1994

BiCMOS technology and some applications in high performance arithmetic structures.

James Christopher. Czilli
University of Windsor

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Canada

BiCMOS Technology and Some Applications in High Performance Arithmetic Structures

by

James Christopher Czilli

A Thesis
Submitted to the Faculty of Graduate Studies through the
Department of Electrical Engineering in Partial Fulfillment
of the Requirements for the Degree of
Master of Applied Science
at the
University of Windsor

Windsor, Ontario
February, 1994.

ISBN 0-315-93262-7

Name JAMES C. CZILLI

Dissertation Abstracts International is arranged by broad, general subject categories. Please select the one subject which most nearly describes the content of your dissertation. Enter the corresponding four-digit code in the spaces provided.

ELECTRONICS AND ELECTRICAL ENGINEERING — SUBJECT TERM

[ ][ ][ ][ ] — SUBJECT CODE

U·M·I

Subject Categories

THE HUMANITIES AND SOCIAL SCIENCES

COMMUNICATIONS AND THE ARTS

Architecture 0729
Art History 0377
Cinema 0900
Dance 0378
Fine Arts 0357
Information Science 0723
Journalism 0391
Library Science 0399
Mass Communications 0708
Music 0413
Speech Communication 0459
Theater 0465

EDUCATION

General 0515
Administration 0514
Adult and Continuing 0516
Agricultural 0517
Art 0273
Bilingual and Multicultural 0282
Business 0688
Community College 0275
Curriculum and Instruction 0727
Early Childhood 0518
Elementary 0524
Finance 0277
Guidance and Counseling 0519
Health 0680
Higher 0745
History of 0520
Home Economics 0278
Industrial 0521
Language and Literature 0279
Mathematics 0280
Music 0522
Philosophy of 0998
Physical 0523
Psychology 0525
Reading 0535
Religious 0527
Sciences 0714
Secondary 0533
Social Sciences 0534
Sociology of 0340
Special 0529
Teacher Training 0530
Technology 0710
Tests and Measurements 0288
Vocational 0747

LANGUAGE, LITERATURE AND LINGUISTICS

Language
- General 0679
- Ancient 0289
- Linguistics 0290
- Modern 0291

Literature
- General 0401
- Classical 0294
- Comparative 0295
- Medieval 0297
- Modern 0298
- African 0316
- American 0591
- Asian 0305
- Canadian (English) 0352
- Canadian (French) 0355
- English 0593
- Germanic 0311
- Latin American 0312
- Middle Eastern 0315
- Romance 0313
- Slavic and East European 0314

PHILOSOPHY, RELIGION AND THEOLOGY

Philosophy 0422
Religion
- General 0318
- Biblical Studies 0321
- Clergy 0319
- History of 0320
- Philosophy of 0322

Theology 0469

SOCIAL SCIENCES

American Studies 0323
Anthropology
- Archaeology 0324
- Cultural 0326
- Physical 0327

Business Administration
- General 0310
- Accounting 0272
- Banking 0770
- Management 0454
- Marketing 0338

Canadian Studies 0385
Economics
- General 0501
- Agricultural 0503
- Commerce-Business 0505
- Finance 0508
- History 0509
- Labor 0510
- Theory 0511

Folklore 0358
Geography 0366
Gerontology 0351
History
- General 0578
- Ancient 0579
- Medieval 0581
- Modern 0582
- Black 0328
- African 0331
- Asia, Australia and Oceania 0332
- Canadian 0334
- European 0335
- Latin American 0336
- Middle Eastern 0333
- United States 0337

History of Science 0585
Law 0398
Political Science
- General 0615
- International Law and Relations 0616
- Public Administration 0617

Recreation 0814
Social Work 0452
Sociology
- General 0626
- Criminology and Penology 0627
- Demography 0938
- Ethnic and Racial Studies 0631
- Individual and Family Studies 0628
- Industrial and Labor Relations 0629
- Public and Social Welfare 0630
- Social Structure and Development 0700
- Theory and Methods 0344

Transportation 0709
Urban and Regional Planning 0999
Women's Studies 0453

THE SCIENCES AND ENGINEERING

BIOLOGICAL SCIENCES

Agriculture
- General 0473
- Agronomy 0285
- Animal Culture and Nutrition 0475
- Animal Pathology 0476
- Food Science and Technology 0359
- Forestry and Wildlife 0478
- Plant Culture 0479
- Plant Pathology 0480
- Plant Physiology 0817
- Range Management 0777
- Wood Technology 0746

Biology
- General 0306
- Anatomy 0287
- Biostatistics 0308
- Botany 0309
- Cell 0379
- Ecology 0329
- Entomology 0353
- Genetics 0369
- Limnology 0793
- Microbiology 0410
- Molecular 0307
- Neuroscience 0317
- Oceanography 0416
- Physiology 0433
- Radiation 0821
- Veterinary Science 0778
- Zoology 0472

Biophysics
- General 0786
- Medical 0760

EARTH SCIENCES

Biogeochemistry 0425
Geochemistry 0996
Geodesy 0370
Geology 0372
Geophysics 0373
Hydrology 0388
Mineralogy 0411
Paleobotany 0345
Paleoecology 0426
Paleontology 0418
Paleozoology 0985
Palynology 0427
Physical Geography 0368
Physical Oceanography 0415

HEALTH AND ENVIRONMENTAL SCIENCES

Environmental Sciences 0768
Health Sciences
- General 0566
- Audiology 0300
- Chemotherapy 0992
- Dentistry 0567
- Education 0350
- Hospital Management 0769
- Human Development 0758
- Immunology 0982
- Medicine and Surgery 0564
- Mental Health 0347
- Nursing 0569
- Nutrition 0570
- Obstetrics and Gynecology 0380
- Occupational Health and Therapy 0354
- Ophthalmology 0381
- Pathology 0571
- Pharmacology 0419
- Pharmacy 0572
- Physical Therapy 0382
- Public Health 0573
- Radiology 0574
- Recreation 0575
- Speech Pathology 0460
- Toxicology 0383

Home Economics 0386

PHYSICAL SCIENCES

Pure Sciences

Chemistry
- General 0485
- Agricultural 0749
- Analytical 0486
- Biochemistry 0487
- Inorganic 0488
- Nuclear 0738
- Organic 0490
- Pharmaceutical 0491
- Physical 0494
- Polymer 0495
- Radiation 0754

Mathematics 0405
Physics
- General 0605
- Acoustics 0986
- Astronomy and Astrophysics 0606
- Atmospheric Science 0608
- Atomic 0748
- Electronics and Electricity 0607
- Elementary Particles and High Energy 0798
- Fluid and Plasma 0759
- Molecular 0609
- Nuclear 0610
- Optics 0752
- Radiation 0756
- Solid State 0611

Statistics 0463

Applied Sciences

Applied Mechanics 0346
Computer Science 0984
Engineering
- General 0537
- Aerospace 0538
- Agricultural 0539
- Automotive 0540
- Biomedical 0541
- Chemical 0542
- Civil 0543
- Electronics and Electrical 0544
- Heat and Thermodynamics 0348
- Hydraulic 0545
- Industrial 0546
- Marine 0547
- Materials Science 0794
- Mechanical 0548
- Metallurgy 0743
- Mining 0551
- Nuclear 0552
- Packaging 0549
- Petroleum 0765
- Sanitary and Municipal 0554
- System Science 0790

Geotechnology 0428
Operations Research 0796
Plastics Technology 0795
Textile Technology 0994

PSYCHOLOGY

General 0621
Behavioral 0384
Clinical 0622
Developmental 0620
Experimental 0623
Industrial 0624
Personality 0625
Physiological 0989
Psychobiology 0349
Psychometrics 0632
Social 0451

Abstract

This work provides a treatment of BiCMOS technology from several perspectives. The manner in which the modern BiCMOS process evolved from a predominantly CMOS processing base is discussed, and a survey of special processing technologies is given. Without these advanced techniques the current level of BiCMOS and CMOS process performance would be unattainable. Examples of such include trench isolation, lightly doped drain structures, and advanced metallization techniques. Additionally, BATMOS, Northern Telecom's BiCMOS process technology is described in detail. Issues pertaining to BiCMOS process scalability, device scalability, and process tradeoffs are discussed. Included are such topics as active device scaling, second order effects which become important in scaled technologies, BiCMOS process tradeoffs, and latchup in BiCMOS. Several high performance arithmetic architectures were implemented in the form of macrocells, and their design is discussed. Toward that end, a survey of hardware multipliers is given, concentrating on the parallel types, and two new recently proposed architectures are described which form the basis for five of the six macrocells. As well, the implementation of a fast adder macrocell is treated. Some details of the design process of these macrocells are also highlighted. The use of BiCMOS in the realization of dynamic, massively pipelined arithmetic structures is explored. A review of pipelining strategies is briefly given, followed by a description of the true single phase clocking (TSPC) technique. Factors effecting the implementation of NMOS transistor switching trees in BATMOS is investigated, and key simulation results are reported. A new latching principle is described which is based on a current steering concept. Several new latch structures based on this principle are described and simulation results from test structures are presented. The performance of these new structures is compared to the TSPC technique, and simulation results of both latching arrangements used with a synthesized switching tree based mod 7 multiplier are presented. Finally, an ultra fast latching structure is introduced, simulation results are presented, and some additional discussion is given.

Dedicated with love
to my family.

Acknowledgments

The support and guidance offered by G. A. Jullien was invaluable in the successful completion of this thesis. His knowledge and enthusiasm provided a source of inspiration, and many of the ideas contained in this work were precipitated from informative discussions with him. W. C. Miller also gave insightful comments and observations which provided important direction for this work. S. Bandyopadhyay is recognized as having served a valuable supervisory role on my committee. Always willing to offer his time, Zhongde Wang provided extensive input in the area of hardware multipliers. Additionally, my family has been a constant source of strength, support, and encouragement which has allowed me to persevere through the difficult times. Finally, my friends are acknowledged for their understanding and patience. Micronet provided funding for a portion of this work.

Table of Contents

List of Figures

List of Symbols

ß	effective current gain of bipolar transistor
β_f	peak common emitter current gain
ΔV	voltage change
ε_{ox}	relative permittivity of silicon dioxide
ε_s	relative permittivity of silicon
ε_o	permittivity of free space
ϕ_s	surface potential
κ	figure of merit for MOS devices
µ	mobility
μ_n	electron mobility
ρ_i	number of partial product rows per stage
v	velocity
v_∞	saturation velocity
BV_{CBO}	collector—base breakdown voltage
BV_{CEO}	collector—emitter breakdown voltage
C_L	load capacitance
C_{gox}	gate oxide capacitance per unit area
C_g	gate capacitance
C_{inv}	inversion layer capacitance
C_D	depletion capacitance
D_g	latch delay
D_{nb}	diffusion constant for electrons in base
D_{pe}	diffusion constant for holes in emitter
E	electric field
E_{br}	avalanche critical field
f_T	cut-off frequency
g_m	transconductance
in_{ij}	inputs to a counter
I_C	collector current
I_{DS}	drain current
I_{avg}	average current

K	number of stages in partial product reduction array
k	scaling factor
k_1	scaling factor
k_2	scaling factor
k_h	horizontal scale factor
k_l	vertical scale factor
k_u	voltage scale factor
k_B	Boltzman constant
L_{pe}	diffusion length of holes in emitter
L	MOSFET channel length
N_C	collector doping concentration
N_E	emitter doping concentration
N_{SUB}	substrate doping
N_B	peak base doping
q	charge
Q_n	inversion charge
Q_B	base Gummel number
R_E	emitter resistance
r_o	output resistance
R_C	collector resistance
R_B	base resistance
S	subthreshold swing
T	temperature
T_D	circuit delay
t_{ox}	gate oxide thickness
t_h	hold time
t_{su}	setup time
V_{CE}	voltage between collector and emitter
V_{DD}	+ supply voltage
V_{DS}	drain to source voltage
V_{GS}	gate voltage with respect to the source
V_{pt}	punchthrough voltage
V_{rt}	reachthrough voltage
V_{SS}	- supply voltage
V_T	MOS threshold voltage
V_{TN}	NMOS threshold voltage

V_{TP}	PMOS threshold voltage
W_C	collector width
W_B	base width
W	MOSFET channel width
W_E	emitter width
X_j	junction depth

Chapter 1

INTRODUCTION

1.1 Introduction

BiCMOS technology constitutes a marriage of CMOS and bipolar technologies. Both types of devices are available on the same wafer, which means that the advantages of both may be reaped at the expense of extra process complexity. CMOS devices offer low power dissipation, wide noise margins, and a very high packing density. Bipolar junction transistors (BJTs) offer high switching speed, high current density per unit area, and flexible I/O levels. Since the mid 1970s, CMOS has been the mainstream semiconductor technology chiefly because of its packing density, and power dissipation. Bipolar technologies were usually used for small but fast system subcircuits, and they had to be manufactured on separate substrates. In a BiCMOS process, both CMOS and bipolar devices are available on the same chip, thus affording the option of using both devices simultaneously in an integrated design.

BiCMOS fills a market gap between very dense, medium speed CMOS chips, and bipolar ECL integrated circuits with high power consumption. It no longer is a niche technology, however, since its utility has been recognized on a broad scale. It has been used in the implementation of gate arrays [1], static RAMs (SRAMs) [2], dynamic RAMs (DRAMs) [3], and complex subsystems such as a phase locked loop clock generator [4]. The Pentium® microprocessor from Intel, and the Super Sparc® RISC processor from Sun Micro-systems/Texas Instruments are both examples of microprocessors which have been implemented with BiCMOS technology [5]. Usually BiCMOS designs are predominantly CMOS based, and the bipolar devices are used only where their qualities are most useful. Common applications for the Bipolar devices include gates with high drive for high capacitance nodes [6], which exploit the device's high transconductance, and sense amps in memory chips [3], which exploit their superior analog qualities. Misuse of the bipolar devices can lead to unnecessarily large designs and compromised performance [7]. The

former is due to the additional space occupied by bipolar transistors while the latter is due to the fact that a CMOS gate can switch faster under light loads than a BiCMOS gate[1].

1.2 Thesis Objectives

This thesis work has three main objectives. The first is to provide a survey of current BiCMOS processing techniques and methodologies, as well as process considerations which determine device behaviour. The second involves the design and characterization of several high performance arithmetic macrocells[2]. The third objective is to investigate the application of BiCMOS technology in the realization of massively pipelined arithmetic structures.

1.2 Thesis Organization

This work is organized into six chapters. The first chapter comprises this introduction, while the second describes BiCMOS processing techniques and methodologies. The manner in which the modern BiCMOS process evolved from a predominantly CMOS processing base is discussed. Special processing technologies are described which are used to form structures without which the current level of BiCMOS and CMOS process performance would be unattainable. Examples include trench isolation, lightly doped drain (LDD) structures, and advanced metallization techniques. Finally, BATMOS, Northern Telecom's BiCMOS process technology, is described in detail.

The third chapter delves into issues pertaining to BiCMOS process scalability, device scalability, and process tradeoffs. A brief description of classical scaling of MOS devices is given, followed by a summary of other scaling methods for MOS and bipolar devices. Many second order effects become prominent performance limiting factors when feature sizes shrink, and some of these factors are discussed for both MOS and bipolar devices. Due to the presence of both device types in BiCMOS, many process tradeoffs must be made, and the main ones are discussed. Finally, a treatment of latchup is given, including a short review of latchup in CMOS, as well as a discussion of new modes which can only occur in a BiCMOS process.

---

[1] This statement is based on a conventional totem-pole style gate [8].

[2] This work was funded under a contract with the Canadian Microelectronics Corporation, and the Micronet Network of Centres of Excellence (with four member universities).

The fourth chapter deals with the design and characterization of several high performance arithmetic macrocells. A survey of hardware multipliers is given, concentrating on the parallel types. Booth encoding is briefly explained, and two new recently proposed architectures are described. The implementation of these architectures in the form of macrocells is discussed, as well as the implementation of a fast adder macrocell. Some details of the design process of these macrocells are also described.

Chapter five investigates the use of BiCMOS in the realization of pipelined arithmetic structures. A review of pipelining strategies is briefly given, followed by a description of the true single phase clocking (TSPC) technique. Factors effecting the implementation of NMOS transistor switching trees in BATMOS is investigated, and key simulation results are presented. A new latching principle is described which is based on a current steering concept. Several new structures based on this principle are described and simulation results are presented. The performance of these new structures is compared to the TSPC technique, and simulation results of both latching arrangements used with a synthesized switching tree based mod 7 multiplier are presented. Finally, an ultra fast latching structure is introduced, simulation results are presented, and some additional discussion is given.

The final chapter summarizes this work including pertinent results, and also offers suggestions for future work.

Chapter 2

PROCESS TECHNOLOGY

2.1 Introduction

BiCMOS is the name given to semiconductor fabrication technologies which are designed to include both CMOS and bipolar devices on the same die. Usually the process is optimized for performance, cost, or analog compatibility but, unfortunately, the paths to these ends usually diverge. A process which has been optimized for speed is usually characterized by small feature size (< 1μm in 1993), thin gate oxide, the presence of an epitaxial layer, 2 or 3 layers of low resistance interconnect and advanced isolation techniques. In cost optimized processes, a reduced number of mask levels, and no epitaxial layer yield a simpler fabrication process at the cost of compromised device performance. An analog process will have less aggressive design rules, a thick epitaxial layer and oxide layers, and may be designed for higher operating voltages (10V to 15V). It is important to realize that the development of BiCMOS technology has been driven from a CMOS processing base, which explains why low end BiCMOS processes may have, in total, only one or two additional mask levels when compared to the CMOS parent process. In this chapter, typical CMOS fabrication techniques and practices are examined briefly, and then successive process changes are described to illustrate the evolution of a high performance BiCMOS technology from the basic CMOS process flow. Advanced processing techniques will be covered, and BATMOS, Northern Telecom's BiCMOS process, will be discussed.

2.2 Silicon and Process Basics

The electrical conductivity of silicon is between that of an insulator and that of a conductor, which is why it is called a semiconductor. Through a process called doping, impurities are purposely introduced into the crystal lattice structure of the silicon which alter its electrical

behaviour. A typical substrate doping[1] concentration is approximately 10^{15} atoms/cm^3, while the particle density of silicon is on the order of 10^{22} atoms/cm^3. This means that the addition of only 1 dopant atom per 10^7 silicon atoms almost completely controls the electrical properties of the silicon. Impurities are classified as either donors or acceptors. Donors are penta-valent elements whose atomic structure results in a loosely bound electron when the atom is embedded in the silicon lattice. Acceptors, on the contrary, result in an empty electron position being created in the lattice. This space, or hole, behaves in many circumstances as a positively charged particle which is free to move through the lattice. Both electrons and holes are called charge carriers because their movement constitutes current through a semiconductor. For the above reasons, silicon which has been doped with an acceptor dopant such as boron, is called p type since the majority charge carriers (holes) possess a positive charge. Similarly, silicon doped with a donor impurity such as phosphorus is called n type since the charge carriers (electrons) have a negative charge. Integrated circuits are formed on substrates which are part of a larger wafer. These wafers are sliced from large, single crystal silicon cylinders. There are several fundamental processing techniques which are then applied to the wafers. These basic techniques are elemental to any silicon semiconductor fabrication process.

Semiconductor fabrication is comprised essentially of an iteration of several fundamental processes including oxidation, deposition, etching, photolithography, diffusion and ion implantation. The first three are used at different times in the process flow for different reasons, while the latter two are employed chiefly for the purpose of doping. Silicon oxide may be grown or deposited. Oxidation, or oxide growth is the process by which a layer of silicon dioxide (SiO_2) is formed on the surface of the wafer. It is important for device isolation as well as stress relief and dopant masking. When grown on the surface, it consumes silicon, and the oxide grows into, as well as out of the silicon substrate surface. Deposition is the means by which thin films of various materials are established on the surface of the wafer. There are many different forms of deposition, including chemical vapour deposition (CVD), vacuum deposition and sputtering. Oxide and nitride are usually deposited by CVD, and metal interconnect such as aluminum is usually sputtered. Etching is the process of selective removal of materials from the wafer, which may include portions of layers or films which have been deposited on the substrate, or some of the substrate itself. Wet etching is accomplished with liquid chemicals while dry etching, or plasma etching, is implemented with ionized gases. Photolithography is a process which renders

---

[1] This will vary widely depending on the substrate architecture.

geometrically specific depositions generated from design data which are used as masks for etching as well as other processes such as implantation. The chemical agent used is called a photoresist, due to its unique sensitivity to light and its resistance to etching agents. The photolithographic process, in the traditional sense, involves the light projection of design mask information onto a deposited layer of photoresist. The light chemically alters the resist, and the unaltered areas are removed with a solvent, leaving behind the required geometry. This process is called developing the photoresist. During the etching procedure, the areas underneath the resist are protected from the etching agent, or etchant. In recent years, the need for smaller and smaller line widths in the photolithographic process have prompted the development of techniques using lasers [9] and X rays [10] instead of normal light. Some recent techniques use a process called electron beam lithography (EBL) to generate chip geometries on the wafer with a high energy electron beam.

In order to introduce impurities into the silicon substrate two main techniques are used which can be classified as diffusion and implantation. Diffusion is a process which achieves impurity placement within the silicon crystal lattice by providing a concentration gradient of dopant atoms. Under high temperature, typically 800 °C to 1400 °C, the atoms diffuse into the silicon, creating the dopant profile for the region. Different profiles can be achieved by using a finite or infinite dopant source. In either case, temperature and time are the critical factors, and thus high temperature steps are usually confined to the front end of the fabrication sequence in order to limit uncontrolled diffusion after dopant profile establishment. Selective diffusion can be achieved by using silicon dioxide as a barrier. Wells are patterned and etched into a thick layer of oxide and the wafer is exposed to a dopant source, with the oxide blocking diffusion in unexposed areas. Ion implantation involves the acceleration of impurity atoms toward the substrate to be doped. The kinetic energy of the atoms is such that they are able to penetrate the crystal structure, thus achieving lattice placement. This technique allows implantation through a thin layer of oxide and very accurate doping profile control, both of which are very desirable traits. Disadvantages include the lattice damage resulting from the collision of dopant atoms with the crystal, and the sophisticated and expensive equipment required to implement the procedure. For the former reason, an implantation step is usually followed at some point by a high temperature anneal to repair the crystal lattice damage. A combination of implantation and diffusion may be used, where a shallow dopant implant is heated to allow the impurity to diffuse into the bulk. Many different improvements and variations to the

above basic practices have been implemented, but all semiconductor fabrication technologies still rely on these fundamental processes.

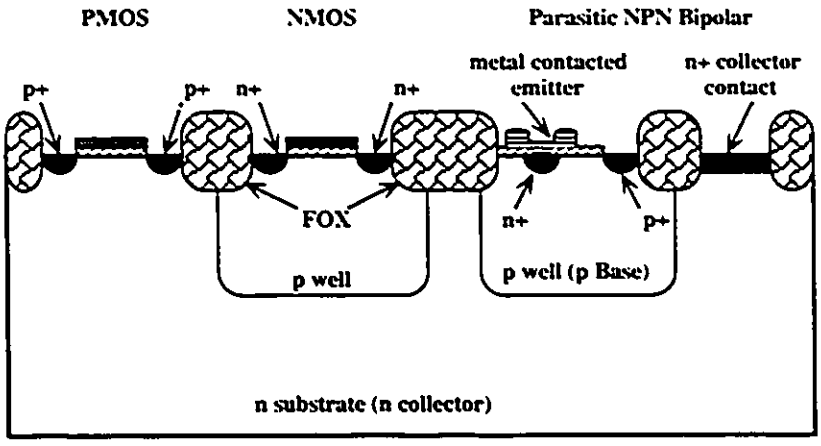
2.3 Typical CMOS Process Flow

As an example of a typical CMOS process flow, consider the CMOS3DLM process [11], which begins by patterning the p wells on the n substrate. These will be areas of p type silicon, and they will allow the creation of NMOS transistors. To form the wells, a layer of thick oxide is grown across the wafer, and a layer of photoresist is deposited. Using PWELL design mask information, the wafer is exposed, developed and etched to selectively remove oxide where p wells are to be implanted. A thin layer of oxide is grown over the exposed areas, and the p wells are implanted through this oxide. The thin and thick oxides are removed and a new layer of thin oxide is grown across the entire wafer. After this step, the resulting wells consist of areas of light p doping. Next, the NWELL mask layer information is used to define regions which will receive an n implant, or channel stopper, to set the threshold of non-active regions in the n substrate. This will assure that the parasitic device thresholds are such that there will not be inversion under the field oxide during normal operation of the chip. The PWELL mask is used to pattern a layer of photoresist over the p well areas to protect them from the channel stopper implant, which is carried out through the thin oxide covering the wafer. The photoresist is then removed, and a layer of silicon nitride (Si_3N_4) is deposited and patterned using the DEVWELL (device well) mask. Thus, regions which are to become parts of MOS transistors, diffusion interconnects, and capacitors are masked by the Si_3N_4. Photoresist is deposited and developed using the PGUARD mask, and a p implant into the p wells is performed next as a channel stopper. Since the device wells are still covered by nitride, they are protected from this implant, as are all regions outside of the areas defined by the PGUARD mask.

The photoresist is removed and a thick oxide is grown over the entire wafer, except in the areas covered by Si_3N_4. This oxide is called field oxide (FOX), and does not form in the device well regions since nitride inhibits oxide growth. The nitride is removed, and a thin gate oxide is grown over the device well areas. The NWELL mask is again used to pattern photoresist, and a p implant is used to adjust the threshold of the device wells in the n substrate which will later form PMOS transistors. This is necessary to counteract the previous n type channel stopper implant which was performed near the beginning of the process. If capacitors are to be created, the CAPDOP mask is used to define device wells

which will undergo a heavy p+ implantation through the gate oxide to form the bottom capacitor plates. At this stage, all device well regions are covered with gate oxide, and all other regions of the wafer are blanketed with FOX. A layer of polycrystalline silicon referred to as polysilicon, or simply poly, is deposited over the entire wafer, doped n+, and a layer of photoresist is then deposited. The POLY1 mask is used to develop the resist, thus defining MOSFET gates, interconnections, and capacitor top plates, and the wafer is then etched. The NPLUS mask is used to photolithographically define areas of n+ doping which will form the source/drain structures of the NMOS transistors, and n substrate contact areas. Similarly, the PPLUS mask is used to define the p+ regions corresponding to similar structures. The n+ poly involved in the source/drain doping step for the PMOS devices remains n+, despite the counter doping, due to their high n+ concentration.

Figure 2.1: P Well CMOS Process Cross-section



A thick layer of oxide is deposited over the entire wafer as an isolation layer on which the first level of metal interconnect will be patterned. The CONTACT mask is used to define contact windows which are etched away to allow the underlying layers to be contacted. A layer of aluminum is deposited across the entire wafer, and the METAL1 mask is used to define its geometry. This involves similar steps as before, i.e. photoresist development and etching. Another layer of SiO_2 is deposited over the wafer to isolate the metal interconnect. VIA mask information is used to etch contact windows through the oxide to the underlying metal, thus allowing contact between the two metal layers. Another layer of metal is deposited over the wafer, and patterned using the METAL2 mask. Bonding and pad contacts are made to this level of interconnect. Finally, a passivation layer is deposited across the entire wafer. The GLASS mask information is used to etch openings to the

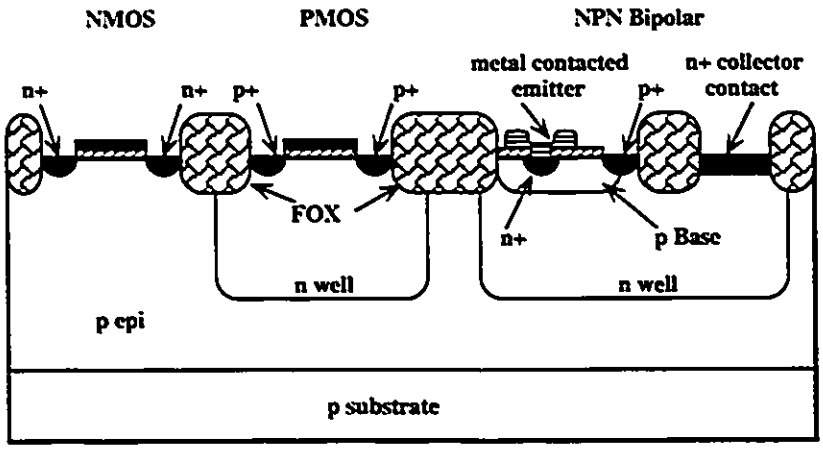
metal bonding pads, thus allowing connection by bonding wires. A cross-section of this process showing the important diffusions is shown in Figure 2.1.

2.4 Evolution of the BiCMOS Process Flow

Considering the previous process description, it can readily be seen that a bipolar device is present in parasitic form, as shown in Figure 2.1. These devices have been effectively used as optical sensors in neural network designs [12], even though they are not an intentionally fabricated device. In such an application, the p well in which NMOS transistors are normally formed serves as the base. The substrate acts as the collector and a source/drain diffusion constitutes the emitter. This is a very poor quality transistor, however, since it possesses a very large effective base width. The chief difference in a BiCMOS process is that the BJT is intentionally designed into the process flow, and thus there are added features which increase the performance of these devices.

Since the main limitation in the above example of a bipolar device is the base width, a low end BiCMOS process can be envisioned with the addition of only one additional mask layer to form a thin base. Consider a simple n well CMOS process with the conceptual cross-section illustrated in Figure 2.2.

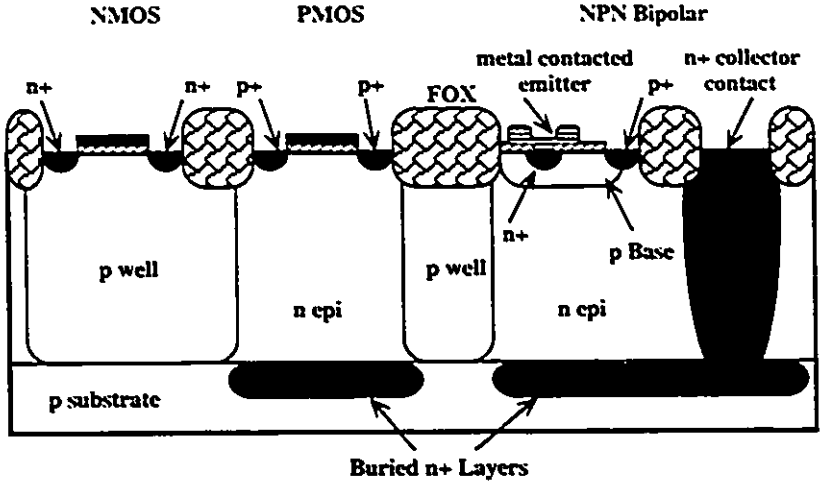
Figure 2.2: Low End BiCMOS Process



An NPN transistor has been formed with the n well acting as the collector, an n+ NMOS source/drain diffusion acting as the metal contacted emitter and a PMOS source/drain p+ diffusion acting as the base contact. With the addition of one extra mask level, a thin base has been formed by a single p implantation step, thus yielding a bipolar structure with

much higher performance than in the parasitic case. Typical base regions in early BiCMOS technologies were about 1μm deep with a doping of about 10^{17} atoms/cm^3. Even though this is a substantial improvement, the quality of the bipolar device has still been compromised in order to maintain common fabrication steps with its CMOS parent process. The main limitation inherent to this bipolar structure arises from the lightly doped n well, which has a correspondingly high resistance. This will cause the collector resistance to be high, which in turn will yield a device with low cutoff frequency, poor current drive, and a high collector-emitter saturation voltage. All these factors will limit the performance of the BJT.

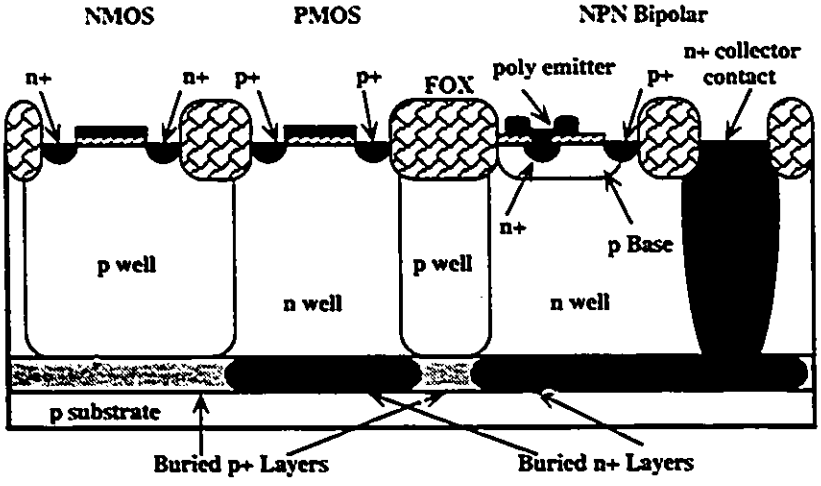
Figure 2.3: BiCMOS Process With Buried Layers and Deep Collector Contact



The collector resistance can be reduced in several ways, all of which are at the cost of increased process complexity. Consider Figure 2.3 which shows a conceptual cross-section of an improved BiCMOS process. Buried n+ layers have been implanted into the p substrate and an n doped epitaxial layer, or epi, has been grown on top of this (typically a few microns thick). Since the buried n+ layers and the n wells are aligned, an additional mask level is not required. An extra mask level has been added to define a deep n+ contact which traverses the depth of the epi in order to form a low resistance contact to the underlying buried layer. This contact can be implanted after FOX formation. In addition to reducing the collector resistance, the buried layers also reduce the susceptibility of the process to latchup by effectively reducing bulk resistance. There are, however, several drawbacks to this scenario, one of which is that the packing density of the devices is limited due to isolation considerations. Two adjacent n+ layers corresponding to two

separate bipolar devices must be separated by a large distance of lightly doped p substrate to avoid collector to collector punchthrough from one device to the other. This is due to the fact that since the substrate is lightly doped, and the buried layers are heavily doped, the depletion regions resulting from the pn junctions formed where the two regions meet will extend a large distance into the p substrate. Unfortunately, simply raising the substrate doping is not an optimal solution. Although this will allow closer device spacing, the increased doping will cause a corresponding increase in the collector—substrate capacitance, thus reducing bipolar performance. Another problem with the above bipolar structure is that the n epi region must be counter doped in order to isolate n well regions, and to form p wells for NMOS devices. Counter doping the n type epi layer causes processing difficulties as well as performance reduction in NMOS transistors due to mobility degradation. Since the n+ emitter and the source drain diffusion share the same process step, a low quality emitter is rendered which is contacted by metal. This limits device performance[1]. Even with these shortcomings, the bipolar devices resulting from this process are far superior to those of the process illustrated in Figure 2.2. The improvement has been realized by the addition of only two mask levels to the baseline CMOS process. The first one defines the thin p base implant, and the second defines the n+ deep collector implant.

Figure 2.4: BiCMOS Process With Self Aligned Buried Layers and Poly Emitter



Examining Figure 2.4, an improvement to the above process can be made by providing self aligned buried n+ and buried p+ layers. This effectively reduces the minimum collector to

[1] Emitter considerations will be discussed in chapter 3.

collector spacing, but increases the collector sidewall capacitance because an extra, highly doped, pn junction sidewall capacitance component is present. Since there is a buried p+ as well as a buried n+ layer, however, the substrate doping can be reduced, thus shrinking the net collector capacitance without compromise of packing density. A near intrinsic epi is grown, and self aligned n wells and p wells are defined. This eliminates the counter doping problem encountered with the n-type epi mentioned previously, and is what is referred to as a twin-tub process, since both well regions are doped individually.

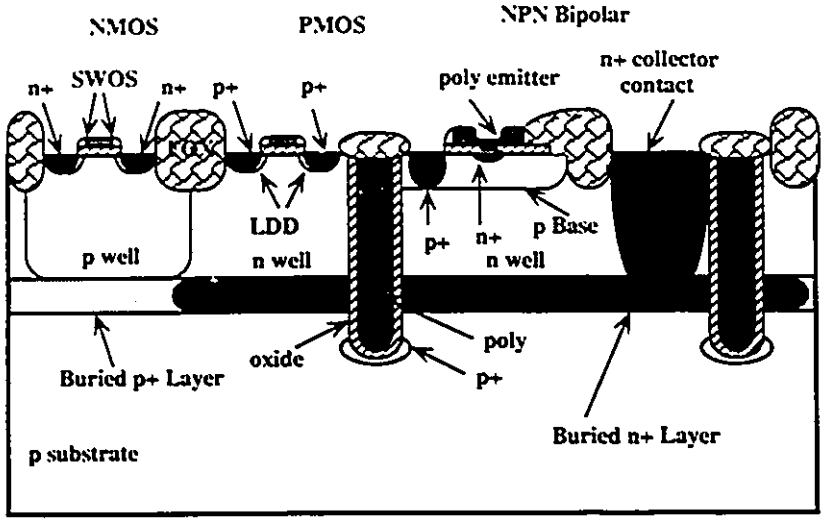
An additional mask level can be added by defining a special implant to form a shallow emitter which can be contacted with polysilicon. These types of emitters allow much higher device performance. For this process, which yields high quality MOS and bipolar devices, an additional 3 mask levels are necessary compared to the baseline CMOS process flow. There are no extra mask levels due to the buried layers since they are not only self aligned, but they are also aligned to their respective n and p wells. One additional mask level is required for the p implant to form the base, while the extrinsic base contact is provided by the shared process step of p+ source/drain island implantation. The second additional mask level defines the n+ deep collector implant, while the third additional mask level is necessary for emitter formation.

It should be stated that even though the number of masks required in addition to those of a baseline CMOS process are minimal, the actual process flow is substantially different from the first simple p well process described. Many of the process improvements, however, are already currently used in high performance CMOS, and thus both processes benefit.

A modern high end BiCMOS process cross-section is shown in Figure 2.5. Notice the many additional features contributing to device performance enhancement. MOS gates are silicided, providing much reduced parasitic resistances. Sidewall oxide spacer (SWOS) technology allows the formation of lightly doped drains (LDDs) in MOS devices, and self aligned intrinsic bases in bipolar devices. A p+ extrinsic base, which is self aligned to the sidewall oxide spacer on the edge of the poly emitter, provides a smaller base area, and thus a lower collector—base capacitance. Bipolar trench isolation increases the packing density of bipolar devices and, as well, reduces the collector sidewall capacitance; this, in turn, increases bipolar circuit performance. Local interconnect (LI) technology provides efficient interconnections between poly gates, poly emitters and diffusions. There are no contact cuts required for these connections, and thus they allow for very compact layouts. Tungsten metallization is used to plug submicron contacts and vias to alleviate metal step

coverage and electromigration problems. A thin epitaxial layer reduces the parasitic sidewall collector capacitance of bipolar devices, thus increasing performance.

Figure 2.5: High Performance BiCMOS process Cross-section



2.5 Advanced Processing Techniques

As was mentioned previously, many of the advances in CMOS processing technology over the last several years are performance enhancing in a BiCMOS process as well. These advances include such things as shrunken feature size, buried layers, epitaxial layer, silicidation, as well as improved planarization and metallization techniques. The performance of the integrated circuits of today would be impossible to attain without these advancements in fabrication technology and methodology. These techniques are still based fundamentally on the basic processing steps described in the early portion of this chapter; however, they warrant special attention due to their importance in modern BiCMOS fabrication technologies. Some of the major processing techniques are discussed below.

2.5.1 Epitaxy

Epitaxy is a process by which a layer of single crystal silicon, called an epitaxial layer, or epi, is grown on a single crystal silicon substrate [13]. The new layer continues the crystal lattice of the substrate, and its formation is achieved by using the substrate as a crystal seed, with new material precipitating out of a reaction with a gaseous compound at temperatures far below the melting point of silicon. Epi layers allow the formation of

buried layer structures as well as individual well doping profile optimization. They are used in virtually all high performance CMOS and BiCMOS processes of today. Epitaxial layer thickness is a crucial parameter in determining the performance level of the technology, since both CMOS and bipolar device behaviour is strongly influenced by it. A very high performance BiCMOS technology might have an epitaxial thickness of less than 1μm [14].

2.5.2 Buried Layers

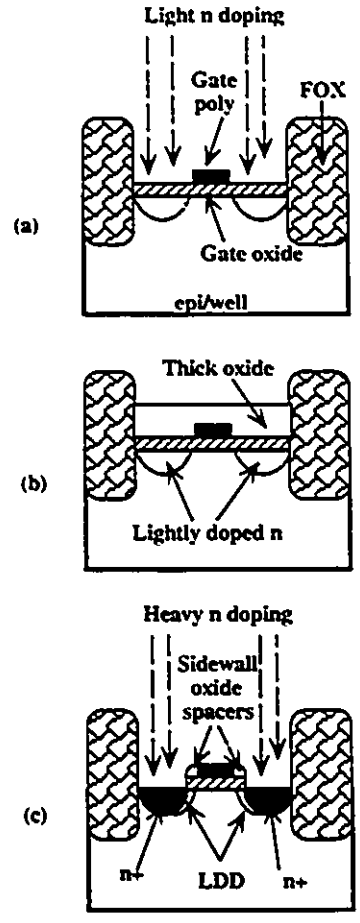
Highly doped buried layers are formed in the very early stages of wafer fabrication by implanting high impurity concentrations in areas of the substrate before the growth of the epitaxial layer. They provide many desirable process features including reduced latchup susceptibility through reduced bulk resistance, low resistance collectors, and increased packing density of bipolar devices. A typical doping level for buried layers would be on the order of 10^{19} *atoms/cm*^{3} [15] and a thickness of less than one, to several microns. A self aligned process with both types of buried layers is desirable, not only from the standpoint of reduced mask count, but also due to physical limitations. For example, in a process with only buried n+ layers, and an n epitaxial layer, counter doped wells will suffer from mobility degradation which will cause NMOS device performance to suffer. In a self aligned process, with a near intrinsic epi, optimization of individual well doping profiles is possible, which results in much higher device performance. Packing density is also increased due to reasons mentioned in section 2.4. The limiting factor in buried layer spacing, or well to well spacing, is usually the degree of lateral diffusion between the heavily doped regions. This is aggravated by high temperature processing steps.

2.5.3 Lightly Doped Drain (LDD)

Lightly doped drains are essential as feature sizes shrink into the submicron range [16]. These structures are necessary in order to control hot carrier effects which cause reliability problems by degrading device threshold voltages and transconductance over time [17]. Sidewall oxide spacer (SWOS) technology facilitates the formation of a lightly doped region, just in front of the drain diffusion. Thus, the drain depletion region extends further into the effective drain area, which reduces the local electric field intensity. This, in turn, reduces the probability of hot carrier generation, and increases the punchthrough voltage.

The process used to form an LDD structure in an NMOS transistor is depicted in Figure 2.6. After the gate oxide has been grown, and the gate poly has been deposited, patterned, and doped, a light doping is implanted through the gate oxide into the source/drain areas, as illustrated in Figure 2.6a. A thick oxide is then deposited by CVD, as illustrated in Figure 2.6b, and an anisotropic[1] etch is performed which removes most of the oxide, but leaves sidewall oxide spacers (SWOSs) on each side of the gate poly. These are the structures which make LDD formation possible. A heavy n+ implant is performed, with a small area being protected from the dopant by the remaining thick oxide, rendering a lightly doped drain region directly under the spacer. This is shown in Figure 2.6c. Note that the lightly doped region is present at both the source and drain, but it only serves a useful purpose at the drain end.

Figure 2.6: Steps in Forming LDD Structures



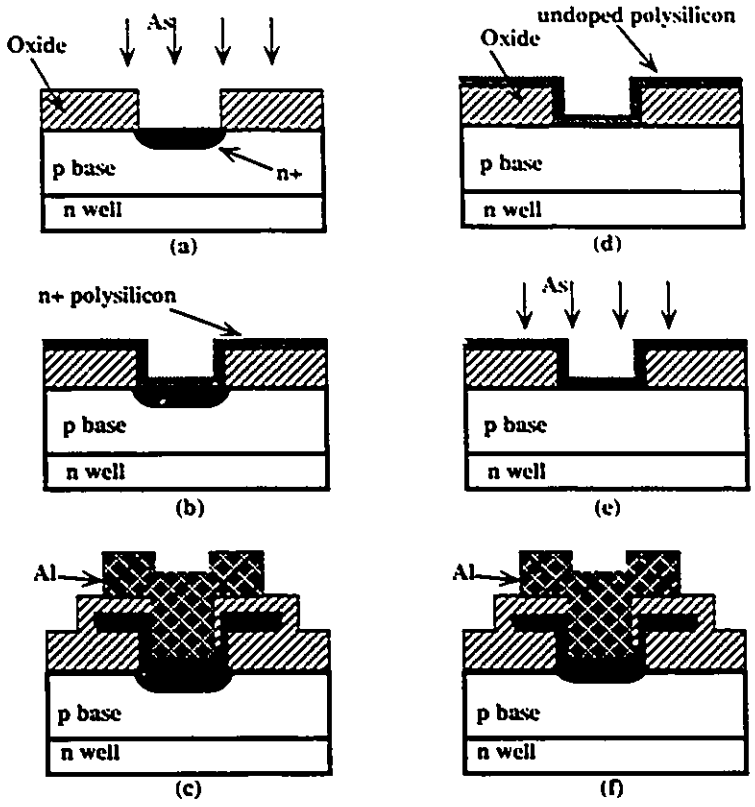
2.5.4 Polysilicon Emitters

Most high performance BiCMOS technologies utilize polysilicon to form the emitter. There are two main processes used for poly emitter formation. A distinction will be made here between a poly contacted emitter, and a poly emitter. The former uses an implant step to form the emitter region directly, while the latter relies on diffusion. Major processing steps for poly contacted emitter formation are illustrated in Figure 2.7 (a)-(c), while steps for poly emitter formation are shown in Figure 2.7 (d)-(f). The former begins with a masked, heavy implant of arsenic which forms the emitter region, as shown in (a). Polysilicon is deposited, (b), and patterned, followed by deposition of isolation oxide and a layer of metal, (c). Thus, the emitter region is contacted by polysilicon, which in turn is contacted by metal. Poly emitters generally produce shallower junctions. A layer of polysilicon is deposited, (d) and implanted with arsenic to dope it n+, (e). The wafer is subjected to a thermal step which allows the dopant to diffuse into the poly,

[1] An anisotropic etch provides a ratio for vertical to horizontal etching action which is less than one. See [18], pg. 736.

and into the emitter region. Thus, in this process, the poly is the source of diffusion dopant. Steps then proceed as for the other type, with the end result illustrated in (f). Polysilicon emitters will be discussed further in chapter 3.

Figure 2.7: Poly Emitter and Poly Contacted Emitter Formation



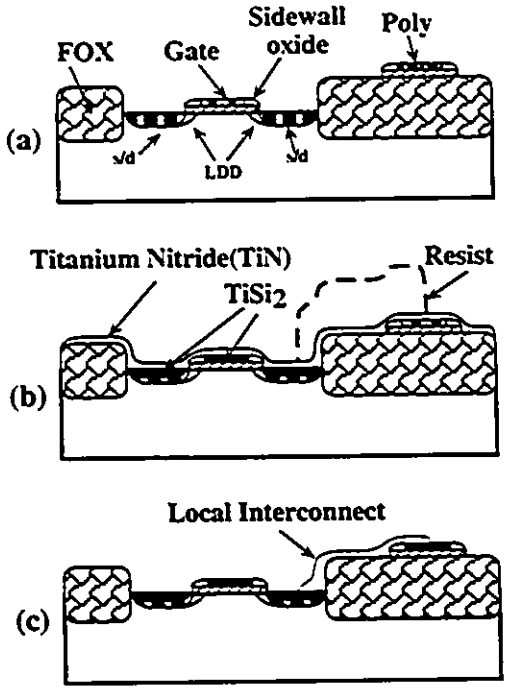
2.5.5 Silicidation

Silicidation is a process which reduces parasitic resistance of gates, poly emitters, poly interconnect, and diffusion regions. For example, doped polysilicon typically has a resistance of approximately, 30 Ω/square[1] but it can be reduced to below 5 Ω/square by using silicidation[2]. Essentially, this process relies on the chemical combination of silicon with a refractory metal such as molybdenum (Mo), tantalum (Ta), or titanium (Ti). The latter is by far the most popular, and when combined with silicon, it forms titanium disilicide ($TiSi_2$), which is highly conductive. The silicon is effectively clad with the metal, and its resistance is reduced because of the chemical combination. The process

[1] From details of the CMOS4S 1.2μ process [11].
[2] In the BATMOS process, silicidation is used, and the resistance of gate poly is quoted as 4 Ω/square [19].

proceeds as follows. Titanium is deposited across the wafer, and a silicon—titanium reaction is carried out in a nitrogen ambient. A reaction takes place between all exposed silicon, including polysilicon and titanium, thus lowering the sheet resistance of these areas. This reaction must be carefully controlled to prevent excessive reaction which can consume entire diffusion regions. Additionally, a layer of titanium nitride (TiN), which is also highly conductive, is formed over the wafer, due to the combination of titanium with gaseous nitrogen. When the reaction is complete, the TiN is removed, along with any unreacted Ti. The SWOSs prevent shorts from developing from source/drain diffusions to gates, and are crucial for reliability reasons. The Ti which has chemically combined with the silicon reduces the resistance of these areas.

Figure 2.8: Process Steps In Forming LI

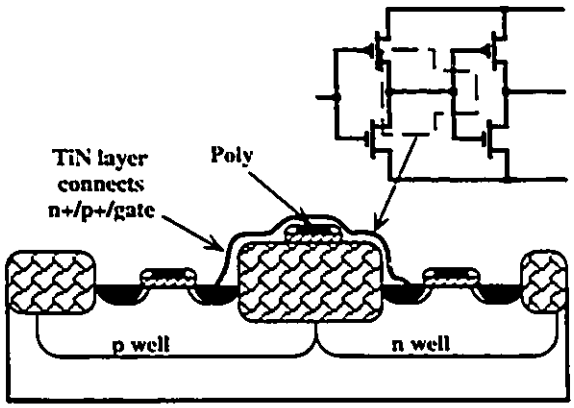


2.5.6 Local Interconnect (LI)

Local interconnect technology uses a byproduct of the above silicidation process. Since the reaction between the Ti and the nitrogen atmosphere in which the reaction takes place produces a conductive layer of TiN, it can be used as an interconnect layer. A separate mask level can be used to etch the TiN after the silicide reaction. LI formation is illustrated in Figure 2.8.

The LI formation process is essentially an extension of the silicidation process. It starts with a structure as shown in Figure 2.8a. After the TiN is deposited and the Ti-Si reaction is complete, photoresist is deposited across the wafer. LI mask information is used to develop the photoresist, resulting in the situation shown in Figure 2.8b. Here, the $TiSi_2$ has been formed on the poly and diffusions, and resist covers areas of TiN which are to be used as interconnect. The TiN is etched, and the wafer is annealed to cause the TiN to set. LI is very efficient in connecting regions of diffusion and polysilicon together. Sources and drains can connect to each other or to other gates and emitters without the need for contact cuts or vias. In addition, this interconnect does not interfere with the normal routing of the conventional metal layers.

Figure 2.9: Inverter Connections Implemented with LI



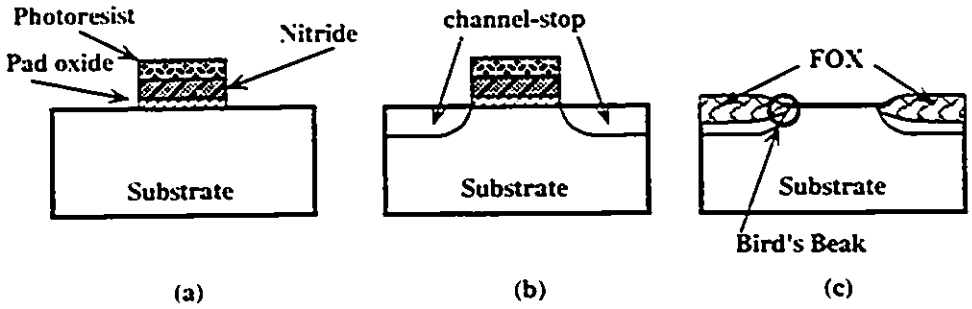
An example of the utility of LI is shown in Figure 2.9. This illustration shows a cross-section of the structures which would be present in a practical situation where LI could be used. Two inverters are connected together, with the output of the first feeding the input of the second. The connection implemented by LI is indicated by a dotted polygon on the inset schematic and involves two drain/source diffusions and a polysilicon layer. It can be seen that LI can considerably increase the packing density of layouts, especially those containing many instantiations of simple cells containing local connections similar to the one shown here. An obvious example is a static RAM chip.

2.5.7 Local Oxidation of Silicon (LOCOS)

Isolation in integrated circuits is primarily accomplished by reverse biasing pn junctions. For example, in the BiCMOS process shown in Figure 2.3, adjacent n wells are separated by heavily counter doped islands of p silicon to prevent punchthrough. These islands are tied to the lowest potential in the circuit, thus isolating adjacent n wells. The area of these regions is relatively large compared to the device size, and they create large collector parasitic capacitances. Another isolation concern is that MOS devices may be connected by

unwanted inverted conduction channels which form under the layers of interconnect. Local Oxidation of Silicon (LOCOS) is a broad class of isolation techniques which all rely on the selective formation of thick FOX.

Figure 2.10: Simple LOCOS Isolation Structure Formation

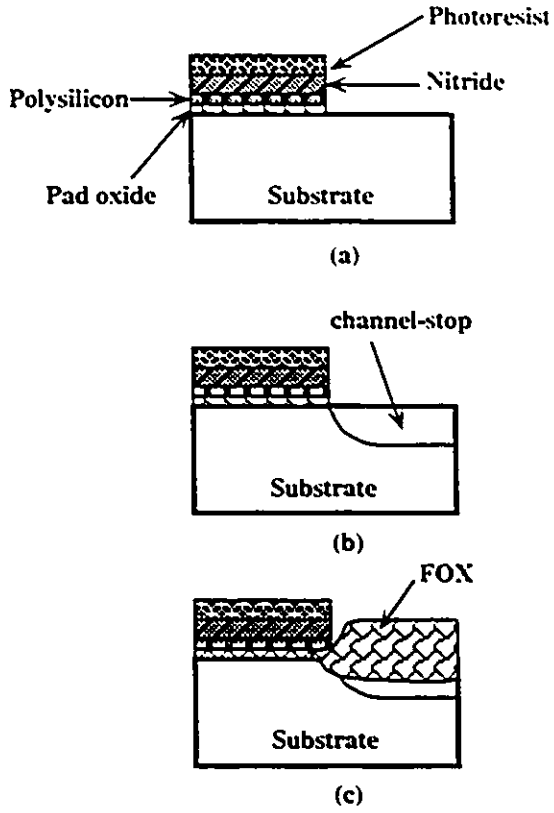


The basic steps in a LOCOS process are illustrated in Figure 2.10. A thin layer of pad oxide, about 30 nm or 50 nm thick [16], is grown over the wafer, followed by deposition of a thicker layer of silicon nitride. The pad oxide relieves the surface stress caused by the nitride in future high temperature processing steps, and thus it is called stress relief oxide (SRO). Photoresist is deposited and developed so that device well areas are protected from the next etching step. The unprotected areas are etched, removing the nitride and pad oxide layers, as illustrated in Figure 2.10a. A channel stopper implant is then performed affecting only areas not covered by nitride. This implant sets a high threshold voltage outside of active areas to eliminate the formation of parasitic channels. Figure 2.10b illustrates the situation immediately following this implant step. A thick layer of field oxide is then grown across the wafer. This oxide grows into as well as out of the silicon, with 54% of its thickness lying above the original surface [20]. The combination of the thickness of the oxide as well as a channel stopping implant inhibit channel formation under interconnections which are routed over the area. Figure 2.10c illustrates one of the problems with this process: a structure called a *bird's beak*. It is an oxide encroachment into the device well area under the nitride layer which introduces imprecise channel widths. This encroachment may be on the order of .5μm per side [16], which poses serious problems when scaling devices.

A poly buffer LOCOS (PBL) process, which minimizes bird's beak encroachment, is illustrated in Figure 2.11. In this LOCOS process, an additional layer of polysilicon is deposited on top of the pad oxide to provide extra stress relief, and a much thicker nitride layer is deposited. This is illustrated in Figure 2.11a. The channel stopper and oxidation

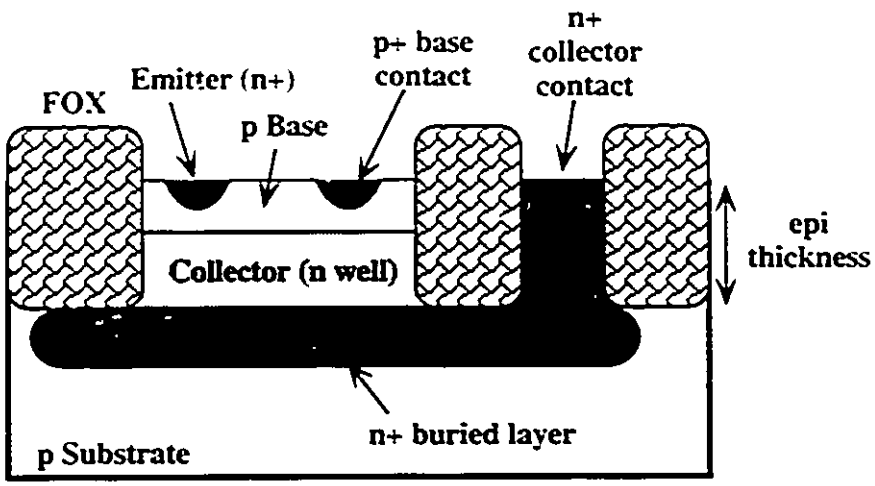
steps proceed as before, and are shown in (b) and (c) of the figure. Notice that the arrangement prevents the deep lateral extension of oxide under the nitride thus inhibiting *bird's beak* formation.

Figure 2.11: Poly Buffer LOCOS Isolation Structure Formation



LOCOS can also be used for bipolar isolation. Recessed LOCOS is particularly effective, but it involves some complicated processing steps due to the necessarily deep penetration of the FOX. Figure 2.12 shows recessed LOCOS isolation implemented for a bipolar device. Notice that the FOX extends down to the buried layer. Essentially, areas to be oxidized are recessed by plasma etching, the bottoms of the crevasses receive a channel stopper implant, and the FOX is grown. Layers of FOX are 2.2 times the thickness of the silicon that they consume. The islands of FOX grow laterally as well as vertically, and thus as thinner and thinner epitaxial layers are used to increase process performance, recessed LOCOS isolation for bipolar devices becomes more and more attractive because of increased packing density. Also, due to the absence of large depletion regions, the collector sidewall capacitance is reduced.

Figure 2.12: Example of Bipolar process with LOCOS Isolation

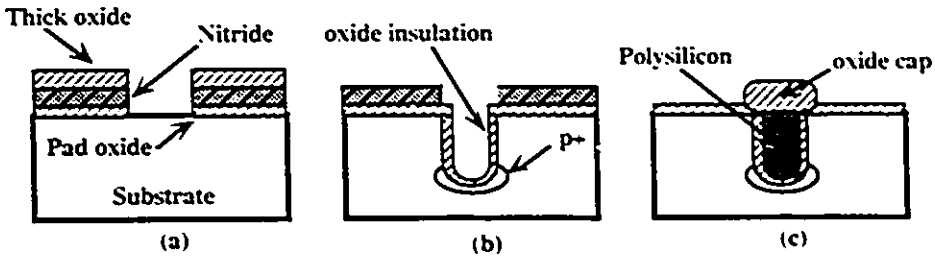


2.5.8 Trench Isolation

Trench isolation affords high packing density while, at the same time, it reduces sidewall collector capacitance. This isolation technique has no *bird's beak* encroachment, is latch up free, renders fairly planar surface relief, and is compact. Unfortunately, it also requires complicated processing steps to achieve the necessary structure which is essentially a deep trench reaching down into, or past, the buried layers. The general processing steps are illustrated in Figure 2.13. A layer of pad oxide is grown across the wafer to relieve surface stress caused by subsequent processing steps. A layer of nitride is deposited, as well as a layer of thick oxide which will act as a masking layer in future steps. Photoresist is deposited, developed, and the wafer is etched to expose areas which will become trenches. This is illustrated in Figure 2.13a. Reactive ion etching, a form of dry etching, is performed to create the trenches. A channel stopper implant is applied to the bottom of the trench to halt channel formation, and a layer of insulation oxide is grown over the inner walls of the trench. This is illustrated in Figure 2.13b. A layer of polysilicon is deposited over the entire wafer, which fills the trenches, and the surface of the wafer is etched so that only the poly filling the trenches remains. The wafer is oxidized using the nitride as a mask, which results in oxide caps forming over the trenches. Finally, the nitride is removed. Poly is used to fill the trenches because of its ability to flow into deep narrow holes. The finished structure is shown in Figure 2.13c. This method of isolation yields

very compact structures, however this is at a high cost due to the large number of extra processing steps which are required.

Figure 2.13: Process Steps For Trench Isolation Structure



2.5.9 Metallization and Planarization

As silicon chips with larger and larger dies are fabricated, and feature sizes are scaled, interconnect technology is challenged. Aluminum (Al) is by far the most common low resistance interconnect material used in modern integrated circuits. It has a low resistivity, good adherence to silicon and silicon dioxide, and it is easy to bond, pattern and deposit. There are several problems associated with its use, however, including contact failure, electromigration, and poor step coverage. Contact failure can result from either improper metal coverage at the time of fabrication, which may be linked to step coverage, or by contact electromigration which happens over time. The latter phenomenon occurs when silicon diffuses into aluminum at the interface of a silicon-aluminum contact, with resulting diffusion voids being filled with Al. This can result in a conductive spike forming which, for example, may short an NMOS n+ drain diffusion to the underlying substrate. Electromigration is the electric current induced transport of metal atoms along grain boundaries in metal traces. Large grain regions tend to minimize this effect, as do alloying agents such as copper which precipitate along grain boundaries and inhibit metal transport.

The thickness of many of the layers, such as field oxide and metal interconnect, are not usually scaled by the same factor as feature sizes are. This, combined with the presence of extra layers to implement such things as polysilicon emitters and passive devices such as resistors and capacitors, create exceptional surface relief which deposited interconnect layers must traverse. Step coverage problems of aluminum interconnections on high relief surfaces can pose severe reliability concerns. The metal can either fail to cover a step completely, thus causing an open circuit fault, or may unevenly cover it, leaving a high resistance thin spot which may fail later due to electromigration. One technique of easing

these problems is called planarization. In this method, the oxide deposited between interconnect layers is partially etched away, leaving a surface with reduced relief. It is a type of "chemical sanding" which smoothes the surface in preparation for the next layer of interconnect. Another related problem involves submicron vias with very steep walls. It is very difficult to cause aluminum to flow into these small holes. For this reason, chemical vapour deposition (CVD) is used to deposit tungsten plugs in the bottoms of the vias and contacts. This reduces the necessity for aluminum to completely cover the via interior. Some processes abandon aluminum altogether and use another metal, such as tungsten[1], which does not suffer as badly from the above problems. These are just some of the techniques used in order to implement reliable, low resistance interconnect in today's scaled technologies.

2.6 BATMOS: Northern Telecom's BiCMOS Process

2.6.1 Introduction

BATMOS is the name given to Northern Telecom's BiCMOS semiconductor process. A description of this process can be found in [21], [19], [22], and a discussion of this material follows. BATMOS is an acronym for Bipolar Analog Telecom MOS. It is a .8μm minimum feature size, 5V process aimed at both analog and digital telecommunications applications. In this process, NPN transistors can be realized with an f_T of 11 GHz. No PNP bipolar devices are available; this is under current development. Benchmark circuits for this technology have yielded impressive results. CMOS inverter delays of 150 ps with 950 ps/pf loading effects have been observed [21]. BiCMOS inverters have achieved 190 ps delay with 510 ps/pF loading effects [21]. ECL inverters have achieved speeds of 71 ps [21]. BATMOS uses many of the advanced fabrication techniques described previously.

2.6.2 Process Synopsis

BATMOS is a twin tub process employing a very lightly doped n epitaxial layer which was chosen to allow individual well optimization. It uses such advanced processing techniques as self aligned buried layers and wells, LDD, SWOS technology, as well as silicided poly and diffusions. Three levels of metal interconnect are provided, with the first two being

[1] The first two levels of metal in BATMOS are comprised of tungsten [19].

comprised of tungsten, and the top level consisting of aluminum alloyed with silicon and copper. Passive components available within the process include 3 types of resistors, and a capacitor. The passive components have a high degree of linearity. Three levels of polysilicon deposition are possible, with the first forming MOS gates and capacitor bottom plates, the second forming the capacitor top plates, and the third forming the polysilicon emitter for the NPN bipolar transistor.

The process flow is set up in a modular fashion, as shown in Figure 2.14. In this diagram, pairs of structures such as NPN Collector—n well formation, p base implant—p LDD implant, and extrinsic base—source/drain implant are indicated. These structures are realized by the same process steps, resulting in a simpler process flow. The figure also indicates modules which can be included or excluded from the baseline process flow without any change in the electrical properties of the devices rendered. This is a very important and valuable strength of the BATMOS process, since it provides a large degree of flexibility in wafer manufacture. Bipolar transistors, or passive resistors and capacitors can be excluded from fabrication to produce different technology variations at reduced cost.

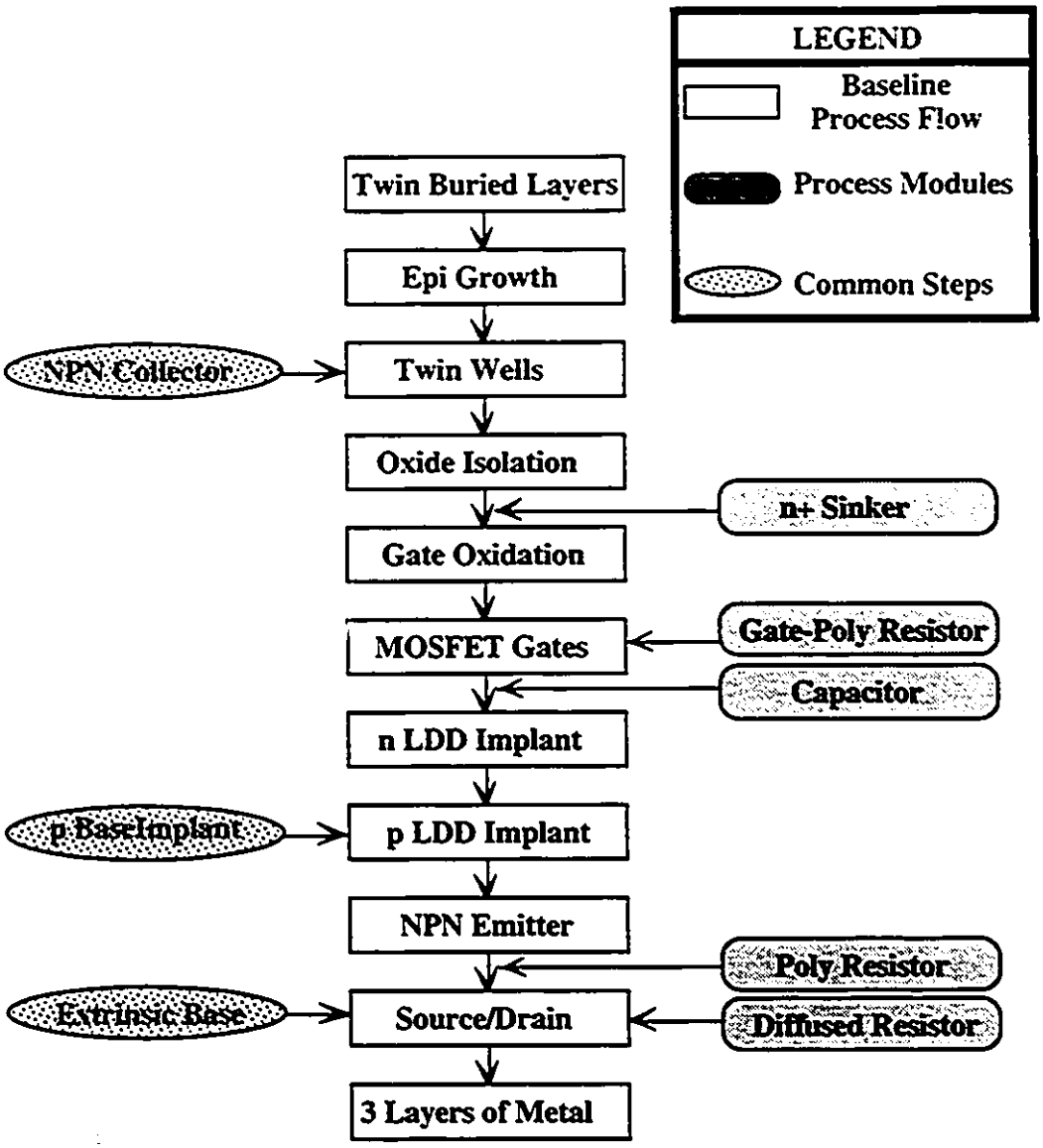
2.6.3 Detailed Process Description

The process starts with a p type substrate on which self aligned buried layers are implanted using the NWELL design layer information. The n+ areas receive an antimony implant while the p+ regions receive a boron implant, each of which are followed by an annealing step to repair implant damage to the crystal lattice and to drive the dopants in. A thin epitaxial layer is grown on the substrate which has been lightly doped with phosphorus to make it slightly n type. Autodoping by the n+ buried layers is avoided due to the fact that antimony was used in lieu of a dopant such as arsenic, which has substantial lateral and vertical autodoping tendencies. Self aligned well formation is carried out by forming oxide and nitride layers, and patterning the layers with the NWELL mask to expose n well regions. After a high energy phosphorus implant to set the well doping, oxide is grown across the n well areas. This oxide is inhibited from growing in p well regions by the nitride layer present. The nitride is then etched, and the n wells are left covered with oxide, supplying a p implant mask and effectively providing self aligned wells. The p well doping level is set by a surface p implant as well as upward diffusion from the underlying p+ buried layer, rendering a "V" shaped doping profile[1].

---

[1] This type of profile is very desirable and its benefits will be discussed in section 3.4.

Figure 2.14: Flow Chart Showing Major Processing Steps



After well formation, layers of oxide, polysilicon, and nitride are deposited in order to implement poly buffer LOCOS (PBL) isolation, as described earlier, and the wafer is patterned and etched such that only device well areas are covered. Here, a device well is considered to be those regions defined by any of the following design layers: NDEV, PDEV, PBASE, RDIF, NPLUG. Photoresist is deposited, and a derivative of the NWELL design mask layer is used (called PG for p guard) to pattern the photoresist such that n wells are covered. A p implant is performed in the p well regions in order to raise the

threshold in the areas which will later be covered by FOX[1]. Note that this implant is blocked by photoresist in the n wells and by nitride in the device wells. The photoresist is stripped, and thick field oxidation is carried out. Only areas not covered by nitride are oxidized. A minimum separation of 1.4μ between these wells is achieved with this isolation technology. The layers associated with the PBL are removed.

The next step forms the collector contact, or sinker, which provides a low resistance link to the n+ buried layer. Photoresist is deposited, and a derivative of the NPLUG design layer mask is used to develop it. A high energy, high dose phosphorus implant is performed which forms a low resistance heavily n+ doped path through the epi down to the buried layer.

In preparation for the gate oxide, a sacrificial oxide is grown to condition the silicon surface in the device wells. The oxide is removed and a high quality, 175 Å thick gate oxide is grown. This thin oxide allows a blanket threshold adjustment implant through the oxide for both types of MOS transistors, and also facilitates good subthreshold control of the devices. Amorphous silicon is deposited on the wafer, and implant doped n+ with phosphorus. If the capacitor option is not being implemented, an annealing step recrystallizes the amorphous silicon into polysilicon and distributes the dopant more uniformly into the poly. If the capacitor process option is being used, a layer of oxide $\approx 300\mathring{A}$ thick is grown, during which time the amorphous silicon is annealed, and another layer of amorphous silicon is deposited and doped. The top plate of the capacitor is patterned, followed by the bottom plate and the MOSFET gates. Amorphous silicon is used to provide a smooth surface allowing very thin capacitor oxide growth. Capacitor and MOS gate definition are accomplished with derivatives of the GATE and CAP design layers. The LRPOLY design layer can be used to define a 30 Ω/square resistor which is formed from the gate poly. The separate mask is necessary to avoid silicidation of the resistor body in later processing steps.

As a first step in LDD formation, moderate phosphorus doping is applied to device well regions in the p well. The gate polysilicon shields the channel region, while the rest of the areas on the wafer are protected by a layer of photoresist previously patterned with a derivative of the NDEV design mask. A thermal drive step allows the doping to diffuse slightly under the gate. A moderate p doping is also implanted into the source/drain regions

[1] An error on page 2-5 of [22] is noted. The implant specified as a p+ guard ring is actually a channel stopper implant defined by a processing mask called pgrdex-PG [19].

of the PMOS transistors in a similar way, using a derivative of the PDEV mask. The p type base implant is performed self aligned to the device well edge, using a derivative of the PBASE design layer mask, and a layer of oxide approximately 800 Å thick is deposited. This latter step is the first in a two step process to form the SWOSs, as well as provide insulation for the poly emitter which will be formed. A rapid thermal anneal (RTA) is performed to repair implant damage and densify the deposited oxide. This anneal must be kept short to limit uncontrolled diffusion of implanted regions. The base oxide is etched to expose active emitter regions, using information derived from the EMITTER design mask.

The polysilicon—emitter interface is very important in determining device quality and predictability, and for this reason it is cleaned very carefully. Amorphous silicon is deposited and a photolithographic step using data derived from the RPOLY design mask is used to pattern a layer of photoresist. A heavy arsenic implant to dope the emitter is performed, while the resistor regions are protected by the previous photoresist layer (RPOLY). Another photomasking and implant step dopes the ends of the resistors p+ which facilitates ohmic contacts. To set the resistor value, an unmasked implant step is performed. The heavy doping levels attained in early steps are not shifted significantly by this implant. The amorphous silicon layer is then patterned to form the emitters and resistors, and annealed to convert them into polysilicon and distribute the dopants[1]. This is the third and last layer of polysilicon which is used in this process.

Another oxide layer is deposited over the wafer, constituting the second portion of the material which will form the sidewall spacers. An anisotropic etch is used to expose gates and diffusions, but leave small amounts of oxide on the edges of poly regions. These are the structures which form the SWOSs. A source/drain—p base boron implant is performed using a photomask constructed from the PDEV, RDIFF and PBASE design masks in order to provide heavy p+ doping in these regions. This is necessary to provide ohmic contacts to the devices, and achieve the necessary value for the diffusion resistor. Even though the emitter and gate regions are open to this implant, significant counter doping does not occur due to concentration differences. The gate poly protects the channel from this implant. The portion of the base external to the emitter, called the extrinsic base, is self aligned to the edge of the SWOS on the emitter, and the source/drain regions are self aligned to the SWOSs on the gate poly edges. The source/drain regions of the NMOS transistors are etched to expose them, using mask information from the NDEV design mask, and they are then doped n+. This concludes the processing steps which form the active devices.

[1] Note that a polysilicon emitter, as opposed to a poly contacted emitter, is implemented in BATMOS.

A self aligned silicidation is next carried out to reduce the resistance of diffusions and polysilicon. A thin oxide is grown over the wafer and areas to be silicided are exposed by etching. The remaining oxide prevents resistors from being silicided. Titanium is deposited, and reacted with the exposed silicon to form a titanium disilicide layer. Unreacted Ti is removed and the wafer is now ready for the interconnection layers[1].

A thick layer of Boro-Phosphosilicate Glass (BPSG) is deposited over the wafer and annealed, to form an insulation layer, and also to form a barrier between contaminants and the active devices. Another layer of glass is deposited, and partially etched back to planarize the topography. Contact masking is performed, and the contact windows are etched with a highly silicide selective etchant in order to prevent contact region damage. A layer of titanium nitride is deposited in order to prevent diffusion of the first level of metal into the silicide layer. CVD is used to deposit tungsten in order to form the METAL1 interconnect layer. This layer is patterned, covered with a layer of undoped glass, and planarized. Similar processing to the first metal layer is carried out for the second level of metal interconnect, defined by the design mask METAL2. Aluminum which has been doped with silicon and copper is used to form the top level of interconnect, which is defined by the TOPMET design layer. Finally, a passivation layer is deposited over the wafer, and pad contact holes, defined by the design layer PAD, are etched to facilitate connection to the bonding wires.

2.7 Summary

This chapter has discussed BiCMOS process technology from the perspective of a CMOS processing base. A short review of fundamental concepts in semiconductor technology was given, followed by a discussion of the evolution of a typical CMOS process into that of a high performance BiCMOS type. The performance of today's integrated circuits would be impossible without very specialized methods and processing techniques. Some of these issues were discussed including epitaxial layers, lightly doped drains, local interconnect, isolation techniques, silicidation, and high performance interconnect issues. Finally, BATMOS, Northern Telecom's BiCMOS technology was discussed, and a detailed process description was given.

---

[1] LI is not presently implemented in BATMOS.

Chapter 3

SCALING, DEVICE, AND PROCESS CONSIDERATIONS

3.1 Introduction

Each year technologies with smaller and smaller feature sizes are developed allowing ever increasing levels of integration. Memory chips now contain in the neighborhood of $10^7 \sim 10^8$ transistors incorporated on a single die [23]. Microprocessors routinely are designed with over 1½ million transistors [24]. Feature sizes of .05μm [25] and below have been reported. When technology is scaled to these levels, process techniques must be developed to control second order effects in active devices which tend to counteract the benefits of scaling. This situation is complicated further when the process must be optimized for both MOS and bipolar devices, and many design tradeoffs must be made.

In the highly integrated chips of today, parasitics associated with the interconnections are starting to dominate while just a few short years ago, most parasitic effects were due to gate capacitance. This is due to two main reasons. First of all, as feature sizes in a given technology are scaled, the parasitic capacitance corresponding to the individual MOS devices also decrease due to reduced physical size. Thus, the contribution of each device to total logic gate capacitance is reduced. Second, the overall size of silicon dies is increasing. This is due both to advances in processing which allow high yield production of larger wafers, and also to the integration levels which are required in advanced chips of today. With larger chips, global interconnections are longer, and thus interconnect capacitance is greater for long lines on these larger dies. It is true that local connections will benefit from the reduced device capacitance, however, global lines are inevitably involved in the critical delay paths of these large, integrated systems.

To illustrate this concept, consider the CMOS1B process [26][1] which was available in the early 1980s. It possesses a 5μm minimum feature size, and one level of metal interconnect. An inverter in this technology, comprised of a minimum sized NMOS device and a PMOS device with twice the minimum width, will have a total gate oxide capacitance[2] of approximately 3.045×10^{-2} pF. This is equivalent to .264 mm of 5μm wide metal interconnect[3]. A typical MSI die in the early 1980s was only a few millimeters on a side, and thus the calculated length would represent a fairly long connection. When a modern process such as BATMOS is considered, the situation changes. An inverter comprised of a minimum sized NMOS device, and a PMOS device of twice minimum width possesses a total gate oxide capacitance of approximately 3.864×10^{-3} pF which is equivalent to .123 mm of .8μm wide METAL1 interconnect[4]. Current integration levels produce die sizes in excess of 10 mm per side, much larger than the 0.123 mm equivalent metal interconnect. In fact, a moderate number of connections on a large die would be of this length, and thus the interconnect capacitance begins to dominate.

In a scaled technology, intrinsic interconnect capacitance remains relatively constant[5]. Even though the metal line widths shrink, which suggests reduced capacitance, the oxide layers are also scaled, which tends to increase capacitance. Parasitic interconnection resistance, which varies inversely with the cross-sectional area of the wire, will increase due to reduced interconnect width, and thus, the resulting RC constant of a length of wire actually increases in value. Therefore, as die sizes increase in a scaled technology, global interconnections play a major part in determining overall system performance.

3.2 Scaling of Active Devices

Scaling is a common means of achieving increased circuit speed and packing density. It attempts to increase circuit performance, in the ideal case, by the minimization of device parasitic capacitances and potential swings. The former is accomplished by reduction of physical dimensions, and the latter chiefly by reduction of voltages such as the power supply level. Scaling approaches for MOS devices and bipolar devices are different. In the

---

[1] This preceded the CMOS3DLM process.

[2] It is recognized that this is not a precise reflection of gate capacitance, but it will do for this simple analysis.

[3] Parameters for this calculation were obtained from [26].

[4] Parameters for this calculation were obtained from [19]. This calculation takes fringing into account.

[5] Consider that the metal—substrate capacitance parameter is $2.3\times10^{-5}\ pF/\mu m^2$ in CMOS1B and $2.78\times10^{-5}\ pF/\mu m^2$ in BATMOS.

ideal sense for MOS devices, all physical parameters are altered such that the electric field strength and shape within, and surrounding, the active device are maintained as they exist in the non-scaled device. This is the key to avoiding effects which compromise performance and reliability of the devices. In the case of MOS devices, the above scaling method is referred to as ideal scaling, or constant field (CE) scaling. There are several other approaches, including constant voltage (CV), and quasi-constant voltage (QCV) scaling. For the case of bipolar devices, there are two main approaches which are generally referred to as constant collector current (CIC) and constant collector current density (CJC) scaling. Bipolar scaling, in general, is more complex than its MOS counterpart due to the strong three dimensional component of the processes which induce device behaviour.

To highlight the performance increase gained from scaling, consider a simplified example [27] of ideal MOS scaling. An elementary relation expressing simple circuit delay is given by:

$$T_D = \frac{C_L \Delta V}{I_{avg}} \tag{3.1}$$

where T_D is the circuit delay, ΔV is the voltage swing, C_L is the complete output load, and I_{avg} is the average current drive of the circuit. Thus, the fundamental goals of ideal scaling are to reduce T_D by reducing C_L and ΔV while maintaining a high I_{avg}. This should be attained without compromise of the physical integrity of the devices which could result from high electric fields. In the case of ideal MOS scaling, horizontal and vertical device dimensions, including width, W, length, L, gate oxide thickness, t_{gox}, and source/drain junction depth, X_j, are scaled by a factor of $1/k$ $(k>1)$, while substrate doping, N_{SUB}, is scaled by a factor of k. Voltages, including power supply voltage, V_{DD}, and threshold voltages, V_{TN} and V_{TP}, are reduced by a factor of $1/k$. This scaling method renders field strengths and shapes which are largely the same as in the non-scaled device, and thus oxide breakdown, hot carrier effects, and velocity saturation will not be a problem. The MOS gate capacitance per unit area, C_{gox}, given by the relation:

$$C_{gox} = \frac{\varepsilon_{ox}}{t_{gox}} \tag{3.2}$$

is increased by a factor of k, since the value of t_{ox} has been reduced by $1/k$. As shown below in equation (3.3), the drain current through a scaled device, I_{DS} is reduced by a factor of $1/k$.

$$\begin{aligned} I_{DS} &= \mu C_{ox} \frac{W}{L} \frac{(V_{GS} - V_T)^2}{2} \\ &\propto 1 \times k \times \frac{1/k}{1/k} \times (1/k)^2 = \frac{1}{k} \end{aligned} \tag{3.3}$$

The gate capacitance, C_g, is also reduced by a factor of $1/k$, as shown in equation (3.4).

$$\begin{aligned} C_g &= WLC_{ox} \\ &\propto \frac{1}{k} \times \frac{1}{k} \times k = \frac{1}{k} \end{aligned} \tag{3.4}$$

Considering a CMOS gate which drives a purely capacitive load, then let I_{avg} represent the average current of the pull-up or pull-down section of the gate. Since the gate is CMOS, the potential swing will be V_{DD}. Assume that the load is composed of other CMOS gates, and thus comprised of the sum of the individual MOSFET gate capacitances making up these gates, then let this load be represented by nC_g. Thus, the circuit delay for this case can be estimated using equation (3.1), and a dimensional analysis yields the result that the scaling process has reduced the gate delay by $1/k$. This analysis is shown in equation (3.5).

$$\begin{aligned} T_D &= \frac{nC_gV_{DD}}{I_{avg}} \\ &\propto \frac{1/k \times 1/k}{1/k} = \frac{1}{k} \end{aligned} \tag{3.5}$$

Additionally, since power is the product of voltage and current, it is obvious that the power dissipation per scaled gate will be reduced by $1/k^2$, since both the current and voltage are reduced by a factor of $1/k$.

The area of each device will be reduced by $1/k^2$, since both the length and width of the devices are scaled, and thus the packing density, which is proportional to the inverse of the area of the individual devices, will increase by a factor of k^2. Finally, the power dissipation density, or the ratio of power dissipation to chip area, remains the same since

both the area and power terms are scaled by the same factor, thus canceling out yielding a scale factor of one. The above results and relations are summarized in Table 3.1.

Table 3.1: Summary of Ideal or Constant Field (CE) Scaling

Parameter—Quantity	Scaling Factor
Dimensions (W, L, t_{ox}, X_j)	$1/k$
Voltages (V_{DD}, V_{TN}, V_{TP})	k
Substrate Doping (N_{SUB})	$1/k$
Device drain current (I_{DS})	$1/k$
Gate capacitance (C_g)	$1/k$
Simple gate delay (T_D)	$1/k$
Power dissipation per simple gate	$1/k^2$
Packing density ($\propto 1/Area$)	k^2
Power dissipation density (*Power/Area*)	1

Thus, ideal scaling results in faster, lower power, and more spatially dense chips which have the same power dissipation density as their non-scaled counterparts. These are all desirable qualities, with the latter allowing higher levels of integration with the same thermal conduction requirements. Disadvantages, however, occur due to reduced subthreshold swing in a scaled device. This is due to the fact that voltage swing does not scale [23].

Ideal scaling is a natural progression from a device physics point of view. It is obviously preferable to preserve the desirable device characteristics by maintaining field profile over scaling. Industry cannot, however, change supply voltages every time technology advances a generation, or every time feature sizes are reduced. This would involve tremendous costs to system designers trying to implement designs with integrated circuits from different vendors and different technologies [28], [29]. One difficulty would lie in providing expensive, multi-level power supplies. Driven by this practical scaling limitation, several approaches have been developed which are derivatives of the above fundamental procedure. These methods can be categorized, in addition to the previous ideal, or constant field (CE) method, as constant voltage (CV) and quasi-constant voltage (QCV). They attempt to address the voltage scaling problem.

Constant voltage scaling, as its name suggests, does not scale the relevant voltages. Dimensions are selectively scaled. Horizontal dimensions may be scaled by a certain factor, k_1, while vertical dimensions such as gate oxide thickness may be scaled less aggressively, by a factor of $\sqrt{k_1}$. This is necessary to avoid such problems as oxide breakdown due to excessive field intensities across the scaled gate oxide. The CV method is only possible down to about .6μm or .5μm channel lengths with a 5V standard supply voltage [16], after which internal field levels will pose serious reliability concerns due to hot carrier effects. This has been the prime mover behind the recent and ongoing switch by industry to a 3.3 V standard supply voltage [29], [28].

Table 3.2: Comparison of Different MOS Scaling Methods

Parameter—Quantity	Constant Field (CE)	Constant Voltage (CV)	Quasi-Constant Voltage (QCV)
Horizontal Dimensions (W, L)	k	k_1	k_2
Gate oxide thickness (t_{ox})	k	$\sqrt{k_1}$	k_2
Substrate Doping (N_{SUB})	k	k_1	k_2
Supply voltage (V_{DD})	k	1	$\sqrt{k_2}$

Quasi constant voltage scaling is a compromise between constant field scaling and constant voltage scaling. In this method, the device dimensions are scaled more aggressively, say by a factor of k_2, than the voltage levels, which are scaled by $\sqrt{k_2}$. This allows more aggressive scaling of dimensions than with the CV method because of the reduced fields which are rendered. Less aggressive scaling of voltages addresses two concerns. Firstly, threshold voltage is effected by many factors such as channel length, drain voltage, process variability, and temperature, and thus should be scaled less aggressively. Also, the built-in junction potential does not scale, and thus, the depletion width becomes a larger portion of the channel length for small devices. Table 3.2 compares scaling factors of major parameters for all three scaling methods. A general method which scales dimensions independent from voltages is presented in [30]. In this method, the shape of the electric field is maintained, although local fields increase by a factor equal to the ratio of the dimension scaling factor and the voltage scaling factor.

Scaling bipolar devices is generally more complicated than MOS devices due to their inherent complexity. Structurally, the BJT is more complex than a MOS transistor. In the latter, all of the physical processes which constitute its behaviour occur essentially in a thin layer at the substrate surface. In a long channel device, the behaviour can be described with reasonable accuracy using a two dimensional model. This is not the case for a BJT, since it is a vertical device. The physical processes which induce its behaviour involve all layers, from the thin, highly doped emitter layer just under the silicon surface, through a series of complex doping profiles which make up the base and collector, and finally, involving the deep buried layer forming the lower collector. Both diffusion and drift currents play important roles in determining device behaviour. For these reasons, scaling is a more complex process for bipolar devices. The two main procedures which are popular, called constant collector current (CIC) and constant collector current density (CJC) scaling, are discussed in [16], and will be treated briefly here. Both methods attempt to maintain a current quantity constant as the device size and dopings are altered.

Unlike MOSFET's, which are essentially field controlled devices, BJTs are controlled by current, and thus bipolar scaling methods attempt to maintain the current or current density within the device constant. In the CIC scaling method, the knee current of the device is kept constant. This method assumes that the high frequency gain roll off is due to base pushout, or the Kirk effect[1], and the resulting equation for knee current is used as a starting point for the technique. Thus, if CIC scaling is used, the same current flowing through a smaller device will require increased doping in the collector to delay the onset of the Kirk effect. This will be unnecessary in the CJC method, since the same current density will occur in the scaled device which is a direct result of the intentional scaling of the collector current by a factor of k^2, if k $(k<1)$ is the horizontal scaling factor. The ensuing lower collector current, however, will require high quality emitters (polysilicon) and, as well, parasitic capacitances will be more important. This latter point is placated, since the strong requirement to increase the collector doping due to the Kirk effect is not present. Recall that increased doping levels increase the parasitic capacitances. A decrease in base width in either method will require an increase in base doping due to punchthrough considerations. Voltage is either maintained at the non-scaled level, or scaled less aggressively than other quantities. The former case may pose reliability problems as feature sizes continue to shrink, since highly doped emitter and base regions can induce hot carrier generation in the emitter—base depletion region, leading to instability in device parametrics [31]. In general, CIC scaling is preferable because leakage current due to stress of the emitter—base junction

[1] This will be discussed in section 3.3.2.3.

has less impact on current gain [32]. A sample of scaling values is illustrated in Table 3.3. In this table, k_h is the horizontal scale factor, k_v is the vertical scale factor, and k_u is the voltage scaling factor. All factors are less than one.

Table 3.3: Comparison of Different Bipolar Scaling Methods

Parameter	Constant Current (CIC)	Constant Current Density (CJC)
Horizontal Dimensions	k_h	k_h
Vertical Dimensions	k_v	k_v
Base Doping	$k_v^{-2}k_m$	$k_v^{-2}k_m$
Voltage	k_u	k_u
Knee current	1	k_h^2
Collector current	1	k_h^2
Collector current density	k_h^{-2}	1
Epi doping	k_h^{-2}	1
Collector depletion capacitance	k_h	k_h^2

A key limitation in bipolar scaling is that the voltage necessary to turn the device on, sometimes referred to as the bipolar threshold voltage, is determined by the bandgap of the semiconductor material. Thus, it is not a parameter which is variable in a scaling process. This is in contrast to the threshold voltage of a MOS device, which is sensitive to both process parameters and geometry, and can be adjusted. For these reasons, the degree to which power supply voltages can be reduced and still yield high performance BiCMOS circuits is limited [33][1].

Scaling of devices in a BiCMOS technology is essentially a combination of the above methods. As feature sizes are scaled to submicron levels, short channel effects mandate heavy subsurface doping in MOSFET channels, while high well doping is required to delay the onset of the Kirk effect in bipolar devices. As a result, process requirements tend to converge in some areas for bipolar and MOS devices as feature sizes shrink, which tends to relax the design tradeoffs which must occur between the two types of devices.

[1] This statement is based on conventional totem-pole style gates utilizing no special circuit techniques.

BiCMOS gate performance is effected significantly by the scaling of supply voltages. This is due to bipolar device related issues outlined above, as well as other reasons. In the conventional totem pole driver configuration [8], the output is not comprised of a full voltage supply swing, or rail-to-rail swing. This may not be a major concern at 5V supply levels, but as levels fall, the percentage of supply swing which is not covered increases due to the non-scalability of the semiconductor bandgap. These degraded signals cause leakage currents to flow in MOS devices which increase as the MOS devices are scaled, leading to increased power dissipation, and unacceptable performance. Circuit techniques to provide rail-to-rail swing in BiCMOS totem pole type circuits have been studied [34], [35], [36], [37], [38], and usually involve either the use of MOS devices to pull the final output to one of the supply rails, or some sort of level shifting technique [39]. BiCMOS integrated circuits have an advantage in that critical circuit paths can be implemented in pure ECL logic to increase performance [40]. Voltage scaling will effect this as well, since as supply levels shrink, three level gating will no longer be possible. Thus, there are additional considerations when devices are to be scaled in a BiCMOS process.

3.3 Active Device Issues

The benefits of scaling a BiCMOS process are counteracted by many second order device effects which are unfortunately, aggravated by the scaling procedure. In MOS devices, mobility degradation, velocity saturation, and increased parasitic source and drain resistances degrade the current drive of scaled devices. Such things as non-scalability of the silicon bandgap cause subthreshold MOS characteristics to suffer, and also cause threshold levels to vary with device geometry. Failure to scale the supply voltage causes gate oxide breakdown, as well as hot carrier injection into the gate oxide, which shifts device threshold voltage over time. Additionally, scaling feature sizes tends to increase the likeliness of the occurrence of latchup. The scaling of bipolar devices also has its share of problems. Thin, lightly doped epitaxial layers which are necessary for high device f_T, cause high collector resistance, and are more susceptible to gain roll off due to the Kirk effect. If doping is increased in the epitaxial layer, collector—emitter breakdown voltage will suffer. Very thin, lightly doped base regions are susceptible to collector—emitter punchthrough. Important device issues which are particularly relevant in scaled BiCMOS technologies will be discussed, as well as processing issues related to them. MOS and bipolar devices will be treated separately, and then BiCMOS will be treated in general.

3.3.1 MOS Device Issues

There are many design issues which must be dealt with so that a process can render high quality MOS transistors of small feature size. These design considerations are necessary to control performance compromise by second order effects. The major effects will be discussed, and some processing considerations and tradeoffs which minimize them will be mentioned.

3.3.1.1 Channel Length Modulation

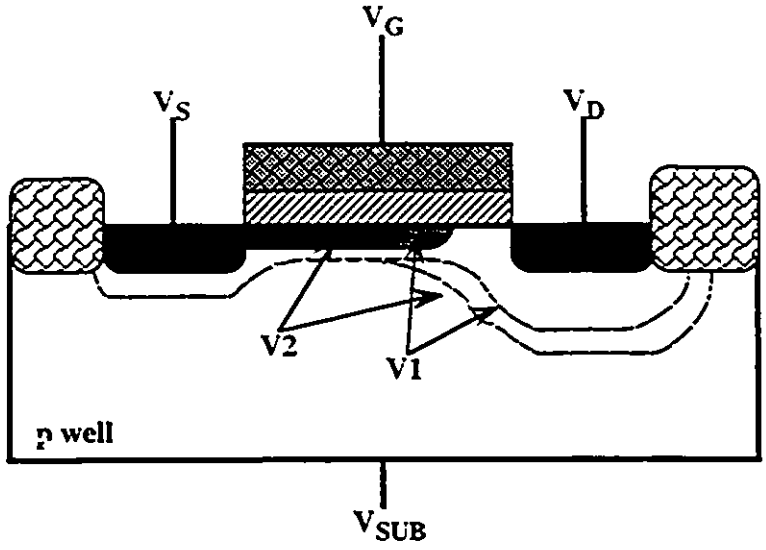
In simple MOS theory, the drain current I_{DS} remains constant with increasing drain voltage, V_{DS}. In reality, there are physical processes which cause the drain current to deviate from this ideal behaviour. As drain voltage is increased, the reverse voltage applied to the drain—substrate junction is increased. This has the effect of increasing the depletion width of that junction, which in turn moves the drain depletion boundary closer to the source. The effective channel length of the device is reduced, which in turn, increases the drain current. This is illustrated in Figure 3.1, where the channel and depletion edge are shown for low (V1), and high (V2) drain voltages. The net effect is that a drain voltage increase will be accompanied by a rise in the drain current, which in turn will cause the MOS device to have a finite output resistance. In a static inverter, this has the effect of making the transfer curve less sharp. The impact of this effect is more severe in short channel devices, since the depletion widths, which are doping and bandgap dependent, occupy a larger percentage of the channel length.

3.3.1.2 Threshold Voltage

The threshold voltage of a MOSFET must be low enough to provide high current drive, yet high enough to provide acceptable subthreshold characteristics and leakage current. It is affected by a variety of different phenomenon when scaling is executed, including hot carrier effects, device geometry, doping profiles, and voltage levels. Some of these factors will be discussed.

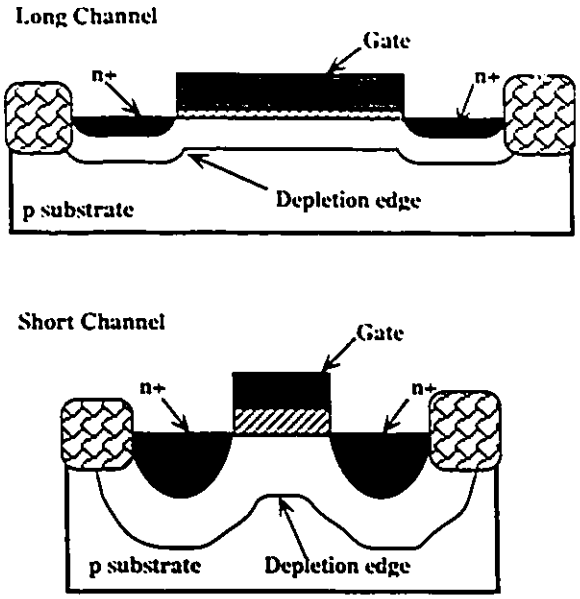
The body effect, or back bias effect, is a well known cause of threshold increase in MOSFETs. It is caused by the presence of a source—bulk voltage which must be counteracted by the applied gate voltage in order to induce channel generation, and thus has the effect of increasing threshold voltage. Scaled technologies commonly employ increased substrate doping which aggravates the body effect.

Figure 3.1: Illustration of Channel Length Modulation



As geometries are scaled, the threshold voltage of MOS devices tends to change due to the correspondingly smaller channel dimensions. Decreased channel lengths tend to decrease the threshold voltage: a phenomenon usually referred to as the short channel effect. A long and a short channel device are pictured in Figure 3.2 which also shows depletion region edges associated with the devices. Charge associated with the drain and source depletion regions accounts for a larger percentage of channel charge in the short channel device. In a long channel device, the channel properties are controlled by gate—bulk interaction. In the short channel device, the charge accumulation supported in the channel is also due to gate voltage, however the contribution from the source and drain depletion regions becomes more significant as the channel length is scaled. Since the gate voltage is required to support less charge than if the depletion charge was absent, the threshold voltage is reduced. The reduced quantity of charge is illustrated in Figure 3.3. The trapezoidal area illustrated represents the bulk charge controlled by the gate, and it is smaller due to drain and source infringement than the rectangular section indicated in the lower right of the diagram. This latter area represents the charge in a channel in which drain and source regions have no effect, and may be a valid approximation for long channel devices. This concept is closely related to drain induced barrier lowering, which will be discussed in a subsequent section. The result of the above phenomenon is the reduction of the threshold voltage in short channel devices.

Figure 3.2: Long and Short Channel MOS Devices



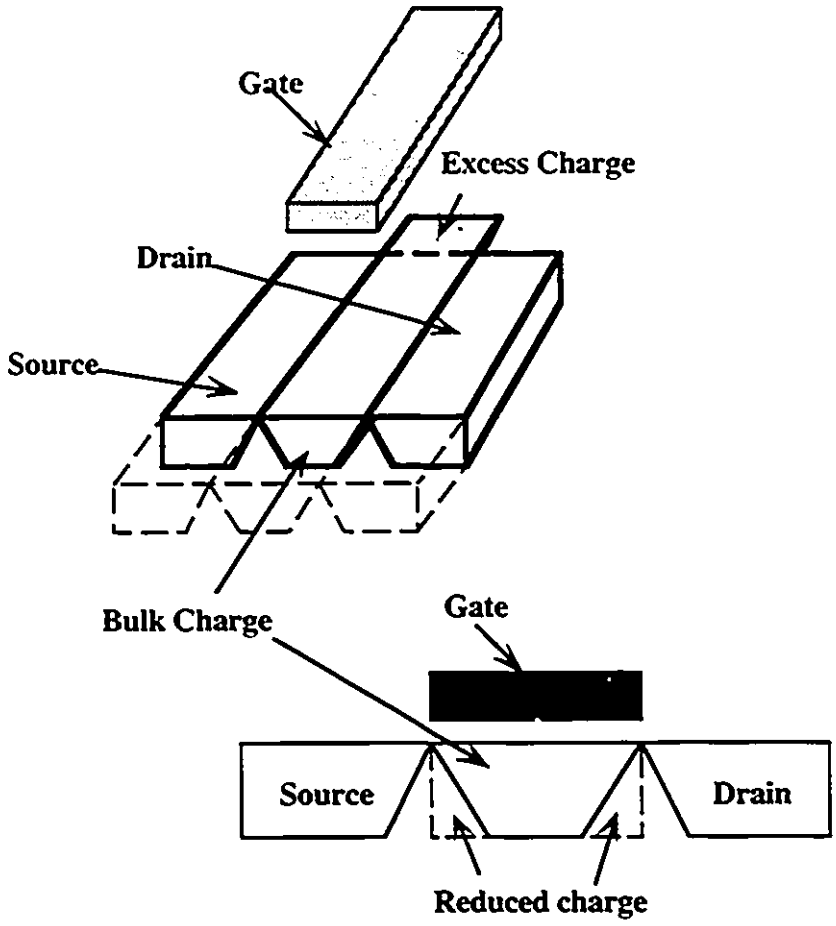
In an inverted channel, some charge is accumulated under the field oxide bordering the length which does not contribute to the device drain current. This charge is supported by the gate voltage, and it is conceptually illustrated in Figure 3.3 as excess charge along the rear length of the channel. As a channel is scaled to very narrow dimensions, the proportion of total gate charge, that this excess charge accounts for, grows, thus a larger proportion of gate voltage is used in supporting it. This factor tends to increase the threshold voltage as channel width is reduced. Threshold voltage is more strongly effected by the scaling of channel length than by the scaling of channel width.

The background doping is key to determining device behaviour, and many short channel effects are reduced by increasing its value, however this is at the expense of increased body effect and junction capacitances. The compromise usually taken is to provide one or two implants to increase surface and immediate subsurface doping, thus allowing a more lightly doped tub. Retrograde wells with "V" shaped profiles are a very effective compromise, and they will be discussed in section 3.4. The channel doping profile is the single most important factor which determines device characteristics, including of course, threshold voltage.

3.3.1.3 Subthreshold Current

In fundamental MOS transistor theory, it is assumed that no drain current flows when the gate to source voltage is less than the threshold voltage of the device. In a physical device, the drain current decreases exponentially to zero as the gate to source voltage falls below the threshold voltage.

Figure 3.3: Conceptual Illustration of Charge in a Short Channel Device



Subthreshold behavior is characterized by a quantity called subthreshold swing, denoted by S, which indicates the quality of switch that the MOSFET provides. It is defined as the amount of gate voltage change which is necessary to effect a decade drop in drain current. A small value of swing is desirable, since this means that the device will turn off quickly. A typical value of swing for transistors in a modern BiCMOS process is 84 mV/decade[1].

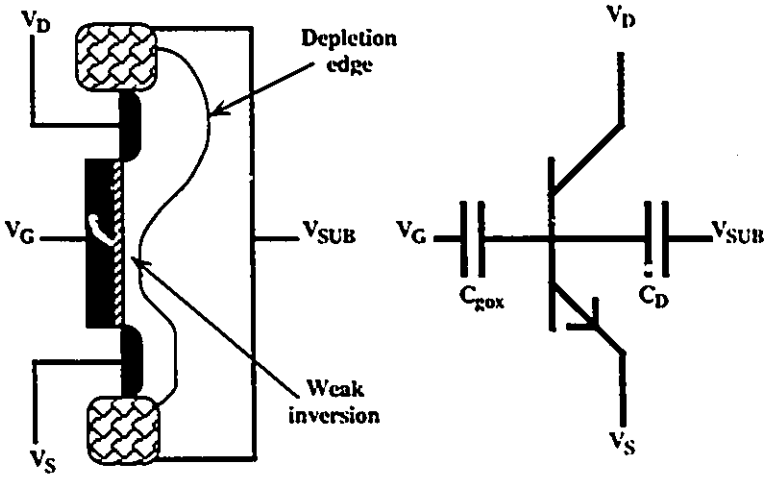
[1]Northern Telecom's BATMOS process presently yields NMOS transistors with a swing of 84 mV/decade [21].

Consider an NMOS transistor conducting 1 μA when the gate voltage is equal to the threshold voltage. If a current of 1 pA is an acceptable leakage current, then it will require a gate voltage drop of .504 V to attain this. The swing of a MOS transistor can be expressed as [16], [27]:

$$S = \ln 10\left(\frac{k_B T}{q}\right)\left(1+\frac{C_D}{C_{gox}}\right) \tag{3.6}$$

where C_{gox} is the gate oxide capacitance, and C_D is the depletion capacitance. Note that $\ln 10(k_B T/q) \approx 60mV$ / *decade*, and thus there is a fundamental limit to how much the swing can be reduced. This has serious implications to scaling practices, since threshold voltages which are scaled excessively will yield very poor quality switches. The leakage current in these devices will be unacceptably high due to the inability of the reduction of gate voltage to completely turn the device off.

Figure 3.4: Weakly Inverted MOSFET—Bipolar Analogy



At gate voltages below the MOS threshold voltage, the device ceases MOSFET-like behaviour, and begins acting more like a bipolar transistor. The channel is not strongly inverted, and there is no low conductance path between the drain and source, thus minority carriers and diffusion currents are dominant as in a bipolar device. This concept is illustrated in Figure 3.4. As is shown, the weakly inverted channel of the MOS device in the subthreshold region acts like a bipolar base, while the source and drain act as the emitter and collector respectively. A capacitive voltage divider is formed by the gate oxide and

depletion capacitances, thus only a portion of the gate voltage is seen by the base region. This is the origin of the $\left(1 + C_D/C_{ox}\right)$ term in equation 3.6.

As mentioned previously, the channel doping profile is key in determining subthreshold characteristics. The subthreshold swing of a device varies directly with the channel depletion capacitance, and inversely with the oxide capacitance. For this reason, the swing is decreased (improved) with a decrease in the channel doping, as well as a decrease in the oxide thickness. Actually, subthreshold swing requirements place an upper limit on the maximum surface channel doping which can be used to reduce short channel effects. This has implications in BiCMOS process design.

3.3.1.4 Velocity Saturation

Simple FET theory assumes that carriers in the channel, with an associated mobility factor of μ[1], move at a velocity, v, which is a linear function of the transverse electric field, E, present. This relationship is described in the following equation:

$$v(E) = \mu E \tag{3.7}$$

When devices are scaled, and the power supply is not, fields develop which cause this relationship to fail, and carriers reach a maximum velocity, termed the saturation velocity, v_{sat}. Once the carriers reach this speed, increase in field strength does not increase velocity. The critical field at which this occurs is dependent on carrier scattering, and is about 1.5×10^6 V/m [41]. This corresponds to a maximum velocity of about 10^5 m/s. One velocity saturation equation, first proposed in [42], is given as:

$$v(E) = \frac{\mu E}{1 + \frac{\mu}{v_{sat}} E} \tag{3.8}$$

This relation is used in [41] to derive an expression for drain current in a saturated MOS transistor where the carrier velocity in the channel is saturated. This expression is written as:

$$I_{DS} = W C_{ox} (V_{GS} - V_T) v_{sat} \tag{3.9}$$

[1] It is recognized that electrons and holes have different mobilities, however, this is temporarily ignored in the interest of simplicity.

It is important to note that the drain current no longer varies with the square of the gate voltage but follows a linear dependence, thus scaled devices operating with velocity saturated carriers will have reduced current drive. Also, notice that the velocity saturated current does not depend on channel length. This has important implications in scaling, since it means that further reduction in channel length will not afford increased current drive.

3.3.1.5 Mobility Degradation

Mobility of carriers in the inverted channel of a MOSFET is a decreasing function of the normal field [43], and this is attributed to carrier scattering at the surface oxide—silicon interface which has the net effect of decreasing the carrier mobility. This reduction of mobility is exacerbated by scaling practices, especially if the voltage is not diminished, or if it is reduced less aggressively than the spatial dimensions, because the average normal field is increased. Even when proportional voltage scaling is applied, the average normal field still increases due to non-scalability of material work functions [30], and the mobility is still reduced. Thus, carrier scattering in short channel, scaled devices can have a negative effect on current drive.

There are three main scattering mechanisms [43]. Phonon scattering is caused by lattice vibrations, and has little effect at low temperatures. Coulomb scattering is a result of the presence of charge centers which are comprised of fixed oxide charge, interface state charge, or localized charge due to ionized impurities. This mechanism is important in the weakly inverted channel. Finally, there is surface roughness scattering which is a result of deviations of the Si—SiO_2 interface from the surface plane. The important factors which determine the dominant scattering mechanism are field strength and temperature. At low temperatures and low fields (weak inversion), Coulomb scattering is dominant, while at high fields, surface scattering is the main mechanism. At room temperature, however, important low field scattering mechanisms are the Coulomb and phonon type, while at high fields, surface roughness and phonon scattering are the predominant mechanisms.

3.3.1.6 Effect of Very Thin Gate Oxide

In ideal scaling, the oxide thickness is reduced, but since the field is constant, the inverted channel thickness is the same. The inverted channel charge can be expressed in terms of a capacitance, C_{inv}, through the expression [30]:

$$C_{inv} = -\left(\frac{\partial Q_n}{\partial \phi_s}\right) \tag{3.10}$$

where Q_n is the charge accumulated in the inversion layer, and ϕ_s is the surface potential. This is an incremental capacitance dependent on the amount of charge induced. Scaling gate oxide thickness achieves greater channel control due to the increased voltage available for inversion charge accumulation. The voltage drop across the thin inversion layer is usually neglected in MOS analysis[1]. The traditional model is comprised of a series connection of capacitors representing the oxide capacitance, and the depletion capacitance. When the gate oxide is made thinner, the gate oxide capacitance increases. This means that less voltage will drop across it, and thus, there will be more voltage available for charge accumulation which results in higher device current drive [44]. This is why modern technologies are implementing thinner and thinner gate oxides[2]. With very thin gate oxides, C_{inv} increases to a point where there is very little voltage dropped across it. Since the proportion dropped across the depletion/accumulation region is that much larger, the voltage drop across the inversion layer can no longer be ignored. This leads to the sublinear behaviour reported in [45]. Modern device modeling is thus more complex since this effect must be incorporated into the fundamental analysis.

3.3.1.7 Source and Drain Resistance

As devices are scaled, junction depths and contact window sizes are also reduced. This effectively increases the resistance of the respective drain or source with which they are associated. The net effect is a reduction in device transconductance. The contact resistance associated with the drain or source connection can be broken up into two major components which are caused by different physical processes. The first component is due to the restriction of current flow through the contact opening, while the second is due to material properties.

When current issues through a narrow contact window into a semiconductor, resistance results from the restricted flow. The actual pattern of flow, and the speed with which the flow spreads out is dependent on such things as the layout geometry, type of contact (n+ or p+), whether the contact is part of an active device or simply a substrate or tub contact, and, also, on the substrate architecture. The latter point includes such factors as thick or

---

[1] See [18], page 317.

[2] BATMOS has a gate oxide thickness of 175Å. An inversion layer is typically 100Å thick.

thin epitaxial layer, doping levels in the substrate, and existence of buried layers. The resistance resulting from restricted current flow will vary greatly depending on the above mentioned factors.

Material properties also have a strong effect on contact resistance. Interface imperfections, and the nature of the metal—semiconductor junction are the main determining factors. A junction between a metal and a semiconductor forms a diode. The band structures of the metal and semiconductor align in such a way as to cause charge transfer which results in an equilibrium state, and a depletion region. A metal—semiconductor junction is very similar to a one sided pn junction, in which one side is lightly doped and the other is very heavily doped. The depletion region will extend very far into the lightly doped side, and almost a negligible distance into the heavily doped side. Such is the case with a metal—semiconductor junction, where the allowed states and electron densities in the metal are very much larger than the doping of the semiconductor. A diode makes a very poor contact, however, and thus the diode nature of the contact must be minimized to guarantee an ohmic contact. This is accomplished by increasing the semiconductor doping near the interface to the level of degeneracy[1]. The potential barrier, and depletion width formed by the junction are then so small, the latter being only a few angströms, that carriers can tunnel through easily in both directions.

In scaled technologies, shallow junctions not only increase resistance, but they also create reliability concerns due to metal spiking, as well as hot carrier generation. Spiking involves diffusion of metal into and through the heavily doped contact area, and it can create a short circuit. Alloying techniques can reduce this problem, however this does not address the hot carrier issue.

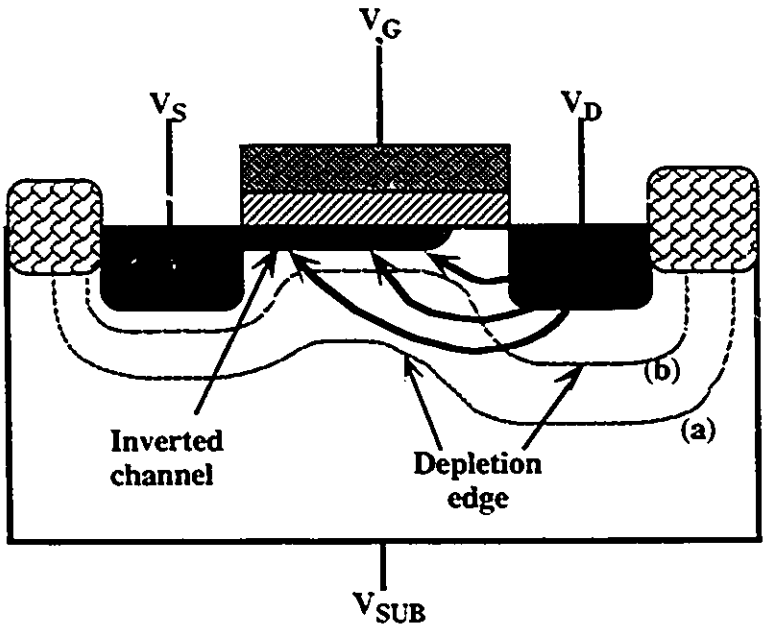
3.3.1.8 Drain Induced Barrier Lowering (DIBL)

The normal subthreshold current in a MOSFET is a surface current, flowing close to the silicon—oxide interface. If the drain voltage is high, leakage current occurs below the surface of the channel. This current is due to a phenomenon called drain induced barrier lowering (DIBL), and it is exacerbated as channel length is decreased. The DIBL effect is illustrated in Figure 3.5. As channel length shrinks, the drain and source depletion layers are spaced closer together. Field penetration from the drain to the source, indicated by the

[1]Degeneracy here is taken to mean that doping concentrations are so high that Maxwell-Boltzman statistics no longer apply, and Fermi-Dirac statistics must be used. The name degenerate comes from the fact that under these conditions, different quantum states can have the same energy [13].

curved arrows in the diagram, lowers the potential barrier at the source end, allowing carriers to cross the junction. This increases the leakage current, and thus increases the subthreshold swing of the device. The diagram shows two depletion edges, labeled (a) and (b). They correspond to low and high bulk doping respectively. As is illustrated, with low substrate doping, the depletion edge is far from the device, allowing field penetration from the source to the drain. This is one of the reasons why aggressively scaled technologies require heavier substrate dopings. When the sum of the drain and the source depletion region widths is greater than the channel length, the structure no longer behaves like a MOSFET, and the resulting condition is termed punchthrough.

Figure 3.5: Mechanism for Drain Induced Barrier Lowering



3.3.1.9 Hot Carrier Effects

If supply voltage levels are not scaled with other parameters such as spatial dimensions, field intensities within the devices will inevitably increase. Hot carrier generation occurs at field strengths of approximately 5×10^6 V/cm [16] or above, and it can shift device threshold voltage, alter subthreshold swing, and reduce device transconductance over time. It embodies the generation of very high energy carriers which are injected into the silicon—oxide interface, and the gate oxide itself. When the former occurs, interface traps (states) may be formed which can bind free carriers and form stationary charge, and when the latter

occurs, existing oxide traps acquire excess charge which also forms stationary charge. These processes develop an embedded charge over time which causes the negative effects mentioned above. Additionally, high energy carriers can cause impact ionization at the drain end. This involves the creation of free carrier pairs by the impact of energetic carriers with lattice atoms. When this occurs in an NMOS device, a minority hole current flows into the substrate [17]. Hot carrier effects are worse in NMOS devices because holes have a much smaller mean free path than electrons and are therefore much cooler in a given electric field [46]. For this reason they are less effective at forming interface traps, and reliability concerns are not as great.

The chief means of combating hot carrier effects in modern BiCMOS processes is with the formation of lightly doped drains (LDD). These structures were described earlier, and consist of lightly doped extensions of the drain (and source) diffusions into the channel. The extension on the source diffusion is essentially a byproduct of the process, while the one on the drain is very important. It causes the depletion region which faces the channel to extend further into the effective drain region. This has the consequence of spreading out the electric field, thus reducing its intensity. Hot carrier generation is much less likely with the resulting reduced fields. This structure increases the drain and source resistance, but this is traded off against the advantage of allowing smaller channel lengths without the normal reliability concerns. Anomalous behaviour in LDD PMOS devices has recently been reported in [47] where results suggest that LDD structures actually exacerbated hot carrier damage as compared to an abrupt junction device. The results were measured from surface channel, p+ poly gate devices however, and thus are not broadly applicable[1].

3.3.2 Bipolar Device Issues

Scaling of bipolar devices tends to aggravate many of the performance compromising factors which are present in non-scaled devices, and special processing techniques are necessary to control them. Thin, highly doped epitaxial layers which are necessary for providing small collector resistance, and delaying the onset of the Kirk effect are limited by junction breakdown considerations and parasitic capacitance. Polysilicon emitters are necessary in scaled devices for acceptable current gain; however, their realization requires special processing steps. More compact device footprints are possible through self aligned processes; however, this requires special techniques such as sidewall oxide spacer formation. Behaviour of bipolar transistors is determined by many different parameters,

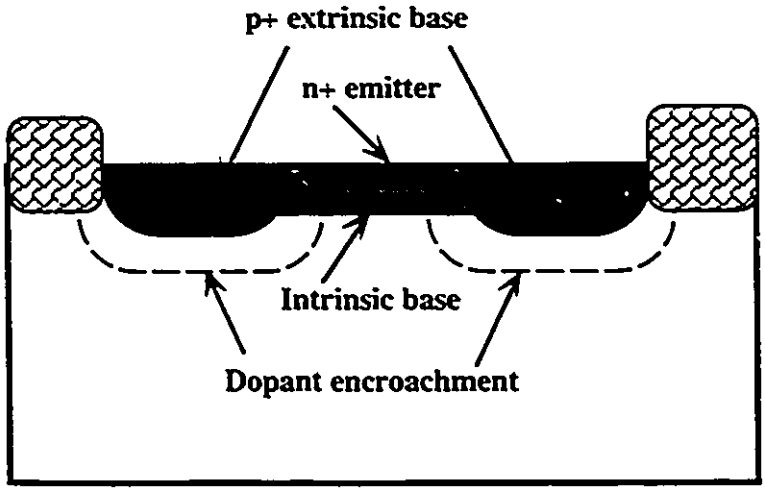
[1] BATMOS uses n+ gate poly, and buried channel PMOS devices are rendered.

some of the key ones being cutoff frequency, f_T, peak current gain, β_f, breakdown voltages, BV_{CBO} and BV_{CEO}, reachthrough voltage, V_{rt}, punchthrough voltage, V_{pt}, emitter, collector and base resistances, R_E, R_B, and R_C respectively. The relationship of these parameters to process parameters, such as collector doping concentration, N_C, collector width, W_C, base Gummel number, Q_B, peak base doping, N_B, and base width, W_B, will be discussed.

3.3.2.1 Extrinsic and Intrinsic Device

An important idea in the scaling of bipolar device geometries involves the concept of intrinsic and extrinsic devices. The intrinsic device is considered to be the portion of the active device immediately under the emitter. The extrinsic device, with a very high doping, is used to contact the intrinsic base region, and thus is spaced as close as possible to it. A double contacted base structure is pictured in Figure 3.6. Sidewall oxide spacer technology is very important in achieving minimum separation, but encroachment of dopant into the intrinsic base region is a major limitation as feature sizes are scaled.

Figure 3.6: Intrinsic Base Region With Extrinsic Base Dopant Encroachment



This encroachment is indicated in Figure 3.6. It effectively yields a heavily doped junction at the sidewall of the emitter and thus reverse leakage current increases due to tunneling. In a bipolar structure such as this, carriers are injected from the bottom as well as the sidewall of the emitter, but the effective base width and Gummel number is significantly larger for

the extrinsic base than for the intrinsic one. The result is a reduction in current gain, and f_T of the device which increases markedly as the area to perimeter ratio of the emitter decreases [48]. This is because the percent of sidewall injected carriers begins to account for a significant amount of the injected base current, and since the extrinsic base is of much lower quality, the gain of the device suffers. Actually, for small extrinsic base spacing and moderate encroachment, the extrinsic base current does not significantly modulate the collector current at all [49]. For the above reasons, extrinsic base resistance must be traded off against device performance.

3.3.2.2 Conductivity Modulation (Webster Effect)

Simple bipolar device theory assumes that the number of electrons[1] injected from the emitter into the base is small compared to the doping concentration in the base. Under conditions of moderate current levels this assumption may hold, but under very high current levels, the concentration of injected electrons may actually become larger than the doping concentration of the base. This condition is compensated for by an increase in hole concentration in the base so that charge neutrality is maintained, and this high level injection situation is referred to as conductivity modulation, or the Webster effect. Increased injection levels increase the effective base doping, and base charge, and thus the performance of the device is compromised. The simplified expression below describes the maximum common emitter current gain [50]:

$$\beta_f = \frac{D_{nb} L_{pe} N_E}{D_{pe} W_B N_B} \tag{3.11}$$

From this equation, it is apparent that an increase in base doping, N_B, reduces the current gain of the device. Note that this relation does not take recombination of carriers within the base into account, thus, it would be expected that as effective base doping increases, increased recombination will result, and the current gain will decrease more than is predicted by the above equation.

The onset of the Webster effect is most discernible in a device possessing a very lightly doped base. Most modern bipolar transistors, which have been aggressively scaled, have moderately doped bases and thus this effect is masked by other factors.

---

[1] An NPN device will be considered.

3.3.2.3 Base Pushout (Kirk Effect)

Another phenomenon which accounts for reduced current gain at high levels of collector current is called base pushout, or the Kirk effect [51]. Normally it is assumed that the number of carriers reaching the collector is small compared to the doping level of the collector itself[1]. This assumption holds true at low to medium currents, however at high levels of current, there is a minority carrier charge buildup in the collector—base depletion region. When this occurs, the depletion region moves outward, away from the base and into the bulk of the collector. Depending on current level, and the epitaxial layer thickness and doping, the depletion region may migrate all the way to the buried layer. The net effect of this high current phenomenon is the effective widening of the neutral base region. As is indicated in equation 3.11, a larger base width results in a reduction in the device current gain.

The Kirk effect places constraints on the minimum acceptable doping in the collector, which is equivalent to background doping of the n tub. Devices which have been CIC scaled required substantial collector doping to retard the Kirk effect. Increased collector doping, however, increases the collector parasitic capacitance and lowers the breakdown voltages associated with it. Actually, it is these breakdown voltages which will place a limit on the maximum doping allowable in the collector.

3.3.2.4 Base width Modulation (Early Effect)

Simple bipolar theory suggests that the collector base voltage has no effect on the current gain. The reverse potential across the collector—base junction serves to collect carriers in the base which have been injected by the emitter and transported by diffusion. The reverse voltage can have controlling action when the transistor has a very thin base, or when the base is very lightly doped. Under these circumstances, the width of the depletion region extending from the collector into the base may become significant. The result is that the collector voltage causes a considerable variation in the depletion width which in turn changes the width of the active base region. The base width is effectively modulated by the collector voltage, which in turn results in a corresponding modulation of the collector current and causes the device to have a finite output resistance. The effect is modeled with a voltage value, referred to as the Early voltage, in the expression for collector current. If the slope of the collector current curves of a device in the forward active region of operation

[1] The doping here is sited as an indicator of the fixed charge concentration in depletion regions formed within the collector.

were traced backward, they would intersect at a single point on the V_{ce} axis which is defined as the Early voltage.

The Early effect places constraints on the minimum base doping level. In general, modern scaled bipolar devices have moderately doped bases, which tends to reduce this effect. A typical modern BiCMOS process would render NPN bipolar devices with an Early Voltage of 60V or greater[1].

3.3.2.5 Current Gain

The effective common collector current gain, or β, of a bipolar transistor is not constant with collector current, which is contrary to simple bipolar theory. At very low current levels, the base current is dominated by recombination in the base—emitter depletion layer, and thus the ratio of base current to emitter current rises, thus decreasing the current gain. At high levels of collector current, the Webster effect and the Kirk effect reduce the current gain of the device. In modern devices, with lightly doped collectors, the Kirk effect is by far the major contributor to gain roll off under high current conditions. The base Gummel number, Q_b, which is defined as the integral of dopant concentration in the base between the two space charge regions, sets the collector saturation current, as well as the peak current gain, or β_f. The latter relationship is:

$$\beta_f \propto \frac{1}{Q_b} \tag{3.12}$$

The required current gain of a bipolar device sets a limit on the maximum base doping, and it is traded off against base resistance, which decreases with increased doping. The minimum base doping, on the other hand, is set by punchthrough and Early Voltage requirements.

3.3.2.6 Punchthrough

In normal operating circumstances, the depletion regions associated with the emitter—base and collector—base junctions are separated by a neutral base region. Under high voltage conditions, however, the depletion regions may extend an abnormal length into the base and actually touch. Once this occurs, the emitter and collector are connected by a single

[1] This is the value for BATMOS, Northern Telecom's BiCMOS process [21].

depletion region, and a large current flows between the emitter and collector. This condition is referred to as punchthrough, and the base—collector voltage necessary to induce this effect is called the punchthrough voltage. The punchthrough phenomenon places constraints on the minimum width and doping level of the base.

3.3.2.7 Reachthrough

In modern bipolar transistors, the base doping is usually higher than that of the collector, and for this reason the depletion region associated with the base—collector junction extends further into the collector than the base. If collector voltage is increased, the depletion widths in both the base and the collector increase, with the latter width increasing to a larger extent. The collector—base voltage which causes the depletion edge in the collector to reach down to the buried layer is called the reachthrough voltage. Avalanche breakdown can readily occur under the above conditions. The reachthrough voltage places constraints on the minimum collector width and doping.

3.3.2.8 Breakdown Voltages

The different pn junctions inherent to the structure of a bipolar transistor present design constraints due to the respective reverse voltage levels at which they breakdown. The term breakdown is frequently used as a general term for junction failure due to several different mechanisms including punchthrough, Zener breakdown, and avalanche breakdown. The first mechanism has already been discussed. Zener breakdown is supported by a tunneling mechanism in which large numbers of electrons penetrate the potential barrier created and imposed by the bandgap of a semiconductor. Electric fields on the order of 10^6 V/cm [50] are necessary to provide carriers with the energy necessary to tunnel. In modern devices and voltage levels, this field is only likely to occur near pn junctions formed from very heavily doped semiconductor regions. These junctions will produce depletion region widths small enough to produce very high fields across them, but narrow enough so that kinetic energy sufficient for impact ionization cannot be attained[1]. In most practical transistors, Zener breakdown is not the chief mechanism of junction breakdown.

Avalanche breakdown, driven by impact ionization, is by far the most common mode of junction breakdown in modern bipolar transistors due to the smaller fields required for triggering. This means that the phenomenon can occur at junctions which are not as

[1] Typically, junction widths must be around 10 nm to support Zener breakdown [13].

heavily doped as those mentioned the previous case. High reverse voltages provide thermally generated carriers with sufficient kinetic energy to shatter the silicon—silicon bonds between lattice atoms and generate a free electron—hole pair. These carriers can, in turn, gain sufficient energy to engage in a similar collision, thus creating a multiplicative effect which generates a large number of carriers and a correspondingly large reverse current. This process is known as avalanche multiplication.

The collector—base breakdown voltage, BV_{CB0}, for a bipolar transistor, with the emitter open circuit, is given by [50]:

$$BV_{CB0} = \frac{\varepsilon_s E_{br}^2}{2qN_C} \tag{3.13}$$

It is not strictly true that the breakdown voltage is inversely proportional to the collector doping, since E_{br}, the critical field, varies slightly around 2×10^5 V/cm [16] with the doping concentration. Note that the equation is a modification of that for a one sided pn junction [13], with N_C being substituted for the impurity concentration of the lightly doped side. The BV_{CB0} value places a constraint on the maximum doping concentration of the collector, which means that breakdown voltage considerations are traded off against collector resistance.

The collector—emitter breakdown voltage, or common emitter mode breakdown voltage, denoted by BV_{CE0}, is approximated by the expression [50]:

$$BV_{CE0} \approx \frac{BV_{CB0}}{\sqrt[n]{\beta_f}} \tag{3.14}$$

For modern NPN devices, β_f has a typical value of 100[1], and n can vary between 3 and 6 [50], thus BV_{CE0} is significantly lower than BV_{CB0}. An explanation of this is given in [50]. Expression 3.14 highlights an important design constraint. Since the common emitter breakdown voltage is inversely proportional to the common emitter current gain, it is very difficult to simultaneously achieve both high gain and high breakdown voltage in a bipolar transistor. BV_{CE0} sets the minimum epitaxial layer thickness for a given doping level.

---

[1] Northern Telecom's BiCMOS process renders NPN devices with b_f = 96.78

3.3.2.9 Polysilicon Emitters

Polysilicon was first used as a bipolar emitter in 1976 [52], and the fabrication processes involved with polysilicon and polysilicon contacted emitters has been discussed earlier. Major advantages associated with polysilicon emitters include their compatibility with self aligned processes, their suitability in forming shallow emitter—base junctions, and the high levels of common emitter current gain which are attainable. The physical processes which occur at the polysilicon—silicon interface and grain boundaries in the polysilicon are not yet fully understood, but they are crucial to emitter operation. For this reason, they are very difficult to model. The chief advantage of poly and poly contacted emitters over metal contacted ones lies in the reduction of base current through the reduction of hole current which is injected into the emitter[1]. This will be discussed as well as other issues of emitter operation.

The polysilicon—silicon interface of a poly or poly contacted emitter strongly influences current gain of the device. There are two main processing techniques involving this interface which are used to increase device current gain, with one involving oxide growth and the other involving interface cleaning. The former can increase gains by a factor of 10 or more, while the latter affords gain increases of 2 or 3 [50]. Within this section, the term poly emitter will be used in reference to both polysilicon and polysilicon contacted emitters.

No matter how a poly emitter is formed, there is always a very thin layer of oxide at the interface between the silicon and polysilicon. In one type of processing, a high quality, ultra thin oxide with a thickness in the neighborhood of 10 Å [50] is purposely grown on the poly—silicon emitter interface. The silicon dioxide has a wider bandgap than silicon and thus the interface causes an energy barrier to both electrons and holes. Carriers cross the interfacial oxide by quantum mechanical tunneling, and holes are inhibited more than electrons due to the fact that they are cooler[2] in the electric fields present. This has the effect of both reducing base current and increasing emitter resistance. The latter effect is only serious at high current levels, while the reduction of base current results in very significant gain increases. The effects of the oxide are limited by such things as emitter thickness and doping, and the configuration of emitter surface states. In regards to the former, a shallow emitter is necessary for the oxide to be effective. This is because with large emitter depths, holes recombine before reaching the interface, thus rendering its hole

---

[1] NPN devices are considered in this section.

[2] They attain less energy due to a shorter mean free path.

barrier qualities ineffectual. The emitter depth, therefore, must be small with respect to hole diffusion length.

Another method which increases bipolar gain involves the chemical cleaning of the silicon surface before polysilicon deposition. Emitter depth is the sum of the emitter junction depth, and the poly thickness, which means that the transport properties of both materials influence the base current. This is in contrast to a metal contacted emitter where recombination of injected holes takes place almost exclusively at the metal—silicon interface. In a shallow emitter device, where the depth is small compared to the hole diffusion length, a linear hole distribution in the emitter is formed. In a poly emitter device, the holes are not forced to recombine at the polysilicon—silicon interface, since they may diffuse into the poly and combine there. For this reason the concentration gradient of minority carriers in the emitter is much smaller in the case of a poly emitter. Since the minority hole flow in an emitter is a diffusion driven process, the reduced gradient will amount to a reduced hole flow, and thus a reduced base current. Key to the realization of the above situation is a very clean interface, called nominally clean, prior to poly deposition, and a hydrofluoric acid (HF) etch is usually used. If the interface is not cleaned properly, heavy recombination will occur at the interface and the advantages of the poly emitter will be degraded. Grain boundaries, and the polysilicon—silicon interface which acts as a pseudo grain boundary, contain large defect density concentrations, as well as dangling bonds, and thus act as recombination centers for minority holes. Also, the boundaries can block hole transport due to the effective reduced hole mobility in their vicinity. This is a major concern in devices with nominally clean interfaces.

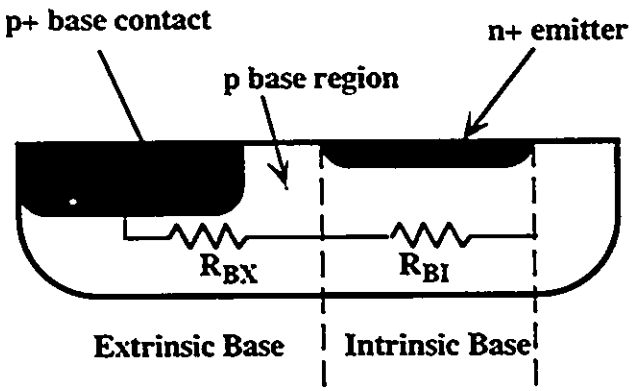
The increased current gain provided by poly emitters allows a tradeoff to be made with some other parameter, such as base doping. Increased base doping can be accomplished, thus reducing the base resistance, without the gain penalty which normally would be exacted in a metal contacted emitter device.

3.3.2.10 Parasitic Resistances

The parasitic resistances inherent to the bipolar device structure have an effect on the performance of the device, and their minimization is of key importance. Resistances associated with the base, emitter, and collector have different factors effecting them.

The base resistance is one of the most important factors effecting bipolar operation. In digital circuits, it limits the rate at which the base associated capacitance can be charged and discharged. Base resistance is made up of an intrinsic component and an extrinsic component. The intrinsic portion is comprised of the resistance of the active base region directly under the emitter, while the extrinsic part is comprised of the resistance between the edge of the active base region and the base contact. These two resistances are illustrated in Figure 3.7. The total base resistance is very dependent on device geometry, and may be non-linear due to such things as current crowding under the emitter [50]. Reduction of base resistance can be accomplished by multiple contacts and increased doping. The negative aspect of these solutions is larger device area, and higher base associated capacitances, respectively.

Figure 3.7: Concept of Intrinsic and Extrinsic Base Resistance



The emitter resistance depends heavily on the type of emitter being considered. Strictly speaking, metal contacted emitters have very low resistances associated with them, however, other degrading qualities make the low resistance an unacceptable compromise. As mentioned earlier, metal emitters tend to induce large hole currents due to the concentration gradient internal to the emitter, set up by the forced recombination at the metal—silicon interface. Polysilicon emitter resistance is very complex to model for reasons discussed previously. It depends on such things as interfacial oxide layer thickness, electron effective barrier height, and doping concentration. Advanced processes such as silicidation can reduce poly emitter resistance significantly.

Collector resistance can be considered the sum of the epitaxial resistance below the active transistor, as well as the resistance of the buried layer and deep contact if they are present.

It has significant effects in high current circuits, including conventional BiCMOS gates under a heavy capacitive load.

3.4 Bipolar and CMOS Devices in a BiCMOS Process

One of the goals in designing a BiCMOS process flow is to provide high quality devices of both MOS and bipolar types, and in fact, this is what separates BiCMOS from CMOS. The process flow may be optimized more towards a certain device type, which is generally indicative of the parent process from which the BiCMOS version sprung. For example, in early BiCMOS technologies the relative quality of the MOS devices was much higher than that of the bipolar device. This was due to the fact that BiCMOS development has essentially been driven from a CMOS processing base.

There are process goals common to the realization of both types of devices, and many process based performance improvements for one type also have a positive impact on the other. Minimization of diode capacitance is beneficial to both MOS and bipolar devices. This capacitance is strongly associated with the n+ and p+ source and drain regions in the MOS device as well as the junctions comprising the BJT. Buried layers accommodate low collector resistances, and improve latchup immunity of the process as well. Silicidation of poly and diffusions provides low resistance contacts for both types of devices. Sidewall oxide spacer technology allows LDD structures to be formed which reduce hot carrier generation in MOS transistors. Additionally, this technology allows self aligned emitters and extrinsic bases to be formed, which reduce bipolar parasitic capacitances, and shrinks the device's footprint. Advanced isolation techniques such as trench isolation allow high device packing factors, and also improve latchup immunity. Thus, there is some symbiosis in BiCMOS process design.

PNP bipolar devices are uncommon in present BiCMOS process flows. The quality of the vertical device which is rendered is inferior to that of the NPN type due to processing difficulties which manifest themselves as low current gain, high collector resistance, and high base transit time. Lateral PNP devices are possible, but they also suffer from poor performance. Realization of PNP devices adds considerable complexity to the already

complicated BiCMOS process flow. Development of vertical PNPs is already gaining momentum[1] driven by their demonstrated utility in low voltage BiCMOS circuits [36].

High performance BiCMOS requires the presence of an epitaxial layer. Specification of this layer, along with buried layers, such that acceptable MOS and bipolar devices are rendered is a very complex optimization problem, since the strongest coupling of BJT and FET characteristics occurs at the front end of the process. A large number of design tradeoffs must be made between process parameters. The specification of these steps even for one type of device involves a large number of compromises. For example, to minimize bipolar collector resistance, a thin, highly doped epitaxial layer would be desirable. However, to reduce the effects of junction capacitances, and breakdown voltages, a thick, lightly doped epitaxial layer would be beneficial. In MOS devices DIBL, as well as other short channel effects benefit from a highly doped, thin epitaxial layer. Junction breakdown, source/drain capacitance and body effect considerations, however, mandate a lightly doped, thick epi. Bipolar devices will also set a minimum thickness due to BV_{CE0} requirements. As well, a minimum thickness is mandated by NMOS requirements, since diffusion up from the p+ buried layer underneath the p well causes an unacceptable increase in body effect and junction capacitances. Too thick an epi layer will decrease the f_T of bipolar devices and increase collector resistance. Buried layer p+ implant dose is determined by collector to collector punchthrough spacing requirements, while the n+ buried layer is usually doped to the solid solubility limit to minimize collector resistance and latchup susceptibility. As a general rule of thumb, for CMOS devices, a thicker lightly doped epi is preferable, while for bipolar devices, a thinner higher doped epi offers better performance [53].

Optimization of the doping profile of the respective wells, including background doping and surface implants, is at the heart of BiCMOS process design. It is here that most compromises will be made. The actual well doping profiles depend strongly on the choice of process technology, i.e. n well—p substrate, p well—n substrate, or twin tub. Separate background, channel stop, and channel threshold adjust implants add a great deal of flexibility and allow a higher level of device optimization. For example, a separate subsurface implant to stop drain—source punchthrough in NMOS transistors will effectively decouple the DIBL and junction capacitance issues. In a modern BiCMOS process, typical background (well or epi) doping lies between 1×10^{16} and 3×10^{16}

[1] A PNP device is under development for the BATMOS process.

$atoms/cm^3$, while typical channel implants render surface doping concentrations as high as 3×10^{16} $atoms/cm^3$ [53].

Processing involved with BJT base formation is largely independent of CMOS design concerns. The base doping profile, determined by the energy and dose of the base implant, will be a compromise between base resistance, device f_T, emitter—base capacitance, and current gain demands. Early voltage requirements will place a constraint on the minimum acceptable base doping.

Finally, it is very important to control the thermal cycles within the process flow. High temperature cycles after several critical implants are usually undesirable because of the uncontrolled dopant diffusion which will take place. Long duration, high temperature, process steps are confined to the front end of the process. This is when the main character of the well doping profile is formed. Wells can be retrograde, with increased doping as depth increases. This is accomplished by performing a high temperature cycle, called a drive, which induces diffusion of dopants in the buried layer up into the well. This is an important technique in a BiCMOS process since it offers benefits of both a lightly and heavily doped epitaxial layer. Lower regions are highly doped, providing low collector resistance, but upper levels are lightly doped, providing low junction capacitances within the MOS devices. Retrograde "V" profile wells are formed in a similar fashion, along with an additional surface implant. This provides heavy but shallow subsurface doping which is sufficient to control short channel effects, but does not appreciably increase the source and drain island capacitances. As well, the vertical diffusion provides low collector resistance, and improves latchup immunity. Thus, a great deal of the character of both types of devices in a BiCMOS process is determined in the first few steps of the processing.

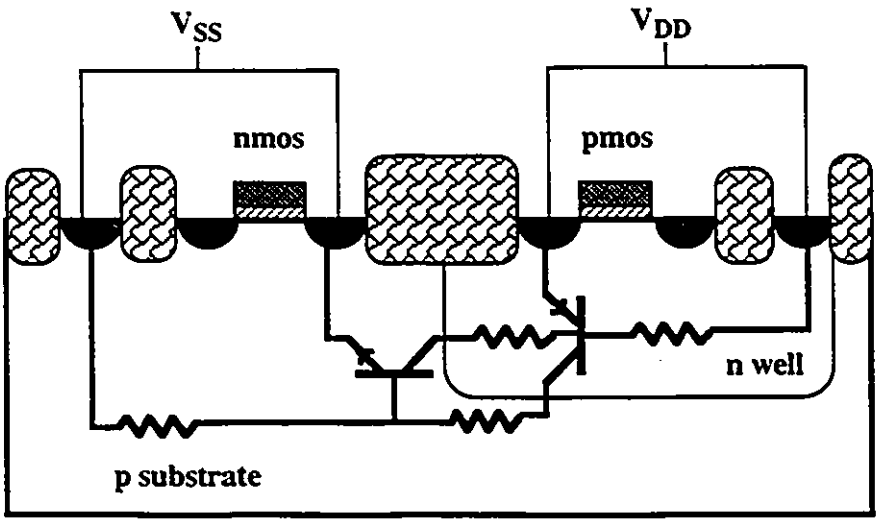
3.5 Latchup

Latchup is a phenomenon which occurs in CMOS technology as a result of positive feedback between parasitic bipolar devices, and it has been well studied [54], [55], [56], [57]. It was discovered early in CMOS development, and continues to be an important factor in design. BiCMOS technology presents a number of new concerns which are not present in CMOS technologies, and thus must be addressed.

The phenomenon of latchup arises from regenerative feedback between a parasitic NPN transistor and a parasitic PNP device. These parasitic devices are present in any CMOS

process, and are illustrated for a simple n well technology in Figure 3.8 which shows the source of a PMOS transistor connected to V_{DD}, and the source of an NMOS transistor connected to V_{SS}. This is a very common circuit connection in CMOS. The wells are connected to the supplies as shown, but unfortunately, this sets up a feedback loop involving the parasitic BJTs. The lumped element model for latchup is indicated in the figure, and is redrawn in Figure 3.9. As is apparent, if either one of the bipolar bases is forward biased, collector current will flow. This will bias the other parasitic bipolar device by causing a voltage drop to develop across the resistance present, and that device will start conducting current which reinforces the bias on the first. Thus, the process is regenerative, and the power and ground rails are effectively shorted. The resulting condition is called latchup, and it can quickly destroy the integrated circuit through overheating. Any one of several inducements can cause the initial forward biased condition including spurious noise, voltage overshoots, static discharges, and signal application before power up. It is interesting to note that before latchup and its prevention was fully understood in the early days of CMOS development, it was of paramount concern, since integrated circuits which self-destructed for no apparent reason were not of much utility. Fortunately, the phenomenon is now well characterized, and prevention methods are integrated into design practices and fabrication processes.

Figure 3.8: Latchup in an N Well CMOS Process

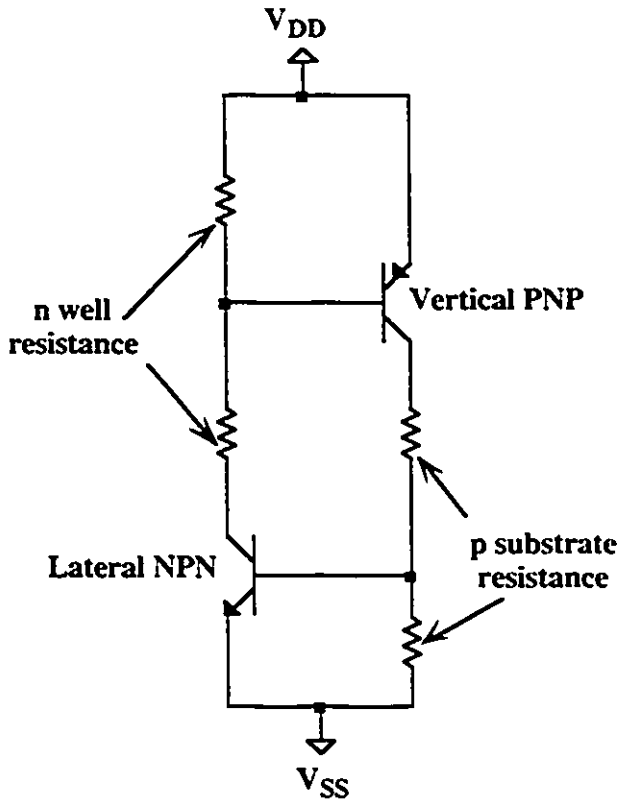


Aggressive scaling is common in modern semiconductor technologies to provide higher packing density. This means that the space between devices has become very small and consequently a reduction of the effective base widths of the parasitic devices has resulted,

providing correspondingly higher gains. This increased gain more than offsets reduced resistances due to reduced current path distances. Actually, there is also the tendency of resistance to increase due to a more restricted current path in a shallower well. Scaling, in general, tends to increase latchup susceptibility [41].

BiCMOS technology is generally considered more resistant to latchup than CMOS [31]. This seems counter-intuitive considering the fact that there are several latchup modes which arise exclusively in BiCMOS technologies. These situations are due to the well structures and the presence of high quality bipolar devices. The fact that BiCMOS is latchup hard despite these facts can be attributed to advanced fabrication process design.

Figure 3.9: Lumped Element Model For Latchup

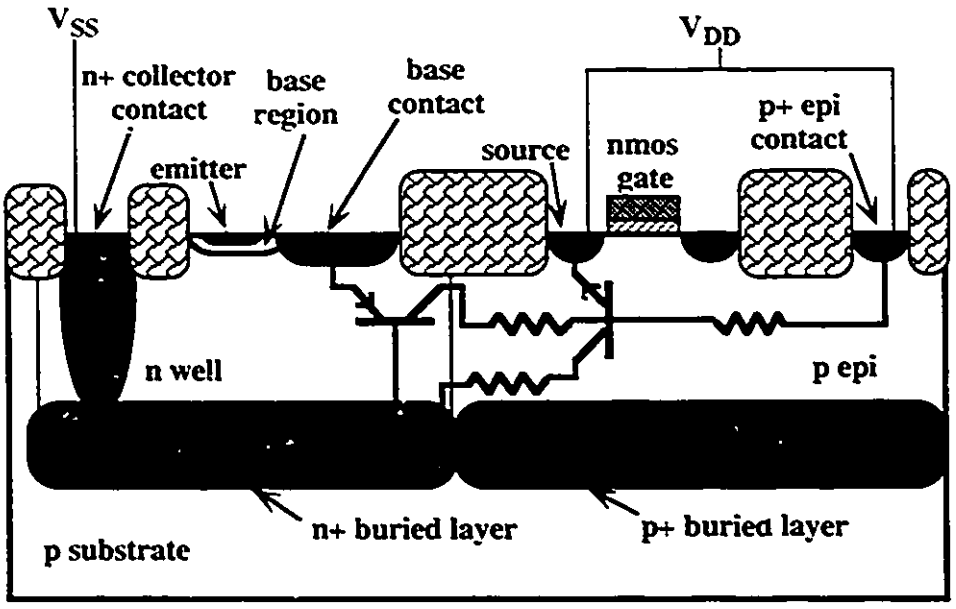


The presence of an indigenous NPN device (the "Bi" in most "BiCMOS" processes) can create conditions which are favourable for latchup. If the collector of this device debiases at high current, and saturates, the collector—base junction becomes forward biased and large numbers of carriers are injected directly into the well. This provides a source of base current which could easily turn on a parasitic vertical PNP device. In a modern process,

the presence of a buried n+ layer would cause virtually all of the injected minority carriers to recombine before reaching the p substrate collector, thus eliminating the possibility of vertical PNP devices triggering latchup.

Figure 3.10 portrays a situation which could occur in modern BiCMOS processes. Notice the large, highly doped extrinsic base contact, the deep n+ collector contact, and the presence of two buried layers. The figure illustrates a structure which could bring about latchup. Both bipolar transistors involved are parasitic, and both are lateral devices. If, in the course of circuit operation, V_{DD} is applied to the base of the indigenous NPN transistor and it saturates, injecting minority carriers into the n well, latchup could be triggered. In BiCMOS logic circuits, BJT saturation is a common occurrence, and this source of carriers is not normally present in a CMOS technology.

Figure 3.10: Latchup in an Advanced BiCMOS Technology

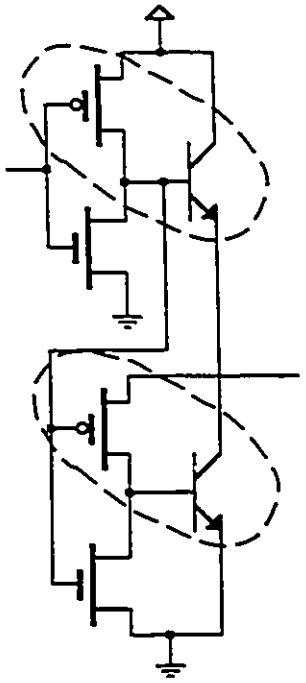


BiCMOS offers the possibility of merged bipolar and CMOS devices [58], [31], [59]. These merged devices provide reduced area due to well and contact sharing, as well as increased performance through a marked reduction of parasitic capacitances. Figure 3.11 illustrates where these devices might be used in an inverter circuit design [58]. Unfortunately, these structures are also much more susceptible to latchup due to device proximity and lack of isolation [60]. Figure 3.12 illustrates conceptually the structure of one of these BiPMOS devices. The latchup action takes place between the indigenous high

quality NPN and the parasitic lateral PNP formed under the PMOS transistor. This latter device is also of fairly high quality due to the fact that the short gate lengths of modern technologies effectively provides a thin parasitic base region.

Residence in the same well of both devices provides extremely close proximity, and thus carriers injected by one structure have a short distance to travel before they can influence the other. When the PMOS transistor turns on, the p base of the NPN is pulled high, base current is sourced to the bipolar, and collector current begins to flow. Some of this current may flow in through the well contact, and thus the source—well junction, or the parasitic emitter—base junction, may become forward biased, turning on the lateral PNP. Once this occurs, it sources current to the base of the NPN, thus increasing the collector current drive. Under this condition, the circuit is latched, and even if the PMOS transistor is shut off, the NPN continues to conduct. When a low value is applied, the PMOS transistor is turned off, the lower BJT in Figure 3.11 is turned on, and the power rails are effectively shorted. Careful layout of this structure can go far in avoiding latchup in this circumstance. By sourcing current to the PMOS transistor via the collector electrode, which effectively restricts current flow to a specific path within the well, it has been shown that latchup can be avoided [58].

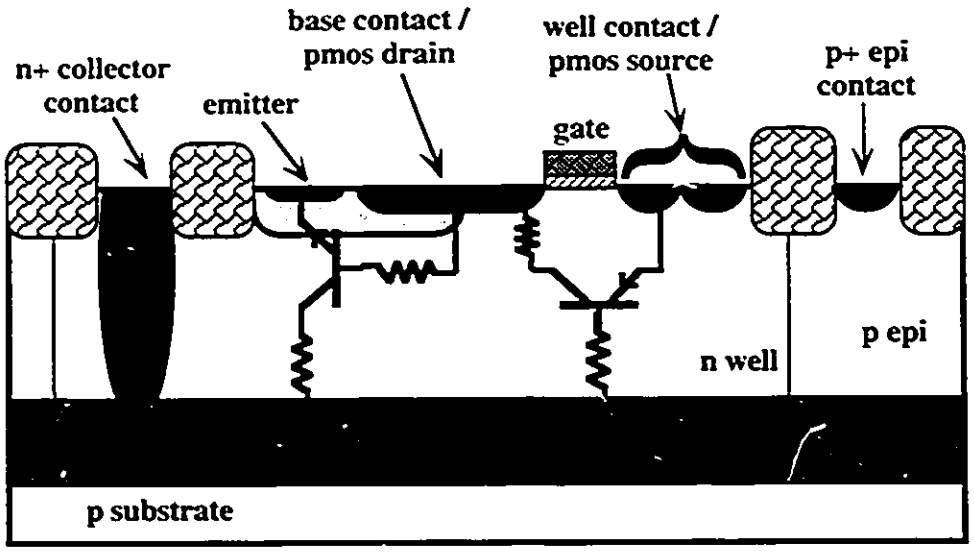
Figure 3.11: Merged Devices in a BiCMOS Buffer



Other well structures which are prevalent in BiCMOS technologies are shown in Figure 3.13. The first and second have buried n+ layers under the p well, and both wells respectively, while the third has self aligned buried n+ and p+ layers. The first two structures are common in technologies which must implement ECL or low voltage TTL circuits, while the latter is representative of more mainstream high performance processes. It has been shown that isolated p well structures such as the first two in the figure are much more susceptible to latchup [15]. This is due to the fact that these structures tend to increase parasitic device gain and increase bulk resistances (base shunting resistance) due to an effectively thinner well. The depletion region extends much further into the light to moderately doped well than into the heavily doped buried layer, thus decreasing the actual

active well depth. The third structure illustrated has excellent latchup hardness. The buried layers here are of the same doping type as the wells, thus they considerably reduce the well/bulk resistance, as well as the parasitic transistor gain by forcing recombination of minority carriers before they can induce activity in a parasitic device.

Figure 3.12: Structure Of Merged Device and Associated Latchup Paths

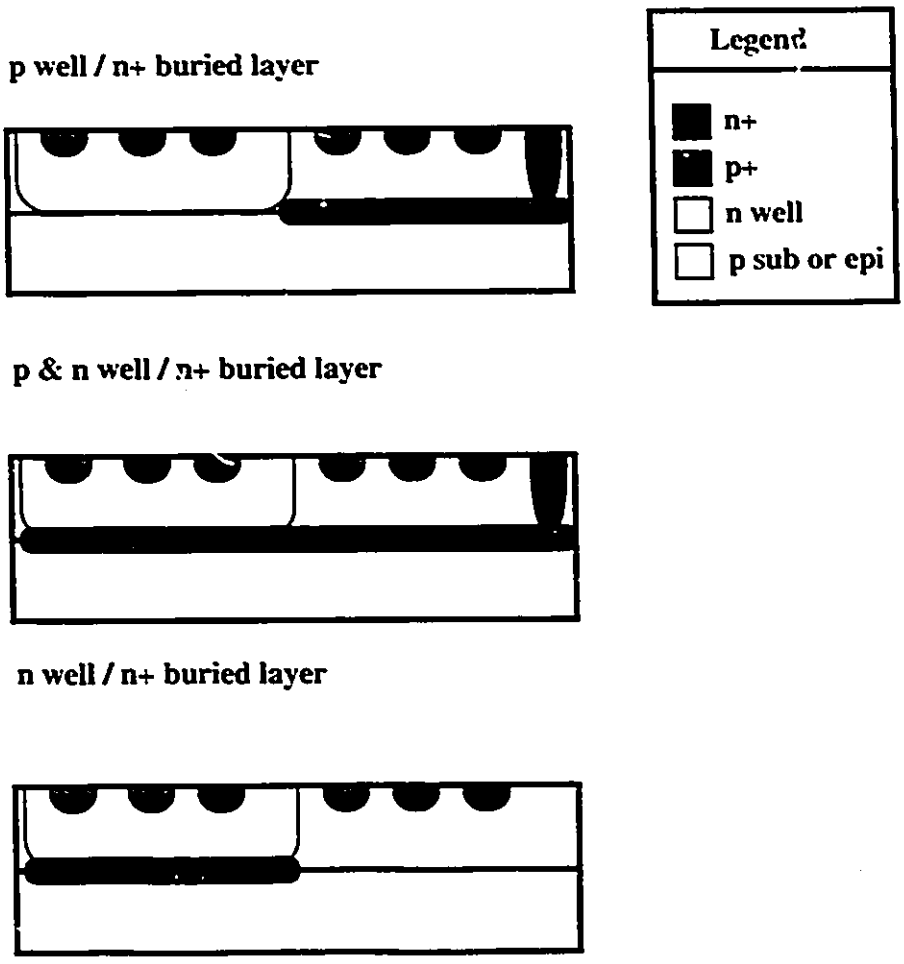


Latchup requires that three conditions be fulfilled: (i) sufficient voltage must be sustained across the emitter—base junctions of both of the parasitic bipolar transistors, (ii) the loop gain of the parasitic device structure must be greater than unity in order to achieve regeneration, (iii) the bias supply must be capable of sourcing current sufficient to sustain the latchup state. The first point can be addressed by reducing the parasitic resistances of concern, while the second point can be treated by reduction of the individual ß values for the respective parasitic devices. The last point depends on the circuit design, layout, and technology. Thus, addressing any of these three areas individually will decrease overall susceptibility to latchup.

Several structures, called guard rings, are useful in latchup prevention. Majority carrier guard rings reduce bulk resistance and capture majority carriers before they reach parasitic devices. They consist of rings of n+ diffusion around the interior of n wells, and rings of p+ diffusion around the interior of p wells. In an n well—p substrate technology, they would be manifested by a ring of n+ around the inside of the n wells, and a ring of p+ around the exterior of the n well, separated by a small distance. The n+ ring is connected

to V_{DD}, while the p+ ring is connected to V_{SS}, accomplishing two things. They reduce the effective bulk resistance by providing a high area, low resistance contact to the well and the substrate, and thus reduce potential variations over the bulk. Also, they provide a sink for excess majority carriers, thus reducing the ß of the parasitic transistors. Minority carrier guard rings, on the other hand, consist of rings of n+ diffusion in p wells or rings of p+ diffusion in n wells which are connected to V_{DD} and V_{SS} respectively. These rings, with their associated reverse biased junctions, serve to collect minority carriers before they can reach a parasitic device, thus reducing the effective gains of those devices. Guard rings have been shown to be less effective in preventing latchup initiated by vertical devices, as opposed to lateral parasitic transistors [61].

Figure 3.13: Well Structures In Advanced BiCMOS Processes



This is due to the fact that lateral devices are activated by surface currents, while vertical device currents tend to occur deeper in the substrate. Other measures to prevent latchup

include design rules which mandate regular substrate and well contact spacing[1] to maintain uniform bulk potential, as well as split or butted contacts which assist in achieving the same goal. Other, more exotic methods of latchup prevention include the reduction of minority carrier lifetime in the substrate by neutron irradiation or gold doping. This effectively lowers the ß of the parasitic devices.

Recently, an interesting approach to latchup prevention was reported [62] which addresses point (iii) in the above discussion. It utilizes floating wells, and deep p+ diffusions. Essentially, the well is not hard wired to power or ground and is allowed to float. This allows the well potential to follow the transient, thus avoiding the activation of the parasitic vertical bipolar. Even if the bipolar is turned on, it will only have a brief, and limited supply of base current. This technique is aimed at a specific CMOS/DMOS[2] technology, and is not broadly applicable.

Although the methods of latchup prevention implemented by physical layout design are very effective, the reason BiCMOS is highly resistant to latchup lies fundamentally in the process design. Tradeoffs are made to provide latchup resistance. A higher well doping may not be optimal for PMOS performance, but may be necessary to reduce well resistance. Buried layers can cause autodoping complications in well profiles, but are probably the single most effective latchup prevention feature with which a process can be equipped. Thus careful process design has produced BiCMOS technologies which are more latchup resistant than standard CMOS technologies, despite a number of factors which seem counter-intuitive.

3.6 Summary

This chapter has discussed many issues dealing with device scaling, second order effects in scaled devices, process factors used to control these effects, BiCMOS process tradeoffs, and latchup. Five different methods of scaling were mentioned. Three were associated with MOS devices and two pertained to bipolar devices. Many second order effects which are of concern in scaled BiCMOS technologies were discussed. Both types of devices are present, thus both types were considered. Some of the MOS issues discussed included channel length modulation, threshold voltage shift, subthreshold current, velocity

[1] See BATMOS design rules: numbers 32.0-32.2 [22]

[2] DMOS stands for Double-diffused MOS, a structure commonly used to implement power transistors with high breakdown voltages.

saturation, mobility degradation, the effect of very thin gate oxide, source and drain resistance, drain induced barrier lowering, and hot carrier effects. Bipolar device issues discussed included extrinsic and intrinsic device concerns, the Webster effect, the Kirk effect, the Early effect, punchthrough, reachthrough, breakdown voltages, polysilicon emitters and parasitic resistances. Tradeoffs which must be made in a BiCMOS process to render both types of devices were mentioned, and finally, a treatment of latchup and its prevention in CMOS and BiCMOS was given.

Chapter 4

HIGH PERFORMANCE ARITHMETIC CELLS

4.1 Introduction

Modern microprocessors are increasing in complexity every year, driven by a demand for higher and higher levels of performance. Scientific software applications demand very fast arithmetic operations since much of the run time is spent solving equations numerically. These types of computations invariably involve a large number of additions and multiplications, and thus architectures have been introduced to speed up these operations. Multiplication takes much longer to execute than addition, which is intuitively correct since the multiplication operation is simply a series of many additions. To address this issue, more and more general purpose processors are including hardware multipliers in their architectures [63], [24], [5], [64]. Historically this was implemented on a separate chip called a math co-processor, which was controlled by the main CPU. In recent years, the trend has been more towards including the multiplier on the same die. This is a practice which has been common in the design of special purpose digital signal processors [65] for many years[1]. Interest in the design of hardware multipliers and adders has increased greatly in recent years.

This chapter will give a survey of various types of multipliers and adders, and describe several high performance macrocells which were implemented in a .8μ BiCMOS technology. These cells were rendered as macrocells, and were designed in partial fulfillment of the CMC/Micronet contract. The two main goals of the contract work were to provide high level VLSI primitives for integrated system design, and to test and verify the viability of the BiCMOS Edge™ design environment. The final implementation of these

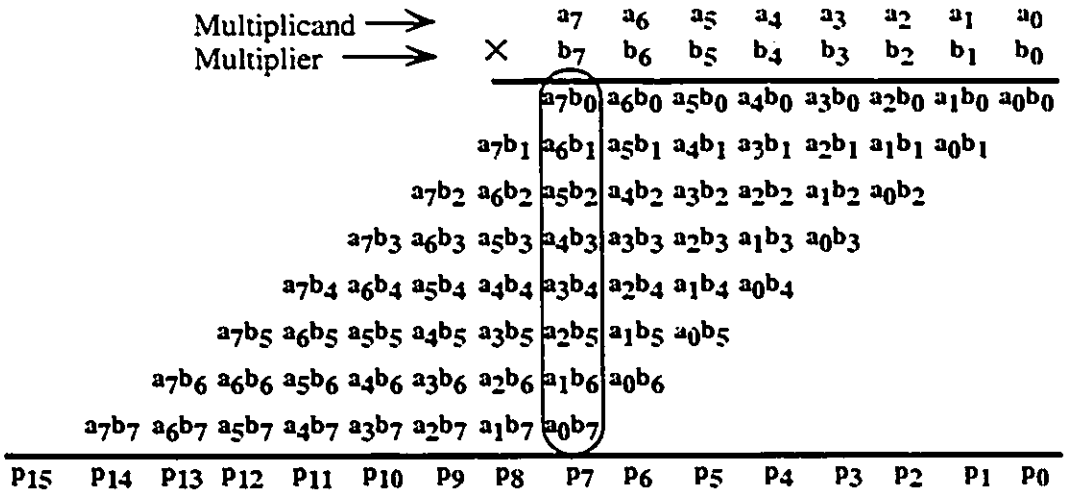
[1] The term special purpose processor is used loosely, since virtually all commercial DSP processors are still essentially general purpose processors with some hardware added to speed up certain operations.

designs within the Edge™ design framework demonstrates that a non-trivial real world design process can be carried out within this environment. Finally, a chip will be described which was submitted for fabrication.

4.2 Multipliers

Multiplication consists of two main steps, termed partial product (PP) formation and PP accumulation. The multiplication process between two binary numbers is illustrated in Figure 4.1. PP formation, in the case of binary numbers, consists of a simple logical AND of the multiplier[1] digits and the multiplicand. Each successive row, comprising a unique summand, is shifted to the left so that each column of digits has the same binary weight, as indicated in the diagram. The majority of time in the multiplication process is occupied by the summation of these partial products. Many different schemes have been developed to streamline this portion of the algorithm, including sequential and parallel methods.

Figure 4.1: The multiplication process

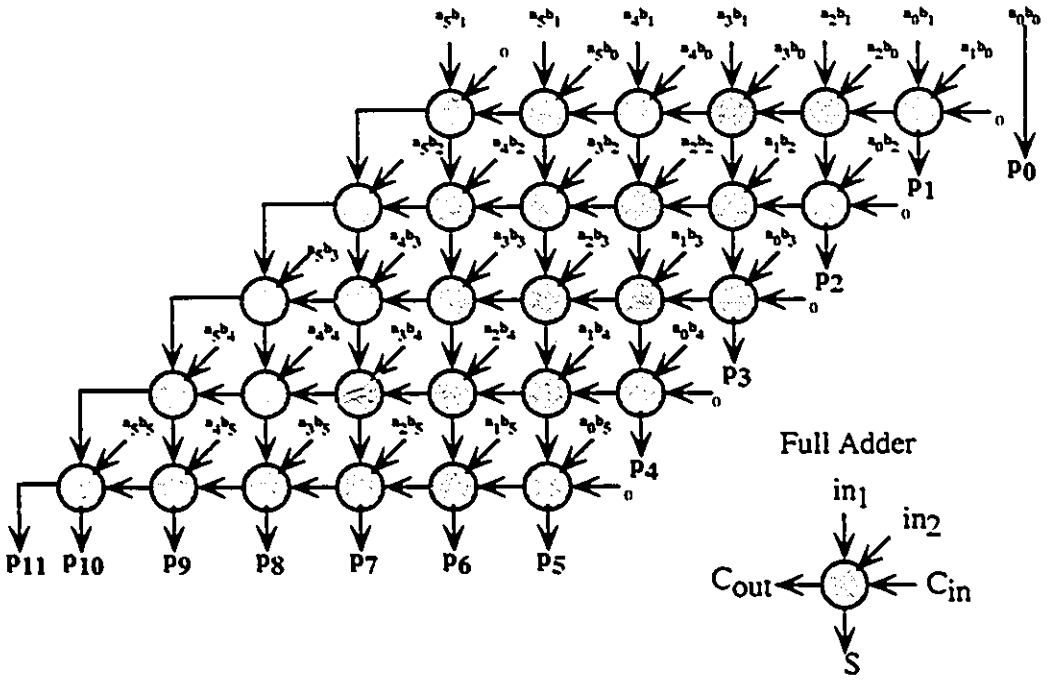


Sequential, or shift—add multipliers perform the multiplication operation with a minimum of hardware, however, performance suffers as a result. For two n bit operands, n partial products are generated and accumulated sequentially, one PP per clock cycle. The multiplier bits are processed, one per clock cycle, by performing a logical AND with the

[1] Henceforth, the word "multiplier" in underlined text will refer to the first operand in the multiplication operation, while the word "multiplier" in normal text will refer to the architecture and/or hardware which performs the operation.

multiplicand, thus producing a new partial product. The result is accumulated in a register twice as wide as the operands. The accumulated result is shifted once, the next multiplier digit is shifted in, and the process is started again. This method is very slow compared to other methods, but it is hardware efficient. Several techniques have been developed to speed up the architecture [66], and more recently [67]. They essentially annex a second adder onto the architecture which performs a fast addition in the latter clock cycles to improve overall performance. Although this provides significant increase in speed, it still falls short of the performance of parallel architectures.

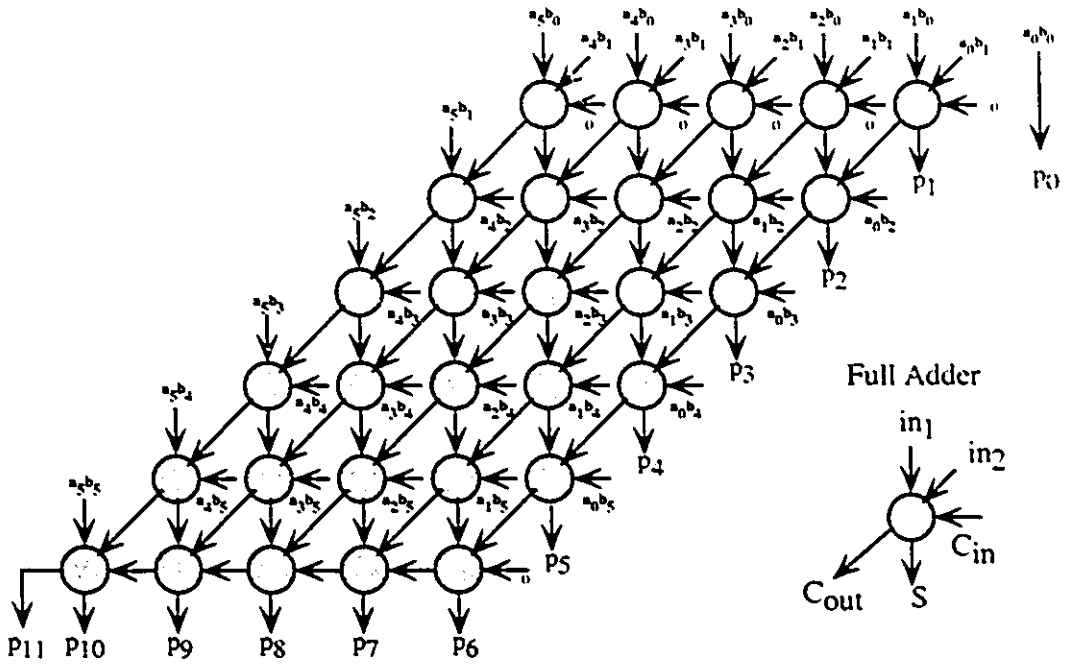
Figure 4.2: Simple Array Multiplier



Parallel multipliers evolved out of a need for faster computations. Partial products are summed in parallel, thus offering a significant speed increase. Consider the array pictured in Figure 4.2, which performs the multiplication of two six bit operands. It consists of 5 rows of full adders (FAs), each row forming a ripple carry adder. This array performs accumulation of the partial products which originate from the PP formation circuitry, typically called the AND plane. The critical path of this circuit is dominated by the carry propagation delay along each row of adders. A great improvement can be made by connecting the adders in a carry—save fashion, as illustrated in Figure 4.3. In this style of connection, each row of carry—save adders (CSAs) sums up one additional partial

product, but carries are fed into following stages instead of into the next adder in the same stage.

Figure 4.3: CSA Array Multiplier



This architecture has been implemented in [68]. Since the accumulated PPs are kept in carry—save form, embodied by a separate sum and carry vector, there is no carry propagation in the traditional sense. Note that a ripple carry adder is necessary to absorb the sums and carries in the final stage. Recently an interesting method of reducing this delay overhead has been proposed [69], where the delay associated with the conversion from carry—save form is overlapped with the delay associated with carry—save vector formation. The critical path of the multiplier in Figure 4.3 lies along the right and bottom edge, equating to a delay of $2(n-1)$ CSA delays[1]. Multipliers which are comprised of many identical cells and have a very regular structure and connectivity are sometimes called array multipliers, however, the term is not well defined in the literature. These types of multipliers have been very popular for VLSI implementations due to their regularity in both layout and interconnection. They offer a significant performance improvement over shift—add styles, but their disadvantages include a delay and component count which are linearly

[1] Mistakes are noted in [70] and [71] which report the delay of this array as 2n-1.

dependent on operand length. This becomes a problem when implementing multipliers for large operand sizes.

Many improvements have been made to the basic linear parallel array described above. Nakamura [71] has proposed several iterative array architectures, including a radix 4 multiplier, a (6,3,4) counter multiplier, and a 5-counter multiplier. The first one decreases the number of cells needed in the CSA array by performing the PP accumulation within a higher radix number system. If we express the radix as 2^r, incrementing r by 1 reduces the number of adders required by ¼, while the delay is reduced by ½ . This is based on the assumption that the delay is the same for a full adder regardless of the radix, which is not a realistic assumption. The hardware complexity of higher radix adders becomes prohibitive as the radix is increased. The other iterative architectures use elements called counters and compressors.

Counters are circuits whose outputs represent the weighted binary number of values equal to a logical one on the input. In other words, they count the number of ones in the input vector. In general a (c,d) counter has c inputs of the same weight and produces d weighted outputs, obeying the following constraint:

$$2^d - 1 \geq c \tag{4.1}$$

Equivalently, the maximum number representable with the output bits must be at least as large as the number of inputs to the counter. A full adder is sometimes called a (3:2) counter. Multi-input counters are also possible, as demonstrated in [72], and they can have inputs of different weights. A $(c_{k-1},c_{k-2},\ldots,c_0,d)$ counter has $\sum_{i=0}^{k-1} c_i$ inputs and d outputs. The input bit groups, with in_{ij} representing input j of group i, have a weight of 2^i, and the counter output is described by:

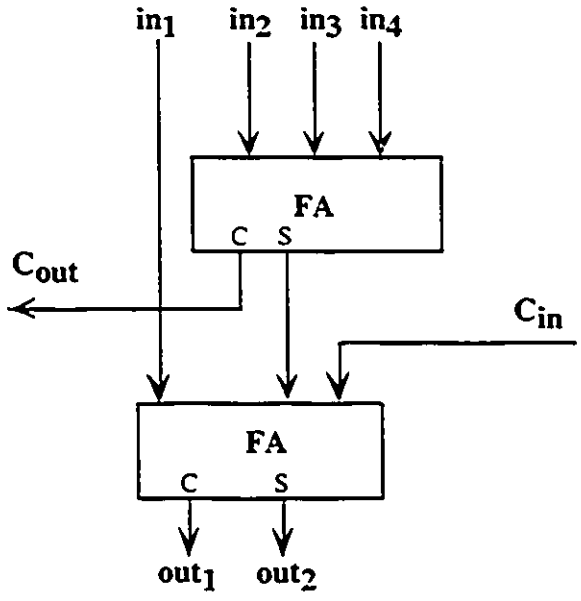
$$v = \sum_{i=0}^{k-1} \sum_{j=0}^{c_i-1} in_{ij} 2^i \tag{4.2}$$

Because d is the number of output bits, the following inequality should hold:

$$2^d - 1 \geq \sum_{i=0}^{k-1} c_i 2^i \tag{4.3}$$

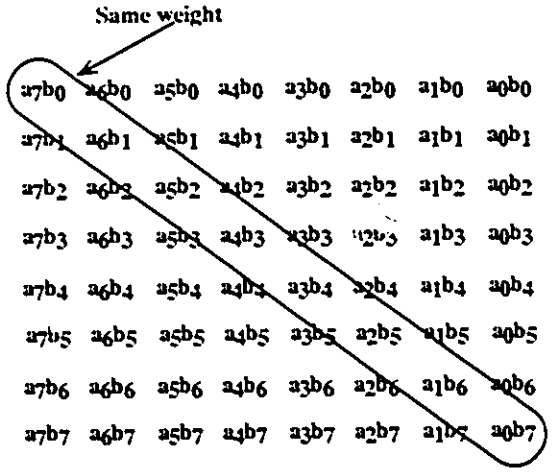
If the constraint for counters given in equation 4.1 is violated, the resulting circuit is sometimes called a compressor. For example, (4:2) compressors, or (4:2) adders as they are sometimes called, have been effectively used in several multiplier realizations [73], [74], [75]. They are so named because of the manner in which they "compress" 4 equally weighted input bits into two weighted output bits. This is not strictly true, since technically, a (4:2) compressor has five inputs, consisting of four equally weighted bits and an input carry. Even though [76] reports (4:2) compressors to actually be (5,3) counters, this is incorrect, since the output carry, and intermediate output carry are of the same weight. Figure 4.4 illustrates a (4:2) compressor constructed from full adders, or (3,2) counters. Notice that the carry cannot propagate for more than two compressors due to the two stage structure, and hence long carry propagate chains are avoided. Higher level counters and compressors are possible, with a (7:3) compressor possessing two input and two output intermediate carries, and a (7,3) counter possessing 7 identically weighted inputs and 3 weighted outputs. Both of these structures have been used in multiplier implementations [76], [77]. There is, unfortunately, no standardized terminology which is universally accepted in the literature for these different processing elements; however, a rule of thumb is that compressors possess intermediate carries and counters do not.

Figure 4.4: A (4:2) Compressor



The (6,3,4) counter iterative array architecture [70] achieves its speed increase over the conventional CSA array due to multi-direction sum and carry propagation, and the use of complex counters. The latter serve to absorb the various sums and carries along the bottom-left—top-right diagonal of the square array yielding a delay of n cells, assuming a counter has the same delay as an FA. If this assumption is not true[1] then the worst case delay would be (n-1) complex counter delays, plus one FA delay. This architecture suffers from irregularity along the diagonal which effectively eliminates the possibility of its layout in a single regular structure.

Figure 4.5: Square Partial Product Array



The 5-counter multiplier [70], [71] has a regular structure, and its architecture will be elaborated upon. Suppose the partial product array shown in Figure 4.1 were rearranged into a square matrix by "pushing" the top right and bottom left corners in opposite directions. This would mean that partial product bits possessing the same weight would lie on diagonals, as indicated in Figure 4.5. Now if the array were folded over on itself, with the bottom-left—top-right diagonal serving as the "folding crease", the resulting arrangement would be as shown in Figure 4.6. The diagonal weights have been preserved, and pairs of partial products having identical weights have been formed. It is this array which is the inspiration for the 5-counter multiplier. Figure 4.7 shows the interconnection for this architecture. Note that the triangular array of PPs has been flipped vertically, and rotated right by 90°. PPs are applied pairwise to the inputs of the (5,3) counters, and they are accumulated within the array except for counters on the diagonal edge, which only

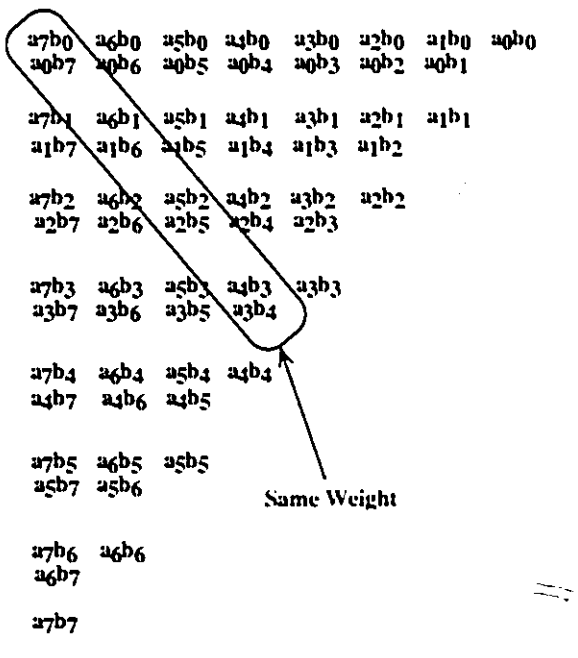
[1] It is realized that this is probably the case.

receive one PP. As can be seen from the figure, counters along any diagonal receive the least significant bit from the output of the counter diagonally above it, as well as the higher order bits from one and two diagonals below. This is in accordance with the binary weight distribution within the triangular array. The structure has super carries along the top edge which seem to skip a level of counters. This is necessary to feed the carries to counters with appropriately weighted inputs. There is irregularity along the diagonal edge of this structure, and there is only one partial product input to the (5,3) counters lying on this edge. As well, carries must be accumulated with a special adder which adds two bits of the same weight and two bits of ½ weight. The adder output is expressed as:

$$\begin{aligned} &SUM: && out1 = w \oplus x \oplus (yz) \\ &CARRY: && out2 = wx + wyz + xyz \end{aligned} \qquad (4.4)$$

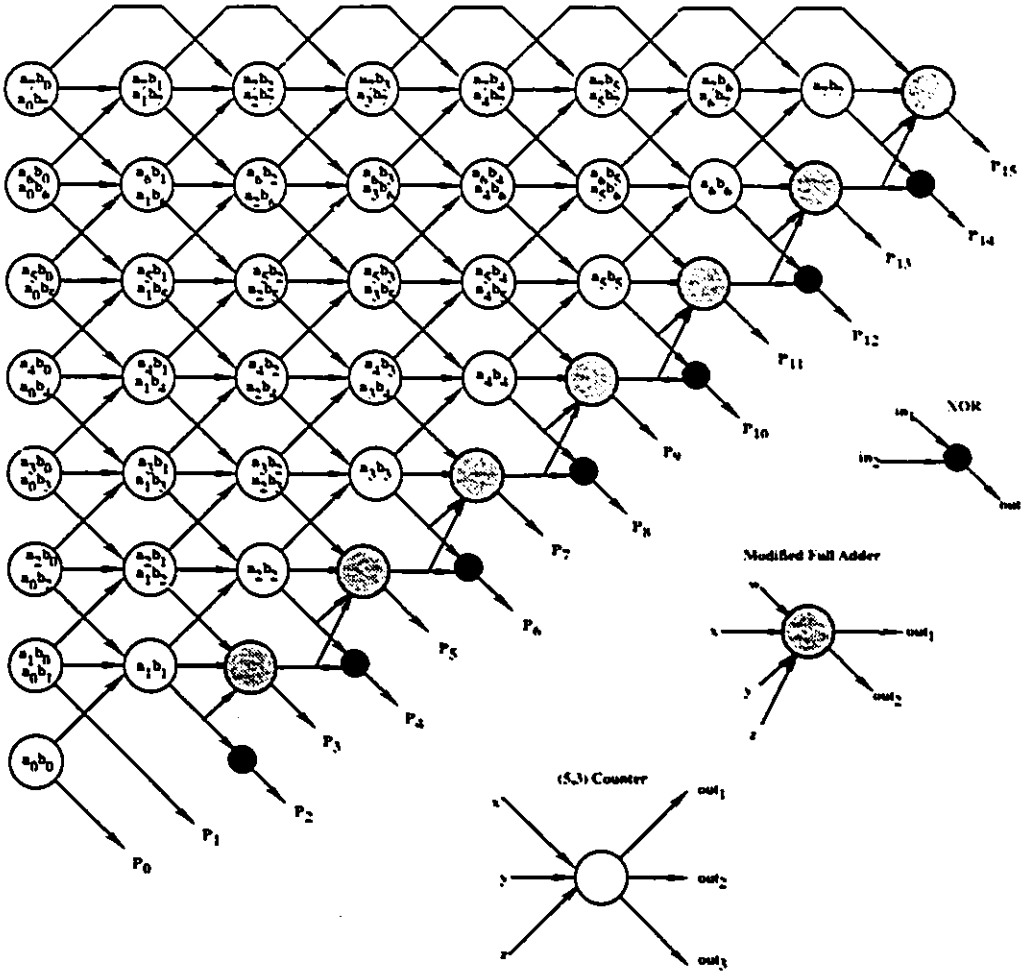
The necessity for such an operation can be explained as follows. Outputs of the multiplier formed by the XOR gates progress from bits with equal weight. Notice that the inputs are from the counter diagonally above it, and the second highest weighted bit from the counter which is one diagonal to the left. These two inputs are of the same weight, and they are summed without carry generation by the XOR gate. Carry generation is effectively performed in the next diagonal of higher weight by the modified adder. A logical AND is performed on the two digits, as shown in equation set 4.4, and the result is added to the other bits. This has the effect of adding in the lower order carry. This is the basis for claiming that two of the inputs to the adder have ½ weight. If we treat the (5,3) counter delay and the modified adder delay as being equal, the overall multiplier delay is (n+1) cell delays. Thus, as stated previously, the (5,3) counter multiplier has a linear delay dependence on operand length. A 32 bit version of this architecture has been implemented in CMOS [71].

Figure 4.6: Folded Partial Product Array



The advantage of most linear parallel multipliers lies in their regularity, but with large operands the delay can become unacceptable. Consider that IEEE Standard 754 double precision floating point numbers possess a 52 bit mantissa. Thus, although these multipliers offer very regular structures, they are not suited for many of today's high precision processors.

Figure 4.7: (5,3) Counter Architecture



Multiplication delay can be minimized by reducing the number of series adds required to accumulate the partial products. Some well known methods for achieving this include the recoding of multiplier digits, and the use of tree structures. The recoding method attempts to maximize the number of zeroes in a number. A zero digit in the multiplier operand results in a shift of the accumulated PP, with no accumulation necessary. It is thus desirable to have as many zeroes as possible to reduce the computational load. This idea

was proposed by Booth [78]. Consider a multiplier possessing a field of m consecutive ones, which can be represented as $[00\cdots0(11\cdots11)0\cdots]$. This number would require the generation and accumulation of at least m partial products; however, it can alternatively be expressed as the difference of two numbers:

$$[00\cdots0(11\cdots11)0\cdots]=[00\cdots1(00\cdots00)0\cdots]-[00\cdots0(00\cdots01)0\cdots] \quad (4.5)$$

If we use signed digit notation[1], the above number can be written as $[00\cdots1(00\cdots0\bar{1})0\cdots]$. Instead of generating m partial products, we need only generate two numbers, i.e. the terms on the right side of equation 4.5. If there are eight consecutive ones in a multiplier, then m=8, and we need only generate two numbers. One of these numbers will be in two's complement form, and their summation yields the same result as that of the summation of all eight PPs.

Booth encoding in its simplest implementation examines successive groups of two bits at a time to determine the start and end of a string of ones. Disadvantages to this scheme include a variable number of required additions between successive shift operations, as well as an inefficiency in dealing with isolated ones and zeroes in the number to be recoded. The radix 4 modified Booth encoding algorithm is a very popular derivative of the above procedure [80]. It produces ½ partial products from an $n \times n$ bit multiplication by examining three bits at a time. Other more complicated modifications to the above algorithm have been used, and a fairly recent work unifies the theory behind higher radix modifications [81]. Radix 4 modified Booth encoding is by far the most popular recoding algorithm, and it has been successfully used in many multiplier implementations [73], [82], [74], [83], [75].

As was mentioned before, tree structures are also used to reduce the number of series adds required to accumulate the partial products. Wallace [84] proposed the use of what he

---

[1]Signed digit number systems utilize digit sets containing both positive and negative members. A signed digit (SD) number $N=(\eta_{k-1},\eta_{k-2},\ldots,\eta_0)$, with $\eta_i \in \{1,0,-1\}$ and $N=\sum_{i=0}^{k-1}\eta_i 2^i$, is expressed in weighted form, with digits bearing a bar indicating a negative value. For example, $100\bar{1}$ would be equal to $1\times2^3+0\times2^2+0\times2^1+(-1)\times2^0=2^3-1=7$. Converting from such a number, to binary is simply a matter of separating the negative and positive components e.g. $[100\bar{1}]_{SD}=[1000-0001]_2$. In a two's complement system, this would involve the addition of the first binary number to the two's complement of the second term. For further information, see [79].

termed pseudoadders in a tree structure to achieve logarithmic PP accumulation delay, neglecting the delay of the final carry propagating adder which is always necessary. His pseudoadders are now called multi-bit carry—save adders (MCSA)[1]. In his proposed algorithm, the rows of n bit partial products are grouped into threes and applied to the inputs of MCSAs. The logarithmic delay is a marked improvement over the linear delay of the other type of parallel multiplier, and is a direct result of the tree structured approach. The MCSAs essentially accept three numbers (partial products) as inputs, and produce two numbers as outputs. Thus, each stage of MCSAs reduces the number of partial products by ⅔. In an $n \times n$ bit multiplication where n rows of PPs are generated, and each stage of the K stage Wallace tree reduces the number of partial products as stated, we would expect the following expression to hold:

$$n\left(\frac{2}{3}\right)^{K} = 2 \qquad (4.6)$$

If this is the case, then we can perform the following manipulations:

$$\log_2 n + K \log_2 \frac{2}{3} = \log_2 2$$

$$\log_2 n - \log_2 2 = K \log_2 \frac{2}{3}$$

$$K = \frac{\log_2 n/2}{\log_2 2/3}$$

Thus, the number of stages, K, is logarithmically dependent on the operand length. Since the delay is proportional to the number of stages that the PPs must progress through to be reduced into two vectors which can be summed, the delay is also logarithmic. The above discussion is the basis for the logarithmic delay often quoted for tree based multipliers. One problem exists in this analysis, however, which involves the number of PP rows per stage. Equation 4.6 assumes that n rows can be reduced to 2 rows by K successive multiplications by ⅔. This is not strictly true, since there can only be an integer number of partial products. This is taken into account, if we express the number of PP rows in each stage of the multiplier as p_i, where i is the stage number, starting from the final stage,

[1] A distinction must be made here. In much of the literature today, the term carry—save adder (CSA) is used to describe a simple full adder connected in carry—save fashion. In the paper by Wallace, pseudoadders are referred to, which are multi-bit CSAs. These structures perform the carry—save summation of three n bit operands. This notion of a CSA will be referred to as a multi-bit CSA (MCSA) to avoid confusion.

called stage 0, and progressing upwards through the tree. Thus the corrected interpretation is :

$$\rho_0 = 2$$
$$\rho_{i+1} = \lfloor 3/2 \rho_i \rfloor \qquad (4.7)$$

where $\lfloor x \rfloor$ is the floor function[1], and the recursion rule yields the series {2, 3, 4, 6, 9, 13, 19, 28, ...}. It was recognized by Dadda [85] that the number of full adders, or (3,2) counters as he refers to them, required to perform the PP accumulation could be minimized if PP row reduction proceeded according to the above series. This effectively exploits the geometry of the partial product array resulting from the multiplication operation.

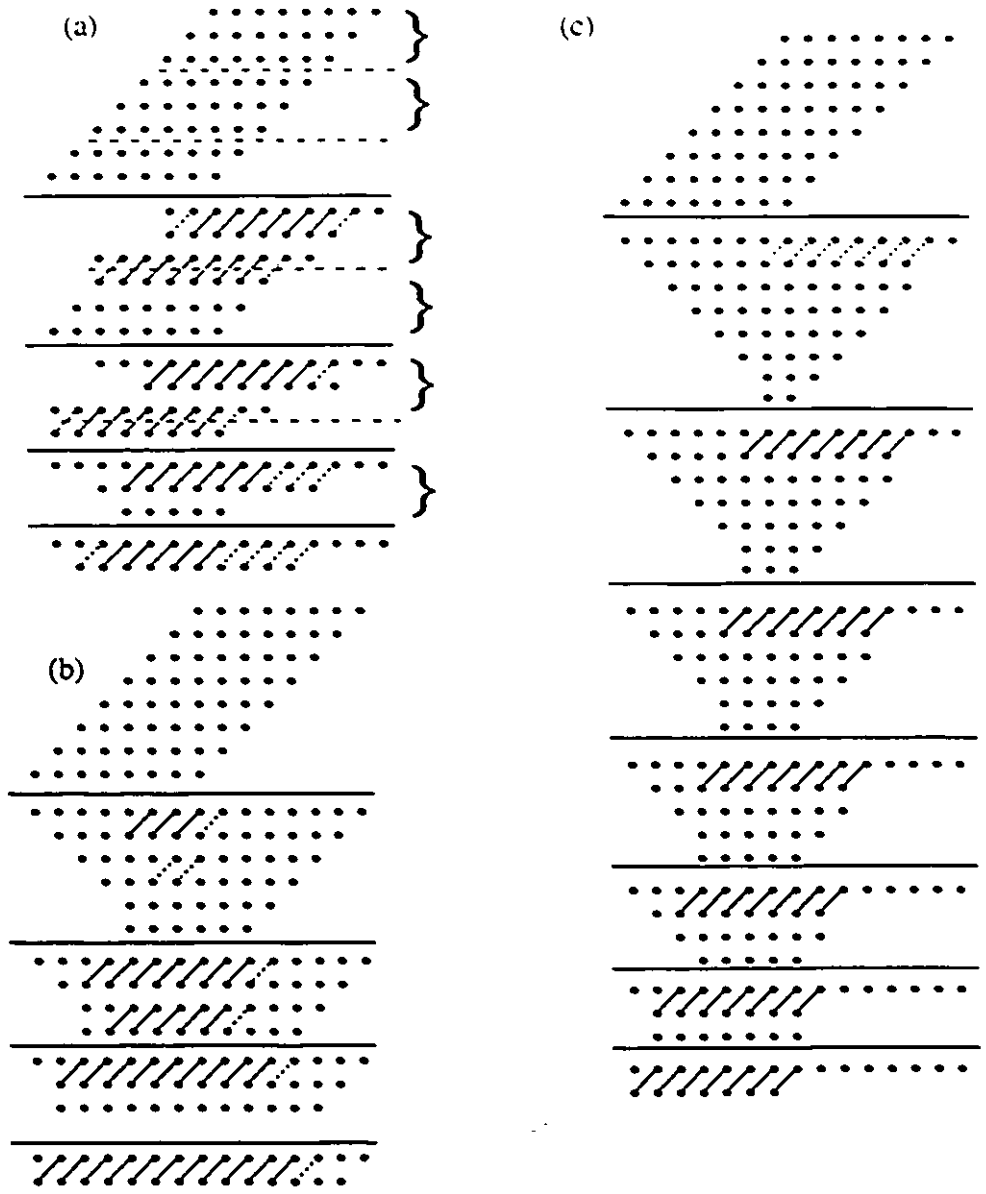
Figure 4.8 shows arrangements for several types of multiplier architectures which illustrates the algorithm by which the partial products are accumulated. The figure is for an 8×8 bit multiplication, and only the carry—save portion of the accumulation is illustrated. Multipliers of the following types are shown: (a) a standard Wallace tree, (b) Dadda's proposed architecture, and (c) a simple CSA array multiplier. The dots represent binary digits, and the upper parallelogram shaped arrangement of dots in each scheme represents the original PP array, with columns of bits sharing the same weights. Since full adders, and half adders are being used in all the cases pictured, the successive accumulation of partial products is indicated in the diagram as follows. Three bits from the same column (same weight) which are fed into a full adder will produce a sum bit and a carry bit, with the former remaining in the same column, and the latter being issued into the column to the immediate left with a correspondingly higher weight. This relationship is indicated in the diagram by a solid diagonal line segment connecting two dots. The dots represent the two output bits from a half adder if the line segment is dotted. This figure illustrates some key features of the different multiplier types.

In the Wallace tree, bits are grouped into threes in each stage and applied to the inputs of MCSAs. This grouping is indicated by the brackets in the diagram. The Wallace scheme uses 51 adders (36 full adders and 15 half adders) to reduce the PPs to a carry and sum vector in four stages, which corresponds to a delay of 4 full adders. The Dadda scheme uses the same number of stages and thus achieves the same delay, with only 42 adders (36 full adders and 15 half adders).

---

[1] The floor function $\lfloor x \rfloor$ yields the integral portion of x.

Figure 4.8: Comparison of Several Multiplication Algorithms



The CSA array adder uses 49 adders (42 full adders and 7 half adders) in 7 stages to accumulate the products, thus having an inferior delay to the other two schemes. Both the Wallace tree and the Dadda scheme achieve high speed by reducing the number of required series adds to a minimum. In the Dadda scheme, this is accomplished by distributing adders based on which columns of equal weight bits need to be reduced in order to yield a specific number of rows per stage. Because of this column—oriented algorithm, these

types of multipliers are called column compression multipliers. All multipliers of this type require a final carry propagating fast adder which usually employs some form of accelerated carry generation. The structure of the Dadda multiplier has the added feature of overlapping fast adder delay with higher order sum/carry vector generation, since the lower order bits progress through fewer stages.

Dadda also proposed the use of higher order counters in column compression multipliers to reduce the number of stages required [85]. For example, if we assume that (7,3) counters are available, the number of partial products are no longer reduced by ⅔ as in the case for (3,2) counters, but by ⅞. The final sum and carry vector will still be produced from 3 rows of partial products, but these three rows can result from the compression of 7 rows. The 7 rows progress from 15 rows, and we can generate a series in a similar way as for the Dadda multiplier, consisting of the set: {2, 3, 7, 15, 35, 79,...}. Thus, a 54×54 bit multiplication PP array will be reduced in 5 stages if (7,3) counters are available, where it would take 8 stages if only full and half adders were available. The tradeoff is that (7,3) counters are much more complex than (3,2) counters, and will generally have greater delay.

Implementation of high order counters has been studied in [86]. As well, Wang [87] claims that more complex cells bring with them a correspondingly larger area, which means that cross cell and cross stage connections will be longer and the associated parasitics will be larger. Multiplier designs using (7,3) counters and compressors have been presented in [76], where it was reported that the multipliers performed similarly, yet the (7,3) counter based architecture was easier to layout due to less interconnection. A notable example is in the IBM RS/6000 Floating Point Unit design [83], [77], where the (7,3) counter was chosen due to the reduction of interconnection requirements associated with it. Stenzel suggested the use of multi-input counters which accept input from several columns, and they were adapted to the Dadda scheme [72]. Implementation of these counters with combinational logic results in circuits which suffer from carry propagation delay. Many multipliers have been designed based on (4:2) compressors. Santoro [75] used the (4:2) compressor pictured in Figure 4.4 to implement a multiplier which was based partially on a Wallace tree, while Nagamatsu [74] used the same compressor in a pure Wallace tree design. A recent work [88] argues that the use of (4:2) compressors actually renders modified Booth encoding [80] obsolete since an equivalent reduction is accomplished with less hardware and in less time.

Tree (column compression) multipliers tend to be difficult to physically realize due to their irregularity in structure and interconnection. The number of adders in each stage varies, and there are cross stage connections, as well as long interstage connections. This makes their physical realization difficult. Despite this, Wallace trees have been the basis for many multiplier implementations [73], [82], [74], [83], [75], [89], and the Dadda architecture has also been successfully implemented [90], [91]. Indeed, layout is not nearly as serious a problem today as it was in the past due to the advanced computer aided design (CAD) tools now available, however, connection specific capacitive loading will still be a prime mover behind the development of regular structures which maintain the superior delay characteristics of the above multipliers.

4.3 Macrocells

The work involved with the macrocells to be described was in partial completion of a contract to develop a viable Cadence Edge™ design environment. The two main goals of this contract work were to provide high level VLSI design primitives for high performance arithmetic systems, as well as to test and verify the BiCMOS Edge design environment. To attain these goals, several macrocells were realized using fast architectures. The final implementation of these designs within the Edge design framework demonstrates that a non-trivial real world design process can be carried out within this environment. In total, six different macrocells were implemented.

4.3.1 Two Bit Full Adder Multiplier

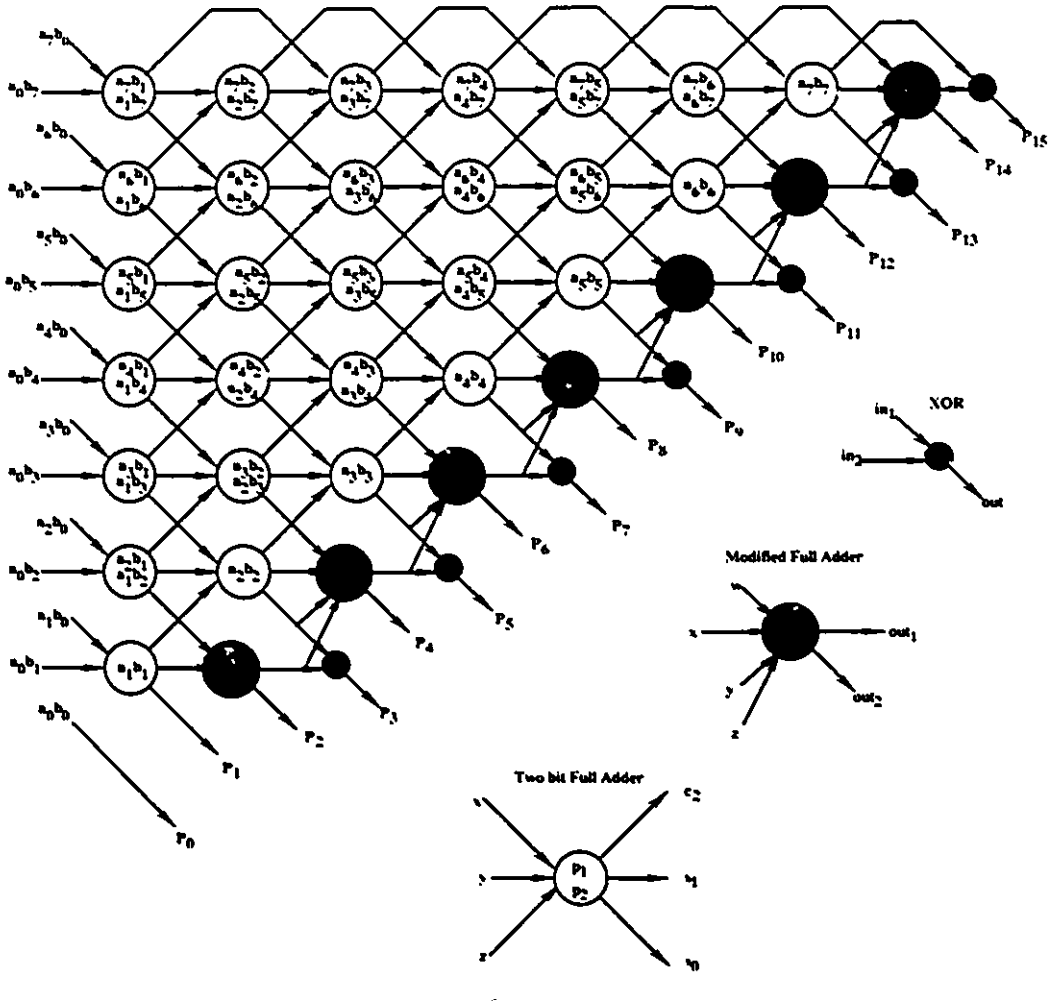
A new parallel multiplier with linear delay has been proposed by Wang [92], and the structure is shown in Figure 4.9. It is similar to the 5 counter multiplier [70] discussed earlier, but uses two bit full adders as a basic cell. The behaviour of such an adder is described by the following equation:

$$x + y + z + 2(p_1 + p_2) = 4c_2 + 2s_1 + s_0 \qquad (4.8)$$

where x, y, and z are inputs of equal weight, and p_1 and p_2 are inputs with twice the weight. In other words, this adder will accept a pair of two bit binary numbers and an input carry, and sum them up. As shown in the figure, partial products are applied to the p_1 and p_2 inputs of the two bit full adders. Inputs applied along the left edge of the array are also partial products, however, they are of a weight lower than the PPs shown inside

the associated adders. This can be verified by consulting the triangular array in Figure 4.6, and using the diagonal weighting concept mentioned earlier. Inputs positioned in different diagonals possess different weights. Recall that in Nakamura's structure, only 2 of the 5 possible inputs to the (5,3) counters along this edge were used, while in Wang's structure, 4 of the 5 possible inputs of the adders are used.

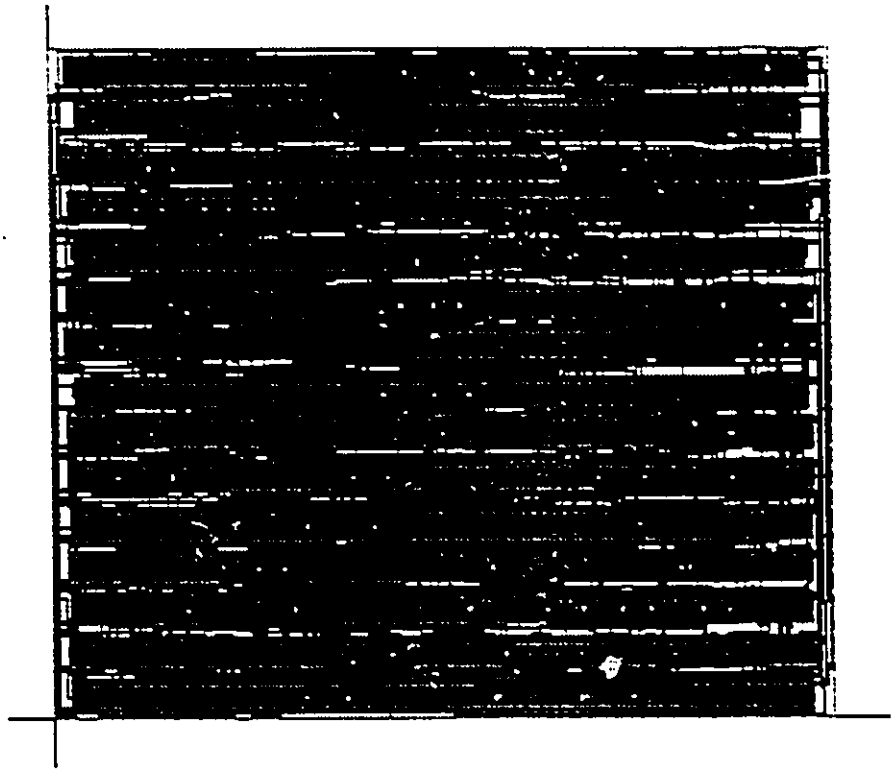
Figure 4.9: Wang's Linear Parallel Multiplier Architecture



Three different versions of this architecture were created as BATMOS macrocells. Two different structures for the two bit full adder were used in two separate multiplier designs. One realized the adder directly with logic gates [93], while in the second implementation, full adder standard tcells were used to construct the two bit adder [93]. Designs were

accomplished with 100% CMOS, and also with BiCMOS tcells[1] and standard cells[2]. The Edge™ Place and Route tool was used for layout creation, although a significant amount of post processing was required[3]. A heuristic method of distributing BiCMOS gates was used in the BiCMOS macrocells, where an attempt was made to place cells containing bipolar driver transistors at nodes within the circuit which were most heavily loaded. The layout for the BiCMOS implementation of the architecture using simple gates is shown in Figure 4.10. It contains 624 standard cells, and has a size of $1705\,\mu m \times 1469\,\mu m$. Further details of this cell, and of the other linear parallel multiplier macrocells are provided in Appendix B and Appendix C.

Figure 4.10: Layout of Wang's Linear Parallel Multiplier



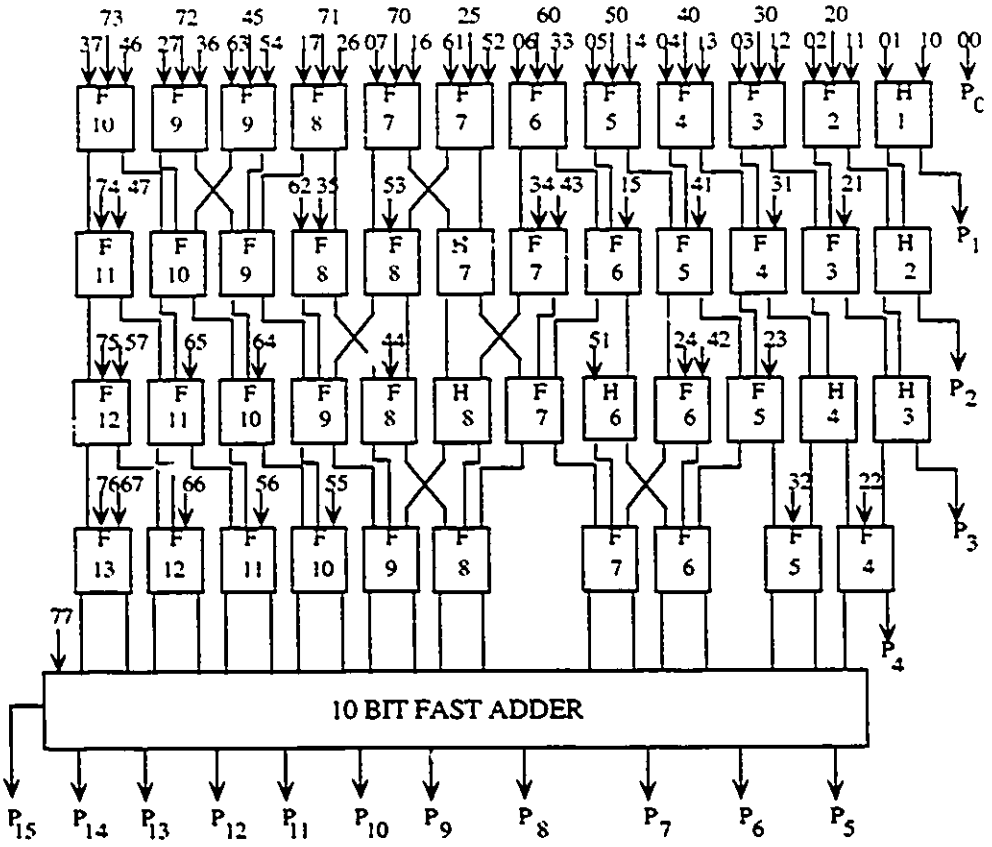
4.3.2 Column Compression Multiplier

The column compression architecture proposed by Wang [87] was implemented as a macrocell. This architecture is pictured in Figure 4.11[4] where Wang's notation has been

[1] In BATMOS technology, tcells comprise a special standard cell library supplied by Northern Telecom.
[2] Some standard cell designs were supplied by other parties involved with the contract.
[3] See Appendix A.
[4] This figure was taken from [87].

used to indicate partial products i.e. a PP termed 34 corresponds to a3b4 in the notation used earlier.

Figure 4.11: Wang's Column Compression Multiplier Architecture

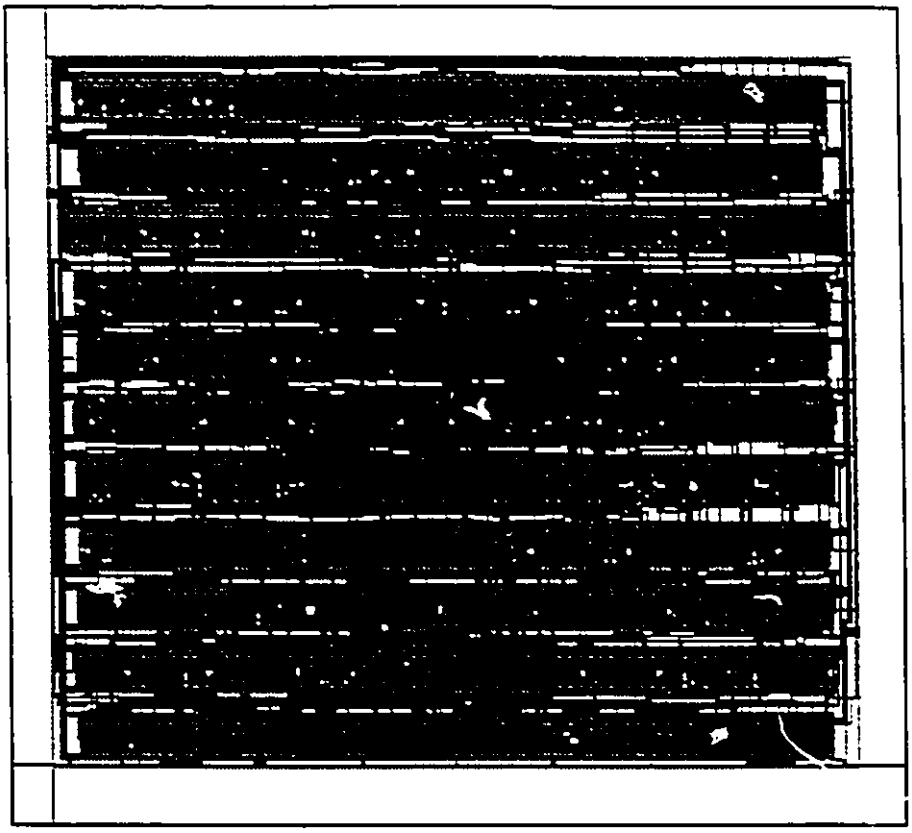


Wang introduces an area efficiency metric which gives an indication of the regularity of the multiplier layout. An 8 bit Dadda multiplier (see figure in [90][1]) implemented using full and half adders is quoted as having an area efficiency of 75%. This is due to the irregular distribution of adders within the various multiplier stages which produce uneven row lengths and thus, cause the physical layout of the architecture to possess an irregular footprint. Interconnect parasitics are increased if the Dadda multiplier is laid out rectangularly due to the longer wires necessary to connect adders which have been displaced from their respective rows. Wang uses a heuristic method of redistributing the adders within the various stages to achieve maximum local connectivity, as well as a much more rectangular footprint than the traditional Dadda multiplier. The architecture pictured

[1] There are several errors in this figure, however the overall structure is correct.

has the added advantage of possessing a shorter fast adder than the traditional 2n-2 bit adder in the conventional Dadda scheme

Figure 4.12: Layout of Wang's Column Compression Multiplier



BATMOS was used in the realization of two different macrocell versions of this architecture. Designs were accomplished in 100% CMOS, and also in BiCMOS. A heuristic method of distributing BiCMOS gates was used in the BiCMOS macrocells, as was explained previously. Physical representation generation was accomplished in a similar manner as for the previous multiplier. The layout for the BiCMOS implementation of this architecture is shown in Figure 4.12. It contains 265 standard cells, and has a size of $1182\,\mu m \times 1043\,\mu m$. Further details of this cell, and of the other column compression multiplier macrocell are provided in Appendix B and Appendix C.

4.3.3 High Performance Adder

The conventional method of adding binary numbers suffers from a long delay due to carry propagation. Consider the addition of two n bit numbers of the form $a_{n-1}, a_{n-2}, \ldots, a_0$,

where the carry into bit position i is c_i, and the carry out of bit position i is c_{i+1}. We can obtain the sum $s_n, s_{n-1}, \ldots, s_0$ by using the following algorithm:

$$\begin{aligned} c_0 &= 0 \\ c_{i+1} &= a_i b_i + a_i c_i + b_i c_i \\ s_i &= a_i \oplus b_i \ominus c_i \\ s_n &= c_n \end{aligned} \tag{4.9}$$

Unfortunately, this scheme is sequential in a sense, since the lower bit positions must be added before the higher ones are processed. This results in a reduction of performance due to a ripple carry effect. There have been many methods developed to speed up the addition process [79], with the carry look-ahead technique being one of the more popular. This algorithm involves the definition of a carry generate, g_i, and a carry propagate, p_i, term for each bit position in the operands. If g_i is logically evaluated as true, then a carry will be generated from the respective bit position, while if p_i is true, the value of the carry into the bit position will be propagated through. Both variables cannot be considered true simultaneously. The algorithm is expressed below:

$$\begin{aligned} c_0 &= 0 \\ g_i &= a_i b_i \\ p_i &= a_i \oplus b_i \\ c_{i+1} &= g_i + (p_i c_i) \\ s_i &= a_i \oplus b_i \oplus c_i \\ s_n &= c_n \end{aligned} \tag{4.10}$$

Due to the parallel formation of generate and propagate terms, the speed of the addition process is greatly increased. This method is implementation limited due to fanin considerations. The generation of the c_i term requires a large number of gate inputs for large i. The solution usually employed is the division of the operands into blocks of bits, and the formation of generate and propagate terms for the respective blocks. Thus, the blocks of bits are treated in the same way as the bit positions were treated previously. This unfortunately renders irregular layout structures.

Brent and Kung suggest an architecture for a fast adder which they claim has a very regular layout [94]. To accomplish this they reformulate the carry chain algorithm by defining an operator, denoted by $\circ$, which behaves as:

$$(a,b)\circ(c,d)=(a+(bc),bd) \tag{4.11}$$

A proof of the associative property of this operator is given in [94], and the algorithm for carry generation using it is summarized below[1]:

$$\begin{aligned}
&c_0 = 0\\
&g_i = a_i b_i\\
&p_i = a_i \oplus b_i\\
&(G_{i-1},P_{i-1}) = \begin{cases}(g_0,p_0) & \text{if } i=0\\ (g_i,p_i)\circ(G_i,P_i) & \text{if } 1\le i\le n-1\end{cases}\\
&c_i = G_i \quad \text{for } i=0\ldots(n-1)\\
&s_i = p_i \oplus c_i\\
&s_n = c_n
\end{aligned} \tag{4.12}$$

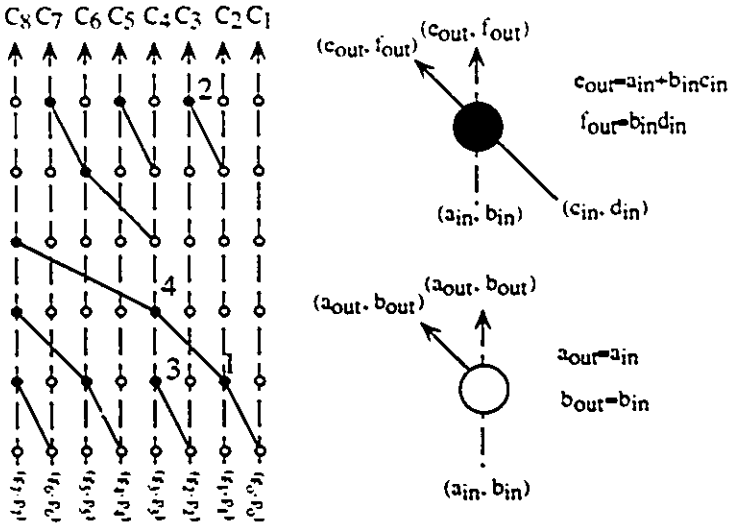
A proof for the equality $c_i = G_i$ is given in [94]. Since the algorithm is associative, a tree type algorithm can be used to compute the recursive segment, and therefore a logarithmic delay will result. Figure 4.13 illustrates the tree based carry computation section of the algorithm[2], implemented with an array of simple processing elements which renders a regular layout. There are two types of processing elements, one which performs the $\circ$ operation, and one which simply transmits its inputs to its outputs. The generate and propagate terms are formed in parallel, and applied to the lower edge of the array. The (g_0,p_0) term directly provides c_1. This term is combined with (g_1,p_1) at position 1 in the diagram to provide c_2. Carry c_3 results from the combination of the result from position 1 with (g_2,p_2) at position 2. The term resulting from position 1 is also combined at position 4 with $(g_4,p_4)\circ(g_3,p_3)$, formed at position 3, to produce $c_4=((g_4,p_4)\circ\big((g_3,p_3)\circ\big((g_2,p_2)\circ(g_1,p_1)\big)\big)$. The computation for the rest of the carries progresses similarly. This tree based approach is only possible due to the associative nature of the $\circ$ operation. When the carries are available, all that is necessary is the implementation of the second last line of the algorithm given in equation (4.12). This is accomplished with a bank of XOR gates. The overall architecture can be divided into three stages, as shown in Figure 4.14. The first stage forms the generate and propagate terms, while the second

[1] Note that indexing is different than in [94] since in the author's opinion, this notation is less confusing, and more consistent with that in the literature.

[2] Mistakes were found in a similar diagram given in [94]

stage consists of the tree structure discussed above, and computes the carry values for each bit position. The final stage is comprised of a bank of XOR gates which performs the final summation, and produces the total.

Figure 4.13: Tree Nature of Function Computation



Brent and Kung argue that their scheme renders a regular layout, as suggested in the Figure 4.13. This may be true at the diagrammatic level, but in actual VLSI implementations, this is not strictly the case. The dark processor in the diagram is simple to implement with logic gates, and has comparable complexity to a binary half adder. The light processors are trivial, and consist of wire interconnects. For this reason, the two different elements will occupy largely disparate areas in an actual layout, and thus the structure's regularity is largely contrived.

This adder was implemented as a macrocell with ECL standard cells[1] in BATMOS. Figure 4.15 shows the circuit designed to implement the dark processor units in Figure 4.13. Logic design of the macrocell was constrained by the available gates [22], thus limiting design flexibility. The entire adder cell is composed of 231 standard cells and is 2581 μm $\times 2545\,\mu m$. The layout is pictured in Figure 4.16, and further details are provided in Appendix B and Appendix C.

[1] This macrocell is comprised exclusively of cells provided by other contract participants.

Figure 4.14: Block Diagram of Adder

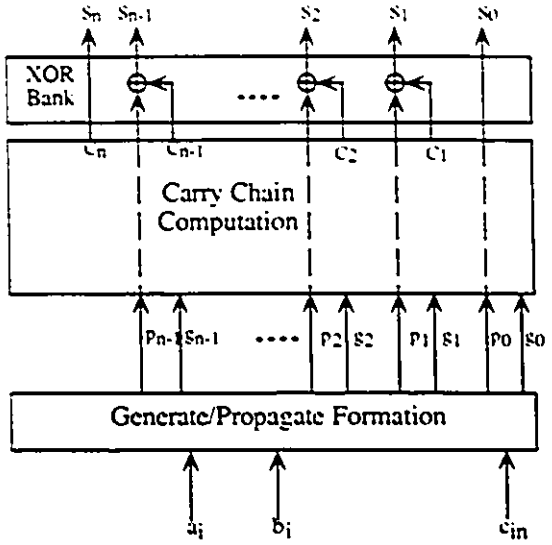
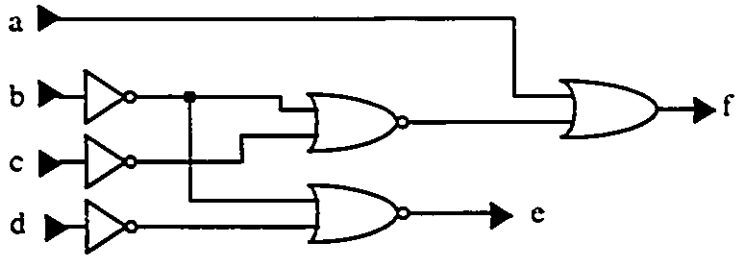


Figure 4.15: Special Bit Processor Circuit

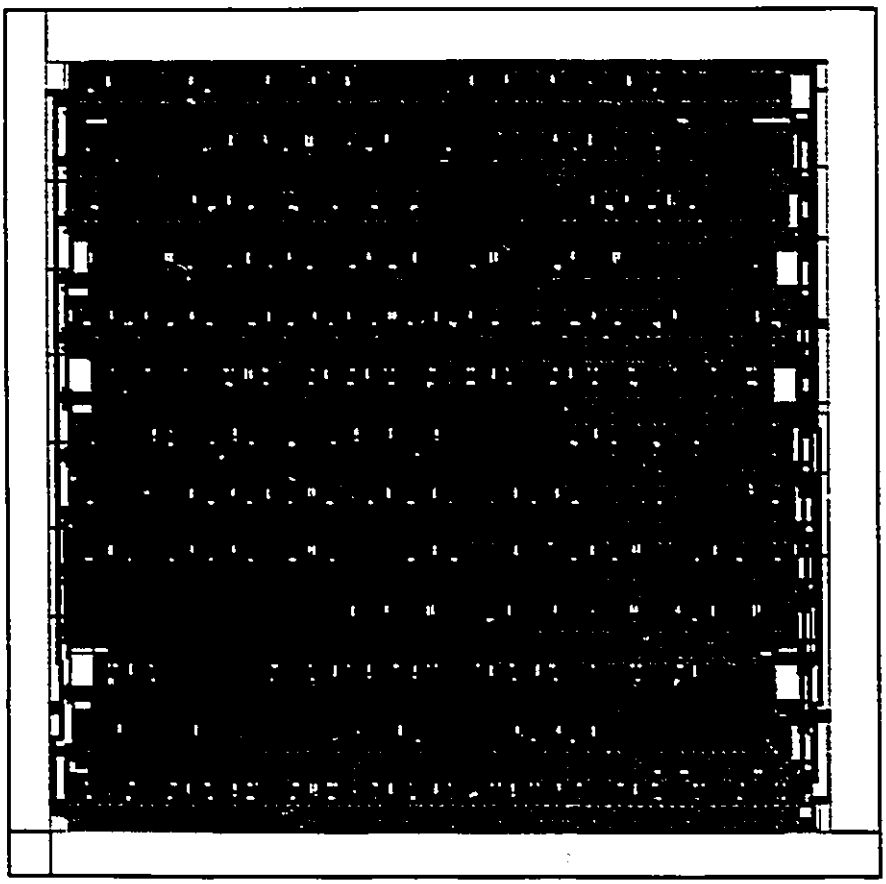


4.4 Additional Discussion of Work Completed

As outlined in the previous section, several architectures were implemented as macrocells in Northern Telecom's .8μ BiCMOS technology. This undertaking was part of a contract [22] and the macrocells were implemented using the Cadence Edge™ tool set. This work had a two fold purpose. First, it was desired to provide a high performance library of macrocells which would be useful in high level arithmetic system design. Second, this work was intended to provide a stringent conformance test for the Edge™ design environment by carrying out high and low level design processes within it.

The macrocells were laid out using the Edge™ automatic Place and Route tool. Most of the standard cells were tcells, and they belong to a hybrid standard cell/gate array cell library supplied by Bell Northern Research/Northern Telecom. Some modifications were performed by other contract participants. Other standard cells, including the ECL cells, the i/o pads, and some of the BiCMOS gates, were designed by other contract participants [22]. It is recognized that architectural advantages of some of the designs, such as regularity of interconnection and spatial organization, are not fully exploited with the chosen design methodology; however, due to the dual goals of this work, as well as time constraints, the manner chosen was warranted.

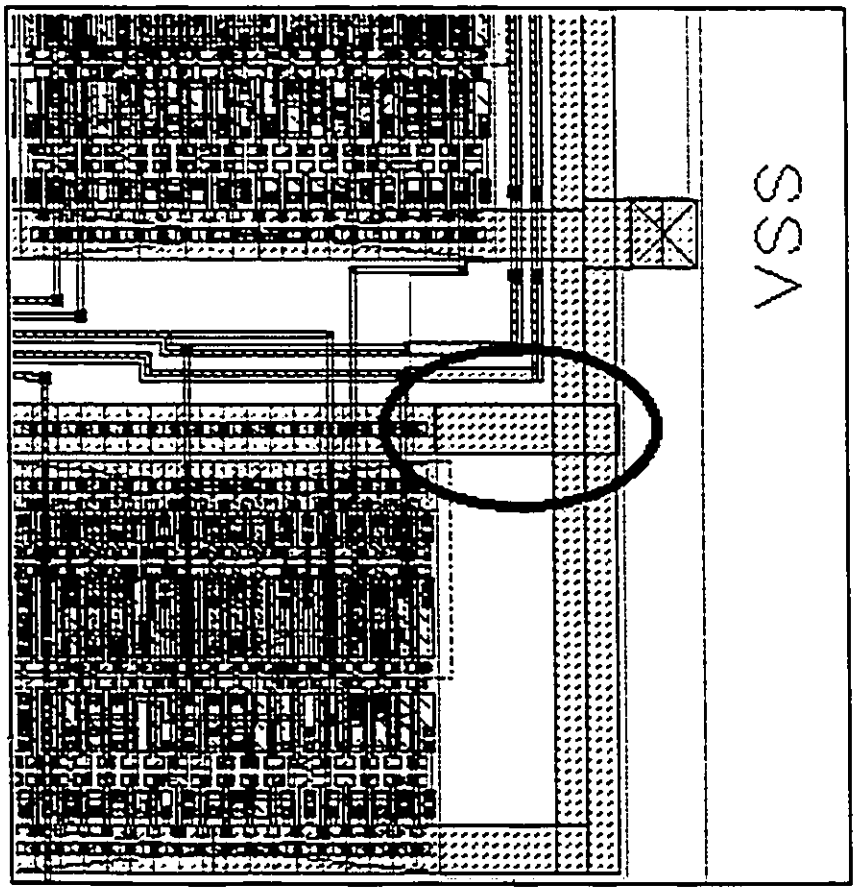
Figure 4.16: Layout of High Performance Adder



Contract deliverables associated with this work consisted of several types of Edge™ representations for each macrocell. The representation types included layout, abstract, extracted, lvs, and symbol. Additionally, Verilog™, Veritime™, and Verifault™ descriptions of all macrocells were supplied, along with relevant Cadence Simulation and Test Language (STL) files. The designs were iterated upon many times due to changes in

the environment initiated by Northern Telecom, and other contract participants. An example of this involved a process design rule change midway through the contract period which necessitated a complete physical redesign. Latchup rule #32.2 [22] required the addition of extra substrate contacts along the top section of all tcells. Because of this revision, all physical representations for the macrocells had to be redone. As well, there was a large amount of post processing required on the placed and routed layouts. The rule change required that a manual connection be made to V_{ss} for every row of standard cells. One such connection is pictured in Figure 4.17. Also, extraneous pin creation after channel explosion made extraction impossible, so a SKILL routine was written to remove the pins automatically[1].

Figure 4.17: The Manual Connection of a Cell Row

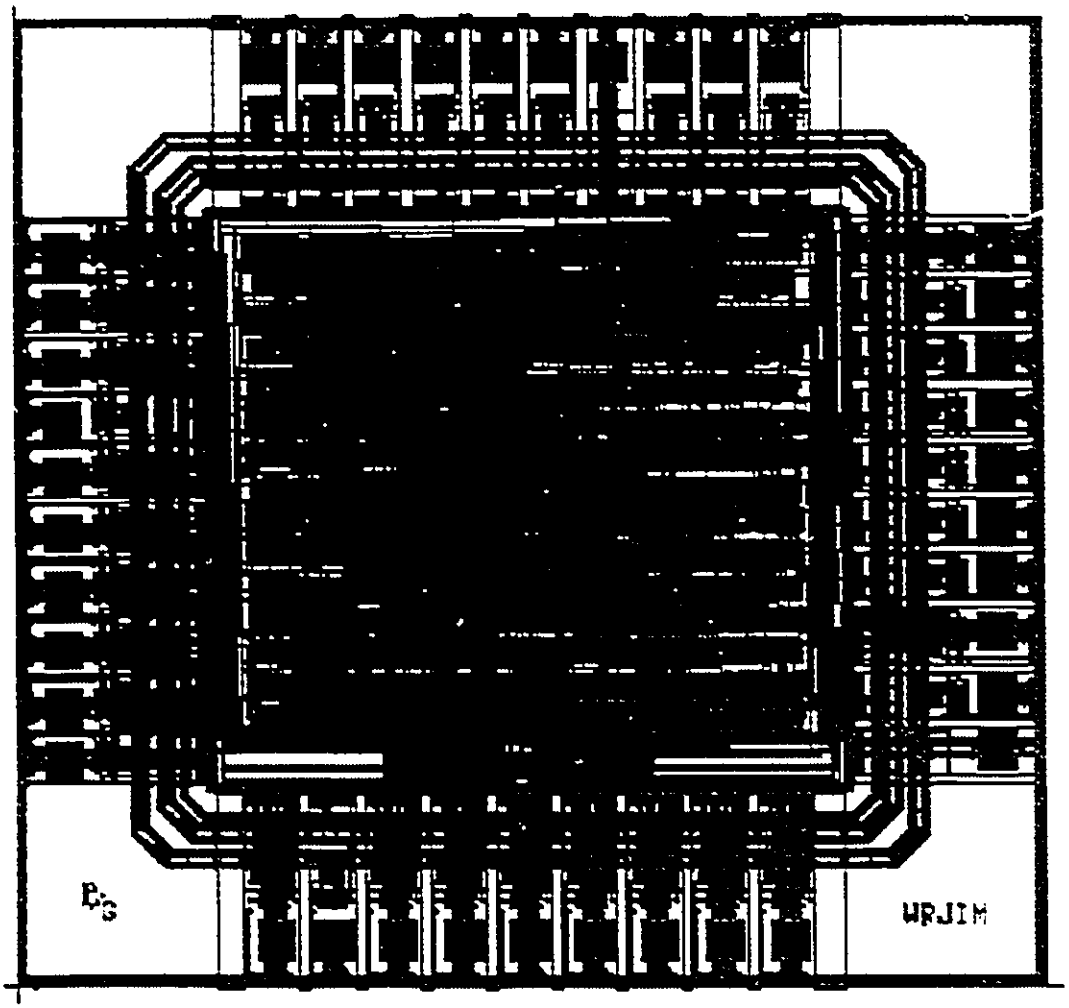


Design Verification took the form of several different types of simulations using Verilog™, Veritime™, and Verifault™, as well as the usual physical design verification procedures

[1] See Appendix E.

involving such things as design rule checking. Verilog™ was used to check functional design correctness. Verifault™ was used to observe the fault coverage of the set of test vectors which were used, while Veritime™ was used to observe the timing properties of the designs. Since all three of these tools are driven by the Verilog™ hardware description language, files obtained from the Verilog™ netlister were used for all three of the simulators. The Cadence Simulation and Test Language (STL) was used to drive the simulation process, which allowed application and automatic output verification of thousands of test vectors.

Figure 4.18: Multiplier Chip Submitted for Fabrication



Although Veritime™ is very useful for estimating clock timing by providing critical path information, actual clock speeds for these architectures were difficult to predict accurately based on these simulations due to the fact that accurate timing information was not available

at the time of the contract. When complete timing information becomes available, the same netlists and files that were provided with the macrocells can be used to simply regenerate simulation results.

The work described above was delivered to the contractor[1], and in combination with some other cells [95], and a report [22], constituted a very successful contract execution by the University of Windsor[2]. As a proof of concept, the BiCMOS version of the two bit full adder multiplier using discrete gates to construct the adder was submitted for fabrication as a standalone chip. The layout is pictured in Figure 4.18, and implementation specific details are given in Appendix F.

4.5 Summary

This chapter has discussed the theoretical basis for several multipliers and adders. Linear parallel multipliers, sometimes referred to as array multipliers depending on the structure, possess a delay which is linearly dependent on operand length. Their regular architecture is very favourable from a VLSI layout perspective, however delay and size become prohibitive with larger operands. Column compression multipliers, sometimes loosely referred to as tree multipliers, achieve a logarithmic delay characteristic with operand length, but possess very irregular structures, thus making them difficult to physically realize. Two special architectures were described. One was a linear parallel multiplier which achieves lower delay than comparable architectures, while the other was a column compression type, which possesses a higher degree of regularity than the Dadda scheme. Additionally, a fast adder architecture proposed by Brent and Kung was described. Contract work which involved the implementation of these architectures as macrocells in BATMOS, Northern Telecom's BiCMOS technology, was discussed, and certain aspects of it were elaborated upon. This portion of the thesis work was in partial fulfillment of a contract which was successfully executed by the University of Windsor. Additionally, one of the contract macrocells was submitted for fabrication as a proof of concept chip.

---

[1] Micronet.

[2] See letter of thanks (email) in Appendix D.

Chapter 5

CLOCKING STRATEGIES FOR PIPELINED ARITHMETIC STRUCTURES

5.1 Introduction

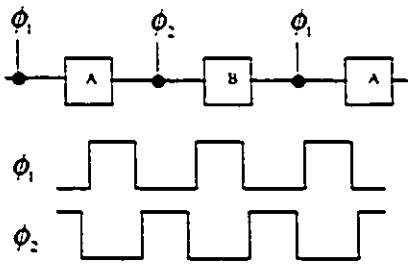
When considering pipelined arithmetic structures, the clocking strategy has a profound impact on the design and implementation of the system. The logic design style is also intimately linked to the clocking scheme. High performance technologies usually offer several levels of low resistance interconnect, but as integration levels continue to increase every year, interconnect congestion is an important spatial concern. Also, as die sizes increase, global interconnections on average become longer, and clock skew becomes a major design problem. A pipelined system has different considerations than a general system due to its nearest neighbor connections, as well its spatial organization. Clocking strategy, and latch configuration can have a major impact on system throughput. Primary consideration is given here to pipelined structures implemented with dynamic logic.

5.2 Pipeline Clocking and Circuit Techniques

Figure 5.1 illustrates the concept of pipelining. Boxes labeled A and B represent processing networks which perform a logic function while the dark circles represent latches. The busses connecting the elements are represented by the solid lines. The system pictured uses a timing strategy called nonoverlapping two phase clocking. This technique was used in NMOS technologies, where the latching elements consisted of simple NMOS pass transistors with the respective clock signal driving their gates. The analogy to this method in CMOS is called nonoverlapping pseudo two phase clocking. In this technique, ϕ_1 and ϕ_2 are used, as well as their complements, $\overline{\phi}_1$ and $\overline{\phi}_2$. The pass transistors are

replaced with transmission gates, which pass both high and low values equally well. Theoretically, four signals must be routed to the clocked elements, but in practice, two clock signals are usually distributed, with their complements being provided by local inversion. This scheme allows the use of both static and dynamic logic in the processing blocks. Doubling up the latches, placing a static inverter between them, and applying alternate clocks to the first and second latch increases throughput if static logic is used [96]. Two phase clocking schemes are susceptible to a phenomenon called clock skew.

Figure 5.1: Pipeline Concept



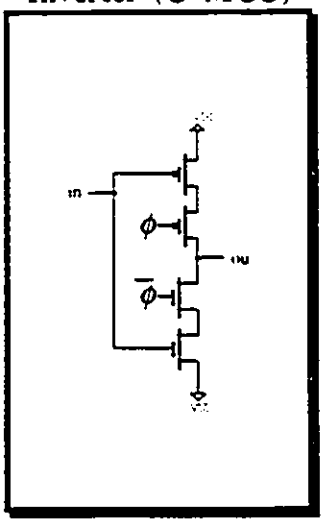
Clock skew can be defined as the difference in arrival times of a clock signal at different points on a chip, and it can occur between clock phases, or it may involve only one phase. This concept is different from clock delay, since a clock signal arriving from two different paths with equal delay will have zero skew. Clock skew is usually caused by such things as unequally loaded clock lines and unequal lengths of clock interconnect. In heavily pipelined systems, local connectivity is predominant, and thus the clock skew associated with the long global clock lines may be a chief performance limiting factor. Clock skew worsens as integration levels (chip sizes) increase due to longer interconnection paths, and as feature size shrinks, due to a higher RC constant per unit length of interconnect [27]. Pseudo two phase clocking in CMOS is therefore, very much effected by clock skew, since separation of phases must be maintained to avoid data race. Extra dead time is usually inserted between the clock signals to compensate for skew, but this is an inefficient solution.

Pseudo single phase clocking has been very popular, and has spurred the development of many dynamic logic design techniques which are suited for pipelining. The latches used are either the transmission gate style, mentioned previously, or a clocked inverter style, which is illustrated in Figure 5.2; it is usually given the acronym C^2MOS. Consecutive latches in a pipeline have the ϕ and $\overline{\phi}$ connections reversed, which provides the necessary latching action for the logic circuitry which is between them.

Several dynamic logic techniques have developed which are useful in pipelined structures. Domino logic [97] uses a single clock phase as well as a static inverter to eliminate the internal race condition encountered with basic dynamic logic gates, but has several

limitations. Only non-inverting structures are possible, and charge sharing is a serious problem with large numbers of series connected gates.

Figure 5.2: Clocked Inverter (C^2MOS)



The former difficulty limits design flexibility, while the latter problem has been effectively addressed by the use of sample and hold techniques [98], [99] which rely on a four phase clock. Even though [99] claims to have very low susceptibility to clock skew, the routing resources necessary for four distinct clock phases make this technique unattractive for large, pipelined designs. NORA dynamic logic [100] uses pseudo single phase clocking, and implements logic with both p and n MOSFETs. This allows the clock cycle to be utilized more fully due to the presence of ϕ and $\overline{\phi}$ blocks, and increases logic functionality within each block due to the presence of p logic. While one logic block is precharging, the other is evaluating. This is an attractive method for implementing pipelined structures, and it is compatible with domino logic. There are, however, several drawbacks including speed degradation due to the use of lower mobility PMOS devices, and charge sharing. As well, only an even number of inversions is permitted between successive p and n type blocks. This latter point limits design flexibility. True single phase clocking (TSPC), first suggested in [101], provides an interesting alternative to the above mentioned pipeline schemes.

5.3 True Single Phase Clocking (TSPC)

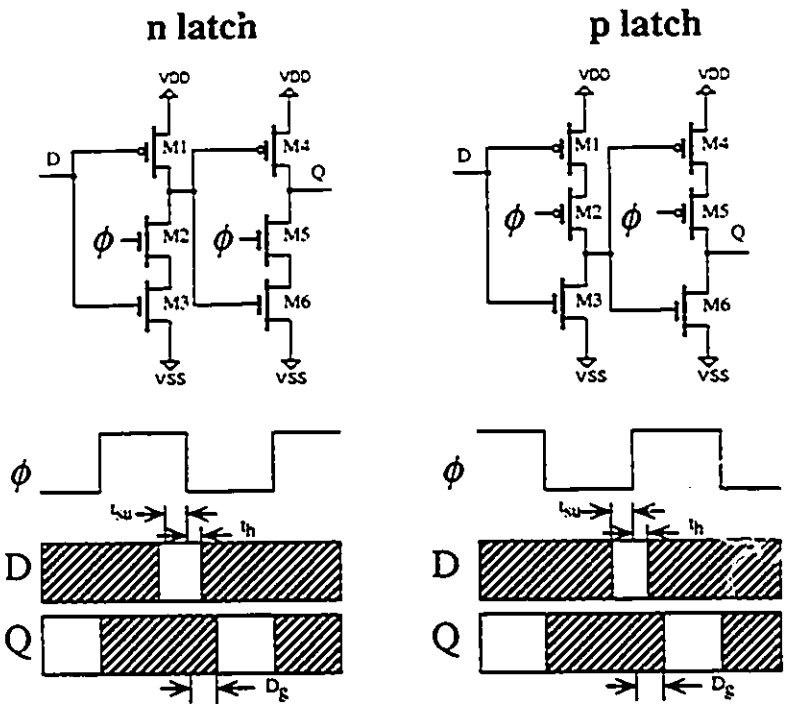
The true single phase clocking technique uses a single clock signal, without the need for a complement, and has several advantages. It requires a minimum of routing resources, uses simple clock generators, and skew between different clock phases is eliminated[1]. Other advantages will become apparent shortly. Large scale designs have been implemented using this clocking strategy with great success [102]. The example which lends most credence to the claimed advantages of this technique is the Alpha microprocessor, by Digital Equipment Corporation, which uses a form of this clocking scheme [24]. As first suggested in [101], this technique is similar to NORA logic in that it uses n and p blocks

[1] Self skew is still possible, however.

which precharge and evaluate on different portions of the clock signal, but the key difference, obviously, is that it requires only a single clock phase. In [103] the work is generalized, and further developments are made in latching configurations. Also, reverse clock distribution is recognized as an effective method of making the TSPC structures resistant to clock skew, and several design examples are given. Timing restrictions and constraints in TSPC systems are given rigorous treatment in [104], and the effect of clock skew on performance is examined more closely. The robustness of the TSPC strategy against such things as noise, clock skew, and clock slope is given a thorough treatment in [105]. High test clock rates of 700 MHz with non-trivial systems in a 1.2μ CMOS process have been reported using this clocking technique [102]. The philosophy of the TSPC technique will now be explained.

The motivation behind using a pseudo two phase nonoverlapping clock in CMOS is to create two points in each clock cycle at which to latch data. The non-overlapping feature of the clock signals creates set up and hold times for latches which prevents data flowthrough. The basic philosophy of TSPC is to provide these conditions, not with the clock signals but with the latch design. Two types of latches, termed n latches and p latches, are transparent during different parts of the clock signal, and this is demonstrated in Figure 5.3 with what are termed in [103] as N-C^{2}MOS and P-C^{2}MOS latches. In the figure, D is the data input to the latch, and Q is the latch output. The graphs below the latches indicate respective timing behaviour. Hatched areas indicate periods in time when the signal can vary, while unhatched sections indicate areas of constant, or stable signal. At the critical clock transitions, the input must be held constant for a certain time before and after the transition. These times are termed setup time, t_{su}, and hold time, t_h, respectively. The time between the clock transition causing the latch to change from the transparent state to the hold state and the time at which the output becomes stable is termed the latch delay, D_k. These latches operate by becoming transparent when the clocked transistors are turned on. When this happens the latch is effectively two inverters connected together, thus accounting for its transparency. When the clocked transistors are turned off, the output node is isolated from any input changes. Consider the n latch driven by a "1" input until after the clock goes low; this clock transition puts the latch in the hold state. Now let the input change to "0", which turns on M1. This charges up the internal node, which turns on M6, but the output node cannot be discharged because the clock has turned off M5. Other input changes can be similarly traced through for both types of latches, and it will be discovered that the output retains the value it possessed just prior to the clock transition. The split output latch [103] finds its origins with this style of latch.

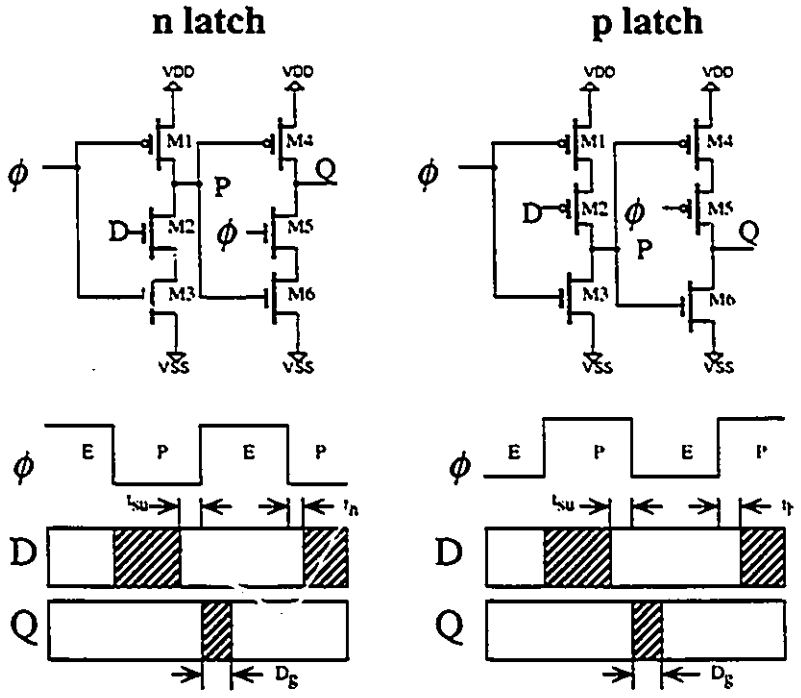
Figure 5.3: N-C^{2}MOS and P-C^{2}MOS Latch structures



A more useful set of latches for precharged dynamic pipelined structures is shown in Figure 5.4. These are called precharged latches, due to the fact that they possess a precharged node, indicated as node P in the schematic portion of the figure. Even though their topology looks very similar to the previous latches, their timing behaviour is different, as indicated in the figure. The latches go through precharge and evaluate phases, indicated by a P and E respectively in the lower graph, just as in conventional dynamic logic. The principle of operation of these types of latches will be explained for the n type latch. The first stage behaves as a conventional dynamic circuit, with precharge and evaluate phases occurring in the same fashion. The second stage is essentially a clocked inverter. During precharge, $\phi = 0$ and the drain node of M2 is charged high. This turns on M6, turns off M4, and since M5 is in the off state, due to the clock level, the output node is not discharged. This constitutes the hold state of the latch. When the latch enters the evaluate state, $\phi = 1$, and the first stage evaluates accordingly while M5 in the second stage is turned on, thus effectively causing this stage to act as an inverter. The logic level of the internal dynamic node, after evaluation, is inverted and passed to the output node, indicated by "Q". Thus, when the clock is high, after latch delay, D_g, the new value is present on the output. Two important features make these latches very useful. First, the transistors

labeled M2 in the figure can be replaced with logic networks, which is a very compact method of implementing pipeline stages. Second, when an n type latch is in the precharge phase a p latch will be in the evaluate phase, and vise versa, thus they are very suitable for pipeline applications, since the required latching action is inherent to their operation. The second point is very important, since it implies that inputs will change only at the beginning of the precharge cycle for successive pipeline stages. This point will be discussed later.

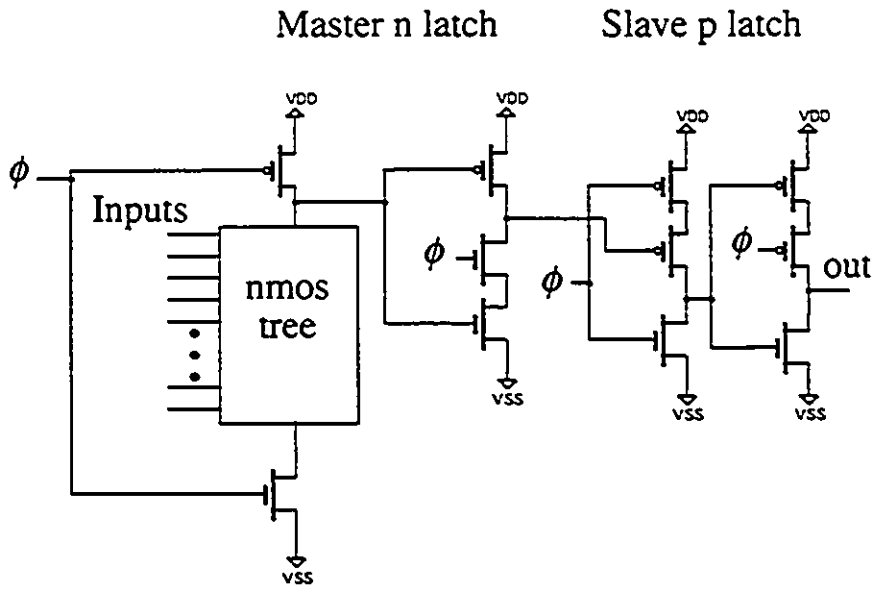
Figure 5.4: TSPC Precharged n and p Latch structures



Any truth table for a logic function can be mapped into a binary decision tree, which is effectively an n dimensional ROM. Jullien has suggested a graph based reduction technique which allows efficient implementation of these trees with reduced transistor count [106], [107]. The combined technique of dynamically pipelining reduced binary trees is referred to as *Switching Trees*. The reduction technique is essentially driven by two rules which minimize the number of transistors needed to implement a multi-output logic function through merging of subtrees and deletion of common edges [106]. This technique provides a minimized structure which lends itself to automatic logic function synthesis and layout. Exclusive use of n transistors in the logic trees has allowed spatially efficient tree realizations due to their high current drive per unit area compared to p transistors. Thus, TSPC dynamic latches offer an attractive technique for implementing structures which have

high functionality per pipeline stage, high throughput, small spatial costs and are pipelined at every bit. The n transistor trees are embedded in the precharged n latch, which drives a slave p latch. The slow PMOS devices have thus been kept to a minimum. Also, this master—slave arrangement is very tolerant to high levels of clock skew as long as reverse clock distribution is used [105], [103]. The arrangement is illustrated in Figure 5.5.

Figure 5.5: Switching Tree embedded in a Master—Slave Latch



The above technique has been successfully used in a mature 3μ CMOS process[1] [106], attaining clock rates of 40 MHz[2] with 9-input trees with a maximum tree height of 6. Key performance limiting factors include charge sharing among the internal nodes of the tree of NMOS devices, and node P in Figure 5.4, and also the pulldown of node P. Due to the clocking technique and latch arrangement, inputs only change at the beginning of the precharge cycle. This is very important in avoiding worst case charge sharing conditions in the NMOS tree. It has been observed that acceptable performance can be obtained with trees as high as 6 transistors in the above technology. This limits logic design flexibility, as well as the functionality per pipeline stage. Thus, although the TSPC scheme has many advantages, the ability to support larger tree heights would be desirable. BATMOS, with its .8μ feature size, should afford better performance; the investigation is reported below.

[1] The technology was CMOS3DLM provided through the Canadian Microelectronics Corporation.
[2] This clock rate was reportedly limited by the i/o pads which were used.

5.4 NMOS Trees In BATMOS

Through tree synthesis of logic functions, useful pipelined arithmetic systems can be formed. The tree height sets the degree of complexity of logic functions which can be implemented, and it is desirable to have large heights in order to facilitate logic design partition flexibility, and to maximize functionality in each pipeline stage. Unfortunately, tree height is limited by the enabling technology due to charge redistribution effects within the tree in a standard TSPC latch. The .8μm feature size available in BATMOS promises reduced parasitic capacitances associated with the active devices; however, the small feature size raises some concerns. Although many process techniques, discussed in Chapter 3, have been used to reduce second order effects, the current drive of the scaled MOS devices may still be reduced from that of a square-law dependence on gate voltage. For this reason, there may be contention between the reduced capacitance, which tends to speed the circuits up, and possible reduced drive, which tends to slow the circuit down. It is necessary to assess which effect is dominant.

Estimation of *switching tree* performance was accomplished by the simulation of single tree paths, as shown in Figure 5.6, with near minimum sized transistors[1]. All simulations were carried out with netlists obtained from mask extracted data. It was found that for these circuits, schematic simulations were potentially misleading due to underestimation of network and node capacitances[2]. Increased loading effects due to subtree merging inherent to the tree reduction algorithm were simulated with a load of additional transistors connected in parallel at either the bottom of the tree, as shown in the figure, or at the top of the chain, to the source of the highest NMOS device. The number of these load devices was chosen to be close to the value of the tree height[3]. Worst case conditions for charge sharing and pull down were simulated in all cases. Three inputs were controlled, and are indicated in Figure 5.6 as top input, tree input, and bottom input. For the remainder of this chapter, these signals will be represented by a vector of the form $\{a, b, c\}$, where a is the logic value of the signal applied to the top input of the test structure, b is the value applied to the tree input, and c is the value applied to the bottom input. The worst case charge sharing effects take place when the following sequence of inputs is applied sequentially:

---

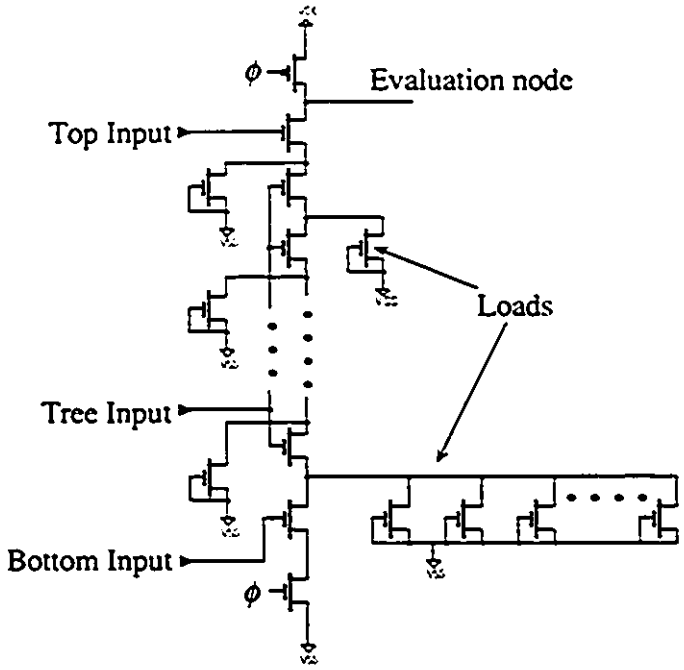
[1] Transistors were used with widths which afforded rectangular device well regions taking into account contact cut requirements. In this case, 1.8μ devices were used.

[2] Although extra capacitance could have been added to the schematic netlists to more closely simulate circuit conditions, the mask extraction was opted for due to its accuracy.

[3] This is based on the observation of several tree synthesized designs, and embodies a fair amount of pessimism.

{0, 1, 1}, {1, 1, 0}. The worst case pulldown occurs when the following inputs are applied sequentially: {1, 1, 0}, {1, 1, 1}. Thus, it is reasoned that if the test circuit functions properly under these input conditions, then a *switching tree* of the same height should operate correctly without compromise due to charge sharing or pulldown.

Figure 5.6: Test Structure For Switching Trees Embedded in TSPC Latch



Realistic simulation conditions were created by supplying all signals through inverters, and inserting a slight delay after the negative going edge of the clock before inputs were applied. This latter point simulates the minor delay associated with p latch evaluation in a pipelined system, and is very important in providing realistic results at high clock rates. A tree height of 16 transistors, loaded at the bottom by 12 more devices as in Fig. 5.6, operated correctly in simulation with a clock period of 12 ns (≈ 83 MHz) and a rise time of 1 ns. The results of this simulation are shown in Figure 5.7. When the same structure was clocked at 100 MHz with a 1 ns clock rise time it failed, as shown in the simulation results presented in Figure 5.8; however, the structure functioned correctly when the rise time was decreased to .5 ns. The tree was also loaded exclusively at the top, by connecting the extra load to the source of the first NMOS device in the chain. This circuit simulated correct operation at 62.5 MHz. It should be noted that this arrangement would not occur with a real switching tree since the majority of merging is carried out in lower stages. It has been found that even though there is substantial charge sharing effects at these tree heights and at

these clock frequencies, as shown in Figure 5.7 (second trace from the bottom), the circuit inevitably failed due to insufficient pulldown. This latter point is illustrated in Figure 5.8, where the output (bottom trace) fails due to pulldown failure (second trace from bottom). With this in mind, it is convenient to use a special measure of comparison.

Figure 5.7: 16 High, Bottom Loaded Test Structure at 83 MHz

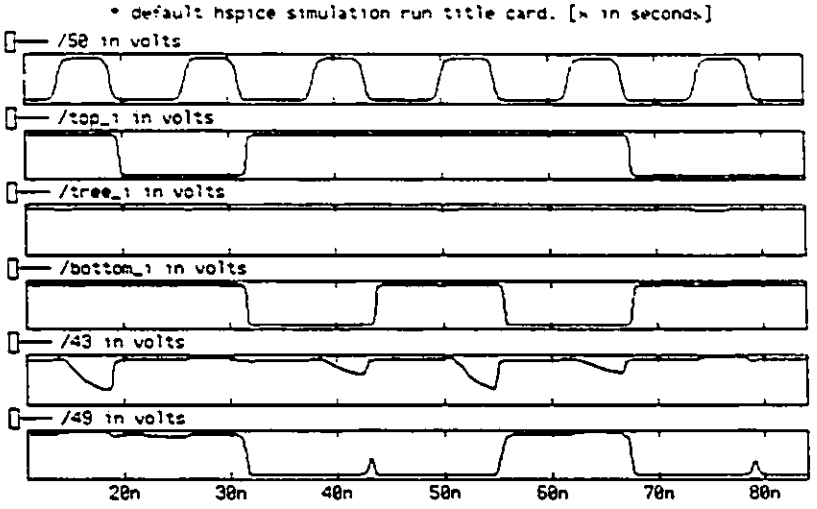
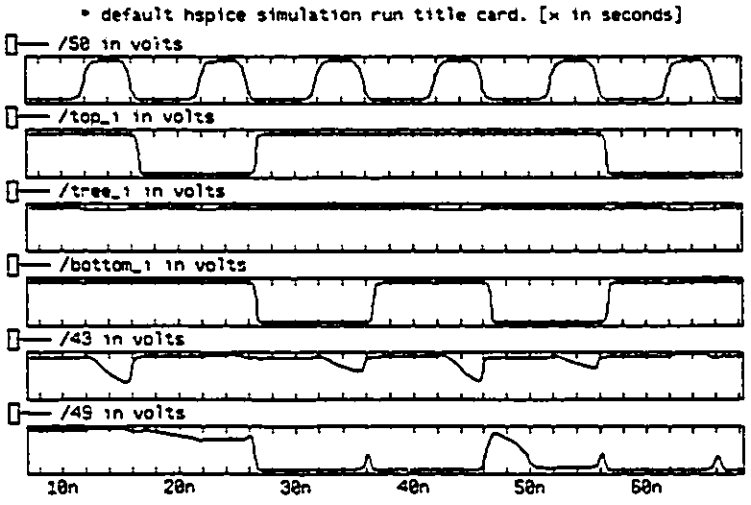


Figure 5.8: 16 High, Bottom Loaded Test Structure at 100 MHz



The need for a special metric of performance for these specific circuits is necessitated by several factors. First of all, the simulations indicate that failure is consistently due to

insufficient pulldown. When clocking limits are approached, signal slew rate can be difficult to accurately obtain due to the deformed shape of the waveforms. Also, due to the difficulty in obtaining an accurate value for the maximum clock rate, a measure which can be obtained easily with finer granularity is desirable. For these reasons, the comparative pulldown value (CPV) is introduced. We define CPV as the maximum pull-down voltage on the dynamic evaluation node during the evaluation phase, throughout the entire test set of input stimuli and responses. Obviously, the lower the CPV the better. The CPV of the bottom loaded structure at 83 MHz was 1.45V, and the CPV of the top loaded structure at 62.5 MHz was 1.35V.

The above results suggest that in a scaled technology such as BATMOS, very large functionality in each pipeline stage is possible with Jullien's tree synthesis technique [106]. In order to lend some credence to the assumption concerning the characterization of an entire switching tree by a single, appropriately loaded n transistor path, an actual synthesized logic block was simulated and compared to the results from test structures. Due to the fact that the BATMOS technology has only recently become available to the university community, there were no such blocks previously synthesized. As a result, programs were written that would convert the design from another technology[1], and for this reason, minimum feature sizes were not used for transistors in the tree[2]. A mod 7 multiplier, which is representative of a 6 high tree structure, was converted and simulated with the same considerations as outlined previously. Unfortunately, it was very difficult to predict which inputs would cause the worst case conditions in the actual tree, so an "educated" guess was made. Also, many random inputs were included as a precaution. Test structures loaded individually at both the top and bottom were also simulated for worst case conditions. Transistor sizes were the same in the tree and the test circuits, and all buffering circuitry was identical. The mod 7 multiplier was successfully clocked during simulation at 125 MHz with a .5 ns rise time on the clock waveform. A clock with a 1 ns rise time and a frequency of 100 MHz also simulated correctly, yielding an CPV of .9V. A portion of these latter simulation results is illustrated in Figure 5.9. It is interesting to note that the charge sharing dip at this frequency was only 300 mV. For comparison purposes, a 1 ns clock rise time will be considered.

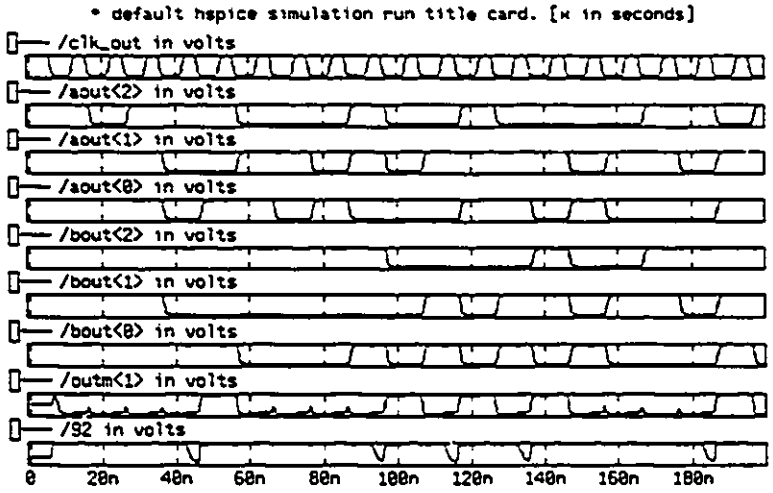
---

[1] See Appendix E.

[2] The design was converted from Canadian Microelectronics Corporation's 1.2μ CMOS4S process. The transistors possessed gate widths of 4.1μ and gate lengths of 1.6μ.

Test structures containing 6-high transistor chains were simulated with top and bottom loading to compare the results to those of the multiplier. The bottom loaded test circuit showed an CPV of .94V, and a charge sharing dip of 700 mV. Both results are worse than the actual switching tree, which suggests that the single-path test structure offers a pessimistic estimate of performance. The top loaded test structure only had an CPV of 1.7V, but had minimal charge sharing dip. Considering the above results, it is apparent that there is pessimism in the simulated charge share value of the bottom loaded structure, and pulldown value of the top loaded one. These results suggest that if both structures yield acceptable results at a given clock frequency, it is a reasonable assumption that a switching tree[1] will function correctly if it is of a height equal to or less than the height of the series connected n transistor chain in the test structure, excluding the ground switch, of course. It is also noted here that virtually all output failures in the simulations were due to insufficient pulldown. This suggests that testing only the top loaded structure may be sufficient, since simulations clearly yield pessimistic results for pulldown. Thus, the single equivalent path assumption for simulating binary tree paths is reinforced by these results.

Figure 5.9: Mod 7 multiplier at 100 MHz



Corner models for the BATMOS process were used to observe the sensitivity of the TSPC latching structures to process variations The top loaded, 6 transistor test structure used above, was simulated at 100 MHz to observe the change in CPV. In going from best to

[1] That is, a switching tree of the type being treated here.

worst process parameters, the CPV changed from 1V to 2.5V. Note that this was the worst case structure for this type of latch given the above mentioned failure mechanism.

The TSPC strategy appears to benefit greatly from the reduced feature size of a scaled process. The results presented in this section support the notion that the effects of parasitic capacitance in these circuits is reduced at a faster rate than the current drive is reduced, as feature sizes are scaled.

5.5 BiCMOS in Dynamic Circuits

Since BiCMOS offers both bipolar and MOS devices on the same substrate, it is of interest to investigate the application of this technology to pipelined systems. The conventional use for bipolar devices is usually within a totem-pole driver configuration [8]. The logic function is implemented with a separate CMOS network, and the bipolar devices are used as low impedance drivers for high capacitance loads [7], [108].

This technique suffers from a full swing problem, which was discussed in section 3.2. A comparatively small amount of work has been done in the area of BiCMOS dynamic circuits. This is not to say that the idea has been ignored. BJTs have been very successfully used in the sense amps of memory chips [3], where they typically are used in a differential pair configuration. A ROM and arithmetic unit [109], as well as several macrocells [110], and a complete microprocessor [108] have been implemented with clocked BiCMOS circuits. In this two phase circuit technique, logic functionality is achieved with CMOS circuits placed between the base and emitter of a bipolar device, which acts as a sensing circuit. An example of this circuit technique is illustrated in Figure 5.10. During the evaluation phase, the MOS transistors may or may not shunt current away from the bipolar base. If they do, the BJT remains off, and the output is high, while if they do not, base current turns on the bipolar device, and the output is discharged low. The speed advantage here is gained through the reduction of the CMOS voltage swing to that found between the base—emitter junction of the bipolar transistor. This technique uses two clock phases, it is not aimed at pipelined structures, and minimal logic functionality has been implemented in the CMOS network.

Figure 5.10: Two Phase Dynamic Circuit

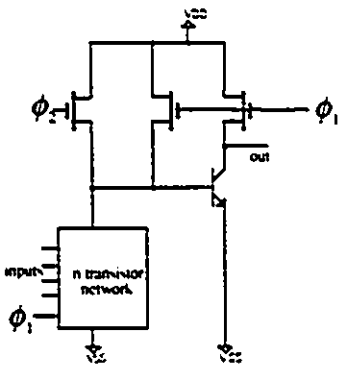
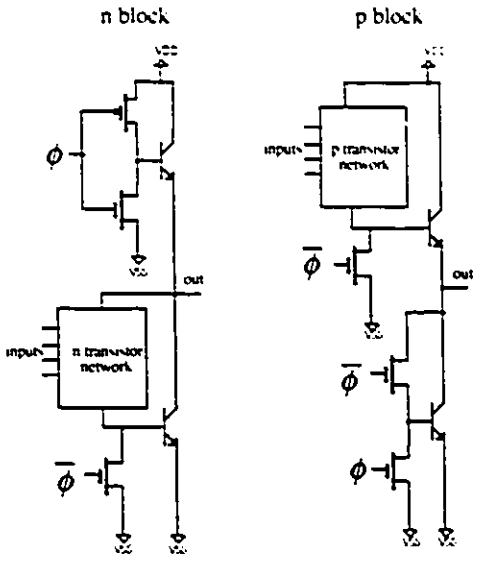


Figure 5.11: Example of Kuo's Circuits



Kuo proposes some interesting dynamic BiCMOS techniques. In [111] and [112] the implementation of carry look ahead circuits are presented. The circuit technique can be likened to a bipolar version of NORA [100], where a pseudo single phase clock is used with alternating n and p blocks possessing opposite precharge-evaluate timing to avoid data race. An illustration of this circuit technique is given in Figure 5.11. Logic is implemented with networks of NMOS and PMOS transistors. In an n block, the NMOS logic network is placed between the evaluation node at the bipolar collector, and the bipolar base. The p blocks are connected similarly. The technique has been applied to fast adders [112], multipliers [113], and fuzzy controllers [114]. The more recent version of this technique used in [115] and [113] is aimed at low voltage operation and uses PMOS transistors in the logic network which allows the realization of non-inverting gates. It replaces the pullup BJT with a PMOS transistor, and has a feedback static NAND gate to control the evaluation of the logic network. This latter point allows the circuit to require only a single clock phase, and is similar to local clock inversion schemes. The BJT pulls down the dynamic node when static current is momentarily switched into the base by the combination of the logic network, and a gated PMOS transistor[1].

A new latching technique will be presented which provides for tremendous logic complexity per stage, uses minimal PMOS logic, possesses no data race, and is free from most charge sharing problems.

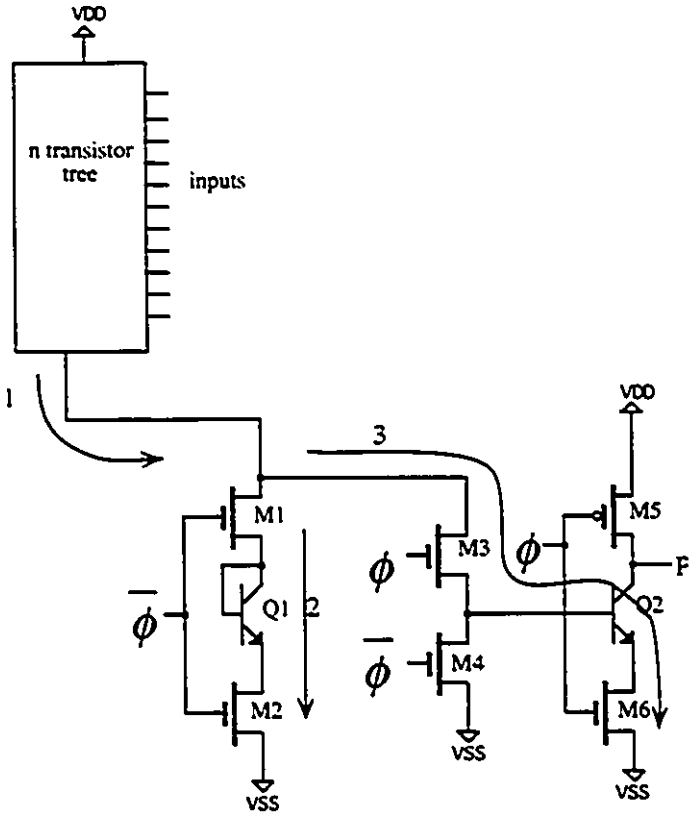
5.6 Current Steering (CS) Latch

Although switching trees, embedded in TSPC latches, do not suffer from true worst case charge sharing, their performance is still limited by voltage sag and insufficient pull down, as discussed in section 5.4. The former is a result of insufficient charge being transferred

[1] This device is gated with the static NAND gate.

to the tree, thus creating a low voltage condition on the evaluate node during the evaluation phase. This can always be corrected in the case of a TSPC latch by increasing the precharge time, however this places constraints on clock shape and system frequency. Insufficient pulldown is due to the fact that a low level on the evaluate node during the evaluate phase is always accomplished by discharging a high value on the precharge node through the tree. Again, slowing down the clock will remedy this problem; this, unfortunately, reduces the throughput rate. To address these problems, a new latching structure is proposed which exploits the high quality bipolar transistors available in a BiCMOS process, such as BATMOS [116].

Figure 5.12: Current Steering Concept



The basic circuit is shown in Figure 5.12. The top of the tree is connected to V_{DD}, and there is no precharge transistor. If the inputs create a conducting path through the NMOS tree, static current will flow. This current is indicated by the arrow numbered 1 in the figure. If $\phi = 0\ (\overline{\phi} = 1)$ then the current follows the path indicated by the arrow labeled with a 2. This is because transistors M1 and M2 are turned on, and transistor M3 is turned off. Note also that M4 is turned on, effectively connecting the base of Q2 to ground and

keeping the device turned off. M5 is turned on, and node P is charged to V_{pp}. These actions constitute the precharge phase. When $\phi = 1$ ($\overline{\phi} = 0$) then the current flows in the path indicated by arrow number 3. Note that transistors M1 and M2 are turned off, and M3 is on, effectively directing the current into the base of Q2 and turning the device on, thus discharging node P. This is the evaluate phase. The above description assumes that the inputs cause a conducting path to be formed through the tree. If there is no such path established, then no current flows[1]. When $\phi = 0$ ($\overline{\phi} = 1$) node P precharges high and when $\phi = 1$ ($\overline{\phi} = 0$), Q2 remains off due to insufficient base current, thus node P remains high. It can be seen that the timing behaviour of this structure is similar to the first stage of a precharged TSPC n latch[2]. When an embedded tree in this latter circuit provides a conducting path, it will evaluate low. In this new structure, if the tree provides a conducting path, the bipolar transistor, Q2, will turn on, and the output node will also evaluate low. Notice the symmetry in the current paths, with each one containing two MOSFETs and a bipolar base—emitter junction.

The new latching structure senses static current flowing in the tree and latches its output accordingly, which is inherently a faster mode of operation for large trees. This can be appreciated by considering the following two cases. For a low value to appear on the evaluation node of a tree embedded in a TSPC latch, it must be discharged through the tree. Thus, a delay is incurred which is dependent on the time it takes to remove sufficient charge to reduce the evaluation node voltage, and this node cannot be sampled before this delay has passed. With the new structure, however, as soon as current begins to flow, it can be detected, and the latch can react accordingly. Another way of looking at this is that the dynamic evaluation node of the current latch is implicitly decoupled from charge redistribution within the tree, and the same small capacitance must be discharged low regardless of the tree height. The bipolar device performs the evaluation node pulldown, not the transistors in the tree. The main disadvantage is a voltage swing that is not rail-to-rail[3]; however, this can easily be remedied by the inclusion of static inverters as will be discussed.

So far only one stage of the latch has been discussed. It would be desirable for the structure to be compatible with the TSPC strategy, and philosophy, in order to utilize the

---

[1] Excluding the normal leakage current associated with MOSFETs, of course.

[2] The obvious difference being that this structure requires ϕ and $\overline{\phi}$.

[3] The bipolar device will saturate, and the node will pull down to within a few hundred millivolts of ground.

clocking advantages mentioned previously. Recall that we have previously used switching trees embedded within a master n latch, with a slave p latch inserted into the pipeline (this allows exclusive use of n transistor logic). We wish to retain this philosophy for the new latch structure. Many different configurations have been explored, through extensive simulation, and the major results are outlined below.

First, consideration must be given to clock compatibility. The situation is subtly different from that of a conventional latch design, in which the first stage is transparent when the second is in the hold state, and vise versa. This is because there is logic embedded in the first stage, so the notion of latch transparency is not identical here. Consider a conventional TSPC pipeline, with a master—slave latch arrangement. In this system, the tree is precharged when $\phi = 0$, and evaluated when $\phi = 1$. The clocked inverter following the tree is transparent when $\phi = 1$, and is in the holding state when $\phi = 0$. Consequently, the new output from the n latch is valid shortly after the rising edge of the clock, with the delay primarily being the evaluation delay of the tree. The first stage of the slave p latch is in evaluation when $\phi = 0$ and precharge (actually, predischarge) when $\phi = 1$, while the final clocked inverter is transparent when $\phi = 0$, and in the hold state when $\phi = 1$. Thus the only time at which inputs to a logic tree which are fed from TSPC pipeline latches can change is at the beginning of the precharge cycle. This is very important in avoiding worst case charge sharing conditions. It would be desirable for the new latch to have similar functionality to that described above, thus making it compatible with the TSPC approach[1].

In this section, the structure shown in Figure 5.13 will be considered for reasons to be discussed later. Since this structure provides precharge and evaluate compatibility with the TSPC latches, but does not constitute a complete latch, it will be referred to as the precharge—evaluate stage, or PE stage. This stage replaces the first stage of the n latch used in the TSPC master—slave arrangement discussed previously.

Several different topologies were investigated in an attempt to implement an n latch with the new current steering structure. Figure 5.14 shows one such structure, which will be referred to as a Double Inverter Current Steering (DICS) n latch. It contains static inverters followed by a clocked inverter. The two static inverters alleviate the partial swing problem.

---

[1] It is recognized that 100% compatibility is impossible with the present structure due to the two clock signals required.

and increase the noise margins of the circuit. During precharge, $\phi = 0$ ($\overline{\phi} = 1$), and node P is precharged high. This high state is transferred through the inverters[1] and turns off M6.

Figure 5.13: Revised Current Steering Circuit

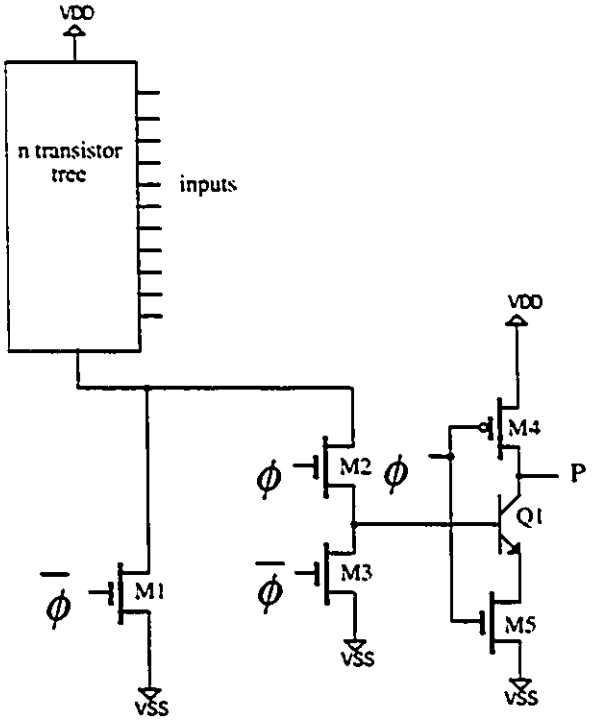
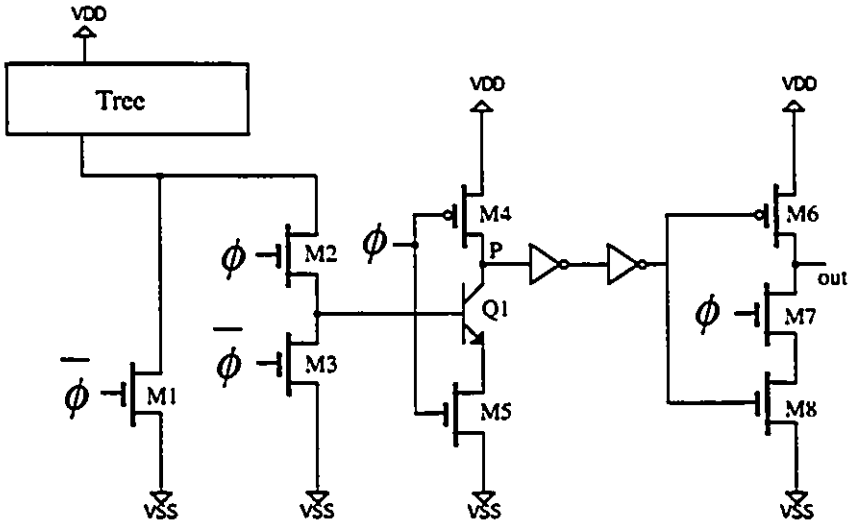


Figure 5.14: DICS n Latch

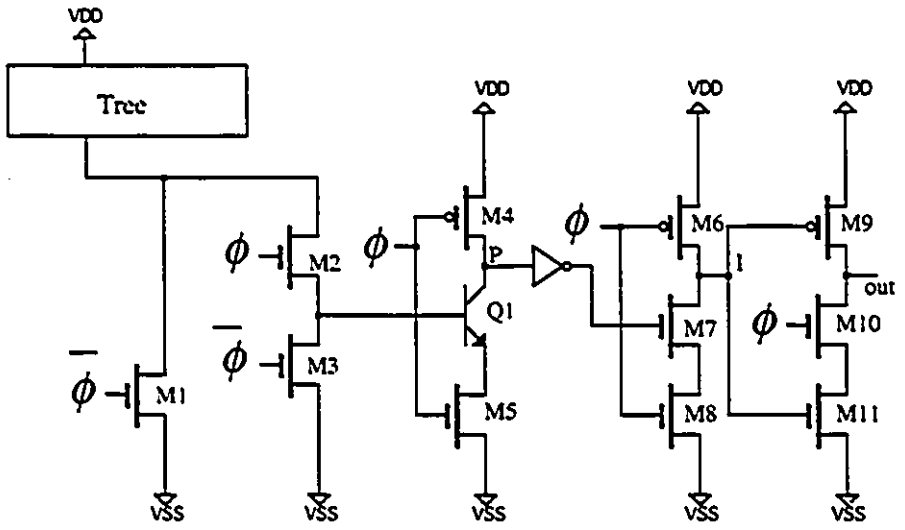


[1] Two inversions is logically the same as no inversion.

Since the clock is low, M7 is also off, and the n latch is in the hold state. During evaluation, $\phi = 1$ ($\overline{\phi} = 0$), and node P takes on the logic value dictated by the implemented function. This value is propagated through the inverters, inverted by the last clocked inverter, and passed to the output. This latter inversion is possible since the clock turns M7 on.

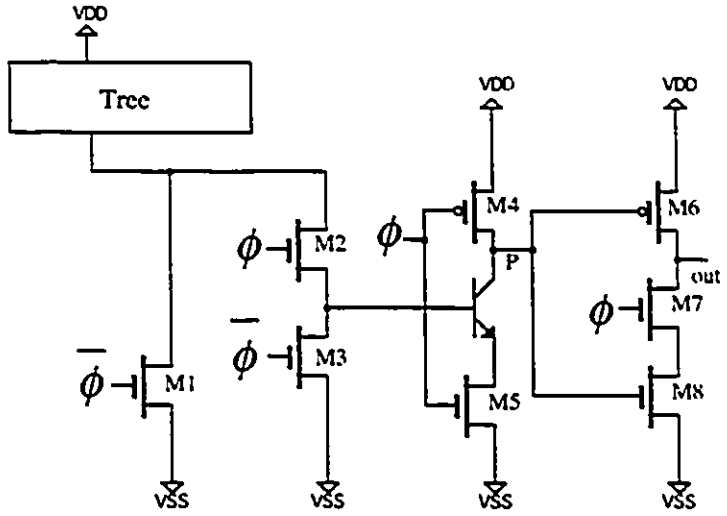
A domino logic compatible configuration was also explored, and is illustrated in Figure 5.15. The circuit style, which we will refer to as a Domino Style Current Steering (DSCS) n latch, is an adaptation of the domino compatible structure suggested in [101]. It contains a static inverter which has the beneficial effects mentioned above. Since node P is charged high during precharge ($\phi = 0$), the output of this inverter drives M7 to the off state, allowing internal node I to be precharged high, thus turning M9 off. Since the clock also maintains M10 in the off condition during precharge, the output is held at a constant value. During evaluation, ($\phi = 1$), node P may or may not pull down, depending on the evaluation of the tree. If it does pull down, then M7 turns on, node I is discharged, and the output node is charged high. If P does not pull down, M8 remains off, node I remains high, and the output node is pulled down. Jullien [106] reports the possibility of reducing switching tree complexity by using complex tree input decoders implemented with domino logic. This gives utility to the above latch structure.

Figure 5.15: DSCS n Latch



The n latch shown in Figure 5.16 has no static inverters, and is referred to as the Completely Dynamic Current Steering (CDCS) n latch[1]. It bears the most resemblance to the types of latch structures used previously with tree synthesized pipelines. During precharge, node P is pulled high which turns off M6, while the clock maintains M7 in the off state. This constitutes the hold state for the latch. During the evaluate phase, M5 is turned on, and node P assumes its value based on tree evaluation, while M7 turns on, which allows values to propagate to the output. This is the most compact structure, but it also lacks the robustness of the others due to the lack of static logic.

Figure 5.16: CDCS n Latch



The above three structures realize n latch functionality and they require a pseudo single phase clock, consisting of a signal and its complement. All of these latches can be paired with a slave p latch of the dynamic TSPC type, shown in Figure 5.4, thus allowing the construction of highly pipelined data paths.

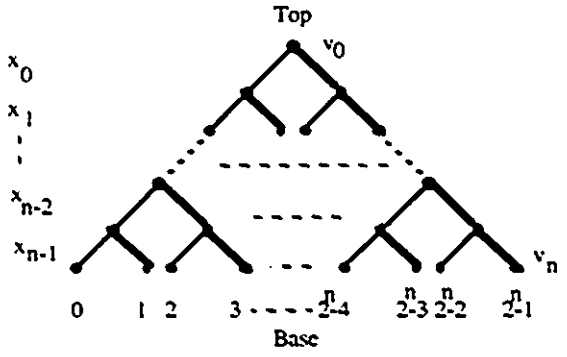
5.6.1 Simulation of Current Steering Latch Structures

The assumption of a single appropriately loaded chain of n transistors characterizing a switching tree was also employed in the simulation of the new latch. There are some subtle differences which should be highlighted. A diagrammatic illustration of a binary tree is

[1] It is recognized that these latch structures operate using static current through the n transistor tree, however the name "Completely Dynamic Current Steering" will be taken to refer to circuitry following the evaluation node P in the diagram.

shown in Figure 5.17[1]. For a silicon implementation, transistors receiving input signals are placed at positions with thick lines, while transistors receiving the complement of these signals are placed at positions indicated by thin lines. If the resulting tree is embedded in a TSPC latch, the evaluation node is at the top, and the ground switches[2] are at the bottom. During a low evaluation of the tree, the dynamic node is discharged by current flowing into the top of the tree, and out of the ground switches at the bottom of the tree. If neighboring trees are merged, current may flow into one tree, and out of the bottom of a shared portion. Because of this fact, there will be a problem in unambiguously determining which tree possesses the conduction path if current is used as the indicator, and if it is sensed at the bottom of the tree.

Figure 5.17: A Binary Tree

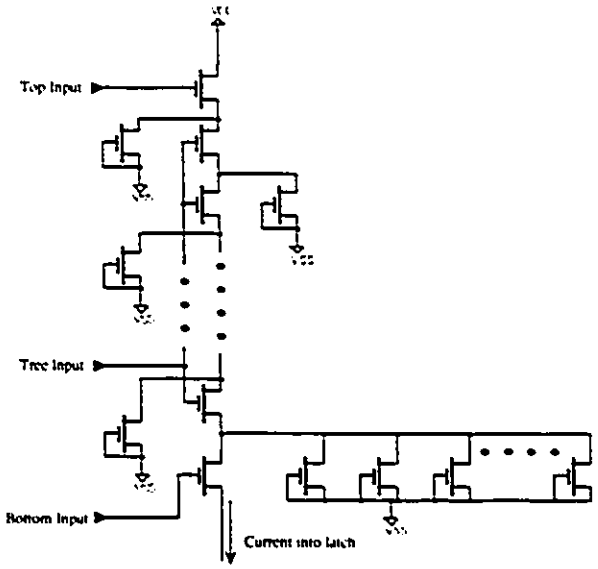


The evaluation failure mechanism is different from the sneak path phenomenon reported in [106], [107]. Sneak paths are a direct result of the graph based reduction algorithm which replaces transistors with wires. They remain a concern with the new latch structures, however, they are separate from the issue being treated here. In the case being discussed, incorrect tree evaluation may result due to current flowing in an entirely different tree (subtree) if current is sensed from the bottom of the tree in the TSPC configuration. The solution is to invert the trees when embedding them in these new current mode latches. The bottom transistors are connected to V_{DD} and the latch detects static current flowing out of the top of the respective trees. There is no precharge transistor, and merged nodes are closer to the top of the resulting tree structure, whereas in the TSPC latch, they are closer to the bottom. The test structure used to study this new latch is pictured in Figure 5.18, which illustrates the bottom loaded scenario. For some simulations, the load was placed

[1] This figure was taken from [107].
[2] The final tree may have merged subtrees and several ground switches.

near the top of the transistor chain, connected to the source of the top most NMOS device. For all simulations to be discussed, mask extracted data was used to realistically portray parasitic effects. Schematic simulations were found to be misleading in the case of these circuits, due to underestimation of parasitic capacitances.

Figure 5.18: Test Structure For New Latch

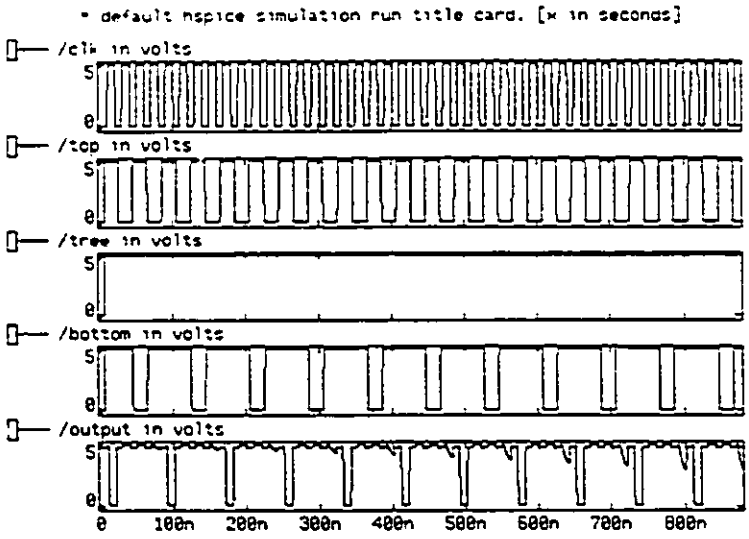


For test structure simulations, three inputs were controlled as before, and are indicated in the figure as top input, tree input, and bottom input. First, precharge and evaluate functionality was verified, with the same input patterns applied for the TSPC latches, and a new worst case input pattern was discovered. If inputs {1, 1, 0} and {0, 1, 1} are applied sequentially over two clock cycles near the beginning of the precharge phase, behaviour specific to this latch results. The first input vector causes the parasitic capacitances associated with MOSFETs above the bottom device in the chain to be charged, but no static current flows because the lowest transistor is turned off. When the next vector is applied, the bottom transistor is turned on and the top one is turned off, thus no static current flows[1], and the bipolar device should not be activated. The parasitic capacitances, however, must be discharged, and do so, during precharge, through path 1 in Figure 5.12. If the latch begins to evaluate before the tree is discharged, the parasitic current will be switched into the base of Q2 in the figure, and this current has been found to be sufficient to momentarily turn on the bipolar transistor. The dynamic node will be correspondingly

[1] Assume for the moment that the static current due to the two transistors changing state simultaneously is zero.

discharged, and a logic error will result. This effect is demonstrated in the simulation results presented in Figure 5.19. The lower trace indicates the voltage on the dynamic evaluation node, labeled P in Figure 5.12, as the discharge time, after the worst case transition, is successively reduced. It has been found that this parasitic discharge can take in the neighborhood of 4 ns to 5 ns to discharge to equivalent voltage levels which do not effect the bipolar device[1].

Figure 5.19: Failure Mode Unique to Initial Version of New Latch



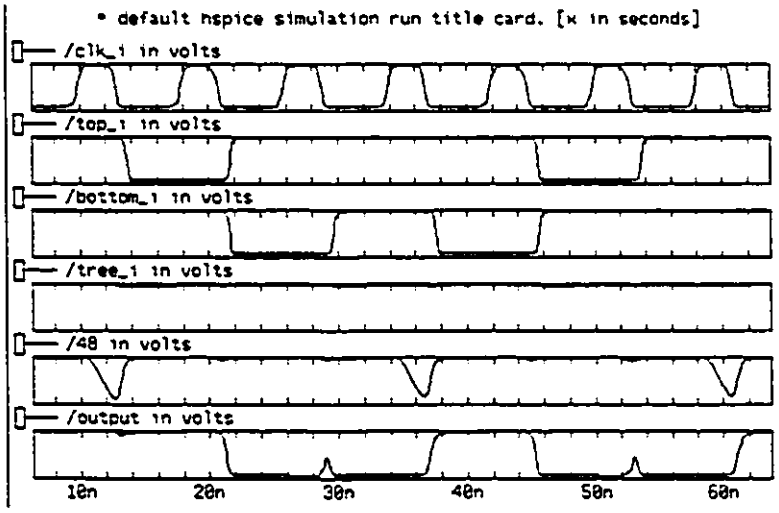
Thus, the parasitic capacitance discharge current limits the clock frequency by placing a constraint on the minimum precharge time. The evaluate time is very fast, since the high drive bipolar device need only pull down the dynamic node, and thus it is found that the above constraint is dominant in the structure pictured in Figure 5.12. For this reason, the topology shown in Figure 5.13 is preferred, since it discharges the parasitic capacitance must faster. Simulations have shown that failure due to this mechanism is not dominant in the structure of Fig. 5.13.

As was practiced for the TSPC latch simulations, all signals were supplied through inverters, and a slight delay after the negative going edge of the clock, prior to applying inputs, was used in order to obtain reasonable results. A transistor chain height of 16 transistors, loaded at the bottom by 12 more devices, as in Figure 5.18, produced correct results when simulated with a CDCS master n latch/slave TSPC p latch using a 125 MHz

[1] This was for a 16 high test structure of 3μm/.8μm transistors.

clock and a rise time of 1 ns under all worst case inputs[1]. When the same structure was loaded at the top with the identical load, the simulation indicated correct operation at the same frequency. In the former case, an CPV of 1.8V was achieved, while in the latter case, the CPV was .5V. This is encouraging since an actual tree would more closely resemble the latter, because it would be upside down as indicated previously. A portion of the simulation results from the latter case are reproduced in Figure 5.20. The trace at the bottom, and second from the bottom, represent voltage on the output node of the p latch and the evaluation node in the n latch, respectively. These results suggest that the bottom loaded case gives a very pessimistic estimate of performance. Both the DSCS n latch and the DICS n latch were simulated with 16 high n transistor chains and 12 transistor bottom loads. They both simulated correct operation using a 9 ns (≈ 111 MHz) clock with a 1 ns rise time. It is expected that these clock rates would be substantially higher for top loaded structures, as the previous simulations would indicate. As well, it would be expected that substantially faster clocks are possible if they posses smaller rise times, based on earlier results. A valuable comparison can be made between these results and the simulations of the TSPC n latches. Identical test structures were used[2], and thus we can compare results from the worst case structures for each type of latch. The simulations indicate that the current steering latches can be successfully clocked at nearly twice the frequency of the TSPC latches.

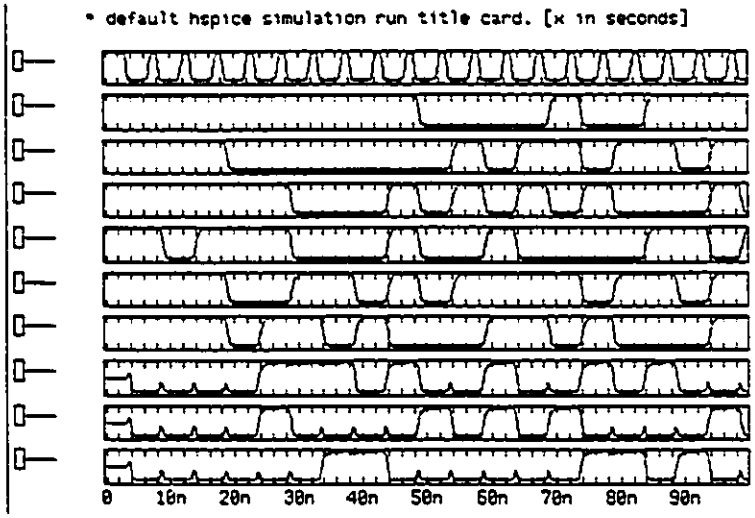
Figure 5.20: CDCS n Latch At 125 MHz



[1] Worst case inputs for both the TSPC latch and the new latch were used in all tests.
[2] Transistor chain height, transistor size, and buffering circuitry were all identical.

As before, in order to test the assumption concerning the characterization of an entire switching tree by a single, appropriately loaded, n transistor path, the same mod 7 multiplier was fitted with the new style of latch. The master latches were CDCS n latches, while the p latches were dynamic TSPC style. Test structures with a tree height of 6, which were individually loaded at both the top and bottom, were also simulated for worst case input conditions. Transistor sizes were all identical in the tree and the test circuits, as was all buffering circuitry. The mod 7 multiplier demonstrated correct simulated operation with a 5 ns clock (200 MHz) possessing a 1 ns rise time[1]. A portion of the results from this simulation are reproduced in Figure 5.21. At this frequency, the CPV was 1.1V. Both test structures performed worse than this, which is encouraging since it suggests that they represent pessimistic models of the switching tree structure. As before, the bottom loaded structure performed the worst, failing in the simulation due to insufficient pulldown with an CPV of 2.34V. The top loaded structure passed the test, but had a CPV of 1.6V, which is approximately 45% worse than the CPV for the multiplier. These results suggest that the test structures actually model the switching tree behaviour very pessimistically, and thus, the earlier results for 16 high trees are very encouraging.

Figure 5.21: Mod 7 Multiplier at 200 MHz

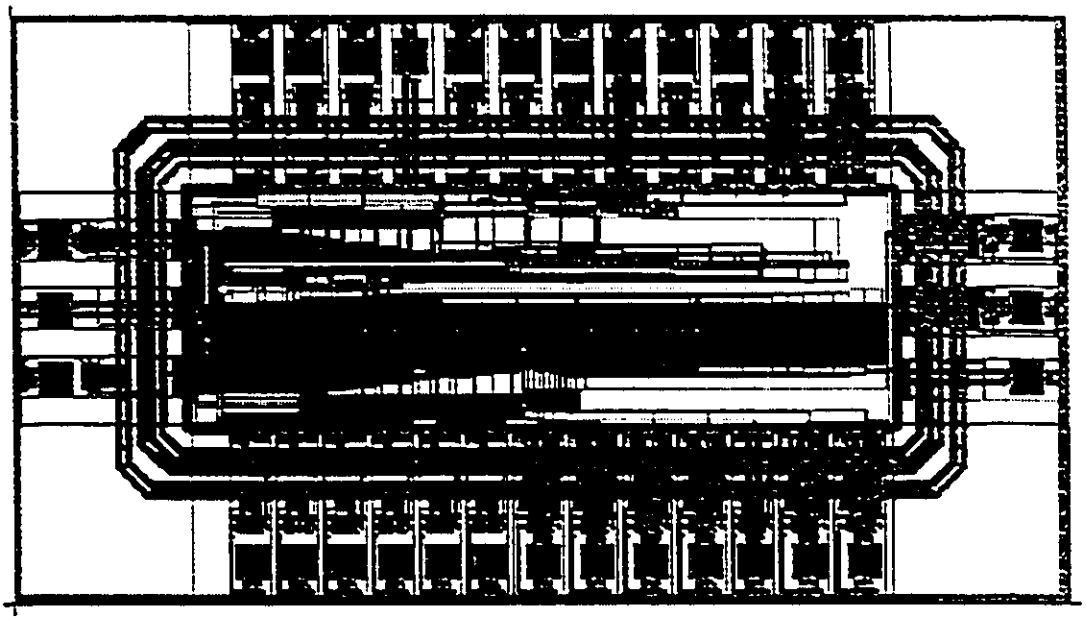


Corner models for the BATMOS process were used to observe the sensitivity of the new latch structures to process variations. The top loaded, 6 transistor test structure used above was simulated at 100 MHz to observe the change in CPV. In going from best to worst case

[1] The clock was stretched out to 6 ns with a 1 ns rise time to achieve correct simulated operation using worst case process parameters.

process parameters, the CPV changed from 2.36V to 4.13V, a difference of 1.77V. This is compared to the 1.5V change observed with the TSPC latches. These differences were observed at different frequencies, and thus may constitute an unfair comparison. A set of 4 simulations was carried out with the worst case structures for each type of latch under best and worst case process parameters, using a clock with a period of 30 ns (= 33 MHz) in order to obtain well defined pulldown wave shapes in both cases. The pulldown slew rate (SR) was measured for all cases. It was found that the percentage change[1] in the slew rate for the TSPC latch was 45.85%, while that of the CDCS latch structure was 52.8%. Thus it appears that the new latch is more sensitive to process variations than the TSPC type.

Figure 5.22: Mod 17 Multiplier Chip



A chip was submitted for fabrication, pictured in Figure 5.22, which contains several of the above test structures discussed in previous sections, as well as a mod 17 multiplier. The design for the multiplier core was created using the tree synthesis technique mentioned earlier, and the layout was generated automatically [117] in another technology. The conversion procedure outlined in Appendix E was executed, and the converted core was the starting point for the chip partitioning. Latches, inverters, and buffers were added

1 $\%\,change = \frac{SR_{BEST} - SR_{WORST}}{SR_{BEST}}$

manually, and the design was placed in the pad frame using the Edge™ Place and Route tool. Implementation specific details of this chip are given in Appendix G.

5.6.2 Ultra Fast Current Steering Latch Structure

A modification to the basic PE stage of the latch structures previously described is proposed here which attains even higher performance; it is also completely compatible with all previously mentioned topologies. Latches based on this structure will be referenced by including a preceding U in the acronym. For example, a DSCS n latch containing this modified structure will be referred to as a UDSCS n latch.

Figure 5.23: Modified PE Section

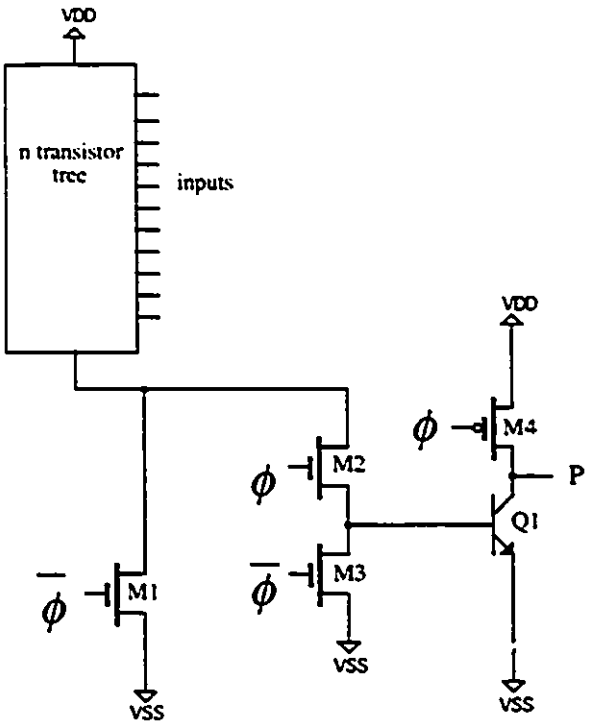
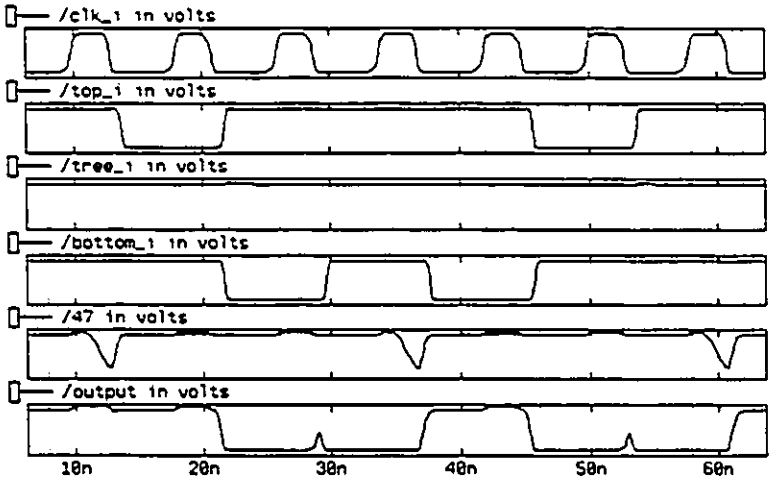


Figure 5.23 shows the new circuit. The modification consists of the removal of the clocked MOSFET which was previously connected between the emitter of Q1 and ground. The operation of this structure is identical to the previous one, but its speed is considerably higher due to a much quicker pulldown of the dynamic node labeled P. This latter point was confirmed by simulation. The structure was simulated as before, taking previously outlined precautions to obtain realistic results. A UCDCS n latch in a master—slave configuration with a TSPC p latch was simulated in a test structure containing a 16 high chain of near minimum sized transistors. The structure was loaded at the bottom by 12 more devices of the same size. Recall that this is the worst case configuration for these current steering structures. The circuit was simulated at a clock rate of 125 MHz with the clock possessing a 1 ns rise time, and the CPV rendered was .81V. A portion of the results from this simulation are reproduced in Figure 5.24. These results can be compared with previous simulations of the other structures. Recall that under identical simulation

conditions[1], the previous current steering design only had an CPV of 1.8V. This constitutes a very significant improvement. The mod 7 multiplier was fitted with latches using this new PE section, and it performed correctly when simulated with a clock of 250 MHz, and a 1 ns rise time. A portion of these results are given in Figure 5.25, showing from the bottom, the output voltage of bit 0, the dynamic evaluation node for this bit, two input bits, and the clock signal. Note the strong pulldown even at these frequencies. When the clock rate was increased beyond this point, the large inverters used to buffer the clock signal began to fail.

Figure 5.24: Test Structure at 125 MHz



5.7 Additional Discussion

Since the operation of these new latches is based on static current flowing through the tree[2], they benefit from factors which are somewhat orthogonal to those in the TSPC case. Charge sharing is not nearly as big a concern, and precharge transistor size is not an issue. For quick pulldown, it is desirable to switch a large current into the base quickly, thus turning the bipolar device on quickly. This suggests that minimum sized transistors are not necessarily optimal in high trees, since the equivalent channel resistance of the structure will be high. It can be postulated that for very high trees, larger transistors may be necessary to operate these latches optimally.

---

[1] Everything was identical including the buffering circuitry, transistor sizes and chain length, load devices, clock rate and clock rise time.

[2] In all cases for the near minimum sized transistor trees, the static current was less than 100μA.

Figure 5.25: Mod 7 Multiplier at 250 MHz

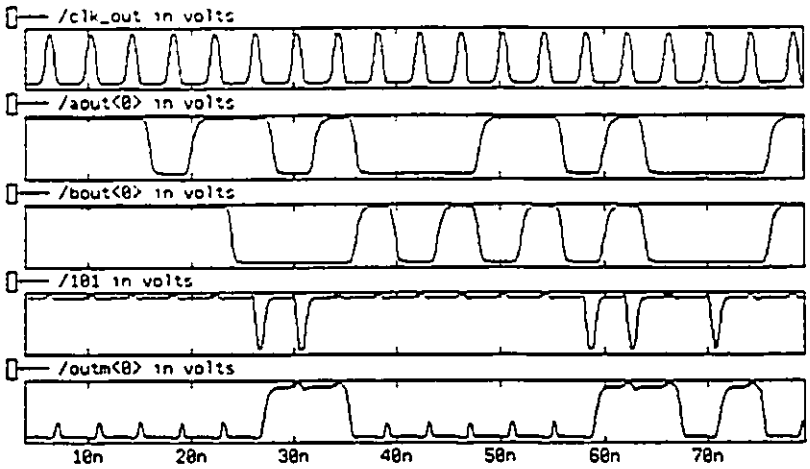
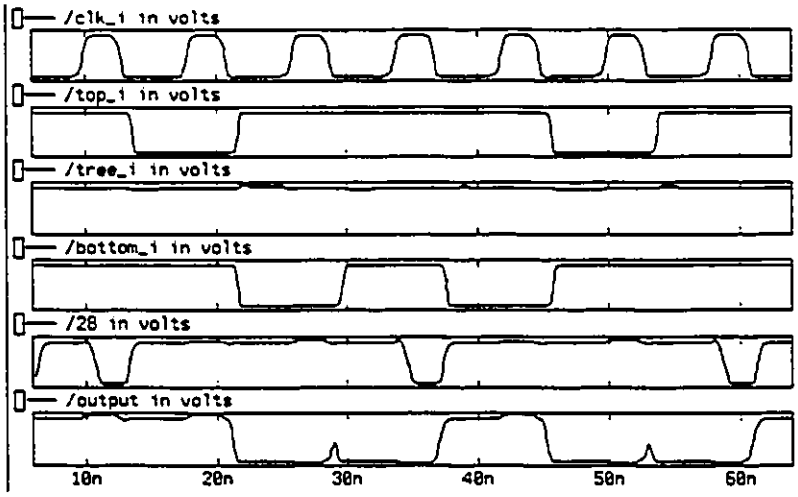


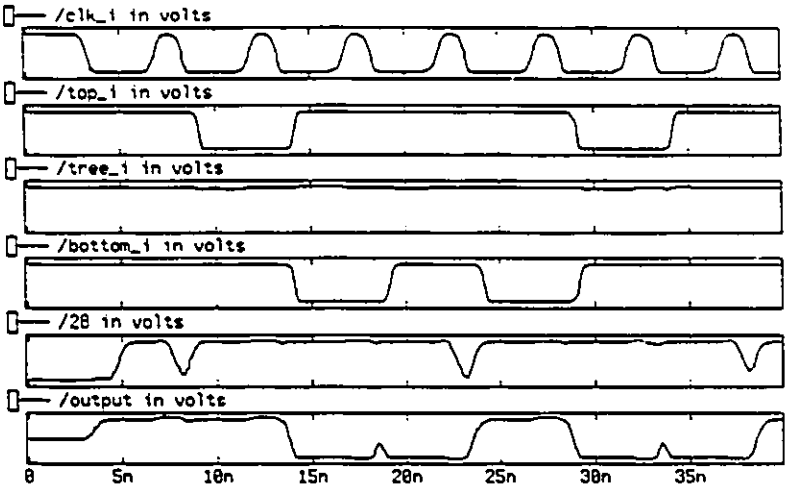
Figure 5.26: UCDCS Test Structure With Large Transistors at 125 MHz



This hypothesis was reinforced through simulation. A 16 high chain of near minimum sized NMOS transistors, loaded at the bottom by 12 more, was simulated at 125 MHz and a 1 ns clock rise time using the new PE section in a UCDCS structure. Recall that this was the worst case load positioning for this structure. The CPV was .81V under these conditions, and the structure failed when a 7 ns clock was used. When the transistors in the tree were replaced by 5μm/.8μm devices, the CPV at 125 MHz was reduced to .58V. This constitutes a significant performance increase. A portion of the results from this simulation are presented in Figure 5.26; note that the dynamic node, whose voltage is

represented second from the bottom, pulls down to its minimum in a fraction of the clock cycle. This structure, using the larger transistors, actually simulated correct operation up to a clock rate of 200 MHz, where it rendered a CPV of 1.29V. A portion of the results from that simulation are shown in Figure 5.27. Note the deterioration in the pulldown waveform compared to the previous simulation.

Figure 5.27: UCDCS Test Structure With Large Transistors at 200 MHz



Since earlier simulations indicated insufficient pulldown as the failure mechanism in the TSPC structures at high clock rates, these circuits should also benefit from a decreased equivalent channel resistance. A 16 high chain of transistors, loaded at the top with 12 more, was simulated. This was the worst case configuration for pulldown in this structure; the transistors had W/L ratios of 5μm/.8μm. The simulation used a clock period of 12 ns with a 1 ns rise time, and a portion of the results are reproduced in Figure 5.28. The CPV is .53V, as compared to a value of 1.45V in the circuit with near minimum transistors discussed previously, under identical simulation conditions. Charge sharing was comparable in both cases.

Since PMOS transistors have been used in the past to implement logic functions in pipelinable dynamic logic structures [112], [101], [100], it is natural to investigate the viability of this idea in the new latch. In Jullien's technique [106], PMOS logic is confined to the simple slave p latch; however, it is useful to investigate if this is a wise practice with the new latch structures. A simulation was run with a 16 high chain of near minimum sized PMOS transistors, loaded with 12 more at the bottom, in a similar configuration to the

NMOS chains. Note that this is the worst case pulldown configuration for the NMOS chains. A portion of these results are presented in Figure 5.29.

Figure 5.28: TSPC Test Structure With Large Transistors at 83 MHz

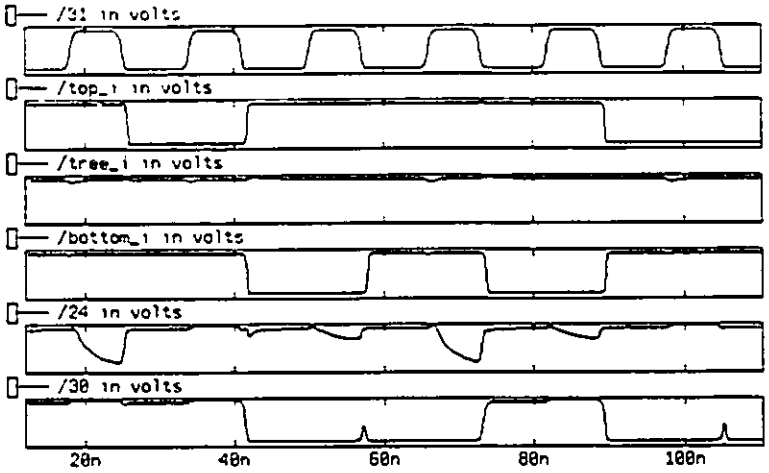
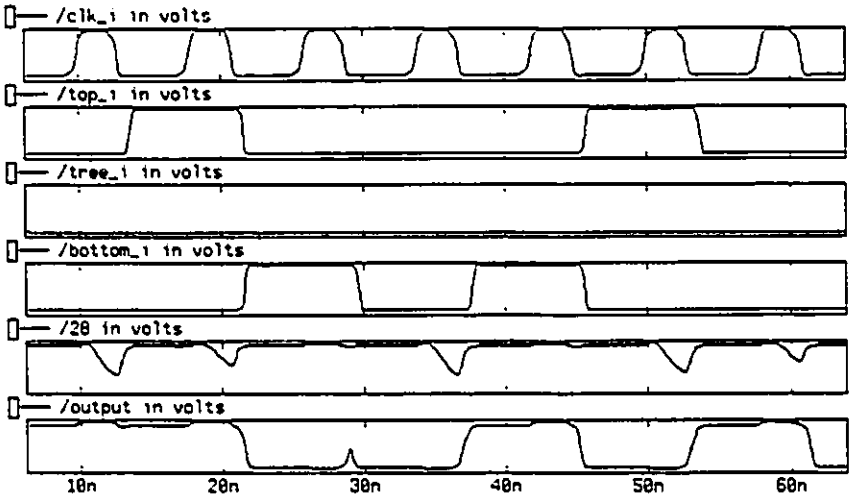


Figure 5.29: Latch With PMOS Chain



The clock has a period of 12 ns and a 1 ns rise time. The results indicate that both pulldown and charge sharing are much worse for this configuration. Note the error at the output node, indicated by the bottom trace, at time $\approx 52ns$. This error is due to the unique failure mode for this style of latch. Recall that the problem was all but eliminated in the NMOS case. This simulation suggests that direct substitution of PMOS trees for NMOS would be imprudent.

The BATMOS performance levels suggested by the simulations in this chapter are much higher than those obtained in more mature technologies, such as CMOS3DLM. Some justification for this can be obtained by performing a rudimentary estimate. The expression for the transconductance of an NMOS device is [118]:

$$g_m = \mu_n C_{ox} {}^{W}\!/\!_{L} (V_{GS} - V_{TN}) \qquad (5.1)$$

Thus, $\kappa = \mu_n C_{ox} {}^{W}\!/\!_{L}$ can be used as a figure of merit for MOS devices in a certain technology, since it is a relative indicator of performance. Since both technologies are 5V, and the threshold values for each are within one or two millivolts, they can be compared on this basis. Calculations were performed using HSPICE™ Level 3 model parameters for both BATMOS and CMOS3DLM. The low field bulk mobility was used for the κ calculations. The mobility in equation 5.1 is actually the effective surface mobility, which takes into account surface scattering caused by normal fields[1]. The parameter which accounts for this[2] is almost twice as large in CMOS3DLM [11] than in BATMOS [22], and thus mobility degradation is more severe for the former. For this reason, consider the following discussion to be optimistic for CMOS3DLM. The equations for effective gate length and width are given in [119], as are other pertinent equations. Consider a minimum sized transistor in CMOS3DLM which has a gate length and width of 3μm[3]. For this device, $\kappa = 3.49 \times 10^{-5}\ {}^{A}\!/\!_{V^2}$ from calculated values, and $\kappa = 2.61 \times 10^{-5}\ {}^{A}\!/\!_{V^2}$ from measured data[4]. The same transistor in BATMOS will have $\kappa = 5.16 \times 10^{-5}\ {}^{A}\!/\!_{V^2}$, almost a 1.5 times increase in current drive simply from changing technologies. If we consider the near minimum devices used extensively in the simulations of this chapter, they have ${}^{W}\!/\!_{L} = {}^{1.8\mu m}\!/\!_{.8\mu m}$, and, $\kappa = 1.01 \times 10^{-4}\ {}^{A}\!/\!_{V^2}$. This is nearly 4 times the empirically based figure for CMOS3DLM, and almost 3 times the calculated value. Combine this with the fact that parasitic capacitances are dramatically reduced, due to the small feature size, and it becomes apparent why the tested circuits perform so well in BATMOS compared to CMOS3DLM.

The new latching structures presented in this chapter differ in several ways from the BiCMOS techniques outlined in section 5.5. These structures do not use BJTs in a differential pair as is common on memory chips [3]. The new latches do not require two

---

[1] See Chapter 3.
[2] Parameter THETA, see page 7-17 of [119].
[3] Actual microns, not design scale microns are being referred to.
[4] Two model parameter values are presented in [11], with one being obtained empirically.

phase clocking as in [108], [110], and [109]. They also differ from the structures suggested by Kuo in significant ways. Minimal logic functionality per stage has been demonstrated with Kuo's structures, while the potential for much improved functionality with these new structures is evident. The structures that Kuo recommends for use in pipelined applications [112], [111] use two bipolar devices to pullup and pulldown the dynamic node. The simulations presented previously suggest that in large tree structures, insufficient pulldown is the major cause of circuit failure. This suggests that the bipolar pullup is unnecessary, and thus the new current steering latches only use one pulldown bipolar device. The designs in [113] and [115] are aimed at low voltage operation, and extensive use of PMOS logic is present. The new structures proposed here are aimed at massively pipelined arithmetic structures, and PMOS logic is avoided. The prudence of this latter point is underscored by a simulation which was presented above. The pulldown of Kuo's structures in [113] and [115] are activated by a momentary current pulse controlled by a static NAND gate which is connected in a feedback configuration. The new latches are activated by the presence of static current flowing through the tree, and no gating of the clock, or feedback is present. Finally, the structures proposed here combine logic functionality inside of a latch circuit. This is not explicitly done elsewhere using a similar circuit.

5.8 Summary

This chapter has outlined clocking techniques applicable to pipelined arithmetic structures. The true single phase clocking strategy was elaborated upon, including explanation of the functionality of basic latch structures associated with the technique. The concept of embedding complex trees of n transistors, called switching trees, inside an n latch was explained. Also, it was discussed how pipelined structures can be realized by employing this technique in a master—slave latching arrangement. Implementation of this technique in BATMOS was explored through simulation. It was found, through the use of mask extracted data from test structures, that 16 high switching trees of near minimum sized transistors operate correctly at clock rates of 62.5 MHz as a very pessimistic estimate, and 82 MHz as a more optimistic estimate. Comparison between results obtained from test structures and those obtained from a mod 7 multiplier were found to correspond well. Several new latch structures were suggested, based on a current steering principle which exploits the bipolar devices present in a BiCMOS process. Simulation of test structures indicated that these latches can be successfully clocked at twice the frequency of the TSPC latches. Test structure results were compared to those from a mod 7 multiplier, and were found to give pessimistic performance indication in all cases. Corner models for the

process were used, and it was found that the current steering structure is more sensitive to process variation than the TSPC structures. An ultra fast latch structure was suggested, based on a modification of the previous current steering structure and simulations indicate its performance to be substantially higher than the previous structures. Overall, the simulation results suggest that these new latching structures are very attractive for pipelined systems.

Chapter 6

CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORK

6.1 Conclusions

This thesis work has accomplished the goals set forth at the beginning. The first objective was to provide a survey of current BiCMOS processing techniques and methodologies, as well as process considerations which determine device behaviour, and this was accomplished in chapters 2 and 3. Silicon and silicon processing fundamentals were treated briefly, and the evolution of BiCMOS from a CMOS processing base was discussed. Several important advanced processing techniques were treated, and Northern Telecom's BATMOS process was described in detail. Issues pertaining to BiCMOS device scaling, device design, and process factors were then discussed in chapter 2. Scaling techniques for both MOS and bipolar devices were considered. Second order active device issues which become important in scaled BiCMOS technologies were discussed in depth. The topics were divided into issues pertaining to MOS devices and those pertaining to bipolar transistors. Process tradeoffs necessary due to inclusion of both types of devices on the same substrate were then treated. Finally, latchup in a CMOS process was briefly reviewed, and the phenomenon was then discussed from a BiCMOS perspective. Latchup modes specific to BiCMOS technologies were covered, and it was explained why, even though there are more mechanisms for latchup in BiCMOS, the process is generally more resistant.

The second thesis objective was to design several high performance arithmetic macrocells. Toward this end, theory was discussed with regards to multiplication and addition hardware architectures. Two major types of hardware multipliers were discussed, namely, the linear parallel type, and the column compression type. Two new architectures which have been recently proposed were explained in detail, since they were the designs which

were implemented as macrocells. This work was in partial fulfillment of a contract, which was successfully executed[1]. The contract required the delivery of several representations and files associated with the cells. Six different macrocells were designed and verified. These designs included implementations in CMOS, BiCMOS, and ECL. A chip containing one of these multiplier macrocells was submitted for fabrication.

The third thesis objective was to investigate the application of BiCMOS technology in the realization of highly pipelined arithmetic structures. This was accomplished, and the related work was summarized in Chapter 5. Clocking strategies for pipelined systems were briefly reviewed, and the true single phase clocking methodology was elaborated on, since it is the foundation for a powerful synthesized, pipelined design style for arithmetic structures. The technique was evaluated in Northern Telecom's BATMOS technology, and it was found through simulation that performance of these structures should dramatically increase in this technology. Tree test structures rendered correct simulation results with clock rates in excess of 62.5 and 83 MHz, depending on the load configuration. A new latching structure based on a current steering concept was introduced and several new structures which are compatible with the precharge—discharge timing of the TSPC latches were proposed. A new measure of performance, called the Comparative Pulldown Value (CPV) was introduced, and this metric was used to compare the different simulation data. Simulation of test structures revealed possible clock rates of almost double those possible with the TSPC method. Results of test structures and a mod 7 multiplier were compared, and it was found that in all cases, the test structures performed more poorly than the actual tree structure they were designed to simulate. This suggests that these structures offer pessimistic performance estimates for the corresponding switching trees. Process corner parameters were used in simulations, and the results suggest a slightly higher process variation sensitivity with the new latch. A chip was submitted for fabrication containing a mod 17 multiplier, as well as several test structures. Finally, a new ultra fast latching structure was presented which performed substantially better than the other current mode latches. A tree synthesized mod 7 multiplier was successfully clocked at 250 MHz in simulation using these ultra fast latches. It was discovered that tree test structures using these current mode latches, as well as the TSPC latches, actually benefited from the use of larger transistors in the series chain. This suggests even higher performance is possible from the structures which were treated. Finally, rough calculations were presented which lend credence to the simulation results observed.

---

[1]See letter of thanks (email) in Appendix D.

The results obtained from simulations carried out in the course of this thesis work suggest that switching tree heights of 16 are possible with TSPC latches using near minimum sized transistors when they are clocked between 50 and 100 MHz. The new current mode latches should operate with the same tree height at substantially more than twice this frequency, depending on the latch structure which is used.

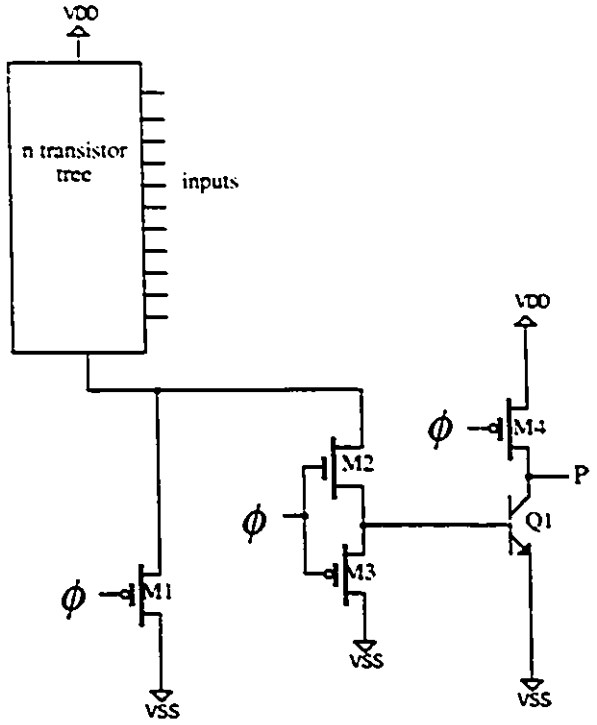
6.2 Future Work

There are several extensions to this work which would be useful. The macrocells should be re-characterized with up to date timing information, since at the time of contract completion, the information available was incomplete. This is a simple matter of re-running the simulations using existing netlists, and the new library information. As well, the chip which was submitted should be tested to verify functionality, as well as to provide a valuable feedback of information on the reliability of the timing information.

The new latches which have been suggested here offer many possibilities for future development. They provide potential for greatly increased functionality per pipeline stage. Further characterization of these structures is necessary, and test results from the chip which was submitted will be valuable in this respect. More exhaustive comparisons of switching tree blocks and the test structures are necessary to increase confidence in the results that were predicted by the simulations in this work. Also, a study of the power dissipation of these latches is necessary.

It would be convenient if the current steering latches could be clocked with a true single phase clock. Toward this end, an investigation of the sensitivity of the structures to slight phase lags caused by local clock inverters should be investigated. Alternately, the structure could be modified by replacing the NMOS devices M1 and M3 in Figure 5.13 with PMOS devices which would effectively eliminate the need for the $\overline{\phi}$ clock phase, thus making this structure an authentic true single phase clocked circuit. The new circuit is pictured in Figure 6.1. Although PMOS devices suffer from poor drive, the resulting performance compromise may be an equitable tradeoff for the reduction to one clock phase. Finally, the circuits suggested in [113] and [115] warrant further attention. They should be investigated to see if they provide a more optimal solution for the pipelining challenges discussed in this thesis.

Figure 6.1: Suggestion For TSPC Version of Current Steering Latch.



References

1. Gallia, J., et. al. "High Performance BiCMOS 100K-Gate Array." *IEEE Jnl. Solid State Circuits*. vol. 25 No. 25 pp. 142-149, 1990.

2. Matsui, M., et. al. "An 8-ns 1-Mbit ECL BiCMOS SRAM with Double-Latch ECL-to-CMOS-Level Converters." *IEEE Jnl. Solid State Circuits*. vol. 24 No. 5 pp. 1226-1231, 1989.

3. Watanabe, S., K. Sakui, T. Fuse, T. Hara, S. Aritome, K. Hieda. "BiCMOS Circuit Technology for High-Speed DRAM's." *IEEE Jnl. Solid State Circuits*. vol. 28 No. 1 pp. 4-8, 1993.

4. Kurita, K., T. Hotta, T. Nakano, N. Kitamura. "PLL-Based BiCMOS On-Chip Clock Generator for Very High-Speed Microprocessor." *IEEE Jnl. Solid State Circuits*. vol. 26 No. 4 pp. 585-589, 1991.

5. Geppert, L. "Not Your Father's CPU." *IEEE Spectrum*, vol. 30 No. 12 pp. 20-23

6. Kubo, M., I. Masuda, K. Miyata, K. Ogiue. "Perspective on BiCMOS VLSI's." *IEEE Jnl. Solid State Circuits*. vol. 23 No. 1 pp. 5-11, 1988.

7. Wissel, L., E. Gould. "Optimal Usage of CMOS Within a BiCMOS Technology." *IEEE Jnl. Solid State Circuits*. vol. 27 No. 3 pp. 300-306, 1992.

8. Deierling, K. "Digital Design." BiCMOS Technology and Applications. Alvarez ed. 1989 Kluwer Academic Publishers.

9. Flores, G., B. Kirkpatrick. "Optical Lithography Fends off X Rays." *IEEE Spectrum*. vol. 28 No. 10 pp. 24-27, 1991.

10. Smith, H., M. Schattenburg. "X-Ray Lithography From 500 to 30nm: X-Ray Nanolithography." *IBM Journal of Research and Development*. vol. 37 No. 3 pp. 319-327, 1993.

11. "Guide to the Integrated Circuit Implementation Services of the Canadian Microelectronics Corporation.", Canadian Microelectronics Corporation, report #GICIS Version 4:0, March, 1989

12. Lei, K., G. A. Jullien, W. C. Miller. "An intelligent Optical Sensor Realization." *The 36th Midwest Symposium on Circuits and Systems*. In Press, 1993.

13. Bar-Lev, A. "Semiconductors and Electronic Devices." 1979 Prentice-Hall, Inc.

14. Liu, T., et. al. "A Half-Micron Super Self-Aligned BiCMOS Technology For High Speed Applications." *IEDM Tech. Digest*. pp. 23-26, 1992.

15. Aoki, T. "A Discussion on the Temperature Dependence of Latch-Up Trigger Current in CMOS/BiCMOS Structures." *IEEE Trans. on Electron Devices*. vol. 40 No. 11 pp. 2023-2028, 1993.

16. Embabi, S., A. Bellaouar, M. Elmasry. "Digital BiCMOS Integrated Circuit Design." 1993 Kluwer Academic Publishers.

17. Leblebici, Y., S. Kang. "Modeling and Simulation of Hot-Carrier-Induced Device Degradation in MOS Circuits." *IEEE Jnl. Solid State Circuits*. vol. 28 No. 5 pp. 585-595, 1993.

18. Dillinger, T. "VLSI Engineering." 1988 Prentice Hall.

19. "Primer on CMC 0.8-Micron BiCMOS, a Version of NTE BATMOS.", Canadian Microelectronics Corporation, ICI-039R00, 1992

20. Allen, P., D. Holberg. "CMOS Analog Circuit Design." 1987 Holt, Rinehart and Winston, Inc.

21. Hadaway, R., P. Kempf, P. Schvan, M. Rowlandson, V. Ho, J. Kolk, B. Tait, D. Sutherland, G. Jolly, I. Emesh. "A Sub-Micron BiCMOS Technology for Telecommunications." *21st European Solid State Device Research Conference (ESSDERC)*. pp. 513-516, 1991.

22. "Micronet/CMC BiCMOS Design Environment.", Micronet, April, 1993.

23. Nagata, M. "Limitations, Innovations, and Challenges of Circuits and Devices into a Half Micrometer and Beyond." *IEEE Jnl. Solid State Circuits*. vol. 27 No. 4 pp. 465-472, 1992.

24. Dobberpubl, D. W., et. al. "A 200-MHz 64-bit Dual-Issue CMOS Microprocessor." *IEEE Jnl. Solid State Circuits*. vol. 27 No. 11 pp. 1555-1567, 1992.

25. McLeod, J., ed. "Changing the Chemistry Makes 0.05 Micron Transistors Possible." *Electronics*. vol. 66, pp. 3

26. "Guide to the Integrated Circuit Implementation Services of the Canadian Microelectronics Corporation.", Canadian Microelectronics Corporation, report #GICIS Version 3:0, January, 1987

27. Bakoglu, H. B. "Circuits, Interconnections, and Packaging for VLSI." 1990 Addison-Wesley Publishing Company Inc.

28. Prince, B., R. Salters. "IC Voltage Dives." *IEEE Spectrum*. vol. 29 No. 5 pp. 22-25, 1992.

29. Harrington, J. "The 3V Transition." *ASIC & EDA*. November pp. 36-40, 1993.

30. Baccarani, G., et. al. "Generalized Scaling Theory and its Application to a 1/4 Micrometer MOSFET Design." *IEEE Trans. on Electron Devices*. vol. ED-31 No. 4 pp. 452-462, 1984.

31. Lahri, R., S. Joshi, B. Bastani. "Process Reliability." BiCMOS Technology and Applications. Alvarez ed. 1989 Kluwer Academic Publishers.

32. Tang, D., et. al. "Junction Degradation in Bipolar Transistors and the Reliability Imposed Constraints to Scaling and Design." *IEEE Trans. on Electron Devices.* vol. ED-35 No. 12 pp. 2101-2107, 1988.

33. Arnborg, T. "Performance Predictions of Scaled BiCMOS Gates Using Physical Simulation." *IEEE Jnl. Solid State Circuits.* vol. 27 No. 5 pp. 754-760, 1992.

34. Hiraki, M., et. al. "A 1.5-V Full-Swing BiCMOS Logic Circuit." *IEEE Jnl. Solid State Circuits.* vol. 27 No. 11 pp. 1568-1573, 1992.

35. Chik, R., C. Salama. "1.5V Bootstrapped BiCMOS Logic Gate." *Electronics Letters.* vol. 29 No. 3 pp. 307-308, 1993.

36. Shin, H. "Full-Swing BiCMOS Logic Circuits with Complementary Emitter-Follower Driver Configuration." *IEEE Jnl. Solid State Circuits.* vol. 26 No. 4 pp. 578-584, 1991.

37. Embabi, S., A. Bellaouar, M. Elmasry, R. Hadaway. "New Full-Voltage Swing Buffers." *IEEE Jnl. Solid State Circuits.* vol. 26 No. 2 pp. 150-153, 1991.

38. Shin, H. "Performance Comparison of Driver Configurations and Full-Swing Techniques for BiCMOS Logic Circuits." *IEEE Jnl. Solid State Circuits.* vol. 25 No. 3 pp. 863-865, 1990.

39. Chen, C. "Level-Shifted 0.5-um BiCMOS Circuits." *IEEE Jnl. Solid State Circuit.* vol. 25 No. 5 pp. 1214-1216, 1990.

40. Boudon, G., F. Wallart, E. Maillart. "Internal ECL-BiCMOS Translator Circuits in Half Micron Technology." *IEEE International Conference on Computer Design: VLSI in Computers and Processors.* pp. 314-317, 1989.

41. Shoji, M. "CMOS Digital Circuit Technology." 1988 Prentice-Hall, Inc.

42. Canali, C., G. Majni, R. Minder, G. Ottaviani. "Electron and Hole Drift Velocity Measurements in Silicon and Their Empirical Relation To Electric Field and Temperature." *IEEE Trans. on Electron Devices.* vol. ED-22 November pp. 1045-1047, 1975.

43. Sun, S., J. Plummer. "Electron Mobility in Inversion and Accumulation Layers on Thermally Oxidized Silicon Surfaces." *IEEE Trans. on Electron Devices.* vol. ED-27 pp. 1497-1508, 1980.

44. Pao, H., C. Shah. "Effects of Diffusion Current on Characteristics of Metal-Oxide (insulator) Semiconductor Transistors." *Solid State Electronics.* vol. 9 pp. 927-937, 1966.

45. Hayashi, Y. "Static Characteristics of Extremely Thin Gate Oxide M. O. S Transistors." *Electronics Letters.* vol. 11 No. 25/26 pp. 618-620, 1975.

46. Hu, C., et. al. "Hot-Electron-Induced MOSFET Degradation-Model, Moniter and Improvement." *IEEE Transactions on Electron Devices.* vol. ED-32 No. 2 pp. 375-384, 1985.

47. Boyle, B., K. Mistry. "Anomalous Hot-Carrier Behavior for LDD p-Channel Transistors." *IEEE Electron Device Letters.* vol. 14 No. 11 pp. 536-538, 1993.

48. Tang, D., et. al. "Design Considerations of High-Performance Narrow-Emitter Bipolar Transistors." *IEEE Electron Device Letters.* vol. EDL-8 No. 4 pp. 174-175, 1987.

49. Verret, D., J. Brighton. "Two-Dimensional Effects in the Bipolar Polysilicon Self-Aligned Transistor." *IEEE Trans. Electron Devices.* vol. ED-34 November pp. 2297-2303, 1987.

50. Ashburn, P. "Design and Realization of Bipolar Transistors." 1988 John Wiley & Sons, Ltd.

51. Kirk, C. "A Theory of Transistor Cut-off Frequency Falloff at High Current Densities." *IRE Trans. Electron. Devices.* vol. ED9 pp. 164, 1962.

52. Graul, J., A. Glasl, H. Murrmann. "High-Performance transistors with arsenic-implanted polysil emitters." *IEEE Jnl. Solid State Circuits.* vol. SC11 pp. 491, 1976.

53. Teplik, J. "Device Design." BiCMOS Technology and Applications. Alvarez ed. 1989 Kluwer Academic Publishers.

54. Estreich, D., R. Dutton. "Modeling Latch-Up In CMOS Integrated Circuits and Systems." *IEEE Trans. on Computer-Aided Design of Integrated Circuits.* vol. CAD-1 No. 4 pp. 157-163, 1982.

55. Troutman, R. "Recent Developments in CMOS Latchup." *IEDM Tech. Digest.* pp. 296-299, 1984.

56. Payne, R., W. Grant, W. Bertran. "Elimination of Latchup in Bulk CMOS." *IEDM Tech. Digest.* pp. 248-251, 1980.

57. Troutman, R. "Recent Developments and Future Trends in Latch-Up Prevention in Scaled CMOS." *IEEE Trans. on Electron Devices.* vol. ED-30 pp. 1564, 1983.

58. Ritts, R., et. al. "Merged BiCMOS Logic to Extend the CMOS/BiCMOS Performance Crossover Below 2.5-V Supply." *IEEE Jnl. Solid State Circuits.* vol. 26 No. 11 pp. 1606-1613, 1991.

59. Kuo, J., G. Rossel, R. Dutton. "Two-Dimensional Analysis of a Merged BiPMOS Device." *IEEE Trans. on Computer-Aided Design.* vol. 8 No. 8 pp. 929-932, 1989.

60. Momose, H., T. Maeda, K. Inoue, I. Kamohara, T. Kobayashi, Y. Urakawa, K. Maeguchi. "Characterization of Speed and Stability of BiNMOS Gates With a Bipolar and PMOSFET Merged Structure." *IEDM Tech. Digest.* pp. 231-234, 1990.

61. Menozzi, R., et. al. "Layout Dependence of CMOS Latchup." *IEEE Trans. on Electron Devices.* vol. 35 No. 11 pp. 1892-1901, 1988.

62. Bafleur, M., J. Buxo, M. Vidal, P. Givelin, V. Macary, G. Sarrabayrouse. "Application of a Floating Well Concept to a Latch-up-Free, Low-Cost, Smart Power High-Side Switch Technology." *IEEE Trans. on Electron Devices.* vol. 40 No. 7 pp. 1340-1342, 1993.

63. Gwennap, L. "Cyrix Describes Pentium Competitor." *Microprocessor Report*, vol. 7, pp. 1

64. "M68040 User's Manual.", Motorola Inc., doc. no. 68040UM/AD, 1992.

65. "TMS320Cx User's Guide.", Texas Instruments, doc. no. 2558539-9721 rev. E. 1991.

66. Gnanasekaran, R. "A Fast Serial-Parallel Binary Multiplier." *IEEE Trans. on Computers*. vol. C-34 No. 8 pp. 741-744, 1985.

67. El-Desouky, A., M. Salem, A. Abd El-Gwad, L. Labib. "A New Technique For Binary Multiplier." *Int. Jnl. of Mini and Microcomputers*. vol. 14 No. 2 pp. 68-76, 1992.

68. Pezaris, S. D. "A 40ns 17-bit by 17-bit Array Multiplier." *IEEE Trans. on Computers*. vol. C-20 February pp. 442-447, 1971.

69. Montuschi, P., L. Ciminiera. "nXn Carry-Save Multipliers Without Final Addtion." *11th Symposium on Computer Arithmetic*. pp. 54-61, 1993.

70. Nakamura, S. "Algorithms for Iterative Array Multiplication." *IEEE Trans. on Computers*. vol. C-35 No. 8 pp. 713-719, 1986.

71. Nakamura, S., K. Chu. "A Single Chip Parallel Multiplier by MOS Technology." *IEEE Trans. on Computers*. vol. 37 No. 3 pp. 274-282, 1988.

72. Stenzel, W., W. Kubitz, G. Garcia. "A Compact High-Speed Parallel Multiplication Scheme." *IEEE Trans. on Computers*. vol. C-26 pp. 948-957, 1977.

73. Goto, G., T. Sato, M. Nakajima, T. Sukemura. "A 54X54-b Regularly Structured Tree Multiplier." *IEEE Jnl. Solid State Circuits*. vol. 27 No. 9 pp. 1229-1235, 1992.

74. Nagamatsu, M., S. Tanaka, J. Mori, K. Hirano, T. Noguchi, K. Hatanaka. "A 15-ns 32X32-b CMOS Multiplier with an Improved Parallel Structure." *IEEE Jnl. Solid State Circuits*. vol. 25 No. 2 pp. 494-497, 1990.

75. Santoro, M. "SPIM: A Pipelined 64X64-bit Iterative Multiplier." *IEEE Jnl. Solid State Circuits*. vol. 24 No. 2 pp. 487-493, 1989.

76. Mehta, M., E. Swartzlander Jr. "High Speed Multiplier Design Using Multi-Input Counter and Compressor Circuits." *10th Symp. on Computer Arithmetic*, pp. 43-50, 1991.

77. Montoye, R., E. Hokenek, S. Runyon. "Design of the IBM RISC System/6000 Floating Point Execution Unit." *IBM Journal of Research*, vol. 34 pp. 59-70, 1990.

78. Booth, A. D. "A Signed Binary Multiplication Technique." *Quart. Jnl. Mech. and Applied Math*, vol. IV Pt. 2 pp. 236-240, 1951.

79. Koren, I. "Computer Arithmetic Algorithms." 1993 Prentice-Hall, Inc.

80. MacSorley, O. "High-Speed Arithmetic in Binary Computers." *Proc. of the IRE*, January pp. 67-91, 1961.

81. Sam, H., A. Gupta. "A Generalized Multibit Recoding of Two's Complement Binary Numbers and Its Proof with Application in Multiplier Implementations." *IEEE Trans. on Computers*, vol. 39 No. 8 pp. 1006-1015, 1990.

82. Mori, J., et. al. "A 10-ns 54X54-b Parallel Structured Full Array Multiplier with .5um CMOS Technology." *IEEE Jnl. Solid State Circuits*, vol. 26 No. 4 pp. 600-605, 1991.

83. Hokenek, E., R. Montoye, P. Cook. "Second-Generation RISC Floating Point with Multiply-Add Fused." *IEEE Jnl. Solid State Circuits*, vol. 25 No. 5 pp. 1207-1213, 1990.

84. Wallace, C. S. "A Suggestion for a Fast Multiplier." *IEEE Trans. Electronic Computers*, vol. EC-13 February pp. 14-17, 1964.

85. Dadda, L. "Some Schemes For Parallel Multipliers." *Alta Frequenza*, vol. 34 pp. 349-356, 1965.

86. Swartzlander, E. Jr. "Parallel Counters." *IEEE Trans. on Computers*, vol. C-22 No. 11 pp. 1021-1024, 1973.

87. Wang, Z., G. A. Jullien, W. C. Miller. "New Design Techniques for Column Compression Multipliers." *IEEE Trans. on Computers*. In Press 1994.

88. Villeger, D., V. Oklobdzija. "Evaluation of Booth Encoding Techniques For Parallel Multiplier Implementation." *Electronics Letters*. vol. 29 No. 23 pp. 2016-2017, 1993.

89. El-Gamal, A., D. Gluss, P. Ang, J. Greene, J. Reyneri. "A CMOS 32b Wallace Tree Multiplier-Accumulator." *IEEE Solid State Circuits Conference*. vol. THPM 15.5 pp. 194-195, 1986.

90. Crawley, D., G. Amaratunga. "8X8 Bit Pipelined Dadda Multiplier in CMOS." *IEE Proceedings*. vol. 135 Pt. G pp. 231-240, 1988.

91. Cappello, P., K. Steiglitz. "A VLSI Layout for a Pipelined Dadda Multiplier." *ACM Trans. on Computer Systems*. vol. 1 No. 2 pp. 157-174, 1983.

92. Wang, Z., G. A. Jullien, W. C. Miller. "An Architecture For Parallel Multipliers." *25th Asilomar Conf. on Signal, System, and Computers*. pp. 403-407, 1991.

93. Mesfin, B. "Implementation of High Performance Floating Point Unit Multiplier." M. A. Sc., University of Windsor, Windsor, Canada, 1992.

94. Brent, R., H. Kung. "A Regular Layout For Parallel Adders." *IEEE Trans. on Computers*. vol. C-31 No. 3 pp. 260-264, 1982.

95. Chan, H. "BiCMOS Implementation on DSP Arithmetic Blocks." M. A. Sc., University of Windsor, Windsor, Canada, 1993.

96. Summerfield, S. "Comparison of Two-Phase Latch Configurations for Pipelined Processors in MOS VLSI: Case Study: A CMOS Systolic Multiplier." *IEE Proceedings*. vol. 137 Pt. G No. 4 pp. 261-265, 1990.

97. Krambeck, R., C. Lee, H. Law. "High-Speed Compact Circuits with CMOS." *IEEE Jnl. Solid State Circuits*. vol. SC-17 No. 3 pp. 614-619, 1982.

98. Myers, D., P. Ivey. "A Design Style for VLSI CMOS." *IEEE Jnl. Solid State Circuits*. vol. SC-20 No. 3 pp. 741-745, 1985.

99. Wu, C., K. Cheng, J. Wang. "Analysis and Design of a New Race-Free Four-Phase CMOS Logic." *IEEE Jnl. Solid State Circuits*. vol. 28 No. 1 pp. 18-25, 1993.

100. Gonclaves, N., H. De Man. "NORA: A Racefree Dynamic CMOS Technique for Pipelined Logic Structures." *IEEE Jnl. Solid State Circuits*. vol. SC-18 No. 3 pp. 261-266, 1983.

101. Yuan, J., I. Karlsson, C. Svensson. "A True Single-Phase-Clock Dynamic CMOS Circuit technique." *IEEE Jnl. Solid State Circuits*. vol. 22 No. 5 pp. 899-901, 1987.

102. Lu, F., H. Samueli, J. Yuan, C. Svensson. "A 700-MHz 24-b Pipelined Accumulator in 1.2um CMOS for Application as a Numerically Controlled Oscillator." *IEEE Jnl. Solid State Circuits*. vol. 28 No. 8 pp. 878-885, 1993.

103. Yuan, J., C. Svensson. "High-Speed CMOS Circuit technique." *IEEE Jnl. Solid State Circuits*. vol. 24 No. 1 pp. 62-70, 1989.

104. Afghahi, M., C. Svensson. "A Unified Single-Phase Clocking Scheme for VLSI Systems." *IEEE Jnl. Solid State Circuits*. vol. 25 No. 1 pp. 225-233, 1990.

105. Larsson, P. "Robustness of Digital CMOS Techniques With Special Emphasis on the True Single Phase Clocking Strategy." PhD, Linkoping University, Linkoping Sweden

106. Jullien, G. A., W. C. Miller, L. Del Pup, R. Grondin, D. Zhang. "Synthesis of Dynamic Computational Blocks for Bit-Level Systolic arrays." *IEEE Jnl. Solid State Circuits*. In Press 1994.

107. Jullien, G. A., W. C. Miller, R. Grondin, Z. Wang, D. Zhang, L. Del Pup, S. Bizzan. "WoodChuck: A Low-Level Synthesizer for Dynamic Pipelined DSP Arithmetic Logic Blocks." *IEEE International Symposium on Circuits and Systems*. vol. 1 pp. 176-179, 1992.

108. Hotta, T., T. Bandoh, A. Hotta, T. Nakano, S. Iwamoto, S. Adachi. "A 70-MHz 32-b Microprocessor with 1.0um BiCMOS Macrocell Library." *IEEE Jnl. Solid State Circuits*. vol. 25 No. 3 pp. 770-777, 1990.

109. Hotta, T., I. Masuda, H. Maejima, M. Ueno, M. Iwamura, K. Kurita, A. Hotta. "CMOS/Bipolar Circuits For 60MHz Digital Processing." *IEEE Jnl. Solid State Circuits*. vol. SC-21 No. 5 pp. 808-813, 1986.

110. Hotta, T., K. Kurita, H. Maejima, M. Iwamura, S. Tanaka, T. Bandoh, T. Yamauchi, A. Hotta. "1.3-um CMOS/Bipolar Standard Cell Library for VLSI Computers." *IEEE Jnl. Solid State Circuits*. vol. 23 No. 2 pp. 500-505, 1988.

111. Kuo, J., H. Liao, H. Chen. "BiCMOS Dynamic Manchester Carry Look Ahead Circuit For High Speed Arithmetic Unit VLSI." *Electronics Letters*. vol. 28 No. 5 pp. 477-478, 1992.

112. Kuo, J., H. Liao, H. Chen. "A BiCMOS Dynamic Carry Lookahead Adder Circuit For VLSI Implementation of High-Speed Arithmetic Unit." *IEEE Jnl. Solid State Circuits*. vol. 28 No. 3 pp. 375-378, 1993.

113. Kuo, J., K. Su, J. Lou. "1.5V BiCMOS Dynamic Multiplier Using Wallace Tree Reduction Architecture." *Electronics Letters*. vol. 29 No. 24 pp. 2097-2098, 1993.

114. Chen, S., C. Chiang, K. Su, J. Kuo. "BiCMOS Dynamic Minimum Circuit Using A Parallel Comparison Algorithm For Fuzzy Controllers." *Electronics Letters*. vol. 29 No. 6 pp. 551-553, 1993.

115. Kuo, J., J. Wang, Y. Chen. "Low-Voltage BiCMOS Dynamic Minimum Circuit Using a Parallel Comparison Algorithm For Fuzzy Controllers." *Electronics Letters*. vol. 30 No. 1 pp. 31-32, 1994.

116. Zhou, P., Personal Communication, August, 1993.

117. Siddiq, S., Personal Communication, August, 1993.

118. Sedra, A., K. Smith. "Microelectronic Circuits." 1987 Holt, Rinehart and Winston.

119. "HSPICE User's Manual, Elements and Models.", Meta Software, Inc., Volume 2, HSPICE Version h92, 1992.

Appendix A

EDGE™ BiCMOS ENVIRONMENT HINTS

A.1 Introduction

The following is a collection hints for carrying out design work in the BiCMOS Edge™ environment. It is not an exhaustive list, and the information contained is subject to change as the environment is developed. An environment description is contained in the Micronet document "Micronet/CMC BiCMOS Design Environment" available from CMC, however, there are some details included here which are not treated in that document.

A.2 List of Hints

- Global power and ground nets are not used in BATMOS due to the possibility of multiple supplies or grounds in mixed signal designs. Use "vdd" and "vss" for power and ground, respectively.

- Custom bipolar transistor layout should be avoided. The models provided for bipolar transistors presently are not scalable. It is recommended that the bipolar layouts provided in the utility cell directory be used to obtain accurate simulations.

- Design scale microns = real microns in BATMOS.

- Layout and circuit extraction in BiCMOS requires a post processing skill program to be run. This program performs many functions, and it is called "fixExtract()". It calls another program named "finishExtract()". Among other things, these routines fix floating nodes caused by jumper pins, and attach the correct model property to the extracted transistors so the appropriate model call is present in the netlist. At the time of writing, the routine can be automatically run from the "WINDSOR PDV" menu by using "FLAT (with pcap)" in the "Batmos" section of extraction options. Presently, this menu entry calls finishextract(), then fixExtract(), and finally calls finishExtract() again. Extraction should only require a single call to fixExtract(), but at the time of writing, the above procedure is necessary. It may be necessary to quit and restart Cadence if parasitics are not being extracted when the extract parasitic command is entered. Schematic extraction is performed from the "BATMOS-Lib" menu. This extraction uses the same skill routines as above.

- HSPICE™ simulations can be run with the following considerations. A statement in the "control" file of the simulation run directory must point to a file containing device models. An example statement would be:

.LIB '/path_to_my_model_file/hspice.models'

The present environment for BiCMOS does not use "gnd!" to indicate ground in the cells. This is due to the possibility of multiple grounds in mixed mode designs. The equivalent is "Vss" or "vss". Other nodes may still be netlisted as ground by the Cadence netlister. For example, if "vss" has been used in a design, and the netlist contains nodes called "0!" and "gnd!", simply connect a dc voltage source of zero value between these nodes, and the "Vss" node you have used. This is done in the control file. For example, the following lines would be used in the above case:

vfix1 [#vss] [#gnd!] dc 0

vfix2 [#vss] [#0!] dc 0

- Physical representation checking and verification procedures, such as DRC and PDextract, on large designs can take a huge amount of computer resources. A design rule check of one of the BiCMOS multiplier macrocells took as much as 100 Megabytes of disk space to complete due to the large temporary files that Edge™ writes to disk during the analysis. As well, the analysis took several hours to complete. Use the environment variable "DRCTEMPDIR" **before** starting Edge™ to set a disk area with sufficient space. An example command at the UNIX prompt would be:

  setenv DRCTEMPDIR "/scratch1"

  where "/scratch1" is a large disk partition. Multiple disk areas can be designated. Check the Edge™ manuals.

- Verilog™ library directories must be included in the statement defining the "simVerilogOptions" variable in the .simrc or .simlocal files. The "-y" option must be used to force Verilog™ to search the appropriate directories in a certain order. Directories containing the most accurate models should be listed first, with the less accurate ones being listed in order of descending precision. At the time of macrocell design, only certain tcells had been characterized to properly provide accurate delay information. For this reason, results from the Veritool simulations should be looked upon as estimates, and not accurate predictions. As more cells are characterized, all that is necessary is to replace the new Verilog™ model file for the specific cell into the

appropriate library directory, and re-run the simulation with the netlist and associated STL or simulator input files which are supplied.

- Due to a process design rule change, the tcell library has been altered by including a contact strip along the top of each cell. This strip should be connected to vss to avoid possible substrate currents and latchup. In an automatically placed and routed design, there must be at least one connection per standard cell row.

- Net names are referred to in the Place and Route tool, and in representations generated by that tool, preceded by a "l" character. For example, if the net is named "my_net" in the schematic, during and after placement and routing, the physical net will be called "lmy_net". Consequently, power and ground are referred to as "lvdd", and "lvss" respectively.

- If routing channels are exploded so that a level of hierarchy is removed, many extraneous pins appear in the top level layout representation, and the presence of these pins will cause PDextract to fail. They must be removed prior to extraction. Appendix E contains a program written in SKILL™ which will remove these unwanted pins. After the routing channels have been exploded in the layout rep, simply load and run the skill routine in the active window. The program will remove the pins, and the layout can then be saved and extracted.

- Large macrocells require substantial post place and route processing. Pins must be named and placed on top of all ports at the top level of hierarchy in order for the names to carried to the extracted rep. Connections must be made from vss to the contact on top of the tcells at least once for each row of cells. If channels are to be exploded , all resulting extraneous pins must be removed before extraction. These procedures take a substantial amount of time, and this should be budgeted for.

- Vectored nets in the symbol and/or lvs reps cause problems when using the PDcompare tool. Use of single nets, e.g., use a<0> and a<1> explicitly instead of using a<1:0>, sometimes helps.

- Jumper pins in tcells sometimes cause problems when using PDcompare.

- Place and Route with i/o pads is complicated. Significant questions can be directed to Andrew Scott at CMC.

Appendix B

MACROCELL LAYOUTS

B.1 Introduction

This appendix contains layout images of the arithmetic macrocells which were designed. The name in brackets is the name of the top level Cadence Edge™ block. Appendix C contains implementation specific details of these cells.

B.2 Macrocells Implemented in Pure CMOS

Figure B.1: Dadda Style Multiplier (dadda_top)

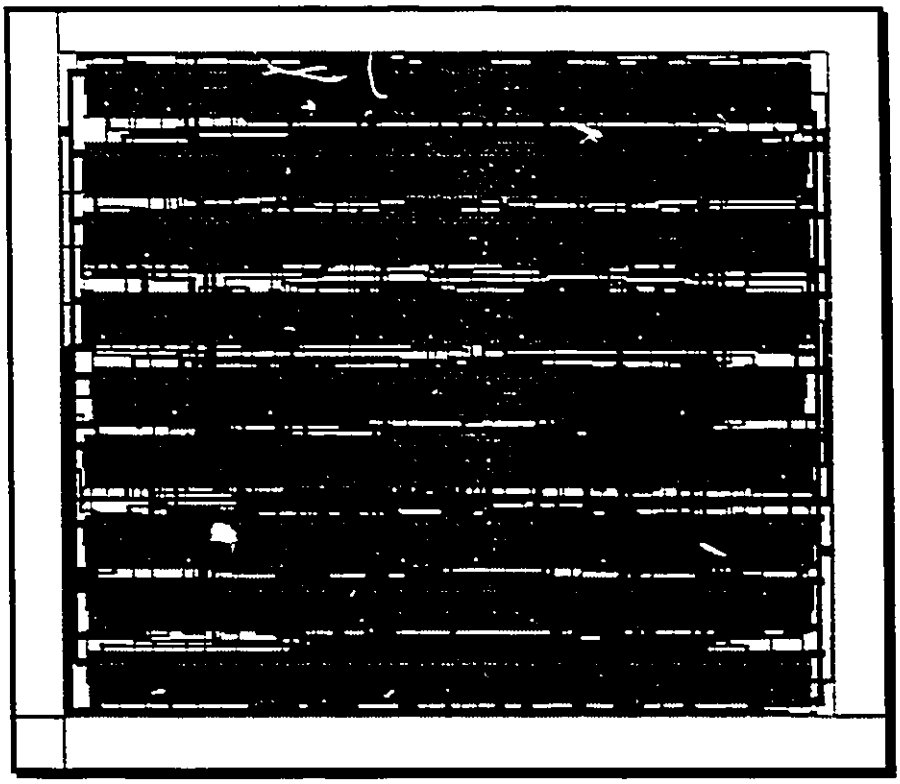


Figure B.2: Two Bit Full Adder Multiplier Using Full Adder Cells (tbfa_fa_top)

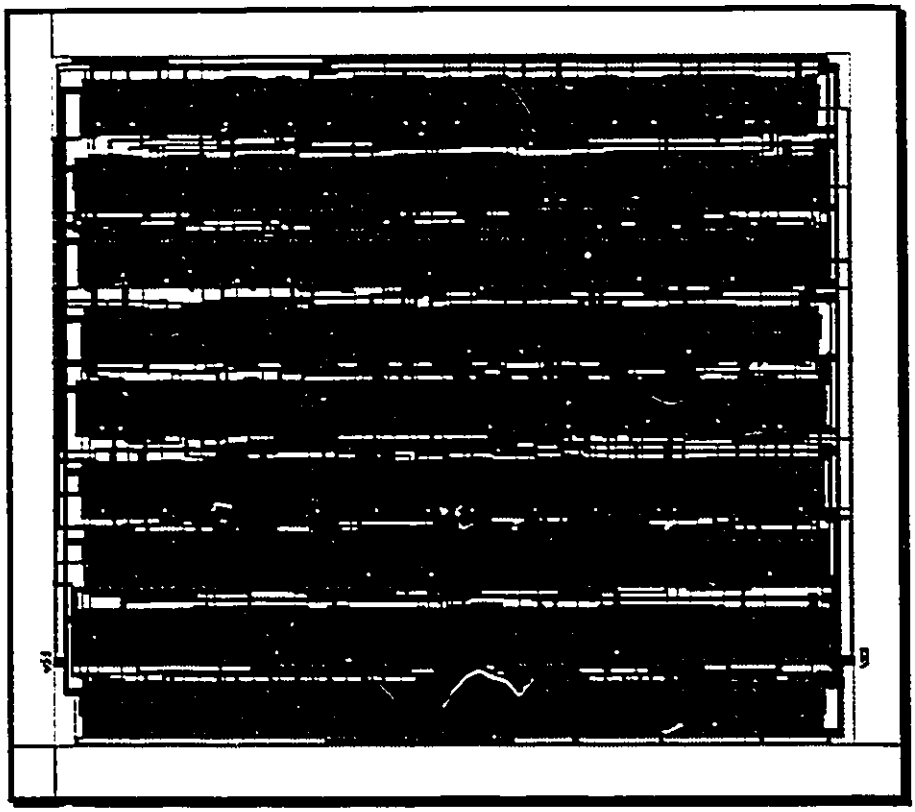
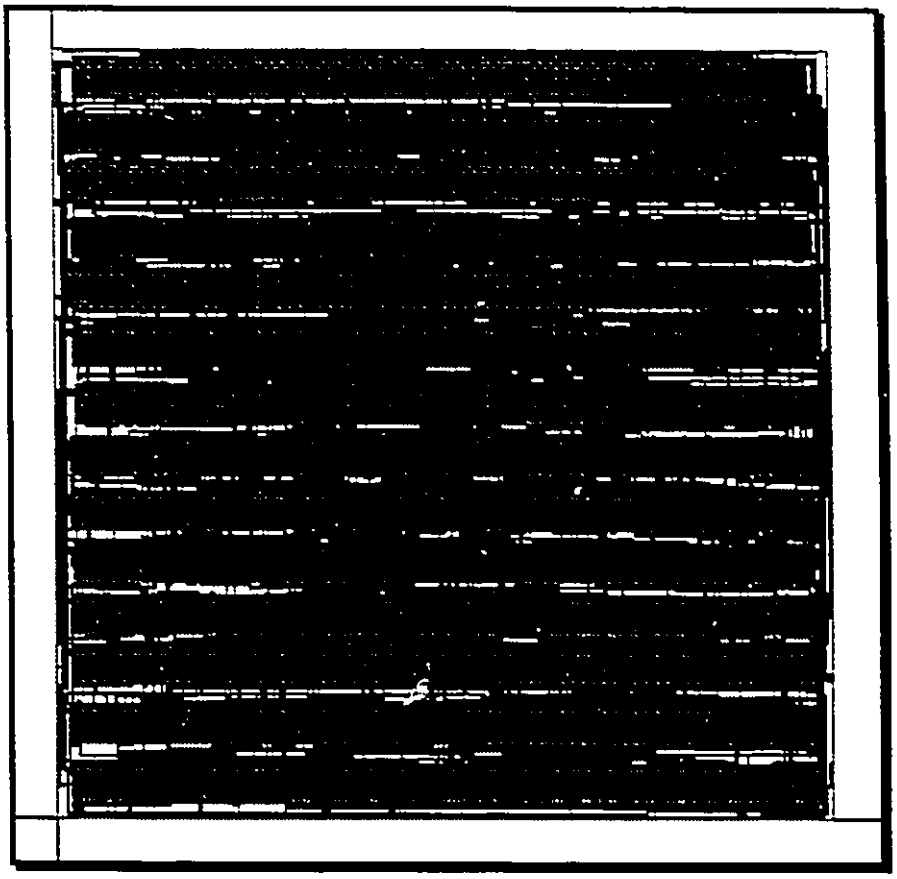


Figure B.3: Two Bit Full Adder Multiplier Using Simple Gates (tbfa_g_top)



B.3 Macrocells Implemented in BiCMOS

Figure B.4: Dadda Style Multiplier (dadda_top_bic)

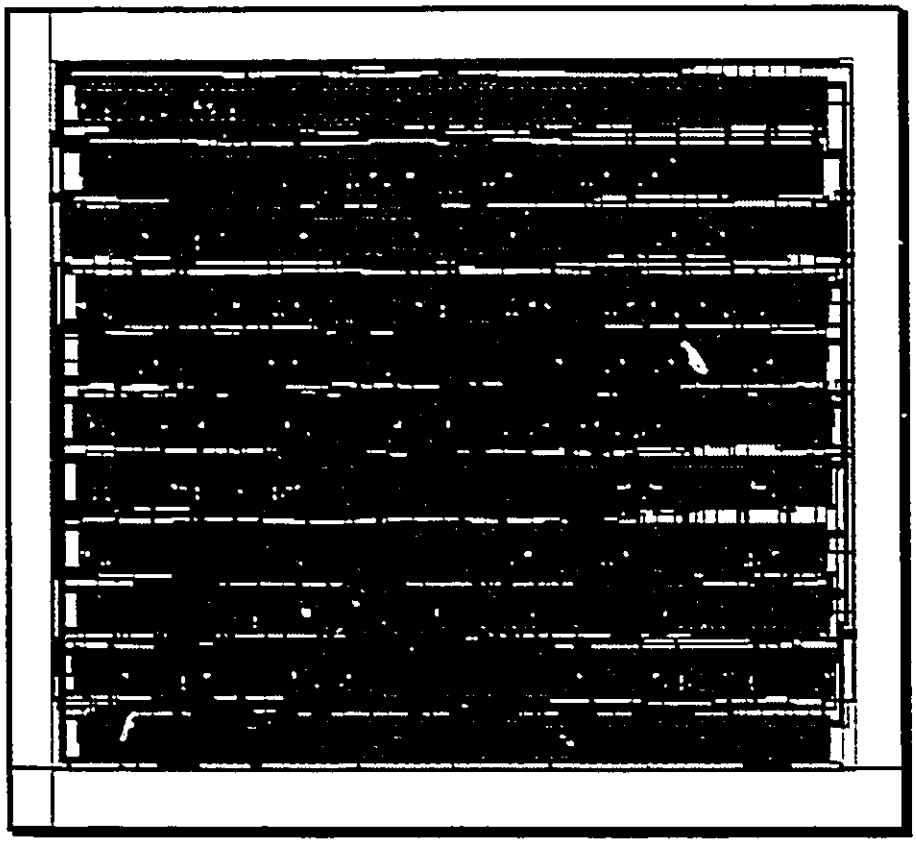
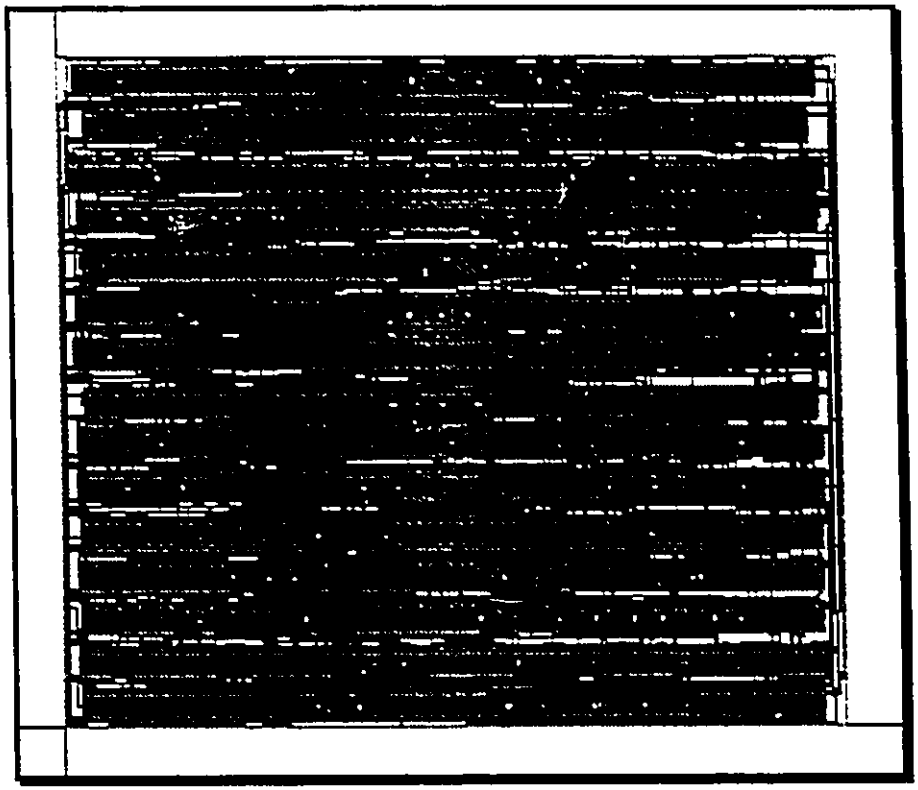
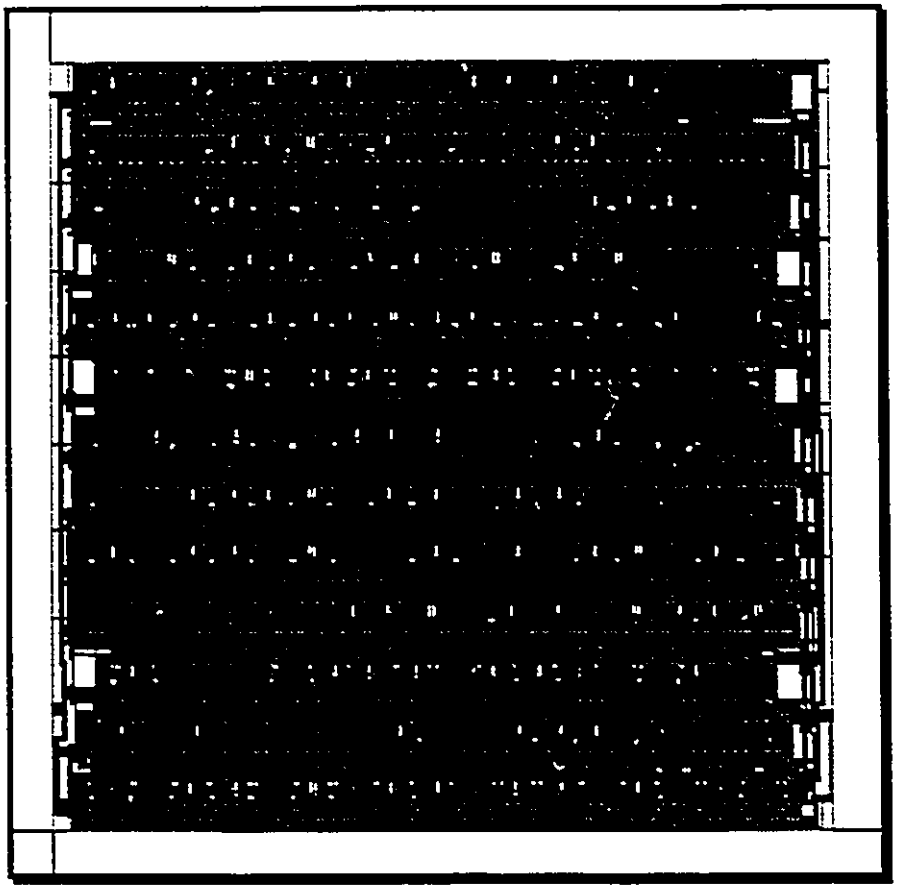


Figure B.5: Two Bit Full Adder Multiplier Using Simple Gates (tbfa_g_bic_top)



B.4 Macrocell Implemented in ECL

Figure B.6: Fast Adder (adder_top)



Appendix C

MACROCELL SPECIFIC DETAILS

C.1 Introduction

The designed macrocells take the form of various Edge™ representations (reps). The layout rep is hierarchical, that is, it is composed of flat channels containing routing metal traces surrounding instances of standard cells. Thus, in order to view the layout, the Cadence block path must include a path to the various standard cell libraries associated with the BiCMOS environment release. The extracted reps were produced by performing a macrocell extraction on the layout. In addition to the above representations, abstract, lvs, and Verilog™ reps were also created, along with associated files. Directories containing Verilog™ runs which were used to verify the architecture are available, along with the associated STL source files. As well, Verifault™ and Veritime™ analysis results are available.

Clocking of the macrocells is straightforward. All of them, except the ECL fast adder, use single phase clocking. Data is clocked into the multiplier macrocells on the negative edge of the clock, and results are clocked out of the multipliers on the positive edge of the clock pulse. The storage elements in the multipliers are scannable D flip flops. The ECL fast adder uses a master slave ECL d latch arrangement, with the data being clocked into the macrocell on the rising edge of the clock pulse, and results clocked out on the following rising edge. Note that for this macrocell, both true and complement clock signals are necessary.

Details of each of the cells are summarized in the tables below.

C.2 Macrocells Implemented in Pure CMOS

Table C.1: Dadda Style Multiplier

category	info
top level block name	dadda_top
standard cell types	CMOS
size (approx. $L \times W$)	1030 μm ×1030 μm
number of standard cells	201
inputs	a<7:0>, b<7:0>
outputs	pr<15:0>
clock(s)	clk
scan input	sm

Table C.2: Two Bit Full Adder Multiplier Using Simple Gates

category	info
top level block name	tbfa_g_top
standard cell types	CMOS
size (approx. $L \times W$)	1391 μm ×1378 μm
number of standard cells	684
inputs	a<7:0>, b<7:0>
outputs	pr<15:0>
clock(s)	clk
scan input	sm

Table C.3: Two Bit Full Adder Multiplier Using Full Adder Cells

category	info
top level block name	tbfa_fa_top
standard cell types	CMOS
size (approx. $L \times W$)	1016 μm ×875 μm
number of standard cells	180
inputs	a<7:0>, b<7:0>
outputs	pr<15:0>
clock(s)	clk
scan input	sm

C.3 Macrocells Implemented in BiCMOS

Table C.4: Two Bit Full Adder Multiplier Using Simple Gates

category	info
top level block name	tbfa_g_bic_top
standard cell types	CMOS/BiCMOS
size (approx. $L \times W$)	1705 μm $\times$1469 μm
number of standard cells	624
inputs	a<7:0>, b<7:0>
outputs	pr<15:0>
clock(s)	clk
scan input	sm

Table C.5: Dadda Style Multiplier

category	info
top level block name	dadda_top_bic
standard cell types	CMOS/BiCMOS
size (approx. $L \times W$)	1182 μm $\times$1043 μm
number of standard cells	265
inputs	a<7:0>, b<7:0>
outputs	pr<15:0>
clock(s)	clk
scan input	sm

C.4 Macrocell Implemented in ECL

Table C.6: Fast Adder

category	info
top level block name	adder_top
standard cell types	ECL
size (approx. $L \times W$)	2581 μm $\times$2545 μm
number of standard cells	231
inputs	a<7:0>, b<7:0>
outputs	S<7:0>, c_out
clock(s)	clk, clk_bar
scan input	none

Appendix D

Email

>From arlene@terminus.ic.cmc.ca Tue Aug 31 10:06:22 1993
Return-Path: <arlene@terminus.ic.cmc.ca>
Received: from terminus (terminus.IC.CMC.CA) by galadrial.engn.uwindsor.ca (4.1 SMI-4.1)
id AA10072; Tue, 31 Aug 93 10:06:17 EDT
Received: from stargazer.ic.cmc.ca ([130.15.52.13]) by terminus (4.1 SMI-4.0)
id AA00855; Tue, 31 Aug 93 09:54:13 EDT
Date: Tue, 31 Aug 93 09:54:13 EDT
From: arlene@terminus.ic.cmc.ca (Arlene Seale)
Message-Id: <9308311354.AA00855@terminus>
To: jullien@Engn.Uwindsor.Ca
Subject: BiCMOS Contract
Cc: arlene@terminus.ic.cmc.ca
Status: R

Dear Graham,

We have completed our review of the BiCMOS contract deliverables and are pleased with the results. The availability of the bulk of this material was publicized in late June and there has been a high level of interest in it from the university community. The remainder of the deliverables are expected to be released in the next couple of weeks. Because of the positive response to the contract material, we are anticipating strong interest in the first CMC-coordinated fabrication run on September 29.

Tony Marsh, our President, has written to Micronet to formally express our appreciation and has asked that Micronet convey our gratitude for your very significant contributions to the contract. I would also like to express my personal thanks to you, and to those at your university who contributed to the contract, and look forward to working with you in the future.

Best Regards,

Andrew

P. Andrew Scott
CAE Specialist
Canadian Microelectronics Corporation
Queen's University
210A Carruthers Hall
Kingston, Ontario
Canada K7L 3N6

Phone: (613) 545-2914
Fax: (613) 548-8104
e-mail: scott@cmc.ca

Appendix E

PROGRAMS

E.1 Introduction

Several utility programs were written, mainly in SKILL™ and STL, to accomplish various aspects of the thesis work described. They are grouped in the following sections according to functionality.

E.2 Removal of Extraneous pins

If routing channels are exploded so that a level of hierarchy is removed, many extraneous pins appear in the top level layout representation, and the presence of these pins will cause PDextract to fail. They must be removed prior to extraction. After the routing channels have been exploded in the layout rep, simply load and run the skill routine given below in the active window. The program will remove the pins, and the layout can then be saved and extracted.

```
;*************************************************************************
:Procedure to remove extraneous pins from routing channels
;
:University of Windsor
;
:April 1993                J. C. Czilli
;
;*************************************************************************
procedure( fix_chan()
   prog( ( x workingrep term )

        ; try to open the rep in the window:
        ;
        workingrep=getWindowRep()
        if( workingrep == nil                then
                printf( "Unable to get rep in Window\n" )
                return(nil)
        )
                        foreach(x workingrep->shapes
                                if( ( ( x->layer == 20 ) || ( x->layer == 18 ) )
                                        && (x->shape == "rectangle")
                                        && (x->purpose == "pin") &&
                                        (strncmp("sda_UnconnectedPins"
                                                        x->net->name 19) == 0)
                                        deleteObject(x))
                        )
                        foreach(term workingrep->terminals
                                if(term->io=="jumper" then
                                        deleteObject(term->pins->inst)
                                )
                        )
)
)
```

E.3 Conversion of CMOS4S to BATMOS

The following procedure was used to convert CMOS4S designs into BATMOS. The CMOS4S design was converted to STREAM using "strmout" and the layers file given below:

```
#*********************************************************
#       layers Map File for converting cmos4s layouts into intermediate
#       form before conversion to BATMOS
#
#       James C. Czilli   1993


nwell     drawing       1  0
nwell     pin          1  0
nwell     interconnect   1  0
devwell   drawing       2  0
devwell   pin          2  0
devwell   interconnect   2  0
pguard    drawing       3  0
cappoly   drawing       5  0
cappoly   interconnect   5  0
poly      drawing       6  0
poly      pin          6  0
poly      interconnect   6  0
ndope     drawing       7  0
pdope     drawing       8  0
contact   drawing       9  0
contact   interconnect   9  0
metal     drawing      10  0
metal     pin          40  0
metal     interconnect  10  0
#metal     stub         10  0
via       drawing      11  0
via       interconnect  11  0
dmet      drawing      12  0
dmet      pin          43 0
dmet      interconnect  12  0
#dmet      stub         12  0
pyrox     drawing      13  0
DRC       drawing      35  0
```

Layers which are the same as in BATMOS are mapped accordingly while layers that are not are mapped to intermediate drawing layers. The STREAM file is then converted back into Cadence format, using "strmin" and the layers map file given below:

```
# strm2cdsMap (for converting cmos4s intermediate STREAM file to BATMOS)
#
# This file is used with strmin to map STREAM layer numbers to
# Edge layers.
#
#       James C. Czilli          Sept. 1993
#
y3              drawing         8       0

CON             drawing         9       0
MET1            drawing         10      0
MET1            pin             40      0
VIA1            drawing         11      0
MET2            drawing         12      0
MET2            pin             43      0
NWELL                   drawing         1       0
y0              drawing         2       0
GATE            drawing         6       0
PAD             drawing         13      0
DRC             drawing         35      0
```

Post processing is accomplished by running the next two skill routines on the layout. The first one, given below, derives the proper BATMOS layers from the extra drawing layers created during STREAM conversion.

```
;*****************************************************************
;
;Procedure to convert cmos4s designs to BATMOS
;
;       James C. Czilli          sept 1993
;
;
;
;*****************************************************

procedure( convert()
   prog( ( x workingrep term )

        ; try to open the rep in the window:
        ;
        workingrep=getWindowRep()
        if( workingrep == nil                   then
                printf( "Unable to get rep in Window\n" )
                return(nil)
        )

        geom=workingrep->shapes
        orig_contacts=setof( x_temp geom (x_temp->layer == 17) )

        setLayer("CON")
        foreach(big_contact orig_contacts
                o_ll=lowerLeft( big_contact->bBox )
                new_contact=relRectangle( o_ll .8 .8 )
                deleteObject( big_contact )
```

```
        )
        orig_vias=setof( x_temp geom (x_temp->layer == 19 ) )
        foreach(big_via orig_vias
                o_ll=lowerLeft( big_via->bBox )
                new_via=relRectangle( o_ll .8 .8 )
                deleteObject( big_via )
        )
; y3 is p+ in 4s
; y0 is devicewell in 4s
        devicewells=getLPP("y0")
        pplus=getLPP("y3")
        ndevarea=layerAndNot(workingrep "NDEV" devicewells->shapes pplus->shapes
)
        setLayer("NDEV")
        foreach( n_temp ndevarea
                rectangle(n_temp->bBox)
        )
        pdevarea=layerAnd(workingrep "PDEV" devicewells->shapes pplus->shapes )
        setLayer("PDEV")
        foreach( p_temp pdevarea
                rectangle(p_temp->bBox)
        )
        poly_tracks=getLPP("GATE")
        ntrans=getLPP("NDEV")
        met=getLPP("MET1")
        transistors=layerAnd(workingrep "y1" poly_tracks->shapes ntrans->shapes )
        metal_cross=layerAnd(workingrep "y1" poly_tracks->shapes met->shapes )
        forbidden=append(transistors metal_cross)
        via_regions=layerAndNot(workingrep "y2"  poly_tracks->shapes forbidden)
        setMaster("via_con layout current" )
        foreach(via_r_temp via_regions
                top=topEdge( via_r_temp ) - 2
                bottom=bottomEdge( via_r_temp ) + 2
                height=fix(top - bottom)
                num=height/2
                nesbitt=list(xCoord( lowerLeft(via_r_temp->bBox)) bottom)
                if( num > 0 then
                        corner=nesbitt
                        for( loop 1 num
                                new_vias=instance( corner )
                                corner=list(xCoord(corner) (yCoord(corner)+2))
                        )
                        setLayer("MET1")
                        rectangle(nesbitt list(rightEdge(via_r_temp) top))

                )
        )
        foreach(x_temp geom
                if(((x_temp->layer == 143) || (x_temp->layer == 140) ||
(x_temp->layer == 141) || (x_temp->layer == 142))
                        deleteObject( x_temp )
                )
        )
)
)
```

The second routine, given below, places substrate contacts throughout the layout in order to conform to latchup rules. This routine was written to be used with switching tree layout cores, so care should be exercised in using it for other designs. Before the program is run, a rectangle should be drawn on the "prboundary" drawing layer to indicate the layout area to be processed. The layout should be thoroughly checked afterward to make sure contacts have not been placed in unwanted areas.

```
;*************************************************************************
;Procedure to insert substrate contacts into ntran trees
;
;        James C. Czilli              Sept. 1993;
;
;*************************************************************************
procedure( subcon()
   prog( ( x workingrep term )

        ; try to open the rep in the window:
        ;
        workingrep=getWindowRep()
        if( workingrep == nil                   then
                printf( "Unable to get rep in Window\n" )
                return(nil)
        )

        metal2=getLPP("MET2")
        metal1=getLPP("MET1")
        devicewells=getLPP("NDEV")
        bound=getLPP("prboundary")
        temp_bound=head(bound->shapes)
        ll=lowerLeft(temp_bound->bBox)
        ur=upperRight(temp_bound->bBox)
        setLayer("y7")
        rectangle(list(list((xCoord(ll)+20-2.5-2.8) (yCoord(ll)+20-1.4-2.8) )
        list((xCoord(ur)-20+2.5+2.8) (yCoord(ur)-20+1.4+2.8))))
        bound=getLPP("y7")
        temp1_geom=layerOr(workingrep "y1" devicewells->shapes metal1->shapes )
        temp2_geom=layerOr(workingrep "y1" temp1_geom->shapes metal2->shapes )
        temp_y1=getLPP("y1")
        temp1_geom=layerXor(workingrep "y2" temp_y1->shapes bound->shapes)
        tiled=layerTile(workingrep "y3" temp1_geom)
        foreach(temp_reg tiled
                if( ((rightEdge(temp_reg)-leftEdge(temp_reg)) < 3.8)
                        || ( (topEdge(temp_reg)-bottomEdge(temp_reg)) < 4.2 )
                        deleteObject(temp_reg)
                )
        )
        temp_y3=getLPP("y3")
        tiled=temp_y3->shapes
        setLayer("bkgnd")
        subAll()
        setLayer("y2")
        foreach(temp_reg tiled
                setLayer("y2")
                subAll()
                top_edge=topEdge(temp_reg)
```

```
        bottom_edge=bottomEdge(temp_reg)
        add(list((leftEdge(temp_reg)+1) (bottom_edge+1)))
        regenerate=head(selectedSet())
        subAll()
        left_side=leftEdge(regenerate)
        right_side=rightEdge(regenerate)
        setLayer("y5")
        rect=rectangle( list(left_side bottom_edge) list(right_side top_edge))
        height=top_edge-bottom_edge
        num=fix(height/4.2)
        ll=list(left_side bottom_edge)
        y=bottom_edge
        for(loop 1 num
                y=y+4.2
                cut(list(ll list(right_side y)))
                subAll()
        )
)
subc_areas=getLPP("y5")
foreach(temp_shape subc_areas~>shapes
        if( (topEdge(temp_shape)-bottomEdge(temp_shape)) < 4.2 ||
                        (rightEdge(temp_shape)-leftEdge(temp_shape)) < 6.6
                deleteObject(temp_shape)
        )
)
temp_y5=getLPP("y5")
remove=getLPP("y1")
foreach(temp_shape remove~>shapes
                deleteObject(temp_shape)
)
remove=getLPP("y2")
foreach(temp_shape remove~>shapes
                deleteObject(temp_shape)
)
remove=getLPP("y3")
foreach(temp_shape remove~>shapes
                deleteObject(temp_shape)
)
remove=getLPP("y4")
foreach(temp_shape remove~>shapes
                deleteObject(temp_shape)
)
remove=getLPP("y7")
foreach(temp_shape remove~>shapes
                deleteObject(temp_shape)
)


setMaster("via_sub layout current" )
subc_areas=temp_y5~>shapes
setLayer("bkgnd")
subAll()
setLayer("y5")
while(subc_areas != nil
        subAll()
        current_area=head(subc_areas)
        x=leftEdge(current_area)+2.4
        y=bottomEdge(current_area)
```

```
                    setLayer("NDEV")
                    add(list(x (y-1)))
                    direction=selectedSet()
                    subAll()
                    setLayer("y5")
                    if( direction == nil then displacement=1.1
                    else displacement=1.4
                    )
                    y=y+displacement
                    new_con=instance(list(x y))
                    center=centerBox(new_con->bBox)
                    x=xCoord(center)
                    y=yCoord(center)
                    rule_leftedge=x-25
                    rule_topedge=y+25
                    rule_rightedge=x+28
                    rule_bottomedge=y-28
                    addArea(list(list(rule_leftedge rule_bottomedge ) list(rule_rightedge
rule_topedge)))
                    deleteShapes()
                    update=getLPP("y5")
                    subc_areas=update->shapes
          )


)
)
```

E.4 Sample STL file

Cadence Simulation and Test language (STL) was used extensively for simulation and verification. The "simdiff" program was used for automatic comparison of results.

The example program below controls a Verilog™ simulation of one of the multiplier macrocells:

```
;
; STL test program for multiplier macrocells
;
; To chage the number of patterns applied, change the limit in the
; "for" loop.
;
;          James C. Czilli              1993


stlinit
:stltrace
defpin vdd          in
defpin vss          in
defpin pr<15:0> out
defpin a<7:0>       in
defp ר b<7:0>       in
```

```
defpin clk          clk
defpin sm           in

defformat a b sm vdd vss pr
deftiming 1ns 10ns 100ns  ; define basic time units
defclock "11111....." clk
deftest
        result=0
        prev_result=0
        ain=0
        bin=0
        limit=2**8
        for(j 0 100
                ain<7:0>=random(limit)
                bin<7:0>=random(limit)
                prev_result=result
                result<15:0>=ain<7:0> * bin<7:0>
                xv( ain bin 0 1 0 prev_result )
        )

endtest
```

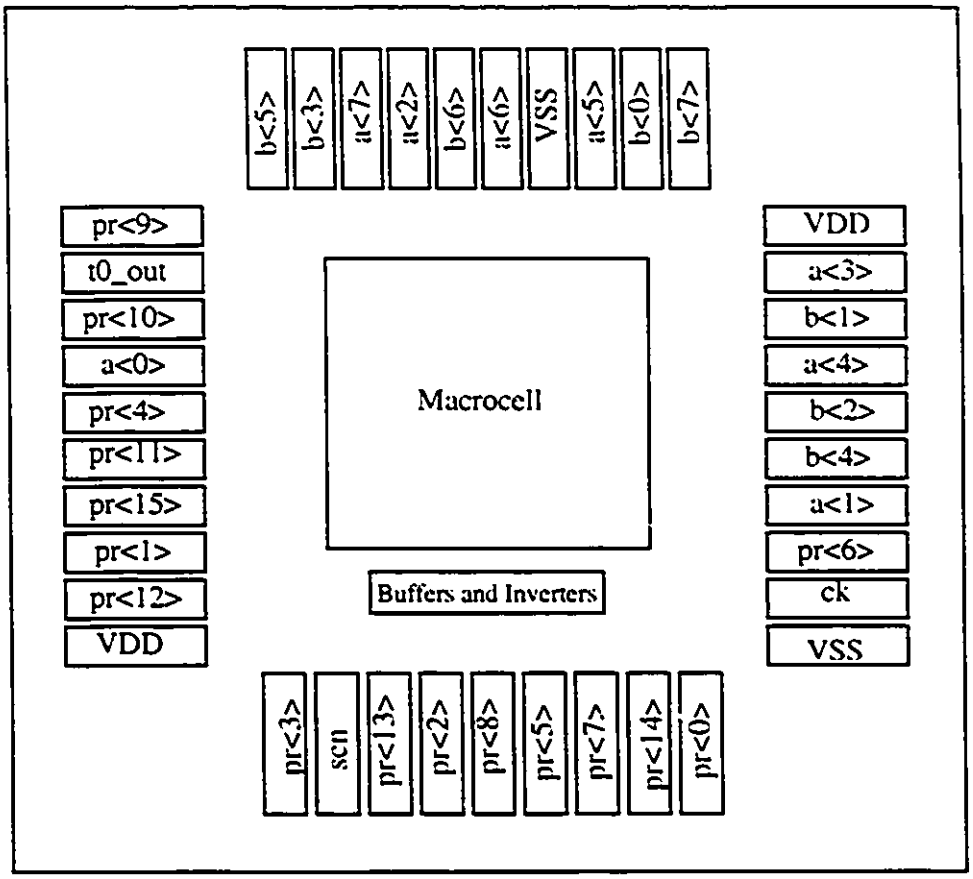
Appendix F

MULTIPLIER CHIP

F.1 Summary and Pinout

The BiCMOS version of the two bit full adder multiplier macrocell was submitted for fabrication as a stand-alone chip of size 3006.6μm × 2833.9μm. This appendix contains implementation specific details of that chip. In this section, signals associated with the chip are referenced in boldface type. Figure F.1 shows the pinout diagram for the chip.

Figure F.1: Pinout Diagram



The two 8 bit input operands are denoted as follows as **a<7>a<6>...a<1>a<0>**, and **b<7>b<6>...b<1>b<0>**, and they produce a 16 bit product denoted as **pr<15>pr<14>...pr<1>pr<0>**. The outputs are buffered through BiCMOS inverters before being sent through the output pads offchip.

The inputs are clocked into the macrocell on the negative edge of the clock input, **ck**, and the outputs are clocked out of the cell on the positive edge of the same signal. The clock signal is buffered through a BiCMOS buffer[1].

The scan enable input, **scn**, for the scan latches is buffered through a BiCMOS buffer.

A simple test circuit is included on the chip. The signal **a<0>** is applied to an inverter and the result, **t0_out**, is sent off chip. The chip has two VDD pads and two VSS pads. The enable pins on all of the output pads are tied to VDD.

The top level layout representation for this chip is "tbfam_g_bchip1/layout"

[1] Note that both buffers and inverters were used in this design.

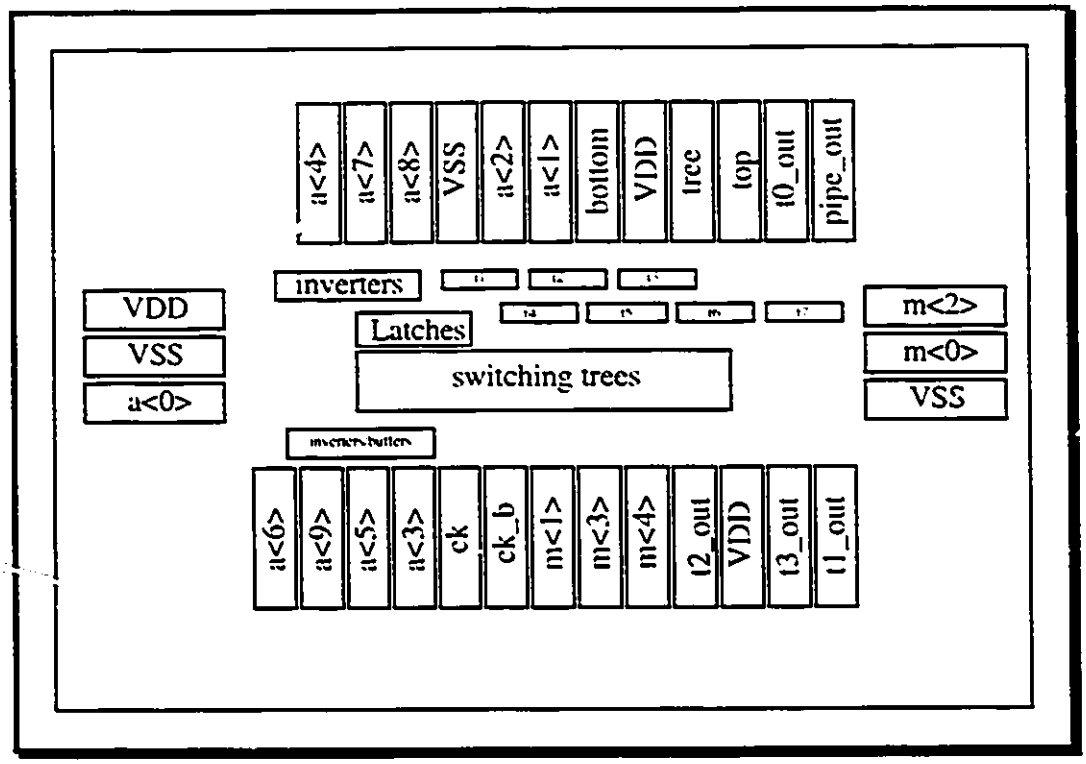
Appendix G

MOD 17 CHIP

G.1 Summary and Pinout

A design was submitted for fabrication containing a mod 17 multiplier and several test structures, and this appendix contains implementation specific details of that chip. The size of the integrated circuit was 3489.4μm × 1935.2μm, and a pinout diagram is given in Figure G.1, which also shows the general partitioning of the design. In this section, boldface type indicates the name for a chip associated signal.

Figure G.1: Pinout Diagram



The mod 17 multiplier and associated latches is contained in the blocks labeled "switching tree" and "Latches". DSCS latches were used, and their outputs were buffered with small static inverters, and inverted again by tcell inverters before being sent to the output pads. The inputs to the multiplier consist of two 5 bit numbers. If the operands are expressed as $x_4x_3x_2x_1x_0$, and $y_4y_3y_2y_1y_0$, then they are mapped to the chip inputs as follows:

$x_4x_3x_2x_1x_0 \Rightarrow$ **a<9>a<8>a<7>a<6>a<5>**

$y_4y_3y_2y_1y_0 \Rightarrow$ **a<4>a<3>a<2>a<1>a<0>**

These chip input signals are applied to the switching tree inputs with BiCMOS buffers to drive these heavily loaded nodes. Input complementation is provided by internal BiCMOS inverters. The product is denoted as **m<4>m<3>m<2>m<1>m<0>**[1]. These outputs are fed through small, local isolation inverters, through tcell inverters, and applied to the output pads.

Several test structures were included on the chip as well, and they are indicated in Figure G.1 as blocks t1 through t7. The inputs to the test structures are obtained from chip inputs **top**, **tree**, and **bottom**. The cell structures, as well as the input signal correspondence, are as illustrated in Figure 5.18. All test structures are based on the DSCS master—slave latch. The t1 block's NMOS chain is 12 transistors high, and the bottom load consists of 8 more transistors. They all possess a W/L ratio of 3μm/.8μm. The output of the master slave latch is fed into a local static inverter to isolate it from heavy loads, and then through a heavier tcell inverter before being fed to the output pad as signal **t1_out**.

The test structure labeled t2 is a DSCS master—slave latch with the same configuration as above, but with different transistor sizes. All NMOS devices in the tree and the n latch portion of the master—slave latch are 1.8μm/.8μm devices[2]. The output is inverted twice before being applied to the output pad as signal **t2_out**.

The t3 test structure is a DSCS master—slave latch with a 14 high transistor chain and a 12 transistor bottom load. These devices are of size ratio 1.8μm/.8μm. As before, the output is inverted twice before being applied to the output pad as signal **t3_out**.

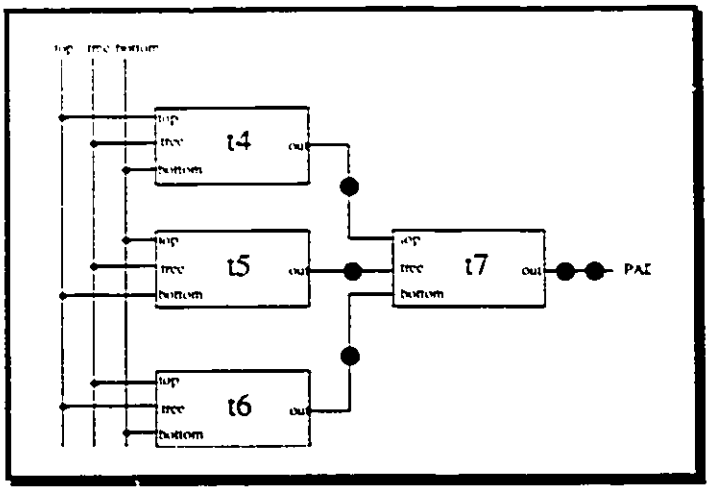
The test structures t4, t5, t6, and t7 are intended to simulate a two stage pipeline. Individually, they are identical test structures to t1. The same three chip input signals are applied to different cell terminals, thus allowing variation. The structure is illustrated in Figure G.2. The large dark circles represent static inverters. The output of this structure is applied to an output pad as signal **pipe_out**.

---

[1] The weights of these output digits are assigned arbitrarily in this case.

[2] As mentioned before, minimum size devices were not used due to contact considerations.

Figure G.2: Simulated Pipeline



A trivial test circuit was also included consisting of a static inverter. The chip input signal **a<0>** is inverted and applied to an output pad as signal **t0_out.**

Clock signals are applied with signals **ck** and **ck_b** corresponding to the clock signal and its compliment. Both are delivered to the chip through BiCMOS buffers placed between the input pad and the chip circuitry.

The top level layout representation for this chip is "mod17_chip/layout"

Vita Auctoris

James C. Czilli was born on June 28, 1968 in Buffalo, New York. He completed his high school education in Windsor, Ontario at Assumption College School in 1987, at which time he entered the Bachelor of Applied Science program at the University of Windsor in Electrical Engineering. He held several scholarships throughout his undergraduate studies, and was named on the Dean's List in 1991, his graduating year. He entered the Master of Applied Science program in Electrical Engineering at the University of Windsor, and received an Ontario Graduate Scholarship. He is currently employed by Northern Telecom Electronics in Nepean, Ontario, working in an ASIC design group. His interests include CMOS/BiCMOS digital circuit design techniques, computationally specific VLSI architectures, microprocessor design, computer arithmetic and VLSI algorithms. James C. Czilli is a member of the IEEE.