillar height as detailed in Section 2.1.1.

Figure 2.15 Detail view of source or drain junction of MOSFET [55].

For the I_{on} and I_{off} current, CACTI-3DD assumes $I_{on} = 20 \mu A$, $I_{off} = 1 \text{ fA}$ for every node as an ideal value. In DATE roadmap, we have tried to keep I_{off} below 5 fA with the lowest possible channel doping density while the threshold voltage met threshold voltage trend within the standard deviation (0.0665 V). With these conditions, I_{on} of SRCAT scaled to $3.4 \mu A$ at 16 nm node. For VCAT, we only have tried to meet I_{off} below 5 fA as mentioned above. With this condition, I_{on} of VCAT is above $20 \mu A$ for every technology node. The I_{on} simulation results of the recessed transistor are smaller compared to the CACTI-3DD assumption since I_{on} is inversely proportional to effective channel length. However, the lower current flow could be compensated by the smaller die size according to technology scaling. As a result, DRAM specification could be satisfied.

2.3.2 High Voltage and Peripheral Transistor

Table 2.5 shows capacitance and turn-on current roadmap of high-voltage (HV) transistors. To calculate the capacitance of a single high voltage (HV) transistor, we assume that the HV transistor in all roadmaps (i.e., Rambus, CACTI-3DD, and DATE) have a minimum

gate width of $3 F$ and also have a junction length of $3 F$ as depicted in Figure 2.15. As a result, CACTI-3DD roadmap expects most optimistic capacitance projection, comparing DATE to Rambus since CACTI-3DD expects least gate capacitance even though Rambus does not include side-wall and overlap capacitance. DATE exhibits the most conservative capacitance because DATE adopts the most conservative side-wall capacitance from ITRS MASTAR and assumes the most conservative gate capacitance mainly due to longer gate length expectation of Rambus roadmap. For the turn-on current, CACTI-3DD uses a fixed number on each node even through temperature changes. DATE follows ITRS roadmap at 25°C and reflects turn-on current change due to temperature changes based on MASTAR calculation. Overall, CACTI-3DD expects about two times larger current than DATE.

Table 2.6 shows capacitance and turn-on current roadmap of peripheral transistors. For the capacitance of a single device comparison, we assume that the peripheral transistors of all roadmaps have a gate width of $3 F$. Peripheral transistors are also assumed to have a junction length of $3 F$. Rambus roadmap expects most optimistic capacitance projection, comparing DATE to CACTI-3DD since Rambus does not include side-wall and overlap capacitance. Between CACTI-3DD and DATE, DATE exhibits more device capacitance because DATE adopts higher side-wall capacitance as discussed in the case of HV transistor and also expects more gate capacitance mainly due to longer channel length expectation of Rambus roadmap. For the turn-on current, CACTI-3DD also uses a fixed number on each node even though temperature changes. DATE follows ITRS roadmap at 25°C and reflects turn-on current change due to temperature changes based on MASTAR calculation. Over all tech nodes, DATE exhibits turn-on currents that are similar or lower than CACTI-3DD.

²The Rambus and CACTI-3DD projection was derived and calculated based on the source code or data provided by the author.

³The Rambus and CACTI-3DD projection was derived and calculated based on the source code or data provided by the author.

Table 2.5 High voltage transistor roadmap²

Technology (nm)	90	75	65	55	44	40	36	31	27	24	22	18	16
Capacitance (aF/device *)													
Rambus	765.4	637.1	536.3	445.9	350.6	313.1	276.9	234.2	204.8	175.0	150.1	125.9	109.3
CACTI-3DD	780.1		509.0		321.4			200.2			127.5		
DATE	1309.5	1120.3	955.2	820.5	653.0	570.9	503.7	438.8	379.0	327.8	282.6	238.6	209.0
Ion current ($\mu\text{A}/\mu\text{m}$)													
CACTI-3DD	1094.3		1031.0		999.4			1024.5			910.5		
DATE [†]	440.5	500.6	500.5	465.6	450.8	410.9	410.6	400.3	400.2	400.4	450.2	450.0	450.0
ITRS (Spec.) ^{†‡}	440.0	500.0	500.0	465.0	450.0	410.0	410.0	400.0	400.0	400.0	450.0	450.0	450.0

* Transistor minimum width is assumed as 3 times of the minimum feature size of each technology.

† When temperature at 25°C

‡ 90 nm follows MASTAR 90 nm LSTP projection. 75 nm and 65 nm nodes follow 68 nm node of ITRS 2007. 55 nm node follows ITRS2007 58 nm node. 44 nm and 40 nm nodes follow ITRS2009. 36 nm, 31 nm and 27 nm nodes follow ITRS2011. 24 nm and below nodes follow ITRS2013.

Table 2.6 Peripheral transistor roadmap³

Technology (nm)	90	75	65	55	44	40	36	31	27	24	22	18	16
Capacitance (aF/device *)													
Rambus	629.5	482.6	375.9	307.3	229.5	198.3	164.7	130.4	111.4	88.6	75.9	64.2	56.3
CACTI-3DD	912.1		559.0		346.6			222.7			141.9		
DATE	976.4	812.5	778.3	653.2	509.6	435.6	382.2	312.2	272.6	233.1	202.7	171.0	153.6
Ion current ($\mu\text{A}/\mu\text{m}$)													
CACTI-3DD	503.6		519.2		666.2			683.6			727.6		
DATE [†]	500.0	500.0	500.1	465.0	450.0	410.2	410.8	400.1	400.5	400.2	450.2	450.4	450.4
ITRS (Spec.) ^{†‡}	500.0	500.0	500.0	465.0	450.0	410.0	410.0	400.0	400.0	400.0	450.0	450.0	450.0

* Transistor minimum width is assumed as 3 times of the minimum feature size of each technology.

† When temperature at 25°C.

‡ 90 nm, 75 nm, and 65 nm nodes follow 68 nm node of ITRS 2007. 55 nm node follows ITRS2007 58 nm node. 44 nm and 40 nm nodes follow ITRS2009. 36 nm, 31 nm and 27 nm nodes follow ITRS2011. 24 nm and below nodes follow ITRS2013.

Table 2.7 DATE wire roadmap

Technology (nm)	90	75	65	55	44	40	36	31	27	24	22	18	16
Capacitance (fF/ μm)													
Poly-WL	0.762	0.722	0.684	0.632	0.568	0.549	0.531	0.502	0.476	0.429	0.401	0.346	0.330
Poly-BL	0.762	0.722	0.684	0.632	0.568	0.549	0.531	0.502	0.476	0.429	0.401	0.346	0.330
M1	0.326	0.326	0.326	0.327	0.332	0.332	0.332	0.338	0.343	0.310	0.301	0.281	0.278
M2	0.351	0.351	0.351	0.352	0.342	0.341	0.341	0.332	0.343	0.303	0.293	0.275	0.269
M3	0.350	0.350	0.350	0.351	0.344	0.344	0.344	0.337	0.345	0.303	0.295	0.275	0.268
Resistance ($\Omega/\mu\text{m}$)													
Poly-WL	44.893	64.646	86.068	120.21	189.113	227.273	280.584	378.394	498.815	631.313	760.153	1122.334	1420.455
Poly-BL	44.893	64.646	86.068	120.21	189.113	227.273	280.584	378.394	498.815	631.313	760.153	1122.334	1420.455
M1	2.997	4.109	5.865	8.570	14.267	18.563	25.559	37.191	52.297	72.393	88.295	148.321	205.873
M2	0.100	0.203	0.292	0.431	0.770	1.004	1.383	2.106	1.595	4.231	5.245	8.948	12.352
M3	0.032	0.053	0.075	0.109	0.135	0.195	0.274	0.395	0.545	0.780	0.980	1.675	2.391

2.3.3 Wire

Wire capacitance and resistance per μm have been calculated using the Horowitz equation for DATE. Table 2.7 shows the detailed result.

Figure 2.16 Wire resistance roadmap.

Figure 2.16 shows DATE wire resistance roadmap. As an absolute value, poly-wires (i.e., poly-wordline and poly-bitline) have the same highest resistance at 16 nm with $1420.46 \Omega/\mu\text{m}$ since DATE assumes the same physical dimension between the wordline and bitline. Poly-wires also have high resistance in the order of poly-wires, M1, M2, M3 wire. This sequence lasts from 90 nm. As for the relative change, the M3 wire has a difference of about 75 fold the difference of $2.391 \Omega/\mu\text{m}$ at 16 nm and $0.032 \Omega/\mu\text{m}$ at 90 nm. On the other hand, the

poly-wordline difference is about 32-fold. It is assumed that the conductor effective copper resistivity from the ITRS roadmap is gradually increased from $2.2 \mu\Omega\cdot\text{cm}$ at 90 nm to $6.88 \mu\Omega\cdot\text{cm}$ at the 16 nm node while polysilicon resistivity is assumed as $80 \mu\Omega\cdot\text{cm}$ in all technology nodes.

Figure 2.17 shows DATE wire capacitance roadmap. Overall, wire capacitance decreases as technology advances. The poly-wire has the highest capacitance in all node, mainly due to the wire pitch is the smallest. The M2 and M3 has a similar aspect ratio as 1.5 and 1.75 respectively, the same distance ratio (i.e., half of the wire pitch), and a similar dielectric between wires. Thus, the M2 and M3 have similar capacitance on all nodes as shown in Table 2.7.

Figure 2.17 Wire capacitance roadmap.

Figure 2.18 compares the capacitance of M2 and M3 with the roadmap of ITRS and CACTI-3DD. M2 of DATE corresponds to the semi-global metal, and M3 corresponds to the global metal of ITRS and CACTI-3DD. Since DATE adopts material properties from ITRS roadmap, the difference between ITRS and DATE is due to geometry prediction differences.

Figure 2.18 Metal capacitance comparison⁴.

CACTI-3DD follows original CACTI wire projection [12]. CACTI assumes different material properties and physical dimensions from ITRS. In M2 and M3 layer, DATE has the most conservative projection. The M2 and M3 projections in Figure 2.18 are approximately twice the capacitance value compared to the ITRS-aggressive predictions.

Figure 2.19 compares the resistance of M2 and M3 with the roadmap of ITRS and CACTI-

⁴The ITRS roadmap was calculated from the ITRS physical dimension and material roadmap. The CACTI-3DD roadmap is derived from the source code.

⁵The ITRS roadmap was calculated from the ITRS physical dimension and material roadmap. The CACTI-3DD roadmap is derived from the source code.

(a) Metal 2.

(b) Metal 3.

Figure 2.19 Metal resistance comparison⁵.

3DD. Among M2 and M3 layers, ITRS M2 perspective is the most conservative. Except for ITRS M2 and ITRS M3-aggressive cases, all nodes have resistances of less than $15 \Omega/\mu\text{m}$. In M2 layer roadmap, DATE expects the smallest resistance in all nodes except CACTI-3DD aggressive projection. In M3 layer roadmap, DATE expects the smallest resistance in all nodes mainly due to it have the largest physical dimension compare to the ITRS and CACTI-3DD.

For comparison, we choose three commodity logic design processes. The normalized values of wire capacitance and resistance across three anonymous processes with those of DATE, ITRS, and CACTI-3DD, are presented in Table 2.8.

CACTI-3DD has about 5% to 35% more capacitance than anonymous processes even though CACTI-3DD assumes SiO_2 (dielectric constant: 3.9) is used as a dielectric material while ITRS project different dielectric constant in each technology node [49]. CACTI also

⁶The ITRS roadmap was calculated from the ITRS physical dimension and material roadmap. The CACTI-3DD roadmap is derived from the source code. The values of ITRS and CACTI-3DD are the average values of aggressive and conservative roadmap.

Table 2.8 Wire comparison with commodity logic design process⁶

Wire	DATE				ITRS			CACTI-3DD		
	Poly, Wordline	M1	M2 (Semi.)	M3 (Global)	M1	Semi-Glob.	Global	M1	Semi-Glob.	Global
Capacitance (%)										
45 nm	384.7	146.5	160.6	92.1	88.3	94.2	92.1	121.4	134.2	136.9
65 nm – A	404.7	155.8	144.9	158.2	93.2	78.4	97.2	136.2	117.7	137.9
65 nm – B	353.7	128.4	141.5	108.5	76.8	76.6	66.6	112.3	114.9	105.8
Resistance (%)										
45 nm	57.7	333.0	85.5	57.5	261.8	1180.1	642.5	66.4	117.9	107.5
65 nm – A	70.0	178.6	71.9	67.6	139.0	964.3	959.5	40.7	121.9	103.6
65 nm – B	84.8	336.7	20.2	21.4	262.1	271.3	303.4	76.6	34.3	32.8

uses constant bulk copper resistivity instead of using effective resistivity for the resistance calculation. This affects the most optimistic resistance prediction on the M1 layer.

On the other hand, ITRS assumes to use the different dielectric material and relative resistivity on each node. In the same node, ITRS assumes that the dielectric surrounding the wire with all metal layers is the same high-k material, whereas some actual processes use low-k material for the semi-global and global wire. With these differences, ITRS expects about 6% to 12% less capacitance on 45 nm node. In 65 nm A and B, ITRS projects about 3% to 33% less capacitance. Since ITRS has a different physical dimension with different resistivity due to different dielectric materials, resistance is 6- to 10-fold different in 45 nm and 2- to 9-fold different in 65 nm processes in M2 and M3 layer.

The anonymous processes are for the general logic design. However, DATE assumes DRAM process. Even though it is not an apple to apple comparison, in the case of M2 and M3 layers, it is meaningful to compare DRAM and general process wires. The M1 layer is also comparable: the M1 layer has a similar physical width and is surrounded by a

dielectric of high-k or SiO₂. The polysilicon layer is assumed to have larger aspect ratios than anonymous processes at DATE. Thus, DATE expects about three- to four-fold greater capacitance, with 16% to 42% less resistance than anonymous processes. In M1 layer, DATE exhibit about 1.3- to 1.6-fold greater capacitance with about 1.8- to 3.4-fold greater resistance. From this, we could expect DATE assumes smaller geometry and higher dielectric material than anonymous processes in M1 layer when M1 layer resistivity is equal. DATE expects about 14.5% to 80% less resistance than anonymous processes in M2 and M3. On the other hand, the capacitance is about 1.5-fold greater. When dielectric material between M2 and M3 layer are similar to anonymous processes, larger physical dimension assumption results in smaller resistance with larger capacitance.

2.3.4 Through Silicon Via

Through silicon via (TSV) is mainly made by etching or laser drilling. When the TSV is formed by etching, it is hard to achieve high etch rates, smooth sidewalls with controllable sidewall angle, and minimal mask undercut. When making a TSV with a laser, the masking and etching steps are not needed. Although this method has the advantage of reducing process step, it causes debris or splatters due to laser ablation [49].

Because of these challenges, the ITRS conservatively predicts scaling of the TSV. Table 2.9 shows ITRS TSV roadmap by year. In Table 2.9, despite technology advance, the recent ITRS roadmap assumes a size of TSV that is still similar or larger than previous years.

CACTI-3DD, on the other hand, assumed that the physical dimension of the TSV decreases as the technology advances. However, the assumptions of CACTI-3DD do not go beyond the premises of the ITRS. Table 2.10 shows CACTI-3DD TSV roadmap.

⁷Global interconnect level TSV size [49].

⁸Calculation based on most conservative prediction shown in Table 2.9.

Table 2.9 ITRS TSV physical dimension roadmap⁷

ITRS version (year)	ITRS 2009		ITRS 2011		ITRS 2013		ITRS 2015	
Expecting Year	2009 ~ 2012	2012 ~ 2015	2011 ~ 2014	2015 ~ 2018	2012 ~ 2014	2015 ~ 2018	2013 ~ 2014	2015 ~ 2018
Technology Included (nm)	54 nm ~ 32 nm	32 nm ~ 21 nm	36 nm ~ 25 nm	23 nm ~ 16 nm	45 nm ~ 32 nm	32 nm ~ 22 nm	28 nm ~ 26 nm	24 nm ~ 18 nm
Minimum Diameter (μm)	4 ~ 8	2 ~ 4	4 ~ 8	2 ~ 4	4 ~ 8	2 ~ 4	5 ~ 10	2 ~ 4
Minimum Pitch (μm)	8 ~ 16	4 ~ 8	8 ~ 16	4 ~ 8	8 ~ 16	2 ~ 8	10 ~ 20	4 ~ 8
Minimum Depth (μm)	20 ~ 50	20 ~ 50	20 ~ 50	20~50	20 ~ 50	20 ~ 50	40 ~ 100	30 ~ 50

Table 2.10 CACTI-3DD TSV physical dimension roadmap

Technology (nm)	90	70	50	40	30	21	16
Diameter (μm)	11.3	11.3	7.5	5	3.8	3.2	2.6
Pitch (μm)	90	90	60	40	30	25	20
Depth (μm)	75	75	63	50	38	32	17

Table 2.11 DATE TSV area, capacitance, and resistance roadmap

Technology (nm)	90	70	50	40	30	21	16
Area (mm^2)	0.0081	0.0081	0.0036	0.0016	0.0009	0.0006	0.0004
Capacitance (fF)	127.2	127.2	70.2	38.7	22.5	16.5	7.3
Resistance (Ω)	0.213	0.213	0.225	0.246	0.260	0.271	0.257

Table 2.12 ITRS TSV area, capacitance, and resistance roadmap⁸

Technology (nm)	90	70	50	45	30	21	18
ITRS year (version)	N.A.	N.A.	2009	2013	2013	2015	2015
Area (mm ²)	N.A.	N.A.	0.0003	0.0003	0.0001	0.0001	0.0001
Capacitance (fF)	N.A.	N.A.	158.8	158.8	115.0	115.0	115.0
Resistance (Ω)	N.A.	N.A.	0.118	0.118	0.172	0.172	0.172

In DATE, we adopt CACTI-3DD TSV roadmap since we assume TSV size would scale due to technology advancement. The Table 2.11 shows the DATE TSV roadmap calculated as described in Section 2.2.2. Compared to Table 2.12 DATE TSV roadmap exhibits larger area and smaller capacitance projection mainly due to larger pitch. The DATE TSV roadmap also exhibit larger resistance due to smaller diameter projection. This makes DATE area predictions more conservative and latency predictions faster than they would be if the ITRS TSV roadmap were used.

CHAPTER

3

DRAM CIRCUIT LEVEL MODELING

System level power models calculate system power by using the values shown in the vendor specification, while circuit level models calculate the resistance, capacitance, and area values of a single transistor. Circuit level models can be expanded upon to calculate the resistance, capacitance, and area of the logic composed of multiple transistors. The DRAM Area Timing and Energy (DATE) model is a circuit level modeling method. By adding more logic blocks with interconnects, the circuit level model calculates the area, speed, and energy of the system.

The circuit level model can perform unpredictable area and latency modeling with the system level model. However, the circuit level model can propagate the error of a smaller

module to the system level of error while the system level model calculates accurate results based on the vendor specification. Precise modeling is essential for each module in order to take advantage of the circuit level model.

Examples of circuit level modeling of the DRAM memory system are CACTI, CACTI-D, CACTI-3DD, and Rambus models introduced in Section 1.3. Rambus released a planar DRAM energy model in 2010 [13]. The Rambus model computes the capacitance and energy consumption based on a physical dimension scaling projection of a single device as well as the layout of the peripheral circuitry. The Rambus model also provides detailed DRAM architecture with peripheral circuitry including hierarchical wordline and bitline.

CACTI [12] is six-transistor (6T) SRAM cell based cache memory model that is widely used in computer architecture community to model SRAM cache memory. CACTI-D introduces one-transistor-one-capacitor (1T1C) DRAM cells and DRAM subarrays on top of the CACTI [56]. CACTI 5.1 inherits CACTI-D's DRAM model. However, the memory control path and data path were inherited from the SRAM. Thus, CACTI-D or CACTI 5.1 is more appropriate for modeling embedded DRAM than off-chip DRAM. CACTI-3DD [14] inherits CACTI 5.1 and adds banks and buses with TSVs for modeling three-dimensional DRAM. However, CACTI-3DD uses ideal device models, as shown in chapter 2, and does not support the emerging structure, such as $4F^2$ cell layout. DATE inherits benefits of CACTI with architectural assumptions of Rambus.

Figure 3.1 shows the program flow of the DATE circuit level model. After DATE reads the user configuration and technology roadmap, physical dimension and properties of the subarray are established and calculated. With the subarray geometry, bank size is calculated along with speed and energy of other bank components such as wordline driver and column select decoder. After calculating the bank properties, DATE computes the positions of the banks with the floor plan received from the user. TSVs are also inserted

Figure 3.1 DATE program flow.

at this time according to the user configuration. This produces the die size and signaling length of the DRAM. After these calculations, DATE obtains DRAM area, energy, and speed by adding all delays and energy for each part.

In this chapter, we first present component modeling such as transistor, driver, repeater, address decoder, and sense amplifier model. Along with the components, architectural level, subarray, bank and die level layout models are presented. For verification of the modeling method, the DATE model results are compared to energy and performance projections published in the data sheets of several commodity DRAMs. The DATE model results will also

be compared to energy and performance results from several state-of-the-art publications.

3.1 Component Modeling

3.1.1 General Layout and Drain Capacitance

Figure 3.2 Wide width transistor layout.

In the circuit level model, the logic gate area, turn on resistance, and capacitance is derived from the physical geometry obtained from a transistor layout. DATE assumes ideal layout design rules as shown in Figure 3.2 and Figure 3.3. Figure 3.2 shows a wide width transistor. In Figure 3.2, the poly length is given as a physical gate length technology projection in each technology. The poly width can be changed according to the user's

Figure 3.3 Folded transistor layout.

definition. The minimum source or drain width is defined as $1.5 F$, where F stands for the minimum feature size of the technology node. For a more conservative model, it is assumed that DATE occupies more of the $3 F$ source and drain width at the edge of the transistor. A contact is a square of length F on one side. When a contact exists edge of a source or a drain region, the contact is located $1 F$ away from the edge. The distance between contacts is also assumed to be $1 F$.

Figure 3.3 shows folded transistor layout. As in Figure 3.3, if contact is needed between folded polys, DATE admits contact among the active region. When the transistor is folded, DATE takes conservative capacitance values by adding a width of $2 F$ to one side of the edge region.

The drain region must be defined before calculating the drain capacitance, The drain region of the transistor shown in Figure 3.2 has the same width as the source. In this case, only one of the two regions are considered. DATE assumes a drain region of a folded transistor as shown in Figure 3.4. As shown in Figure 3.4a, when the total number of polysilicon

Figure 3.4 Drain region of the folded transistor.

is an odd number, the number of the drain region is given as:

$$\text{number of drain region} = (\text{number of poly} + 1)/2. \quad (3.1)$$

As shown in Figure 3.4b, when the total number of polysilicon is an even number, the number of the drain region is given as:

$$\text{number of drain region} = \text{number of poly} - 1. \quad (3.2)$$

In Figure 3.4, the gray color part represents the drain region. In both cases, DATE takes a conservative drain capacitance by including the gray colored, $2F$ more region as the drain. After determining the drain area, DATE calculates the drain capacitance by adding all of the capacitance as shown in Figure 2.15.

For automated logic block layout, DATE assumes several internal reserved spaces as shown in Figure 3.5. When the user sets a custom height, DATE reserves $2F$ for power rails, $5F$ as the region needed for inter-cell routing. Except for the reserved $7F$, DATE assigns the P-type active region and N-type active region according to user-defined P/N ratio. Within a given height, DATE determines whether or not to fold the transistor according to the

Figure 3.5 Internal layout height assumption.

transistor width. After determining the layout of a single transistor, the type of logic gate and the number of inputs determine the number of total transistors and whether to connect the transistors in parallel or series. DATE automatically allocates the area so that all the transistors can fit in a given height, and returns the width to the user for calculating overall area.

Figure 3.6 shows an example layout of two input NAND gates. In Figure 3.6, the PMOS

Figure 3.6 Two input NAND gate schematic and layout example.

Figure 3.7 Drain region of series-connected transistor.

Figure 3.8 DATE logic design assumption.

transistors are connected in parallel, and the NMOS transistors are connected in series. In the case of parallel connections, each transistor is placed side by side, with a spacing of $1.5 F$. In the case of a serial connection, transistors are integrated into one layout. The M3 and M4 transistors on the left side of the schematic are merged into a single layout connected to the bottom of the right side of Figure 3.6.

Figure 3.7 shows the drain region as a gray area when three transistors are connected in series. For the series-connected transistors, the number of the drain region is the same number of the transistor. The drain capacitance of the series-connected transistors is also calculated by adding all the capacitance shown in Figure 2.15 for the gray areas.

3.1.2 Digital Logic and Driving Buffer

DATE assumes that the buffer followed by the digital logic is driving the following logic or wire as shown in Figure 3.8. To calculate the size of the transistors utilized to build the logic and buffers, DATE utilizes the logical effort method that calculates the best number of stages in a multistage logic network with the balanced transistors size of each stage [54]¹. The logical effort method measures a gate delay in τ units. τ stands for the delay of an inverter which drives identical inverter without parasitic capacitance. The delay of logic,

¹All of the equations presented in this section are taken from Chapter 1 of Logical Effort [54].

Table 3.1 The logical effort of logic gates [54]

Gate type	NAND [†]	NOR [†]	pNOR [‡]	Inverter
Equation	$\frac{n+\gamma}{1+\gamma}$	$\frac{1+n\gamma}{1+\gamma}$	4/9	1

[†] n is number of input, γ is inverter's pullup and pulldown transistor width ratio.

[‡] Falling case, where the mobility ratio, μ is equal to 2.

measured in unit τ , is consist of summation of two-part: The parasitic delay, p and the effort delay, f :

$$d = f + p \quad (3.3)$$

The effort delay or the stage effort, f is the product of two factors:

$$f = gh \quad (3.4)$$

The logical effort, g is defined as the ratio of between the input capacitance of the logic gate and the input capacitance of an inverter which driving the same output current. Table 3.1 shows logical effort for inputs of logic gates. The electrical effort, h is defined by:

$$h = \frac{C_{load}}{C_{in}} \quad (3.5)$$

where C_{load} is the capacitance loaded at the output of the gate and C_{in} is the capacitance of the input of the gate.

In a multistage network, the path delay, D is the sum of each logic stage along the path:

$$D = \sum d_i = D_F + P \quad (3.6)$$

where D_F is the path effort delay and P is the path parasitic delay. The path effort delay is given by:

$$D_F = \sum g_i h_i \quad (3.7)$$

and the path parasitic delay is

$$P = \sum p_i \quad (3.8)$$

The path effort is given by:

$$F = G B H = \prod g_i h_i = \prod f_i \quad (3.9)$$

where G is path logical effort, H is path electrical effort, B is branching effort of the entire path, and f_i is an effort delay of a random stage i . The detail explanation and derivation is given in Appendix C. The total delay of multistage is the least when each stage has the same stage effort [54]. In N-stage logic, to achieve minimum delay, any stage effort is given by:

$$\hat{f} = g_i h_i = F^{1/N} \quad (3.10)$$

We use hat over the stage effort to indicate expression which achieves minimum delay. From Equation 3.6, the minimum delay of the N-stage logic network, D_{min} is driven by:

$$D_{min} = N F^{1/N} + P \quad (3.11)$$

In an inverter buffer chain as in Figure 3.8, the path effort is derived as:

$$F = \frac{C_{load,last}}{C_{in,first}} \quad (3.12)$$

since there is no branch, the path branching effort, B is one, and the path logical effort is also one since the logical effort of an inverter is one. For the driving buffer chain, Nils et al. [57] showed the optimum fanout of each inverter, that is, the optimum stage effort to achieve the least delay is within a range of 2.7 to 5.3 according to the technology dependency. Within the fanout range of 2.7 to 5.3, the delay difference is generally within 5% [58]. Since the study [57] showed the optimum stage effort was 3.98 for the large load driving stage buffers, DATE selects the optimum stage effort number of 4. This simplifies the optimum stage number calculation. From Equation 3.10, the stage effort can be represented as,

$$\hat{f} = F^{1/N} = 4 \quad (3.13)$$

$$\log_F 4 = \frac{1}{N} \quad (3.14)$$

$$N = \frac{\log F}{\log 4} \quad (3.15)$$

with Equation 3.12, the optimum stage number is derived as,

$$N = \frac{\log\left(\frac{C_{load,last}}{C_{in,first}}\right)}{\log 4} \quad (3.16)$$

The size of each inverter stage could be recursively derived from the loaded capacitance of the last stage. At stage i , the input capacitor, $C_{in,i}$ is simply derived from the equation 3.10 as,

$$\hat{f} = g_i h_i = \frac{C_{out,i}}{C_{in,i}} \quad (3.17)$$

where $C_{out,i}$ is equal to $C_{in,i+1}$ when there is no parasitic capacitance. We can rewrite the

equation as:

$$C_{in,i} = \frac{C_{in,i+1}}{\hat{f}} \quad (3.18)$$

At the last stage, $C_{in,i+1}$ is equal to C_{load} as shown in Figure 3.8. When the inverter's pullup and pulldown transistor width ratio is γ , and the NMOS input capacitance, $C_{in,i,NMOS}$ is given by:

$$C_{in,i,NMOS} = \frac{1}{1+\gamma} C_{in,i} \quad (3.19)$$

Therefore, we can calculate transistor width of each stage since the input capacitance is proportional to the width of the transistor.

Figure 3.9 Transistor size of inverter, NAND, and NOR gate.

DATE provides the inverter, NAND, and NOR gate as the base logic gate to support larger functional blocks like address decoder and driving buffer. The transistor size for the logic in Figure 3.8 is decided by the number of inputs and also the kind of logic. Figure 3.9 shows transistor size of reference inverter, two input NAND, and NOR gate which have same driving

Figure 3.10 Horowitz gate model (amended from Horowitz [59], Figure 5.5).

force as a reference inverter. In Figure 3.9, W represents the minimum NMOS transistor width. To simplify, pullup and pulldown transistor ratio (γ) is assumed of 2. In Figure 3.9, the transistor size of NAND and NOR gate is adjusted to have the same driving force as the reference inverter. In the NAND gate, PMOS transistors are connected in parallel, and NMOS transistors are connected in series. In the NOR gate, PMOS transistors are connected in series, and NMOS transistors are connected in parallel. When the logic gate has N inputs, adding one more input leads to transistor size change. According to Ohm's law, adding one more NMOS transistor in series from N to $N + 1$ changes the transistor size as NW to $(N + 1)W$ in the NAND gate since the resistance of a transistor is inversely proportional to the width. For the NOR gate, the PMOS size would be $2NW$ to $2(N + 1)W$. For the driving buffer, DATE assumes a reference inverter for the first stage of the driver. Once the width of transistors is decided, the total area would be calculated as in Section 3.1.1

For the energy calculation of the gate, DATE accounts for drain-out charges by adding

Figure 3.11 Two input NAND gate.

capacitance of every node since the dissipated energy is given by the equation [55]:

$$E = C_L \times V^2 \times P_{0 \rightarrow 1} \quad (3.20)$$

where C_L is the sum of the intrinsic capacitance of the gate and loaded capacitance of the output. $P_{0 \rightarrow 1}$ is the probability that the device would consume energy. DATE fully accounts for power dissipation when a capacitor is charged and ignore the discharge event as CACTI [12].

A gate can be simplified as shown in Figure 3.10. At the node A, there is an intrinsic capacitance in the X direction. To calculate the intrinsic capacitance, we accounted for every drain capacitance connected in node A. For example, for a two NAND gate as shown in Figure 3.11, we added drain capacitance of T1 and T2 as discussed in Section 3.1.1. For the series connect transistors, T3 and T4, we accounted for the drain region as depicted

in Figure 3.7 and calculated the drain capacitance accordingly. In the Y direction, there is a load capacitance. Since the load capacitance is the sum of the gate capacitance of the transistor which connected node A, we added every gate capacitance of connected node A.

For the speed calculation, DATE uses Horowitz approximation as seen in [59] eq. (5.5).

$$t_d = \tau_f \sqrt{(\ln 0.5)^2 + 2\alpha\beta(1 - 0.5)} \quad (3.21)$$

where β is the gate turn-on voltage which is normalized range 0 to 0.5. DATE adopts number of 0.5 for the β for conservative approximation. The α is the normalized rise time constants when input change zero to logical one and defined as

$$\alpha = \tau_{in} / \tau_f \quad (3.22)$$

where τ_{in} is input rise time and τ_f is intrinsic gate delay which is $\tau_f = R_d(C_{intrinsic} + C_{load})$. The total delay of the gate and buffer chain shown in Figure 3.8 is the sum of the delays of each node, which is calculated using Equation 3.21.

3.1.3 Repeater for Wire

When wire length linearly increases, the delay of wire increases quadratically since both resistance and capacitance increase linearly. Also, large wire load on the driver leads to excessive short-circuit power dissipation on the last stage of the driver, which is due to the degrading of the waveform shape [60]. The general design approach is to introduce a repeater to resolve the problems caused by large wire loads. In the DATE model, Rabaey's approach is adopted for the repeater model [55]. Figure 3.12 shows interconnect line with the repeater. R_d , C_{out} , and C_{in} represent the resistance, output capacitance and input

Figure 3.12 Interconnect line with repeater [55].

capacitance of a minimum size inverter respectively. r and c represent the resistance and capacitance per unit length, respectively. The optimum size inverter ratio, S_{opt} , for the minimum size inverter is given by

$$S_{opt} = \sqrt{\frac{R_d c}{r C_{in}}} \quad (3.23)$$

as seen in [55], eq. (9.10). The optimum wire length which is segmented by the inverter is given by:

$$L_{crit} = \sqrt{\frac{t_{p1}}{0.38 r c}} \quad (3.24)$$

as seen in [55], eq. (9.11). Here, $t_{p1} = 0.69 R_d C_{in} (1 + \gamma)$ represents the delay of a fanout of one inverter. γ stands for the ratio of input capacitance and output capacitance of a minimum size inverter. Thus, Equation 3.24 can be simplified as

$$L_{crit} = \sqrt{\frac{t_{p1}}{0.38 r c}} \simeq \sqrt{\frac{2 R_d C_{in} (1 + C_{out}/C_{in})}{r c}} = \sqrt{\frac{2 R_d (C_{in} + C_{out})}{r c}} \quad (3.25)$$

By using Elmore delay approach [61], propagation delay of the interconnect is given as:

$$t_{p,crit} = m(0.69 \frac{R_d}{s} (s \gamma C_{in} + c L/m + s C_{in}) + 0.69 (r L/m) (s C_{in}) + 0.38 (L/m)^2) \quad (3.26)$$

as seen in [55], eq. (9.9). Here, s , m , and L represent relative repeater size compared to the minimum inverter, the number of the repeater in the interconnect and the interconnect length respectively. When the interconnect length is the number of 1, the number of optimized repeater m is equal to $1/L_{crit}$. Thus, the unit length interconnect propagation delay with the optimized repeater is derived as:

$$t_{p,opt,unit} \simeq 0.69/L_{crit}(R_d(C_{out} + C_{in}) + R_d c L_{crit}/S_{opt} + r L_{crit} S_{opt} C_{in} + 1/2 r c L_{crit}^2) \quad (3.27)$$

DATE supports delay penalty to save energy by sacrificing speed; after calculating the optimum size repeater interconnect delay, wire length and repeater size are matched to the user specified delay penalty by decreasing the size of the repeater slightly as well as increasing the wire length slightly. To calculate energy, DATE accounts for the capacitance at the wired node and multiplies it by the voltage square as discussed in Section 3.1.2

3.1.4 Address Decoder

DATE provides a two-stage address decoder for both row and column address decoding as shown in CACTI5.1 [62] and described in Rambus model [13]. Figure 3.13 shows the nine-bit row address decoding path from the input to each main word line (MWL) as an example. After the MWL, there is a sub-wordline (SWL) which is driven by the inverter buffer. The row address path consists of the predecoder stage and following second stage decoder blocks.

In the row address path, the predecoder stage includes two operable predecoder blocks operable to generate output signals in response to the input address signals. Each predecoder block consists of two-level decoding logic. In the first stage of predecoder, addresses are decoded using up to three 2-to-4 or 3-to-8 base decoders in parallel. The outputs of these base decoders generate the final predecoder signal output by using NAND gates. Two

Figure 3.13 Nine bit row address decoding path.

input NAND gates are used when there are two base decoder blocks, and three input NAND gates are used when there are three base decoder blocks. Figure 3.14 shows a 2-to-4 base decoder and a 5-bit address input predecoder block as an example. Since one predecoder could have three 3-to-8 base decoders, the predecoder stage is operable to handle 18-bit addresses.

The second decoder stage in the row address path consists of a plurality of decoding logic which is comprised of NOR gate and inverter as shown in Figure 3.13². The inputs of NOR gate are connected to the output signals of two predecoder blocks, and the following inverter driver drives the MWL. Figure 3.15 shows schematics of two input NOR gates. In the row address path, The NOR gate drains out the stored internal charge when the driving output is not selected. To reduce drain charge at the second stage, DATE adopts a pseudo NOR gate as shown in Figure 3.15b and assumes to control *pEN* pin. The pair of NOR gate and inverter driver of the second stage are grouped by subarray and the *pEN* pin is enabled

²The Rambus model describes it as a pair of pNOR gate and inverter driver in source code [13].

Figure 3.14 Predecoder structure (amended from CACTI5.1 [62], Figure 14).

Figure 3.15 Two input NOR gate.

per subarray. The subarray is the minimum cell group unit present in the DRAM bank and is described in detail in Section 3.2.

The column address path also has a two-decoding stage to generate column select (CS) signals. The difference from the row address decoder is that the driving buffer of the predecoder block is an inverter driver and uses the static NAND gate for the second stage. Since the static gates are used in the column address decoding path, the power consumed is less than the row path. In each address path, the first and the second stage is assumed to be located at the side of a bank. Address decoding delay is considered to be equal to critical path speed.

3.1.5 Bitline and Bitline Sense Amplifier

Figure 3.16 Bitline sense amplifier [13, 63].

Figure 3.16 shows the schematic of the bitline sense amplifier. The bitline and complement bitline are precharged at half the voltage storage capacity of the storage capacitor by using an equalizer. When wordline is enabled during the read operation, voltage differences occur in the bitline pair due to the current flowing from the capacitor. The voltage difference acts as an input to the cross-coupled inverter sense amplifier. DATE adopts the method used by CACTI5.1 and Horowitz to analyze delay of bitline and sense amplifier [59, 62]³.

As an ideal case, the voltage difference without wordline rise time effect is limited by following Equation:

$$V_{sense-max} = \frac{V_{storage-cap}}{2} \frac{C_{storage}}{C_{storage} + C_{bitline}} \quad (3.28)$$

In practical, the sense amplifier input voltage is given by

$$V_{SA-input} = \frac{V_{storage-cap}}{2} \frac{C_{storage}}{C_{storage} + C_{bitline} + (C_{gate,drain} \times \text{number of cell in bitline})} \quad (3.29)$$

The delay for the differential signal to develop is proportional to the RC time constant with the voltage ratio of practical and ideal case. The delay is given by the equation:

$$T_{step} = R_{in} \frac{C_{storage} C_{bitline}}{C_{storage} + C_{bitline}} \frac{V_{SA-input}}{V_{sense-max}} \quad (3.30)$$

where R_{in} is the series resistance of the gate transistor and the bitline. The T_{step} is equal to the time to store the data in the DRAM cell because, during write operation, data is coming from the I/O buffer, which is not affected from the wordline rise time. As a result, the transfer charge between the bitline and the cell capacitor in the read operation and the write operation is the same.

³All the equations in this section are taken and derived from the Section 6.1 and 9.3 of CACTI5.1.

According to the CACTI [12], the bitline delay with the effect of the wordline rise time is given by the equation:

$$T_{bitline} = \begin{cases} T_{step} + \frac{V_{WL} - V_{th,gate}}{2m} & \text{if } T_{step} > 0.5 \frac{V_{WL} - V_{th,gate}}{m} \\ \sqrt{2T_{step} \frac{V_{WL} - V_{th,gate}}{m}} & \text{if } T_{step} \leq 0.5 \frac{V_{WL} - V_{th,gate}}{m} \end{cases} \quad (3.31)$$

$$(3.32)$$

where wordline rise signal slope, m is given as:

$$m = \frac{V_{WL}}{T_{WL,rise}} \quad (3.33)$$

Equation 3.31 shows the case where the wordline rises rapidly.

To estimate the sense amplifier area, DATE adopts the sense amplifier layout of Samsung DRAM depicted in [11]. In the study, the cross-coupled inverter sense amplifier is derived to be about $60 F$ and a short length of $6 F$. Based on the derivation, DATE adds $30 F$ on the long side for the equalizer and the column select transistor. Therefore, the length of the entire long side is estimated to be $90 F$.

For the energy calculation, DATE calculates the drain capacitance of the transistors which consist of the sense amplifier, connected to the bit line and multiplies it by the bitline voltage and the supply voltage as discussed in Section 3.1.2

3.2 Architecture Level Modeling

Figure 3.17 shows the floor plan of a typical 8-bank double data rate (DDR) DRAM [13, 64]. A bank is set of independent memory arrays inside DRAM which has different channels in response to the command. In Figure 3.17, gray box corresponds to the bank. The bank

Figure 3.17 Eight bank DDR DRAM floor plan [13, 64].

comprises a group of subarrays. Row logic is placed in between banks to decode row addresses and to drive the main wordline. Column logic is also placed at the other edge of the bank to decode column address and to drive column select signal. The secondary sense amplifier exists along with the column logic to drive data line. Data, address, and control I/O pads exist on I/O pad area along with the transceiver circuits. The center stripe has control logic and the power system such as charge pump and voltage regulator to support bank operation. When TSVs are needed to support a 3D floor plan, DATE assumes that the TSV is located on the center stripe as introduced in the study [65]. Since banks share the logic and I/Os, concurrent operation of banks is limited. In addition to the DDR DRAM shown in Figure 3.17, several architectures such as hybrid memory cube (HMC) and Wide IO have been proposed [2, 4]. All of these architectures add I/O channels to avoid congestion or adjust the size of the banks according to the application. The basic floor plan concept of DDR DRAM is not different, which places the banks and shares the control logic among the banks.

DATE allows the user to determine the placement of the banks within the die. The user is required to input how many banks are arranged in the WL direction and the BL direction, or in the top and side directions of the die, respectively. For example, in Figure 3.17, there are four banks in the WL direction and two banks in the BL direction in the case of the 8-bank DRAM. Based on the size of the bank and the layout of the bank entered by the user, DATE prefers by placing the center stripe in the center of the die where the length is short so that the die shape approaches the square.

Figure 3.18 shows the enlargement of subarray area. A subarray is a bit-cell array surrounded by a row of sub-wordline drivers and sense-amplifiers with wordline control bus. Main wordline, column select, and master array data lines play a role of interface between the cell array and outside of the bank. During a read operation, the main wordline is se-

Figure 3.18 Subarray and related peripheral circuits [66].

lected by the row decoder. Inside the subarray, the main wordline is connected to four sub-wordline drivers. One of the sub-wordline drivers is selected by a wordline control signal and then drive selected wordline by the row address. The conceptual schematics of a sub-wordline driver and wordline control signal driver are shown at right side of Figure 3.18. After bitcells are enabled by sub-wordline, a voltage difference occurs due to the charge stored in the bitcell capacitor. Column select signal is enabled by the column address and select data to read. The bitline sense-amplifiers sense the difference and drive out the selected data through master array data lines. During a write operation, after sub-wordline is enabled, a sense amplifier drives charge to write data to a bitcell capacitor. The size of subarray is determined by performance and the density of the memory. Typically, sub-wordline and bitlines are connected with 256 to 512 cells, and the column select line and main wordline go across 16 to 32 subarrays [11, 13]. As discussed in Section 3.1.4, the

predecoder model supports maximum 18-bit address. Thus, DATE supports up to 20-row address bit due to 4-bit wordline control signal and support maximum 18 column address bit since there is no additional control signal for column select line.

DATE reads cell-layout type, subarray size, row and column address size, total memory size, and a number of banks from user input. The DATE support subarray floor plan supported by commodity DRAM - the user can select the floor plan between $8 F^2$ and $6 F^2$ cell layout. DATE utilizes user input to place banks - the user assigns number of banks in the row direction and column direction. DATE follows general floor plan between banks and peripheral logic as shown in Figure 3.17. The address path is assumed to receive input from approximately half the distance from the center of the DRAM to the edge. The data path is also assumed to have an I/O pads and transceiver in the same part as the address path. Each address path is routed through the control logic located at the center. The row address decoder located between the banks, and the column address decoder adjacent to the center stripe. To calculate the delay and energy in the worst case scenario, DATE assumes access to the farthest memory cell from the center as shown in Figure 3.17. It is assumed that the data path goes through the same path as the column address.

DATE supports additional subarray floor plans for the emerging device, referred to as vertical channel access transistor (VCAT). Figure 3.19 shows the schematic diagram of subarray for the conventional $6 F^2$ DRAM. The yellow box corresponds to the sense amplifier. The bitline pitch is $3 F$, thus, sense amplifier could have layout width of $6 F$ and could be arranged in the same column side by side with the bitline. The left side of Figure 3.19 shows the aligned layout between bitline and the sense amplifier. The right side of Figure 3.19 corresponds to the simplified version of Figure 3.18. The sense amplifiers are aligned with bitline as a line. Figure 3.20 shows the schematic diagram of subarray for the $4 F^2$ DRAM. In $4 F^2$ DRAM, it is hard to achieve a $4 F$ width sense-amplifier layout

Figure 3.19 Schematic diagram of primitive core array for the conventional $6 F^2$ DRAM [11].

without changing the basic circuitry considering general design rules such as minimum required number of contacts, space between active regions, and poly spaces. [11]. In a $4 F^2$ cell layout DRAM case, DATE assumes a sense amplifier layout that is tilted by 90 degrees with wordline strapping as proposed in the publication [11]. The right side of Figure 3.20 represent 90 degrees tilted sense amplifier array. Metal 1 (M1) is used for the connection between buried bitline and the assigned sense amplifier. The left side of Figure 3.20 shows the brief schematic of VCAT based subarray. Compared to Figure 3.19, the cell area is reduced due to the cell area factor transition from $6 F^2$ to $4 F^2$. Instead of the sub-wordline driver, $4 F^2$ based subarray has global wordline driver which drives M1 global wordline. DATE assumes the global wordline driver had the same schematic and logical function as the sub-wordline driver and located in between subarrays as well. DATE also assumes

Figure 3.20 Schematic diagram of primitive core array for the 4 F^2 DRAM [11].

wordline strapping model as shown in the publication [11]. Thus, M1 global wordline is connected to the poly gate line.

To calculate area, energy, and speed of entire DRAM, DATE first performs calculations on the subarray and calculates the cell area of the bank based on the subarray. After calculating the energy and speed for the required address decoder with the row and column addresses input by the user, the area of the bank is calculated by adding the width of each decoder to the edge of the bank. DATE then estimates the floor plan of the entire die from the user input, as shown in Figure 3.17. The total energy usage and latency of the DRAM is the sum of all charge and discharge events and latency of each component along the path shown in Figure 3.17. The area of the DRAM die is also calculated as the sum of the areas of all components. Since DATE does not have an appropriate model of I/O pads, power systems,

and control logic, DATE accounts for the estimated center-stripe area derived from several documentation and publications of 80 nm to 30 nm technology along with the calculated bank area [11, 64, 65, 67, 68]. In a commodity planar DRAM, the estimated additional center stripe area is about 20% of the sum of the entire bank area.

3.3 Validation

For the validation, the DATE model results are compared to energy, and speed published in the data sheets of several commodity DRAMs across various technologies and different DRAM generations [43, 69–83]. Instead of reporting the operating energy directly, the datasheet reports current values for various operating scenarios according to the JEDEC standard [3]. To calculate the precise operating energy according to the datasheet, we used the system level model, DRAMPower, proposed by Chandrasekar et al. [16, 84]. Since the area information is not available in the datasheet, the area result calculated at DATE is compared with the estimated area derived from the die photos of several documentations of 80 nm to 30 nm technology [64, 67, 68]. For validating 3 D DRAM and VCAT based DRAM, the DATE model results are compared to the area, energy, and speed from several state-of-the-art publications [11, 65, 85].

DATE calculates each energy event for active, burst-read, burst-write, and precharge command. Active command energy is the energy that enables circuits belonging to the path until a row address is decoded and a wordline is selected. Both the read and write command energy includes to the energy that enables the circuits which belong to the path until the desired bit decoded by reading the column address, among the selected pages by the active command. In addition, the write energy includes received data path until to write the selected bit cells with input data and the read energy includes read out data path

Table 3.2 Validation of energy calculation

Tech. (nm)	Validation Target DRAM Part Name [Ref.]	Active (nJ) DATE/Spec.†	Burst Read (nJ) DATE/Spec.†	Burst Write (nJ) DATE/Spec.†	Precharge (nJ) DATE/Spec.†
80	K4T1G084QC-ZCE6 [43]	2.008/2.025	0.763/0.756	0.758/0.756	1.051/1.080
	K4T1G084QA-ZCD5 [69]	1.895/1.856	1.070/0.945	0.880/0.945	1.015/1.080
	K4T51083QE-ZCE7 [70]	1.993/2.024	0.689/0.765	0.508/0.495	1.049/1.013
68	K4T2G084QA-HCF7 [71]	1.359/1.215	0.972/1.080	0.791/0.810	0.904/0.945
	K4T1G164QQ-HC(L)E6 [72]	2.611/2.838	1.203/1.135	0.667/0.703	1.334/1.351
	K4T1G084QQ-HC(L)E6 [72]	1.585/1.622	0.786/0.703	0.609/0.541	0.836/0.946
	K4T1G084QD-ZCE6 [73]	2.069/2.138	0.926/0.912	0.747/0.798	1.063/1.140
50	MT42L64M32D1KL [74]	2.653/2.815	1.086/1.156	0.982/0.991	1.461/1.310
	MT47H64M16HR-25E [75]	3.203/3.240	1.034/1.035	1.189/1.125	1.531/1.575
	MT47H128M8CF-25E [75]	2.184/2.304	0.725/0.783	0.938/0.828	1.085/0.999
44	EDJ2108BCSE [76]	0.258/0.270	0.671/0.721	0.817/0.766	0.424/0.405
	H5TQ2G83BFR [77]	0.791/0.811	0.590/0.568	0.507/0.568	0.378/0.405
42	MT41K256M8DA-125:M [78]	0.346/0.378	0.587/0.594	0.464/0.493	0.463/0.464
35	K4B2G0846D-HYK0 [79]	0.503/0.525	0.433/0.413	0.461/0.450	0.616/0.516
30	MT41J512M8RH-093:E [81]	1.217/1.316	0.818/0.827	0.731/0.636	0.692/0.650
	MT41J256M16HA-107:E [81]	1.217/1.234	1.447/1.305	0.928/0.906	0.702/0.794
	MT41K512M8RH-125:E [80]	0.835/0.893	0.820/0.893	0.733/0.653	0.520/0.474
	MT41K256M8DA-125:K [82]	0.465/0.368	0.510/0.488	0.513/0.510	0.369/0.392
	EDJ2108BDBG-GN-F [83]	0.605/0.630	0.458/0.503	0.524/0.540	0.411/0.454
Mean Percentage Error (%)		-1.39	0.20	1.74	-0.29
Standard Deviation (%)		7.88	7.56	7.38	8.30

† The Specification results are calculated by DRAM-power [84].

including transmission circuit from the selected bit cells. The precharge command energy is the energy for the circuits that are enabled to charge the bitline to the half of $V_{bitcell}$ after the active, read, or write command is executed.

Table 3.2 shows the comparison of DATE energy results with the calculated energy from the specification, based on the system level model [84]. Listed commodity DRAMs in Table 3.2 are planar DRAMs. The source of the process nodes for each DRAM is described in detail in Appendix D. The DATE energy calculation results yield an accuracy with the mean percentage errors of -1.39% to 1.74% and the standard deviation of less than 8.30%.

Table 3.3 Validation of key timing parameter calculation

Tech. (nm)	Validation Target DRAM Part Name [Ref.]	t_{RCD} (ns) DATE/Spec.	t_{CAS} (ns) DATE/Spec.	t_{RP} (ns) DATE/Spec.	t_{RRD} (ns) DATE/Spec.	t_{WR} (ns) DATE/Spec.
80	K4T1G084QC-ZCE6 [43]	11.67/15.00	13.28/15.00	11.79/15.00	5.27/7.50	14.45/15.00
	K4T1G084QA-ZCD5 [69]	11.52/15.00	14.79/15.00	11.43/15.00	5.31/7.50	15.40/15.00
	K4T51083QE-ZCE7 [70]	11.63/12.50	12.20/12.50	11.75/12.50	5.24/7.50	13.41/15.00
68	K4T2G084QA-HCF7 [71]	11.96/15.00	13.74/15.00	10.60/15.00	5.89/7.50	11.38/15.00
	K4T1G164QQ-HC(L)E6 [72]	14.51/15.00	11.31/15.00	10.66/15.00	7.30/10.00	12.73/15.00
	K4T1G084QQ-HC(L)E6 [72]	11.65/15.00	12.97/15.00	10.19/15.00	5.04/7.50	13.38/15.00
	K4T1G084QD-ZCE6 [73]	11.65/15.00	12.84/15.00	12.11/15.00	5.04/7.50	15.06/15.00
50	MT42L64M32D1KL [74]	15.72/18.00	11.84/15.00	24.83/21.00	6.68/10.00	12.13/15.00
	MT47H64M16HR-25E [75]	11.94/12.50	9.51/12.50	11.39/12.50	5.26/7.50	13.45/15.00
	MT47H128M8CF-25E [75]	11.12/12.50	11.03/12.50	14.35/12.50	4.53/7.50	16.32/15.00
44	EDJ2108BCSE [76]	10.64/13.50	9.63/13.50	8.68/13.50	4.52/6.00	9.46/15.00
	H5TQ2G83BFR [77]	10.51/13.50	9.71/13.50	8.33/13.50	4.51/6.00	9.04/15.00
42	MT41K256M8DA-125:M [78]	10.30/13.75	9.04/13.75	9.81/13.75	4.81/6.25	9.40/15.00
35	K4B2G0846D-HYK0 [79]	9.91/13.75	7.81/13.75	8.89/13.75	4.44/6.00	8.42/15.00
30	MT41J512M8RH-093:E [81]	11.48/13.13	7.48/13.13	14.25/13.13	4.38/5.60	14.88/15.00
	MT41J256M16HA-107:E [81]	11.74/13.91	8.40/13.91	14.00/13.91	4.68/5.35	15.69/15.00
	MT41K512M8RH-125:E [80]	10.97/13.75	7.99/13.75	8.81/13.75	4.49/7.50	10.53/15.00
	MT41K256M8DA-125:K [82]	8.98/13.75	7.78/13.75	4.03/13.75	4.39/5.00	5.93/15.00
	EDJ2108BDBG-GN-F [83]	9.80/13.75	7.45/13.75	5.87/13.75	4.24/5.00	7.37/15.00
All	Mean Percentage Error (%)	-19.25	-24.87	-23.06	-26.68	-19.38
	Standard Deviation (%)	8.43	14.90	23.02	7.87	20.46
Out of Spec.	Mean Percentage Error (%)	N.A.	N.A.	10.55	N.A.	5.36
	Standard Deviation (%)	N.A.	N.A.	7.73	N.A.	3.14

Table 3.3 shows the comparison of key timing parameter calculation results with the reported numbers from the specifications [43, 69–83]. t_{RCD} represents row address to column address delay - the period between the issuing of the active command and the read/write command. During t_{RCD} , the charge flows from the bit cells by enabled wordline, and the sense amplifier on bitline settles sufficiently to amplify the signal. t_{RCD} also includes the latency of row address decoding path. t_{CAS} represents column address strobe delay - t_{CAS} is the delay between receiving a read command and the moment the first piece of data is available on the output port. t_{CAS} includes the latency of column address decoding path and data out path. t_{RP} represents row precharge delay - after the page open, t_{RP} is the minimum period to issue another the active commands in the same bank. t_{RP} includes data write back delay from the sense amplifier to bit cell with bitline precharge delay and wordline reset delay. t_{RRD} is the minimum delay of successive row active to row active delay between different banks. t_{WR} is the write recovery time after the write command.

The mean percentage errors of DATE latency calculation for listed DRAMs, are -19.25%, -24.87%, -23.06%, -26.68%, and -19.38% for t_{RCD} , t_{CAS} , t_{RP} , t_{RRD} , and t_{WR} , respectively, with a standard deviation of less than 24%. From the results, DATE seems to predict each latency faster than actual DRAM speed. The measured shmoo plots in the latest publications [11, 65] show that the measured latency of each timing parameter is much faster than the timing specification in DRAM manuals. The timing numbers in the specification are not actual time delay; they are slower than the actual latency of the DRAM. Thus, the results comparing the specification and DATE calculation in Table 3.3 are not sufficient conditions to verify that the calculated latency from DATE model is the same or similar to the actual DRAM latency. However, the negative percentage errors are the necessary conditions, which indicate the calculation is in the right direction. The results highlighted in gray in Table 3.3 are the values outside of the specification. The mean percentage errors for the out of the specifications

Table 3.4 Validation of area calculation

Tech. (nm)	Validation Target	Bank Area (mm ²)	Die area (mm ²)
	DRAM Part Name [Ref.]	DATE/Target†/Error	DATE/Target†/Error
80	K4T51083QE-ZCE7 [70]	10.27/9.69/3.38%	50.14/49.00/2.33%
68	K4T1G084QD-ZCE6 [73]	6.88/7.15/-3.78%	67.19/69.60/-3.47%
50	MT42L64M32D1KL [74]	9.04/9.72/-6.78%	95.79/94.50/1.37%
44	EDJ2108BCSE [76]‡	5.71/5.75/-0.66%	51.25/51.60/-0.68%
44	H5TQ2G83BFR [77]‡	6.83/6.88/-0.67%	60.29/60.00/0.48%
35	K4B2G0846D-HYK0 [79]	3.53/3.84/-8.20%	34.69/36.00/-3.63%
Mean Percentage Error (%)		-0.60	-2.82
Standard Deviation (%)		2.49	4.36

† The target areas are derived from various documentations [64, 67, 68].

‡ The cell array is derived as 8 F^2 layout from the die photo [64] and the online article [86].

are about 10.6%, and 5.4% for t_{RP} , and t_{WR} , respectively, and all have standard deviations within 7.8%.

Table 3.4 shows the comparison of DATE model area results with the derived areas of the target planar DRAM designs. The bank area is composed of subarray areas including bitcell, the sense amplifier, and sub-wordline drivers. The mean area errors for the bank area is approximate -3% with a standard deviation of less than 4.4%. As discussed in Section 3.2, the die area includes about 20% of the additional peripheral area of total bank area. The mean area errors for the die area is about -0.6% with standard deviation of 2.49%.

To verify the 3D design and VCAT based design, we used data from several papers and commercial 3D DRAM specifications [11, 65, 85]. Table 3.5 shows timing parameter comparison of VCAT based DRAM and 3D DRAMs. t_{RC} is row cycle time: the delay between successive active commands at the same bank. t_{RC} includes page open, read or write

Table 3.5 Validation of timing parameter of VCAT based and 3D DRAM

Tech. (nm)	Validation Target DRAM Part Name [Ref.]	t_{RCD} (ns)	t_{RP} (ns)	t_{RC} (ns)
		DATE/Target/Error	DATE/Target/Error	DATE/Target/Error
80	50 Mb VCAT DRAM† [11]	8.03/8.00/0.37%	N.A.	30.74/31.00/-0.85%
50	8 Gb 3D DRAM [65]	10.70/15.00/-28.70%	9.01/15.00/-39.97%	N.A.
30	MT41K1G8TRF-125:E Twin [85]	10.84/13.75/-21.13%	13.52/13.75/-1.61%	45.51/47.50/-4.20%

† The VCAT DRAM latency is measured value [11].

Table 3.6 Validation of area calculation of VCAT based and 3D DRAM

Tech. (nm)	Validation Target DRAM Part Name [Ref.]	Bank Area (mm ²)	Die area (mm ²)
		DATE/Target†/Error	DATE/Target†/Error
80	50 Mb VCAT DRAM [11]	1.77/1.88/-5.39%	4.90/4.93/-0.75%
50	8 Gb 3D DRAM [65]	9.14/9.00/1.54%	106.22/98.10/8.28%

operation, and page close latency. The target latencies of 8 Gb 3D DRAM [65] and Micron TwinDie™ DRAM (MT41K1G8TRF) [85] are by the DDR3 specification [3]. For the same reason shown in Table 3.3, most timing comparisons result in negative errors. Only the latency obtained from VCAT design is a measurement value and not a specification [11]. For measured values in the VCAT design, DATE yield the accuracy with errors of 0.37%, -0.85% in t_{RCD} and t_{RC} , respectively.

Table 3.6 shows the comparison of DATE model area results against the derived areas of the VCAT based DRAM and 3D DRAM. The errors for the VCAT based DRAM are -5.39%, -0.75% for bank area and die area respectably. The errors for 8 Gb 3D DRAM are 1.54%, 8.28% for the bank and die area, respectably.

As shown in Figure 3.1, DATE is circuit-level modeling based on technology predictions. The error of technology projection would propagate to the circuit modeling results. To

Table 3.7 Timing parameter change according to process change

Timing Param.		Normal	L_{gate}		Ch.Doping		T_{ox}	
			+10%	-10%	+10%	-10%	+10%	-10%
t_{RCD}	Mean Error (%)	-19.25	-4.76	-25.92	-12.33	-22.84	-4.28	-17.40
	Standard Dev. (%)	8.43	16.02	8.89	9.05	8.61	9.39	8.42
t_{CAS}	Mean Error (%)	-24.87	-14.18	-32.28	-18.61	-28.83	-16.41	-29.01
	Standard Dev. (%)	14.90	15.14	17.82	16.01	16.14	16.38	17.03
t_{RP}	Mean Error (%)	-23.06	-12.21	-31.02	-16.13	-27.12	-12.57	-29.11
	Standard Dev. (%)	23.02	25.83	21.52	24.43	22.23	25.40	12.61
t_{RRD}	Mean Error (%)	-26.68	-11.14	-35.54	-18.47	-31.54	-15.73	-31.4
	Standard Dev. (%)	7.87	8.74	6.38	9.22	6.57	9.69	6.89
t_{WR}	Mean Error (%)	-19.38	-10.52	-18.65	-13.91	-18.65	-11.38	24.40
	Standard Dev. (%)	20.46	22.81	20.05	22.33	20.07	23.24	20.00

examine the impact of propagating error, we varied the L_{gate} , channel doping, and T_{ox} values by $\pm 10\%$, respectively.

Table 3.7 shows timing parameter change according to process change. Increasing gate length, channel doping, and gate oxide thickness reduces turn-on current of the transistor while all the other conditions are unchanged. Compared to the normal case, 10% more gate length, channel doping, and oxide thickness results slower mean percentage error. Except for the t_{RCD} of 10% gate length increase case and t_{RP} of 10% T_{ox} reduce case, the standard deviation in most cases is also within the change of $\pm 4\%$.

Table 3.8 shows energy change according to process change. Increasing gate length by 10% results increased gate capacitance about 10%. As channel doping is changed, the gate capacitance is almost unchanged, and the depletion capacitance is changed less than $\pm 10\%$, according to the calculation by the MASTAR [47]. Reducing oxide thickness 10% results increased gate capacitance about 10%. DATE adjusts the size of the driving buffers with these capacitance changes to achieve optimum speed. In the overall, energy change

Table 3.8 Energy change according to process change

Operating Energy		Normal	L_{gate}		Ch.Doping		T_{ox}	
			+10%	-10%	+10%	-10%	+10%	-10%
Active	Mean Error (%)	-1.39	0.09	-1.28	-0.59	-1.19	-10.08	0.09
	Standard Dev. (%)	7.88	8.18	8.18	8.16	7.92	9.39	8.18
Burst Read	Mean Error (%)	0.20	6.1	1.49	3.81	1.06	-16.41	5.93
	Standard Dev. (%)	7.56	10.32	9.64	10.08	8.07	16.38	10.32
Burst Write	Mean Error (%)	1.74	5.18	2.23	3.77	2.19	-12.57	5.03
	Standard Dev. (%)	7.38	7.46	7.22	7.33	7.23	25.40	7.45
Precharge	Mean Error (%)	-0.29	1.67	0.19	0.92	0.02	-11.38	1.65
	Standard Dev. (%)	8.30	8.22	8.18	8.17	8.28	23.24	8.22
Overall	Mean Error (%)	0.07	3.26	0.66	1.98	0.52	-12.61	3.18
	Standard Dev. (%)	7.78	8.55	8.31	8.44	7.88	18.60	8.54

as follows: a 10% increase in gate length results increased overall mean error but a 10% decrease in gate length also affects buffer sizes to optimum speed, which results slight increase of the mean error by 0.59%; a 10% decrease in gate oxide thickness results in the more overall mean error, and a 10% increase in gate oxide thickness results in decrease of the mean error by 12.54%; and a 10% increase in channel doping leads the increase in energy due to the increased depletion charge but a 10% decrease in channel doping leads slight increasing in energy since the buffer size changes. Except for a 10% increase in T_{ox} case, The standard deviation in all cases is also within the change of 0.8%.

3.4 Comparison with Other Models

Table 3.9 shows circuit level model comparison of CACTI-3DD, Rambus, and DATE. All three models calculate area and energy. Rambus does not support timing. Only CACTI-3DD and DATE support 3D design with TSV model. Of the three models, DATE only supports

8 F^2 , 6 F^2 , and 4 F^2 . Moreover, DATE supports emerging device, i.e., VCAT. DATE supports bank-level custom layout. DATE supports two supply voltage input which makes possible to support LPDDR2 and higher version of LPDDR specification. DATE and CACTI have wire model and support various metal layers for each technology node. DATE, CACTI, and Rambus support 90 nm to 16 nm, 90 nm to 22 nm, and 140 nm to 16 nm transistor roadmaps, respectively. DATE is verified against 22 DRAMs from 80 nm to 30 nm. CACTI-3DD is also verified against three DRAMs from 80nm to 50nm. The Rambus model is not verified against real DRAM. Taken together, DATE is the most verified model that support all known DRAM designs from 90 nm to 16 nm node.

Table 3.9 Circuit level model comparison

Model	DATE	CACTI-3DD [†]	Rambus [†]
Area	✓	✓	✓ [‡]
Timing	✓	✓	×
Energy	✓	✓	✓
3D Support	✓	✓	×
Subarray $8 F^2$	✓	×	×
Layout $6 F^2$	✓	✓	✓
$4 F^2$	✓	×	×
Emerging Device	✓	×	×
Bank-level Custom Layout	✓	×	×
Support two Different Vdd [◇]	✓	×	×
Wire Model	✓	✓	×
Process Node (nm)	90 ~ 16	90 ~ 22	140 ~ 16 [♣]
DRAM used for Validation [♠]	22(2)	3(2)	N.A.
Validated Tech. Node (nm)	80,68,50 44,42,35 30	80,78,50	N.A.

✓ indicates "Supported"

× indicates "Not supported"

[†] Based on the code and data provided by the author.

[‡] Rambus model only calculates the bank size, the user should specify the peripheral area for the energy calculation.

[◇] LPDDR and the higher versions such as LPDDR2 and LPDDR3 have two supply voltages.

[♣] Process projection is not embedded in the source code.

[♠] Numbers in parentheses are published paper rather than commercial DRAM specification.

CHAPTER

4

CASE STUDY: DRAM DESIGN SPACE EXPLORATION

In this case study, we explore DRAM design space regarding energy consumption, area efficiency, and data throughput. We restrict DRAM size to 1 Gb and data I/O size to eight bits with a fixed burst length of eight as in the DDR3 specification [3].

The DRAM has many design components as mentioned in Chapter 3. Variation of the design options such as die count, bank count per die, row and column address, and the subarray size affects the following design components:

- The die size and the bank size, thus affecting the bus length and other wire lengths.
- Logic design.
- Peripheral circuits like sense-amplifier and driving buffer.

To evaluate the effect of the design change on each component, we start from the most basic design case, i.e., 2D single bank. This only incorporates the effects of components related to the bank design and address decoder. We then move to the planar multibank case and then investigate 3D design space using through silicon vias(TSVs). In order to observe the variation according to the difference of cell array layout, we compared the commonly used $6 F^2$ cell layout with the $4 F^2$ cell layout. The section concludes with the design metric comparison of 68 nm, 35 nm, and 16 nm DRAM technologies.

4.1 Planar Design Space Exploration in 35 nm Node

4.1.1 Single Bank Design Space

The single bank is not a practical DRAM design option, but the large bank size such as 1 Gb helps us to understand the tradeoff of the design elements that support and make up the bank. In commodity DRAM, the typical page size is 8 Kb or 16 Kb [3]. We assumed base page size as 8 Kb and decided to expand or reduce the page size to 2, 4, and 8 times to explore the design space. In our case study, the page size will be evaluated from 2^{10} bit to 2^{16} bit. The page size is given as:

$$\text{page size} = 2^{\text{column address bit}} \times \text{data bit.} \quad (4.1)$$

Since we assumed data bit as 8 bit, column address bit can be fixed for each page size

Table 4.1 Address bit and physical dimension of bank matched to each page size at in 1 Gb single bank, 6 F^2 layout

Page Size (Bit)	Column Address (Bit)	Row Address (Bit)	bank size in wordline† (mm)	bank size in bitline† (mm)
1024	7	20	0.124	161.137
2048	8	19	0.247	80.568
4096	9	18	0.495	40.284
8192	10	17	0.989	20.142
16384	11	16	1.978	10.071
32768	12	15	3.956	5.036
65536	13	14	7.912	2.518

† Subarray size in bitline direction is 64 bit and subarray size in wordline direction is 512 bit.

and row address bit was decided as well to form the 1 Gb bank. In detail, Table 4.1 shows the row and column address bits matched to each page size. Table 4.1 also shows physical dimensions of bank in each case. While the page size change, we also change subarray size in each wordline and bitline direction from 2^5 bit to 2^{12} bit. As page size increased, bank size in wordline direction increased and bitline direction decreased. The geometry change caused wordline and bitline length to change along with the shift in driving peripheral circuit size.

4.1.1.1 Area Efficiency

The next design metric is area efficiency. Area Efficiency is the ratio of total cell area over total design area [32], given by the formula:

$$\text{Area Efficiency} = \frac{\text{Total memory size in bit} \times \text{Bit line pitch} \times \text{Word line pitch} \times 100}{\text{Total die area}} [\%] \quad (4.2)$$

Using less peripheral circuit to control memory cells results in higher area efficiency. Table 4.2 shows the best area efficiency on each page size with subarray size. In each design case, the largest bitcell area design option has the best area efficiency since less peripheral logics exists among the cell area. Compared to the $6F^2$ layout design, the reduction in bitcell area of the $4F^2$ layout is mainly responsible for the overall area shrinking while the peripheral area consists of MOSFET. In the $4F^2$ layout, the overall die area is reduced compared to the $6F^2$ layout, and the area efficiency is also reduced.

Table 4.2 Area efficiency of single bank DRAM

Cell Layout	Page Size (Bit)	Subarray Size		Area Efficiency (%)	Area (mm ²)
		Row (Bit)	Column (Bit)		
$6F^2$	1024	4096	512	48.4	16.3
	2048	4096	1024	60.9	13.0
	4069	4096	2048	70.0	11.3
	8192	4096	4096	75.6	10.4
	16384	4096	4096	75.6	10.4
	32768	4096	4096	75.6	10.4
	65536	4096	4096	75.6	10.4
$4F^2$	1024	4096	512	39.1	13.4
	2048	4096	1024	51.7	10.2
	4069	4096	2048	61.5	8.6
	8192	4096	4096	68.0	7.7
	16384	4096	4096	68.0	7.7
	32768	4096	4096	68.0	7.7
	65536	4096	4096	68.0	7.7

Table 4.3 Read energy efficiency of single bank DRAM, 6 F^2 layout

Page Size (Bit)	Subarray Row (Bit)	Column (Bit)	Energy Efficiency (Bit/nJ)
1024	512	512	13.8
2048	256	512	23.4
4069	128	512	35.3
8192	64	512	44.5
16384	64	512	46.9
32768	32	512	40.7
65536	32	512	28.5

4.1.1.2 Energy Efficiency

Table 4.3 shows the best read operation energy efficiency of each page size with corresponding subarray configuration in 6 F^2 cell layout. We assumed read operation was composed of a successive sequence of active (page open), burst read, and precharge (page close) commands. When the page size is 16384 bit, and the subarray size is 64 bit and 512 bit in the direction of the wordline (WL) and the bitline (BL) direction respectively, 1 Gb single bank DRAM has optimal energy efficiency. During the energy efficiency evaluation, the variables are page size, and subarray size in WL and BL direction. We observed the energy changes of each design elements by changing each variable while other conditions were fixed.

Table 4.4 shows read energy change in each design component as page size changes while the subarray configurations are equal. As shown in Table 4.1, bank size in BL direction decreased as page size increased. When the bank size was reduced in the BL direction, the bus energy was reduced because the length of the row address bus was reduced. As

the page size increased, the row predecoder energy decreased because the row address bit decreased, but the main-wordline (MWL) decoder energy and sense-amplifier (SA) activation energy increased as the MWL length increased. The sub-wordline (SWL) driver energy also increased due to the increased the page size. The SWL select energy included the WL control driver, dx signal driver, WL reset driver, and 4 bit decoded address path from the control logic in the center stripe to the corner of the bank as shown in Figure 3.17 and Figure 3.18. The SWL select energy was reduced mainly due to the BL direction length change. In the burst read operation, column predecoder energy increased due to increasing column address bit as page size increased. DATE assumes column select signals are transmitted by the bus. Since the bus has a repeater at the regular interval, the column select decoder energy is unchanged. The bit select energy and SA read-out energy are also do not change due to constant subarray size. Data read-out energy using data bus decreased due to bank size decreasing in bitline direction. In the precharge operation, SA precharge energy increased as the number of the closing page size increased. The energy used to deliver by the precharge command to the corners of the bank was proportional to the sum of the WL and BL lengths as shown in Table 4.1. When the page size increased, the bank size in bitline direction decreased. Thus, the precharge command path had similar length even though the change of the page size. By including loaded SA enable buffers, which is proportional to the page size, the precharge command signal energy was the smallest when the page size is 4096 bit. By summing the tradeoffs of the design components, the lowest energy was obtained when the page size was 16384 bit as highlighted in Table 4.4.

Figure 4.1 shows energy efficiency of the 6 F^2 cell array, single bank DRAM when the page size is fixed at 16384 bit as changing row and column size of the subarray. In Table 4.3, the optimal read energy efficiency is shown as page size change when the subarray size is 64 bit in the row direction and 512 bit in the column direction. Table 4.5 shows the physical

Table 4.4 Read operation energy change in each component as page size change

Page Size (Bit)		1024	2048	4096	8192	16384	32768	65536
Active Energy (pJ)	Bus, Cmd-In	1314.176	625.095	297.448	142.736	71.155	40.417	30.468
	Row Addr. Predec.	1.353	1.021	1.061	0.794	0.476	0.325	0.201
	MWL Dec.	1.043	1.898	3.610	7.033	13.878	27.570	54.953
	SWL Drv.	2.328	4.657	9.314	18.627	37.254	74.509	149.017
	SWL sel.	6.058	5.755	5.452	5.149	4.846	4.544	4.241
	SA act.	12.662	25.324	50.647	101.295	202.589	405.178	810.357
Read Energy (pJ)	Bus, Cmd-In	76.896	44.324	25.723	15.967	12.368	14.408	23.670
	Col. Addr. Predec.	0.029	0.019	0.032	0.044	0.098	0.125	0.167
	Col-Sel Dec.	1.341	1.341	1.341	1.341	1.341	1.341	1.341
	Bit Sel. Sig.	0.214	0.214	0.214	0.214	0.214	0.214	0.214
	SA read-out Drv.	2.788	2.788	2.788	2.788	2.788	2.788	2.788
	Bus, Data-Out and Tx	4423.909	2324.127	1276.136	755.902	503.163	390.991	357.825
Precharge Energy (pJ)	SA precharge	14.544	29.089	58.177	116.354	232.708	465.416	930.833
	Bus, Cmd	293.828	273.624	266.045	267.301	278.014	303.525	356.416
Energy Efficiency (Bit/nJ)		10.404	19.163	32.019	44.535	46.929	36.843	23.410

Figure 4.1 Energy efficiency as subarray size change with $6 F^2$ layout, 16384-bit page size.

Table 4.5 Bank size of single bank DRAM as subarray size change, 6 F^2 layout

Subarray Size		bank size	
Row (Bit)	Column (Bit)	Wordline (mm)	Bitline (mm)
32	512	1.98	13.26
64	512	1.98	10.07
128	512	1.98	8.48
256	512	1.98	7.70
512	512	1.98	7.28
1024	512	1.98	7.08
2048	512	1.98	6.98
4096	512	1.98	6.93
64	32	6.30	10.07
64	64	3.80	10.07
64	128	3.32	10.07
64	256	2.47	10.07
64	512	1.98	10.07
64	1024	1.57	10.07
64	2048	1.37	10.07
64	4096	1.27	10.07

dimension of the bank changes as the subarray bit size changes in WL and BL direction while the page size is fixed at 16384 bit. While subarray column size and page size are fixed, it was observed that the larger the subarray size in the row direction, the smaller the number of the sense amplifier area are inserted between the bitcell areas. Thus, the bank size decreased in the BL direction. As the subarray size increased in the column direction, the number of SWL driver areas, which inserted between the bitcell arrays are decreased. The bank size decreases in the WL direction.

Table 4.6 and Table 4.7 shows the detailed component energy of the read operation while subarray size changed. Table 4.6 shows how energy efficiency changes when the page size

and subarray column size are fixed while subarray row size is changed. Since the address does not change, the row and column predecoder energy do not change. As the subarray size increased in the BL direction, there was a change in the energy of the peripheral circuits driving subarray in the BL direction. It was observed that sense amplifier activation and precharge energies increased. Since the MWL drivers were composed of pseudo-NOR gates as described in Section 3.1.4 and we assumed to be evaluated in subarray units, the larger the size of the subarray row, more energy was consumed. As the bank decreased in the BL direction, the command path energy at the precharge and the data-out bus energy at the read command decreased. The lowest energy was observed when the subarray row size was 64 bit (see highlighted column in Table 4.6).

Table 4.6 Read operation energy change as subarray row size change

Subarray Row Size (Bit)		32	64	128	256	512	1024	2048	4096
Active Energy (pJ)	Bus, Cmd-In	91.977	71.115	60.736	55.524	52.917	51.613	50.960	50.634
	Row Addr. Predec.	0.476	0.476	0.476	0.476	0.476	0.476	0.476	0.476
	MWL Dec.	7.345	13.878	26.945	53.077	105.343	209.874	418.935	837.059
	SWL Drv.	37.254	37.254	37.254	37.254	37.254	37.254	37.254	37.254
	SWL sel.	4.846	4.846	4.846	4.846	4.846	4.846	4.846	4.846
	SA act.	146.174	202.589	315.419	541.079	992.398	1895.036	3700.313	7310.867
Read Energy (pJ)	Bus, Cmd-In	14.760	12.368	11.170	10.596	10.269	10.118	10.043	10.005
	Col. Addr. Predec.	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098
	Col-Sel Dec.	1.341	1.341	1.341	1.341	1.341	1.341	1.341	1.341
	Bit Sel. Sig.	0.162	0.214	0.312	0.507	1.310	2.275	3.657	6.285
	SA read-out Drv.	2.788	2.788	2.788	2.788	2.788	2.788	2.788	2.788
	Bus, Data-Out and Tx	586.452	503.163	461.489	440.640	430.210	424.995	422.386	421.082
Precharge Energy (pJ)	SA precharge	148.086	232.708	401.953	740.442	1417.421	2771.379	5479.294	10895.120
	Bus, Cmd	357.531	278.014	238.255	218.375	208.435	203.465	200.980	199.737
Energy Efficiency (Bit/nJ)		45.658	46.929	40.846	30.298	19.531	11.358	6.175	3.227

Table 4.7 shows the energy change of the read operation while subarray column size is changed. As the subarray column size increased, the length of the bank in the WL direc-

Table 4.7 Read operation energy change as subarray column size change

Subarray Column Size (Bit)		32	64	128	256	512	1024	2048	4096
Active Energy (pJ)	Bus, Cmd-In	82.080	75.766	74.537	72.402	71.155	70.125	69.610	69.353
	Row Addr. Predec.	0.476	0.476	0.476	0.476	0.476	0.476	0.476	0.476
	MWL Dec.	54.652	31.656	25.108	17.930	13.878	10.810	9.277	8.510
	SWL Drv.	36.145	32.657	46.504	41.266	37.254	31.321	28.354	26.870
	SWL sel.	4.846	4.846	4.846	4.846	4.846	4.846	4.846	4.846
	SA act.	219.101	214.827	208.130	204.436	202.589	201.666	201.204	200.973
Read Energy (pJ)	Bus, Cmd-In	22.578	16.678	15.529	13.534	12.368	11.406	10.925	10.684
	Col. Addr. Predec.	0.098	0.098	0.098	0.098	0.098	0.098	0.098	0.098
	Col-Sel Dec.	1.341	1.341	1.341	1.341	1.341	1.341	1.341	1.341
	Bit Sel. Sig.	0.214	0.214	0.214	0.214	0.214	0.214	0.214	0.214
	SA read-out Drv.	0.174	0.348	0.696	1.393	2.788	5.585	11.210	22.576
	Bus, Data-Out and Tx	441.232	418.109	419.944	432.470	503.163	786.197	1901.581	6307.991
Precharge Energy (pJ)	SA precharge	232.708	232.708	232.708	232.708	232.708	232.708	232.708	232.708
	Bus, Cmd	4052.150	2043.644	1035.486	530.546	278.014	151.611	88.409	56.808
Energy Efficiency (Bit/nJ)		12.370	20.733	30.866	41.071	46.929	42.375	24.984	9.16

tion decreased as shown in Table 4.5, and the length of the subarray in the WL direction increased. As the subarray bit size increased in the WL direction, row address bus energy in active operation was reduced, since the bank size in WL direction decreased. The MWL driver energy decreased as the bank size decreased in WL direction. The SWL driver energy decreased because the number of SWL driver decreased due to the decrease in subarray numbers. Sense amplifier active energy reduced slightly since the command path was reduced due to bank size reduction in the WL direction. The SA readout driver energy and the data-out bus energy was increased because data was reading in subarray units. The lowest energy was observed when the subarray column size was 512 bit (highlighted in gray, in Table 4.7).

Table 4.8 shows the best energy efficiency of the read operation of each page size with the corresponding subarray configuration in 4 F^2 cell layout. The best energy efficiency was obtained when the page size was 8192 bit, and the subarray size was 128 bit and 512 bit, respectively in the WL direction and the BL direction, which is highlighted in gray in Table 4.8.

Table 4.8 Read energy efficiency of single bank DRAM, 4 F^2 layout

Page Size (Bit)	Subarray		Energy Efficiency (Bit/nJ)
	Row (Bit)	Column (Bit)	
1024	1024	512	19.7
2048	512	512	32.4
4069	256	512	46.7
8192	128	512	56.3
16384	64	512	55.8
32768	32	512	45.5
65536	32	512	31.4

Figure 4.2 Energy efficiency as subarray size change with 4 F^2 bitcell layout, 16384-bit page size.

Since the bitcell area was reduced, the $4 F^2$ layout exhibited better energy efficiency than the $6 F^2$ layout case. As shown in Section 3.2, the VCAT based DRAM design was assumed to have the same peripheral circuitry as the $6 F^2$ layout DRAM design. Therefore, except for the energy change due to the bitcell area reduction, the energy change of each component has a trend similar to the $6 F^2$ cell array based design case. Figure 4.2 shows energy efficiency of the $4 F^2$ cell array, a single bank DRAM when the page size is fixed at 8192 bit as the row, and column bit size of the subarray is changed. As shown in Table 4.8, the optimal energy efficiency was observed when the subarray size was 128 bit in the row direction and 512 bit in the column direction.

4.1.1.3 Throughput

Table 4.9 shows the best read operation throughput of each page size with the corresponding subarray bit size in $6 F^2$ cell layout. The single bank DRAM showed optimal throughput when the page size was 16384 bit, and the subarray size was 128 bit and 512 bit in the direction of the wordline (WL) and the bitline (BL) direction respectively. In a single bank design, the design variables were page size, subarray WL bit size, and subarray BL bit size. We kept tracking the latency change of the critical path elements among the parallel operation by changing each variable while other conditions were fixed. For example, at the moment of the row address decoding in bank level, the main-wordline (MWL) driver and the sense amplifier (SA) activation signal were driven at the same time. Since the MWL driver loaded more capacitance than the SA activation driver, the MWL driver delay was larger than the SA activation delay. We accounted for the MWL delay when determining the latency of row address path of the bank.

Table 4.10 shows the throughput and the speed change of each critical path component as page size changed for the $6 F^2$ layout, 1 Gb single bank DRAM. As shown in Table 4.1,

Table 4.9 Throughput of read operation, single bank DRAM, 6 F^2 layout

Page Size (Bit)	Subarray Row (Bit)	Column (Bit)	Throughput (GB/sec)
1024	2048	128	0.011
2048	1024	128	0.037
4069	512	128	0.090
8192	256	256	0.148
16384	128	512	0.174
32768	64	1024	0.159
65536	64	1024	0.112

Table 4.10 Speed of each component and read throughput as page size change

Page Size (Bit)		1024	2048	4096	8192	16384	32768	65536
Row Addr.Path (ns)	Access Bus	4.710	2.359	1.186	0.604	0.323	0.201	0.171
	Predec.	1.049	1.052	1.001	0.953	0.868	0.804	0.510
	MWL Dec.	0.485	0.678	1.045	1.893	3.884	9.021	23.592
	SWL Drv.	2.302	2.429	2.630	2.947	3.444	4.254	5.641
Column Addr.Path (ns)	Access Bus	1.576	0.796	0.413	0.235	0.173	0.195	0.305
	Predec.	0.258	0.287	0.244	0.461	0.503	0.563	0.604
	Col. Sel. Dec.	2.731	2.743	2.744	2.745	2.760	2.780	2.781
Data Path (ns)	Bitline	3.701	3.232	3.065	3.202	3.694	4.783	6.923
	Sense Amp.	0.005	0.005	0.005	0.005	0.005	0.005	0.005
	Subarray Out Drv.	0.097	0.097	0.097	0.097	0.097	0.097	0.097
	Data Bus	185.158	53.085	17.111	6.703	3.530	2.765	3.379
	Data I/O Buffer	1.250	1.250	1.250	1.250	1.250	1.250	1.250
Read Throughput (GB/sec)		0.008	0.030	0.080	0.144	0.174	0.148	0.093

the bank size in BL direction decreased as page size increased. When the bank size was reduced in the BL direction, the row address bus latency was reduced. As the page size increased, the row predecoder latency gradually decreases because the row address bit

decreased. Comparing the case of 20-bit and 19-bit address, 19-bit predecoding is slower. With the similar capacitance load, 3 input NAND gate was slower than the 2 input NAND gate; the 19-bit predecoder had 3 input NAND gates to handle odd addresses, while 20-bit predecoder had only 2 input NAND gates. Thus, as the address was reduced, the latency of predecoder was occasionally increased, but was reduced overall. The MWL decoder latency increased as the WL length increased. The sub-wordline (SWL) driver latency also increased due to increased MWL rising time. For the column address path, column address predecoder latency gradually increased as the column address increased. The column select signal decoder also increased due to increasing input latency as column predecoding time was increased.

In the data path, as shown in Equation 3.32, the bitline delay was proportional to the rise time of the wordline. The wordline rise time was affected by the combining delay of the row address path, which is passing through the predecoder, the row address bus, the MWL decoder, and SWL driver. The MWL decoder and SWL driver latency rose due to increased length of the wordline while the predecoder latency and row address bus latency decreased. Reflecting the latency changes in the row address path, the bitline delay had a minimum value when the page size was 4096 bit. The data path bus latency decreased as the bank size in BL direction decreased. As shown in Table 4.10, the optimal throughput was observed when the page size was 16384 bit.

Figure 4.3 shows data throughput of the 6 F^2 cell array, single bank DRAM when the page size was fixed at 16384 bit as the row and column size of the subarray was changed. As shown in Table 4.9, the optimal throughput was observed when the subarray size was 128 bit in the row direction and 512 bit in the column direction. Table 4.11 and Table 4.12 shows the detailed component speed and throughput of the read operation when subarray size is changed.

Figure 4.3 Read throughput as subarray size change at $6 F^2$ layout, 16384-bit page size.

Table 4.11 shows read throughput change as subarray row size change while other conditions were fixed. As the subarray row bit size increased, the length of the bank decreased and the length of the subarray itself increased. The address access bus and data bus speed were reduced due to the bank size reduction. The bitline latency and sense amplifier latency also increased due to the subarray size increase in bitline direction. As shown in Table 4.11, when the subarray bit size was 128 bit, single bank DRAM exhibited optimal read throughput.

Table 4.12 shows the read throughput and speed of each component as subarray column size changes. As the subarray column size increased, the length of the bank in the WL direction decreases, and the MWL decoder latency decreased when SWL driver latency was increased. The bitline latency, which is mainly affected by MWL and SWL raise time, exhibited the optimum delay when subarray column size was 512 bit. Since bank size was

Table 4.11 Speed of each component and read throughput as subarray row size change

Subarray Row Size (Bit)		32	64	128	256	512	1024	2048	4096
Row Addr.Path (ns)	Access Bus	0.490	0.379	0.323	0.296	0.282	0.275	0.271	0.270
	Predec.	0.868	0.868	0.868	0.868	0.868	0.868	0.868	0.868
	MWL Dec.	3.884	3.884	3.884	3.884	3.884	3.884	3.884	3.884
	SWL Drv.	3.444	3.444	3.444	3.444	3.444	3.444	3.444	3.444
Column Addr.Path (ns)	Access Bus	0.229	0.192	0.173	0.164	0.159	0.157	0.156	0.155
	Predec.	0.503	0.503	0.503	0.503	0.503	0.503	0.503	0.503
	Select Decoder	2.760	2.760	2.760	2.760	2.760	2.760	2.760	2.760
Data Path (ns)	Bitline	2.938	3.244	3.694	4.254	4.853	5.458	6.126	7.021
	Sense Amp.	0.004	0.005	0.005	0.006	0.008	0.010	0.013	0.016
	Subarray Out Drv.	0.097	0.097	0.097	0.097	0.097	0.097	0.097	0.097
	Data Bus	5.446	4.124	3.530	3.249	3.113	3.046	3.012	2.996
	Data I/O Buffer	1.250	1.250	1.250	1.250	1.250	1.250	1.250	1.250
Read Throughput (GB/sec)		0.154	0.169	0.174	0.170	0.161	0.143	0.112	0.070

reduced in WL direction, the latency of both row and column address access was reduced. The data bus latency included local bus in the subarray and global bus in the bank. As the subarray size increased in the WL direction and the bank size decreased, the local data bus and the global data bus speed increased and decreased, respectively. The data bus showed balanced minimum latency when the subarray column size was 512 bit. As shown in Table 4.12(see highlighted column), single bank DRAMs displayed optimal throughput when the subarray column was 512 bit.

Table 4.13 shows the data throughput of the read operation of each page size with corresponding subarray configuration in $4 F^2$ cell layout. The best throughput was obtained when the page size was 16384 bit, and the subarray size was 128 bit and 1024 bit in the WL direction and the BL direction, respectively, which is highlighted in gray in Table 4.13. Since the bitcell area was reduced, the $4 F^2$ layout exhibited better data throughput than the $6 F^2$

Table 4.12 Speed of each component and throughput as subarray column size change

Subarray Column Size (Bit)		32	64	128	256	512	1024	2048	4096
Row Addr.Path (ns)	Access Bus	0.381	0.348	0.341	0.330	0.323	0.318	0.315	0.314
	Predec.	0.868	0.868	0.868	0.868	0.868	0.868	0.868	0.868
	MWL Dec.	21.041	10.136	7.754	5.191	3.884	2.956	2.518	2.305
	SWL Drv.	3.719	3.272	2.884	2.949	3.444	5.168	12.235	40.018
Column Addr.Path (ns)	Access Bus	0.331	0.240	0.222	0.191	0.173	0.158	0.151	0.147
	Predec.	0.503	0.503	0.503	0.503	0.503	0.503	0.503	0.503
	Select Decoder	2.760	2.760	2.760	2.760	2.760	2.760	2.760	2.760
Data Path (ns)	Bitline	6.455	4.840	4.357	3.868	3.694	3.862	5.053	8.349
	Sense Amp.	0.005	0.005	0.005	0.005	0.005	0.005	0.005	0.005
	Subarray Out Drv.	0.096	0.096	0.096	0.096	0.097	0.097	0.098	0.010
	Data Bus	5.015	4.043	3.896	3.624	3.530	3.557	3.858	4.701
	Data I/O Buffer	1.250	1.250	1.250	1.250	1.250	1.250	1.250	1.250
Read Throughput (GB/sec)		0.094	0.133	0.147	0.165	0.174	0.172	0.146	0.092

layout case. The VCAT based DRAM design had the same peripheral circuitry as the $6 F^2$ layout DRAM design mentioned in Section 3.2. The throughput change of each component concerning the change of page size and subarray size displayed a similar trend as the $6 F^2$ design case. Figure 4.4 shows data throughput of the VCAT based cell array, single bank DRAM when the page size is fixed at 16384 bit as row and column bit size of the subarray is changed. The optimal data throughput was observed when the subarray size was 128 bit in the row direction and 1024 bit in the column direction.

4.1.2 Multi-bank Design Space in 35 nm Node

4.1.2.1 Area Efficiency

Table 4.14 shows the page size and subarray configuration with optimal area efficiency for $6 F^2$ and $4 F^2$ cell arrays when the number of banks was increased. As shown in Equa-

Table 4.13 Throughput of read operation, single bank DRAM, 4 F^2 layout

Page Size (Bit)	Subarray		Throughput (GB/sec)
	Row (Bit)	Column (Bit)	
1024	2048	512	0.024
2048	1024	512	0.066
4069	512	512	0.133
8192	256	1024	0.186
16384	128	1024	0.199
32768	64	2048	0.173
65536	64	2048	0.120

Figure 4.4 Data Throughput as subarray size change at 4 F^2 layout with 16384-bit page size.

tion 4.2, the smaller the peripheral circuit area and the redundancy area, the higher the area efficiency. The highest area efficiency value at 6 F^2 cell array was 75.6%, and the highest area efficiency value at 4 F^2 cell array was 68.0%. Since 4 F^2 cell array has a smaller area than the 6 F^2 array, VCAT based design has a smaller die area compared to RCAT based design.

Table 4.14 Area efficiency of planar multibank DRAM

Cell Layout	Num. Bank	Page Size (Bit)	Subarray Size Row (Bit) Column (Bit)		Area Efficiency (%)	Area (mm ²)
$6F^2$	1	65536	4096	4096	75.6	10.4
	2	65536	4096	4096	75.6	10.4
	4	32768	4096	4096	74.8	10.5
	8	16384	4096	4096	73.4	10.8
	16	8192	4096	4096	70.0	11.3
	32	16384	1024	4096	63.6	12.4
	64	8192	1024	4096	56.8	13.9
	128	4096	1024	2048	45.9	17.2
	256	4096	512	2048	35.8	22.0
	512	2048	512	1024	26.1	30.2
$4F^2$	1	65536	4096	4096	68.0	7.7
	2	65536	4096	4096	68.0	7.7
	4	32768	4096	4096	67.4	7.8
	8	8192	4096	4096	65.3	8.0
	16	8192	4096	4096	62.1	8.5
	32	8192	2048	4096	55.3	9.5
	64	8192	1024	4096	45.7	11.5
	128	4096	1024	2048	36.4	14.4
	256	2048	1024	1024	26.7	19.7
	512	2048	512	1024	19.0	27.7

4.1.2.2 Energy Efficiency

Table 4.15 shows the best read energy efficiency and page size when the number of bank was increased in planar 1 Gb DRAM. The multibank design displayed the best energy efficiency with 512 banks. As the number of the bank increased, single bank size decreased. Thus, the number of operating logic and buffers for a single bank decreased. The wire energy inside the bank also decreased due to the smaller bank size as the bus energy between bank remained the same. As the number of banks increased, the proportion of wire energy increased.

Table 4.15 Read Energy and Efficiency of 1 Gb 2D Multibank DRAM in 35 nm node, 6 F^2 layout

Num. Bank	Page Size (Bit)	Energy Efficiency (Bit/nJ)	Read Energy (nJ)	Wire Energy (nJ)	Wire Percent (%)
1	16384	46.9	1.364	0.963	67.4
2	8192	65.3	0.980	0.676	76.9
4	8192	79.5	0.805	0.514	71.7
8	4096	105.1	0.609	0.442	80.8
16	4096	128.4	0.499	0.387	76.6
32	2048	152.3	0.420	0.322	85.4
64	1024	165.5	0.387	0.313	84.3
128	1024	177.0	0.362	0.288	90.9
256	1024	182.2	0.351	0.277	90.7
512	1024	183.3	0.349	0.277	90.8

Figure 4.5 shows the detail of the wire energy change as the number of bank increased. During the burst read operation, DATE assumes operation scenario as serial commands as "Active-Read-Precharge," and there were two column address path attempts at "Read"

Figure 4.5 Energy sum of wire component in multi-bank 2D DRAM, 6 F^2 Layout.

and "Precharge" command. The column address path has larger energy than row address path. In Figure 4.5, data path energy is dominant since the wire energy is proportional to the multiplication of Wire length and number of switching signals. Although the column address path is used twice in the operation scenario, the data size read in the burst operation is 64 bit. Thus, the data path energy was the largest.

Table 4.16 shows the best read energy efficiency with VCAT based cell array design. The VCAT based multibank design exhibited the best energy efficiency with 256 banks. As in the RCAT based design, bank size decrease as the number of banks increased and the number of operating logic and buffers decreased as well. The wire energy inside the bank also decreased due to the smaller bank size as the bus energy between bank remained the same. As the number of banks increased, the proportion of wire energy also increased. As shown in Figure 4.6, the wire had minimum energy with a 256 bank configuration. After the point, the wire energy showed diminished return as the number of banks increased because as the bus length connecting the bank to the I/O increases, the data path energy is

Table 4.16 Read Energy and Efficiency of 1 Gb 2D Multibank DRAM in 35 nm node, 4 F^2 layout

Num. Bank	Page Size (Bit)	Energy Efficiency (Bit/nJ)	Read Energy (nJ)	Wire Energy (nJ)	Percent (%)
1	8192	56.3	1.136	0.862	75.9
2	8192	76.7	0.834	0.655	78.5
4	8192	91.4	0.700	0.522	74.5
8	4096	118.4	0.540	0.447	82.7
16	2048	143.2	0.447	0.373	83.5
32	1024	160.6	0.399	0.358	89.9
64	1024	176.7	0.362	0.323	89.1
128	1024	184.1	0.348	0.309	88.9
256	1024	186.5	0.343	0.305	88.9
512	1024	181.9	0.352	0.314	89.3

increased.

4.1.2.3 Throughput

Table 4.17 shows the best throughput as bank numbers increase in planar DRAM. Comparing these data with Table 4.15, the optimum page sizes were slightly different to achieve the best speed results. In planar DRAM design, DRAM exhibited optimal throughput when the number of banks was 512, that is, the size was 2 Mb per one bank. When the bank size is decreased, decoder and wire delay in the bank also decrease because the number of address and size of the bank decreased.

Figure 4.7 shows the sum of each design component delay while the number of bank increases. The decoder indicates the sum of row and column address decoder delay. Wire indicates the sum of the row, column address access bus delay and data out path delay. WL, and BL indicate the sum of wordline and bitline delay. I/O and miscellaneous indicates I/O

Figure 4.6 Energy sum of wire component in multibank 2D DRAM, 4 F^2 Layout.

transceiver and control signal delay. In Figure 4.7, I/O circuits and miscellaneous parts were not changed substantially since we assumed DRAM operates at 800 MHz (I/O transceiver operates at 1600 MHz). For the energy, the capacitance is the only factor that impacts energy consumption. However, for the throughput, resistance and capacitance must be considered together. Thus, comparing Figure 4.5 with Figure 4.7, WL, and BL occupy a larger portion of entire delay since WL, and BL have larger resistance compared to the bus. As shown in Figure 4.7, WL, and BL delay accounts for about 40% to 50% of the sum of each component delay. WL, and BL consist of WSi_2 which has higher resistivity than metal. WL, and BL have minimum wire width. Thus, their resistance is higher than the metal bus.

In the VCAT based design, the optimal throughput is 0.384 GB/sec with 512 banks. The WL and BL also occupy a large portion of entire delay. In the VCAT-based design case, WL, and BL delay accounts for about 37% to 47% of the sum of each component delay.

Table 4.17 Throughput of 1 Gb 2D Multibank DRAM, 6 F^2 layout

Num. Bank	Page Size (Bit)	Read Throughput (GByte/sec)
1	16384	0.174
2	8192	0.197
4	8192	0.225
8	4096	0.247
16	4096	0.270
32	4096	0.288
64	4096	0.315
128	2048	0.333
256	2048	0.351
512	1024	0.366

Figure 4.7 Sum of each component delay in multibank 2D DRAM, 6 F^2 Layout.

4.2 3D Design Space Exploration in 35 nm Node

A DRAM rank is a group of DRAM devices which respond and operate at the same time by the single command. Since it is a logical concept, a rank could be a single die, a single chip or multiple chips on a printed circuit board (PCB) package.

To design three dimensional DRAM, there are two different approaches to die stacking in rank-level. Figure 4.8 shows examples of forming ranks in three-dimensional design. Figure 4.8a shows a traditional planar DRAM, which, in this case, the rank is a single die. Figure 4.8b shows coarse-grained rank-level 3D stacking of planar ranks and Figure 4.9 shows chip micrograph of fabricated DDR3 DRAM die for the coarse-grained rank-level die stacking. In Figure 4.9, the total number of TSVs at the center of the die is about 400 and the center TSV area is about 4.8% of the die [65]. In such an architecture, multiple dies share a single I/O interface and TSVs. When a die is selected, it treats the TSVs as a dedicated bus.

Figure 4.8c shows fine-grained rank-level 3D stacking of planar ranks. Here, a planar rank is divided into multiple "dielets" that are stacked across multiple layers. Each rank has its own dedicated TSV bus and I/O interface, which results in no contention bus access across different ranks. This architecture has an expensive design cost as it requires the number of TSVs be equal to the number of ranks implemented, which is expensive in terms of area and power.

Bank-level 3D stacking is shown in Figure 4.10. A bank split in row or column direction and shares the same addresses between two or more dies. The split bank requires that the number of TSVs equal the number of rows and columns in the bank. Compared to the TSV area in Figure 4.9, this makes single bank splitting more expensive than the fine-grained rank-level stacking in terms of area and power. Therefore, we do not consider 3D bank level stacking case.

In a fine-grained case, as shown in Figure 4.8c, DRAM has an independent logic die at the bottom of DRAM. The logic die could be designed and fabricated by using another process to achieve best operation performance. This is beyond DATE model's evaluation range. Extending DATE to model this stacking approach would require additional research to predict the performance of the logic die, such as in [2]. Thus, in this section, we only consider the coarse-grained rank-level die stacking design. The 35 nm node will be discussed first followed by a comparison with the results of 16 nm and 68 nm nodes.

Figure 4.8 Rank level die-stacking.

4.2.1 Area Efficiency

Table 4.18 show area efficiency of 3D stacked DRAM with $6 F^2$ and $4 F^2$ cell layout. As discussed in Section 4.1.1.1, as space for the peripheral circuit and additional functions is increased, area efficiency is decreased. As the number of layers increases, the memory size to be allocated per die to maintain 1 Gb becomes smaller (i.e., the cell array size per die is reduced). The number of data and control signals is similar even if the number of

¹The figure is used under author's permission.

Figure 4.9 Chip micrograph of the fabricated DRAM die and cross-sectional view of TSVs [65].¹

Figure 4.10 Bank level die-stacking.

Table 4.18 Area efficiency of 1 Gb 3D multibank DRAM in 35 nm node

Cell Layout	Num.of die	Num. Bank/die	Page Size (Bit)	Subarray Row (Bit)	Subarray Column (Bit)	Area Efficiency (%)	Single Die Area (mm ²)	TSV Area (%)
$6F^2$	1	1	65536	4096	4096	75.6	10.44	N.A.
	2	1	32768	4096	4096	47.3	8.35	5.3
	4	1	16384	4096	4096	39.9	4.94	8.8
	8	1	16384	4096	4096	31.4	3.15	13.8
	16	1	8192	4096	4096	22.5	2.19	19.8
	32	1	8192	2048	4096	14.5	1.70	25.2
$4F^2$	1	1	65536	4096	4096	68.0	7.74	N.A.
	2	1	16384	4096	4096	39.8	6.61	6.7
	4	1	16384	4096	4096	32.6	4.04	10.8
	8	1	8192	4096	4096	24.8	2.66	16.4
	16	1	8192	4096	4096	16.8	1.96	22.1
	32	1	8192	2048	4096	10.6	1.55	27.7

layers increases. Thus the number of TSVs for data and control signals is similar even as the number of die increases. Since the TSV area for data and control signal is almost the same, and the area of the cell area is reduced, the area efficiency deteriorates as the number of dies increases in both $6F^2$ and $4F^2$ layout case. As discussed in Section 4.1.2.1, $6F^2$ cell layout 3D DRAM has higher area efficiency than the $4F^2$ cell layout.

4.2.2 Energy Efficiency

Table 4.19 shows the best energy efficiency on each 3D stacked DRAM configuration as the number of layers is increased. In planar DRAM with VCAT, the design with 4 Mb size banks exhibited the best energy efficiency. In the other cases, the best energy efficiency design had 2 Mb bank size which was the smallest bank size examined during the evaluation. As discussed in Section 4.1.2.2, the smaller the bank size, the smaller the logic size, which

makes smaller logic energy. Thus, the wire energy dominates. In Table 4.19, the planar connection energy between each component is represented by wire energy, and the 3D connection energy is represented by TSV energy. The optimal number of layers is eight in each $4 F^2$ and $6 F^2$ cell layout case. The energy efficiency showed diminishing returns when the number of the dies was greater than eight. In all cases, wire energy and TSV energy accounted for approximately 88 to 95 percent of the entire read energy consumption, which indicates that the logic energy was optimized.

Figure 4.11 shows the energy sum of the wire components and TSV in $6 F^2$ cell layout case. As the number of stacked die increased, the TSV energy started to dominate overall wire energy since the TSV energy increased proportional to the number of layers. When the number of dies was 32, TSV energy accounted for more than 52% of the total energy sum. Except for the TSV energy, data path energy dominated in every case, as we depicted earlier in Section 4.1.2.2.

Figure 4.11 Energy sum of wire components and TSV in 35 nm node.

Table 4.19 Energy efficiency of 1 Gb 3D multibank DRAM in 35 nm node

Cell Layout	Num.of die	Bank Size (Mb)	Energy Efficiency (Bit/nJ)	Read Energy (nJ)	Wire Energy (%)	TSV Energy (%)
$6F^2$	1	2	183.3	0.349	90.8	N.A.
	2	2	209.0	0.306	86.8	2.7
	4	2	227.1	0.282	80.6	8.0
	8	2	227.7	0.281	70.9	17.4
	16	2	205.1	0.312	59.5	32.8
	32	2	157.1	0.407	41.8	52.3
$4F^2$	1	4	186.5	0.343	88.9	N.A.
	2	2	206.6	0.310	89.0	2.7
	4	2	222.9	0.287	83.0	7.8
	8	2	226.1	0.283	73.4	17.3
	16	2	202.1	0.316	59.4	32.3
	32	2	155.6	0.411	41.8	51.8

4.2.3 Throughput

Table 4.20 shows the best throughput on each 3D stacked DRAM as the number of stacked die is increased. In 3D DRAM design, both $6F^2$ and $4F^2$ layout cases displayed optimal throughput when eight dies were used. After the eight-layer, the DRAM throughput exhibited diminishing returns. As discussed in Section 4.1.2.3, the smaller the bank size, faster throughput was observed. In all cases, 3D DRAM exhibited the fastest throughput at the smallest bank size of 2 Mb.

Figure 4.12 shows the sum of the delay of each design component in $6F^2$ cell array as the number of DRAM layer increased. I/O and miscellaneous indicates I/O transceiver and control signal delay. Decoder indicates the sum of row and column address decoder delay. Wire indicates the sum of row and column address access bus delay with data out path

delay. WL and BL indicate sum of wordline and bitline delay. TSV represents for TSV delay.

The transceiver delay is the same in every cases because we assume fixed DDR3 I/O clock speed which is 1600 MHz. The decoder delay is also the same since the bank size of all cases was the same as 2 Mb, which maintains row and column decoder design. The WL and BL latency decreased from 3.544 ns in the planar case to 3.487 ns in the 32 die design case. The wire latency also decreased from approximately 2.1 ns to 1.5 ns as the TSV latency increased up to 0.988 ns.

Table 4.20 Throughput of 1 Gb 3D multibank DRAM, 6 F^2 layout in 35 nm node

Cell Layout	Num.of die	Bank Size (Mb)	Read Throughput (GByte/sec)
$6F^2$	1	2	0.366
	2	2	0.370
	4	2	0.374
	8	2	0.376
	16	2	0.375
	32	2	0.369
$4F^2$	1	2	0.383
	2	2	0.386
	4	2	0.388
	8	2	0.390
	16	2	0.388
	32	2	0.382

Figure 4.12 Delay sum of each design component with TSV in 35 nm node.

4.2.4 Product of Design Metric

Figure 4.13 shows the tendency of the best result of the combinations of area efficiency (AE), energy efficiency (EE), and throughput (TH) as the die count is varied. As discussed in Section 4.2.1, when the number of die increases, the memory cell area in a die becomes smaller compared to the overall die area, therefore the area efficiency is decreased. This tendency of the area efficiency also affects the combined design metric. Thus, the design metric products, which include the area efficiency such as Figure 4.13a, Figure 4.13b, and Figure 4.13d show the best metric value when they are planar DRAMs.

When we consider design metric as the combination of AE and EE as shown in Figure 4.13a, the DRAM exhibited higher performance at four layered design, compared to two die layered design in the case of the $6 F^2$ layout. Compared to planar design and two layered DRAM design, the area efficiency decreased approximately 26% while energy efficiency

increases about 6%. Compared to planar design and four layered DRAM design, the area efficiency decreased approximately 48% while energy efficiency increased approximately 50%. Thus, with four dies stacked DRAM has higher $AE \times EE$ combined performance than die of 2 design since it had higher energy efficiency. In the case of VCAT based design, the lower the number of layers, the higher the area and energy efficiency. Thus, the DRAM exhibited higher $AE \times EE$ combined performance with fewer layered design.

Considering the combined design metric of AE, EE, and TH, and for a DRAM designed with both $6 F^2$ and $4 F^2$ cell, the planar design case showed optimal results due to the influence of AE. However, as in the case of the combination of AE and EE in $6 F^2$ cell design, the DRAM design with four layers exhibited higher energy efficiency than the DRAM design with two layers which resulted in higher combined metric performance in 4 layered DRAM than two layered DRAM design in $6 F^2$ layout based design. In $4 F^2$ layout based design, Compared with the 2-layer design, the 4-layer design shows performance differences of -0.94%, 1.67%, and 0.21% for AE, TH, and EE, respectively, which combination exhibited higher $AE \times EE \times TH$ performance in 4-layer design.

Compared to the RCAT- and the VCAT-based cell area design, the RCAT-based design exhibited better area efficiency as discussed in Section 4.2.1. Thus, in the most metric product cases where the area efficiency is included (i.e., Figure 4.13a and Figure 4.13d), the RCAT-based design exhibited better efficiency as shown in Figure 4.13. When AE and TH are combined as shown in Figure 4.13b, $4 F^2$ layout based DRAM design shows better performance for planar and two layered design. The VCAT-base design displayed better throughput than the RCAT-base design in all cases, and the RCAT-based design displayed better area efficiency than the VCAT-based design in all cases. After 4 layered DRAM design, the throughput differences between RCAT- and VCAT-based design was about 0.03% to 14% while area efficiency difference was about 8% to 30%. Thus, the RCAT-based design

(a) $AE \times EE$.

(b) $AE \times TH$.

(c) $EE \times TH$.

(d) $AE \times EE \times TH$.

Figure 4.13 Multiple design metric trend in 35 nm node.

exhibited better performances after that point. For the EE and the TH combinations, VCAT based design displayed better efficiency trend mainly due to better the throughput as shown in Table 4.20.

4.2.5 Design Metric Comparison in Different Technology

Figure 4.14 shows the design metric peak value comparison between 16 nm, 35 nm, and 68 nm node when the die count is increased. In Figure 4.14a, an area efficiency difference of about 1% between 68 nm and 35 nm node was observed. The area efficiency difference between 68 nm and 16 nm node was approximately 3.0% to 5.5% after the second layer since TSV size scaled conservatively. In the 68 nm and 35 nm node design, the area occupied by the TSV were about 5% to 26% of the total die size as the die count is increased, while the 16 nm node design was about 11% to 40%.

Figure 4.14b shows the best energy efficiency as the number of die count is increased, for different technology node. The optimum energy efficiency was obtained when the number of layers is 8 in 35 nm and 16 nm node. At 68 nm node, the energy efficiency was optimal when the number of layers was 4.

Figure 4.14c presents the best throughput for various process node as the die count is increased. The throughput has the best efficiency in design with four dies in every node. As discussed in Sections 4.2.2 and Section 4.2.3, as the number of die layers increases, the share of TSV increases, resulting in lower energy efficiency and throughput.

As shown in Figure 4.14, the 16 nm node displayed the best performance in all respects. Comparing the best cases of 16 nm node and 68 nm node in the energy efficiency and the throughput, 16 nm node exhibited approximately 7.4 fold better in energy efficiency and approximately 3.3 fold better in throughput compared to the 68 nm node.

(a) Area efficiency

(b) Energy efficiency

(c) Throughput

Figure 4.14 Design metric comparison between 68 nm, 35 nm and 16 nm node in $6 F^2$ cell layout.

Table 4.21 shows the best performance metric results with design configuration in 16 nm node. In the case of area efficiency, 4 F^2 cell layout DRAM design exhibited the smaller area with 68.0% area efficiency compared the 6 F^2 layout. The energy efficiency had an optimum value when the die count was eight in both VCAT-based and RCAT-based design. For the throughput, the RCAT-based design exhibited the optimum point when the die count was eight while VCAT-based design also exhibited the optimum with eight dies. Considering both the throughput and the energy efficiency, the VCAT based design displayed the optimal value with eight dies and 2 Mb/bank design - the energy efficiency was 552.8 Bit/nJ, and throughput is 0.685 GByte/sec.

Table 4.21 The best performance metric results in 16 nm node

Performance Metric	4 F^2 Cell Layout		6 F^2 Cell Layout	
	Value	Configuration	Value	Configuration
AE (%)	68.0	Planar, Single bank (Area: 1.6 mm ²)	75.6	Planar, Single bank (Area:2.2 mm ²)
EE (Bit/nJ)	552.3	2 Mb/bank, 8 die	534.2	2 Mb/bank, 8 die
TH (GByte/sec)	0.694	2 Mb/bank, 8 die	0.636	2 Mb/bank, 8 die
TH×EE	0.379	2 Mb/bank, 8 die (EE:552.8, TH:0.685)	0.327	4 Mb/bank, 8 die (EE:528.3, TH:0.619)

CHAPTER

5

CONCLUSION AND FUTURE WORK

5.1 Summary of Contributions

In this dissertation, we have presented the three dimensional DRAM Area, Timing, and Energy model. The DATE model consists of a process roadmap and circuit level model which is using the roadmap. We have presented DATE process roadmap from 90 nm to 16 nm node:

- Using the physical dimensions in Rambus model [13] and the latest device articles, we have proposed transistor roadmaps by TCAD simulation and calculation for VCAT,

RCAT, high-voltage transistor and the peripheral transistor. The DATE roadmap has the most conservative value than the values already presented in Rambus model or CACTI [12].

- We have proposed a wire roadmap using the material parameters provided through ITRS roadmap [20] and the physical dimensions presented in the ITRS roadmap and the cross-sectional die report [51]. Compared to the anonymous logic design processes, poly and the metal layer 1 are conservatively predicted. The metal layer 2 and 3 are predicted larger size, therefore resistance values of DATE roadmap are smaller than the logic processes and capacitance values of DATE roadmap are more significant than the logic process.
- We adopted the physical dimension projection and model of TSV presented in CACTI-3DD [14].

We have implemented and verified circuit level modeling:

- The logic and buffer size was determined using the logical effort [54]. The wire repeater size was determined using Rabaey's model [55]. Sense amplifier and peripheral circuits followed the model proposed by CACTI [62]. The layout and arrangement of subarrays and banks, and the placement of peripheral logic were based on the DRAM architecture introduced in the Rambus model [13].
- We have validated the DATE by comparing commercial DRAM specifications and published data. Energy verification had a mean error of about -5% to 1%, with a standard deviation of up to 9.8%. Speed verification had a mean error of about -13% to -27% and a standard deviation of up to 24%. In the case of the area, the bank had a

mean error of -3% and the whole die had a mean error of -1%. The standard deviation for area was up to 4.2%.

We have explored the change of each design component in energy and speed and also explored the area change according to the design change in 1 Gb DDR3 DRAM. The best throughput we have achieved was about 0.7 Gb/sec and the best energy efficiency was about 589 bit/nJ in 16 nm node with VCAT based layout.

5.2 Future Work

There are several interesting directions in future research. For the device roadmap, we believe it would be interesting to update current roadmap with emerging devices such as FinFET-based gate transistors or emerging materials for the wire, which would impact overall speed. It would also be interesting to extend DATE model for evaluating the fine-grained 3D DRAM design by adding high-performance transistor roadmap.

In this dissertation, we only explored DDR3 based 1 Gb DRAM. There are potentially other applications such as 3D stacked DRAM on top of the processing unit. As an example, we believe it would be interesting to explore the customized DRAM design space for the machine learning hardware that requires high-speed, wider bandwidth memory.

Another interesting direction is implementing memory timing simulator, which could be coordinated with processor core simulator. This could potentially provide accurate timing and energy values according to the user scenario for the customized DRAM.

BIBLIOGRAPHY

- [1] Tsai, Y.-F. et al. "Design Space Exploration for 3-D Cache". *IEEE Transactions on Very Large Scale Integration (VLSI) Systems* **16.4** (2008), pp. 444–455.
- [2] *Hybrid Memory Cube Specification 1.0*. Hybrid Memory Cube Consortium. 2013.
- [3] *DDR3 SDRAM Specification*. Standard. JEDEC Solid State Technology Association, 2010.
- [4] *Wide I/O Single Data Rate*. Standard. JEDEC Solid State Technology Association, 2011.
- [5] *Wide I/O 2*. Standard. JEDEC Solid State Technology Association, 2014.
- [6] Kim, J.-S. et al. "A 1.2 V 12.8 GB/s 2 Gb Mobile Wide-I/O DRAM With 4 128 I/Os Using TSV Based Stacking". *IEEE Journal of Solid-State Circuits* **47.1** (2012), pp. 107–116.
- [7] Meswani, M. R. et al. "Toward Efficient Programmer-managed Two-level Memory Hierarchies in Exascale Computers". *Proceedings of the 1st International Workshop on Hardware-Software Co-Design for High Performance Computing*. Co-HPC '14. Piscataway, NJ, USA: IEEE Press, 2014, pp. 9–16.
- [8] Sekiguchi, T. et al. "1-Tbyte/s 1-Gbit DRAM Architecture Using 3-D Interconnect for High-Throughput Computing". *IEEE Journal of Solid-State Circuits* **46.4** (2011), pp. 828–837.
- [9] Loh, G. "3D-Stacked Memory Architectures for Multi-core Processors". *35th International Symposium on Computer Architecture, 2008. ISCA '08*. 2008, pp. 453–464.
- [10] Rotenberg, E. et al. "Rationale for a 3D heterogeneous multi-core processor". *2013 IEEE 31st International Conference on Computer Design (ICCD)*. 2013, pp. 154–168.
- [11] Song, K.-W. et al. "A 31 ns Random Cycle VCAT-Based 4F DRAM With Manufacturability and Enhanced Cell Efficiency". *IEEE Journal of Solid-State Circuits* **45.4** (2010), pp. 880–888.
- [12] Wilton, S. J. E. & Jouppi, N. P. "CACTI: an enhanced cache access and cycle time model". *IEEE Journal of Solid-State Circuits* **31.5** (1996), pp. 677–688.
- [13] Vogelsang, T. "Understanding the Energy Consumption of Dynamic Random Access Memories". *Proceedings of the 2010 43rd Annual IEEE/ACM International Symposium on Microarchitecture*. '43. Washington, DC, USA: IEEE Computer Society, 2010, pp. 363–374.

- [14] Chen, K. et al. "CACTI-3DD: Architecture-level modeling for 3D die-stacked DRAM main memory". *Design, Automation Test in Europe Conference Exhibition (DATE), 2012.* 2012, pp. 33–38.
- [15] *TN-41-01: Calculating Memory System Power for DDR3.* Tech. rep. Micron Technology Inc., 2013.
- [16] Chandrasekar, K. et al. "Improved Power Modeling of DDR SDRAMs". *2011 14th Euromicro Conference on Digital System Design (DSD).* 2011, pp. 99–108.
- [17] Chandrasekar, K. et al. "System and circuit level power modeling of energy-efficient 3D-stacked wide I/O DRAMs". *Proceedings of the Conference on Design, Automation and Test in Europe.* '13. San Jose, CA, USA: EDA Consortium, 2013, pp. 236–241.
- [18] Chandrasekar, K & on, so. "DRAMPower: Open-source DRAM Power and energy estimation tool". 2012.
- [19] Weis, C. et al. "Exploration and Optimization of 3-D Integrated DRAM Subsystems". *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems* **32.4** (2013), pp. 597–610.
- [20] Akasaka, Y. et al. *Process Integration, Device and Structures.* Tech. rep. International Technology Roadmap for Semiconductors.
- [21] Kim, J. et al. "The breakthrough in data retention time of DRAM using Recess-Channel-Array Transistor(RCAT) for 88 nm feature size and beyond". *2003 Symposium on VLSI Technology, 2003. Digest of Technical Papers.* 2003, pp. 11–12.
- [22] Kim, J. et al. "The excellent scalability of the RCAT (recess-channel-array-transistor) technology for sub-70nm DRAM feature size and beyond". *2005 IEEE VLSI-TSA International Symposium on VLSI Technology, 2005. (VLSI-TSA-Tech).* 2005, pp. 33–34.
- [23] Kim, J.-Y. et al. "S-RCAT (sphere-shaped-recess-channel-array transistor) technology for 70nm DRAM feature size and beyond". *2005 Symposium on VLSI Technology, 2005. Digest of Technical Papers.* 2005, pp. 34–35.
- [24] Yoon, J.-M. et al. "A Novel Low Leakage Current VPT(Vertical Pillar Transistor) Integration for 4F2 DRAM Cell Array with sub 40 nm Technology". *Device Research Conference, 2006 64th.* 2006, pp. 259–260.

- [25] Katsumata, R. et al. "Fin-Array-FET on bulk silicon for sub-100 nm trench capacitor DRAM". *2003 Symposium on VLSI Technology, 2003. Digest of Technical Papers.* 2003, pp. 61–62.
- [26] Lee, C. et al. "Enhanced data retention of damascene-finFET DRAM with local channel implantation and <100> fin surface orientation engineering". *Electron Devices Meeting, 2004. IEDM Technical Digest. IEEE International.* 2004, pp. 61–64.
- [27] Chung, S.-W. et al. "Highly Scalable Saddle-Fin (S-Fin) Transistor for Sub-50nm DRAM Technology". *2006 Symposium on VLSI Technology, 2006. Digest of Technical Papers.* 2006, pp. 32–33.
- [28] Kim, Y.-S. et al. "Local-damascene-finFET DRAM integration with p+ doped polysilicon gate technology for sub-60nm device generations". *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International.* 2005, pp. 315–318.
- [29] Siddiqi, M. A. "Advanced DRAM Cell Transistors". *Dynamic RAM, Technology Advancements.* CRC Press, 2013, pp. 100–154.
- [30] Song, J. Y. et al. "Fin and recess channel MOSFET (FiReFET) for performance enhancement of Sub-50 nm DRAM cell". *Semiconductor Device Research Symposium, 2007 International.* 2007, pp. 1–2.
- [31] Kim, Y.-S. et al. "Fabrication and Electrical Properties of Local Damascene FinFET Cell Array in Sub-60 nm Feature Sized DRAM". *Journal of semiconductor technology and science* **6.2** (2006), pp. 61–67.
- [32] Keeth, B. et al. *DRAM circuit design: fundamental and high-speed topics.* Wiley-IEEE Press, 2007.
- [33] Tran, L. C. "6f2 dram array, a dram array formed on a semiconductive substrate, a method of forming memory cells in a 6f2 dram array and a method of isolating a single row of memory cells in a 6f2 dram array". Pat. US6545904 B2. U.S. Classification 365/149, 257/E21.66, 257/906, 365/63, 257/E21.654; International Classification H01L27/02, H01L21/8242; Cooperative Classification Y10S257/906, Y10S438/981, H01L27/0214, H01L27/10894, H01L27/10873; European Classification H01L27/108M4C. 2003.
- [34] Park, Y. et al. "Fully Integrated 56 nm DRAM Technology for 1 Gb DRAM". *2007 IEEE Symposium on VLSI Technology.* 2007, pp. 190–191.
- [35] Cho, C. et al. "A 6F2 DRAM technology in 60nm era for gigabit densities". *2005 Symposium on VLSI Technology, 2005. Digest of Technical Papers.* 2005, pp. 36–37.

- [36] Ananthan, V. et al. "Derivation of threshold voltage and drain current for cylindrical MOSFET and application to a recessed MOSFET". *2008 IEEE Workshop on Micro-electronics and Electron Devices*. 2008, pp. 9–11.
- [37] Hwang, T. J. et al. "Work function measurement of tungsten polycide gate structures". en. *Journal of Electronic Materials* **12.4** (1983), pp. 667–679.
- [38] Efavi, J. K. et al. "Tungsten work function engineering for dual metal gate nano-CMOS". en. *Journal of Materials Science: Materials in Electronics* **16.7** (2005), pp. 433–436.
- [39] Lee, J. et al. "Improvement of data retention time in DRAM using recessed channel array transistors with asymmetric channel doping for 80 nm feature size and beyond". *Solid-State Device Research conference, 2004. ESSDERC 2004. Proceeding of the 34th European*. 2004, pp. 449–452.
- [40] *Double Data Rate (DDR) SDRAM Specification*. Standard. JEDEC Solid State Technology Association, 2003.
- [41] Ha, D. et al. "Self-Aligned Local Channel Implantation Using Reverse Gate Pattern for Deep Submicron Dynamic Random Access Memory Cell Transistors". en. *Japanese Journal of Applied Physics* **37.3S** (1998), p. 1059.
- [42] Kim, K. et al. "Extending the DRAM and FLASH memory technologies to 10nm and beyond". Vol. 8326. 2012, pp. 832605–832605–11.
- [43] *1Gb C-die DDR2 SDRAM Specification, K4T1G084QC*. Rev. 1.3. Samsung electronics. 2008.
- [44] Mueller, W. et al. "Challenges for the DRAM cell scaling to 40nm". *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*. 2005, 4 pp.–339.
- [45] Tran, T. et al. "A 58nm Trench DRAM Technology". *Electron Devices Meeting, 2006. IEDM '06. International*. 2006, pp. 1–4.
- [46] Kim, K. "Technology for sub-50nm DRAM and NAND flash manufacturing". *Electron Devices Meeting, 2005. IEDM Technical Digest. IEEE International*. 2005, pp. 323–326.
- [47] Skotnicki, T. et al. "MASTAR - A Model For Analog Simulation Of Subthreshold, Saturation And Weak Avalanche Regions In MOSFETs". (1993 VPAD) *1993 International Workshop on VLSI Process and Device Modeling, 1993*. 1993, pp. 146–147.
- [48] Ho, R. et al. "The future of wires". *Proceedings of the IEEE* **89.4** (2001), pp. 490–504.

- [49] Amakawa, S. et al. *Interconnect*. Tech. rep. International Technology Roadmap for Semiconductors, 2009.
- [50] Radi, H. A. & Rasmussen, J. O. *Principles of Physics*. Undergraduate Lecture Notes in Physics. DOI: 10.1007/978-3-642-23026-4. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013.
- [51] No.051130036QP:Samsung 512Mb C-die DDR SDRAM Constructional Analysis. Tech. rep. Materials Analysis Technology Inc., 2008.
- [52] Saraswat, K. C. et al. "Properties of low-pressure CVD tungsten silicide for MOS VLSI interconnections". *IEEE Transactions on Electron Devices* **30.11** (1983), pp. 1497–1505.
- [53] Katti, G. et al. "Electrical Modeling and Characterization of Through Silicon via for Three-Dimensional ICs". *IEEE Transactions on Electron Devices* **57.1** (2010), pp. 256–262.
- [54] Sutherland, I. et al. *Logical Effort*. Morgan Kaufmann, 1999.
- [55] Rabaey, J. M. et al. *Digital integrated circuits: A design perspective*. Pearson, 2003.
- [56] Thoziyoor, S. et al. "A Comprehensive Memory Modeling Tool and Its Application to the Design and Analysis of Future Memory Hierarchies". *Proceedings of the 35th Annual International Symposium on Computer Architecture*. ISCA '08. Washington, DC, USA: IEEE Computer Society, 2008, pp. 51–62.
- [57] Hedenstierna, N. & Jeppson, K. O. "Comments on the optimum CMOS tapered buffer problem". *IEEE Journal of Solid-State Circuits* **29.2** (1994), pp. 155–158.
- [58] Harris, D. et al. *The Fanout-of-4 Inverter Delay Metric*. data retrieved from CiteSeerX, <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.68.831&rep=rep1&type=pdf>. 2014.
- [59] Horowitz, M. A. *Timing Models for MOS Circuits*. Tech. rep. Stanford, CA, USA: Stanford University, 1983.
- [60] Adler, V. & Friedman, E. G. "Repeater design to reduce delay and power in resistive interconnect". *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing* **45.5** (1998), pp. 607–616.

- [61] Elmore, W. C. "The Transient Response of Damped Linear Networks with Particular Regard to Wideband Amplifiers". *Journal of Applied Physics* **19.1** (1948), pp. 55–63. eprint: <https://doi.org/10.1063/1.1697872>.
- [62] Thoziyoor, S. et al. *CACTI 5.1: A Tool to Model Large Caches*. Research Report. HP Labs, 2008.
- [63] Jacob, B. et al. *Memory Systems: Cache, DRAM, Disk*. Morgan Kaufmann, 2008.
- [64] Lin, C. *Techinsights: A Closer Look at Recent DRAM Releases*. Report. TechInsights Inc., 2011.
- [65] Kang, U. et al. "8 Gb 3-D DDR3 DRAM Using Through-Silicon-Via Technology". *IEEE Journal of Solid-State Circuits* **45.1** (2010), pp. 111–119.
- [66] Vogelsang, T. *User Manual DRAM Core Model*. Rambus Inc. 2010.
- [67] Choi, Y. *Under the Hood: DRAM architectures: 8F2 vs. 6F2*. 2008. URL: https://www.eetimes.com/document.asp?doc_id=1281208 (visited on 11/13/2017).
- [68] *Samsung 68nm 1GB DDR2 SDRAM (K4T1G084QD-ZCE6)*. Report. TechInsights Inc., 2008.
- [69] *1Gb A-die DDR2 SDRAM Specification, K4T1G084QA*. Rev. 1.1. Samsung electronics. 2005.
- [70] *512Mb E-die DDR2 SDRAM Specification, K4T51083QE*. Rev. 1.8. Samsung electronics. 2007.
- [71] *2Gb A-die DDR2 SDRAM Specification, K4T2G084QA*. Rev. 1.3. Samsung Electronics. 2008.
- [72] *1Gb Q-die DDR2 SDRAM Specification, K4T1G084QQ, K4T1G164QQ*. Rev. 1.2. Samsung Electronics. 2008.
- [73] *1Gb D-die DDR2 SDRAM Specification*. Rev. 1.0. Samsung Electronics. 2007.
- [74] *2Gb: x16, x32 Mobile LPDDR2 SDRAM S4 Features, MT42L64M32D1*. Rev. N. Micron. 2010.
- [75] *1Gb: x4, x8, x16 DDR2 SDRAM Features, MT47H128M8, MT47H64M16*. Rev. AA. Micron. 2007.
- [76] *2G bits DDR3 SDRAM, EDJ2108BCSE*. Rev. 2.1. Elpida. 2011.

- [77] *2Gb DDR3 SDRAM, H5TQ2G83BFR*. Rev. 0.2. SK Hynix. 2010.
- [78] *1.35V DDR3L SDRAM, MT41K256M8*. Rev. G. Micron. 2010.
- [79] *2Gb D-die DDR3L SDRAM, K4B2G0846D*. Rev. 1.01. Samsung Electronics. 2010.
- [80] *4Gb: x4, x8, x16 DDR3L SDRAM Description, MT41K512M8, MT41K256M16*. Rev. O. Micron. 2009.
- [81] *4Gb: x4, x8, x16 DDR3 SDRAM Description, MT41J512M8, MT41J256M16*. Rev. S. Micron. 2006.
- [82] *2Gb: x4, x8, x16 DDR3 SDRAM Features, MT41J256M8*. Rev. N. Micron. 2009.
- [83] *2G bits DDR3 SDRAM, EDJ2108BDBG*. Ver. 4. Elpida. 2012.
- [84] Chandrasekar, K. et al. *DRAMPower: Open-source DRAM Power & Energy Estimation Tool*. 2014. URL: <http://www.drampower.info> (visited on 11/14/2017).
- [85] *TwinDieTM 1.35V DDR3L SDRAM, MT41K1G8*. Rev. J. Micron. 2011.
- [86] Das, A. *Hynix DRAM layout, process integration adapt to change*. 2012. URL: https://www.eetimes.com/document.asp?doc_id=1280240 (visited on 01/16/2018).

APPENDICES

APPENDIX

A

DERIVATION OF THE LEAKAGE CURRENT EQUATION

Figure A.1 shows the change of bitline voltage against time. According to the charge conservation law, the amount of total charge before opening the gate transistor, and after the opening should be equal. Thus,

$$(V_{array} - \Delta V)C_S + V_{PRE}C_B = V_{BL}(C_S + C_B) \quad (A.1)$$

Where ΔV is data loss, V_{BL} is bitline voltage after charge equilibrium, V_{PRE} is bitline

precharge voltage, C_B is bitline capacitance, and C_S is storage node capacitance. If we assume $V_{PRE} = 1/2 V_{array}$, then the equation is:

$$(V_{array} - \Delta V)C_S + \frac{1}{2} V_{array} C_B = V_{BL}(C_S + C_B) \quad (A.2)$$

Then,

$$V_{BL} = \frac{C_S}{(C_S + C_B)} \times (V_{array} - \Delta V) + \frac{C_B}{(C_S + C_B)} \times \frac{1}{2} V_{array} \quad (A.3)$$

The complementary bitline charged $V_{PRE} = 1/2 V_{array}$. Thus, the differential signal, ΔV_{BL} is:

$$\Delta V_{BL} = V_{BL} - \overline{V_{BL}} = \frac{C_S}{C_B + C_S} \times (\frac{1}{2} V_{array} - \Delta V). \quad (A.4)$$

Since the current is $i = C \frac{dV}{dt}$, the maximum allowable cell storage leakage current is

$$\frac{C_S \times \Delta V_{MAX}}{t_{REF}} = I_{off} \quad (A.5)$$

where ΔV_{MAX} is maximum allowable data loss, t_{REF} is retention time.

If we reorganize equation A.4 with respect to ΔV and substitute it into equation A.5, then it is:

$$I_{off} = \frac{C_S(V_{array}/2 - \Delta V_{BL}) - C_B \Delta V_{BL}}{t_{REF}}. \quad (A.6)$$

Figure A.1 Variation of bitline voltage with time

APPENDIX

B

TCAD SIMULATION CODE

B.1 Sentaurus Structure Editor Code

Listing B.1 44nm SRCAT, for Build Device Structure

```
;-----  
; SRCAT 44nm – ASC  
(sde:clear)  
; fixed parameters  
(define Lg 0.044) ; [um] Gate length  
(define Tox 0.0047) ; [um] Gate oxide thickness  
  
; – lateral
```

```

(define Ltot 1.0) ; [um] Lateral extend total

(define Lsp ( 0.3 Lg)) ; [um] Spacer length
(define Ldia ( 1.5 Lg)) ; [um] diameter
(define Lrad ( 0.5 Ldia)) ; [um] radius

; - layers
(define Hsub 0.0) ; [um] Substrate thickness
(define Hbox 0.0) ; [um] Buried oxide thickness
(define Hepi 1.0) ; [um] EPI thickness

(define Hpol 0.1) ; [um] Poly gate thickness

; - other
; - spacer rounding
(define fillet-radius 0.08) ; [um] Rounding radius

; - pn junction resolution
(define Gpn 0.001) ; [um]

(define TTD @Trench@) ; [um] Total tranch depth
(define IR (- Lrad Tox)) ; inner radius
(define OR Lrad) ; outter radius

;-----
; Doping quantities

(define ChDop @ChDop@) ; Asymetric Channel Doping Peak Value
(define ChPeakPos @ChPeakPos@)

(define WafDop 1e14); [1/cm3] p-type wafer doping

; junction depth

```

```

(define ChXj @ChXj@) ; Channel Doping junction depth at 1e15 Doping density

; Ars; Arsenic for N type S/D doping
; 1eV, 1.5e12, 0.5min 1100 → 5.3e17 0.05; Arsenic for N type S/D doping
(define SDContDop 5.3e17) ; Contact Doping Peak Value
(define SDXj 0.05) ; SD Doping junction depth at 1e14 doping density

;-----
; Derived quantities
(define Xmax (/ Ltot 2.0))
(define Xg (/ Lg 2.0)) ; Half of Gate Size
(define Xsp (+ Xg Lsp)) ; Spacer starting point

(define Yepi Hepi)
(define Ybox (+ Yepi Hbox))
(define Ysub (+ Ybox Hsub))

(define Ygox ( Tox -1.0))
(define Ypol (- Ygox Hpol))

(define Lcont (- Xmax Xsp))

;-----
; Overlap resolution: New replaces Old
(sdegeo:set-default-boolean "ABA")

;-----
; Creating substrate region
(sdegeo:create-rectangle
  (position ( Xmax -1.0) 0.0 0.0 )
  (position Xmax Ysub 0.0 ) "Silicon" "R.Substrate" )

; Create Circular gate oxide
(sdegeo:create-circular-sheet (position 0.0 (- TTD Lrad) 0.0) Lrad "SiO2" "R.Gateox2" )

```

```

; Creating gate oxide region
(sdegeo:create-polygon
  (list
    (position ( Xsp -1.0) 0.0 0.0 )
    (position ( Xg -1.0) 0.0 0.0 )
    (position ( Xg -1.0) (- TTD Lrad) 0.0 )
    (position Xg (- TTD Lrad) 0.0)
    (position Xg 0.0 0.0)
    (position Xsp      0.0      0.0)
    (position Xsp      Ypol 0.0 )
    (position ( Xsp -1) Ypol 0.0 )
  )
  "SiO2" "R.Gateox"
)

; Creating PolySi gate
(sdegeo:create-rectangle
  (position ( (- Xg Tox) -1.0) (- (- TTD Tox) Lrad) 0.0 )
  (position (- Xg Tox) Ypol 0.0 )
  "PolySi" "R.Polygate"
)
(sdegeo:create-circular-sheet (position 0.0 (- TTD Lrad) 0.0) (- Lrad Tox) "PolySi" "R.Polygate2" )

;-----
; - rounding left spacer
(sdegeo:fillet-2d (find-vertex-id (position ( Xsp -1.0) Ypol 0.0 )) fillet-radius)
; - rounding right spacer
(sdegeo:fillet-2d (find-vertex-id (position Xsp Ypol 0.0 )) fillet-radius)

;-----
; Contact declarations
(sdegeo:define-contact-set "source"
  4.0 (color:rgb 1.0 0.0 0.0 ) "##")

```

```

(sdegeo:define-contact-set "drain"
  4.0 (color:rgb 0.0 1.0 0.0) "##")

(sdegeo:define-contact-set "gate"
  4.0 (color:rgb 0.0 0.0 1.0) "##")

(sdegeo:define-contact-set "substrate"
  4.0 (color:rgb 0.0 1.0 1.0) "##")

;-----
; Contact settings
(sdegeo:define-2d-contact
  (find-edge-id (position ( (+ Xmax Xsp) -0.5) 0.0 0.0))
  "source" )

(sdegeo:define-2d-contact
  (find-edge-id (position ( (+ Xmax Xsp) 0.5) 0.0 0.0))
  "drain")

(sdegeo:define-2d-contact
  (find-edge-id (position 0.0 Ysub 0.0))
  "substrate")

(sdegeo:set-current-contact-set "gate")
(sdegeo:set-contact-boundary-edges (find-body-id (position 0.0 0.0 0.0)) )
(sdegeo:set-contact-boundary-edges (find-body-id (position 0.0 (- TTD Lrad) 0.0)) )
(sdegeo:delete-region (find-body-id (position 0.0 0.0 0.0)) )
(sdegeo:delete-region (find-body-id (position 0.0 (- TTD Lrad) 0.0)) )

;-----
; Saving BND file
(define SOI (get-body-list))
(sdeio:save-tdr-bnd SOI "@tdrboundary/o@")

;-----

```

```

; Implant Window Definition
;
; Source/Drain extensions base line definitions
(sdedr:define-refinement-window "BaseLine.SourceExt" "Line"
  (position ( Xmax -2.0) 0.0 0.0)
  (position ( Xg -1.0) 0.0 0.0) )
(sdedr:define-refinement-window "BaseLine.DrainExt" "Line"
  (position Xg 0.0 0.0)
  (position ( Xmax 2.0) 0.0 0.0) )

(sdedr:define-refinement-window "BaseLine.Drain" "Line"
  (position Xsp 0.0 0.0)
  (position ( Xmax 2.0) 0.0 0.0) )
;
; Profiles:
; - Wafer Own Doping
(sdedr:define-constant-profile "Const.SiEpi"
  "BoronActiveConcentration" WafDop )
(sdedr:define-constant-profile-region "PlaceCD.SiEpi"
  "Const.SiEpi" "R.Substrate" )
;
; Profiles
; - Substrate Asymmetric Channel Doping
; ASC implant profile definition
(sdedr:define-gaussian-profile "Impl.ASCprof"
  "BoronActiveConcentration"
  "PeakPos" ChPeakPos "PeakVal" ChDop
  "ValueAtDepth" WafDop "Depth" ChXj
  "Gauss" "Factor" 0.25)

; ASC Drain implants
(sdedr:define-analytical-profile-placement "Impl.Drain"
  "Impl.ASCprof" "BaseLine.Drain" "Symm" "NoReplace" "Eval")
;
; Source/Drain implant definition

```

```

(sdedr:define-gaussian-profile "Impl.SDextprof"
  "ArsenicActiveConcentration"
  "PeakPos" 0 "PeakVal" SDContDop
  "ValueAtDepth" WafDop "Depth" SDXj
  "Gauss" "Factor" 0.25
)

; Source/Drain implants
(sdedr:define-analytical-profile-placement "Impl.SourceExt"
  "Impl.SDextprof" "BaseLine.SourceExt" "Symm" "NoReplace" "Eval")
(sdedr:define-analytical-profile-placement "Impl.DrainExt"
  "Impl.SDextprof" "BaseLine.DrainExt" "Symm" "NoReplace" "Eval")

;-----
; Meshing Strategy:

; EPI -> Substrate
(sdedr:define-refinement-size "Ref.SiEpi"
  (/ Lcont 4.0) (/ Hepi 8.0)
  Gpn Gpn )
(sdedr:define-refinement-function "Ref.SiEpi"
  "DopingConcentration" "MaxTransDiff" 1)
(sdedr:define-refinement-region "RefPlace.SiEpi"
  "Ref.SiEpi" "R.Substrate" )

; Channel Multibox1
(sdedr:define-refinement-window "MBWindow.Channel"
  "Rectangle"
  (position ( Xg -2) 0.0 0.0)
  (position ( Xg 2) (- TTD ( Ldia 1.2)) 0.0) )
(sdedr:define-multibox-size "MBSize.Channel"
  (/ Lg 8.0) (/ TTD 10.0)
  (/ Lg 12.0) (/ TTD 14.0)
  1.2 1.2 )
(sdedr:define-multibox-placement "MBPlace.Channel"
  "MBSize.Channel" "MBWindow.Channel" )

```

```

; Channel Multibox2
(sdedr:define-refinement-window "MBWindow.Gate"
  "Rectangle"
  (position ( Ldia -1.2) (- TTD ( Ldia 1.2)) 0.0)
  (position ( Ldia 1.2) (+ TTD ( Ldia 0.4)) 0.0) )
(sdedr:define-multibox-size "MBSize.Gate"
  (/ Lg 10.0) (/ TTD 10.0)
  (/ Lg 12.0) (/ TTD 12.0)
  1.2 1.2 )
(sdedr:define-multibox-placement "MBPlace.Gate"
  "MBSize.Gate" "MBWindow.Gate" )

;
; Save CMD file
(sdedr:write-cmd-file "@commands/o@")
(sde:build-mesh "snmesh" "-offset" "44-re")

```

B.2 Sentaurus Device Code

Listing B.2 44nm SRCAT, for Device Simulation

```

#-----#
#--Setting global variable:
#set Vdd @Vdd@
#define _EQUATIONS_ Poisson Electron Hole eQuantumPotential
device NMOS{
  Electrode {
    { Name="source" Voltage= 0.0 Resistor= 40 }
    { Name="drain" Voltage= 0.0 Resistor= 40 }
    { Name="gate" Voltage= 0.0 workfunction = 5.12 }
    { Name="substrate" Voltage= 0.0 Resistor= 40 }
  }

  Thermode {
    {name = "source" temperature=@Kelvin@ surfaceresistance=5e-4}

```



```

{name = "drain" temperature=@Kelvin@ surfaceresistance=5e-4}
{name = "gate" temperature=@Kelvin@ surfaceresistance=5e-4}
{name = "substrate" temperature=@Kelvin@ surfaceresistance=5e-4}
}

File {
#--Predefined SWB paraneters
    Grid      = "@tldr@"
    Plot       = "@tdrdat@"
    Current    = "@plot@"
    Output     = "@log@"
}

Physics {
    Thermodynamic
#--Using local variables:
    _TRANSMOD_
    eQCvanDort -- jpark
    eQuantumPotential add jpark
    EffectiveIntrinsicDensity( OldSlotboom )
    Mobility(
        DopingDep PhuMob
        eHighFieldsaturation( GradQuasiFermi )
        hHighFieldsaturation( GradQuasiFermi )
        Enormal
    )
    Recombination( SRH(DopingDep TempDependence) ) add TempDependence jpark
}

Plot {
--Density and Currents, etc
    eDensity hDensity
    TotalCurrent/Vector eCurrent/Vector hCurrent/Vector
    eMobility hMobility
    eVelocity hVelocity
    eQuasiFermi hQuasiFermi jpark erase

--Temperature

```

```

    Temperature eTemperature    hTemperature

--Fields and charges
    ElectricField/Vector Potential SpaceCharge

--Doping Profiles
    Doping DonorConcentration AcceptorConcentration

--Generation/Recombination
    SRH Band2Band    Auger
    AvalancheGeneration eAvalancheGeneration hAvalancheGeneration

--Driving forces
    eGradQuasiFermi/Vector hGradQuasiFermi/Vector
    eEparallel hEparallel eENormal hENormal

--Band structure/Composition
    BandGap
    BandGapNarrowing
    Affinity
    ConductionBand ValenceBand
    eQuantumPotential

--Traps
    eTrappedCharge hTrappedCharge
    eGapStatesRecombination hGapStatesRecombination
}
}
File {
    output = "@log@"
    ACExtract = "@acplot@"
}

System {
    NMOS trans( drain=d source=s gate=g substrate=b )
    Vsource_pset vd(d 0) {dc=0}
    Vsource_pset vs(s 0) {dc=0}
    Vsource_pset vg(g 0) {dc=0}
}

```

```

        Vsource_pset vb(b 0) {dc=0}
    }

    Math {
        Number_Of_Threads=16
        Extrapolate
        Avalderivatives
        RelErrControl
        Digits=4
        ErRef(electron)=1.e10
        ErRef(hole)=1.e10
        Notdamped=100
        Iterations=20
        DirectCurrent
        ExitOnFailure
    }

    Solve {
        NewCurrentPrefix="init "
        Coupled(Iterations=100){ Poisson eQuantumPotential }
        Coupled{ Poisson Electron Hole eQuantumPotential Temperature }

        - Bias drain to target bias
        Quasistationary(
            InitialStep=0.1 Increment=2.0
            Minstep=1e-5 MaxStep=1
            Goal{ Parameter=vd.dc Voltage= @Vd@ }
        ){ Coupled{ _EQUATIONS_ Temperature} }

        - gate voltage sweep
        NewCurrentPrefix="IdVg_"
        Quasistationary(
            DoZero
            InitialStep=1e-3 Increment=1.5
            Minstep=1e-5 MaxStep=0.05
            Goal{ Parameter=vg.dc Voltage= @Vdd@ }
        )
    }

```

```

{ ACCoupled(
    StartFrequency=1e6 EndFrequency=1e6
    NumberOfPoints=1 Decade
    Node(d s g b) Exclude(vd vs vg vb)
    ACCompute(Time=(Range=(0 1) Intervals=@<int(Vdd/0.1)>@ ))
)
{ _EQUATIONS_ Temperature}
}
}

```

B.3 Inspect Code

Listing B.3 44nm SRCAT, for Data Inspection

```

#setdep @node|sdevice@

array set Vth {}

# Automatic alternating color assignment tied to node index
# -----
set COLORS [list green blue red orange magenta violet brown]
set color [lindex $COLORS [expr @node@[length $COLORS] ]]

# INSPECT IdVg plotting
# -----
# Plotting Id vs Vg curves
set projname IdVg_trans_n@node|sdevice@
set curvename IdVg_trans_n@node@

set projname2 IdVg_n@node|sdevice@_ac
set curvename2 IdVg_@node@_ac

proj_load ${projname}_des.plt $projname
proj_load ${projname2}_des.plt $projname2

cv_createDS $curvename \
"$projname gate OuterVoltage" "$projname drain TotalCurrent" y

```

```

cv_abs $curvename y
cv_setCurveAttr $curvename \
    "n@node|sde@: Vds=@Vd@ Kelvin=@Kelvin@" \
    $color solid 2 none 3 defcolor 1 defcolor

gr_setTitleAttr "Idrain versus Vgate"
gr_setAxisAttr X {Gate Voltage (V)} 16 {} {} black 1 14 0 5 0
gr_setAxisAttr Y {Drain Current (A/um)} 16 {} {} black 1 14 0 5 1
gr_setLegendAttr 1 Helvetica 10 {} white black black 1 top

cv_createDS $curvename2 \
    "$projname2 NO_NODE v(g)" "$projname2 NO_NODE c(g,g)" y
cv_setCurveAttr $curvename2 \
    "n@node|sde@: Vds=@Vd@ Kelvin=@Kelvin@" \
    $color solid 2 none 3 defcolor 1 defcolor

gr_setTitleAttr "Total GateCap versus Vgate"
gr_setAxisAttr X {Gate Voltage (V)} 16 {} {} black 1 14 0 5 0
gr_setAxisAttr Y {Gate Capacitance (F/um)} 16 {} {} black 1 14 0 5 1
gr_setLegendAttr 1 Helvetica 10 {} white black black 1 top

#PlotCV

# — Extraction
# —————#
load_library EXTRACT

# —————#
set filename "threshold.txt"

# if @< Vd < 0.6 >@

set SS [ExtractSS SS $curvename 0.01]
set Vth(1) [ExtractVtgmb Vt $curvename]
set gm [ExtractGmb gm $curvename]
set Cgg [ExtractValue "Cgg" $curvename2 @Vdd@]

```

```

#- Make a threshold voltage file
set fileId [open $filename "w"]
puts -nonewline $fileId $Vth(1)
close $fileId

#else

set in [open $filename "r"]
gets $in Vth(1)

set Vth(2) [ExtractVtgmb Vt $curvename]
set gm [ExtractGmb gm $curvename]
set Idsat [ExtractMax Id $curvename]
set Idoff [ExtractIoff Ioff $curvename 1e-4]
set Cgg [ExtractValue "Cgg" $curvename2 @Vdd@]

set DIBL [expr { ($Vth(2) - $Vth(1))/(@Vd@ - 0.05)}]
ft_scalar DIBL [format %.3f $DIBL]

close $in

#endif

```

APPENDIX

C

DEFINITION AND DERIVATION OF THE PATH EFFORT

In the multistage gate, the path effort, F is defined as:

$$F = GHB \tag{C.1}$$

where G is path logical effort, H is path electrical effort, B is branching effort of the entire path. The definition is reminiscent of Equation 3.4 which presents the single stage case.

The path logical effort, G is defined as:

$$G = \prod g_i \quad (C.2)$$

where g_i is the logical effort, g at stage i . The electrical effort along the path, H is simple ratio between input capacitance of the first stage ($C_{in,first}$) and loaded capacitance at the output of the last stage ($C_{load,last}$).

$$H = \frac{C_{load,last}}{C_{in,first}} \quad (C.3)$$

The branching effort accounts for fanout within a network. Branching effort of a random stage i among the multistage, b_i , is defined at the output node of a logic gate. The definition is given by:

$$b_i = \frac{C_{total,i}}{C_{on-path,i}} \quad (C.4)$$

where C_{total} is total capacitance which is loaded at the output of the stage and $C_{on-path}$ is the load capacitance along the path that we analyze. Thus, the branching effort along the entire path, B is defined by:

$$B = \prod b_i \quad (C.5)$$

where i indicates any single stage along the path.

Since the branching and electrical efforts are related to the electrical effort at the output of a logic stage:

$$HB = \frac{C_{out,last}}{C_{input,1st}} \prod b_i = \prod h_i \quad (C.6)$$

Thus, the path effort is driven as:

$$F = GHB = \prod g_i h_i = \prod f_i \quad (\text{C.7})$$

APPENDIX

D

REFERENCE OF COMMODITY DRAM PART

The process node information of each commodity DRAM is obtained from various websites and technical reports. When each title is clicked, the website is accessed. The final link check date was Nov.14th 2017. The details are as follows:

- K4T51083QE-ZCE7: 80 nm 512 Mb DDR2 DRAM is obtained from the EE-Times article, "Under the Hood: DRAM architectures: 8F2 vs. 6F2."
- K4T1G084QA-ZCD5: 80 nm 1 Gb DDR2 DRAM is obtained from Samsung website,

"SAMSUNG Electronics First to Mass-produce 1Gb DDR2 Memory with 80nm Process Technology."

- K4T1G084QC-ZCE6: 80 nm 1 Gb DDR2 DRAM is obtained from Samsung website, "SAMSUNG Electronics First to Mass-produce 1Gb DDR2 Memory with 80nm Process Technology."
- K4T1G084QD-ZCE6: 68 nm 1 Gb DDR2 DRAM is obtained from TechInsights Inc. website, "Multi-temperature Transistor Characteristics of the Samsung K4T1G084QD-ZCE6 68nm 1Gb DDR2 SDRAM."
- K4T2G084QA-HCF7: 68 nm 2 Gb DDR2 DRAM is obtained from Samsung website, "SAMSUNG Introduces 60nm-class Processing for 2Gb DDR2 DRAM".
- K4T1G164QQ-HC(L)E6, K4T1G084QQ-HC(L)E6: 68 nm 1 Gb DDR2 DRAMs are obtained from Samsung website, "Samsung Electronics Begins World's first DRAM Mass Production using 60nm-Class Technology."
- MT42L64M32D1KL: 50 nm 2 Gb LPDDR2 DRAM is obtained from the technical report of TechInsight Inc., named as "A Closer Look at Recent DRAM Releases, August 2011."
- MT47H64M16HR, MT47H128M8CF: 50 nm 1 Gb DDR2 DRAMs are obtained from Micron's "Product Change Notice: 30939."
- EDJ2108BCSE: 44 nm 2 Gb DDR3 DRAM is obtained from the technical report of TechInsight Inc., named as "A Closer Look at Recent DRAM Releases, August 2011."
- H5TQ2G83BFR: 44 nm 2 Gb DDR3 DRAM is obtained from the EE-times article, "Hynix DRAM layout, process integration adapt to change."

- MT41K256M8DA: 42 nm 2 Gb DDR3 DRAM is obtained from the DDR3 module information of Lenovo's "Hardware Maintenance Manual: Lenovo E49 and ZhaoYang E49."
- K4B2G0846D-HYK0: 35 nm 2 Gb DDR3 DRAM is obtained from the DDR3 module information of Lenovo's "Hardware Maintenance Manual: Lenovo E49 and ZhaoYang E49."
- MT41J512M8RH-093:E, MT41J256M16HA-107:E, MT41K512M8RH-125:E : 30 nm 4 Gb DDR3 DRAMs are obtained from Micron's "EOL of Micron's 4Gb DDR3 30nm product: PCN 31722"
- MT41J256M8DA-125:K : 30 nm 4 Gb DDR3 DRAMs are obtained from Micron's "Product Change Notice: 30832"
- EDJ2108BDBG-GN: 30 nm 2Gb DDR3 DRAM is obtained from the DDR3 module information of Lenovo's "Hardware Maintenance Manual: Lenovo E49 and ZhaoYang E49."
- MT41K1G8TRF-125:E Twin : 30 nm 8 Gb DDR3 3 D stacked DRAM is obtained from Micron's "EOL of Micron's 4Gb DDR3 30nm product: PCN 31722"

APPENDIX

E

HOW TO RUN DATE

E.1 Read-me First

DATE was developed for python 2.x version. DATE has three sub-directories in the main directory.

- ckts: DATE Circuit level model files.
- process: Transistor, Wire, TSV, and default voltage roadmap files in 'comma-separated values' (csv) format.
- chipset: The directory contains executables for various DRAMs introduced for verifi-

cation in Section 3.3

To run the DATE, run the following from the chipset directory. For example, to run Samsung 80nm, 800MHz DDR2 DRAM is:

```
> python Samsung_80nm_512M.py
```

E.2 Executable file with comments

Listing E.1 80nm Samsung DDR2 DRAM executable file

```
"""
    jpark17@ncsu.edu

    <<Samsung 2D 80nm 512Mb configuration>>
    K4T51083QE-ZCE7
    Bank Addr   : 2 bit
    Row Addr    : 14 bit
    Col Addr    : 10 bit
    Data bit    : 8 bit
    Burst       : 4
    # die       : 1

"""
import time
import sys
sys.path.append( '../ ckts ' )

from MemDie import
from GlobalVariables import

start_time = time.time()

#Name -----
DRAM_NAME = "Samsung 2D DDR2 512Mb in 80nm, K4T51083QE-ZCE7, 800MHz"
file_name = "Samsung_80nm_2D_DDR3_512Mb_K4T51083QE.txt" #output file name
VENDOR = "Samsung" # Samsung\Micron\Elpida\Hynix
```

```

IO_SPEC = "DDR2"      # DDR\DDR2\DDR3\LPDDR\LPDDR2\LPDDR3

#General Option -----
kelvin = 300           # -operating temperature (K)
tech_node = 80         # -technology , nm 90-16
sys_freq_MHz = 400     # -system frequency (MHz), half of I/O speed
# - LPDDRx vdd1 and vdd2 are not the same and Vdd1 > Vdd2.
# - DDRx case Vdd2 has to be 0.
# In DDRx case if you put 0 at Vdd1 than it will go with default Vdd
# If you put 0 at Vdd1, than DATE follows default Vdd according to model
Vdd1 = 1.8
Vdd2 = 0
Vdd = [Vdd1, Vdd2] #don't touch it.

#3D Option -----
num_die = 1           # - die count , 1 == planar DRAM

#Data Length -----
num_BurstLength = 4    # -burst depth, 4,8,16
bit_data_size = 8      # -IO width , single Wordline size (between 4,8,16 bit)

#Total memory capacity -----
num_of_bank = 4
bit_bank_addr = log(num_of_bank)/log(2) # bank address bit count
bit_row_addr = 14      # single bank row addr
bit_col_addr = 10      # single bank col addr
bit_assist_row_addr = 2 # WL control signal. 2 is default don't change it.
# Size of the subarray (inside a Bank) option -----
NdWL = 16              # -Ndwl : The number which divides the bank by WL direction.
NdBL = 52              # -Ndbl : The number which divides the bank by BL direction.
# Size and configuration of bank.
cellLayout = "6F2"     # 4F2\6F2\8F2
rowDecLayout = "Side"  # MML decoder position.
# Side - side of bank : MML cell
# Center - center of bank : 1/2 cell MML 1/2 cell
# Fold - Similar to Center : MML 1/2 cell MML 1/2 cell
# Support combination. -----
# in 8F2, BL is too long, fold (cut BL 1/2) is default,

```

```

# 8F2, Center — R & C are not changed, col-Dec x 2 (area, energy)#
# 8F2, Side — WL → WL x 2 : R & C are changed , col-Dec x 2 (area, energy)
#
# 6F2, Center — bank size is not adjusted. Put MML Dec at center of the WL. →
#
#           MML length 1/2 , speed reduce, but evaluated MML dec = 2x.
# 6F2, Fold — bank size is adjusted (cut BL 1/2). WL R & C is not changed
# 6F2, Side — bank size is not adjusted, R & C are not changed
# 4F2, Side — default. not adjustable.

is_ecc = False                # -Add ECC bits inside subarray.

# Big layout (Bank layout) option -----
num_bank_BL = 2               #number of bank in BL direction inside die
num_bank_WL = 2               #number of bank in WL direction inside die

#wire -----
wireRepeaterOpt = REPEATER_30  # -Wire signalling
# REPEATER_NONE : no repeater for bus
# REPEATER_OPT  : optimum repeater spacing and sizing
# REPEATER_5    : sacrifice 5% bus speed by adjusting repeater size and spacing.
# REPEATER_XX   : XX → 10,15,20,25,30,35,40,45
#
#           : sacrifice XX% bus speed by adjusting repeater size and spacing.
isLowSwing = False            # not support future option.

#TSV -----
# TSV would not include when num_die is 1.
TSV_communication_bits = 6     # default additional TSV bit.
TSV_layout_type = TSV_COARSE_RANK_LEVEL
# TSV position
# TSV_COARSE_RANK_LEVEL : support, coarse-grained rank level.
# TSV_FINE_RANK_LEVEL   : not support
# TSV_COARSE_BANK_LEVEL : not support
# TSV_FINE_BANK_LEVEL   : not support
TSV_kind = TSV_INDUCONS
# TSV roadmap
# TSV_ITRS_AGGR : ITRS aggressive : not support
# TSV_ITRS_CONS : ITRS conservative : not support

```



```

# TSV_INDU_AGGR : CACTI based, aggressive : not support
# TSV_INDU_CONS : CACTI based, conservative

#local RW buffer size -----
# array - subarray
# half_array - half of subarray
# q_array - quarter of subarray
# data - burst length
# none
RD_BUF = "q_array"

#local WT buffer size -----
# array - subarray
# half_array - half of subarray
# q_array - quarter of subarray
# data - burst length
# none
WT_BUF = "q_array"

INPUTS = [ num_die, bit_bank_addr, bit_row_addr, bit_assist_row_addr, \
    bit_col_addr, num_BurstLength, bit_data_size, \
    cellLayout, NdWL, NdBL, wireRepeaterOpt, is_ecc, \
    kelvin, tech_node, num_bank_BL, num_bank_WL, \
    TSV_communication_bits, TSV_layout_type, TSV_kind, isLowSwing, \
    sys_freq_MHz, Vdd, DRAM_NAME, VENDOR, IO_SPEC, rowDecLayout, \
    RD_BUF, WT_BUF, file_name ]

DRAM = MemDies()
DRAM.Initialize(INPUTS)

print "time", time.time() - start_time

```