The First-E
Charge Pumps for Today's Low-Cost,
High-Performance Mobile Devices

A groundbreaking tool for circuit design engineers, ***Charge Pump Circuit Design*** is the first book to focus solely on the design and implementation of charge pumps used in EEPROMs, Flash memory, White LED drivers, and a myriad of other circuits finding mass applications in PDAs, digital cameras, MP3 players, video recorders, cell phones, USB drives, and more.

Written by two of today's leading circuit designers, ***Charge Pump Circuit Design*** explores the basic operations, design criteria, and newest approaches for designing state-of-the-art charge pumps. The authors explain the different architectures and requirements, providing comprehensive information for each stage in the design process. Filled with 100 detailed illustrations, this time-saving reference also presents a wealth of practical design tips and potential pitfalls to avoid.

Charge Pump Circuit Design features:

- The latest design techniques for creating highly efficient charge pumps for any type of application requirement
- Step-by-step guidelines for completing a charge pump design—from initial concept to implementation of actual layout
- Thorough mathematical derivations and analyses of operations that are applicable to all charge pump requirements, regardless of the system being designed

Inside This Landmark Design Reference:

- History of High-Voltage Charge Pumps
- Basic Operations of Charge Pumps
- Criteria of a Generic Charge Pump
- How to Design a Basic Charge Pump
- How to Design a Better Charge Pump
- Charge Pump Architectures
- Future Design Reference

The McGraw-Hill Companies

books.mcgraw-hill.com

ISBN 0-07-147045-X

59995

ELECTRONIC ENGINEERING

Charge Pump Circuit Design

Pan
Samaddar

McGraw-Hill
ELECTRONIC ENGINEERING

Charge Pump Circuit Design

✔ Architecture

✔ Concept to Implementation

✔ Operation and Analysis

Feng Pa

ABOUT THE AUTHORS

FENG PAN is doing circuit design at SanDisk Corporation, where he is involved in defining the architecture and implementing charge pumps and related high-voltage circuits over several generations of NAND Flash memory products. He is a holder of 18 U.S. design patents. Mr. Pan previously worked at AMD, where he contributed to the company's NOR Flash memory chip designs. He holds an MS degree from Stanford University and a BS degree from U.C. Berkeley.

TAPAN SAMADDAR is a design engineer at SanDisk Corporation, where he leads the company's high-voltage circuit designs. Previously, he was employed by T-RAM, Inc., where he was involved in the design of high-speed SRAM-compatible memory chips. He is a holder of several U.S. design patents. Mr. Samaddar has also worked on high-density NOR Flash memory at Atmel Corporation, and on high-speed cache memory designs at ST Microelectronics.

Charge Pump Circuit Design

Feng Pan
Tapan Samaddar

McGraw-Hill

New York Chicago San Francisco Lisbon London Madrid
Mexico City Milan New Delhi San Juan Seoul
Singapore Sydney Toronto

The McGraw-Hill Companies

Charge Pump Circuit Design

1 2 3 4 5 6 7 8 9 0 DOC/DOC 0 1 9 8 7 6

ISBN 0-07-147045-X

The sponsoring editor for this book was Wendy Rinaldi, the production supervisor was Jean Bodeaux, the editorial supervisor was Jody McKenzie, and the project manager was Samik Roy Chowdhury (Sam). It was set in New Century Schoolbook by International Typesetting and Composition. The art director for the cover was Margaret Webster-Shapiro.

Printed and bound by RR Donnelley.

McGraw-Hill books are available at special quantity discounts to use as premiums and sales promotions, or for use in corporate training programs. For more information, please write to the Director of Special Sales, McGraw-Hill Professional, Two Penn Plaza, New York, NY 10121-2298. Or contact your local bookstore.

Contents

Preface

Charge pumps are finding increased attention and novel diversified usage in the new era of nanometer-generation chips used in different systems, specifically those incorporating nonvolatile memory. Many of the present and future nanometer-generation chips' performance depend heavily on the ability to efficiently generate high voltages on-chip while meeting stringent power and area requirements. And yet, charge pump design, being purely analog in nature and involving high voltage, needs meticulous design techniques, intensive semiconductor device analysis, careful design layout planning, and accurate parasitic extraction process to produce excellent results in real implementation on silicon.

This book is a product of our years of quest for a practical book on charge pump circuit design. Having made significant contributions in different successful projects at various companies, we have always felt the need for a book on the topic of charge pump design. From our early days we constantly felt the challenge of working on a subject where there are no books and our sources of information were limited to only a few pages of description on various text books and different IEEE-published papers and journals. Most of these documents, while giving us skeletal ideas about the basic architecture and enhancements of charge pumps, did nothing to guide us intricately through different design conceptions and implementation processes. Charge pump, being a pure analog design, carries an inherent risk of unanticipated effects, which when overlooked can significantly reduce circuit performance or cripple the circuit operation.

Both of us have individually looked for books or materials on this topic and asked colleagues and veterans for any published materials, but we soon found out that everyone wished there was a book on this particular topic. By writing this book we have tried to bring our combined personal design experiences along with the knowledge assimilated from our study of different journals and research papers. This book covers the basics of charge pump circuits in detail and provides a thorough mathematical derivation and analysis of charge pump operation. It also strives

to explain the different aspects for an excellent charge pump design, and explains each step in detail. Every effort has been made to provide enough hands-on design information, potential pitfalls to avoid, and practical ideas harnessed from our years of designing charge pumps, which are being used in many chips, finding mass scale adoption.

This book assumes a basic knowledge of semiconductor device physics and MOSFET operation and is targeted toward almost every semiconductor chip design engineer who is involved in analog circuit design and memory circuit design. The book takes a relatively novice reader through various aspects and gives sufficient information to enable him or her to complete their design from conception to actual layout implementation. Further audiences include systems designers and board level integrators. Also this book should be essentially helpful to almost all electrical engineering professors and students at all levels.

This book is organized into nine chapters. Chapter 1 starts with a history of charge pump evolution from the Noble award winning work of Cockcroft and Walton to the eventual adaptation of John F. Dickson. Chapter 2 is intended for a quick refresh of basic MOS device physics and different second order effects relevant to charge pump operation. It also discusses SPICE simulators and BSIM models while providing suggestions to avoid the dreaded SPICE convergence issue. Chapter 3 dives straight into the heart of charge pump operation and quantitatively analyzes the pump characteristics. Chapter 4 is where the basic implementation details are discussed along with different charge pump controlling blocks. It analyzes the different pump regulation schemes and quantifies them. Chapter 5, a prelude to Chapter 9, introduces the different parameters for charge pump specifications and discusses various implementation details. Now, once the basic operation of a 2-phase charge pump and its many characteristics and specifications have been understood, Chapter 6 takes it to the next level and discusses how to design a better charge pump. Chapter 7 discusses various charge pump architectures, such as the modified 2-phase charge pump, the 4-phase charge pump, and the CTS charge pump. Chapter 8 provides future design references and discusses different circuit and system effects that affect the performance of the pump. Finally, Chapter 9 provides a practical design example and discusses the influence of different parameters while analyzing the characteristics of a charge pump.

It is believed that if the reader applies appropriate techniques, which have been presented here, he or she should be able to design charge pumps that meet design and performance specifications and produce excellent operating circuits on silicon.

Feng Pan
Tapan Samaddar

Acknowledgments

I wish to acknowledge all the individuals who provided tremendous help and support. A special thanks go to Wendy Rinaldi and Alex McDonald, acquisitions editors, for believing in us as first-time writers, and guiding and supporting us throughout the entire duration of this project. I am thankful for Samik Roy Chowdhury (Sam) and his team at International Typesetting and Composition for all their assistance and hard work in making this timely production.

My co-author, Tapan Samaddar, for sparking the idea to write this book on charge pump design, and for his inspiring work in researching and completing this book.

My wife, Judy, for her patience and understanding. During the last year while she was pregnant, I had to work many late nights and weekends on this book. She encouraged me and supported me in many ways. I am very grateful to her. My lovely daughter, Tiffany, who was born in early April this year, for her trust and understanding in her daddy, and for forgiving her daddy for not being able to give her time. I am indebted to her. I would also like to thank my parents and my brother for their guidance and support throughout my life.

Many thanks to those who supported me in the past to make this book production possible.

Feng Pan

Writing this book began with a lot of energy and excitement. However, after about six months of persistent writing, drawing, and revising, while meeting tough deadlines, schedules, and surprises at the office, we reached midway in the book and immediately began to feel the streaks of insanity, realizing that the book will never finish on time. It was only through the continued efforts, support, and help of many outstanding individuals that we were able to complete this book.

A special thank you goes to Wendy Rinaldi and Alex McDonald, who have given us much needed guidance, helped encourage the progress, and collated the manuscript. Great appreciation is also given to Sam and his team, who did a marvelous job editing the manuscript, shaped it into a book, and helped finish everything on time.

My loving wife, Somali, has made many contributions, from typing characters and drawing figures to finding numerous errors and raising questions that made me reexamine my own understanding. I am very grateful to her for her support. I would also like to thank my mother and my father for providing me with emotional support and for all the help they have given me throughout the years.

Did I forget anyone? Thanks to all of you who inspired and provided support to write this book.

Tapan Samaddar

Introduction

The approaching nanometer generation of large-scale integrated (LSI) circuits requires power-supply voltages of less than 2 V to enable low-power operation and increased battery life. In addition, because low-power technology has become a primary target for the mainstream LSI designs to meet with the nomadic computing era, oxide thickness, transistor dimensions, and voltage-scaling approaches that are most effective for low power are accelerating and spreading at an unprecedented pace. Therefore, low-voltage circuit technologies for processors, memory, and analog circuits are intensively being investigated.

Many of the system blocks—such as EEPROMs, Flash memories, power management blocks, audio and video codecs, image sensor circuits, and displays—require internal voltages higher than the system supply voltage. This internal high-voltage supply needs to be generated in-system or on-chip. The traditional approach of switch-capacitor circuits or inductor-based linear regulators consumes too much power and silicon area to justify today's shrinking needs. An on-chip charge pump design provides an excellent solution and eliminates the need for an inductor. Having no inductor alleviates any potential electromagnetic-interference concerns that could have an impact on sensitive RF receivers or wireless chipsets. Another advantage is to reduce the cost of using discrete off-chip components. The charge pump solution eliminates the need for DC/DC boost converters and expensive low-profile inductors that are required to meet the size limitations of handheld devices and cellphones.

The advent of personal information devices, digital cameras, and MP3 players has fuelled a boom for nonvolatile memories, particularly Flash memories, because of their high-density, moderate power consumption and high endurance for mechanical shock and vibration. Solid state memories, such as Flash, contain no moving parts and allow for easy and fast data storage. EEPROMs and Flash memory have been some of the biggest drivers to create better and efficient charge pumps. Even though the supply voltage is decreasing, a Flash memory will still need a high internal programming/erase voltage, up to 30 V, regardless of the

power supply trend, and this voltage also needs to be controlled precisely to achieve a narrow deviation in the threshold voltages of its memory-cell transistors.

A familiar problem in system engineering is the subsystem whose power requirements are not met by the main power supply. In such cases, the available supply rails are not directly usable, nor is the direct use of battery voltage (when available) always an option. Lack of space can prevent inclusion of the optimal number of batteries, and in other cases the gradually declining voltage of a discharging battery is not acceptable for the application. Voltage converters can generate the desired voltage levels, and charge pumps are often the best choice for these applications requiring some combination of low power, simplicity, and low cost. Charge pumps are easy to use, because they require no expensive inductors or additional discrete components. Further, charge pumps can be the only option for certain applications, such as those in telecom applications, which require generating +5 V from the available –48 V.

With the increasing popularity of color LCD displays in cellphones, PDAs, and digital cameras, white LEDs are becoming popular illumination sources. Whereas monochrome displays can use colored light sources, such as electroluminescent backlights or colored LEDs, color displays require a white light source to properly display color. The ubiquitous red and green LEDs have a typical forward-voltage drop of about 1.6 V to 2.4 V and can be driven by a simple battery pack. White LEDs, however, typically have a forward-voltage drop of 3 V to 4 V and are more likely to need a separate power supply. The traditional direct LED drivers and inductor-based boost converters have their own limitations. Charge pumps are increasingly finding ground in these applications due to their better performance—they provide the smallest and lowest-cost solution because they rely only on small capacitors for high voltage generation. As video features become more integrated into mobile phone use, improvements in power consumption for LCD backlighting are essential for the maintenance and improvement of overall battery life. Chip vendors competing for a share of this red-hot white LED driver market are pitching advanced charge pump sources that deliver greater backlighting power for the color displays in larger and more complex portable/wireless devices.

Therefore, much of the present and future chip performance will depend heavily on the ability to efficiently generate high voltages on-chip while meeting stringent power and area requirements. And yet, charge pumps, being purely analog in nature and involving high voltage, need meticulous design techniques, intensive semiconductor device analysis, careful design layout planning, and an accurate parasitic extraction process to produce excellent results in real implementation on silicon.

To summarize, charge pumps are finding increased attention and novel diversified usage in the new era of nanometer-generation chips used in

different systems. This book strives to explain the different aspects for an excellent charge pump design, and explains each step in detail. It is full of extra hands-on design information, potential pitfalls to avoid, and practical ideas harnessed from the authors' years of charge pump design experience, which are currently being used in many chips, finding mass scale adoption.

Chapter

1

History of the High-Voltage Charge Pump

The quest for generating a high voltage supply from an available lower voltage supply has existed since the discovery of electricity. The invention of the "induction ring" by Michael Faraday, the British physicist and chemist, in 1831, began the process of generating high voltages from an available lower input voltage using transformers. The need for producing even higher voltages was accentuated by the requirements from physicists, using particle accelerators, to create high energy particles for studying subatomic physics. It was only during this time that Cockcroft and Walton invented a novel method for generating extremely high voltages using a unique connection of discrete diodes and capacitors—a technique that was later adopted by John F. Dickson for implementation on a modern integrated circuit. This chapter will start by examining the transformer and its shortcomings and then gradually lead to a discussion of Dickson's implementation of the charge pump.

1.1 Using a Transformer to Generate High Voltages

A transformer makes it possible to convert AC power at a given voltage level to AC power at a different voltage level. It, of course, cannot increase the maximum power that could be delivered to the output load through the transformation process. In the ideal situation, the power delivered at the transformed output is equal to the power consumed at the input port. If the transformed voltage level is raised, the current at transformed node is proportionally lowered, and vice versa. The transformer, shown in Figure 1-1, is constructed using a ferromagnetic core around which two sets of coils, or multiple coils, of insulated wire

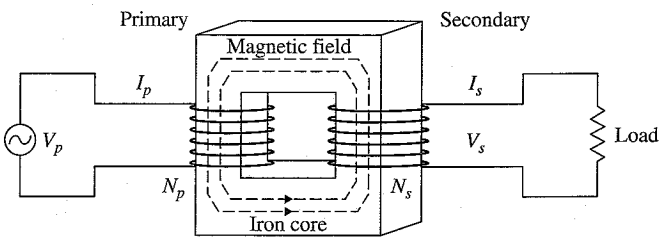


Figure 1-1 A simple transformer.

are wrapped. The input line connects to the "primary" coil, whereas the output lines connect to the "secondary" coils. The alternating current in the primary coil induces an alternating magnetic flux that "flows" around the ferromagnetic core, changing direction during each electrical cycle. The alternating flux in the core, in turn, induces an alternating current in each of the secondary coils. The voltage at the output of the secondary coils is directly related to the primary voltage by the turn's ratio, or the number of turns in the primary coil divided by the number turns in the secondary coil.

For an ideal transformer,

$$\frac{V_s}{V_p} = \frac{N_s}{N_p} \tag{1-1}$$

where V_s and V_p are the voltages at the secondary and the primary nodes, respectively, and N_s and N_p are the number of turns of the secondary and primary coils, respectively. For instance, if the primary coil consists of 100 turns and carries 110 volts, and a secondary coil consists of 1000 turns, the secondary voltage is the following:

$$V_s = 110\text{ V}\frac{1000}{100} = 1100\text{ V}$$

The voltage transformation ratio (primary voltage to secondary voltage) and the current transformation ratio (primary current to secondary current) actually depend on the turns' ratio. Thus, the AC output voltage can either be decreased or be increased by selecting the correct number of turns. A transformer may have multiple secondary coils to feed a number of electrical loads. Yet, a transformer will only work with an input that is alternating voltage source, whereas, in general, electronic circuits require a DC voltage supply to operate. If a high-voltage DC output is required, the stepped-up output AC signal needs to be converted into a DC voltage by using a rectifier—a cumbersome process especially at high voltages. Further, generating high voltages using a transformer makes the transformer very large, heavy, and inefficient—a sure handicap in the modern trend of micro-miniaturization.

1.2 The Cockcroft-Walton High-voltage Charge Pump

Voltage multiplication greater than twice the supply voltage can also be achieved by cascading more than one diode capacitor voltage stage in series. The Swiss physicist Heinrich Greinacher first proposed this kind of voltage multiplier back in 1919. Later, this technique was used by John Douglas Cockcroft and Ernest Thomas Sinton Walton to generate voltage potentials of more than 800,000 volts in their particle accelerator, which in 1951 won Cockcroft and Walton the Nobel Prize in Physics for their research, titled "Transmutation of atomic nuclei by artificially accelerated atomic particles."[1] Cockcroft and Walton, depicted in Figure 1-2, at the Cavendish Laboratory in Cambridge, England, sought a way into the nucleus through a prediction of quantum mechanics. In 1930, Cockcroft and Walton used a 200-kilovolt transformer to accelerate subatomic particles along a straight discharge tube, but they concluded that the particle's energy was not sufficient to trigger any effects and therefore decided to seek higher particle energies.

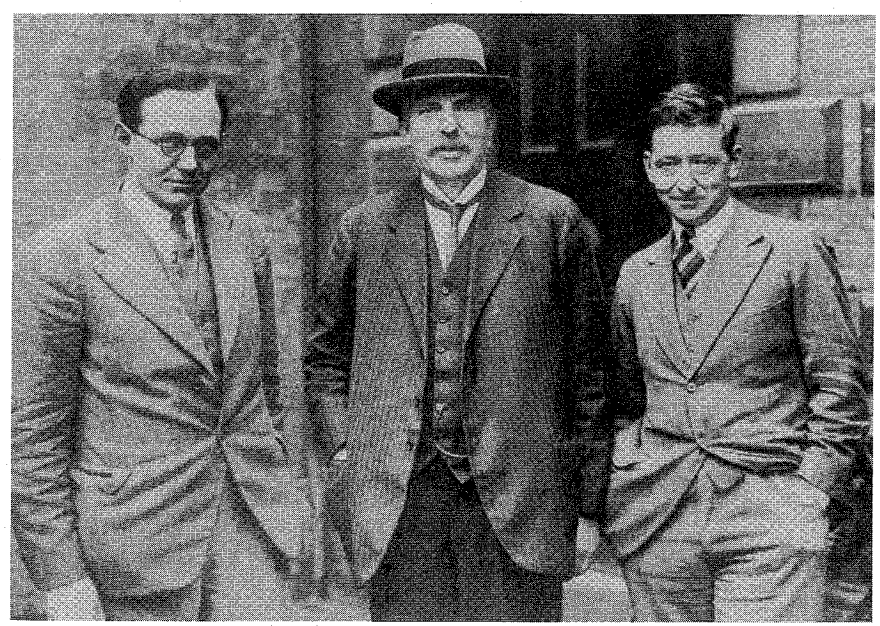


Figure 1-2 John Cockcroft, Ernest Rutherford, and E.T.S. Walton.

To penetrate the atomic nucleus, Cockcroft and Walton built a voltage multiplier that used an elaborate stack of capacitors connected by diodes acting as switches. By using only capacitors and diodes, these voltage multipliers can step up relatively low voltages to extremely high values, while at the same time being far lighter and cheaper than transformers. By activating and deactivating switches in proper sequence, they could build up a potential of more than 800 kilovolts from a transformer, acting as a primary source, generating 200 kilovolts. They used the potential to accelerate subatomic particles along an evacuated tube 8 feet long. In 1932, Cockcroft and Walton put a lithium target at the end of the tube and found that the accelerated particles successfully disintegrated a lithium nucleus into two alpha particles.

The Cockcroft-Walton multiplying circuit is shown in Figure 1-3. Three capacitors (C_A, C_B, and C_C), each of capacity C, are connected in series, and capacitor C_A is connected to the supply voltage, V_{DD}. During phase ϕ, capacitor C_1 is connected to C_A and charged to voltage V_{DD}.

When the switches change position during the next cycle, ϕ_b, capacitor C_1 will share its charge with capacitor C_B, and both will be charged to $V_{DD}/2$ if they have equal capacity. In the next cycle, C_2 and C_B will be connected and share a potential of $V_{DD}/4$, while C_1 is once again charged to V_{DD}. It is thus obvious that if this process continues for a few cycles, charge will be transferred to all the capacitors until a potential of $3V_{DD}$ is developed across the output V_{out}. In general, the switches are replaced by diodes in the actual circuit, and the clocking action is provided by an alternating voltage source, as shown in Figure 1-4.

It can be observed from Figure 1-4 that the output voltage, V_{out}, of each stage is twice the peak input voltage, V_{peak}. Hence, theoretically by using five stages, the input voltage can be stepped up by ten times. This type of circuit can actually be used to generate voltages in the order of megavolts, in some applications, by cascading a larger number of stages.

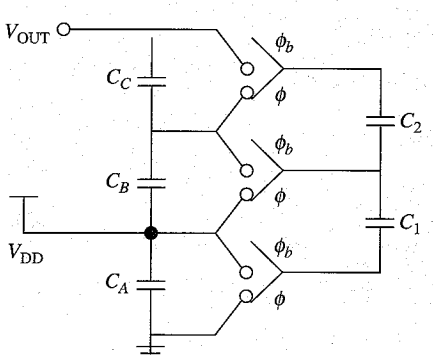


Figure 1-3 Cockcroft-Walton multiplying circuit.

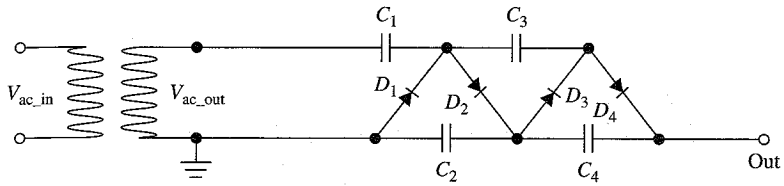


Figure 1-4 Cockcroft-Walton multiplying circuit using diodes and with a step-up transformer in the primary stage.

The Cockcroft-Walton voltage multipliers were, and still are, used in applications such as X-ray tubes, particle accelerators, electrostatic devices, and many other devices, making use of very high DC voltages.

The output voltage of the Cockcroft-Walton voltage multiplier can be expressed as

$$V_{out} = 2 \times n \times V_{peak} - V_{load} \tag{1-2}$$

where V_{load} is the drop in the output voltage when the multiplier is supplying an output current. In the absence of any output load, or at steady conditions, $V_{load} = 0$ V. Even though the number of stages, n, can be very large, efficient voltage multiplication will occur only when the coupling capacitors, C, are much greater than the stray parasitic capacitors, C_S, present at every node. The capacitive division effect in essence will reduce the voltage coupled at every stage. Further, the output impedance increases rapidly as the number of multiplying stages are increased. Because the original Cockcroft-Walton multiplier was built using discrete components, the coupling capacitors could be made sufficiently large for efficient multiplication and adequate drive capability. However, this type of multiplier does not lend itself to integration in monolithic form because, in practice, on-chip capacitors are limited to a few picofarads with relatively high values of stray capacitance to substrate.[2,3]

1.3 The Dickson Charge Pump

In order to overcome the aforementioned limitations,[4] John F. Dickson proposed a voltage multiplier circuit, shown in Figure 1-5. It operates in a similar manner as the classic Cockcroft-Walton multiplier circuit. However, the nodes of the diode chain are coupled to the inputs via capacitors in parallel, instead of in series, so that the capacitors have to withstand the full voltages developed along the chain. This is not a problem here, provided that the integrated circuit process limits are not exceeded. As will be shown later, the advantages of this configuration are that efficient multiplication can be achieved with relatively

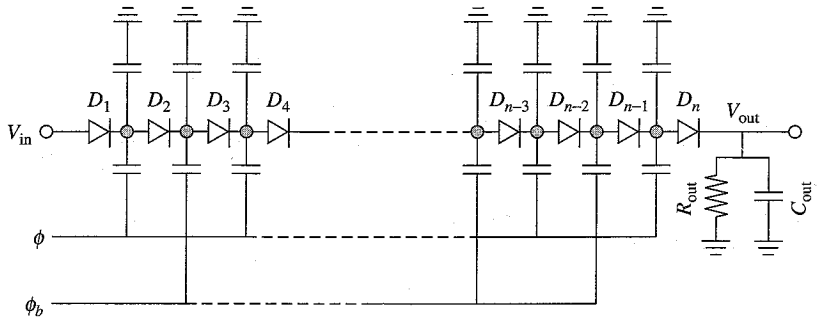


Figure 1-5 Original Dickson charge pump with diode-capacitor implementation.

high value of stray capacitances and that the current drive capability is independent of the number of multiplier stages.[5]

A practical implementation of the Dickson charge pump used in nonvolatile memories is shown in Figure 1-6. In most of the semiconductor logic process, isolated diodes are not available, and hence the multiplier chain is implemented using diode-connected MOS transistors, as shown in Figure 1-6. In this case, because NMOS transistors are used instead of diodes, the diode forward voltage, V_D, is replaced by the NMOS threshold voltage, V_t, which is a function of the node voltage of each stage.

For a few years, the basic Dickson charge pump addressed almost all different high-voltage-generation issues until the advent of submicron design technology. The continuous quest for better CMOS performance has scaled the supply voltage below 1.8 volts, and a new problem has gained the spotlight—the body effect.

1.3.1 The body effect

The threshold voltage of a NMOS transistor can be represented as

$$V_t = V_{t0} + \gamma\left(\sqrt{\phi_s + V_{SB}} - \sqrt{\phi_s}\right) \tag{1-3}$$

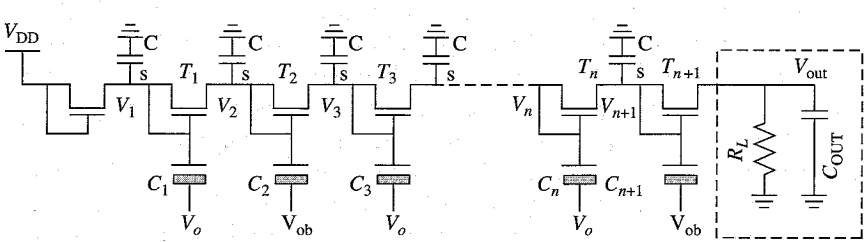


Figure 1-6 MOSFET implementation of Dickson charge pump.

where ϕ_s equals the surface potential at threshold and is represented by

$$\phi_s = 2V_t \ln\frac{N_A}{n_i} \tag{1-4}$$

γ equals the body effect coefficient and is represented by

$$\gamma = \frac{t_{ox}}{\varepsilon_{ox}}\sqrt{2q\varepsilon_{si}N_A} = \frac{\sqrt{2q\varepsilon_{si}N_A}}{C_{ox}} \tag{1-5}$$

V_{t0} equals the zero-bias threshold voltage, and V_{SB} is the source-to-body voltage bias.

It can be seen from the preceding equations that as the source voltage of a NMOS MOSFET increases, the threshold voltage of the MOSFET also rises, which results in decreased MOSFET current, I_{ds}, and hence less charge transfer takes place. The graph in Figure 1-7 shows the variation of V_{SB} versus V_t. As V_{SB} reaches above 15 V (for a particular process with V_{t0} ~ 0.08 V), the actual V_t exceeds 2.5 V, an effect that seriously diminishes the charge pump performance.

1.3.2 Implication of body effect on the Dickson charge pump

In the conventional Dickson charge pump circuit, shown earlier in Figure 1-4, the voltage gained at the nth stage is given by

$$V_n = \frac{CV_{cc}}{(C + C_s)} - V_t[V_{SB}(n)] \tag{1-6}$$

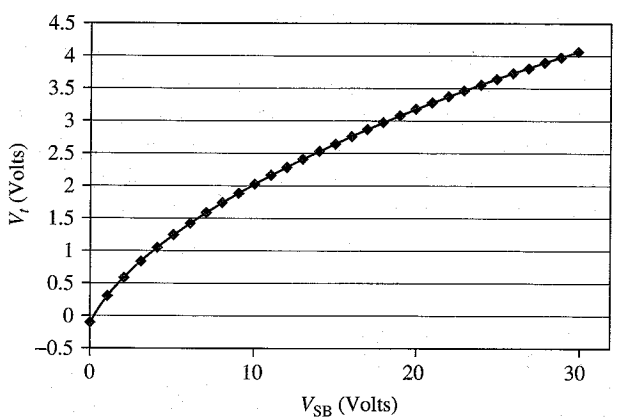


Figure 1-7 Variation of V_t with NMOS V_{SB} voltage.

where C and C_s are clock coupling capacitance and parasitic capacitance at the input node of each unit stage, respectively, V_{cc} is the clock amplitude equal to the power supply voltage, and $V_t[V_{bs}(n)]$ represents the threshold voltage of the nth NMOS transistor with a substrate bias of V_{bs}. This equation indicates that as the output voltage of each stage increases, V_n decreases due to increasing body effect. When the threshold voltage of the last stage's transistor becomes equal to $CV_{cc}/(C+C_s)$, the output voltage will not increase even with the addition of subsequent stages. Because C is much larger than C_s in typical conditions, the maximum output voltage (V_{max}) obtained by a conventional charge pump is given by

$$V_{max} = \left(\frac{V_{cc} - V_{t0}}{\gamma} + \sqrt{2\Phi_F} \right)^2 - 2\Phi_F \tag{1-7}$$

where V_{t0} is the zero-bias threshold voltage, γ is the body effect coefficient, and Φ_F is the substrate's Fermi potential. Evidently, it can be interpreted from equation 1-7 that as the supply voltage, V_{cc}, is lowered below about 2 V, the threshold voltage of the device will start to dominate the output voltage and hence it will limit the maximum output voltage.[6–8]

1.4 Better Solutions

Several methods have been proposed—such as the 4-phase charge pump, the modified 4-phase charge pump, the boosted pump clock scheme, a CTS scheme, and several hybrid versions of these combinations—to get around problems of V_t dependence and to increase the circuit efficiency for chips operating below 2.5 V supply voltages.

In the conventional 4-phase charge pumps, the pumping gain can be increased by increasing the source to gate voltage drop using the special 4-phase clocks, so the gain degradation due to threshold voltage can be alleviated. Also a 2x–4x boosted pump clock source is often used as an easy way to increase efficiency and obtain higher output voltages.

In the CTS scheme, an additional pass transistor is added for each stage; the gate of the pass transistor is controlled by the next stage voltage, which is in opposite phase. Subsequent chapters in this book describe in detail each scheme, along with its advantages and disadvantages.

References

1. Cockcroft, J.D. and E.T. Walton, "Production of high velocity positive ions," *Proceedings of the Royal Society, A*, Vol. 136, pp. 619–630, 1932.
2. Witters, J.S., G. Groeseneken, and A.E. Maes. "Analysis and Modeling of On-Chip High-Voltage Generator Circuits for Use in EEPROM Circuits." *IEEE Journal of Solid-State Circuits*, Vol. 24, No. 5, October 1989.

3. "DC/DC Conversion without Inductors." Maxim application note APP725, December 29, 2000.
4. Dickson, J. "On-chip High-Voltage Generation in NMOS Integrated Circuits Using an Improved Voltage Multiplier Technique." *IEEE Journal of Solid-State Circuits*, Vol. 11, No. 6, pp. 374–378, June 1976.
5. Pylarinos, L. "Charge Pumps: An Overview." http://www.eecg.toronto.edu/~kphang/ece1371/chargepumps.pdf.
6. Pelliconi, R., I. David, B. Andrea, P. Marco, and L.R. Pier. "Power Efficient Charge Pump in Deep Submicron Standard CMOS Technology." *Solid-State Circuits Conference*, 2001. ESSCIRC 2001. Proceedings of the 27th European.
7. "Charge-Pump and Step-Up DC-DC Converter Solutions for Powering White LEDs in Series or Parallel Connections." Maxim application note 1037, April 23, 2002
8. Stratakos, A.J., S.R. Sanders, and R.W. Brodersen. "A low-voltage CMOS DC-DC converter for a portable battery-operated system." *Power Electronics Specialist Conference*, Vol. 1, pp. 619–626, June 1994.
9. Lin, H., H.K. Chang, and C.S. Wong. "Novel High Positive and Negative Pumping Circuits for Low Supply Voltage." *IEEE International Symposium on Circuits and Systems*, Vol. 1, pp. 238–241, May 30-June 2, 1999.
10. Wang, C.C. and C.J. Wu. "Efficiency Improvement in Charge Pump Circuit." *IEEE Journal of Solid-State Circuits*, Vol. 32, No. 6, June 1997.

Chapter

2

Basic MOS Device Physics

In today's evolving chip-manufacturing industry, a solid understanding of semiconductor device physics is essential for the design and implementation of circuits that are analog in nature. In general, all circuits, whether a simple half adder or an operational amplifier, are analog in nature, while the digital abstraction is only an abstraction as long as certain design criteria are met. For its operation, an analog circuit not only relies on basic device operation, but its performance also is dictated by many second-order effects.[1] Because the circuit designer needs to make decisions about which effects need to be considered and which effects can be ignored, a thorough insight into the device physics is important. In this chapter, we will study the basics of MOSFETs (metal oxide semiconductor field-effect transistors) at an elementary level, covering the necessary topics to understand basic circuit operations. We will also talk about secondary effects of MOSFET in detail, specifically those are important for charge pump operation.

The forward- and reverse-biased p-n junctions form the basics of semiconductor technology, and the properties of forward- and reversed-biased junctions have an important influence on the characteristics of many semiconductor devices. For example, reverse-biased p-n junctions, formed by creating an n+ junction or an n-well in the p-substrate, exist in almost every integrated circuit. These junctions contribute as voltage-dependent parasitic capacitances and leakage current components. Further, a number of important characteristics of the active devices, such as the basic operation of the MOS device, breakdown voltage, and output resistance, depend directly on the properties of the depletion region of the p-n junction. To address its importance and usage at many different places, particularly for understanding charge pump operation, we provide a brief analysis of the forward and reverse-biased p-n junction diode device characteristics first.

2.1 The P-N Junction

N-type silicon has many mobile electrons, whereas p-type silicon has many mobile holes. When these two types of semiconductors are brought together in contact, two events occur at the interface due to the diffusions of majority carriers:

1. Electrons migrate from n-type material into p-type material, leaving behind uncompensated donors (+ ions).
2. Holes pour out of p-type material and move toward n-type material, leaving behind uncompensated acceptors (– ions).

The electrons and holes then recombine near the interface, producing a region depleted of free carriers, leaving behind regions with only fixed charges (ionized donors and acceptors).

The diffusion process cannot continue indefinitely because the depleted region creates an electric field whose direction opposes the diffusion of majority carriers (electrons in n-type and holes in p-type). The electric field will sweep minority carriers (holes in n-type and electrons in p-type) across the junction so that there is a drift current of electrons from the p-type side toward the n-type side and of holes from the n-type side toward the p-type side, which is in the opposite direction to the diffusion current. The junction field builds up until these two current flows are equal and equilibrium is established (no net current flow), as shown in Figure 2-1. This process also gives rise to a parasitic (depletion) capacitance.

The energy band diagram for a p-n junction with zero voltage bias can be seen in Figure 2-2.

The zero bias built-in potential can be expressed as

$$\Phi_0 = V_t \ln\left(\frac{N_A N_D}{n_i^2}\right) \tag{2-1}$$

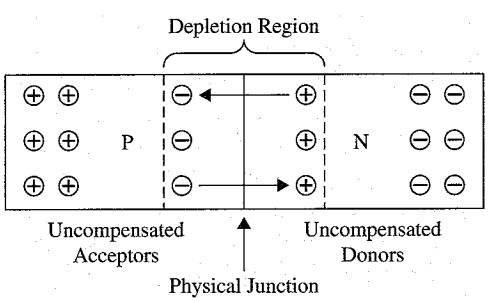


Figure 2-1 P-N junction diode.

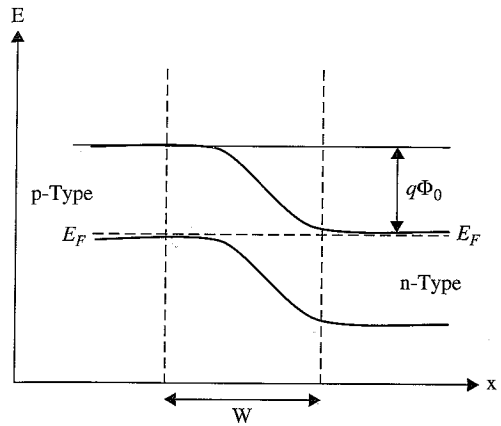


Figure 2-2 The energy band diagram for a p-n junction at 0 V bias.

where $V_t = \frac{kT}{q}$ is the thermal equivalent voltage, which is close to 26 mV at room temperature, N_A and N_D are the doping concentrations of the n-type and p-type semiconductors, respectively, and n_i is the intrinsic carrier concentration of silicon, which is about 1.45×10^{10} atoms/cm^3.

The depletion capacitance C_j of the p-n junction diode can be expressed as

$$C_j = \frac{C_{j0}}{\left[1 - \left(\frac{V_d}{\Phi_0}\right)\right]^m} \tag{2-2}$$

where V_d is forward bias voltage applied across the junction, C_{j0} is the zero-bias capacitance (that is, the effective capacitance when $V_d = 0$ V), and m is the grading profile, which usually varies from 0.2 to 0.5.

2.1.1 Reverse bias

If we apply a bias, $V_B < 0$, by connecting the positive terminal of a DC source to the n-side and the negative terminal to the p-side, the process causes electrons from the n-side to become attracted toward the positive terminal, and vice versa. This changes the depth of the depletion area, which becomes wider, and the potential barrier grows higher.

As can be seen from the energy band diagram in Figure 2-3, the built-in potential is augmented by the applied reverse bias, V_B, and the total voltage across the junction becomes $\Phi_0 + V_B$. The barrier is so high

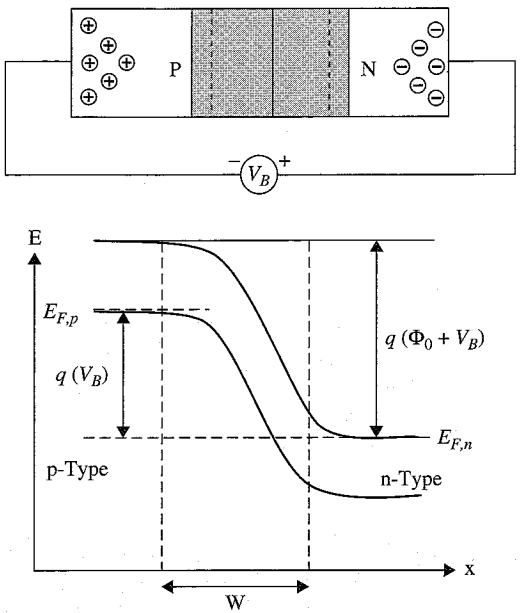


Figure 2-3 P-N junction diode at reverse bias.

that few electrons can cross from the n-type to the p-type region, thus reducing the diffusion current almost to zero.

Nonetheless, enough electron/hole pairs are generated in the depletion region, and these cause a small drift current as they are swept across the depletion region.

The drift current is relatively insensitive to the height of the potential barrier because it is created by random electron/hole pair generation and all the minority carriers generated may diffuse to the depletion region and be swept across it, whatever the size of reverse potential.

2.1.2 Forward bias

Next, if we apply a bias, $V_B > 0$, by connecting the positive terminal of a DC source to the p-side and the negative terminal to the n-side, the barrier height will be reduced to $\Phi_0 - V_B$ and the depletion region will narrow, as shown in Figure 2-4.

Consequently, majority carriers are able to surmount the potential barrier much more easily than in the equilibrium case so that the diffusion current becomes much larger than the drift current. Once the excess holes cross into the n-type region, they recombine with the electrons. This process is the same for electrons crossing into the p-type region.

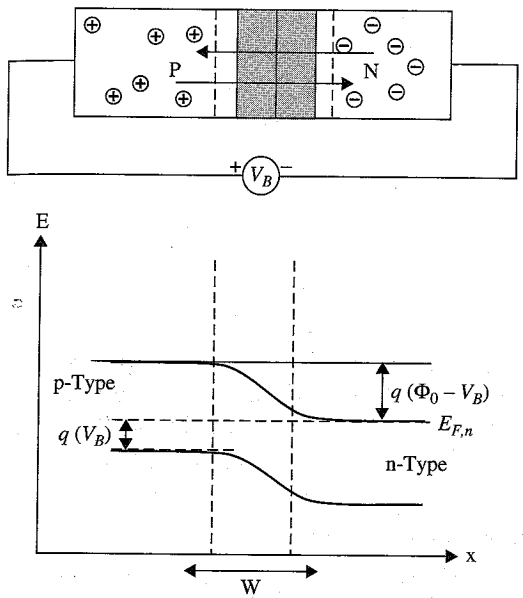


Figure 2-4 P-N junction diode at forward bias.

In equilibrium, as holes diffuse away, they must be met by a constant supply of electrons with which they recombine. Therefore, the current must be supplied at a rate that equals the concentration of holes at the edge of the depletion region. Thus, the current due to hole injection can be expressed as follows:

$$J_p = q\frac{D_p}{L_p}p_{n0}\left(e^{V/V_t} - 1\right) \tag{2-3}$$

Similarly, the current due to electron injection can be expressed as follows:

$$J_n = q\frac{D_n}{L_n}n_{p0}\left(e^{V/V_t} - 1\right) \tag{2-4}$$

The total net current is the sum of the above two equations, 2-3 and 2-4. Hence

$$I = A(J_p + J_n) \tag{2-5}$$

Substituting equations 2-3 and 2-4 in equation 2-5 we get

$$I = qA\left(\frac{D_p}{L_p}p_{n0} + \frac{D_n}{L_n}n_{p0}\right)\left(e^{V/V_t} - 1\right) \quad \text{or} \quad I = I_s\left(e^{V/V_t} - 1\right) \qquad (2\text{-}5a)$$

where I_s is known as the reverse leakage current

$$I_s = qA\left(\frac{D_p}{L_p}p_{n0} + \frac{D_n}{L_n}n_{p0}\right) \qquad (2\text{-}6)$$

As the equation indicates, this current is very strong temperature and bias dependent. We can also see that for forward bias (positive V) the net current increases exponentially with voltage, whereas for reverse bias (negative V) the current is essentially constant and equal to $-I_s$.

2.1.3 P-N junction diode characteristics

Figure 2-5 shows the I-V characteristic of a p-n junction diode. The diode's I-V characteristic can be approximated by two regions of operation. Below a certain bias voltage, the depletion layer has significant width, and the diode can be thought of as an open (nonconductive) circuit. As the bias voltage is increased, at certain conditions the diode will become conductive and allow charges to flow through from the p-side to the n-side, at which point it can be thought of as a closed switch with almost

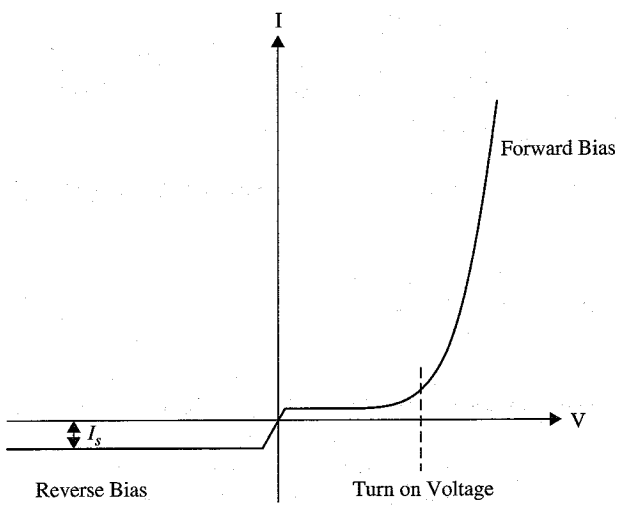


Figure 2-5 I-V characteristics of a p-n junction diode.

zero impendance. Actually, because the transfer function is exponential, a little increase in forward voltage bias causes a large conduction of current. The positive bias voltage above which the diode becomes conducting is actually the built-in potential that needs to be overcome for conduction to occur and it is generally close to 0.6 V.

During the reverse bias, the current through the device is very low (in the nA range) for all reverse voltages in general. Special diodes, such as avalanche and zener diodes, operate in the reverse bias region, such that the reverse voltage is "clamped" to a known value (called "zener voltage") and is often used for voltage regulation.

2.2 The MOS Capacitor

Before we begin our discussion of the MOSFET, it is essential to achieve a satisfactory understanding of the MOS capacitor fundamentals. Consider the n-type MOSFET in capacitor[2–5] configuration in Figure 2-6. The drain and source are grounded, and the gate is connected to a variable-voltage source. The p^+ substrate connection is also grounded. When the gate bias $V_{gs} = 0$ V, there is no potential difference across the gate and the substrate and hence there are no stored charges across the two conducting plates. The MOSFET can be considered to be in a steady state at this time.

2.2.1 The flat band condition

As shown in Figure 2-7, in an ideal MOS capacitor, the metal work function, Φ_m, is equal to the semiconductor work function, Φ_s. Therefore, the Fermi level of the semiconductor, E_{FS}, is aligned with the Fermi level of the gate, E_{FG}. There is no band bending in any region of the MOS capacitor. This assumes that the gate dielectric does not have any trapped charges and the semiconductor is uniformly doped.

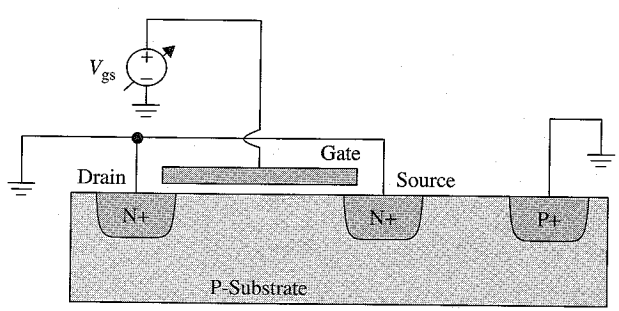


Figure 2-6 MOS transistor structure.

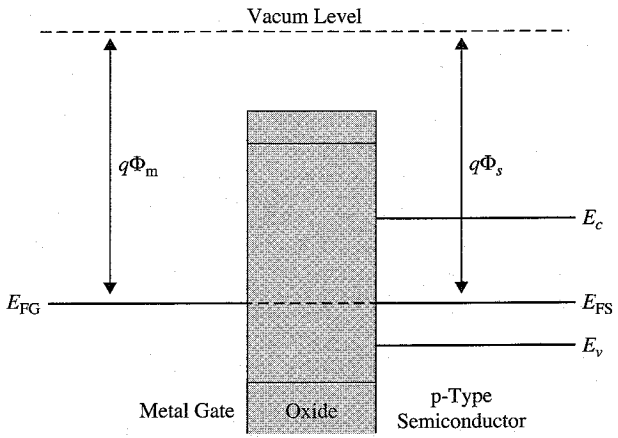


Figure 2-7 Flat band diagram at steady state.

2.2.2 Accumulation

When the gate bias, V_{gs}, is negative, the gate acquires a negative charge. The source of this negative charge is electrons supplied by the voltage source. Because charge neutrality is always maintained across the MOS capacitor, a net positive charge must be available in the silicon substrate to counterbalance the negative charge stored on the gate. This is achieved by an accumulation of majority carrier holes under the gate, as shown in Figure 2-8. (Note: Because the silicon is p-type, the majority carriers are holes.) This condition, during which the majority carrier concentration is greater near the $S_i - S_iO_2$ interface, compared to the substrate bulk, is called *accumulation*.

During a negative gate bias, the gate's Fermi level is increased with respect to the Fermi level of the silicon substrate by an amount equal

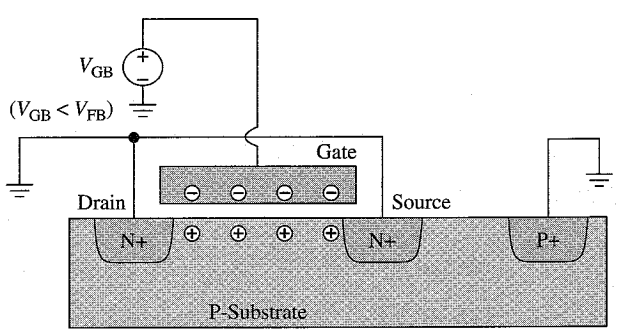


Figure 2-8 MOS transistor in accumulation.

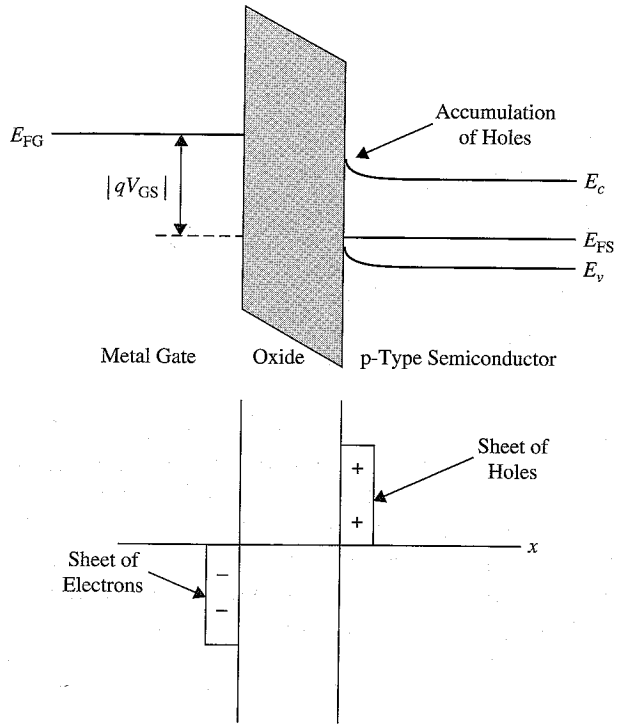


Figure 2-9 Fermi level at accumulation.

to qV_{gs} as shown in Figure 2-9. The accumulation of holes under the gate dielectric causes the energy bands in the silicon substrate to bend upward, and this brings the valence band closer to the Fermi level. Yet, during this whole process, the Fermi level in the substrate is not changed because there is no net flow of current from the gate to the substrate.

2.2.3 Depletion

As shown in Figure 2-10, as V_{gs} is increased from negative to slightly positive (not negative enough to attract a lot of holes and not negative enough to attract a lot of electrons), the surface under the gate is depleted—i.e., the holes under the gate are pushed away, leaving behind ionized, negatively charged acceptor atoms, thus creating a depletion region. The charge in the depletion region, rising from the ionized acceptor atoms, is exactly equal to the charge on the gate in order to preserve charge neutrality.

As opposed to the accumulation phenomenon, with a positive gate bias, the Fermi level of the gate is lowered with respect to the Fermi level of

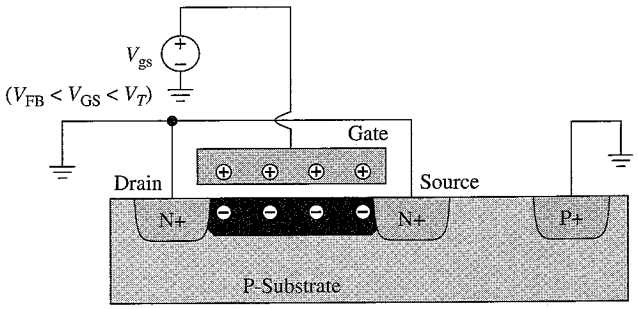


Figure 2-10 MOS transistor in depletion.

the substrate, during the depletion phase, as shown in Figure 2-11. The energy bands bend downward, resulting in a positive surface potential. Under the influence of the positive gate bias, hole depletion under the gate causes the valence band to move away from the Fermi level. As the gate potential is increased slowly, the depletion will slowly cause a band bending at the surface, such that at a particular time the intrinsic level will coincide with the Fermi level, and the surface will resemble an intrinsic material.

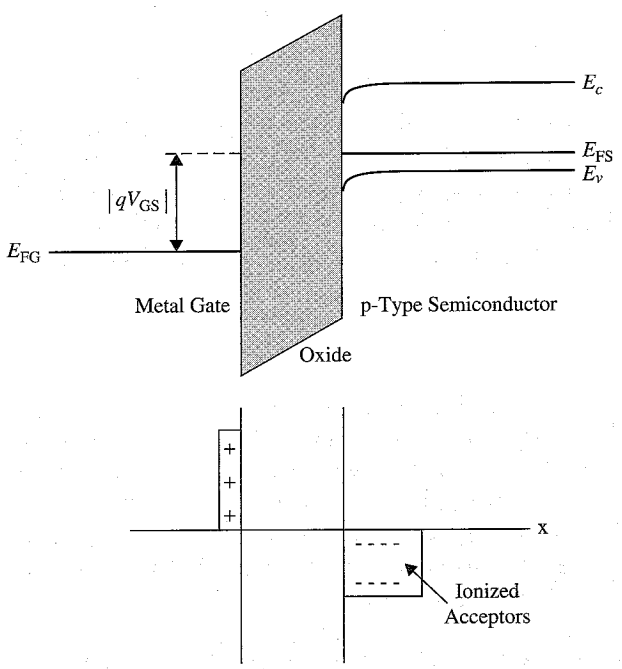


Figure 2-11 Fermi level at depletion.

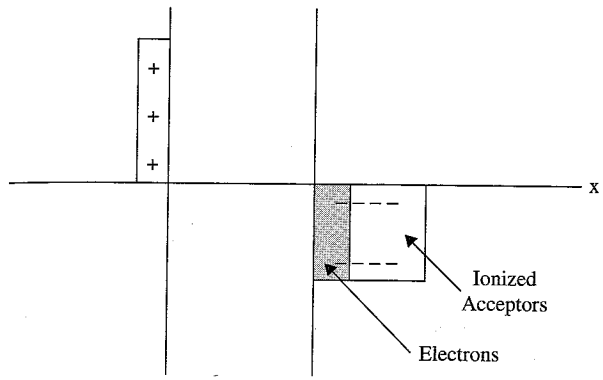


Figure 2-12 Fermi level at weak inversion.

2.2.4 Weak inversion

When the gate potential reaches above a certain threshold, as shown in Figure 2-12, the region under the gate oxide begins to attract more electrons. With electrons accumulated near the surface of the substrate, the surface begins to change gradually from intrinsic type to n-type. Because this makes the surface of opposite polarity, the region under the gate is termed "inverted." The negative charge in the semiconductor is composed of ionized acceptor atoms in the depletion region and free electrons in the inversion layer. At this onset of inversion, the electron concentration at the surface is still less than the hole concentration in the neutral bulk. Thus, this condition is referred to as "weak inversion" because the surface under the oxide is not heavily n^+.

2.2.5 Strong inversion

When V_{gs} is sufficiently large, such that a large number of electrons are attracted under the gate, the surface of p-substrate is said to be "strongly inverted" (that is, no longer p-type, as shown in Figure 2-13).

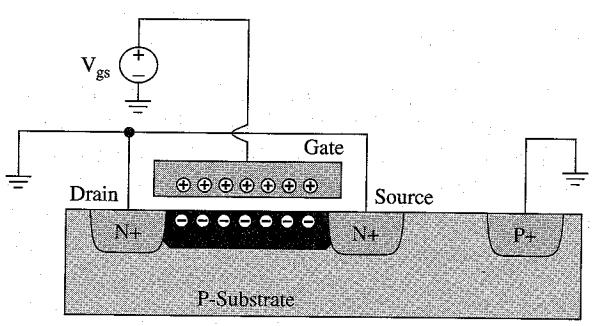


Figure 2-13 MOS transistor in strong inversion.

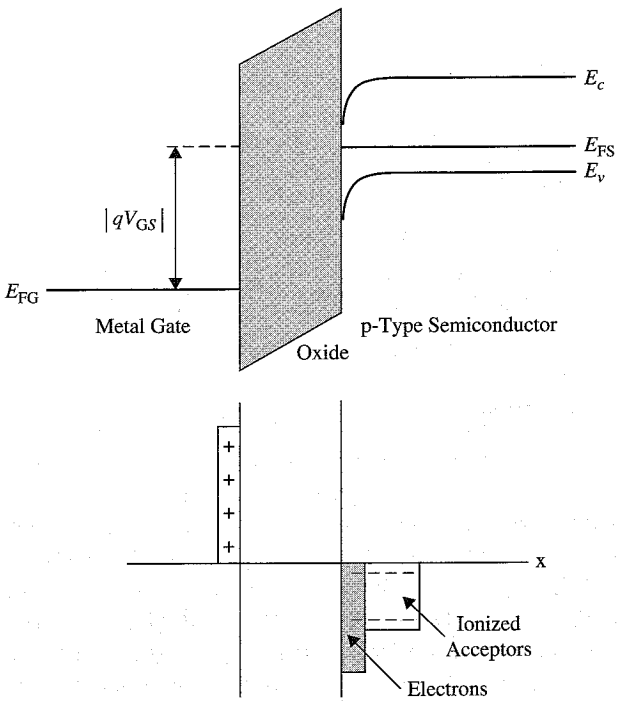


Figure 2-14 Fermi level at strong inversion.

With the increased gate bias, the band bending increases, as shown in Figure 2-14. The depletion region becomes deeper and the electron concentration in the inversion layer increases. When the electron concentration under the gate approaches the hole concentration in the bulk, a strong inversion layer is said to be formed. The surface potential required to achieve strong inversion is defined as the threshold voltage of the MOSFET transistor.

2.3 Capacitance Variation of a MOS

To understand the C-V curve of n-type MOSFET transistor shown in Figure 2-15, let us start with a very negative gate-source voltage bias. The negative potential on the gate attracts a lot of holes from the substrate to under the gate. During this condition, the MOSFET is operating in the strong accumulation region and can be viewed as a capacitor having unit capacitance of C_{ox} because the whole capacitance is essentially formed between the gate and the substrate, and is separated by gate oxide.

It is important to understand that during this time, the capacitance between gate and substrate runs through the large parasitic

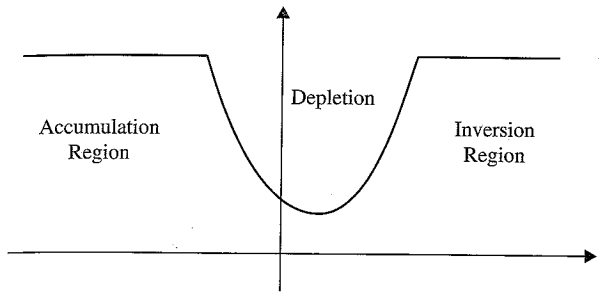


Figure 2-15 C-V curve of n-type MOSFET transistor.

resistance of the substrate, and hence this configuration may not have equivalent unit capacitance of C_{ox} when using the MOSFET at a high frequency.

Next, as V_{gs} rises, the depletion region starts to form under the gate oxide. The concentration of the holes under the gate starts to reduce, and the MOSFET slowly enters into the weak inversion region. Under this condition, the equivalent MOSFET capacitance consists of the gate oxide capacitance in serial with the capacitance from the depletion layer in the substrate. The formation of these serial capacitances causes a reduction in the total equivalent gate to substrate capacitance for the MOSFET.

Finally, as V_{gs} is raised further and above MOSFET threshold voltage V_t, the MOSFET goes from the weak inversion region into the strong inversion region. The substrate under the gate attracts enough electrons to sustain a channel between the source/drain of MOSFET transistor. The equivalent unit MOSFET capacitance once again becomes C_{ox}. This is the best biasing condition for a MOSFET to act as a capacitor.

2.4 The Threshold Voltage

The first parameter of interest that characterizes the switching behavior of the MOSFET device is the threshold voltage, V_t.[2–5] This is defined as V_{gs} voltage at which a MOSFET starts to conduct between source and drain of the transistor. We can graph the relative conduction against the difference in gate-to-source voltage in terms of the source-to-drain current (I_{ds}) versus the different gate-to-source voltage (V_{gs}). These graphs for a fixed drain-source voltage, V_{ds}, are shown in Figure 2-16. It is also possible to make n-channel MOSFET transistors to conduct when the gate voltage is equal to the source voltage (i.e., $V_{gs} = 0$ V), whereas others require a positive difference between the gate and the source voltages to start conduction (negative for PMOS devices). Those devices that are normally non-conducting are classified as "enhancement-mode devices,"

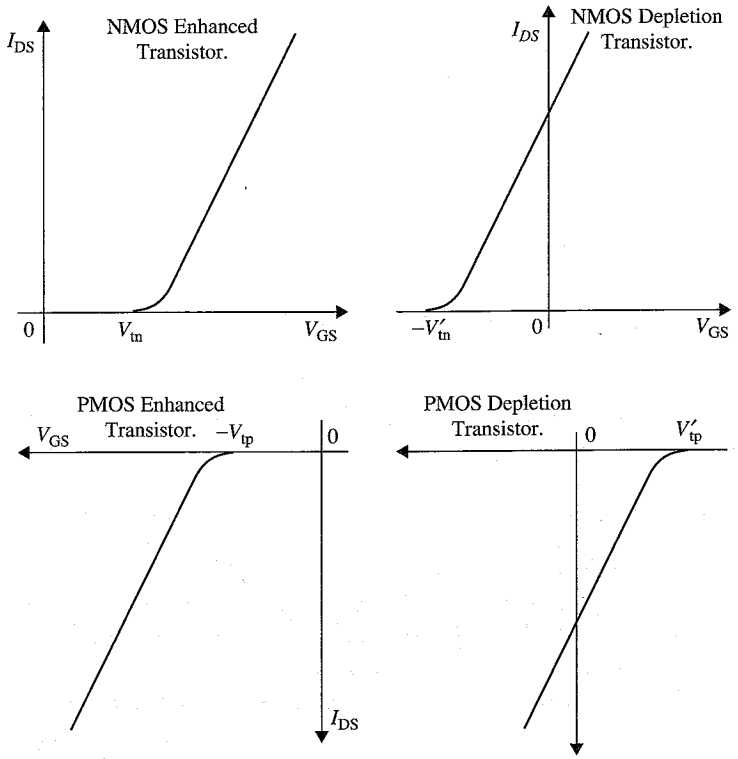


Figure 2-16 Forward bias, reverse bias characteristics of MOSFETs.

whereas those devices that can conduct even with 0 V gate bias are called "depletion-mode devices." The n-channel transistors and p-channel transistors are the dual opposites of each other—that is, the voltage polarities required for correct operation are opposite. The threshold voltages for n-channel and p-channel devices are denoted by V_{TN} and V_{TP}, respectively.

The threshold voltage of a MOSFET capacitor can also be realized as the gate voltage, V_{GB}, required to create and maintain a strong inversion region underneath the gate. Under this condition, the gate can maintain a large concentration of electrons under it and the electrons can flow from the source to the region in substrate underneath the gate and eventually to the drain. Thus, a channel of charge carriers is formed under the gate between source and drain regions when the MOSFET is turned on. During this time, we can also say that the MOSFET is inverted. In reality, the MOSFET turn-on phenomenon is a gradual function of the gate-to-source biasing voltage, making it difficult to determine V_t precisely.

As previously mentioned, the surface underneath the gate of n-type MOSFET is inverted from p-type to n-type when the applied V_{gs} is greater than the threshold voltage, V_t. Under this condition, there is also a depletion region that is formed between the channel and the substrate. The thickness of this depletion region can be expressed as

$$X_d = \sqrt{\frac{2\varepsilon_{si}\Phi}{qN_A}} \tag{2-7}$$

where N_A is the doping density of the uniform doped p-type substrate, ε_{si} is the permittivity of the silicon, and Φ is the potential in the depletion layer at the oxide-silicon interface. Φ can be expressed as $\Phi = |\Phi_S - \Phi_F|$.

Also, Φ_S is the potential, underneath the gate, at the oxide silicon interface, and Φ_F is the potential of the p-type substrate, also known as the Fermi level of the p-type substrate. Φ_F can be expressed as

$$\Phi_F = -\frac{kT}{q}\ln\left[\frac{N_A}{ni}\right] \tag{2-8}$$

where n_i is the intrinsic carrier concentration of silicon($\approx 1.45 \times 10^{16}$ atoms/m^3). The absence of holes in the depletion region leaves a net negative charge due to the immobile acceptor atoms that remain behind. This negative charge is equal to the charge attracted under the gate.

The charge per unit area in the depletion region can be expressed as $Q = qN_A X_d$. Using Equation 2-7, the charge can be expressed as follows:

$$Q = \sqrt{(2qN_A E_{si}\Phi)} = \sqrt{(2qN_A E_{si}|\Phi_S - \Phi_F|)} \tag{2-9}$$

If the surface potential, Φ_S, is equal to Φ_F, then the MOSFET is operating in the accumulation region. As V_{gs} is increased, the surface potential becomes more positive, and when $\Phi_S = 0$ V, the surface under the oxide has become depleted. As V_{gs} is increased further, such that $\Phi_S = -\Phi_F$, the surface near substrate is inverted and becomes n-type material—that is, the electron concentration at the semiconductor-oxide interface is equal to the substrate doping concentration. The value of V_{gs} at which $\Phi_S = -\Phi_F$ is called the threshold voltage of MOSFET. Hence, during the entire transition, the magnitude of change of Φ_S is $2\Phi_F$. In the presence of an inversion layer and without any substrate bias, the charge in the depletion region can be expressed as follows:

$$Q_{b0} = \sqrt{(2qN_A \varepsilon_{si}|2\Phi_F|)} \tag{2-10}$$

If a substrate bias voltage V_{sb} is applied between the source and the substrate, the V_{gs} voltage required to produce the inversion layer becomes $(2|\Phi_F| + V_{sb})$, and the charge stored in the depletion region can be expressed as follows:

$$Q_b = \sqrt{(2qN_A \varepsilon_{si}(2|\Phi_F| + V_{sb}))} \tag{2-11}$$

During the presence of the inversion layer, a parallel plate capacitance is formed between the gate and the inversion region, with the gate oxide acting as the insulator. This capacitor can be expressed as $C_{ox} = Q_b / V_{GI}$, where V_{GI} is the voltage difference between the gate and the inversion region.

Hence, the V_{gs} voltage required to produce and sustain the inversion layer, called the "threshold voltage" (V_t), can be calculated by adding the following components:

- The contact potential that exists between the gate metal and the silicon Φ_{ms}. In fact, Φ_{ms} can be expressed as $\Phi_{ms} = (\Phi_{gate} - \Phi_{oxide}) + (\Phi_{oxide} - \Phi_F) = (\Phi_{gate} - \Phi_F)$.
- The surface inversion potential ($-2\Phi_F$ is required to produce the inversion layer).
- The potential V_{GI} that exists between the gate and the inversion layer.
- The positive charge, Q_{ss}, that exists at the silicon-oxide interface due to imperfections from the growth of gate oxide or as a result of ion implantations used to adjust the threshold voltage of the MOSFET. This charge must be compensated by the gate-source voltage contribution of $-Q_{ss}/C_{ox}$.

Therefore, adding all these components, we can express the threshold voltage as follows:

$$\begin{aligned} V_t &= \Phi_{ms} - 2\Phi_F + \frac{Q_b}{C_{ox}} - \frac{Q_{ss}}{C_{ox}} \\ &= \Phi_{ms} - 2\Phi_F + \frac{Q_{b0} - Q_{ss}}{C_{ox}} - \frac{Q_{b0} - Q_b}{C_{ox}} \end{aligned} \tag{2-12}$$

Using Equation 2-11 and substituting into the preceding equation, we get

$$\begin{aligned} V_t &= \Phi_{ms} - 2\Phi_F + \frac{Q_{b0} - Q_{ss}}{C_{ox}} + \frac{\sqrt{2q\varepsilon_{si}N_A}}{C_{ox}}\left[\sqrt{|2\Phi_F| + V_{SB}} - \sqrt{|2\Phi_F|}\right] \\ &= \Phi_{ms} - 2\Phi_F + \frac{Q_{b0} - Q_{ss}}{C_{ox}} + \gamma\left[\sqrt{|2\Phi_F| + V_{SB}} - \sqrt{|2\Phi_F|}\right] \end{aligned} \tag{2-13}$$

where the parameter γ can be defined as the substrate-bias (or body-effect) coefficient.

$$\gamma = \frac{\sqrt{2qE_{si}N_A}}{C_{ox}}$$

When $V_{SB} = 0$ V (that is, when the source and substrate are at the same potential), we can simplify the preceding equation as follows:

$$V_{t0} = \Phi_{ms} - 2\Phi_F + \frac{Q_{b0} - Q_{ss}}{C_{ox}} \tag{2-14}$$

Next, by substituting Equation 2-15 into Equation 2-14, we can derive a simpler version for the threshold voltage:

$$V_t = V_{t0} + \gamma\left[\sqrt{|2\Phi_F| + V_{SB}} - \sqrt{|2\Phi_F|}\right] \tag{2-15}$$

In general, the value of V_{t0} is usually adjusted during fabrication by implanting additional impurities into the active area. As shown in Equation 2-15, the threshold voltage increases when the back bias, V_{SB}, is applied. A positive bias on the substrate results in a wider depletion region and assists in balancing the gate charge. This causes electron concentration in the inversion layer to decrease. Thus, a higher gate voltage is needed to achieve the onset of inversion resulting in an increase of the threshold voltage. In addition, the doping concentration and oxide thickness can also have impacts on the threshold voltage dependence on back bias. Lower doping concentrations and thinner oxide result in a weaker dependence of back bias on threshold voltage.

2.5 Metal-Oxide Field-Effect Transistor

The MOSFET shown in Figure 2-17 is an n-channel MOSFET, in which electrons can flow from source to drain in the channel induced under the gate oxide. Both n-channel and p-channel MOSFETs are extensively used. In fact, CMOS IC technology relies on the ability to use both devices on the same chip.

2.5.1 Device operation

To understand the MOS device operation, consider the setup shown in Figure 2-18. The nodes of an enhancement-type NMOS transistor have been connected through different voltage sources and ammeters to apply voltages and monitor the device current. Typically, $V_s = V_B$ and $V_D > V$. For simplicity, we can assume that the

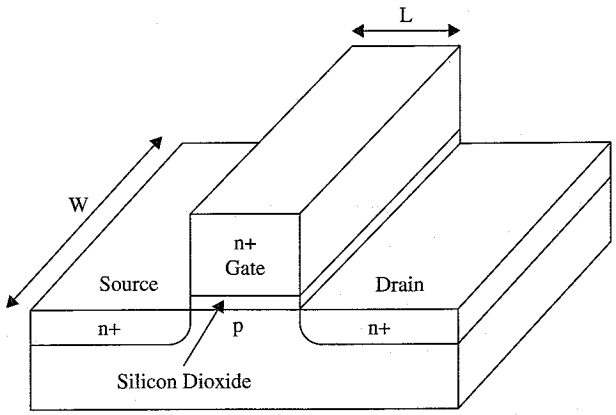


Figure 2-17 MOSFET structure.

source terminals are tied to the ground (i.e., $V_{SB} = 0$ V). These yield the following:

$$V_{GS} = V_{GB} - V_{SB} = V_G$$
$$V_{DS} = V_{DB} - V_{SB} = V_D$$

Cut-off region of MOSFET. When the source and bulk of the n-type MOSFET are biased at 0 V, with a small positive voltage is applied on the drain and 0 V is applied to the gate—i.e., $V_{DS} > 0$ and $V_{GS} > 0$—the drain is a reverse-biased p-n junction. Conduction band electrons in the source region encounter a potential barrier determined by the built-in potential of the source junction. Assuming this is enhancement-type MOSFET with $V_t > 0$, electrons cannot enter into the channel region and, hence, no current flows from the source to the drain. This is referred to as the cut-off state.

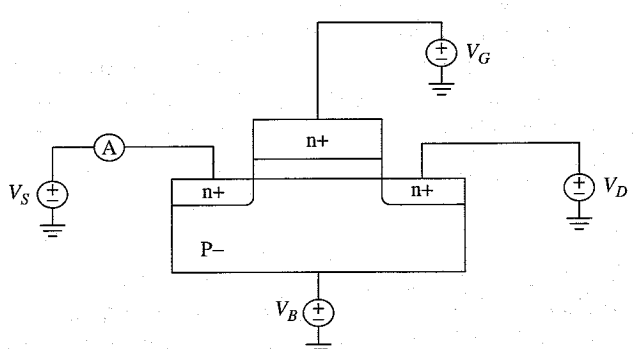


Figure 2-18 Experimental setup for studying MOSFET characterstics.

Linear region of MOSFET. As V_{GS} is increased higher than V_t, with band bending in the channel region ($\psi_s > 0$), it brings the conduction band in the channel region closer to the conduction band in the source region, reducing the height of the potential barrier of electrons. Electrons can now enter the channel, and the current flow from source to drain can be established.

In the low-drain bias regime, the drain-to-source current increases almost linearly with drain bias voltage. Indeed, here the channel resembles an ideal resistor obeying Ohm's law. The channel resistance is determined by the electron concentration in the channel, which is a function of the gate bias. Therefore, the channel acts like a voltage-controlled resistor whose resistance is determined by the applied gate bias. As the gate bias is increased, the slope of the I-V characteristic gradually increases due to the increasing conductivity of the channel. We obtain different slopes for different gate biases. This region where the channel behaves like a resistor is referred to as the linear region of MOSFET. The drain-to-source current in the linear regime is given by

$$I_{DS,lin} = \frac{W}{L}\mu C'_{ox}\left[(V_{GS} - V_T)V_{DS} - \frac{1}{2}V_{DS}^2\right] \qquad (2\text{-}16)$$

where V_t is the threshold voltage, C'_{ox} is the gate capacitance per unit area, and μ is the effective channel mobility, which differs from bulk mobility. We will deal with the concept of effective channel mobility later.

Threshold voltage in the preceding equation is defined as

$$V_t = V_{FB} + 2\Phi_F + \gamma\sqrt{2\Phi_F + V_{SB}} \qquad (2\text{-}17)$$

For very small V_{DS}, the second term in the parentheses can be ignored, and the expression for drain current can be simplified to

$$I_{DS,lin} = \frac{W}{L}\mu C'_{ox}(V_{GS} - V_t)V_{DS} \qquad (2\text{-}18)$$

which defines I-V curve as a straight line with a slope equal to the channel conductance.

$$\sigma_c = \frac{W}{L}\mu C'_{ox}(V_{GS} - V_t) \qquad (2\text{-}19)$$

The MOSFET bias conditions under different modes of operation is shown in Figure 2-19.

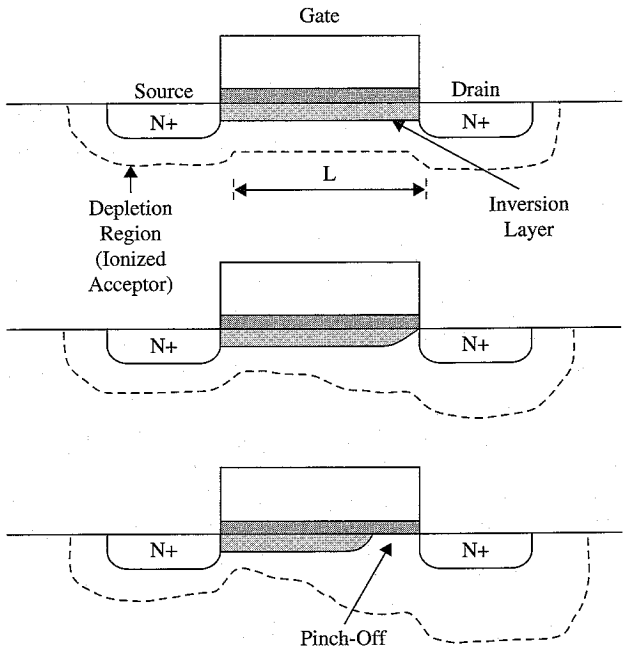


Figure 2-19 MOSFET in different bias conditions.

Saturation region of MOSFET. Next, let us consider what happens when V_{DS} is increased such that $V_{DS} \geq V_{GS} - V_t$. Now, because both the gate and the drain are positively biased, the potential difference across the oxide is smaller near the drain end. Again, because the positive charge on the gate is determined by the potential drop across the gate oxide, the gate charge is smaller near the drain end. This implies that the amount of negative charge in the semiconductor needed to preserve charge neutrality will also be smaller near the drain. Consequently, the electron concentration in the inversion layer near the drain end drops when $V_{DS} \geq V_{GS} - V_t$. This result is reasonable because we know that the gate-to-channel voltage at which there is no channel (before the onset of a weak inversion) is equal to V_t by definition of the threshold voltage. Therefore, at the point where the channel pinches off near the drain side, the channel voltage is $V_{GS} - V_t$, so the horizontal electric field across the channel, during pinch off, does not depend on V_{DS} but instead on the voltage across the channel, which is $V_{GS} - V_t$. Hence, Equation 2-16 is no longer valid when $V_{DS} \geq V_{GS} - V_t$. In this case, the drain current can be derived by substituting $V_{DS} = V_{GS} - V_t$ in Equation 2-16 as

$$I_{DS,lin} = \frac{W}{L}\mu C'_{ox}\left[(V_{GS} - V_t)V_{DS} - \frac{1}{2}V_{DS}^2\right] \qquad (2\text{-}20)$$

As V_{DS} is increased beyond $V_{DS,sat}$, the width of the pinch-off region increases. However, the voltage that drops across the inversion layer remains constant and equal to $V_{DS,sat}$. The portion of the drain bias in excess of $V_{DS,sat}$ appears across the pinch-off region. In a long channel MOSFET, the width of the pinch-off region is small relative to the total length of the channel and the variation can be ignored. Thus, neither the length nor the voltage across the inversion layer changes beyond pinch-off, resulting in a drain current independent of drain bias. Consequently, the drain current saturates. In smaller devices, this assumption falls apart and leads to a phenomenon called "channel length modulation." This can be explained by the fact that as V_{DS} is increased above $V_{GS} - V_t$, the effective channel length of MOSFET is not the physical drawn length "L" anymore because of the presence of a depletion region between the physical pinch-off point in the channel at the drain end and the drain region itself. The effective channel length can be expressed as

$$L_{eff} = L_{drawn} - X_d$$

where X_d is the depletion width. Hence, rewriting Equation 2-20 gives us the following:

$$I_{DS,sat} = \frac{W}{L_{eff}}\mu C'_{ox}\left(V_{GS} - V_t\right)^2 \qquad (2\text{-}21)$$

Now, because L_{eff}, or X_d, is a function of V_{DS}, it follows that $I_{DS,\,sat}$ in Equation 2-21 is actually a function of V_{DS}. In fact, taking channel length modulation into account, $I_{DS,\,sat}$ can be expressed as

$$I_{D,sat} = \frac{W}{L}\mu C'_{ox}\left(V_{GS} - V_t\right)^2(1 + \lambda V_{DS}) \qquad (2\text{-}22)$$

where λ is a channel length modulation parameter, whose typical values are between 0.05 V^{-1} and 0.005 V^{-1}.

From this discussion, it is also evident that the electron distribution is highest near the source and decreases near the drain. To keep a constant current throughout the channel, the electrons travel slower near the source and speed up near the drain. In fact, in the pinch-off region, the electron density is negligibly small. Therefore, in this region, in order to maintain the same current level, the electrons have to travel at much higher speeds to transport the same magnitude of charge.

Figure 2-20 shows I_{DS} vs. V_{DS} plot for an enhanced n-type MOSFET. As shown, at low V_{DS}, the drain current increases almost linearly with V_{DS}, resulting in a series of straight lines with slopes increasing with V_{GS}. At high V_{DS}, the drain current tends toward saturation and becomes almost independent of V_{DS}.

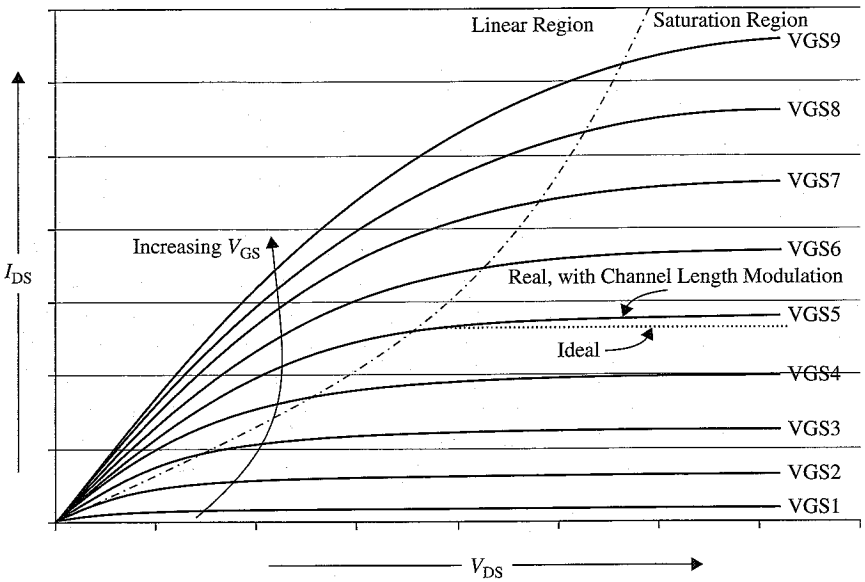


Figure 2-20 I_{DS} vs. V_{DS} for MOSFET.

2.5.2 Second-order effects for MOSFET operation

Our analysis of the MOSFET characteristics so far has entailed making several simplifying assumptions and analyzing an ideal device behavior. The operation of a real MOSFET may show some deviations from the model described which becomes especially acute as the dimensions of MOSFET reach the nano-meter region. It is essential to understand of some of these effects, also termed as second-order effects, and their impact on the device behavior for the design of charge pumps.[2–5] In this section we will describe two second order effects that will be essential in our subsequent circuit analysis.

Punch-through. When the drain of MOSFET is biased at voltage level that is high with respect to the source, the depletion region near the drain could extend to the depletion region near the source side and two of them merge together. It will cause current to flow irrespective of the gate voltage (i.e., even if V_{GS} is 0 V). This phenomenon is known as the punch-through effect. It is sensitive to the effective channel length of MOSFET device, and the applied voltage levels on source and drain of the devices. Since a high voltage charge pump is built based upon CMOS process, and voltage levels seen on internal nodes are much higher than chip supply voltage, extra attention is needed in dealing with any MOSFET transistors that are connected to the high voltage signals.

Impact ionization. As the length of MOSFET transistor is reduced, the electric field near the drain of the transistor will increase (for a fixed-drain voltage) due to saturation. For submicron gate lengths, the electric field near the drain can become so high that electrons are imparted with enough energy to become what is termed "hot." These hot electrons impact the drain, dislodging holes that are then swept toward the negatively charged substrate and appear as the substrate current. This effect is known as "impact ionization." Moreover, the electrons can penetrate the gate oxide, causing a gate current. Over the time this can lead to a degradation of the MOSFET device parameters (threshold voltage, subthreshold current, and transconductance), which in turn can lead to the failure of the circuits eventually. Although the substrate current may be used in a positive manner to estimate the severity of the hot electron effect, it can lead to poor refresh times in dynamic memories, noise in mixed signal systems, and possibly latch-up. Hot holes, on the other hand, do not normally present a problem because of their lower mobility.

2.6 Latch-up in CMOS Technology

The device structures present in the standard CMOS technology are inherently composed of parasitic PNP and NPN bipolar structures that can give rise to a combined SCR-type effect, known as "latch-up." Even though this problem occurs more frequently in regions near I/O interface, a circuit designer contemplating a charge pump design should be well aware of the dangerous latch-up effect when both NMOS- and PMOS-type MOSFETs are used in charge pump circuits handling high-voltage swings in every clock cycle.

Figure 2-21 shows the parasitic bipolar MOSFETs that are responsible for latch-up in CMOS technology. Figure 2-22 shows the schematic of the MOSFET parasitics. The emitter, base, and collector of the bipolar pnp, Q_1, is formed by the source of the PMOS, the n-well, and the substrate, respectively. Bipolar npn Q_2's collector, base, and emitter are formed by the n-well, substrate, and source of the NMOS transistor. The resistors RW_1 and RW_2 are the resistances of the n-well, and resistors RS_1 and RS_2 are the resistances of the substrate. Capacitors C_1 and C_2 are the parasitic capacitors that are present between the drain, source, and the substrate of the PMOS and NMOS transistors.

In a normal operating region, the bias voltages at nodes b_1 and b_2 are such that the parasitic bipolar devices are non-conducting. Because Q_1 is off, the potential at node b_2 is close to 0 V. Since Q_2 is off too, the potential at node b_1 is close to the supply voltage. Now consider what happens when a spurious positive voltage spike appears at node b_2 through the parasitic capacitor C_2. This will cause a high positive bias

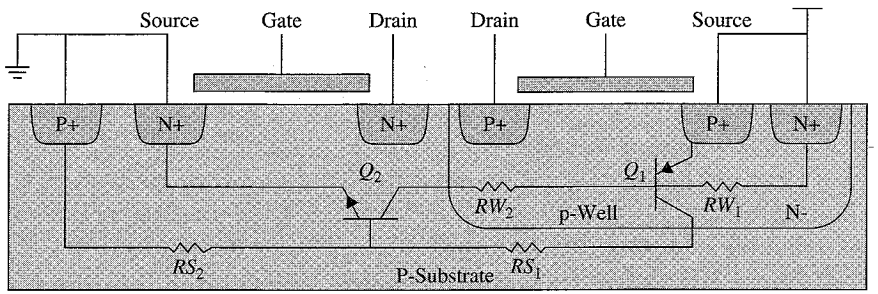


Figure 2-21 Cross-section view of an inverter showing the parasitic components.

at node b_2 and cause Q_2 to turn on. A current through Q_2 will reduce the voltage at node b_1. Now if the gain of the transistors and the input voltage spike amplitude are high enough while resistances RW_1 and RW_2 are low enough, a turned-on Q_2 will cause Q_1 to turn on. As a consequence, Q_1 will supply the necessary current to sustain the bias voltage at node b_2. Hence, this will give rise to a self-sustaining mechanism, resulting in a low-resistance path from the power supply to ground. A similar argument can be made for negative-going input pulses at node b_1 through capacitor C_1.

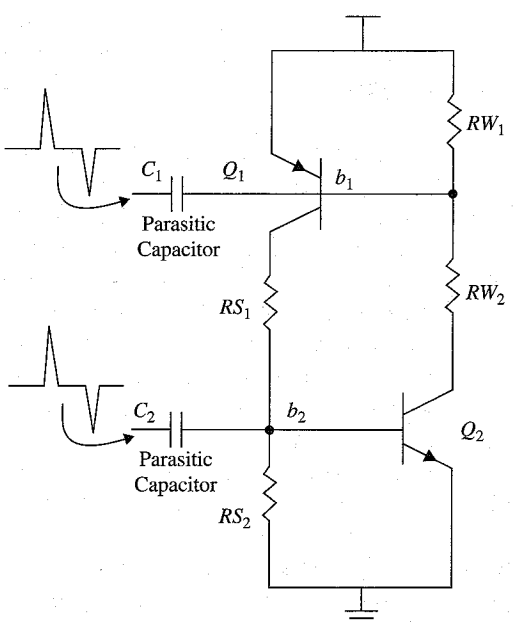


Figure 2-22 Schematic of the parasitic components related to latch-up.

The following common techniques can help to reduce the latch-up problems in real practice:

- **Reduce the parasitic resistances RW_1 and RW_2.** In general, if these resistances are zero, the parasitic PNP and NPN transistors will never turn on. The values of these resistances are strong functions of the distance between substrate near active area and well contacts, and the total number of the contacts. With plenty of these contacts placed close to the NMOS and PMOS transistors, the value of RW_1 and RW_2 can be kept at a minimum.
- **Reduce the noise amplitude at nodes b_1 and b_2 by reducing the size of capacitors C_1 and C_2.** Reducing the size of the NMOS and PMOS transistors will reduce the size of the parasitic capacitors C_1 and C_2, thus lowering the magnitude of the signal fed through.

In general, large MOSFETs, required to drive heavy loads, are especially susceptible to latch-up because of the large drain depletion capacitances. One of the best ways to design latch-up free circuits, especially for charge pumps, is to use only one type of MOSFET in the circuit.

2.7 Merits of PMOS Versus NMOS in Circuit Design

The modern CMOS process uses two types of MOSFETs: NMOS and PMOS. Early MOS IC developments focused on creating LSI chips that used either all NMOS or all PMOS devices. (CMOS, which employs both NMOS and PMOS transistors, came later.) PMOS IC devices were developed before NMOS ICs because they were easier to make. NMOS transistors were susceptible to manufacturing contaminations that caused them to malfunction. However, this contamination did not affect PMOS transistors in quite the same way because they use holes instead of electrons as majority carriers. But products built with PMOS transistors were inherently slow because they relied on lower-mobility holes (electron mobility is about two to three times higher than hole mobility) to carry current instead of electrons, which made the industry focus more on building products with NMOS devices.[6]

Yet, with the advent of technology and the established superiority of the CMOS process, both PMOS and NMOS devices are being used in modern-day circuits. However, analog circuits—and particularly the charge pumps—can be created entirely out of either PMOS or NMOS transistors. A detailed analysis will be done in Chapter 8 of this book. In general, besides NMOS being faster, for given dimensions and bias currents, NMOS transistors exhibit higher output resistance, which equates to higher gain in amplifiers. Further, to conduct the same

magnitude of current, an NMOS transistor can be smaller than the PMOS. But PMOS transistors have one formidable advantage: because modern silicon substrate are p-type, a separate n-well is created to build PMOS transistors, which allows for the independent control of a PMOS transistor's body bias without impacting the neighboring transistors. Hence, even though each MOSFET type has its own set of advantages and disadvantages, it depends on the application and the circuit designer to select the right type of devices for producing the most efficient operation.

2.8 The MOSFET Model

Almost all electronic circuit simulations performed today are done using the SPICE simulator or different variations of SPICE-based simulators. SPICE, an acronym for Simulation Program with Integrated Circuit Emphasis, is a general-purpose circuit simulation program for nonlinear DC, nonlinear transient, and linear AC analyses. Circuits may contain resistors, capacitors, inductors, mutual inductors, independent voltage and current sources, dependent sources, lossless and lossy transmission lines, switches, uniform distributed RC lines, and the five most common semiconductor devices: diodes, BJTs, JFETs, MESFETs, and MOSFETs. SPICE originated at the EECS Department of the University of California at Berkeley.[7,8] Now, to represent the behavior of transistors and other circuit elements in circuit simulations, SPICE requires an accurate model for each device. Over the last few decades, MOSFET modeling has evolved from basic empirical models to really sophisticated levels for accurately modeling many second-order effects. Throughout this chapter we saw the ideal equations that describe the behavior of MOS transistors. Although these incorporate some nonideal effects, they do not accurately model a specific device for a particular process. This is especially true for devices with ever-shrinking dimensions. In the original implementation of SPICE, as released by U.C. Berkeley, three levels of accuracy were provided for MOSFET models for SPICE simulations; these were denoted as Level 1, Level 2, Level 3. The Level 1 model, also known as the Schichman-Hodges model, uses basic device equations (square law) for I_{DS} to calculate current flow, device capacitances, and other parameters. However, the Level 1 model eliminates most of the MOSFET's second-order effects, notable among them the carrier velocity saturation effect. The Level 1 model estimates the basic I-V curves of long channel devices with reasonable accuracy, but it still predicts the output impedance of transistors in saturation quite poorly.

The Level 2 model was then developed to represent many higher-order effects. The Level 2 model, also known as the Grove-Frohman model,

of threshold voltage across the channel. During the derivation of the MOSFET I_{DS} equations, it was assumed that the threshold voltage is constant along the channel. This assumption is not correct because the charge in the depletion region under the channel varies according to the local voltage. The Level 2 model addresses this issue, along with other second-order effects, including the effect of velocity saturation. Silicon data indicate that the Level 2 model provides reasonable I-V accuracy for wide short channel devices in the saturation. But simulating with the Level 2 model is computationally intensive because of the existing 3/2 power equations. The Level 3 model was then developed around 1980 with some simplified equations and some empirical constants added to improve accuracy for short channel length devices. It is computationally more efficient, replacing the 3/2-power terms with a first-order Taylor expansion. The drain-induced barrier-lowering effect was also added. The Level 3 model is impressively physical, modeling two-dimensional effects based on junction depth and depletion depths.

The BSIM series models,[7,8] as the MOSFET models originating from U.C. Berkeley are named, were developed as simple physical models with many new "engineering" parameters added to simplify the equations and yet attain better accuracy. These models have enjoyed widespread adoption and have become a standard for defining CMOS processes. BSIM models have evolved, especially over the last decade, from BSIM1, and BSIM2, and BSIM3 to the latest BSIM4 models. In general, the BSIM models have incorporated the following important features:

- Vertical field dependence of carrier mobility
- Carrier velocity saturation
- Drain-induced barrier lowering
- Depletion charge sharing by source and drain
- Nonuniform doping profile for ion-implanted devices
- Channel length modulation
- Subthreshold conduction
- Geometric dependence of electrical parameters
- Gate tunneling model (BSIM4)
- Intrinsic gate resistance model (BSIM4)
- Substrate resistance model (BSIM4)
- Layout-dependent parasitic model (BSIM4)

It is interesting to note that as the BSIM model has evolved, it has also increased the number of device parameters accordingly, as shown in Figure 2-23. Whereas BSIM1 needs about 50 parameters, BSIM2

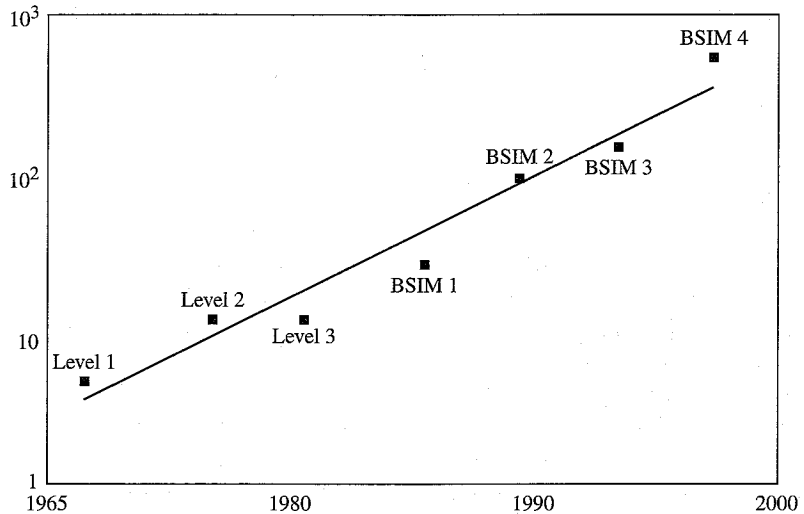


Figure 2-23 MOSFET model parameter variation.

The latest BSIM models strive to deal with frontier effects of the progressing technology, incorporating ultra short channel effects and providing reasonable accuracy for subthreshold and strong inversion operation.

2.9 SPICE Simulation Convergence

Normal SPICE-based simulators solve the MOSFET equations for DC and transient analyses through an iterative process.[9,10] However, for a number of reasons, the iterative process does not always converge on a solution, especially for purely analog circuits involving capacitors, as in the case of charge pump circuits, and this becomes a major source of frustration. In the following paragraphs we will discuss ways to mitigate this major issue.

Consider a typical convergence problem. SPICE makes an initial guess at a circuit's node voltages and, using the circuit conductances, calculates the mesh currents. Next, SPICE uses this current information to recalculate the node voltages and starts the cycle over again. The procedure is repeated until the entire node voltages settle to within certain tolerance limits, which can be altered with OPTIONS parameters, such as RELTOL, VNTOL, and ABSTOL.

Now, if the node voltages do not settle down to the tolerance limits within a certain number of iterations, the DC analysis results in an error message, such as "No convergence in DC analysis." SPICE then exits out of the simulation process because an initial stable operating point is required before running a DC or transient analysis.

When transient analysis is run, this iterative process repeats for each individual time step. If the node voltages do not settle down, SPICE reduces the time step and tries again to determine the node voltages. But, if the time step reduces beyond a certain fraction of the total analysis time, SPICE issues the error message "Time step too small" and quits the simulation.

In general, initial DC convergence problems occur because of incorrect initial voltage guesses, model discontinuities, unstable/bistable operating points, or unrealistic circuit impedances. Further model discontinuities or incorrect circuit connections are usually the causes of transient-analysis failure.

To solve the DC convergence problems, any of the following five steps may be used:

- Make sure all the circuit connections are valid. Check for floating nodes, dangling nodes, and incorrect connections.
- Increase the ABSTOL, RELTOL and VNTOL options.

  ABSTOL is the absolute current tolerance. Its default value is 1 pA. This means that when the node current reaches within 1 pA of its actual value, SPICE assumes that the current has converged and moves to the next step. VNTOL is the node voltage tolerance. Its default value is 1 μV. RELTOL is the relative tolerance parameter, and its default value is 0.1%. RELTOL and VNTOL act together during a simulation. Let us assume during the course of a simulation the actual value of a node voltage is 1 V. The RELTOL parameter will signify a convergence when the node voltage is within 0.1% of 1 V (i.e., within 1 mV of a volt), whereas the VNTOL parameter signifies a convergence when the node voltage is within 1 μV of a volt. SPICE will use the larger of the two numbers (in this case, the RELTOL parameter) to flag the convergence.

  By increasing the value of these parameters, the convergence problems may be avoided at the cost of simulation accuracy. To help circuit convergence, these parameter values may be reduced as follows:

  1. option ABSTOL=1e-6 VNTOL=1e-3 RELTOL=1e-2
  2. option ABSTOL=1e-3 VNTOL=5e-3 RELTOL=1e-1

- Use initial conditions by inserting a UIC keyword in the TRAN statement. For example, using ".tran 1n 100n UIC" causes SPICE to completely bypass the DC analysis. Add ".IC v(X)=" initial-condition statements to assist in the initial stages of the transient analysis.
- Ramp up the power supplies. In general, using a DC voltage source for the power supply may cause convergence problems. Instead, ramp up the supply voltage(s) slowly to get around these problems.

For example, the vvdd vdd 0 pwl 0n 0v 1000n 2.5v command will ramp up the input V_{DD} power supply voltage slowly from 0 V to 2.5 V over 1000n seconds.

Further, convergence during DC sweep can also be helped by avoiding absolute power supply boundaries. Hence, sweeping a voltage source from 0 V to 3 V may prove troublesome, but sweeping it from 0.1 V to 2.95 V may work.

- Reduce the rise/fall time of the input stimuli. During transient analysis, if the circuit uses external input signals, pay attention to the rise/fall times of these input stimuli. Applying signals with sharp rise/fall times may flag "Time step too small" errors. A simple solution is to experiment with longer rise/fall times for the input signals.

2.10 Conclusion

The main purpose of this chapter was to quickly review the basic p-n junction diode, MOSFET capacitor, and MOSFET characteristics and operation to allow us to gain a better perspective into the operations of the charge pump circuits in the coming chapters. This chapter also touched on MOSFET SPICE models and convergence issues during SPICE simulation, which will always be present during charge pump circuit simulations.

References

1. Pelliconi, R., I. David, B. Andrea, P. Marco, and L.R. Pier. "Power Efficient Charge Pump in Deep Submicron Standard CMOS Technology." *Solid-State Circuits Conference*, 2001. ESSCIRC 2001. *Proceedings of the 27th European.*
2. Razavi, B. *Design of Analog CMOS Integrated Circuits.* McGraw-Hill, New York, 2001.
3. Weste, H.E.N. and K. Eshraghian. *Principles of CMOS VLSI Design: A Systems Perspective*, Second Edition. Addison-Wesley, 1994.
4. Baker, J.R., W.H. Li, and E.D. Boyce. *CMOS Circuit Design, Layout, and Simulation.* IEEE Press, New York, 1998.
5. Gray, R.P., J.P. Hurst, H.S. Lewis, and G.R. Meyer. *Analysis and Design of Analog Integrated Circuits*, Fourth Edition. John Wiley & Sons, 2001.
6. Encore desktop calculator at HP. http://www.hp9825.com/html/prologues.html.
7. BSIM SPICE-related documents. <http://www-device.eecs.berkeley.edu/~bsim3/bsim4_intro.html>.
8. "Engineering BSIM for the Nano-Technology Era and Beyond." *Fifth International Conference on Modeling and Simulation of Microsystems.*
9. Quarles, T.L. "Analysis of Reference and Convergence Issues for Circuit Simulation." University of California, Berkeley. ERL Memo M89/42.
10. Hymowitz, Charles. "Step-by-step procedures help you solve Spice convergence problems." EDN design feature. March 3, 1994.

Chapter

3

Basic Operation of a Charge Pump

3.1 Charge Pump System

Charge pumps have been used to generate high voltages for many applications, such as EEPROMs and Flash memories for programming and erasing of the floating-gate. In general, a charge pump is a closed-loop system because the pump output is usually regulated at a predetermined level. To achieve the regulation, the pump needs to be kept turned on, as long as the output voltage is lower than the regulation level. If the pump output level reaches or exceeds the regulation level, the extra charge from pump output must be shunted away or the pump must be completely shut off. Since the system is relying on feedback control to maintain the regulation, obviously there is output noise near regulation level.

A basic block diagram of the charge pump's operation is shown in Figure 3-1. Assuming the pump output voltage is starting from zero at the beginning of operation, the output of the regulator is high (meaning that the pump output voltage is lower than the regulation level), Feedback signal from the regulator would enable pumping clocks to be applied to the charge pump. Charge is transferred from stage to stage and the potential energy of the charge is elevated as it propagates through stages. The output of the charge pump will start to ramp up once the charge reaches the output. During this time the regulator is enabled to constantly sample the output voltage and compare it with the reference voltage. When the output reaches the regulation level, the regulator senses it and deactivates the pumping clock. The charge pump is shut off.

If the output load is purely capacitive in nature, the pump output voltage should remain unchanged for a long period of time and in principle the pump does not need to be turned on again once regulation level

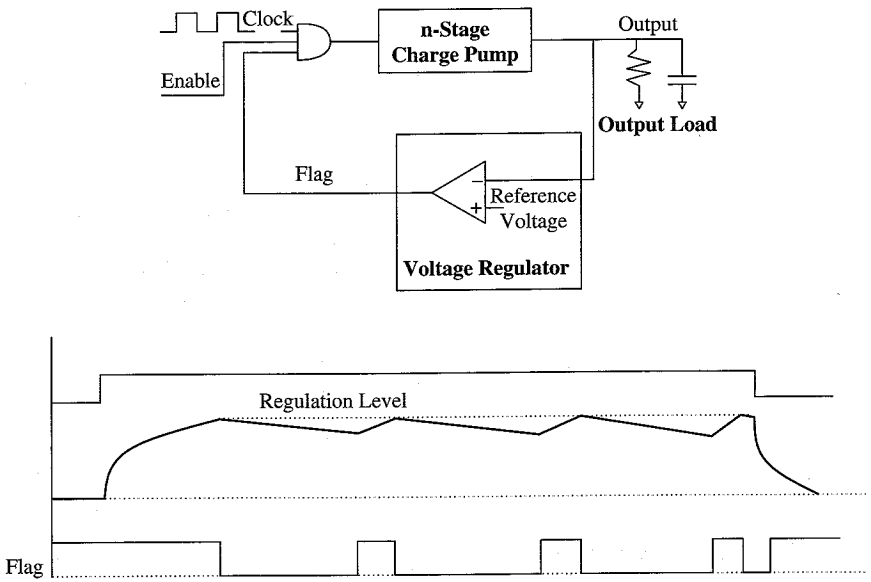


Figure 3-1 A charge pump system.

is reached. In reality, though, this scenario seldom happens. Capacitive coupling from nearby signals, the DC current attributing from different output leakage components, DC current dissipation in the regulator, and myriad other components will obviously cause the output of the charge pump to be discharged over time. Depending on the pump's feedback response time and other implementations, the regulator will eventually detect the deviation of output voltage from the regulation level and enable the charge pump. Since the output is closer to the original preset level, the pump should be able to replenish the lost charges faster and reach regulation level only in a short amount of time. As shown in Figure 3-1, this process will introduce output voltage overshoots and undershoots near regulation level. If the amplitude of noise is large, it may be detrimental to the functionality and device-related reliability of the circuit. Many techniques have been proposed to reduce the output noise for different applications, such as operation-dependent power balancing between pump and output load scheme, and the voltage-controlled oscillator (VCO) approach.

The operation-dependent power balancing between pump and output load scheme relies on matching pump output strength with the loading circuit power consumption at all time. The charge pump may have different output power specifications in many different operations. Even within the same operation the loading circuit may have different current loads over time. Noise occurs if there is mismatch between the

delivered power from the pump and the consumed power from the load. If pump strength can be adjusted based on operation, then the output noise can be minimized. Pump strength can be adjusted as required by varying clock frequency, pump clock amplitude, sizing of effective boosting capacitance, etc., the second approach uses a voltage-controlled oscillator instead of a steady clock generator. The clock frequency of the VCO will be high when the pump is initially starting up. Then this clock frequency will slowly decrease in an almost linear fashion as the output voltage approaches the regulation level and will almost stop when the value is reached. The latter approach is more complex to implement and not suitable for all applications.

At the end of the high-voltage operations, it is required to shut off the charge pump, discharge the high-voltage potential at the output and all internal nodes, and bring them down to normal chip supply levels. This shut-off and discharging process should be carefully managed. If the pump circuit, including the regulator, is shut off before discharging, it may create unwarranted high-voltage stress on some circuits. Further, discharging the high voltage to internal chip supplies may create high-ground bounce or high-V_{cc} bounce without careful timing controls. The noise could create problems for other circuits still in operation. In general, it is better to discharge the internal high-voltage nodes during a pre-defined period for high-voltage recovery purposes only.

Most of the techniques and processes discussed so far will be explained in detail in the following chapters. Throughout this chapter, we will focus on understanding the basics of charge pump operation. We will start with the bucket capacitor model and then discuss the operation of the Dickson charge pump and the different second-order effects, using equations to reinforce the concepts.

3.2 Basic Concept: The Bucket Capacitor Model

Charge pumps are circuits that generate a voltage level higher than the chip supply voltage from which they operate. To see how this is possible, consider the simple circuit shown in Figure 3-2, which consists of a single capacitor and three switches.[1]

During clock phase, ϕ, switches S_1 and S_3 are closed and the capacitor is charged to the supply voltage, V_{DD}. Next, in the second clock phase, ϕ_b, switches S_1 and S_3 are opened and switch S_2 is closed. The bottom plate of the capacitor assumes a potential of V_{DD}, while the capacitor maintains its charge of $Q = C \times V_{DD}$ from the previous phase. This means the following during ϕ_b:

$$Q = C(V_{out} - V_{DD}) = C \times V_{DD} \text{ i.e. } V_{out} = 2 \times V_{DD} \quad (3\text{-}1)$$

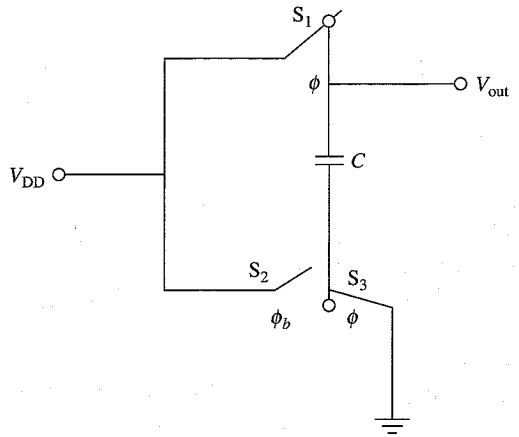


Figure 3-2 Simple voltage doubler.

Thus, in the absence of a DC load, an output voltage can be generated that is twice the input supply voltage. In order to accommodate a load at the output, the circuit should be modified by adding an output capacitance, as shown in Figure 3-3.

In this case, the ideal output voltage is given by

$$V_{out} = \frac{C}{C + C_{out}} \times 2 \times V_{DD} \tag{3-2}$$

Needless to say, the presence of the output capacitor C_{out} will reduce the output voltage, depending on the output load capacitor. If a load, R_L, is present, a ripple voltage, V_R, will also be generated at the output.

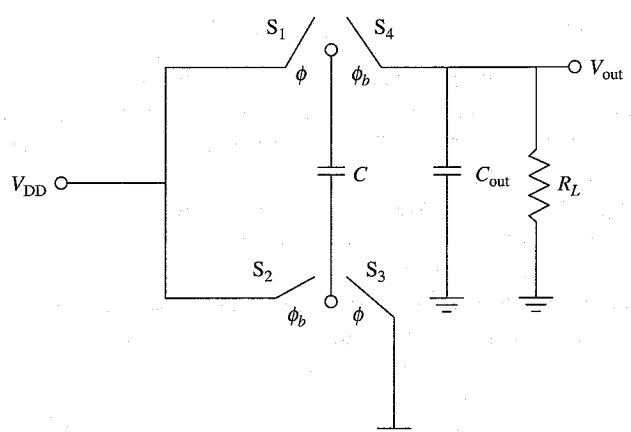


Figure 3-3 Voltage doubler with a load.

The ripple voltage can be reduced by making C_{out} sufficiently large so that V_R is negligible compared to V_{out}, but in doing so the output voltage will also be reduced. It is important to note that the switches mentioned in this circuit can actually be implemented using n-type MOSFETs or p-type MOSFETs, and the capacitor can be a standard metal-metal capacitance or realized using the gate oxide of the MOSFET, with its source-drain substrate connected together to form one end and the gate forming the other end. It is also important to pay considerable attention to the type of MOSFETs used in circuits involving high voltages. We will discuss this point later. Next, we will see a practical method of generating high voltage using the Dickson charge pump.

3.3 The Dickson Charge Pump

The operation of the Dickson pump circuit[2,3] is illustrated in Figure 3-4, which shows the typical voltage waveforms in an n-stage multiplier. As you can seen, ϕ and ϕ_b are two out-of-phase clocks with amplitude V_ϕ and are capacitively coupled to alternate nodes along the diode chain. The two clocks increase the potential voltage of charge at subsequent nodes by pumping packets of charge along the diode chain as the coupling capacitors are successively charged and discharged during each half of the clock cycle.

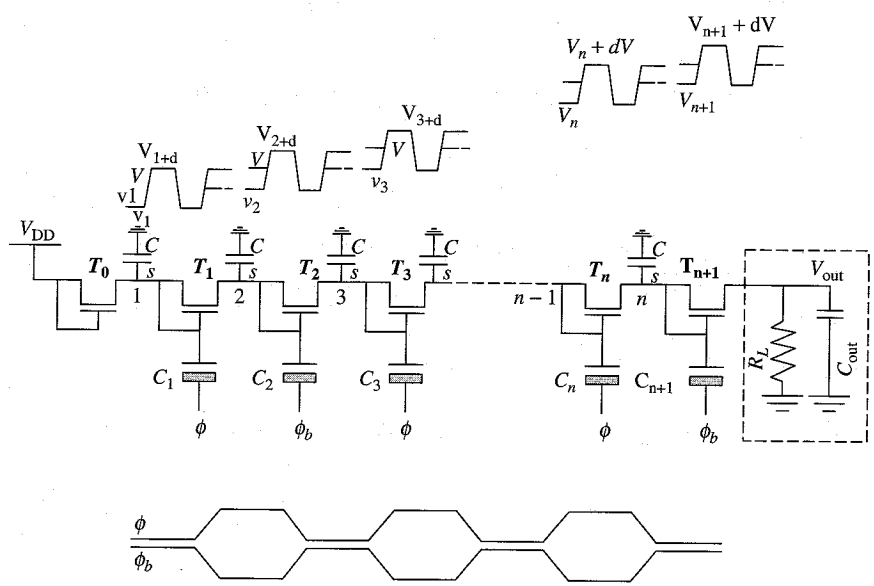


Figure 3-4 Internal waveforms of a 2-phase charge pump.

As shown in Figure 3-4, the difference between the voltages of the nth and (n + l)th nodes at the end of each pumping cycle is given by

$$\Delta V = V_{n+1} - V_n = V'_\phi - V_D \tag{3-3}$$

where V'_ϕ is the voltage swing at each node due to capacitive coupling from the clock, V_D is the forward bias diode voltage, and V_L is the voltage by which the capacitors are discharged down when the multiplier is supplying an output current (I_{out}).[4] For the clock coupling capacitance (C) and stray capacitance (C_S) at each node, the voltage gain can be represented as

$$V'_\phi = \left(\frac{C}{C+C_s}\right)V_\phi \tag{3-4}$$

When the clock ϕ is low and ϕ_b is high, MOSFET T_0 conducts while MOSFET T_1 is off. When T_0 is on, the voltage at node 1 settles at $V_{DD} - V_D$. Next, when the clock ϕ goes high, the voltage at node 1 becomes

$$V_1 = V_{DD} + (V'_\phi - V_D) \tag{3-5}$$

During the time ϕ_b is low and assuming the clock period is sufficiently long, MOSFET T_1 will conduct until the voltage at node 2 charges up to

$$V_2 = V_{DD} + (V'_\phi - V_D) - V_D \tag{3-6}$$

Next, when the clock ϕ goes low and ϕ_b is high, the voltage at node 2 becomes

$$V'_2 = V_{DD} + 2(V'_\phi - V_D) \tag{3-7}$$

Hence, for N stages, the voltage at node "n" will be $V_n = V_{DD} + N^*(V'_\phi - V_D)$. MOSFET T_{n+1} forms an isolating diode, forcing the output voltage to be

$$V_{out} = V_{DD} + N[(V'_\phi - V_D] - V_D \tag{3-8}$$

Equating the value of V'_ϕ from Equation 3-2 we get the following:

$$V_{out} = V_{DD} + N\left[\left(\frac{C}{C+C_s}\right)V_\phi - V_D\right] - V_D \tag{3-9}$$

This equation shows the output voltage in an ideal situation when the pump is not delivering any output load current. Because the pump will

be connected to an output load, which could be drawing a load current, the output voltage will not remain at the voltage expressed in Equation 3-9. Assuming V_L as the voltage drop per stage for supplying the average load current, the charge pumped by each diode per clock cycle is $(C + C_s)V_L$. The current supplied by the pump, at a clock frequency f, is given by

$$I_{out} = f(C+C_s)V_L \tag{3-10}$$

Rewriting Equation 3-10, the output voltage for N stages is reduced by an amount

$$\frac{N \times I_{out}}{(C+C_s)f}$$

Rewriting Equation 3-9 by incorporating the effects of the load current, the output voltage becomes

$$V_{out} = V_{DD} + N\left[\left(\frac{C}{C+C_s}\right)V_\phi - V_D - \frac{I_{out}}{(C+C_s)f}\right] - V_D \tag{3-11}$$

where V_{out} is the maximum peak output voltage at the last stage. Next, when ϕ_b goes low, MOSFET T_{n+1} turns off and the output load resistance, R_L, will discharge the load capacitor. Then, when ϕ_b goes high, MOSFET T_{n+1} will conduct and the output will charge up. This will give rise to a ripple at the output, defined as V_R. Because C_{out} is sufficiently large, V_R will be small compared to V_{out}. V_R can be expressed as

$$V_R = \frac{I_{out}}{f \times C_{out}} = \frac{V_{out}}{f \times R_L \times C_{out}} \tag{3-12}$$

The output ripple voltage is the noise to the loading circuits. Large amplitude of ripple could affect the operation of the chip, such as PLL, EEPROM, or flash memory applications. As can be observed from Equation 3-12, the ripple voltage can be reduced by increasing the clock frequency f or by increasing the output load capacitance. Each of these methods has its own limitations and drawbacks. Frequency cannot be blindly increased without decreasing the pump's efficiency after a certain limit. Also, increasing the load capacitance will reduce the output ramp-up time adversely.

From Equation 3-11, it can be seen that the charge pump will work provided that

$$N\left[\left(\frac{C}{C+C_s}\right)V_\phi - V_D - \frac{I_{out}}{(C+C_s)f}\right] - V_D > 0$$

which simplifies as follows:

$$\left[\left(\frac{C}{C+C_s}\right)V_\phi - V_D - \frac{I_{out}}{(C+C_s)f}\right] > 0 \tag{3-13}$$

Now Equation 3-11 can be rewritten as

$$V_{out} = V_o - I_{out}R_S \tag{3-14}$$

where

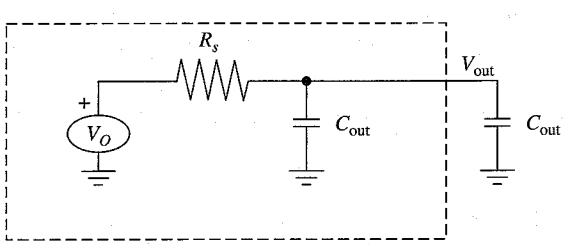
$$V_o = V_{DD} + N\left[\left(\frac{C}{C+C_s}\right)V_\phi - V_D\right] - V_D$$

and

$$R_S = \frac{N}{(C+C_s)f} \tag{3-14a}$$

Here, V_o and R_S are the open-circuit output voltage and output series resistance of the charge pump, respectively.[8] Therefore, Equation 3-14 can be expressed as a simple equivalent circuit of the charge pump, as shown in Figure 3-5.

In deriving this model for the charge pump, it has so far been assumed that the capacitors are fully charged and discharged with the diode connected MOSFET cutoff voltage, V_D. In practice, this is not the case due to the nonlinear voltage-current characteristics and internal series resistance, R_D, of the MOSFETs. This results in a residual voltage in addition to V_D remaining across the diodes at the end of each cycle, causing the multiplier output series resistance, R_S, to increase in a nonlinear manner with load current. However, by making R_D sufficiently small, the increase of R_S due to this effect can be minimized. Proper care should be taken to size the MOSFETs by keeping minimum permissible channel length and allowing sufficient MOSFET width.



3.4 Dynamic Analysis of the Charge Pump

Because the internal nodes of the charge pump are always in a state of flux, we need to make some simple assumptions to quantify its dynamic operation.[5] Assume that the charge pump has slowly risen to the maximum voltage and is maintaining the output voltage constant. Under this scenario, the charge transferred from one capacitor to the next equals the charge transferred to the output. We also need to make the following assumptions:

- The parasitic capacitance at each stage is negligibly small.
- The clock cycle time is sufficiently large, compared to all the RC time constant in the circuit.

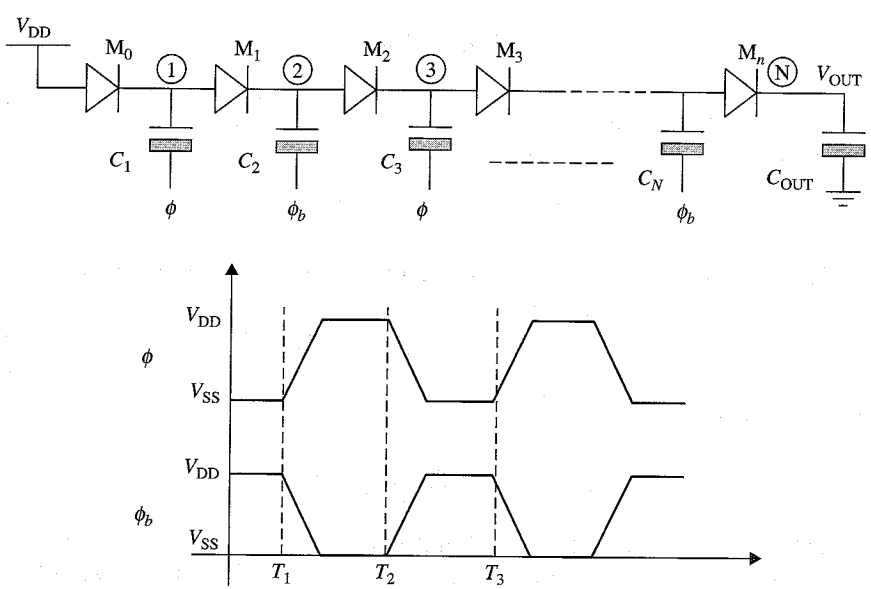
Next, consider the simplified circuit shown in Figure 3-6.

Assuming a complete charge transfer, the charge stored at node 1 at time T_1 is

$$Q_1 = C(V_{DD} - V_t) \tag{3-15}$$

Now the positive edge of the clock ϕ injects a charge at node 1, which causes M_1 to conduct and transfer charges to node 2. At the end of time T_2, assuming the charge transfer is complete, the potential at node 2 will rise to

$$V_2 = (Q_2 + q_{inj})/C \tag{3-16}$$



where Q_2 equals the existing charge at node 2, and q_{inj} is the part of the charge transferred to node 2, after being injected by the clock ϕ at node 1. However,

$$V_2 = V_1 - V_t \tag{3-17}$$

where V_1 is the new potential at node 1 at time T_2, or

$$V_1 = V_{DD} + (Q_1 - q_{inj})/C \tag{3-18}$$

Hence, substituting Equation 3-18 and Equation 3-16 into Equation 3-17, we get

$$V_{cc} + \left(Q_1 - \frac{(Q_1 - q_{inj})}{C}\right) - V_t = \frac{(Q_2 + q_{inj})}{C}$$

or

$$2Q_1 = Q_2 + 2q_{inj}$$

or

$$Q_2 = 2C(V_{DD} - V_t) - 2q_{inj} \tag{3-19}$$

Again, at time T_1, the potential at node 3 can be expressed as

$$\frac{Q_3}{C} = V_{DD} + \frac{Q_2}{C} - V_t$$

Substituting Equation 3-19 in the preceding equation, we get the following:

$$Q_3 = 3C(V_{DD} - V_t) - 2q_{inj}$$

The preceding sequence of charges/stages can be expressed as

$$Q(2n-1) = (2n-1)C(V_{DD} - V_t) - 2(n-1)q_{inj} \tag{3-20}$$

when n is even, and as

$$Q(2n) = 2nC(V_{DD} - V_t) - 2nq_{inj} \tag{3-21}$$

when n is odd. Here, $1 < n < N/2$. Note that the first stage can be defined by $n = 0$, but it is not in the domain of the preceding equations.

Further, $n = 1$ means the physical second stage, and $n = 2$ means the physical third stage, and so on.

Now the voltage equation of the last stage can also be represented as

$$V_{out} = V_{DD} + \frac{Q(N)}{C} - V_t$$

or

$$Q(N) = C(V_{out} - V_{DD} + V_t) \tag{3-22}$$

Assuming an even number of stages, Equation 3-21 can be expressed as

$$q_{inj} = \frac{Q(2n)}{2n} - C(V_{DD} - V_t)$$

Incorporating Equation 3-22 in Equation 3-21, the final output charge at stage N can be expressed as

$$q_{inj} = \frac{C}{N}[(N+1)(V_{DD} - V_t) - V_{out}] \tag{3-23}$$

Equations 3-20 and 3-21 can also be expressed in terms of V_{out} and q_{inj}. Substituting Equation 3-23 in Equation 3-21, we get the following:

$$Q(2n) = 2kC(V_{DD} - V_t) - \frac{2kC}{N}(N+1)(V_d - V_t) + \frac{2kC}{N}V_{out}$$

$$Q(2n) = \frac{2kC}{N}(V_{out} - V_{DD} - V_t)$$

Similarly, substituting Equation 3-23 in Equation 3-20, we get

$$Q(2n-1) = \frac{2(k-1)}{N}C(V_{out} - V_{DD} - V_t) + C(V_{DD} - V_t)$$

As shown in Figure 3-7, the last stage holds the maximum amount of charge. Further, the charge stored in each stage does not increase linearly. This can be quickly explained by the basic operation of the Dickson pump.

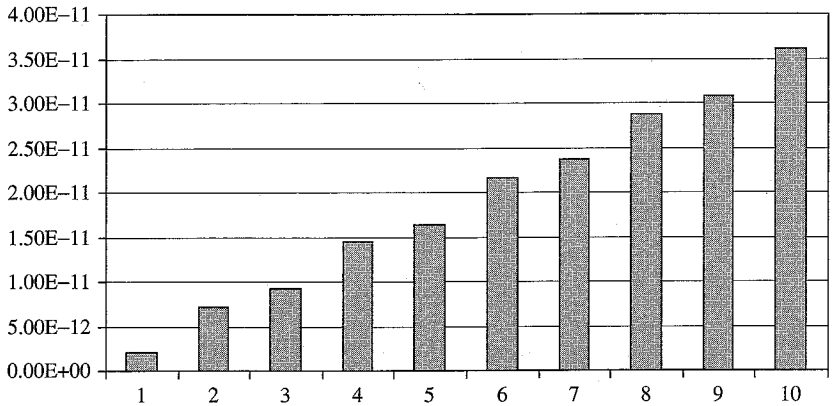


Figure 3-7 Charge stored at each internal stage.

The charge supplied by the power supply, every cycle, in a steady state is equal to the charge transferred to the capacitor, C_1, through the diode M_0, plus the charge transferred by the clocks ϕ and ϕ_b through capacitors C_1 to C_N from one stage to the next.

Solving the equations for the total charge delivered from the power supply and the charge delivered at the output, it can be shown that

$$Q_{VDD}(t)=(N+1)\times C_{load}\times[V_{out}(t)-V_{DD}-V_t] \tag{3-24}$$

where $C_{load} = C_{out} + C_{pump}$, $C_{pump} \sim NC/3$ (assuming $N > 4$), and C_{pump} is the inherent capacitance of the charge pump.

Finally, Tanzawa and Tanaka[5] have shown that the rise time, T_r, for the output voltage to charge up from initial state $(V_{DD} - V_t)$ to a final voltage, V_{fin}, can be expressed as

$$T_r=\frac{\ln\left[1-\dfrac{V_{fin}-(V_{DD}-V_t)}{N(V_{DD}-V_t)}\right]}{\ln\left[\dfrac{1}{1+\dfrac{C}{N(C_{load})}}\right]}T_{osc} \tag{3-25}$$

Thus, the average output current during time T_r can be expressed as

$$I_{out}=\frac{C_{load}[V_{fin}-(V_{DD}-V_t)]}{T_r} \tag{3-26}$$

and the average power supply current consumption during T_r can be derived from Equation 3-24 as

$$I_{VDD}=\frac{Q_{VDD}(t)}{T_r}=\frac{(N+1)C_{load}(V_{fin}-V_{DD}-V_t)}{T_r} \tag{3-27}$$

where $C_{load} = C_{out} + C_{pump}$.

For charge pumps requiring high-output voltages, the number of stages is essentially high ($N > 8$, or if $N \approx 8$, the charge pump may be operating with a boosted clock scheme, such as $2–4V_{DD}$).

Hence, to ensure good output current, $N\,(V_{DD} - V_t) >> V_{out}$. Further, the output load $C_{load} >> C$.

Equation 3-25 can also be expressed as

$$T_r=\frac{\ln\left[1-\dfrac{V_{fin}-(V_{DD}-V_t)}{N(V_{DD}-V_t)}\right]}{\beta} \tag{3-27a}$$

where

$$\beta=\frac{1}{T_{osc}}\ln\left[\frac{1}{1+\dfrac{C}{N(C_{load})}}\right]$$

is a constant.

Now, a continuous function of the form $\ln(x)$ can be expanded easily in terms of a Taylor series. In the current case, even though the output voltage, V_{out}, will be a staircase function as it reaches its destination voltage, V_{fin}, it can be assumed to be a smooth function because the rise time is very large compared to the clock cycle and the output load capacitance is generally very high compared to the pump capacitor. Hence, simplifying Equation 3-27a using a Taylor series expansion for logarithmic functions, Papaix and Daga[6] have simplified the rise time as follows:

$$T_r=\frac{V_{fin}-(V_{DD}-V_t)}{(V_{DD}-V_t)}\left(\frac{1}{3}+\frac{C_{load}}{NC}\right)T_{osc} \tag{3-28}$$

The power efficiency can be calculated like this:

$$\xi=\frac{V_{out}I_{out}}{V_{DD}I_{vdd}}=\frac{V_{out}C_{load}[V_{fin}-(V_{DD}-V_t)]}{V_{DD}(N+1)C_{load}(V_{fin}-V_{DD}-V_t)}=\frac{V_{out}}{(N+1)V_{DD}} \tag{3-29}$$

This shows that the number of stages only determines the power efficiency when V_{out} and V_{DD} are fixed. Further, the rise time, T_r, is proportional to C_{load}, f_{osc}, $1/N$, 1/VDD, and $1/C$. In general, the W/L dimensions of the diodes are not parameters in these equations because the equations were derived assuming a complete charge transfer, every clock cycle. However, in reality, the W/L ratio of the diodes cannot always be very high, because this will increase the layout area, and the frequency of operation is not always tuned for complete charge transfer. Hence, this incomplete charge transfer scenario will cause an anomaly in the rise time between the expected result from Equation 3-28 and the actual simulation results. The exact derivation of the charge pump's output characteristics during incomplete charge transfer is beyond the scope of this book. For initial study purposes, the reader is requested to design and simulate the charge pump with a relaxed clock period to ensure complete charge transfer. After that he/she can reduce the clock period and study the new effects.

The preceding derivations allow us to synthesize an equivalent model of the charge pump, as shown in Figure 3-8.

From Equation 3-24, and referring to Figure 3-8 ,V_{max} equals $(N + 1)$ $(V_{DD} - V_t)$, R_{eqv} equals N/C, and C_{eqv} equals $NC/3$ (assuming $N > 4$).

3.4.1 The body effect revisited

In the previous chapter we saw that the threshold voltage of an NMOS transistor can be represented as

$$V_t = V_{t0} + \gamma(\sqrt{\phi_s + V_{sb}} - \sqrt{\phi_s}) \tag{3-30}$$

where

- ϕ_s equals the surface potential at threshold and is represented by $\phi_s = 2V_t \ln\frac{N_A}{n_i}$.

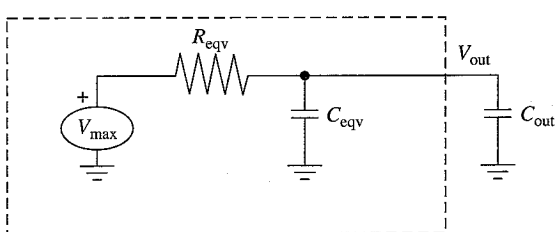


Figure 3-8 Equivalent model of the Dickson pump.

- γ equals the body effect coefficient and is represented by $\gamma = \frac{t_{ox}}{\varepsilon_{ox}}\sqrt{2q\varepsilon_{si}N_A} = \frac{\sqrt{2q\varepsilon_{si}N_A}}{C_{ox}}$.
- V_{t0} equals the zero-bias threshold voltage.
- V_{sb} is the source-to-body voltage bias.

As you can see from the preceding equation, for a particular process corner, V_{t0}, γ and ϕ_s are constants. Hence, as the source voltage of an NMOS MOSFET increases, the threshold voltage of the MOSFET also rises, which results in decreased I_{ds}, so less charge transfer takes place.[4,7]

3.5 Dynamic Analysis of the Charge Pump with Body Effect

The derivation for Equation 3-25 assumes a complete charge transfer and is implemented with diodes. In CMOS technology, because diodes are generally implemented with MOSFETs, as shown in Figure 3-9, with the drain and gate tied together, implementing a Dickson charge pump with a series of NMOS transistors will introduce body effect—the V_t of the MOSFETs will increase as the source voltage increases. Hence, Equation 3-25 needs to be readdressed, incorporating the effect of V_t variation.

In general, for an NMOS with a zero back bias, when the drain and gate are tied together, the source voltage can be expressed as

$$V_s = V_D - V_{t0}$$

Taking body effect into account, this expression can be modified as $V_s = \alpha\,(V_D - V_{t0})$, where α is the body effect factor.[3] Therefore,

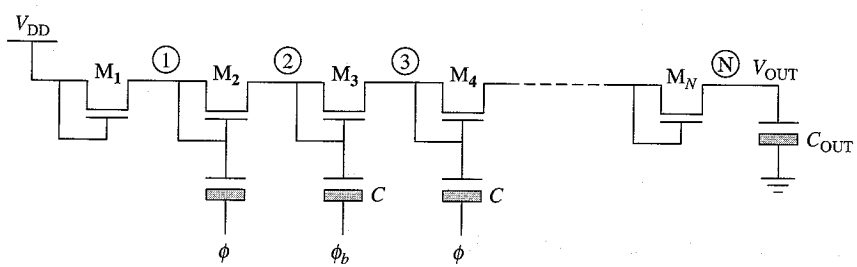


Figure 3-9 Dickson charge pump implemented using MOSFETs.

Equation 3-15 can be expressed as $Q_1 = \alpha C(V_{DD} - V_{t0})$. Equations 3-22, 3-23, and 3-25 can also be restated as follows:

$$Q(2n) = \sum_{j=1}^{j=2n} \alpha^j C(V_{DD} - V_t) - \sum_{j=1}^{j=2n} \alpha^{(j-1)} q_{inj} \tag{3-31}$$

$$Q(2n-1) = \sum_{j=1}^{j=2n-1} \alpha^j [C(V_{DD} - V_t) - q_{inj}] - \alpha^{2(n-1)} q_{inj}$$

$$Q(N) = C\left(\frac{V_{out}}{\alpha} - V_g\right) \tag{3-32}$$

Using Equations 3-31 and 3-32, the output charge q_{inj} can be expressed as

$$q_{inj} = \frac{C\left[\sum_{j=1}^{N+1} \alpha^j (V_{DD} - V_t) - V_{out}\right]}{\sum_{j=1}^{N} \alpha^j} \tag{3-33}$$

which was derived by Kanawaja et al., and the output rise time, T_r, from $(V_{DD} - V_t)$ to V_{fin} can be derived as

$$T_r = \frac{\ln\left[1 - \frac{V_{fin} - (V_{DD} - V_t)}{(V_{DD} - V_t)\left(\sum_{j=1}^{j=n+1} \alpha^j - 1\right)}\right]}{\ln\left[\frac{1}{1 + \frac{C}{N(C_{load})}}\right]} T_{osc} \tag{3-34}$$

As a test, you can see that if there is no body effect—i.e., if $\alpha = 1$, $V_s = V_d - V_{t0}$—then the preceding equation reduces to Equation 3-25.

3.6 Conclusion

We have analyzed the Dickson charge pump operation and quantified its operation while deriving its output voltage, current, and rise time characteristics. We also saw that the charge pump can be represented as a simple voltage source with an internal serial impedance. As the number of stages of the charge pump is increased, the pump is able to deliver higher output voltages, but the internal series resistance also increases, thus decreasing the available output current. In the subsequent chapters, we will build on this understanding of the Dickson charge pump[9] and scrutinize different factors affecting the charge pump's performance and efficiency. We will also examine different varieties of charge pumps designed to increase efficiency and get around the crippling high-threshold voltage problem, such as the 4-phase charge pump, CTC charge pump scheme, and charge pumps designed with PMOS transistors.

References

1. Pylarinos, L. Charge Pumps: An Overview. http://www.eecg.toronto.edu/~kphang/ece1371/chargepumps.pdf.
2. Dickson, J. "On-chip High-Voltage Generation in NMOS Integrated Circuits Using an Improved Voltage Multiplier Technique." *IEEE Journal of Solid-State Circuits*, Vol. 11, No. 6, pp. 374–378, June 1976.
3. Witters, J.S., G. Groeseneken, and H.E. Maes. "Analysis and modeling of on-chip high-voltage generator circuits for use in EEPROM circuits." *IEEE Journal of Solid-State Circuits*, Vol. 24, No. 5, pp. 1372–1380, October 1989.
4. San, H, H. Kobayashi, T. Myono, T. Iijima, and N. Kuroiwa. "Highly-Efficient Low-Voltage-Operation Charge Pump Circuits Using Bootstrapped Gate Transfer Switches." *IEEE Japan Transactions of EIS*, Vol. 120-C.
5. Tanzawa T. and T. Tanaka. "A dynamic analysis of the Dickson charge pump circuit." *IEEE Journal of Solid-State Circuits*, Vol. 32, No. 8, August 1997.
6. Papaix, Caroline and Jean-Michel Daga. High Voltage Generation for Low Power Large VDD Range Non Volatile Memories. PATMOS 2001. http://patmos2001.eivd.ch.
7. Lin, H., J. Lu, and T.Y. Lin. "A new 4-phase charge pump without body effects for low supply voltages." Asia-Pacific Conference on ASICs. August 6-8, 2002.
8. Khouri, O. et al. "Low Output Resistance Charge Pump for Flash Memory Programming." International Workshop on Memory Technology, Design, and Testing. MTDT 2001.
9. Kim, et al. High voltage generating charge pump circuit. U.S. patent number 6,661,682.

Chapter

4

Charge Pump Design Criteria

A charge pump is a common system, the construction of which is based upon simple CMOS transistors and capacitors. Relying on charge conservation theory and the capacitive coupling method, the charge pump generates the final output voltage, which can be either higher or lower than the given chip supply voltages.

Figure 4-1 shows a generic high-voltage charge pump block and the output path network. There are seven components: a high-voltage charge pump, a noise-filtering capacitance (C_{filter}), a decoupling capacitance ($C_{\text{decoupling}}$), a resistor ($R_{\text{regulation}}$) from the resistor divider circuit used in high-voltage regulation path, a high-voltage switch used to connect and disconnect the pump output to the loading circuits, a capacitive output loading (C_{load}), and a current loading (I_{load}). The charge pump is designed to bring its output to the desired voltage level within a fixed amount of time. The load current must be supplied at designed regulation levels.

Many criteria should be considered before the actual design work can start. Normally designers should consider at least two things: the available technology and the product specification. Technology involves the wafer processing for the chips. Models and parameters such as capacitance, sheet resistance, interconnection, and transistors are set by the technology and are used by designers in circuit design. The product specification contains the requirements the final product must achieve. Specifications such as power-consumption requirements, the pump-regulation level, the power efficiency of the design and die size, and so on, are the targets the circuit design has to meet on silicon. The following sections discuss the different major aspects that must be understood before starting the charge pump design.

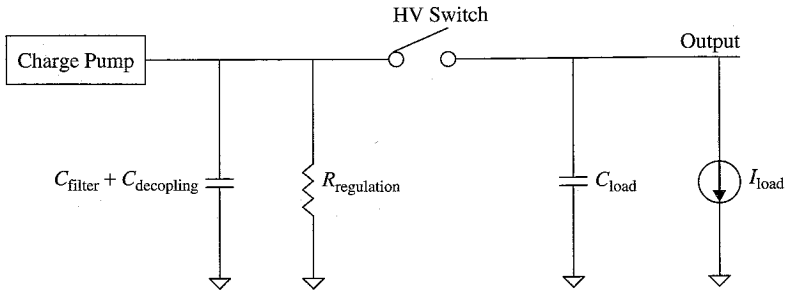


Figure 4-1 Generic high-voltage charge pump block and the output path network.

4.1 Technology

At any time one to many technologies may exist to facilitate the design of the charge pump. Some technologies may be more advanced at a given time, and the final product can be more costly to produce. On the other hand, some technologies may be more advanced at a given time and cheaper in terms of manufacturing cost. It is important to know what is available to designers and the pros and cons for each technology. More advanced technologies may not have the best economic return in the actual production.

Looking back in history, it is clear that technology evolves over time to meet the needs of a design. With more applications being introduced and more demands being created, better technologies are developed to allow more advanced designs to be feasible at lower costs. As mentioned in Chapter 1, the very first charge pump was proposed by Swiss physicist Heinrich Greinacher in 1919. By cascading more than one diode capacitor in series, a voltage greater than the supply voltage can be generated. Later, this technique was used by John Douglas Cockcroft and Ernest Thomas Sinton Walton to generate voltage potentials of more than 800,000 volts in their particle accelerator. This technique was again adopted by John F. Dickson, who proposed the technique of implementing the charge pump on a silicon chip.

Gradually charge pumps became widely used on chips, such as EPROM, EEPROM, Flash memory, PLL design, and so on. They help generate a voltage source that is higher in potential than the available chip power supplies. For example, EPROM/EEPROM uses high voltages potentials to electrically program, or electrically erase, the bits in the nonvolatile memories. The high-voltage potential required is generally higher than the chip supply voltage given. Before 1990, the needed high-voltage potential could not be generated on chips. In order to program or erase the EPROM/EEPROM memories, an external EPROM/EEPROM programmer was used. The desired high-voltage potential has to be generated off-chip. Due to this constraint, the nonvolatile memory chips have to be inserted by

the programmer to be programmed or be erased. Later, they are inserted back into the system. It is inconvenient to use nonvolatile memory in this fashion at system level. With the introduction of Flash EEPROM memory in the early 1990s, the high-voltage potentials could be generated on the chip itself. There is no need to use external power supplies to supply high voltages any more. Flash memory chips can be programmed directly or erased without the need for external high voltage sources. "In-system programming" is the term for on-chip programming or erasure.

In the early days, the high-voltage potentials needed by EPROM/EEPROM[1–3] were near 10~12 V. Later, as the chips migrated from the desktop to handheld devices, the power supply used in the system was scaled down to reduce the power consumption. At the same time, the off-chip high-voltage generation was migrated onto the chip itself. Now the charge pump is an essential component in many chip designs. Without the advance of technology, it is impossible to imagine the proliferation of cell phones, digital cameras, and MP3 players. They all rely on the nonvolatile memories in the system to store critical information. The charge pumps are the essential building blocks that make those applications feasible.

Technology determines the architecture and circuit design. To design a charge pump, designers need to understand the system supply voltage, oxide, and resistor and transistor specifications. All of these play a crucial role in determining circuit performance, total die size, and the power consumption of the design for the final products.

4.1.1 System supply voltage

The power supply for the chip comes from the system board where it is being soldered. The absolute levels of supply voltages are fixed by the system specification. For example, if the chip was designed for boot information storage on the PC main board, the supply voltage would be around 5 V. If the chip was designed to be used in a handheld device, such as a cellphone or PDA, the available supply voltage would be around 1.8 V or 3 V from the batteries. The charge pump uses the given chip power supplies as its charge source. It also uses the chip power supplies as the drives for all supporting pump circuits, such as clock generation and clock buffers. After the initial charge enters into the pump stages, its potential energy is elevated by the work of capacitive coupling. In each clock cycle, the change of potential energy depends on the clock driver strength, boosting capacitance sizes, and parasitic loadings on internal nodes. Ideally, if there is no loss of charge during transfer, and if all circuits are operating at 100% efficiency, then all the power consumed from the chip power supply by the charge pump should be delivered completely to the output of the charge pump. If the pump output power consumption is fixed, then Equation 4-1 should

hold true at all times for the same pump circuit operating at different power supply voltages.

$$P_1 = V_1 \times I_1$$
$$P_2 = V_2 \times I_2 \qquad (4\text{-}1)$$
$$P_1 = P_2 \Rightarrow V_2 \times I_2 = V_1 \times I_1$$

For example, in an ideal case, a 5 V charge pump, which consumes 10 mA of I_{cc} current, has a total power consumption of 50 mW. The output of a charge pump can deliver 50 mW of power at its regulation level. For an equivalent 3 V design to have 50 mW output power capability, the charge pump should consume 16.7 mA of I_{cc} current from 3 V power supply. The total power delivered to the output of the charge pump is still 50 mW. In reality, charge loss occurs during transfer due to parasitic loadings and due to the threshold of NMOS diodes. Also, the circuit efficiency is not 100% and all peripheral supporting circuits consume power.

Figure 4-2 shows the ideal relationship between the I_{cc} of a pump versus the chip supply voltage for an ideal pump with 100% circuit efficiency. If the pump output power requirement is not changed, the total power

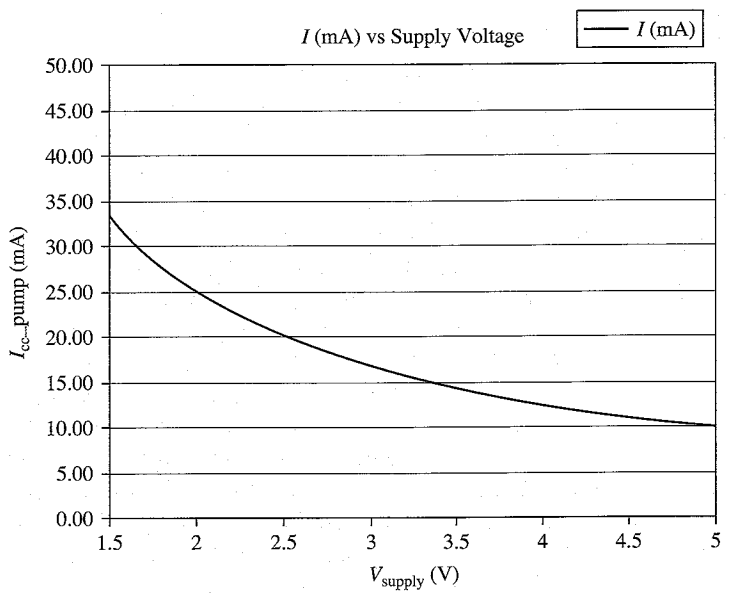


Figure 4-2 I_{cc} of pump vs. chip supply voltage in an ideal case.

consumed from the chip power supply should not change. Stated another way, the product of I_{cc} and V_{cc} should be constant. In reality, Equation 4-1 will not hold because the power supply voltage varies for the identical pump design. There is the loss of charge due to parasitic capacitive loadings on each stage, and there is the loss of charge due to the threshold of NMOS diodes at each pump stage. Peripheral supporting circuits will consume some amount of power in addition to the actual pump. In addition, the circuit efficiency is reduced dramatically as the supply voltage is scaled down.

As the chip power supply voltage is scaled down, as shown in Figure 4-2, the I_{cc} current would increase inversely proportionally. In reality, the efficiency of the transistor and the efficiency of the capacitance would drop as the driving voltage is scaled down. Current drivability would decrease as V_{gs} and V_{ds} are reduced. The threshold voltage of diode-connected NMOS becomes a more significant factor to impede the transfer of charge between stages. In order to meet the same pump output performance at a lower chip power supply voltage, the sizes of transistor and capacitance have to be increased further rather than proportionally with scaling of supply voltage to compensate for the reduction of supply voltage.

In Figure 4-3, a realistic I_{cc} versus V_{supply} curve is plotted on top of the ideal curve discussed in Figure 4-2. As the chip supply voltage is scaled down, the pump has to consume more current and more power in order to meet the same pump output performance. The actual trendlines head upward away from the ideal line as the chip power supply goes lower

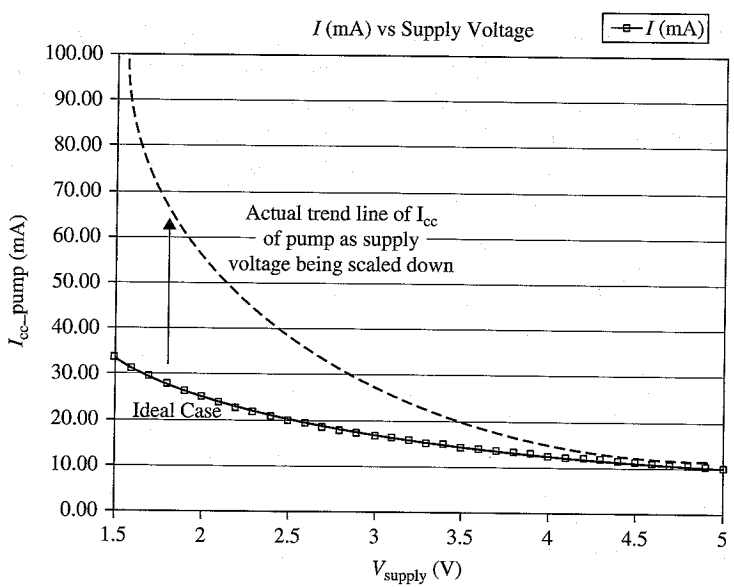


Figure 4-3 I_{cc} of pump vs. chip supply voltage in a realistic case.

and lower. The power efficiency of the charge pump worsens with the scaling of the chip supply voltage. The phenomenon of increased power consumption with the scaling down of the chip power supply has a great impact in terms of the overall chip power specification.

For ASIC chips, the scaling of technology allows the scaling down of oxide thickness as well as the scaling down of the transistor dimension. More transistors can be packed into the same layout area on the silicon. As a consequence, thinner oxide allows the chip power supply voltage to be scaled down to prevent oxide reliability issues. The direct benefit is that the total power consumption for the same number of transistor counts is reduced.

Figure 4-4 shows a simple logic circuit with one inverter driving another inverter. Node *In* switches from high to low, and node *Out* is charged up by the chip supply voltage V_{supply}.

Equation 4-2 represents the total loading capacitance seen by the driving inverter. The gate oxide thickness is assumed to be identical for both NMOS and PMOS transistors. The total loading capacitance is the product of unit gate capacitance C_{ox} and the total transistor gate area, as shown in Equation 4-2. As technology is scaled down, the oxide thickness is also scaled down and C_{ox} increases. At mean time, the dimensions of the transistors are scaled down too and this allows the total gate area to reduce. From Equation 4-2, C_{total} is unlikely to vary strongly with the scaling of technology.

$$C_{load} = C_{ox}(W_pL_p + W_nL_n) \tag{4-2}$$

V_{supply} is scaled down as the gate oxide thickness is scaled down to prevent any oxide reliability issue. Equation 4-3 represents the total charge consumed for one clock cycle for this simple logic gate. Because the loading capacitance is assumed to remain unchanged with the scaling of technology, Q_{total} is scaled down with the technology, as shown in

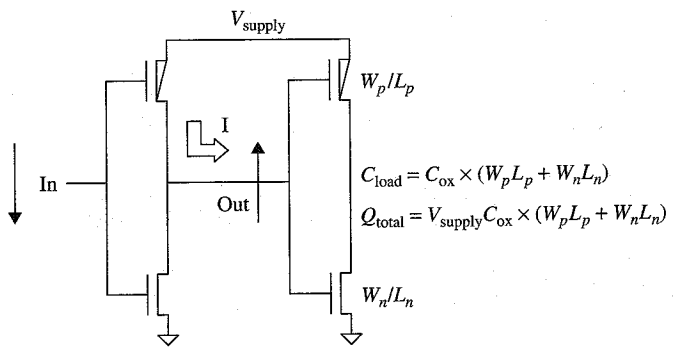


Figure 4-4 Charging of an inverter

Equation 4-3. If the transistor count does not increase, the total power consumption of the same ASIC chip design should be scaled down with the technology. However, this phenomenon is unlikely to occur because the transistor count is usually increased to include more functions on-chip in new designs. The trend of total power consumption of ASIC chips would stay flat or go up against the technology scaling trend.

$$Q_{total} = V_{supply} \times C_{load} = V_{supply}C_{ox}(W_pL_p + W_nL_n) \tag{4-3}$$

For chips (including on-chip charge pumps) to generate high-voltage potentials, it is a different story regarding power consumption. Consider the following equations:

$$Q_{ramp} = \frac{V_{regulation} \times C_{load}}{\#\text{cycles}} \tag{4-4}$$

$$Q_{regulation} = I_{load}(t) \times t_{cycle} \tag{4-5}$$

$$Q_{cycle} = Q_{ramp} + Q_{regulation}$$

$$Q_{cycle} = \frac{V_{regulation} \times C_{load}}{\#\text{cycles}} + I_{load}(t) \times t_{cycle} \tag{4-6}$$

Equation 4-6 denotes the charge per cycle that needs to be delivered to the output of the charge pump as a function of time (t). The first term in Equation 4-6 is represented by Equation 4-4. It denotes the minimum average charge transferred per clock cycle to charge up the output capacitive loading within the specified ramp-up time requirement. If the architecture does not change, C_{load} is unlikely to be changed with the same design. $V_{regulation}$ is a technical parameter specified by design requirements. It is also unlikely to vary too much from generation to generation. The second term in Equation 4-6 is represented by Equation 4-5. It denotes the additional charge consumed by the current load circuits connected to the output of the charge pump. This term is unlikely to be changed unless the architecture or the loading circuit is changed. Because Equation 4-4 and Equation 4-5 are not likely to vary too much with the technology, the total power consumption required on the output of the charge pump is not likely to be scaled down with the scaling of technology. As a consequence, the total power consumption of the entire charge pump is not likely to be scaled down either. This phenomenon is fundamentally different from most ASIC designs

Chips have a power budget determined by system design. As the power supply drops, to meet the same output performance for the charge pump (from Figure 4-3 and Equation 4-6), the I_{cc} current of the chip must increase at a higher rate than the rate of voltage supply scaling. It is important for designers to understand what circuits are being designed in the first place; then they can correctly define the chip specification for overall power consumption and chip performance. Another bad impact due to the lowering chip power supply implies that a larger capacitive area is needed to transfer more charge within each clock cycle to meet the same charge pump performance. If the performance of the pump is not allowed to be scaled down with the lowering power supply, the I_{cc} current needs to increase dramatically, as shown in Figure 4-3. In order to conduct a larger current than designs from previous generation, the overall pumping capacitance must increase at least proportionally to meet the output current required. Larger capacitance translates to a larger layout area, a larger clock driver size, more buffering stages for pump clock drivers, a longer route for the signal to travel, and larger parasitic loadings on all internal nodes of charge pump. All these have a negative effect on the performance of the pump. To further compensate for these losses, the size of the pump must be increased further. Overall, these negative impacts do not favor the charge pump design at all.

Another impact related to chip power supply is power bus planning, which could indirectly lower the supply voltage near the circuits even further. Even with the given chip supplies' specified system, tremendous effort is required to work out the details of power bus planning. Insufficient or bad power bus planning on-chip can cause the internal power supplies to have significant offsets from the voltage levels given the power supply pins. On the system level, bad board design and incorrect noise filtering on the power bus could cause additional voltage drops on the power supplies before even reaching the power supply pins of the chips.

Charge pump designers need to have enough noise margin estimated for the power supplies. Larger noise margins will not kill the designs. Of course, too much noise margin built into a design will increase the die size and total power consumption. Derivative projects can always improve based on silicon debugging from the mother chip. An underestimate of the power margin could kill the design and might require an expensive full-mask revision as well as delay production.

4.1.2 Silicon dioxide (SiO_2)

Silicon dioxide (SiO_2) is one of most important building blocks for semiconductor devices. It is used not only for isolation purpose, such as gate oxide or field oxide, but also as an active circuit component in many

analog circuits. In pump design, the charge pump needs to transfer the charge from lower potentials to higher potentials. The common practice of changing potential energy for charge is based upon charge conservation theory and capacitive coupling technique. The pumping capacitance per stage is determined by the pump output ramp-up speed and the pump output power consumption. Efficient unit area capacitance and a reduction of the total layout area for the pump capacitance required are both crucial in charge pump design.

Let us review the capacitance at a basic level to see how it is used to meet the pump design requirements. As shown in Figure 4-5, two conducting plates, separated by some dielectric material, form the capacitor. In Figure 4-5, the dielectric material is assumed to be SiO_2.

Plate A is connected to the anode of the battery, and plate B is connected to the cathode of the battery. The potential voltage of the battery is V. As the positive charge is accumulated on the top plate, the negative charge is attracted toward the bottom plate. Because two plates are separated by SiO_2, no conduction can occur between nodes A and B. The electric field, E, is built up and points from plate A to plate B.

The calculation for the parallel plate capacitance is given in Equation 4-7 for the sake of this discussion. ε_0 is defined as the permittivity of the dielectric material. SiO_2 is one of the most commonly used dielectric materials for semiconductor processing. ε_r is the relative permittivity of the air. Also, *area* is the total effective overlapping area between the two parallel plates, and *thickness* is the thickness of the dielectric material, or the spacing between the two parallel plates. From Equation 4-7, if the dielectric material is fixed, the total capacitance (C_{total}) is proportional to the area of overlapping capacitance, and is inversely proportional to the distance separating those two plates.

$$C_{total} = \varepsilon_0 \varepsilon_r \frac{\text{area}}{\text{thickness}} \tag{4-7}$$

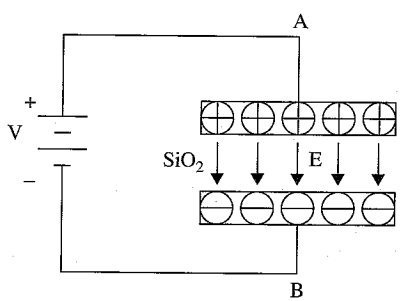


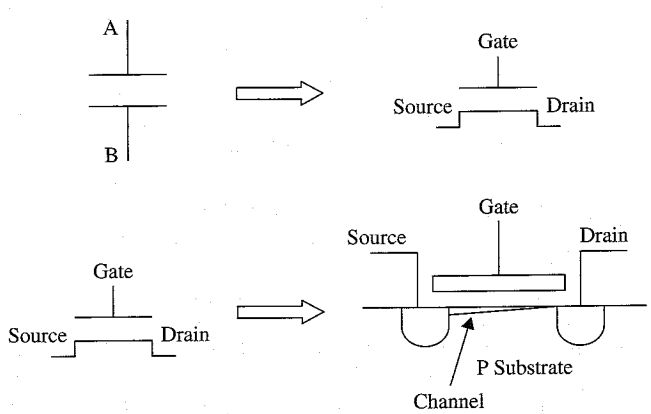
Figure 4-5 Capacitance connected to a battery.

In actual design, the gate capacitance of a MOSFET is usually chosen as the boosting capacitance because of its maximum capacitive efficiency over area. Two of the conducting plates are normally made of polysilicon and the channel formed underneath the gate. In between the two plates is SiO_2, which separates the two conducting nodes. This gate capacitor is commonly used. The dielectric material could be silicon dioxide, nitride, ONO, and so on. The MOSFET transistor has to be biased in either the accumulation region or the inversion region to maximize its equivalent gate capacitance equal to the gate oxide capacitance. The gate oxide is chosen as the boosting capacitance because its thickness is approximately 1/10 ~ 1/20 of the thickness of the field oxide. It is the thinnest dielectric layer that can be found on silicon. It is also well controlled in terms of the silicon process. For the same silicon area, the gate capacitance generates the largest unit capacitance over an area compared with other types of capacitance.

Figure 4-6 shows the generic representation of MOSFET gate capacitance as well as various components related to gate capacitance. The polysilicon gate is one of the terminals for the capacitor. The other terminal is the connection for the source/channel/drain of the NMOS transistor. It is ideal to have MOSFET gate capacitance identically to C_{ox}, with no variation in term of biasing voltages. Also, the design should be simple. However, this capacitance does vary with the biasing conditions of the MOSFET. Figure 4-7 shows the gate capacitance of NMOS versus V_{GB} biasing. The C-V curve of NMOS gate capacitance can be divided into three regions, as shown in Equations 4-8 through 4-10.

$V_{GB} > V_{Tn}$ defines the inversion region.

$$C_{gate} = C_{ox} \tag{4-8}$$



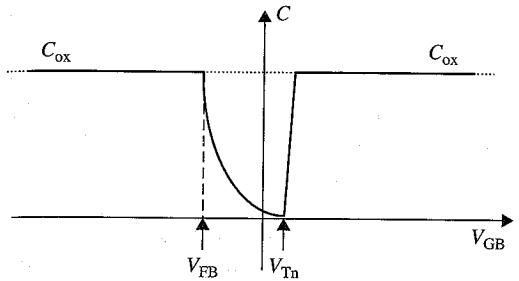


Figure 4-7 C-V curve of NMOS gate capacitance.

$V_{Tn} > V_{GB} > V_{FB}$ defines the depletion region.

$$C_{gate} = \frac{C_{dep}C_{ox}}{C_{dep}+C_{ox}} = \frac{C_{ox}}{1+\frac{\varepsilon_s}{\varepsilon_{ox}}\frac{t_{ox}}{\chi_{dep}}} \tag{4-9}$$

$V_{FB} > V_{GB}$ defines accumulation region.

$$C_{gate} = C_{ox} \tag{4-10}$$

As the biasing conditions change for NMOS, the equivalent gate capacitance varies between inversion, depletion, and accumulation regions. It is important to choose the right devices and operate them in the targeted regions. Otherwise, the gate capacitance cannot be simply treated as C_{ox} in design.

The capacitance is inversely proportional to oxide thickness. The thinner the gate oxide is, the greater the gate capacitance per unit layout area. Stated another way, over the same silicon area, more charges can be stored across the capacitance with thinner dielectric material in between. Chip design demands this unit capacitance over area to be as high as possible to reduce the overall layout area. However, the oxide thickness (or the dielectric thickness) cannot be scaled down forever. The breakdown voltage of the dielectric material constrains the design. Oxide breakdown has always been a serious reliability concern in the semiconductor industry because of the continuous pressure toward smaller and thinner devices. There are three reasons to drive down the gate oxide thickness. The first reason is that smaller device dimensions can be achieved. This allows more and more devices to be packed into the same area compared with previous technologies. The second reason is that the current drivability is increased with the same device dimen-

the power supply voltage to be reduced. Less power consumption should be expected over the same transistor count.

Higher electrical fields in the oxide increase the tunneling of carriers from the channel into the oxide.[4] These carriers slowly degrade the quality of the oxide and lead, over time, to failure of the oxide. This effect is referred to as *time dependent destructive breakdown* (TDDB). The thinnest oxide layers today are already less than 50 angstroms thick. An oxide layer can break down instantaneously at $8-11\times10^6$ V per centimeter of thickness, or 8–11 V per angstrom of thickness. Based on the voltage levels required, the oxide thickness should be carefully designed for high-voltage devices to avoid any reliability issues.

4.1.3 Resistor

The resistor is another critical component in charge pump design. Important characteristics of a resistor are sheet resistance of material, temperature coefficient in design, junction leakage and bias-dependant of resistance.

The effect of the resistor could be good or bad, depending on the application. Most of the time, resistance causes a bad effect in circuit designs. The *RC* delay, the first order delay metric of a circuit, of any wiring is one of the main causes for the slowdown for signal transitions. The charge pump relies on capacitive coupling to transfer charge and to elevate the potential energy of the charge. Because pump clocks are operating at a relative high frequency to drive large boosting capacitance, the metal layers used by pump clocks contribute to the *RC* delay of clock signals. As the charge is transferred from stage to stage, the interconnecting metal layers between the stages cause extra *RC* delay, thus impeding the transfer. The effect of resistance mentioned here is an example causing bad charge pump performance.

Sometimes the circuit designs do use resistors to perform important functions. For example, the simple delay element used in design can be purely a *RC* delay circuit. In terms of pump regulations, the resistor divider scheme is commonly used. In analog circuit design, the resistor also plays many important roles.

The first parameter related to the resistor should be the sheet resistance. Figure 4-8 shows the cross-section view of a conducting wire. The width of the wire is W. The thickness is t, and the length of the wire is L. The total resistance is given in Equation 4-11. The sheet resistance (or the resistance per unit length) of the material is described in Equation 4-12. Choosing the right type of material with the right range of sheet resistance is very important to design. The sheet resistance varies with different material. It is process-dependent. Metal layers have very small sheet resistance. WELL material has very high sheet resistance. Sometimes, if a design needs mega-ohm resistance, it does

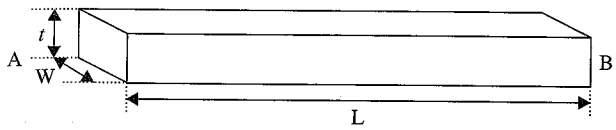


Figure 4-8 Sheet resistance of a conducting material.

not make sense to choose metal layer for the resistor. The layout could explode due to the metal layout's small sheet resistance. On the other hand, if 1 ohm of resistance is needed, it does not make sense to use WELL resistance. The discrepancy due to width and length could cause a much larger discrepancy in terms of the final resistance value.

$$R_{\text{total}} = \frac{\rho L}{tW} = \left(\frac{\rho}{tW}\right)L = R_{\text{sheet}}L \tag{4-11}$$

$$R_{\text{sheet}} = \frac{\rho}{tW} \tag{4-12}$$

The second parameter associated with the resistor is the temperature coefficient of the material. As temperature varies, the effective resistance will vary, as shown in Equation 4-13. R_0 is the initial resistance at temperature T_0, which is normally set at 27 °C. α is the temperature coefficient of the conducting material. The degree of variation is material-dependent. In charge pump design, if the resistor divider scheme is used for pump regulation, the regulation level of high-voltage output is immune to the temperature variation. However, the DC current consumption through the resistor network will vary as the temperature changes. Sometimes the output of the charge pump goes through a low-pass filter before connecting to the loading circuits. *RC* delay would vary with the temperature. For many analog circuit designs, such as band gap reference, reference current generation, and so on, the temperature coefficients of resistors need to be studied.

$$R(t) = R_0\left[1+\alpha\left(t-T_0\right)\right] \tag{4-13}$$

The third parameter associated with resistance is the junction leakage current of the resistor. Resistor types, such as metal resistors and ploy resistors, have no leakage to neighboring dielectric materials. Other type of materials, such as n^+ resistors, p^+ resistors, and NWELL resistors, all have some amount of leakage through the junction. This current could be represented by the reverse-biased diode leakage. What is the concern of current leakage through the resistor? It is all about power consumption of the charge pump. A charge pump design–type of circuit

is not power efficient. To deliver 50 uA of current on the output at a specified regulation level, the pump itself could consume 5 mA of current from its power supply. As the level of regulation increases, or as the power supply voltage scales down, this power efficiency is reduced even further. Minimizing any unwanted current on the output of the pump is becoming a crucial issue. Even if no physical resistor is intended to be connected to the output of the pump, all the source/drain junctions of the transistors and the WELLs connected to the output of the pump act as sources of leakage. They are an addition to the current load on the output of charge pump, and they tend to cause the pump to burn extra power and increase the size of the pump design.

The fourth parameter associated with the resistor concerns bias-dependent resistance. This parameter is mainly related to n^+ resistors, p^+ resistors, and WELL resistors. As the potential voltage changes across the surface of the resistive material, the depletion region underneath changes with the biasing voltage. As a consequence, the sheet resistance may not be the resistance expected from the design manual. This variation could affect pump output regulation levels, timing delay, and so on. Understanding the effect allows the countermeasure to be built into the actual circuits.

4.1.4 Transistor specification

Transistors are the fundamental building blocks for CMOS circuits. Many device characteristics are important to the design, such as I_{ds}, threshold voltage V_t and body bias effect, W_{max} and L_{min} of transistors, junction breakdown voltage, gate oxide breakdown voltage and snapback effect, etc.[5,6]

For high-voltage charge pump design, transistors can be divided into two main categories: low-voltage transistors for logic functions and high-voltage transistors used in high-voltage paths. Low-voltage transistors are mainly used for logic operations in the peripheral of the charge pump. Functional blocks, such as clock generation, the delay element, and buffers are normally implemented by low-voltage type transistors. The potential voltages used by those blocks are typically chip supply voltages or potential voltages being regulated down from those sources. They are no different from the devices used for other logic operations. High-voltage transistors are used for connections in the high-voltage path, not only in the charge pump itself, but also all the peripheral circuits that experience high-voltage potential in operations. The device characteristics of the normal low-voltage transistors, such as I_{ds}, threshold voltage V_t, body bias effect, and so on, are equally important to high-voltage transistors. Due to high-voltage potentials, other aspects need to be specifically watched. Source/drain punchthrough issues, junction breakdown, oxide breakdown, and the snapback effect are all critical in dealing with high-voltage devices.

High-voltage transistors need to conduct current and transfer charges. I_{ds} is an important parameter that can be used to characterize transistors. I_{ds} in the linear region is expressed in Equation 4-14, and I_{ds} in the saturation region is expressed in Equation 4-15. Current drivability is important in terms of charge pump performance, such as the precharging of internal nodes and the charge transferring between stages. The higher the current drivability, the faster the charge can be delivered. Internal nodes can reach equilibrium faster, and faster clock frequency is applicable.

$$I_{ds} = u_n c_{ox}\left(\frac{W}{L}\right)\left[\left(V_{gs} - V_t\right)V_{ds} - V_{ds}^2/2\right] \quad \text{(Linear region)} \quad (4\text{-}14)$$

$$I_{ds} = \frac{1}{2}u_n c_{ox}\left(\frac{W}{L}\right)\left(V_{gs} - V_t\right)^2 \quad \text{(Saturation region)} \quad (4\text{-}15)$$

The threshold voltage V_t of a high-voltage MOSFET is very crucial. It affects the charge transfer, leakage, and size of the charge pump. The effect of body bias further compounds the problem even more severely as the source bias rises. Equation 4-16 measures the V_t of a MOSFET with source bias V_{sb}. The threshold of an NMOS device prevents the charge from being fully transferred between the source and drain of the device. At each pump stage, the efficiency of the charge pump is reduced due to its inability to transfer all the charge. A five-stage charge pump with 3 V power supply and 1 V threshold voltage has only 12.5% output power efficiency in an ideal case. If an eight-stage charge pump is used under similar conditions, the output power efficiency is only 2.4%. These cases do not consider the variation in threshold voltage. If we consider the body bias effect on V_t, later stages would suffer an even bigger voltage drop between source and drain nodes than that of the earlier stages. For charge pump design, it is preferable to have a high-voltage device with a low V_t and the minimum body bias. However, the threshold voltage cannot be too low due to concerns explained later in this chapter regarding device leakage and punch-through issues.

$$V_t = V_{t0} + \gamma\left(\sqrt{2|\Phi_F| + V_{SB}} - \sqrt{2|\Phi_F|}\right) \quad (4\text{-}16)$$

The dimensions of high-voltage transistors can impact the layout size of charge pumps. In terms of the transistor width, the maximum width (W_{max}) allowed is determined by process and device modeling. Devices wider than W_{max} are not allowed by layout rules. Modeling of devices in test chips has a limited dimension. It can cause incorrect projection in simulation for a device that is significantly wider than those used on

the test chip for device modeling. A wider device must be broken down into a device with smaller, multiple fingers, as shown in Figure 4-9. The extra spacing rules, contact rules, and routing of the wire will cause the size of the layout to increase. In terms of the length of the high-voltage transistors, the minimum value of channel length (L_{min}) is specified. In the discussion of the I_{ds} of high-voltage transistors in Equation 4-14 and Equation 4-15, it is understood that decreasing L enables better current drivability for transistors with the same width. This is what designers wish for. However, for high-voltage devices, L_{min} cannot be minimized forever. The punch-through effect is a critical concern for high-voltage design. As the voltage across the junction increases, the width of the depletion region near the source/drain increases under bias. If the channel is too short, the depletion layers under the source and drain regions could merge together. This will form a continuous depletion region and allow current to flow between source and drain with low gate bias, such as $V_{gs} < V_t$. There is rapid increase of I_{ds} with increasing drain voltage. If this situation occurs, the source and drain of the MOSFET can be considered to be shorted together without the control of gate voltage. L_{min} is required for high-voltage transistor design rules under specified biasing conditions.

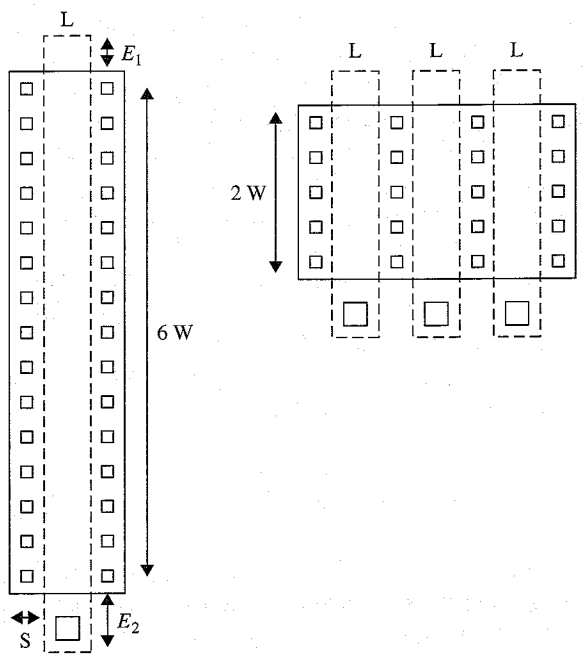


Figure 4-9 Breaking down a wider device into multiple fingers.

For high-voltage transistors, it is possible that either the source or the drain (or both) will connect to some high-voltage potentials. For high-voltage NMOS, the bulk is p^-, and the source/drain is n^+. p^- and n^+ form a simple diode. If the reverse-biased voltage reaches a critical level, there will be a large increase in reverse-biased current as a result of junction breakdown. Once this happens, the transistor could be damaged. This critical level is called *junction breakdown voltage*. In designing high-voltage circuits, high voltage at all nodes should be checked carefully so as not to exceed this critical level.

Oxide breakdown of high-voltage transistor is covered in the section titled "Silicon Dioxide (SiO_2)." Higher fields in the oxide increase the tunneling of carriers from the channel into the oxide. These carriers slowly degrade the quality of the oxide and over time lead to failure of the oxide. This effect is known as time-dependent destructive breakdown (TDDB). Similar to the check for junction breakdown, all gate voltages in a high-voltage path should be checked so as not to exceed the gate oxide breakdown voltage.

Snapback is another phenomenon that can take place within a MOSFET transistor. When the carrier is hot, a high-energy particle could trigger snapback if the field across the drain region is sufficiently high. Snapback occurs when the parasitic bipolar transistor underneath MOSFET transistor is triggered. If the parasitic BJT (Bipolar Junction Transistor) is turned on, the result is a very high current between the drain and source region of the transistor. A special protection scheme should be used to turn on a MOSFET transistor if its drain is exposed to very high voltages.

4.2 Specification

Now that we have looked at the technology issues related to pump design, we will discuss the design specifications for a charge pump.

4.2.1 Output load characteristics

The charge pump is designed to provide high-voltage potential to the loading circuits. The characteristics of a loading circuit can dramatically affect the charge pump implementation. The load of charge pumps can generally be divided into two categories: capacitive load and current load. Figure 4-10(a) shows the generic representation of the capacitive load to a charge pump. As long as the load does not draw any DC current during the regulation phase of operation, it can be simplified into a capacitor. Figure 4-10(b) shows the response of pump output voltage over time for a capacitive load. $T_{regulation}$ is the duration of time for the charge pump output to reach its regulation level. Before the pump output voltage reaches regulation, the required pump output current is

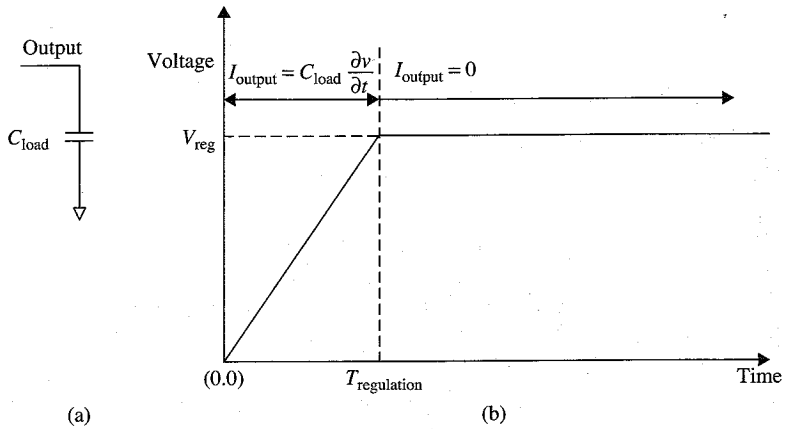


Figure 4-10 Capacitive load to the charge pump.

used to charge up the capacitive load, as shown in Equation 4-17. After the pump output reaches the regulation, no more capacitive charging current is needed. The pump output current in this region is described by Equation 4-18.

At any given time, the minimum current a charge pump design with a capacitive load should deliver to the output is the MAXIMUM function of Equation 4-17 and Equation 4-18. This is rewritten in Equation 4-19.

$$I_{output} = C_{load} \frac{\partial V}{\partial t} \qquad t \le T_{regulation} \tag{4-17}$$

$$I_{output} = 0 \qquad t > T_{regulation} \tag{4-18}$$

$$I_{output} = \text{MAX}\left[C_{load} \frac{\partial V}{\partial t}, 0\right] = C_{load} \frac{\partial V}{\partial t} \tag{4-19}$$

Figure 4-11(a) shows the generic representation of the current load to a charge pump. As long as the load draws DC current during the regulation phase, the load can be considered a current load. Figure 4-11(b) shows the response of the pump over time for the current load.

Before the pump output voltage reaches regulation level, the load current consumed is shown in Equation 4-20. The first term is the average current needed to charge up the capacitive loading, and the second term is the current consumed by the DC current load (in general, due to a resistor divider type regulator) at any given time. After pump output reaches the regulation level, there is no more capacitive current for

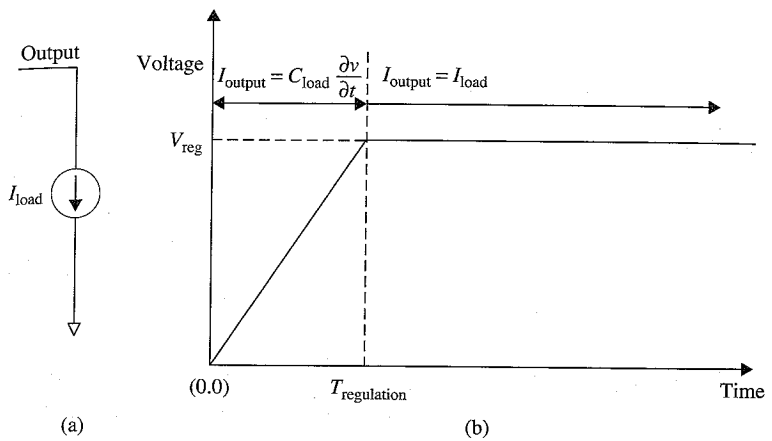


Figure 4-11 Current load to a charge pump.

given time shown in Equation 4-21. At any given time, the charge pump should meet the minimum current requirement delivered to its output. This is the MAXIMUM function of Equation 4-20 and Equation 4-21, and is rewritten in Equation 4-22.

$$I_{output} = C_{load} \frac{\partial V}{\partial t} + I_{load}(t) \qquad t \le T_{regulation} \tag{4-20}$$

$$I_{output} = I_{load}(t) \qquad t > T_{regulation} \tag{4-21}$$

$$I_{output} = \text{MAX}\left[C_{load} \frac{\partial V}{\partial t} + I_{load}(t), I_{load}(t)\right] \tag{4-22}$$

Comparing the capacitive load and the current load for the charge pump, the main difference is the DC load current $I_{output} = I_{load}(t)$ during the regulation phase. It implies there is power consumption by the charge pump during regulation. In the capacitive load case, after the load capacitance is charged up to regulation level, no more current is needed on the load, and the charge pump can be shut off to cut off power consumption. In current load cases, $I_{load}(t)$ is the component that always exists on the output of charge pump. As a consequence, the pump needs to supply current at all time, so the charge pump has to operate continuously, and so provides the power consumption of the whole circuit.

4.2.2 Pump output voltage

Normally, the power supply is defined as the source voltage that can pro-

power that can be delivered by the supply is defined by the design specification. If the loading current exceeds this maximum requirement, the supply voltage will drop below its regulation level. The charge pump is designed to provide current/power at high-voltage potentials. Although the current is drawn from the output of the charge pump, eventually all the charge and power have to come from the given chip supply.

The performance of a charge pump can be broken down into two parts:

- **Pump output ramp-up speed** This is the pump output recovery speed if more capacitive load being connected in the middle of regulations.
- **Pump output power capability** This is the maximum current a charge pump can deliver while maintaining the output-regulated voltage levels.

Figure 4-12 shows the waveform of output voltage over time. It can be divided into four regions. The first region is for charge pump output to ramp up from its initial level to the final regulation level, V_{reg}. During this period, the load capacitance is charged up. The second region is a very short regulation phase. Then some capacitive load is connected to the output. Charge sharing will bring down the voltage at the output. Immediately, the pump responds and tries to recover back to V_{reg}. The last region is when the pump output goes back into regulation once again. This is the first plot designers need to understand. It carries a lot of design specifications that determine the final implementation of a charge pump. The generic high-voltage path from Figure 4-1 has been duplicated in Figure 4-13 for better illustrative purposes. The output waveform of Figure 4-13 would be exactly the same as the one shown in Figure 4-12.

Looking at Figure 4-12, we can extract several key design parameters from the plot. First, this pump is regulated at a level of V_{reg}. This is

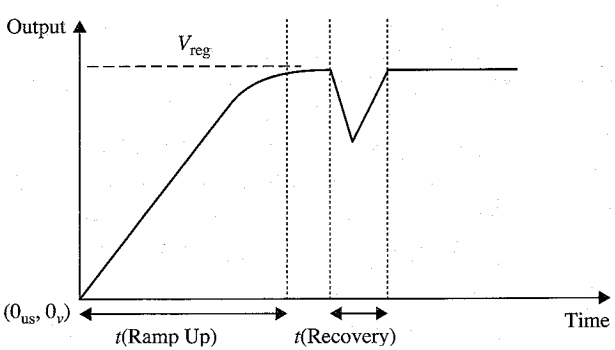


Figure 4-12 Charge pump output voltage versus time.

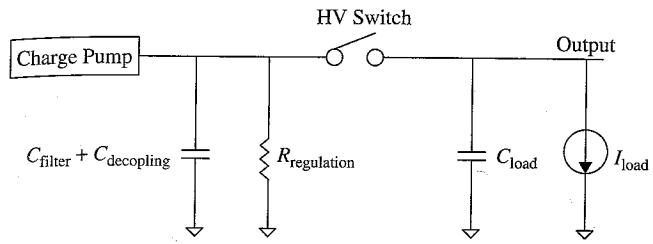


Figure 4-13 Generic high-voltage path with charge pump.

the targeted regulation level in this design. Second, it would take the pump a certain amount of time to ramp up the output node before it reaches regulation level. The time it takes is t_{rampup}, which is the amount of time to charge up C_{filter} and $C_{decoupling}$ to V_{reg} in Figure 4-13. Third, as Figure 4-13 shows, when the HV switch connects capacitive load C_{load} to the pump output, charge sharing will immediately bring down the output voltage of the charge pump. It would take the pump some amount of time to recover back to its original regulation level. This recovery time is specified as $t_{recovery}$.

How do these parameters relate to the charge pump design? Pump output ramp-up (or recovery) speed is critical to the chip's performance. For example, a Flash memory chip has program and erase timing specifications. Usually for programming a byte or a page of information, the maximum programming pulse width is fixed based on program speed. To meet the speed requirement, within one program pulse, voltage on the wordline has to ramp up and stay at the regulation level for some minimum amount of time. Failing to do so causes inefficient programming and pushes out the actual program speed performance. Ramp-up of the pump output voltage requires the pump to charge up all capacitive loadings within the given amount of time. This is one of the design specifications.

Equation 4-23 calculates the total charge that needs to be transferred to the charge pump output to charge up the capacitive load; Equation 4-24 shows within one clock cycle how much charge should be transferred from stage to stage. This is proportional to the boosting capacitance size, and it is also proportional to the boosting clock voltage amplitude. The threshold of the diode-connected transistor inhibits the full transfer of the charge between stages. Equation 4-25 shows the average current the pump has to supply to pump output to meet the t_{rampup} requirement. The shorter the ramp-up time required, the more pump current is needed, and the larger capacitance required per stage.

$$Q_{rampup} = \left(C_{filter} + C_{decoupling}\right) \times V_{reg} \tag{4-23}$$

$$Q_{\text{cycle}} = C_{\text{boost}} \times \left(V_{\text{boost}} - V_t\right) \tag{4-24}$$

$$I_{\text{rampup}} = Q_{\text{rampup}} / t_{\text{rampup}} \tag{4-25}$$

The requirement in the recovery phase is similar to that of the ramp-up of the charge pump. It is assumed that the voltage across C_{load} is 0 V before the HV switch closes.

Because C_{filter} and $C_{\text{decoupling}}$ have been charged up, the only capacitive component that needs to be charged up is C_{load}. Equation 4-26 through Equation 4-28 for the recovery phase are equivalent to Equation 4-23 through Equation 4-25 for the ramp-up phase. The charge pump has to meet the requirements of both cases.

$$C_{\text{filter}}\, Q_{\text{recovery}} = C_{\text{load}} \times V_{\text{reg}} \tag{4-26}$$

$$Q_{\text{cycle}} = C_{\text{boost}} \times \left(V_{\text{boost}} - V_{\text{th}}\right) \tag{4-27}$$

$$I_{\text{recovery}} = Q_{\text{recovery}} / t_{\text{recovery}} \tag{4-28}$$

4.2.3 Pump output current

At fixed output voltage level, the pump needs to deliver current to the load, in addition to the regulation current if the regulation scheme demands DC current.

Figure 4-14 describes the I-V curve of a charge pump. This curve is also called the load line of the charge pump. The pump output current and the corresponding output voltage are plotted. Each point on the curve is the actual operating point of the pump. It represents how much the maximum output current pump is capable of delivering given the regulated voltage, or with the fixed current consumed on the pump output, the maximum regulation level at which the output of charge pump can be maintained. For example, node (V_1,I_1) indicates that if the output of the pump was regulated at V_1, the maximum current pump output that could be delivered is I_1. With any load current less than or equal to I_1, the output voltage can be regulated at V_1. However, if a current larger than I_1 is demanded on the output of the pump, such as I_2 $(I_2 > I_1)$, the pump output could not maintain its regulation at V_1. In Figure 4-14, it is easy to see that V_2 is the maximum voltage level the charge pump can regulate with a load current of I_2. V_2 is less than V_1.

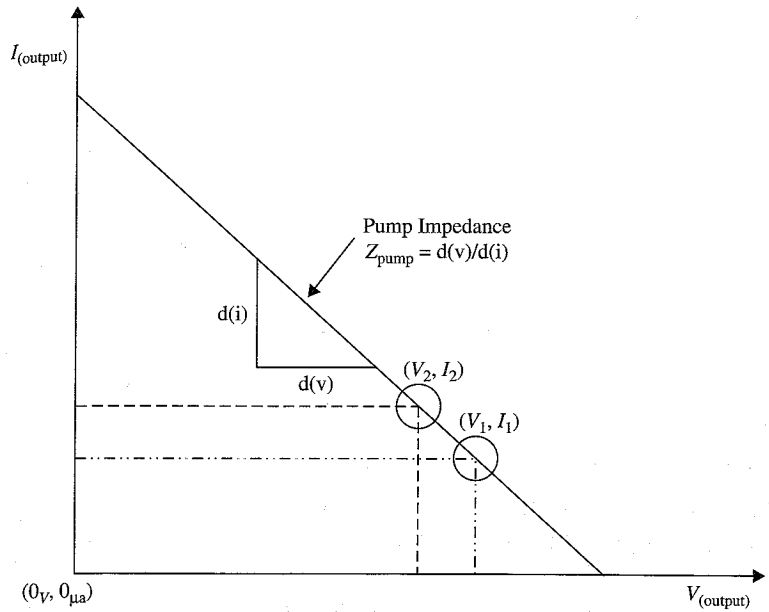


Figure 4-14 I-V curve of a charge pump.

In order to maintain the required output current, the boosting capacitance and the boosting voltage have to meet the requirement shown in Equation 4-29.

$$I_{\text{out}} = \frac{\partial\left[C_{\text{stage}} \times \left\{V_{\text{clock}} - V_t\right\}\right]}{\partial\left[T_{\text{clock\_period}}\right]} \tag{4-29}$$

In practice, the charge transferred per clock cycle from stage to stage has to be greater than or equal to the total charge consumption on the output per clock cycle.

Another important parameter can be derived from Figure 4-14—the slope of the I-V curve. The inverse slope of the I-V curve was described as the impedance of the charge pump. This impedance is related to many parameters in pump design, such as the pump clock frequency, the size of the pump-boosting capacitance, the boosting voltage per stage, the threshold voltage of the diode, the parasitic capacitance within each stage, and so on. The steeper the slope is, the better the current drivability.

Figure 4-15 shows the I-V curve of the charge pump with regulation. The design is regulated at V_{reg}. The maximum the pump can supply

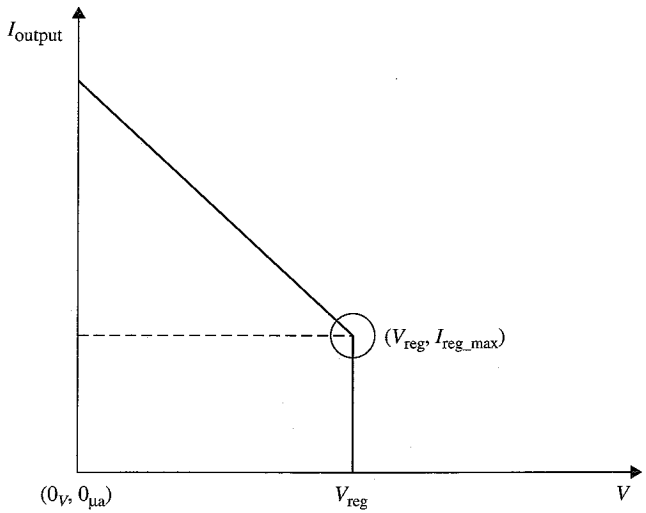


Figure 4-15 I-V curve of a charge pump with regulation.

at V_{reg} is I_{reg_max}. If the load current exceeds this absolute value, the design would fall out of regulation. The new operation point can be found by traversing up the I-V curve based upon load current. The pump design is intended to increase I_{reg_max} to be as large as possible.

4.2.4 Ripple on regulated output voltage

With the regulation applied to the pump output, the output would have ripple (or noise) at the regulation level. The waveform of the ripple is shown in Figure 4-16. This noise is usually due to the finite gain of the amplifier used for regulation sensing as well as the delay in feedback control.

In Figure 4-17, the amplitude of ripple near regulation level is ΔV. t_1 depends on the time to discharge below regulation level, plus the loop delay of the feedback signal and the response of the control circuit. t_2 depends on the characteristic of pump capacitance, pumping clock frequency, feedback loop delay, response of the control circuit, decoupling capacitance, load capacitance, and so on.

Noise reduction is a critical requirement in many applications. Because the regulation is part of the analog circuit design, definitely from the circuit point of view, detailed analysis can be done to improve the gain of the amplifier, to improve the phase margin, and so on. Two unique aspects should be emphasized here that are specific to charge pump noise reduction. One aspect concerns the balancing of power between the output of the pump and that of the loading circuits. The other aspect concerns the capacitive noise-filtering technique.

First, let us analyze the power that can be supplied by the charge

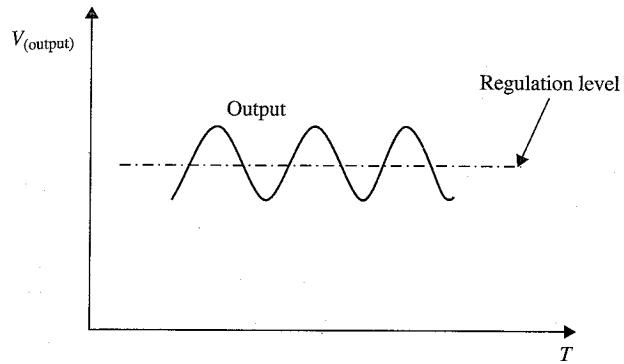


Figure 4-16 Ripple on regulated pump output voltage.

Figure 4-18 shows a generic representation of all the current sources connecting the output of the pump. In the regulation phase, Kirchoff's current law is observed at all times, as shown in Equation 4-30.

$$I_{output} = I_{load} + I_{shunt} + I_{capacitance} \tag{4-30}$$

If at any given time I_{load} and I_{shunt} can be maintained to match I_{output}, and if $I_{capacitance}$ is equal to 0, then the output voltage would not have any noise because the output power of the pump is balanced by the power being consumed by regulation and load circuits.

However, this situation is rare. In many cases, I_{load} is relatively stable. I_{shunt} and $I_{capacitance}$ need to balance the difference $\Delta I = I_{output} - I_{load}$. Due to the delay of feedback response time, $I_{capacitance}$ needs to absorb the difference ΔI right away. The effect is shown for ΔV on the output voltage of the charge pump. If I_{shunt} finally responds, it will contribute to another part of the ΔV change. Ideally, I_{output} should be designed in regulation phase to be near the range of I_{load}. Any variations in

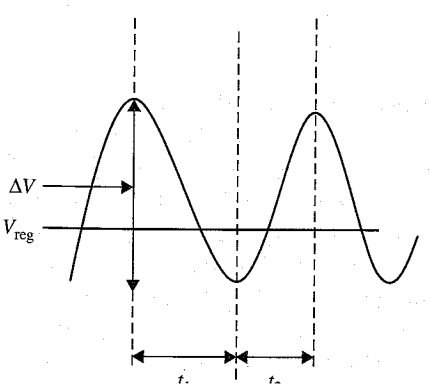


Figure 4-17 Amplitude of ripple.

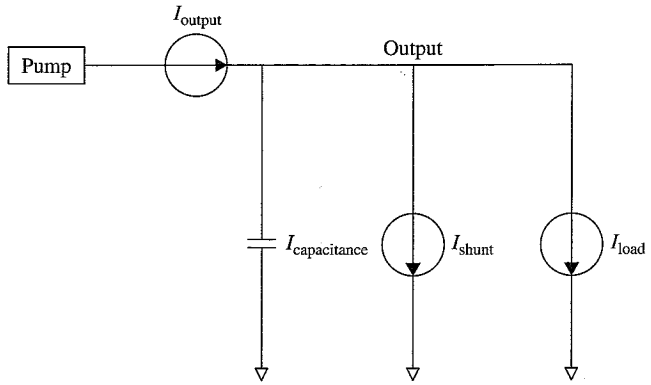


Figure 4-18 Dissection of current consumption on the pump output.

$I_{capacitance}$ and I_{shunt} could be a smaller percentage of I_{output}. This allows the potential noise on the output to be contained within a limited range at first.

In designing the charge pump strength, special design techniques can be applied to vary the output power strength based on the operation demands.[7] For example, in the ramp-up phase, regulation phase, or recovery phase, the output capability could vary to just to meet the demand of the load current. With this approach, the root cause of noise is addressed at first instead of the outcome of the noise due to the mismatching of the powers.

Also, capacitive noise filtering can be a very powerful technique. As shown in Equation 4-30, before I_{shunt} can respond, $I_{capacitance}$ will be the only component that can respond instantaneously. The direct effect is shown in ΔV on the pump output voltage. ΔV has a direct relation to the total filtering capacitance size, as shown in Equation 4-31.

$$\Delta V = \frac{\Delta Q}{\Delta t_{response}} = \frac{\left(I_{output} - I_{load} - I_{shunt}\right)}{\Delta t_{response}} \qquad (4\text{-}31)$$

When the filtering capacitance size is increased, ΔV would be reduced proportionally from the calculation in Equation 4-31. The need of I_{shunt} to reduce output noise in time is becoming less critical. The overall noise on the pump output would be reduced. The drawbacks of this approach include the increasing of the pump ramp-up speed and the possible increasing of the die size. More capacitance on the output node means longer settle time for output. Increasing the filtering capacitance would also increase the layout size. The correct sizing of decoupling capacitance for filtering is important in order to deal with both concerns.

4.2.5 Pump regulation

Regulation of the charge pump is another important concern. Without regulation, the output of the charge pump cannot be determined. Depending on the needs of the application, the implementation of the regulation scheme can be quite different. A common approach of regulation involves some means to sample the pump output voltage and compare the sampled voltage with a known reference voltage through an amplifier. The comparison result is amplified and is fed back to control the pump or supporting circuits in order to maintain the output at regulation level. Many schemes can be used for pump regulation control.

4.2.6 Capacitive divider

One type of commonly used pump regulation scheme is based on the capacitive divider method. This scheme is shown in Figure 4-19. C_1 and C_2 are two known capacitances. Before regulation is enabled, Node *div* has an additional device (not shown in the figure) to initialize the value. It is common to initialize *div* to be 0 V if positive voltage is being regulated on the output.

Once the regulation starts, the initialization devices are shut off. Conservation of the charge is observed on node *div* at all times. Of course, in this discussion, the leakage current on *div* is ignored. After the initialization, node *div* would be coupled up based on capacitive coupling from the pump output. This is shown in Equation 4-32 and Equation 4-33.

$$V_{div}C_2 + C_1\left(V_{div} - V_{output}\right) = 0 \times \left(C_1 + C_2\right) = 0 \qquad (4\text{-}32)$$

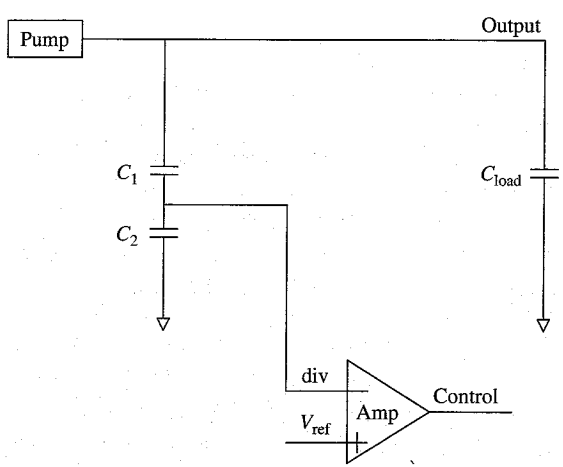


Figure 4-19 Capacitive divider feedback control.

$$V_{div} = V_{output} \times \frac{C_1}{(C_1 + C_2)} \tag{4-33}$$

Equation 4-32 shows the conservation of the charge on node *div* before and after output is being regulated. Equation 4-33 shows how the coupled voltage from V_{output} to *div* can be calculated. In regulation, the divided voltage on node *div* should be equal to V_{ref}, the reference voltage in regulation, as shown in Equation 4-34.

$$V_{div} = V_{ref} \tag{4-34}$$

The final regulation level of the pump output can be calculated based on Equation 4-35. The advantages of this approach are that the feedback speed is very quick and the capacitors do not take any DC current from the output of the charge pump when it is being regulated. As shown earlier, any reduction of the pump output current would allow the pump to be sized smaller, with less overall power consumption. The capacitor divider approach fits into this category.

$$\begin{aligned} V_{ref} &= V_{output} \times \frac{C_1}{(C_1 + C_2)} \\ V_{output} &= V_{ref} \times \frac{(C_1 + C_2)}{C_1} \\ V_{output} &= V_{ref} \times \left[1 + \frac{C_2}{C_1}\right] \end{aligned} \tag{4-35}$$

There are indeed a few disadvantages to using capacitance for feedback control. First, the capacitance value may shift with the process or biasing voltage. If capacitance is built based on oxide, the oxide thickness could vary from die to die, and from wafer to wafer. If capacitance is based on MOSFET gate capacitance, different biasing voltages across MOSFET would cause gate capacitance to go through different regions. In this case, the value of gate capacitance would change. Second, the parasitic capacitance from wiring and connection could change the actual values of C_1 and C_2 from design. Any deviation will cause the final regulated output to be shifted from the target values. Third, the capacitive divider is based on charge conservation on the middle node in between two capacitances. This middle node needs to be initialized to a certain value through devices. Any leakage currents through the junction diode or subthreshold current would make

conservation of the charge invalid and cause the regulation level on output to be off. The maximum regulation time allowed is based on the tolerance of the regulation level. Operations exceeding this maximum regulation requirement should always refresh the capacitor divider circuits.

4.2.7 Resistive divider

Another type of commonly used regulation scheme is based on resistor divider feedback control. Resistors in serial divide the high output voltage, resulting in a lower voltage that can be safely applied to low-voltage transistors in an amplifier. V_{div} is compared with the reference voltage and generates the feedback control signals. This scheme is shown in Figure 4-20. At any given time, V_{div} observes the relation given in Equation 4-36. In the regulation phase, the divided voltage V_{div} should be equal to V_{ref}. The level of pump output, V_{output}, in regulation can be calculated based on Equation 4-37.

$$V_{div} = V_{output} \times \frac{R_2}{(R_1 + R_2)} \tag{4-36}$$

$$\begin{aligned} V_{div} &= V_{ref} \\ V_{ref} &= V_{output} \frac{R_2}{(R_2 + R_1)} \\ V_{output} &= V_{ref} \frac{(R_1 + R_2)}{R_2} \\ V_{output} &= V_{ref} \left[1 + \frac{R_1}{R_2}\right] \end{aligned} \tag{4-37}$$

There are advantages and disadvantages to using a resistor divider in a feedback scheme.

The first advantage of the resistive feedback regulation scheme is that the divided voltage for comparison is based on resistor ratio. This divided voltage does not change if the absolute value of the resistor changes due to doping variation, diffusion variation, and so on. The second advantage of the resistor divider scheme is that the temperature coefficients of resistors are cancelled in Equation 4-36 and Equation 4-37. The temperature coefficient of V_{div} would not have the components from the resistor divider. The third advantage is that this scheme can be easily implemented without too much worry concerning

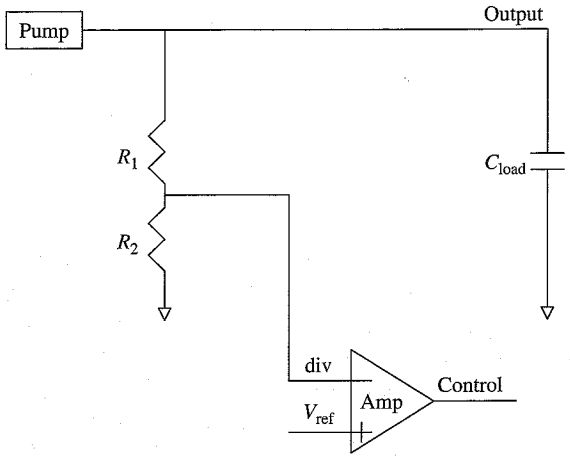


Figure 4-20 Resistive divider feedback control.

parasitic resistance. As long as the sheet resistance of the resistors is large, the impact from parasitic resistance due to metal and contacts is relatively small.

Now the designer should also be aware of some inherent disadvantage associated with the resistor divider type regulator. The first disadvantage of the resistor divider feedback scheme is the power consumption. With resistors connected to the output of the charge pump, it will always consume DC current in regulation. As mentioned earlier, charge pump power efficiency is not very good. If the output current is divided for regulation purposes, the remaining current that can be supplied to the load circuit is reduced. The size of the charge pump may need to be increased to compensate for this extra DC regulation current. Normally the resistance of the divider circuit has to be sized up to allow minimum current consumption while not hurting the layout area due to larger resistance. The second disadvantage is that the resistor divider network is slow in terms of delay in feedback response. The feedback path is based on an *RC* network. All resistors have some parasitic capacitance associated with them. Because resistance needs to be sized up to reduce unwanted DC current being wasted, the *RC* delay through the feedback path is normally very large. That is why the resistor divider feedback scheme is slower in response time compared with that of the capacitive divider scheme and could contribute to large output noise in regulation.

4.2.8 MOSFET biased type regulator

A MOSFET biased type regulator can generally be used for pumps generating relatively lower voltages, and it is best suited for negative voltage

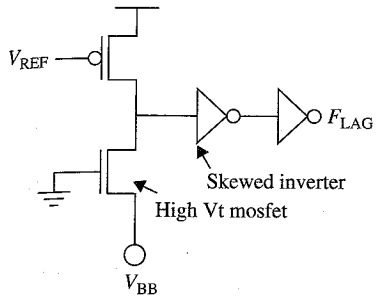


Figure 4-21 A simple biased MOSFET type regulator.

charge pumps. Consider the circuit in Figure 4-21. A high V_t NMOS device's source is tied to the output of the pump generating a negative output voltage. As V_{BB} starts to go low and reaches a preset voltage, the skewed inverter turns on and causes the output to toggle. A controlled V_{ref} voltage source can modify the pump regulation voltage.

4.2.9 Which scheme should be used in design?

Choosing which scheme is best suited for a specific design depends on the requirement of the design. The direct impact from regulation schemes involves regulation speed and the noise on the pump output in regulation.

For example, suppose a charge pump design has the following requirements: to ramp up the output load in 0.3 μs, to supply 1 mA of current at 5 V, and to have only 100 mV ripple on the output node in regulation. The capacitive divider feedback scheme may be a good candidate in terms of regulation speed and noise control. A capacitive divider can sample the output voltage almost instantaneously. The feedback signal may control a shunt current path to channel any extra current away from the output. If output current being delivered is lower than 1 mA, the feedback path can disable the shunt current path. The pump would be fully on because a constant current of 1 mA is required at the output.

As another example, let us say the charge pump has to ramp up the ouput volage to 10 V in 5 μs and does not need to supply any DC current in regulation. In this case, both capacitive and resistive approaches can be chosen to sample and stop the pump when it hits the 10 V regulation level. If a capacitive divider is used, there is a timing concern because intermediate nodes of capacitors would have junctions of transistors. The leakage current over long periods of time can cause the regulation to deviate from the target level. Another issue arises when the accuracy is affected due to extra parasitic capacitances. The size of regulation capacitance needs to be relatively large to minimize this effect. Using the resistive divider scheme in this case will not have

the same concerns. All errors due to parasitic capacitance and leakage current do not exist.

4.3 Pump Power Consumption

Power consumption of the pump itself is another critical specification that needs to be understood. While transferring the charge from input to output, the charge pump has to consume power in clock switching and capacitive boosting. Sometimes the charge pump only delivers a very small fraction of the power to the output; the total amount of power consumption by itself could be 20X or more. For example, let us say a charge pump with a 3.0 V power supply is capable of delivering 100 μA at 10 V to the output by design requirement. To be able to generate this output power, the charge pump in general may be consuming 2 ~ 3 mA of current from the chip power supply. The power consumption specification affects many design considerations, such as charge pump performance, loading circuit, power supply voltage, device threshold voltages, architecture, pump regulation, and so on. It should be studied carefully in the early stages of design to have a realistic power budget on the final chip.

4.4 Die Size of the Charge Pump

Every circuit being put on silicon will take some amount of die area. A charge pump especially consumes area in many cases. As the technology is scaled down, the power supply voltage is scaled down, too. For the pump design, this trend can cause serious impact to the sizing of the charge pump.

As analyzed at the beginning of this chapter, the supply voltage is an important design factor. The area of a pump with a 3 V power supply would be very different from a pump with a 1.8 V power supply while providing the same pump performance. The layout area of the 1.8 V design can be 5X or more than that of the 3 V design. Even if technology is not scaled down, the application will require a die size reduction to maintain profitability. Under this scenerio, every circuit being laid out on chip has to be as small as possible. Based on the optimization theory, if every circuit could be optimized to the minimum size, the total chip area summed should be the smallest in size. Charge pumps need to be optimized.

The size of a charge pump on silicon is determined by many parameters, such as performance requirement, device threshold voltage, architecture, regulation level, and so on. Optimizing the sizing requires the optimization of every parameter to deliver the best sizing, performance, and power.

4.5 Conclusion

This chapter covered some of the fundamental design criteria for charge pump implementation. They are essential to help designers understand the basic requirements, and the capability and the limitations of devices. With all these constraints in mind, circuit designers can have a global view of how the charge pump should be designed as a whole system. It allows most suitable pump architecture to be chosen and realistic implementation to be planned in order to achieve the final design goals. This process is key to any successful charge pump design.

References

1. Umezawa, A., S. Atsumi, M. Kuriyama, H. Banba, Imamiya, K. Naruke, S. Yamada, E. Obj, M. Oshikiri, T. Suzuki, and S. Tanaka, "A 5-V-only operation 0.6-μm flash EEPROM with row decoder scheme in triple-well structure," *IEEE Journal of Solid-State Circuits*, Vol. 27, pp. 1540–1546, 1992.
2. Witters, J. S., G. Groeseneken, and H.E. Maes. "Analysis and modeling of onchip high-voltage generator circuits for use in EEPROM circuits." *IEEE Journal of Solid-State Circuits*, Vol. 24, pp. 1372–1380, October 1989.
3. High-Voltage Generator Circuits for Use in EEPROM Circuits. *IEEE Journal of Solid-State Circuits*, Vol. 24, No. 5, October 1989.
4. Oxide Breakdown http://www.semiconfareast.com/oxidebreakdown.htm.
5. Razavi, B. *Design of Analog CMOS Integrated Circuits*. McGraw-Hill, New York, 2001.
6. Gray, R.P., J.P Hurst, H.S. Lewis, and G.R. Meyer. *Analysis and Design of Analog Integrated Circuits*, Fourth Edition. John Wiley & Sons, New York, 2001.
7. Pan. High voltage ripple reduction. U.S. patent 6,734,718.

Chapter

5

How to Design a Basic Charge Pump

Before starting to design a new charge pump, we need to analyze the design requirements, which will in turn help us define the major pump parameters. In general, the advantage gained by optimizing some of the parameters may be detrimental to other ones, and vice versa. For example, increasing the number of pump stages (n) will result in a higher output voltage, but it also increases the internal impedance of the pump and thus cuts down the output current drivability. For a particular number of stages, the output drive current can be increased by increasing the size of the capacitors and the width of diode-connected MOSFETs. But the larger geometry MOSFETs and capacitors will also increase the wire routing length of the clock signals and other internal connections, which will increase the resistance and the parasitic capacitances on those nodes. As we saw earlier in Equation 3-11, parasitic resistance and parasitic capacitance play a significant role in determining the output voltage and the charge pump efficiency. Hence, the number of pump stages (n) needs to be carefully chosen so as to orchestrate an optimum pump performance.[1]

Take another charge pump specification—the output voltage ramp-up time or the recovery time requirement. In general, charge pumps used in EEPROM and other applications that need a high-output voltage require the regulated output voltage to be available within a stipulated time after the pump is enabled. Further, most of these applications also require that during the time the pump is being enabled, if due to any sudden loading the pump output droops significantly from its steady regulated value, the charge pump should be able to pull up the output voltage back to the regulated level within a defined time, also called the *recovery time*. The factor that directly dictates the output ramp-up time and the recovery time is the

pump's output drive at any given voltage level. As mentioned, increasing the output drive current by sizing up the boosting capacitance and diode-connected devices has a direct effect on the pump's output voltage capability and its output efficiency. Another way to improve charge pump output current drivability is by increasing the pump clock frequency.[2] As could be seen in Equation 3-14a from Chapter 3, increasing the pump frequency can synthetically reduce the internal impedance of the pump. It effectively increases the output current drivability at any given possible regulated voltage level, thus reducing the ramp-up and recovery times. However, it may not be so easy to operate the pump at a very high frequency. In fact, the maximum frequency of pump operation is determined by the minimum time required to transfer charges from one stage to the next, completely or close to 90% of the maximum value. If the pump is operated at a frequency that exceeds the optimum limit, it will cause inefficient charge transfer and therefore poor overall charge pump performance, not to mention higher power consumption $CV_{DD}^2 f_{osc}$. Operating the pump at a frequency that is lower than the optimum limit means the charge pump is not being utilized at its full efficiency level. One of the parameters that limits the higher frequency of operation is the stage-to-stage parasitic capacitance. Optimum capacitor and MOSFET sizes and innovative layout techniques must be used to reduce parasitic capacitances.

In this chapter, we will discuss each of the major parameters in detail and consider how to create trade-offs between one and the others. We will also touch on layout implementation, floor planning, clock distribution, and regulator design—all of which help us start designing a charge pump.

5.1 Charge Pump Specifications

The charge pump, such the one listed in Table 5-1, should have the specifications available before its initial design. In general, part of the specifications will be derived from the system specifications, in which different

TABLE 5-1 Charge Pump Specifications

Specifications	Name	Sample Value
Steady output voltage	V_{out}	16 V
Output voltage ripple	ΔV	0.1 V
Output current requirements	I_{out}	30 μA
Output load capacitance	C_{load}	20 pF
Output voltage ramp-up time	T_{ramp}	10 μS
Output voltage recovery time	T_{reco}	2 μS
Maximum pump layout area	$Area_{pump}$	—
Pump power supply voltage	V_{DD}	2.5 V
Average pump current consumption	I_{pump}	1 mA

budgets are defined for different blocks of the system. The other part of the specification will come from device and circuit requirements.

5.1.1 Output voltage

The first step in designing the charge pump is to determine the required output voltage. The voltage required may be a simple $V_{DD} \pm \delta$ to supply the required back-bias voltage for DRAM memories or a very high output voltage, on the order of 25–30 volts, as required in flash memories. The number of stages (n), along with the pump clock amplitude, directly dictates the final output voltage. The simple Dickson charge pump's equation can be used to determine the number of stages required to attain the required output voltage:

$$V_{out} = NV_{\phi} + V_{in} - (N+1)V_D \tag{5-1}$$

For example, if the clock amplitude (V_{ϕ}) is 2.5 V, the input voltage (V_{in}) is 2.5 V, and V_D is 0.2 V (the V_t of the diode-connected NMOS transistor, with V_t assumed to be constant for now), then to generate a V_{out} of, say 16 V, we need at least six stages ($N = 6$).

However, this is not enough. In general, the MOSFET at the sixth stage will exhibit V_t, which could be 1–2 volts or higher than expected, depending on the MOSFETs' characteristics.[4] Hence, the output voltage that can be delivered may be much less than 16 volts. Another problem is that we have not yet considered the effects of parasitic capacitance, which will consume charge and increase internal impedance. Finally, and most important, because the pump output is very exponential in nature toward its open circuit output voltage, it is never advisable to design the pump such that its required output voltage is very close to its maximum output voltage, in which case the pump will be deemed too weak.[5] Therefore, it is good practice to design a pump in which the required output voltage is about 70%–80% of its maximum output voltage capacity, and to use an active regulator to regulate the charge pump output.

Continuing with this example, we will start with a Dickson charge pump that has eight to nine stages (i.e., a pump that can theoretically deliver the potential output voltage that is about 30%–40% higher than the specification requirement).

5.1.2 Current drivability

The current drivability is the next most important factor to look for. In general, the pump output current decreases linearly with the increased pump output voltage. The power efficiency of the charge pump decreases

with increased pump output voltage. The internal impedance of the charge pump may be represented as follows:

$$R_s = \left(\frac{N}{C + C_s}\right) \times \frac{1}{f} \tag{5-2}$$

To achieve a higher output voltage, the number of pump stages N needs to be increased, but as we can see from the previous equation, this also increases the effective internal resistance. Therefore, to achieve higher output current at a high pump output voltage, two important factors—the boosting capacitor size and the pump clock frequency—need to be increased. Also, another factor that is not accounted for is the simple equation—the MOSFET threshold voltage, V_t, as a function of source bias. As the absolute value of the output voltage increases, the sources of diode-connected devices will increase. As more pump stages are added, more V_t drops through the pump stages would occur. The pump internal impedance will increase. Taking V_t into account, a modified internal resistance may be represented as

$$R_s = \left(\frac{N}{C + C_s}\right) \times \frac{1}{f} \tag{5-3}$$

where R_{vt} represents the effective internal pump impedance imposed due to the gradual increase in V_t of the diode-connected devices along the chain, with the last stage exhibiting the highest absolute V_t.

Higher output drive current results in the faster output ramp-up and the quicker recovery times. Next, we will calculate the initial value of the pump boosting capacitance, C, in a quick two-step process:

1. Because the output load capacitance is already known, with the output ramp-up time requirement (defined in the pump specification), we can easily calculate the linear current required to charge up the output node to the regulation level within the given amount time.

   Continuing with the previous example, assuming $V_{out} = 16$ V, $T_{rampup} = 10$ µs, and $C_{load} = 20$ pF. The linear current required to charge up the pump output can be calculated as follows:

$$i = C_{load} \frac{V_{out}}{T_{rampup}} = 20E^{-12} \frac{16}{10E^{-6}} = 32\,\mu A \tag{5-4}$$

   Hence, the pump should be able to deliver a steady linear current at a minimum of 32 µA. Note that, in general, the output drive current must be a specification for the charge pump.

2. Now, armed with this information, we need to focus our attention on to the last stages of the Dickson charge pump and consider its operation with the two phases of the clock.

At steady state during regulation, assume the pump is delivering a steady current (I_{out}) to the output while maintaining a constant V_{out} (see Figure 5-1). In this situation, I_{out}, the current consumed on the output matches exactly the current delivered by the pump. Extending this concept one step inside, the charge dumped on node 5, through the diode M_5 from preceding stage, is equal to the charge lost through diode M_6 within every clock cycle. This charge gain and loss is manifested as voltage increase and decrease on node 5.[6]

Assuming a long clock cycle, during the second half of the clock, the voltage drop on node 5 due to the output current I_{out} is equal to the voltage gain ΔV due to the positive phase of V_{ob}. Assuming ΔV is 800 mV

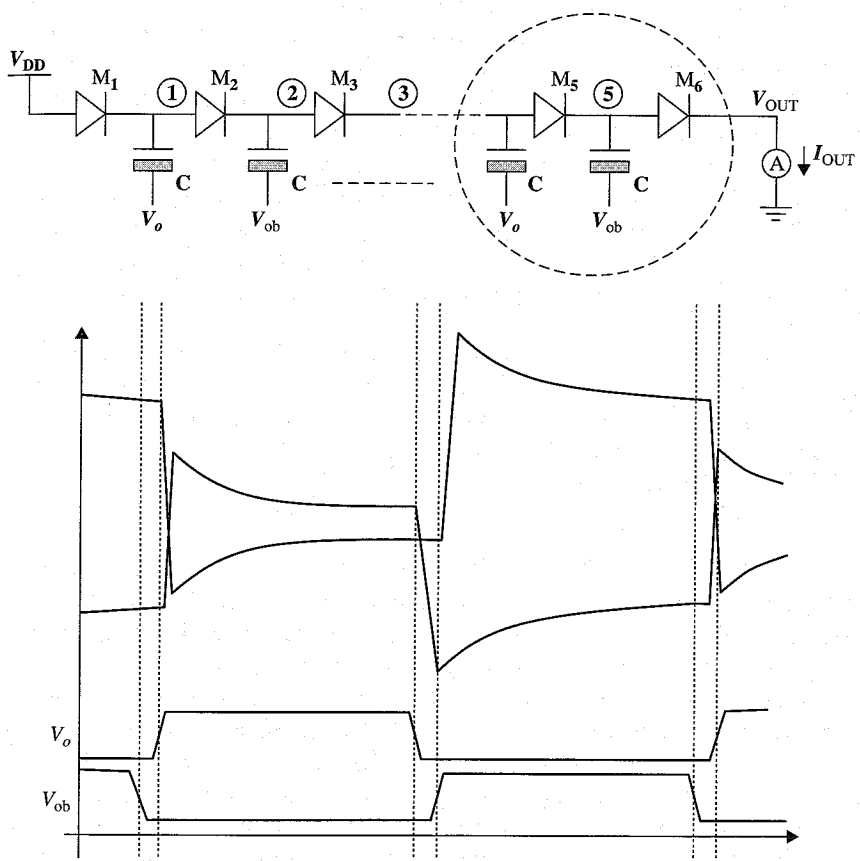


Figure 5-1 Charge pump's internal voltages.

(a typical approximation for the last stage) and the frequency of pump operation is 10 MHz, and because we have already calculated I_{out} from the previous step, we can use the same equation to determine an initial value of C:

$$I_{out} = 32\ \text{uA} = C\frac{\Delta V}{\Delta t} = C\frac{800\times10^{-3}}{100\times10^{-9}} \qquad (5\text{-}5)$$
$$\Rightarrow C = 4\ \text{pF}$$

Therefore, we can start our simulations with our initial assumption of 4 pF for all pump boosting capacitors. After observing the initial set of simulations, the size of the capacitors can be tweaked to meet different specifications. In general, the initial capacitance assumption is on the higher side and will come down, as the design is being optimized. One more thing to keep in mind while determining the output drive current is the size of the diode-connected MOSFET. In general, the MOSFET width should not be too large because that will increase source/drain parasitic capacitances and will also increase signal routing lengths. But the MOSFET width should not be made too small either because that will limit the MOSFET's I_{DS}, which in turn will hinder quick charge transfer from one stage to the next. In general, the initial width of the MOSFET can be calculated from the MOSFET's I_{DS}—versus V_{DS} curves, with a particular width chosen so that it meets the needed I_{out} current at saturation. Again, the MOSFET length should be as minimal as possible, but enough to meet all worst-case, device-related reliability parameters.

The current drivability issue is never complete without mentioning the impact of the circuit's capacitor and diode sizes on the overall pump performance. Care must be taken to keep the device sizes smaller, reduce the layout area, and do creative layout floor planning in order to reduce unwanted parasitic.

5.1.3 Output ramp-up time and recovery time

The pump can be modeled as a voltage source with an internal impedance of R, which in turn limits its output current drivability. The output ramp-up time and voltage recovery time are two important benchmarks of the charge pump performance. Most applications require the output to ramp up fast when a particular operation is initiated—an operation that requires the high-voltage output of the pump. As in an EEPROM, the pump voltage is required when a program operation or an erase operation is chosen. Hence, the pump needs to ramp up the output fast enough to complete the required operation in the shortest possible time in order to meet program or erase speed requirements.

The output voltage recovery time is another important measure of the pump's capability. The pump's output current driving capability is inversely proportional to the regulated output voltage. When the pump output is maintaining a steady output, high-voltage state, sudden output charge sharing due to rapid switching activity or spike of load current will cause a droop in the output voltage, which needs to be replenished within the given "recovery time." Designing the pump to operate at 70%–80% of its maximum capacity will generally guarantee this process.

For a pump with "n" number of stages, the output ramp-up time or recovery time can be controlled by the size of the capacitors and the frequency of operation. A larger capacitor will allow more charge to be transferred in each clock cycle and thus cause a faster increase in output voltage. Higher frequency will expedite this process if the pump is capable. With today's process, in which it is possible to push the ring oscillator frequency to the gigahertz range, the frequency of charge operations is not a speed-critical factor. As mentioned earlier, the maximum frequency of charge operation is determined by the pump architecture, the circuit design and the layout-related parasitic. The range of pump frequency is determined by the application and design requirements. Optimization is needed through design iterations. Further, the clock frequency and the pump performance should be verified with parasitics extracted from the layout (i.e., with back annotated simulation results).

5.1.4 Power consumption

With the rising demand in handheld devices and a requirement for low-power operation, power consumption (or power efficiency) has been one of the significant factors in charge pump design. Traditionally, charge pumps have been the culprit in the chips that incorporate them because charge pumps exhibit lower efficiency at higher output voltage, which makes them take a significant bite in the overall chip power consumption. In reality, the efficiency for charge pump (generally the ones generating very high voltages) could vary between 5%–50% depending on the particular designs.

Even though increasing the capacitor size or the operation frequency can directly increase the pump performance, the power consumption will create the most significant drawback. As the capacitance (C) is increased, the parasitic capacitance (C_s) also increases, which will effectively limit the highest frequency of operation and also cause a large $CV_{DD}^2 f_{osc}$ dissipation. Also, driving this large capacitor requires large drivers, which effectively increase the operating current. There are many design tricks for reducing the $CV_{DD}^2 f_{osc}$-related power consumption, such as a multistep frequency control, where the pump is operated at a higher frequency, initially to ramp up to the required level,

and then switched over to a lower frequency to maintain that voltage level under regulation. A more sophisticated method may be to use a VCO-type of oscillator. The oscillator frequency varies as a function of pump output voltage.

A final factor is the contribution of the voltage regulator in power consumption. It is common that resistive voltage dividers are used to sense the output voltage and regulate the charge pump output. To have a fast regulation response time, a lower resistor value may be chosen, which in turn has larger DC regulation current. Therefore, the pump has to deliver this extra current in addition to the load current. In this case, the charge pump will consume more DC power.

5.2 Pump Clock Source

The clock source for a charge pump can be derived from the system clock source, if available, or a new clock source may be built to support the pump activities. A simple clock source may be built using a ring oscillator, as shown in Figure 5-2.

It consists of an odd number of inverting gates connected in a ring pattern. The NAND gate has been used to enable the oscillation. The oscillation frequency can be expressed as

$$f = \frac{1}{N\left(t_{phl} + t_{plh}\right)} \tag{5-6}$$

where t_{phl} is the signal propagation delay for an input rising edge and t_{plh} is the signal propagation delay for an input falling edge. N is the number of stages of the oscillator. For a properly sized gate, $t_{phl} = t_{plh}$. Hence, the frequency of oscillation can be expressed as

$$f = \frac{1}{2NT_D} \tag{5-7}$$

where T_D is the propagation delay, which equals t_{phl}, which in turn equals t_{plh}.

Because the ring oscillator needs a higher number of stages to produce a lower frequency, specifically with faster semiconductor processes, a standard practice is to allow the oscillator to operate at a high frequency

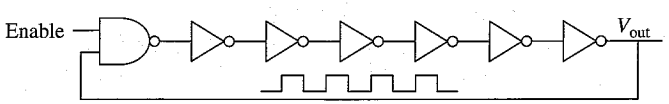


Figure 5-2 A seven-stage ring oscillator.

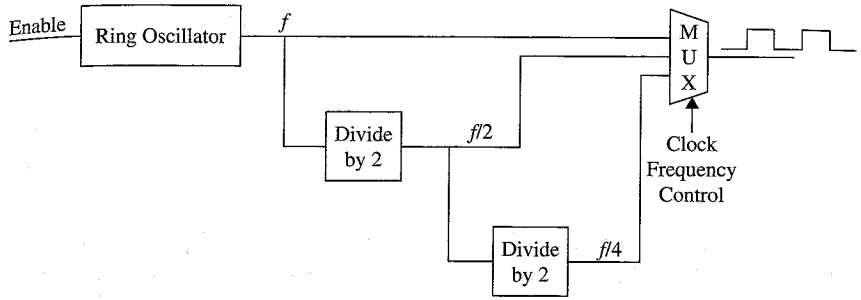


Figure 5-3 Variable frequency clock generator.

and then use a series of dividers to step the frequency down to a required level, as shown in Figure 5-3.

Even though the simple ring oscillator can do the job, it has huge variation—over the process, temperature, and supply voltage variations. The oscillation frequency may vary as much as 30% over the typical operating condition. Keeping the oscillation frequency constant is important because for applications requiring precise pump output voltage, a changing oscillation frequency will cause a change in output voltage noise, power consumption, and output recovery time, etc. To get around this problem, a more-sophisticated, current-controlled ring oscillator may be used.

As shown in Figure 5-4, the current-controlled ring oscillator uses the same inverting gates connected in a ring pattern, with the exception that the current through these inverting gates is controlled by a reference voltage source. This voltage source may actually be derived from a band gap reference voltage generator to produce extra stability across

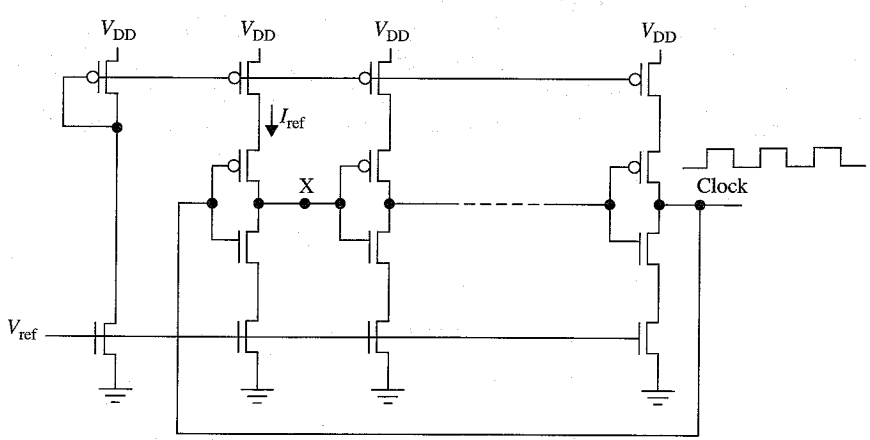


Figure 5-4 Current starved ring oscillator.

temperature and process variations. Assuming the capacitance on node X in Figure 5-4 is C_{stage}, the frequency of operation can be derived as follows:

$$f = \frac{I_{ref}}{NC_{stage}V_{DD}} \tag{5-8}$$

Another significant advantage of this oscillator is that compared to the simple ring oscillator, this oscillator requires a fewer number of stages to attain the same oscillation frequency, and the oscillation frequency can also be controlled linearly by varying the V_{ref}.

5.3 Regulator Design

Even though some simple low-voltage charge pumps generating voltages in the range of $2V_{DD}$ or $V_{DD} + V_t$ may not need a regulator, most of the high-voltage charge pumps almost always need one. Two major types of regulator are generally used: the resistor divider–type regulator, and the capacitor divider–type regulator. Details about these types of regulators were provided in the previous chapter, so we will not discuss them here. Based on the application, the circuit designer should choose the type of regulator that is most suitable for the design.[7–9]

5.4 Non-overlapping Clock Generator

In general, the pumps will need a non-overlapping clock for optimum functioning. For a 2-phase Dickson charge pump, the two clock phases may be generated by using one clock source and its inversion as another clock source, as shown in Figure 5-5.

With this type of clock source, consider what happens during time t_2 for a Dickson charge pump in which ϕ_b is connected to stage $n-1$ and ϕ is connected to stage n. During the time $t_1 - t_2$, stage n-1 is charging up stage n, and if the clock time period is relatively small, the positive going edge of ϕ will cause a quick cutoff of the charge transfer from stage

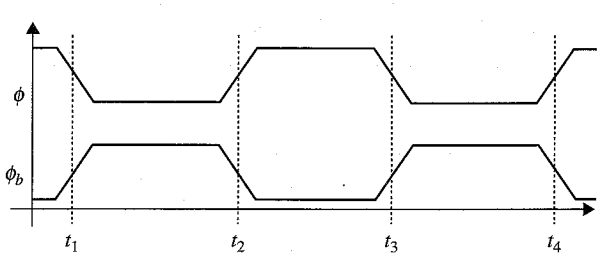


Figure 5-5 A non-overlapping clock.

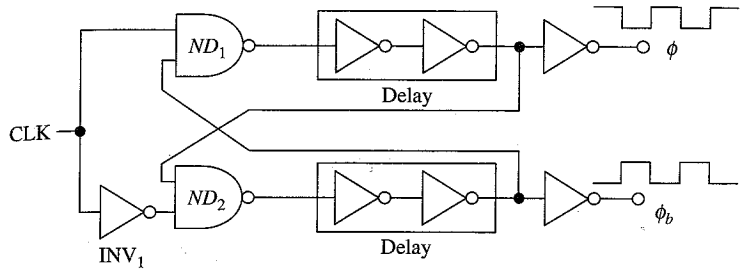


Figure 5-6 A non-overlapping clock generator.

$n-1$ to stage n, thus resulting in incomplete charge transfer and, hence, reduced pump performance. To avoid this scenario, a non-overlapping clock is better suited for a Dickson charge pump. A non-overlapping clock can be generated from the original clock source by a simple circuit (see Figure 5-6 and Figure 5-7).

This circuit takes a clock signal and generates a 2-phase non-overlapping clock. The operation is based on the fact that the falling edge of the input clock passes immediately through the NAND gate ND_1, while the rising edge has first to propagate through the other NAND gate and the cascaded delay element. The resulting signals, ϕ and ϕ_b, have a non-overlapping time equal to the sum of the delays at the NAND gate of the delay element. The delay element is generally made using an even number of inverters. When driving long clock lines, additional buffer stages need to be used to maintain sharp output clock rise and fall times.

Even though the non-overlapping clock generator shown in Figure 5-6 is used extensively in many circuit implementations, it has a small imperfection—the one phase of the clock passes through an extra inverter, INV_1, and the other phase does not. This problem is generally manifested as a nonsymmetric pulse width of the complementary clocks when the pump is operated at a very high frequency. Even though this is not a problem for a normal Dickson charge pump, it may be a problem for

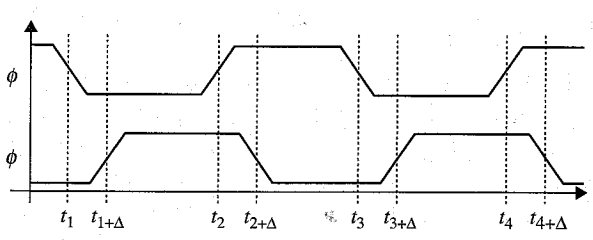


Figure 5-7 A non-overlapping clock generated from the non-overlapping clock generator.

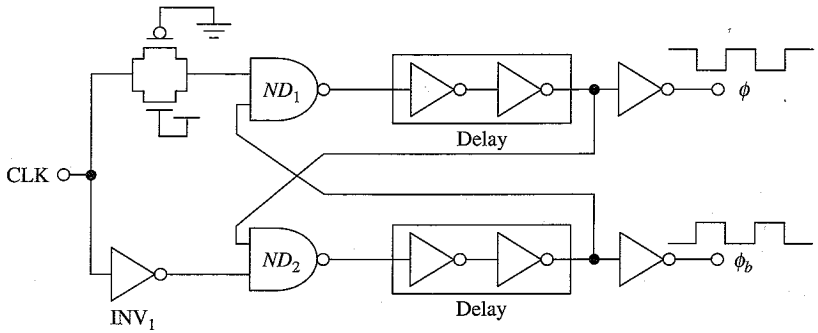


Figure 5-8 A perfected non-overlapping clock generator.

more sophisticated pump designs, such as a 4-phase charge pump, which will be discussed in later chapters. One quick way to fix this issue is to use a passive matching delay element, as shown in Figure 5-8. In this case, the PMOS-NMOS combination pass gate has the same geometry as the PMOS-NMOS MOSFETs in the INV_1 gate to produce symmetric clock output signals.

The initial amount of non-overlap can be kept between 1 ns to 5 ns and later tweaked to provide better results after RC back-annotated simulations are performed. Because the non-overlap time is generally derived from gate delays, this non-overlap will vary with different process, voltage, and temperature conditions and therefore different corner simulations need to be done before optimizing the non-overlap.

5.5 Cross-coupled Voltage Doubler Design

During the course of the design of the charge pumps, clocks must be created with amplitudes higher than the power supply voltage of V_{DD}, such as $2 \times V_{DD}$, $3 \times V_{DD}$, and $4 \times V_{DD}$. As explained later, using clocks of higher amplitude allows for fewer pump stages to be in serial and thus reduces the internal impedance of the charge pump. This approach makes the pump design much more efficient when the power supplies are scaled down. It also allows the charge pump to be capable of high frequency operation when the dynamic threshold voltage of the diodes is very high at fast clock rate.[3,10–12]

A simple cross-coupled voltage doubler design is shown in Figure 5-9. V_o and V_{ob} are the non-overlapping clock inputs, while *Out* and Out_b are the $2 \times V_{DD}$ amplitude clock outputs. To understand the functionality, let us look at what happens when V_o goes high. First, it causes MOSFET p_{1b} to shut off. Then, N_{1b} turns on, pulling down the output to ground, while MOSFET N_{2b} turns on, which precharges node X_{1b} close to $V_{DD} - V_t$. In the next clock phase, when clock V_o goes low, MOSFET N_{1b} and

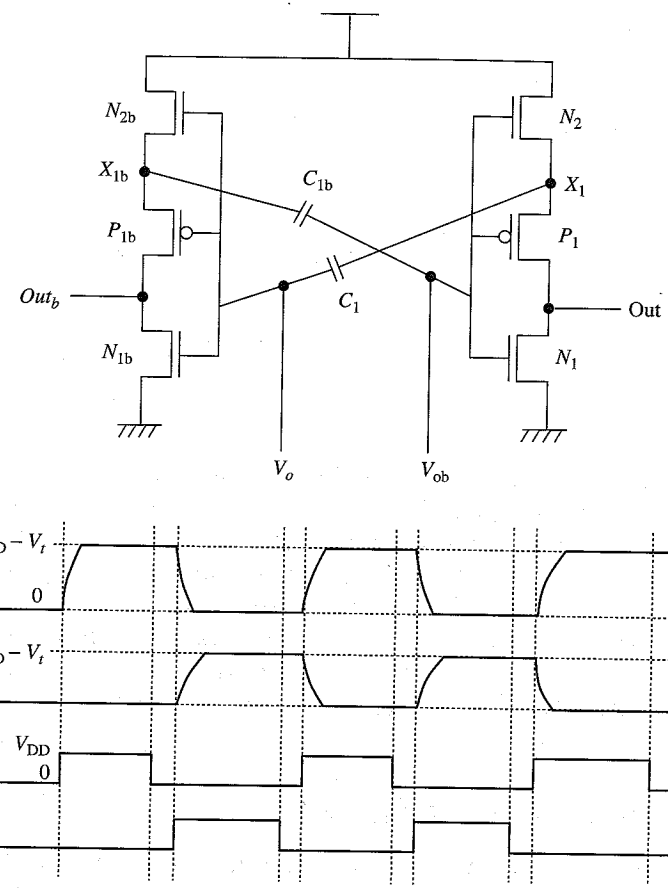


Figure 5-9 A cross-coupled voltage doubler.

N_{2b} turn off and MOSFET p_{1b} turns on. Next, the clock V_{ob} goes high, pushing node X_{1b} to $2V_{DD} - V_t$, which is passed through p_{1b} to the output. A symmetrical operation happens on the right side, resulting in two non-overlapping clocks of higher amplitude. Note that the $-V_t$ term will always remain, which results in a small inefficiency. Using MOSFETs of lower threshold voltage for N_2 and N_{2b} will result in improved performance.

If clock amplitudes of higher potential than $2V_{DD}$ are required, the cross-coupled voltage doubler shown in Figure 5-9 may be cascaded together to produce a clock of amplitude $3V_{DD} - 2V_t$, and three such stages may be cascaded together to produce a clock of amplitude $4V_{DD} - 3V_t$. Figure 5-10 shows the schematic of two stages cascaded together to produce a $3V_{DD} - 2V_t$ clock. Even though this process seems simple, the capacitances at nodes X_1 and X_{1b} need to be kept at a minimum, because

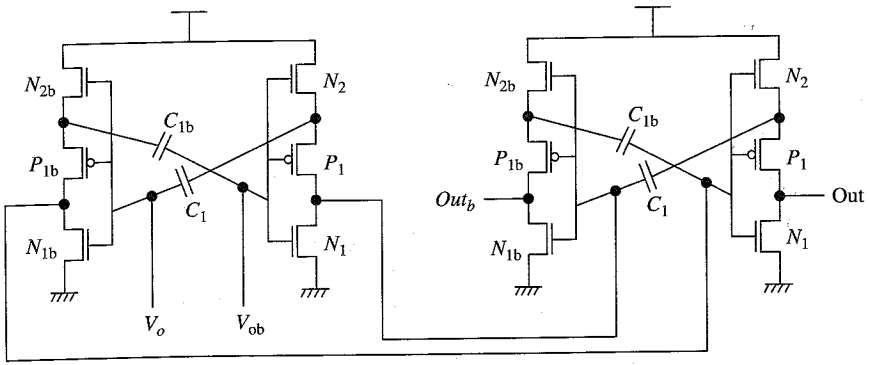


Figure 5-10 A cascaded clock amplitude multiplier.

when the switches from low to high and couples up node X_1 close to the $2V_{DD}$ level, the charge is actually being shared between capacitor C_1 and the parasitic capacitances. As two stages are cascaded together to produce higher voltages, this charge-sharing problem will really create inefficiency, thus reducing the magnitude of the actual output voltage. Further, NMOS transistors at higher potential will exhibit higher threshold voltage due to high source bias, adding to the circuit's inefficiency. Proper care should be taken to select the actual clock amplitude and the circuit layout.

5.6 Logical Effort for Clock Buffer Sizing

Everyone who has gone through a chip design phase has faced the same problem: How to size a signal propagation path through different gates to drive a heavy load or to drive a signal over a long signal line? In other words, how to properly size a series of logic gates to produce the least propagation delay? The *logical effort method*[13] allows us to find optimum solutions for these issues. The logical effort method has the following merits:

- It uses a simple model of delay.
- It allows back-of-the-envelope calculations.
- It helps make rapid comparisons between alternatives.
- It emphasizes remarkable symmetries.

In this chapter, we will quickly detail a simple logical effort methodology and find ways to use logical effort to help make better decisions for clock buffer sizing. You will see that the best performance can be achieved by using optimum fanout for each stage in the logic chain.

You will also see that even though the widely published optimum fanout is "e" (=2.71), the real optimum for CMOS circuits ranges from 3.0 to 4.5, depending on the type of gate.

Logical effort expresses delays in a process-independent unit:

$$D_g = D_{abs} / \tau \tag{5-9}$$

Here, D_{abs} is the actual delay of the gate in units of seconds, τ is the process multiplication factor (in seconds), which represents the delay of a minimum inverter driving another minimum inverter in some process. $\tau \approx 50$ ps for a 0.8-μm process, and $\tau \approx 12$ ps for a 0.18-μm process. D_g is the absolute process-independent delay.

Next, the gate delay can be expressed as

$$D_g = FG + p_{gate} \tag{5-10}$$

where G is the logical effort. In general $G = 1$ for an inverter, by definition. For a 2-input NAND gate $G = 4/3$ and for a 2-input NOR gate $G = 5/3$.

Next, F is the electrical effort $= C_{out}/C_{in}$. The term F can also be expressed as the ratio of output to input capacitance, which is also termed *fanout*.

Finally, p_{gate} is the parasitic delay of the gate driving no load. p_{gate} is set by internal parasitic capacitance. $p_{gate} = 1$ for an inverter. For a 2-input NAND gate $p_{gate} = 2$ and for a 2-input NOR gate $p_{gate} = 2$.

Logical effort, G, is defined as the ratio of the input capacitance of a gate to the input capacitance of an inverter delivering the same output current. Figure 5-11 shows the logical effort calculation for the three most common gates. Note that the product of electrical effort and logical effort, FG, is known as the *stage effort*.

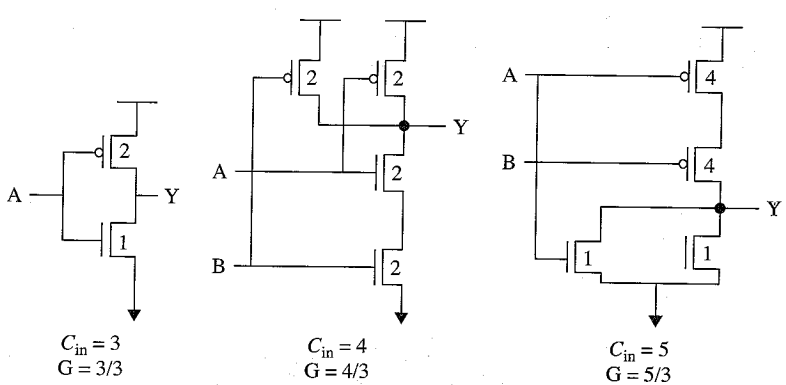


Figure 5-11 Logical effort of common logic gates.

The inverter has a logical effort of 1, by definition. Logical effort for any gate can be determined in the following ways:

- It can be measured from delay versus fanout plots.
- It can be estimated by comparing the gate input area (i.e., by finding the sum of widths, provided the channel lengths are equal) of all the MOSFETs connected to the input and dividing this by the sum of widths of the PMOS and NMOS transistor of an inverter. Figure 5-11 shows the calculation of logical effort by using the gate input area comparison method to calculate the gate delay of an inverter in the following circuit configuration.

Next, let us apply the logical effort concept to find the gate delay of an inverter driving a few gates. By definition, the gate delay will be a function of the fanout, its logical effort, and the intrinsic parasitic capacitance of the gate. Figure 5-12 shows a typical structure.

Using Equation 5-10, $D_g = FG + p_{\text{gate}}$, we have the following:

- Logical effort for inverter: $G = 1$
- Electrical effort: $F = C_{\text{out}}/C_{\text{in}} = 4$
  = input capacitance of four inverters/input capacitance of one inverter.
- Parasitic delay of an inverter: $p_{\text{inv}} = 1$

Therefore, gate delay, D_g, is $1 \times 4 + 1 = 5$.

Multiplying D_g with the process-dependent multiplier, τ, will produce the actual gate delay. The process multiplication factor can be derived easily by plotting a delay-vs.-fanout plot for any gate. Figure 5-13 shows such a plot for an inverter with an approximate trendline and the values of the y-intercept and slope automatically calculated by Excel.

Using $D_{\text{abs}} = \tau(fg + p_{\text{inv}})$ and substituting the values of g, p_{inv}, and f, we can calculate the value of τ by looking up the values of Figure 5-13, as shown in Table 5-2.

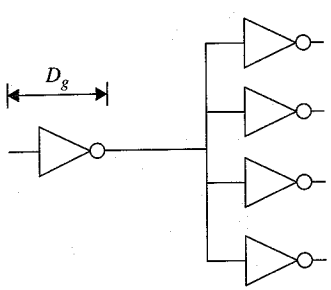


Figure 5-12 A typical circuit setup for calculating gate delay.

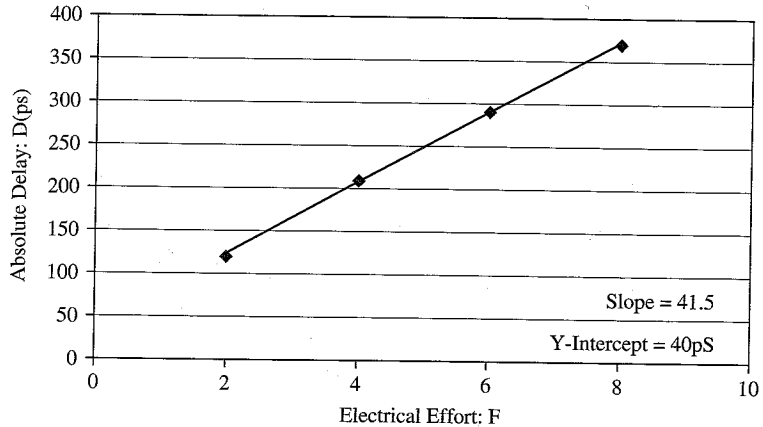


Figure 5-13 Simulated delays of an inverter driving loads with different fanouts for a particular process.

Note that for the inverter, $g = 1$ and $p_{\text{inv}} = 1$ by definition. As can be seen from Table 5-2, the average value of τ is 41 ps.

For designing large circuits that need several stages of gates or buffers, the total signal path delay can be expressed as

$$D_{\text{tot}} = \sum D_g = \sum fg + \sum p \tag{5-11}$$

Next, it can be proved that the delay is smallest when each stage bears the same stage effort:

$$f_1 g_1 = f_2 g_2 = f_3 g_3 = \ldots \tag{5-12}$$

Applying this concept on a series of "n" inverters, the following can be shown:

$$F^n = \frac{C_{\text{load}}}{C_{\text{stage1}}} \tag{5-13}$$

TABLE 5-2 Process Multiplication Factor τ Calculation

f	g	p_{inv}	D_{abs}	$\tau = \frac{D_{\text{abs}}}{(f \times g + p_{\text{inv}})}$
2	1	1	120	40
4	1	1	210	42
6	1	1	290	41.4
8	1	1	370	41.1

Hence, taking log on both sides, we have the following:

$$n = \frac{\ln\left[\frac{C_{load}}{C_{stage1}}\right]}{\ln[F]} \tag{5-14}$$

Because each stage bears the same stage effort and assuming all gates are of a similar type (so we can have the same p_{gate}), the total path delay can also be expressed as follows:

$$D_{tot} = \sum_{1}^{n} Dg = \sum_{i}^{n} fg + \sum_{1}^{n} p = nDg \tag{5-15}$$

Substituting n from Equation 5-14, we have

$$D_{tot} = M\frac{Dg}{\ln[F]}$$

where $M = \ln\left[\frac{C_{load}}{C_{stage1}}\right]$

Because $D_g = FG + p$, substituting D_g in the preceding equation we get

$$D_{tot} = M\frac{(FG+p)}{\ln[F]} \tag{5-16}$$

Next, to find the minimum delay for this path, we should differentiate the preceding equation with respect to F and equate it to 0.

Using the quotient rule of differentiation, we can write

$$\frac{dD_{tot}}{dF} = M\left[\frac{G\ln F - (FG+p)\left(\frac{1}{F}\right)}{(\ln F)^2}\right] = 0 \tag{5-17}$$

$$G\ln F - G = \frac{p}{F}$$

$$F[\ln F - 1] = \frac{p}{G}$$

Because for an inverter $G = 1$, the preceding equation can be written as

$$F[\ln F - 1] = p \tag{5-18}$$

Neglecting parasitics ($p = 0$), we find the familiar result $F = 2.718$ (e). For $p = 1$, we can solve the equation numerically to get $F = 3.5$. Similarly, for a two-input NAND gate and two-input NOR gate, it can be shown that the minimum delay can be achieved by $F = 3.7$. Next, we can do a little sensitivity analysis to see how the delay varies if F is not minimum.

Using

$$D_{tot} = M\frac{Dg}{\ln[F]} \tag{5-19}$$

and representing the minimum delay as $\hat{D}_{tot} = M\frac{D_{min}}{\ln[F_{min}]}$, we can plot

$$\frac{D_{tot}}{\hat{D}_{tot}} = \frac{\frac{Dg}{\ln[F]}}{\frac{D_{min}}{\ln[F_{min}]}} \tag{5-20}$$

with respect to F, as shown in Figure 5-14.

Through further analysis, we can see that for different types of standard logic gates, the minimum fanout, F_{min}, will vary between $2.5 < F < 4.5$. Next, from Figure 5-14, you can see that keeping the fanout per stage $2.4 < F < 6$ gives a delay within 15% of the optimal delay. In practice, keeping a uniform $F = 4$ per-stage number allows for

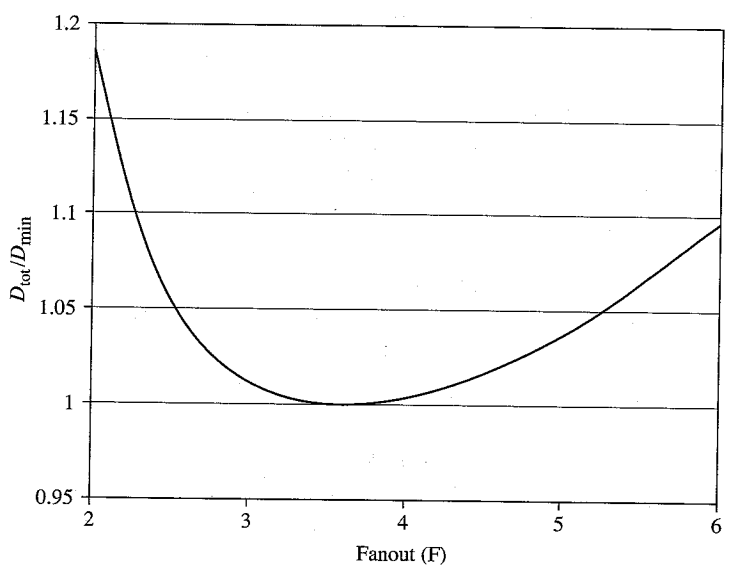


Figure 5-14 Delay sensitivity with fanout variation.

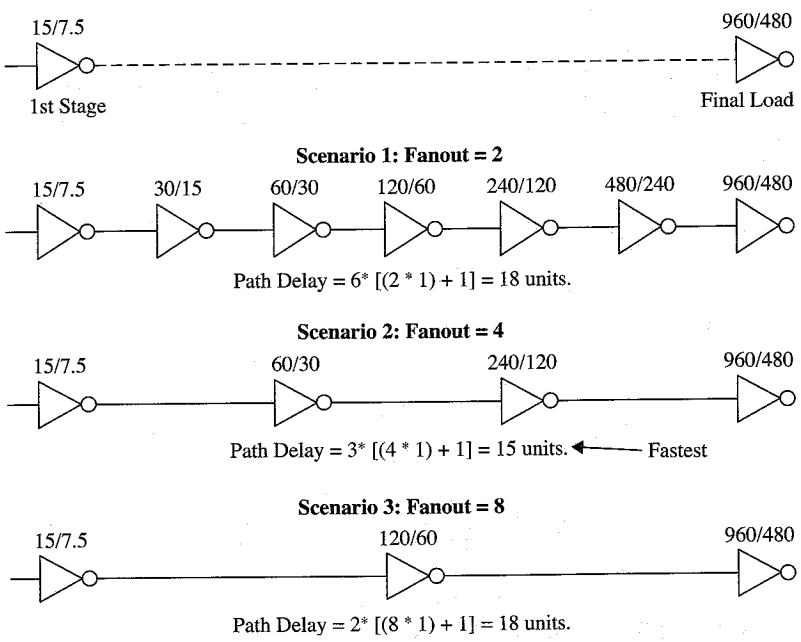
easy calculation and optimum results. An easy way to use the results from logical effort in practice is to first determine the output load capacitance (in terms of gate loading) and then work back by following $F = 4$ fanout backward through each stage.

Take the following example in Figure 5-15 to determine the best fanout/stage and the number of stages for a clock driver whose final equivalent loading is represented with an inverter of size 960 µ (PMOS) and 480 µ (NMOS). Also, for simplicity, let's imagine the initial driver size is 15 µ (PMOS)/7.5 µ (NMOS).[14,15]

From the preceding example, we can easily infer the following:

- Signal propagation paths are fastest when fanout is close to 4.
- Path delay is weakly sensitive to the number of stages and gate sizes.
- Using fewer stages does not mean faster propagation.

Keep in mind that logical effort uses a simplistic delay model and that it neglects input rise/fall time effects. In the preceding calculations, interconnect parasitic capacitance loadings were not accounted for. In practice, this will create erroneous results, especially for long interconnecting wires. An easy way to account for parasitic capacitances is to estimate the loading parasitic capacitance and convert it into an equivalent gate capacitance. This should be a one-step process as soon as the parasitic capacitance-to-gate capacitance ratio is determined.



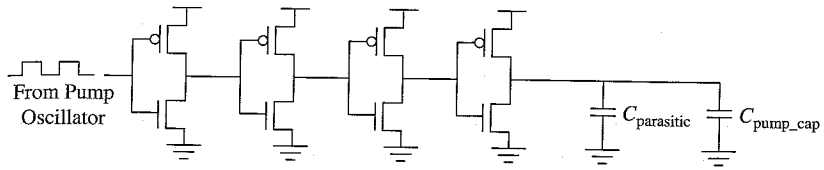


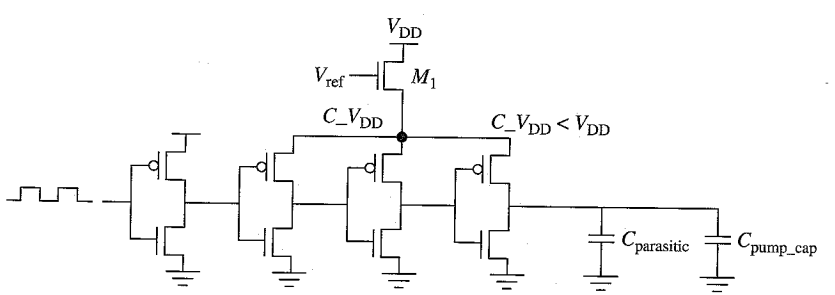
Figure 5-16 A clock signal driving buffer.

Next, connect this equivalent gate capacitance to the output of the driver that is driving the long line (i.e., this driver will see the equivalent gate capacitance as an additional fanout element). This process allows for more accurate estimation of fanout and the total path delay.

A simple clock design can be a series of inverters, sized according to the FO4 ratio, driving the capacitors. Remember that because in most cases the final capacitive load is heavy, an FO4 ratio might not be maintained. Instead, the load may be 10–100 larger than the preceding stage. Nevertheless, a fast ramp up of node V_0 V_ϕ is preferred, and extra boosting techniques may be applied to hasten it (more about this later).

Even though the simple clock buffer shown in Figure 5-16 is sufficient in many applications, it has an inherent shortcoming for charge pump applications. The clock buffer output voltage, V_0, is dependent on the supply voltage, V_{DD}. This means that when the supply voltage variation is on the higher side, the clock amplitude will be higher and the charge pump output ramping will be faster, and vice versa. In some applications, this may create a problem. To solve this problem, a V_{DD}-independent clock source should be created. A quick-and-easy way to create such a clock source is shown in Figure 5-17.

MOSFET M_1 is generally a low V_t transistor. When a constant reference voltage, V_{ref}, is applied at the gate of NMOS M_1, the source voltage at node $C - V_{DD}$ will always be at the level $V_{ref} + V_t$, regardless of V_{DD} variation, provided V_{DD} is higher than $V_{ref} + V_t$. V_{ref} voltage is usually derived from a bandgap reference voltage generator. It must be noted that regulating supply voltages through an NMOS-type transistor has some drawbacks. Referring to Figure 2-20 in Chapter 2, consider the



normal operating point of the present NMOS M_1 corresponding to V_{gs4}. When there is a huge current demand from the $C - V_{DD}$ node, the voltage at node $C - V_{DD}$ will invariably dip down significantly. This happens because as M_1's I_{ds} increases, V_{gs} has to increase. However, because the gate voltage is tied to a constant V_{ref}, $C - V_{DD}$ has to dip down to accommodate the high current demand. In this current scenario, to get around this problem, ample decoupling capacitors should be used on the node $C - V_{DD}$ to filter out voltage dips.

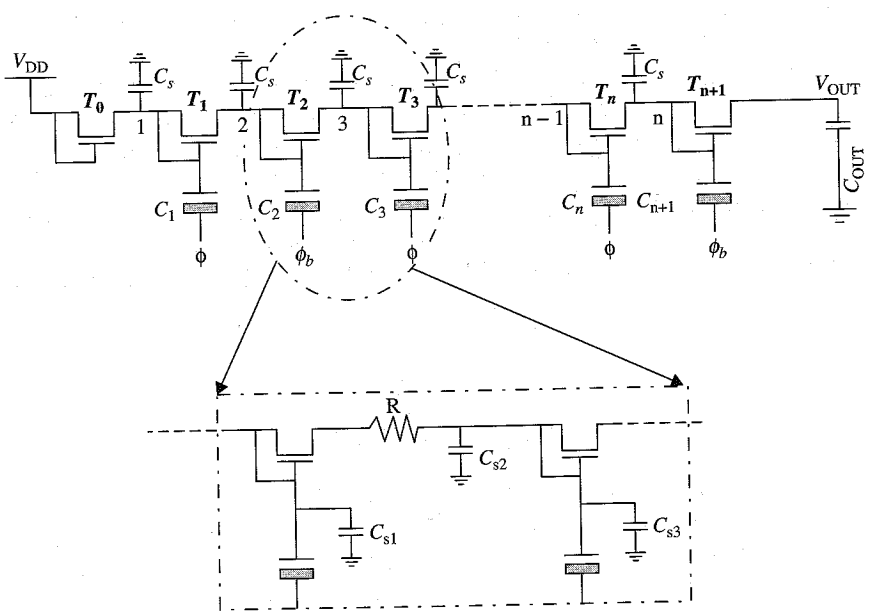
5.7 Parasitic R and C

The parasitic capacitance in the individual Dickson charge pump stages consumes charge during the transferring. It is one of the major factors that can increase the internal impedance of charge pump. Recall the following from a previous equation:

$$V'_{\phi} = \left(\frac{C}{C + C_s}\right) V_{\phi} \tag{5-21}$$

Here, V'_{ϕ} is the voltage induced at each stage, which is a function of the parasitic capacitance C_s. The larger the parasitic C_s, the lower the voltage that could be coupled during each clock cycle.

Figures 5-18 and 5-19 show the primary location of the parasitic capacitance C_s. As capacitor and MOSFET sizes get bigger, they need to be



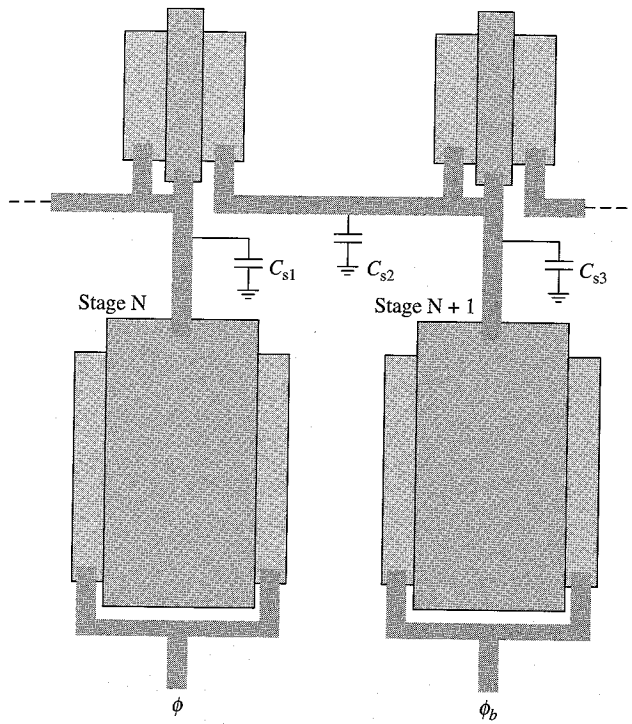
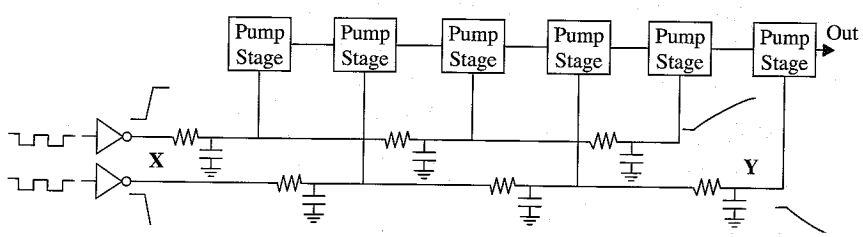


Figure 5-19 Interconnect parasitics, layout view.

split apart into smaller parallel sections. Partitioning large devices into smaller components will increase the total area and the parasitic capacitance associated with each internal node. Proper effort must be made to create an optimum layout floor plan and interconnect metal lengths and widths to make sure the parasitics are kept to a minimum.[16–18]

As shown in Figure 5-20, because of the inherently large size of the capacitors, the total layout area tends to be large. If global clock drivers are used to drive the clocks near X, the clocks will have a large rise/fall times at the end, Y. This will significantly reduce the maximum frequency of operation. In addition, the clock drivers have to be sized up to achieve tolerable clock slopes. This trend tends to consume more power.



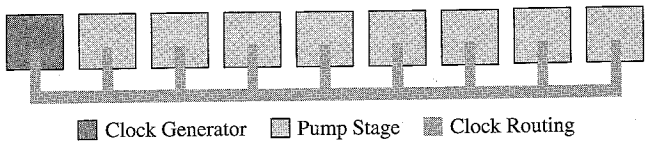


Figure 5-21 Linear layout floor plan.

The clock distribution scheme is also important in reducing the clock line parasitics, which allows faster operation and efficient driver sizing. A linear layout is shown in Figure 5-21. This scheme, though simple, has a lot of clock line parasitics associated with it. Further, the clock rise/fall time at the last block will invariably be longer than those at the initial blocks.

A better way is to use the floral layout pattern shown in Figure 5-22 to reduce clock-signal line-parasitic capacitance. The main idea is to design a floor plan in such a way so as to reduce the parasitic on each signal line.

5.8 Power Bus and Bower Bus Capacitance

The charge pump's inherent inefficiency, coupled with the large amount of required switching current for the drivers, is itself the power hog on the chip. On many occasions the chip exhibits poor performance because of the huge current demand from the charge pump on the internal supply bus V_{DD}. Large *IR* drop on V_{DD} bus could make the actual power supply near the pump deviate significantly from worst case scenario.

In general, it is a better idea to place the charge pump closer to the supply pin pad, or to dedicate a special power supply pad just for the charge pump. The width of power bus from pad to the pump blocks has to be wide enough to address the worst case *IR* drop. Appropriate layout area must be allocated for supply decoupling capacitors in order to smooth out the peak current from the clock buffer switching.

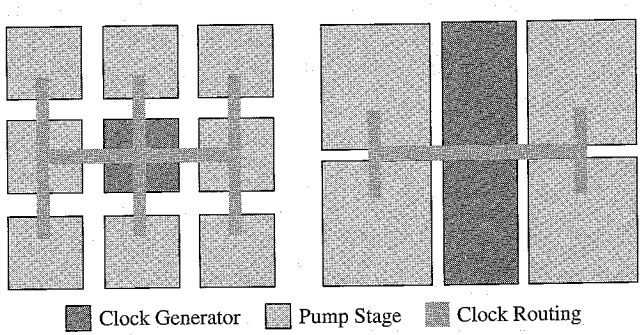


Figure 5-22 Patterned layout floor plan.

5.9 Conclusion

This chapter introduced the various charge pump specifications that circuit designers need to determine as an initial step before designing the charge pump. It introduced nine main design specifications and analyzed each of the major parameters in detail and discussed how to create trade-offs between each one of them. Through these steps readers can determine the initial values of the number of pump stages, the pump capacitor size, the pump clock frequency and various other parameters. This chapter also showed how to design the various supporting charge pump circuit blocks. Starting with a balanced non-overlapping clock source generator, the chapter introduced clock amplitude doubler and tripler circuits. Next, the concept of Logical Effort for clock buffer sizing was introduced, along with a detailed quantitative analysis, as well as the impact of parasitic resistors and capacitances on circuit performance. This chapter concluded with a discussion on layout implementation, floor planning and clock distribution—all of which aid in the design of an efficient charge pump.

The following chapter will take your understanding of the basic operation of the 2-phase charge pump to the next level and illustrate how to design a better charge pump.

References

1. Witters, J.S., G. Groeseneken, H.E. Maes. "Analysis and modeling of on-chip high-voltage generator circuits for use in EEPROM circuits." *IEEE Journal of Solid-State Circuits*. Vol. 24, No. 5, pp. 1372–1380, October 1989.
2. Pelliconi, R., et al. "Power Efficient Charge Pump in Deep Submicron Standard CMOS Technology." *Solid-State Circuits Conference*, 2001. ESSCIRC 2001. Proceedings of the 27th European.
3. Favrat, P., P.H. Deval, M. Declercq. "A New High-Efficiency CMOS Voltage Doubler." *IEEE Custom Integrated Circuits Conference (CICC'97)*, pp. 259–262, May 5–8.
4. Choi, Ki-Hwan et al. "Floating-well Charge Pump Circuits For Sub-2.0 V Single Power Supply Flash Memories." Digest of Technical Papers. Symposium on VLSI Circuits. pp. 61–62, June 12–14, 1997.
5. Naso, et al. "Negative-voltage charge pump with feedback control." U.S. Patent 5,168,174.
6. Papaix, Caroline and Jean-Michel Daga. "High Voltage Generation for Low Power Large VDD Range Non Volatile Memories." PATMOS 2001. http://patmos2001.eivd.ch.
7. Young et al. "Power supply insensitive substrate bias voltage detector circuit." U.S. patent 6,172,554.
8. Sung, Ha Min. "Substrate voltage detection control circuit." U.S. patent 6,281,742.
9. Lee et al. "Circuit for sensing back-bias level in a semiconductor memory device." U.S. patent 5,262,989.
10. Brugler, J.S. "Theoretical Performance of Voltage Multiplier Circuits." *IEEE Journal of Solid-State Circuits*, pp. 132–135, June 1971
11. Favrat, Pierre, Philippe Deval, and Michel J. Declercq. "A High-Efficiency CMOS Voltage Doubler." *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 3, pp. 410–416, March 1998.

12. Tanzawa, T. and A. Shigeru. "Optimization of Word-Line Booster Circuits for Low-Voltage Flash Memories." *IEEE Journal of Solid-State Circuits*, Vol. 34, No. 8, pp. 1091–1095, AUGUST 1999
13. Sutherland, I., F.R. Sproull, and D. Harris. Logical Effort: Designing Fast CMOS Circuits. Morgan Kaufmann, 1999.
14. Bateman, Bruce. "High-Speed SRAM Design." ISSCC'98 Tutorial.
15. Horowitz, M. "VLSI Scaling for Architects." Presentation slides. Computer Systems Laboratory, Stanford University.
16. Chern, J. et al. "Multilevel metal capacitance models for CAD Design Synthesis Systems." *IEEE Electron Device Letters*, Vol. 13, No. 1, pp. 32–34, January 1992.
17. Barke, Erich. "Line-to-Ground Capacitance Calculation for VLSI: A Comparison." *IEEE Transactions on Computer-Aided Design*, Vol. 7, No. 2, pp. 295–298. February 1988.
18. Wong, Shyh-Chyi, Gwo-Yann Lee, and Dye-Jyun Ma. "Modeling of Interconnect Capacitance, Delay, and Crosstalk in VLSI." *IEEE Transactions on Semiconductor Manufacturing*, Vol. 13, No. 1, pp. 108–111, February 2000.

Chapter

6

Designing a Better Charge Pump

After the introduction of basic charge pump design concept and parameters associated with the design, how can one design a better charge pump? What criteria should be used to judge if one design approach is better than another? These are very valid questions the circuit designers will always ask. Charge pumps are used in many different applications. Different chips may have different requirements for the pumps needed. It does not matter what function the pump needs to serve—the size of the pump has to be small in terms of the total layout area. The charge pump should be able to deliver more current at a given regulation level. It should have better power efficiency as a whole system. The output of a charge pump should have less noise while within regulations. These four criteria should be universal rules for judging pump designs.

Die size is always expensive in the chip industry. Especially for commodity chips selling with low profit margins, extra die size savings means higher profitability in production. The charge pump is normally one of the circuit blocks that consume a large die size in many applications. It is always one of the best candidates for size optimization. So how does one design a charge pump with a layout area that is as small as possible? In order to understand what charge pump parameters could be optimized to the size of the pump, let us first revisit the design of the Dickson charge pump[1] in detail. Almost all modern charge pumps are derived from the basic concept of the Dickson charge pump. Understanding the limitations of the Dickson charge pump automatically unveils the path to better designs.

Figure 6-1 is a schematic view of a generic Dickson charge pump. It consists of N total pump stages. Each individual pump stage could be simply divided into a capacitor and NMOS diode-connected transistor. The capacitor is used for receiving and transferring charge and also is used for elevating the potential energy of the charge being delivered

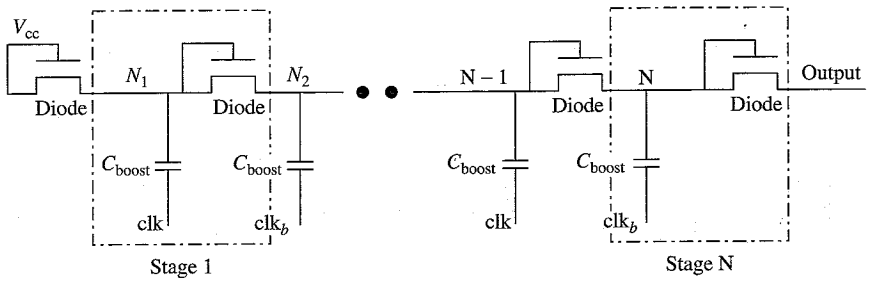


Figure 6-1 Generic Dickson charge pump.

to the output. The diode-connected NMOS in between two stages is used to allow charge transferring in only one direction, and not in the reverse direction. Because the stages of the charge pump are similar in structure, to simplify our analysis, a single pump stage is chosen for close examination. Various parameters associated with this single pump stage could be studied to analyze their effects on the overall charge pump performance, pump layout size, pump clock frequency, and total power consumption. With a detailed knowledge of these key parameters, understanding the tradeoff between these various parameters for optimization purposes should be easier.

Figure 6-2 shows a single stage of generic Dickson charge pump. The capacitor associated with the next stage is also included. The 2-phase clocking scheme used by the Dickson charge pump is shown in Figure 6-3.

6.1 Parameters Associated with Pump Performance

A few basic parameters can be used to characterize the single pump stage in Figure 6-2. Equation 6-1 lists the parameters associated with the operations of a single pump stage. The operations of the pump can be analyzed in two ways—from the charge-transfer point of view and from

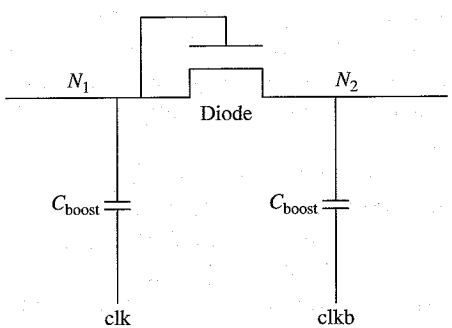


Figure 6-2 A single stage of the Dickson charge pump.

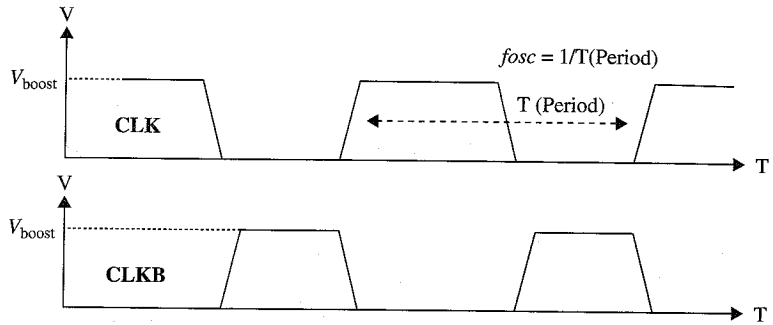


Figure 6-3 A 2-phase clocking scheme for the Dickson charge pump.

the voltage point of view—to see how the parameters in Equation 6-1 affect the charge pump's performance.

C_{boost}	Boosting capacitance per stage
V_{boost}	Boosting clock amplitude
V_t	Threshold voltage of NMOS transistor
T_{period}	Pumping clock period
f_{osc}	Pump boosting clock oscillation frequency

$$f_{osc} = 1/T_{period} \tag{6-1}$$

Q_{stage}	Charge transferred per clock cycle

6.1.1 Charge transfer point of view

First, this single stage is analyzed from the point of view of the charge transfer. While the charge pump is in operation, the charge is being transferred from one stage to the other in each clock cycle. The charge pump has to meet the output load demand in two operating phases—one is for the ramp-up phase and the other is for the regulation phase.

In the ramp-up phase, the output load needs to be charged up from an initial level to the final regulated voltage within the output settle-down time required by the design specification. This statement has two inherited meanings: The first meaning is related to the actual initial output ramp-up speed at the very beginning of operation; the second meaning is related to the output voltage recovery speed during regulation. If extra capacitive load is connected to the pump output, or spike of load current for duration of time while the pump is in regulation, charge sharing or lost charge on the existing load capacitance will immediately occur and bring down the output voltage. The scenario of the load current spike can always be transformed to an extra switching load capacitance.

To simplify our analysis it will not be discussed here. Once the output voltage drops, the pump has to recover and charge up this additional capacitance to the regulation level. In both scenarios, the capacitive loads need to be charged up to a fixed level within the fixed amount time required. These two cases should be grouped into one category. To ramp up the output of the charge pump from an initial level, V_1, to the final regulation level, V_2, within the limited time, T_{rise}, the charge pump has to deliver the minimum charge to its output.

First, a fixed amount of charge, Q_1, needs to charge up the output load, C_{load}. Q_1 is defined in Equation 6-2 by the loading capacitance and voltage difference. During this period of time, I_{load} is the current consumed by all circuits connected to the output of the charge pump. In Equation 6-3, Q_2 is defined as the total amount of charge consumed by circuits within the ramp-up phase. Within the ramp-up phase, the total charge has to be delivered by the charge pump to its output. This is the summation of Equation 6-2 and Equation 6-3. Q_{total} is defined in Equation 6-4. Because charge is being delivered within each clock cycle, it is important to determine the minimum amount of charge that has to be transferred per clock cycle from one stage to the other.

$$Q_1 = (V_2 - V_1)C_{load} \tag{6-2}$$

$$Q_2 = I_{load}(t)T_{rise} \tag{6-3}$$

$$Q_{total} = Q_1 + Q_2 \tag{6-4}$$

The number of clock cycles needed can be calculated using Equation 6-5. T_{period} is the pump clock period. With Equation 6-4 and Equation 6-5, the minimum charge, $Q_{required\_1}$, per clock cycle in the ramp-up phase can be calculated using Equation 6-6. If the load current, $I_{load}(t)$, is small compared with the capacitive current, Q_2 can be ignored in Equation 6-6 for simplicity's sake.

$$N = T_{rise}/T_{period} \tag{6-5}$$

$$Q_{required\_1} = Q_{total}/N$$
$$Q_{required\_1} = \left[(V_2 - V_1)C_{load} + I_{load}(t)T_{rise}\right]T_{period}/T_{rise} \tag{6-6}$$

In the regulation phase, the pump only needs to sustain the output current at the regulated level. No more DC current is needed to charge the load capacitance. The minimum current delivered to the output by the pump has to be larger than the load current. Otherwise, the output is

the load current, $I_{load}(t)$, at the regulated voltage, V_2, or to deliver the minimum output power, $V_2 I_{load}(t)$, to the output, Equation 6-7 has to be satisfied. $Q_{required\_2}$ is the minimum charge per clock cycle that needs to be delivered to the pump output in the regulation phase:

$$Q_{required2} = I_{load}(t)T_{period} \tag{6-7}$$

In order for the pump to meet the requirements of all operations, the minimum charge transferred per clock cycle by the charge pump to its output has to meet the maximum of $Q_{required\_1}$ and $Q_{required\_2}$, which is shown in Equation 6-8. During the charge-transferring phase, the transferred charge is lost along the path. First, the transferred charge is lost to the internal node parasitic capacitance, such wirings, MOSFET source/drain junction, and so on. Second, the charge is lost due to the inability to transfer the full amount of charge between stages. This could be due to the V_t of the diode-connected transistor, longer RC delay, and so on. The actual charge, Q_{stage}, that needs to be transferred between stages has to be larger than Q_{cycle}, given in Equation 6-8:

$$Q_{cycle} \geq \mathrm{MAX}\left[Q_{required\_1}, Q_{required\_2}\right] \tag{6-8}$$

Equation 6-9 provides the relationships of Q_{stage} with respect to the different parameters in Equation 6-1. After looking at the basic requirements from the charge-transfer point of view, we come back to the original design question: How does one design a smaller charge pump that meets the same requirements as before? To reduce the layout area for a charge pump, we have to reduce the total boosting capacitance associated with each stage. This allows the supporting signal drivers and signal line width to be resized proportionally, too. So how do we trade off one parameter for another to allow for smaller boosting capacitance per pump stage?

$$Q_{stage} \propto C_{boost}$$
$$Q_{stage} \propto V_{boost}$$
$$Q_{stage} \propto C_{boost} V_{boost} \tag{6-9}$$
$$Q_{stage} \propto 1/V_t$$

Q_{stage} is reduced by parasitic capacitance

From Equation 6-9, we know that Q_{stage} is proportional to the product $C_{boost}V_{boost}$. It is natural instinct to increase the amplitude of V_{boost} to compensate for the reduction of C_{boost}. Because in many designs the clock drivers are commonly biased by the system supply V_{cc}, a higher chip supply voltage V_{cc} would allow the design to be smaller. The change in size of a charge pump design is obvious as the chip power supplies are scaled down. For similar designs, a 5 V charge pump is smaller than a 3 V design, and 3 V charge pump is much smaller than a 1.8 V charge pump. The chip supply voltages are fixed by a given technology and chip specification, which is something that cannot be changed by designers. However, there are no requirements that limit the designers to using only the system power supply V_{cc} to supply the clock drivers. If there is an additional high-voltage pump that can generate potential higher than the system supply V_{cc}, then all clock drivers can use this new high-voltage supply to drive the pump circuit. The derived pump clock amplitude, V_{boost}, would be higher than V_{cc}. As a consequence, C_{boost} can be reduced using this approach. Indeed, this approach is achievable by circuit design. However, there would be extra circuits due to the new high-voltage clock generation. In practice, this tradeoff has to be studied to see if there is an overall benefit in die-size savings.

What if the clock driver supply is fixed to supply voltage V_{cc}? Could C_{boost} still be reduced by other means? Reducing C_{boost} per stage means reducing the total amount of charge that can be transferred per clock cycle. In the ramp-up phase, according to Equation 6-4, a minimum amount of charge needs to be transferred to the output node per clock cycle. If C_{boost} is reduced, $C_{required\_1}$ per clock cycle would be reduced too. If all other parameters in Equation 6-2 and Equation 6-3 are unchanged, T_{period} should be scaled down proportionally to make the scaling of the boosting capacitance achievable. The higher clock frequency would compensate for the reduction in the boosting capacitance.

Stated another way, during the initial ramp-up phase, a fixed amount of charge needs to charge up the output capacitance. If the charge transferred per clock cycle is reduced, in order to charge up the load within T_{rise}, more clock cycles are needed within the given ramp-up period to complete the job. Clock frequency will be higher. If the ramp-up time or the recovery time is not the maximum in Equation 6-8, it is necessary to check the output power requirement in the regulation phase. Can increasing the pump clock frequency and reducing the boosting capacitance work the same in the regulation phase?

In the regulation phase, according to Equation 6-7, the minimum charge required to deliver to the pump output per clock cycle is proportional to the product of $I_{load}(t)T_{period}$. Assuming the load characteristic, $I_{load}(t)$, is unchanged during optimization and the pump clock frequency is increased, T_{period} will be reduced. The minimum amount of charge, $Q_{required\_2}$, that needs to be delivered in a given clock cycle will be reduced proportionally. The proposal of increasing the pump clock frequency to compensate for the reduction of boosting capacitance works in principle.

6.1.2 Voltage point of view

After examining the charge pump performance from the charge-transfer point of view, we will analyze the charge pump from the voltage point of view to see how the optimization should be done. The charge pump is used to transfer charge from stage to stage, and to elevate the potential energy of the charge through the successive stages.

The single stage of the Dickson charge pump, shown in Figure 6-2, will still be used here for analysis. In Figure 6-3, the 2-phase clocking scheme was presented to drive the Dickson charge pump. In practice, Clk_a and Clk_b are usually non-overlapped clocks, as shown in Figure 6-4. The high phases of the clocks are never overlapped. They are designed to allow better charge transfers because the source voltage of the diode-connected transistor is always coupled down first before the charge transfer starts. The clock amplitude of Clk_a and Clk_b is at V_{boost} level, which is normally the chip supply voltage, V_{cc}, or the on-chip voltage generation supply.

As shown in Figure 6-5, the charge is transferred from stage N_1 to stage N_2. In the first half clock cycle, N_2 is coupled down by the clock at time T_1. At time T_2, node N_1 is coupled by the boosting capacitance. The amplitude of voltage change is given as V_1. Node N_2 starts taking charge from the preceding stage, N_1, after time T_1. Between T_2 and T_3, nodes N_1 and N_2 tend to converge to the same level as the charge is being transferred. Due to the threshold voltage of the transistors and the RC delay of the internal nodes, the full amount of charge transfer cannot be realized. At time T_3, the voltage difference between N_1 and N_2 is V_2 in Figure 6-5. The difference here shows that this pump is unable to transfer the full amount of charge between stages due to V_t and RC delay. The total trapped charge is $\Delta Q = C_{boost}V_2$. The separation of T_2 and T_1, or T_4 and T_3, allows the voltage on the internal nodes to be reset

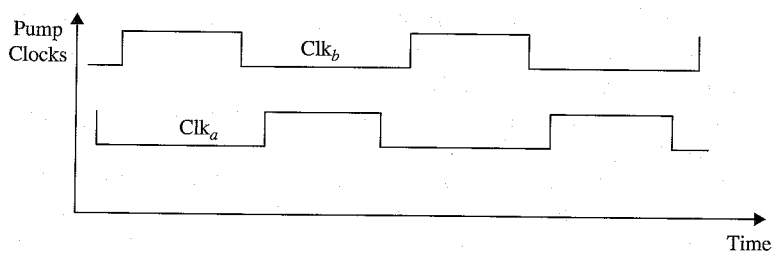


Figure 6-4 A 2-phase phase-clocking scheme.

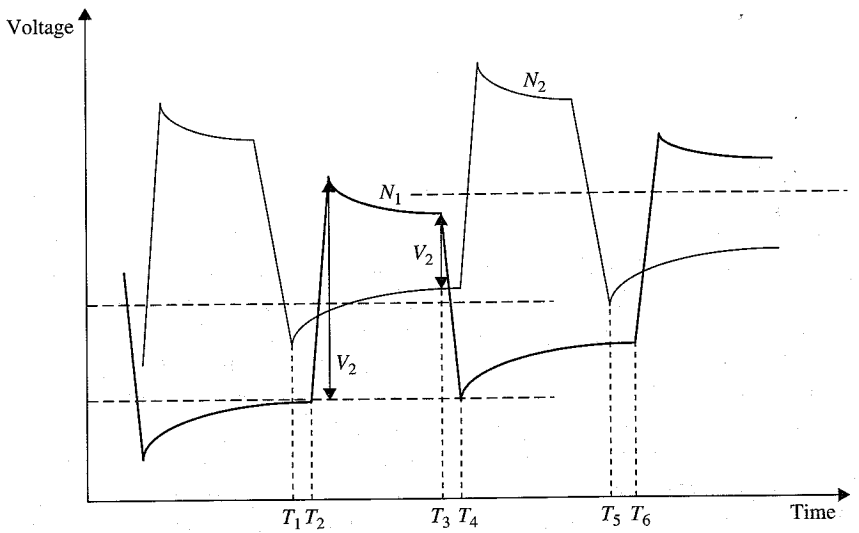


Figure 6-5 Internal voltages and timing for two successive stages.

properly before the next boosting happens. This separation (or non-overlapping) of clocks does not make much difference in real design. In the next half clock cycle, N_1 will be boosted down at T_3 to take charge from the previous stage, and at T_4 N_2 is boosted up to transfer the charge to the next stages.

With the Dickson charge pump, due to the threshold voltage of diode-connected devices and the *RC* delay per stage, the charge rarely can be fully transferred from stage to stage per clock cycle. With the increasing source biasing along the chain, the higher the regulation level needed, the less efficiently the charge can be transferred at late stages in later stages. In order to meet the design requirement, usually a much large boosting capacitance is needed to compensate for this loss of efficiency.

Based on the voltage point of view, in order to design a smaller charge pump, if somehow the device V_t of the diode-connected MOSFETs can be reduced, or somehow be fully cancelled by some means, the charge transfer efficiency can be improved. At the same time, no extra boosting capacitance is wasted to address the inability of charge transferring. Let us put the charge pump parameters together to see if smaller coupling capacitance is achievable.

As shown previously, two parameters can be controlled to optimize the pump size: clock frequency and V_t cancellation. Clock frequency has a direct impact on pump performance and the overall pump size. At a lower clock frequency, pump output power at regulation is directly

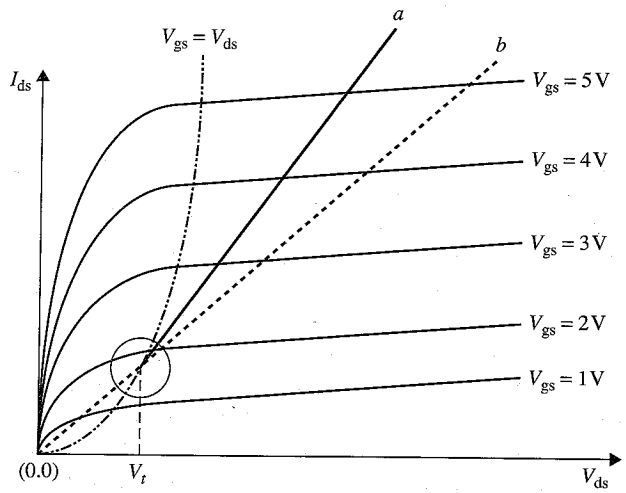


Figure 6-6 I_{ds} versus V_{ds} at different V_{gs}.

proportional to clock frequency. At a higher clock frequency, this relationship will not hold for some architectures or some designs. The maximum clock frequency at which the charge pump can operate depends on the V_t of NMOS and the *RC* delay of the individual pump stage. A diode-connected NMOS device has $V_{ds} = V_{gs}$, and the conduction of current stops when $V_{gs} = V_t$. With this kind of connection, the diode-connected NMOS could be simplified into an equivalent resistance, R_{eq}.

As shown in Figure 6-6, two ways are available to find the equivalent resistance of the NMOS device. First, the slope of the trajectory at each operating point of the actual I_{ds} curve represents the inverse of the equivalent resistance of R_{eq} of the pump stage at that particular operating point. At operating point $V_{ds} = V_{gs} = V_t$, line *a* is the trajectory in Figure 6-6. A crude method would be to connect the operating point and the origin (0,0) with a straight line directly. As line *b* shows in Figure 6-6, the slope could also represent the inverse of the equivalent resistance of R_{eq}. Both methods lead to a similar conclusion.

Figure 6-7 show the equivalent view of a single pump stage. Each stage can be viewed as a simple *RC* network. The delay through the

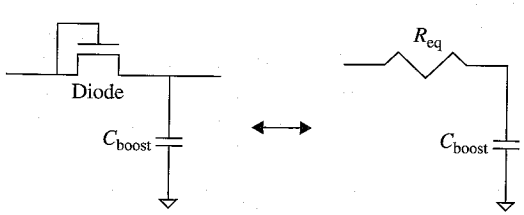


Figure 6-7 Equivalent view of a single pump stage.

stage can be described using Equation 6-10:

$$\tau = kR_{eq}C_{boost} \tag{6-10}$$

k is a coefficient determined by how much of the charge is being transferred up to a satisfied percentage in a given clock cycle. In an RC network, the delay on the output signal swing has an exponential relationship over the time. As the output tries to reach its final target value, it takes an infinite amount of time. There must be a cutoff point in time to determine the pump clock period. One practical approach is to use the amount of time it takes to transfer nearly 70%–80% of charge as the guideline for the cutoff point. If the actual clock period is shorter than τ in terms of the charge transfer, the gain from higher clock frequency and the loss from less charge being transferred per clock cycle will be a wash. Pump performance will not be increased with the increasing clock frequency. τ determines how fast the pump clock can be designed for a real circuit. With the constraint shown in Equation 6-10, it would be difficult to improve the clock frequency with the Dickson pump. However, many other approaches based on the Dickson charge pump allow faster clocking frequency to be realized.

6.2 How to Improve Charge Pump Efficiency

As shown in the previous discussion, the limitations of the Dickson charge pump are V_t drop per pump stage and R_{eq} associated with each diode-connected device. V_t drop per stage reduces the charge transferring efficiency. At each pump stage, for every clock cycle there is always charge that cannot be transferred from stage to stage, which is represented as $\Delta Q = C_{boost} \times V_t$. The effect gets worse at later stages with the body bias effect, or as the output voltage required moves higher. The R_{eq} of each pump stage is another critical factor in determining the pump performance. It is related to the conductivity of the transferring diode. A wider device and shorter channel improves R_{eq}. Larger V_t for the diode-connected device would increase R_{eq}. τ determines how fast the Dickson pump clock could be operated.

In designing a smaller charge pump, the frequency of the pump clock is one of the keys.[2] The other key is to make the V_t of the diode-connected device less effective. In the Dickson charge pump, the clock frequency can only be increased up to a limit. The V_t of the diode-connected device would increase as the source voltage rises. To design a better charge pump means to break the barrier of the clock frequency limitation in

the Dickson charge pump, and also to overcome the reduction of charge-transferring efficiency due to the V_t drop per pump stage.

Two types of approaches are commonly used in the industry to design charge pumps with better efficiency than the conventional Dickson charge pump. The first approach can be generalized as the V_t cancellation charge pump. The second approach can be generalized as a charge pump design using high-amplitude pump clocks.

6.2.1 V_t cancellation scheme

In the first approach, various pump design proposals use different techniques to boost the gate of transferring NMOS in each stage.[3–7] This allows a near full-charge transfer in each clock cycle. As the effective V_t of the transferring NMOS device is cancelled, the equivalent resistance, R_{eq}, of each pump stage is significantly reduced.

Figure 6-8 shows the plot of R_{eq} for the V_t cancellation scheme. Because the charge can be fully transferred from stage to stage, at the end of the charge transferring, V_d and V_s are close to equal potential. V_{ds} is near 0 V after the full transfer of the charge. On the IV curve of NMOS, the slope of all I_{ds} curves near the operating point (0V,0A) is very steep, which is the inverse of R_{eq}. The slope of the thick line is the inverse of R_{eq} for the Dickson charge pump near its operating point, $V_{gs} = V_{ds} = V_t$. It is obvious that equivalent resistance for the V_t cancellation scheme is much smaller than that of the Dickson charge pump. This is the fundamental reason why the V_t cancellation scheme can operate at a much higher pump clock frequency. A faster pump clock can reduce the size of

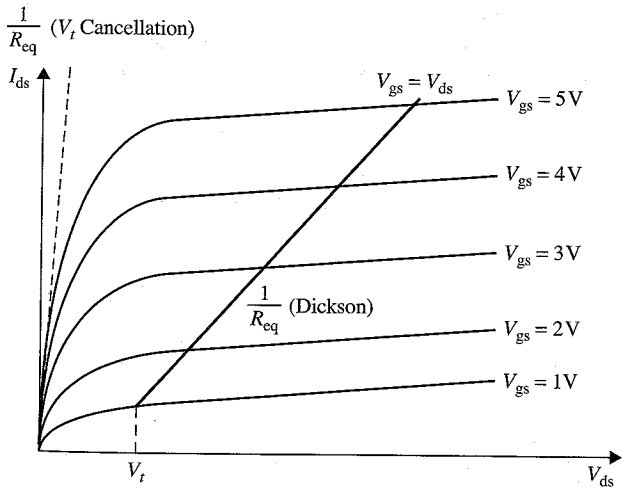


Figure 6-8 R_{eq} for V_t cancellation scheme.

the pump capacitance, which in turn further reduces the time constant, $\tau = R_{eq} \times C_{boost}$. Then the pump clock frequency can be further increased.

One of the great examples of a V_t cancellation scheme is the 4-phase charge pump design. The pump architecture has a very unique approach for canceling the threshold voltage of the diode-connected device through a bootstrapping technique. The details are discussed in Chapter 7.

The maximum clock frequency at which the pump can operate depends on the *RC* delay of the individual pump stage, plus the extra delay through clock generation and the clock driver. With the unknown factor of power supply and parasitic *RC*, it is better to have enough design margin built in. If we were to plot the inverse of performance-versus-pump clock frequency for the V_t cancellation charge pump in Figure 6-9, the curve could be divided into two regions. For $f_{osc} > f_0$, this region could be called the *exponential decay region*. In this region, the delay through either clock driver or the delay through the pump stage becomes significant. Any more reduction of the clock period would cause the entire pump to malfunction. For $f_{osc} < f_0$, because the charge can be fully transferred within the clock cycle, any increment of the clock frequency will translate to more charge being delivered to the output of the charge pump in a fixed amount of time. The performance of the charge pump will change linearly with the clock frequency. This is called the *linear region*. For $f_{osc} < f_0$, the performance change is relatively flat. It is safer to choose the operating clock frequency in this region for a realistic design.

6.2.2 Pump design using high-amplitude pump clocks

The second approach for improving charge pump efficiency is to design the charge pump with pump clocks with higher amplitude. The characteristics of the Dickson charge pump can be described using Equation 6-11. Charge loss per pump stage can be defined using Equation 6-12.

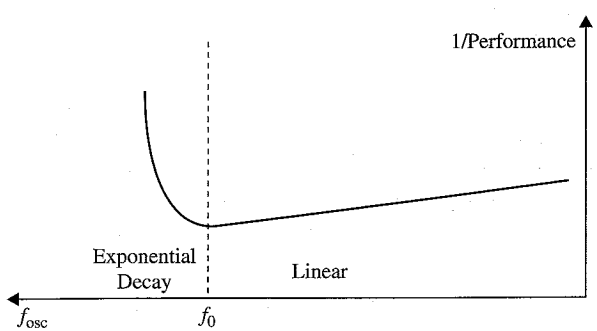


Figure 6-9 Performance versus pump clock frequency for the V_t cancellation pump.

Assuming V_t is unchanged, if the amplitude of the pump clock can be increased, the loss per stage due to the threshold voltage of the diode will be decreased. As a consequence, the charge transfer efficiency can be improved with this new design approach:

$$R_s = \frac{n}{(C+C_s)f_{osc}} \tag{6-11}$$

$$V_{out} = V_{in} + n\left[\left(\frac{C}{C+C_s}\right)V_{clock} - V_{in} - \frac{I_{out}}{(C+C_s)f_{osc}}\right] - V_{in}$$

$$\text{Loss} = \frac{V_t}{V_{clock}} \tag{6-12}$$

Because more charge can be transferred in each clock cycle, and because of the higher elevation of charge potential energy through each pump stage, fewer stages are needed in a serial connection. Based on Equation 6-11, this design approach reduces the equivalent impedance of the charge pump seen from its output port.

Figure 6-10 shows the I-V curves of two different designs. The curve with the starting point of $V_{gs} = 2.5$ V shows normal Dickson charge pump operation. The curve with the starting point of $V_{gs} = 5$ V represents the pump clock being just doubled from 2.5 V to 5.0 V. Because V_t is not cancelled, the charge transferring stops when $V_{gs} = V_{ds} = V_t$. If we draw the trajectories of I-V curves for both designs at the operating point

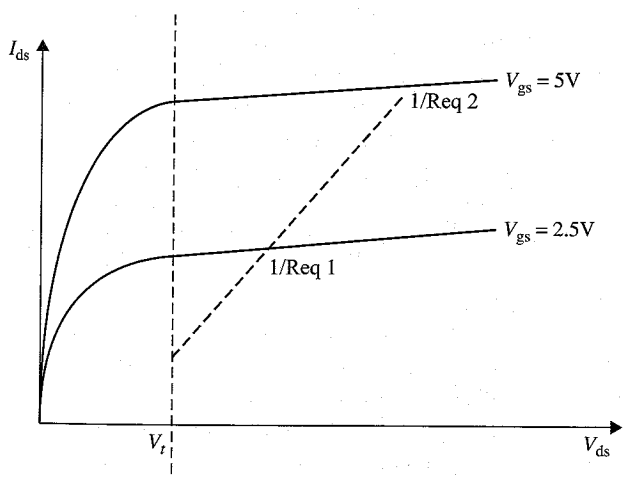


Figure 6-10 R_{eq} of multiplied clock charge pump.

$V_{gs} = V_{ds} = V_t$, both trajectories would overlap each other. The equivalent resistances of the two designs, R_{eq1} and R_{eq2}, should be equal. For a pump design with high-amplitude pump clocks, it is not the equivalent resistance per pump stage that is smaller. Rather, fewer stages are needed in serial to reduce the total impedance of the charge pump. Higher amplitude pump clocks at first hand do not allow the design to operate at a faster frequency than a normal Dickson charge pump. However, because it can transfer a larger percentage of charge with higher potential elevation per pump stage, the performance of charge pump is improved over the traditional approach. If the boosting capacitance can be reduced due to better pump efficiency, the clock frequency may be improved too. There are many great examples of charge pumps using high-amplitude pump clocks. The simplest one involves the use of clock doublers for clock generation, while keeping the rest of the pump circuit unchanged. More details are provided in Chapter 7.

6.3 Regulation of the Pump

Regulation of the charge pump is one of the aspects in determining pump performance. Pump output noise in regulation, accuracy of regulation, power consumption, and layout size are partially affected by the style of regulations. The following subtopics address these issues.

6.3.1 Resistive divider versus capacitive divider

Regulation is another important aspect of designing a charge pump. Feedback and regulation are interrelated. Feedback is the process of sampling the pump output potential and transferring back a control signal to the voltage-generation circuit. Regulation involves using the feedback signal to stabilize the output signal near the target level. Regulation schemes used by charge pumps can be generally divided into two main categories: resistive dividers and capacitive dividers. Of course there are still many other types of regulation schemes existing in industries, such as the Zener diode approach, etc. The analysis of those schemes is not covered in this book. The interested reader should consult other reference books. The approach used here can always be easily extended to other schemes.

Figure 6-11(a) shows a generic representation of a resistive divider, and Figure 6-11(b) shows a generic representation of a capacitive divider. There are pros and cons for both schemes. For the resistive divider feedback scheme, the pros can be generalized into two points: First, it is very simple in terms of implementation. The divided value, V_{div1}, is only dependent on the ratio of resistors, R_1 and R_2, as shown in Equation 6-13.

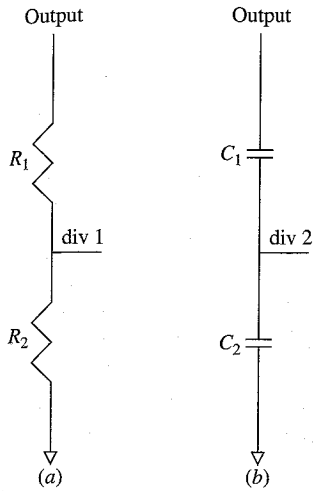


Figure 6-11 Resistive and capacitive dividers.

It is less susceptible to the variations of the process and the temperature. Second, the layout of the resistive divider can be relatively small in layout area if a high sheet resistance material is available:

$$V_{div1} = \frac{R_2}{R_2 + R_1} \times V_{output}$$
$$V_{output} = \left(1 + \frac{R_1}{R_2}\right) \times V_{div1} \quad (6\text{-}13)$$

However, the cons for the resistive divider feedback scheme can also be generalized into two points: First, the resistive divider always has *RC* delay through the feedback path, as shown in Equation 6-14. Because the resistor is used, it has parasitic capacitance associated with it. The *RC* time constant causes a phase shift between the output and feedback control signal. It makes the control of output regulation accuracy a challenge. Second, using a resistive divider always consumes DC power in regulation. As shown in Equation 6-15, current is always flowing from the output of the charge pump to the ground:

$$T_{delay} = \frac{1}{2} R_1 C_{parasitic} \quad (6\text{-}14)$$

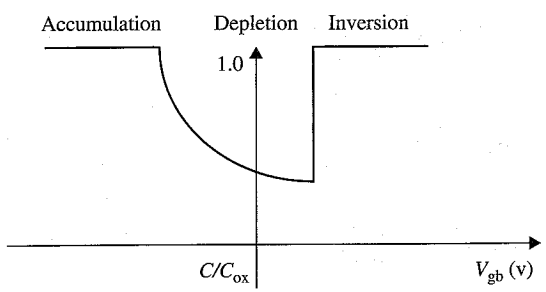
$$I_{regulation} = \frac{V_{output}}{R_1 + R_2} \quad (6\text{-}15)$$

The counterpart of the resistive divider is the capacitive divider. It is based on the theory of conservation of charge and assumes that the middle node between two capacitances can be initialized properly to 0 V when the output voltage also starts at 0 V in this analysis.

The benefit of the capacitive divider is the faster feedback control. In the resistive feedback control, any ΔV change on the output node will take time proportional to $\tau = RC$ for the divided node V_{div1} to respond. In the capacitive divider feedback path, because the resistance associated with each capacitance is generally very small, the divided node V_{div2} will respond to ΔV change on the output almost instantaneously. However, the cons of the capacitive divider are the accuracy of capacitance and the area penalty. On a chip, there are many different types of capacitance. A few of them that can be used are gate capacitance, diffusion capacitance, and metal capacitance. All of them have an issue with the accuracy. Because the final regulation level is strongly based on the capacitance ratio, as in Equation 6-16, any change in the value of capacitance will cause offset to occur. In terms of gate capacitance, such a chip has the highest capacitance per unit area. The CV curve of any transistor can be divided into depletion, inversion, and accumulation regions, as shown in Figure 6-12.

$$V_{div2}C_2 + (V_{div2} - V_{output})C_1 = 0$$
$$V_{div2}(C_2 + C_1) = V_{output}C_1 \qquad (6\text{-}16)$$
$$V_{output} = \left(1 + \frac{C_2}{C_1}\right) \times V_{div2}$$

With MOSFET gate capacitance, the effective capacitance between gate and substrate varies with the MOSFET biasing conditions.[8] In order to use the gate capacitance of the MOSFET for regulation purposes, we need to consider capacitance, initial, intermediate, and final conditions. As shown in Figure 6-12, this involves either avoiding the depletion region or calculating accurately the effective capacitance based on the total charge



stored in the entire operating range. However, the latter part is still not a guarantee of accuracy due to process and layout variations.

Diffusion capacitance is another type of capacitance that can be found on silicon. The calculations are given in Equation 6-17 and Equation 6-18. The junction capacitance is a function of process and biasing voltage. It is not suitable for regulation purposes either due to nonlinearity characteristics.

$$C_j = \frac{C_{j0}}{\left[1 + \frac{V_{back}}{V_{bi}}\right]^m} \qquad (6\text{-}17)$$

$$C_{j0} = A\sqrt{\left(\frac{\varepsilon_s q}{2}\right)\left(\frac{N_A N_D}{N_A + N_D}\right)\left(\frac{1}{V_{bi}}\right)} \qquad (6\text{-}18)$$

The last category is based on the metal capacitance. One of the drawbacks of metal capacitance is the dielectric material thickness. It is common for the field oxide thickness to be 10 to 20 times larger than the thickness of gate oxide. The penalty is the layout area. A relatively large layout size is needed for capacitance with some good accuracy. Another drawback is that metal capacitance varies with the process. Dielectric material thickness for the interconnection can vary easily more than ±20%. The capacitance varies from chip to chip and lot to lot. For a capacitive divider in general, the parasitic capacitance of the interconnection could have a significant impact in terms of the final regulation level. No matter how accurate the post-layout *RC* extraction or the manual estimate of *RC* during design phase, the final silicon can always require layout mask changes to take care of offset due to any unexpected parasitic capacitance on real silicon.

6.3.2 Regulation controls

With the feedback signal available, how would the regulation scheme control the noise on the output of charge pump? Many kinds of regulators are available. For charge pump design, there are several commonly used types in the industry.

The first type is the shunt regulator. This type of regulation is used commonly when large output current is required from the charge pump. The pump would be operating continuously. If the output exceeds the regulation level, the shunt path is turned on to discharge any extra charge. If the output falls below the regulation level, the shunt path is turned off, and the output can be continuously charged up by the pump.

The second type is the voltage regulator. Either an NMOS or a PMOS device is used to connect the pump output and the actual load circuits.

from the pump output to the load circuit based on the feedback control signal. This type of regulation is commonly used where the pump output current required is low.

The third type is an on/off regulator. Whenever the pump output falls below the regulation level, the pump is turned on to deliver charge to its output. Once the regulation level is exceeded on the output node, the feedback control signal turns off the pump and stops the supply. This type of regulation targets low pump output current design.

6.3.3 Noise control for regulation

In addition to the regulation schemes mentioned so far, one more factor should be considered regardless of which regulator is chosen. The noise on the output of the charge pump near regulation definitely depends on the feedback control and the regulation speed. This is similar to the implementation of any other analog circuits. However, for the charge pump design specifically, a large percentage of this noise is due to the imbalance of output power of the charge pump and the active load power consumptions.[9]

For example, charge pump operation has two phases: ramp-up and regulation. When a high-voltage pump is in ramp-up phase, it needs to charge up the decoupling capacitance (plus parasitic capacitance) as well as supply the active load current and regulator current (if any). In regulation phase, the pump needs to supply leakage current, regulator current, and active load current. The difference in power consumption between these two phases is the DC current used to charge the decoupling capacitance.

Ripples are minimized if the pump output power and the power consumed on the output node can be balanced. During the ramp-up phase, the pump is operating at full strength. It requires the output to reach the regulation in the required output settle-down time. Once the output approaches its regulation level, either part of the pump boosting capacitance, the clock driver strength, or the clock frequency can be adjusted dynamically to scale down the power of the pump in regulation to match the magnitude of power consumed. If multiple regulation levels are needed by design, the pump capacitance, clock driving strength, or clock frequency can be adjusted accordingly to allow the pump to always deliver the power just consumed by the output. This technique should always be considered first over other approaches.

6.4 Power Consumption versus Pump Performance

Up to this stage, all the topics covered so far are about how to design the charge pump to maximum its output power capability. There is no mention of the power consumption of the charge pump itself. As handheld devices have proliferated in consumer electronics, the demand for

low-power or power-efficient design has been strong. Devices need to deliver higher performance, occupy smaller die size, and even consume less power. This presents many challenges for pump designs. The power consumption issue is even more serious for low-voltage supply design, such as when the chip supply migrates from 5 V to 3 V, or from 3 V to 1.8 V. So how does one reduce the total power consumption of the charge pump? To refresh your memory, let us look at pump I-V characteristics.

Figure 6-13 shows the load line for a charge pump. Treating the charge pump as a black box, as shown in Figure 6-13(a), when we look at the charge pump output port, the Thevenin equivalent circuit of the charge pump can be represented by Figure 6-13(b). It acts as a battery with supply voltage V_{reg} and with internal impedance R_{eq}. The I-V characteristic of the charge pump is plotted in Figure 6-13(c). The load line of the charge pump can be generated by sweeping a voltage source on the output of charge pump and measuring the current going through the voltage source simultaneously. This is the I-V characteristic of the charge pump.

Figure 6-13(c) shows three I-V curves representing three different pump designs. Each point on the curves represents an operating point of the pumps. At the specified regulation voltage, V_{reg}, three curves give three different output currents. Curve 3 has the largest output

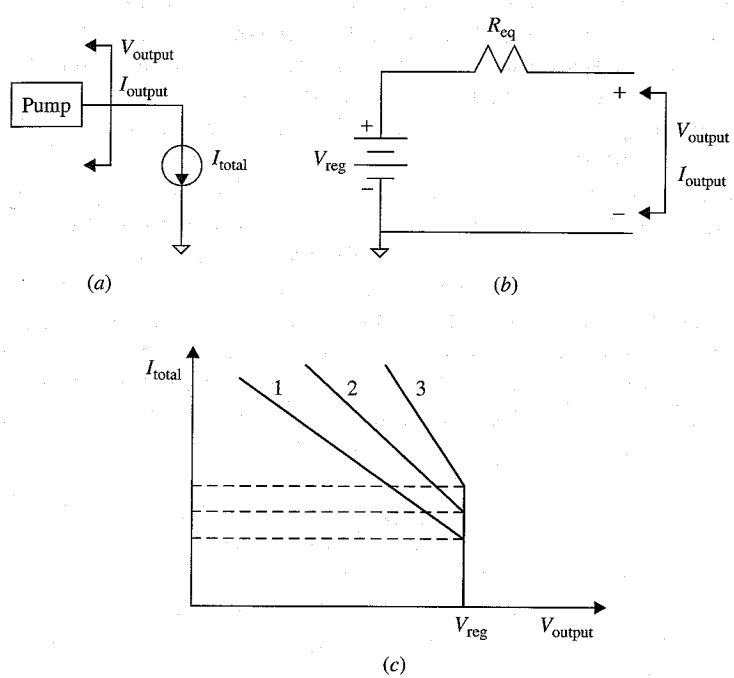


Figure 6-13 The load line for a charge pump.

current, whereas Curve 1 has the lowest output current. The slope of the pump I-V curve depicts the output impedance of the charge pump. This slope for the Dickson charge pump can be described using Equation 6-19.

$$R_s = \frac{n}{(C+C_s)f_{osc}} \tag{6-19}$$

The efficiency of the charge pump is inversely proportional to the impedance of the charge pump. From the architecture point of view, if the number of pump stages can be reduced to meet the same design requirement, this results is less parasitic capacitance along the stages, less voltage drop due to threshold voltage through each stage, and lower pump impedance. This means that more charge can be used to do the work on the output. Fundamentally, to have less power consumed, the V_t cancellation scheme is preferred over the Dickson charge pump approach. A higher pump clock scheme is preferred over Dickson's charge pump approach, too. Therefore, it is important to choose the pump architecture in the early stages of the design.

Next let us review the effect on the I_{cc} current of the charge pump from the approach of using a faster pump clock and a smaller pump capacitance. In the earlier chapters, the goal was to achieve the smallest die size for the pump design. One approach is to increase clock frequency and reduce pump boosting capacitance. However, this optimization can cause the pump power consumption to increase. Figure 6-14 shows the last stage of the charge pump on the left, with its associated waveform near regulation level on the right.

As the clock switches from 0 V to V_{clock} to pass additional charge to the output through the diode-connected NMOS, the pump output is boosted higher by ΔV in Figure 6-14. This change allows the output voltage shift from point a, to the point that just exceeds the regulation, to point b. By discharging the output through either active load current, leakage current, or regulation current, the output is brought down from point b to point c. Then the previously described operations of charging and discharging repeat in regulation. As shown in Figure 6-14, the charge

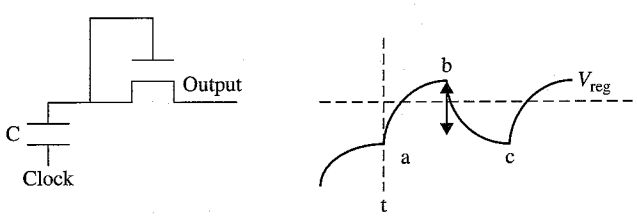


Figure 6-14 Impact of smaller pump capacitance on I_{cc}.

transferred to output between point a and point b is $\Delta Q = C \times \Delta V$. The amount of time it takes to discharge the output from point b to point c is given in Equation 6-20.

$$t = \frac{\Delta Q}{I_{total}} = \frac{\Delta Q}{I_{active} + I_{regulation} + I_{leakage}} \tag{6-20}$$

The power consumption of the charge pump is inversely proportional to the pump output discharging time between point b and point c. To reduce the overall pump power consumption in operations, the pump should be active less frequently. The first factor in Equation 6-20 that could increase t is ΔQ. ΔQ is the total charge being transferred in one clock cycle. The product of $\Delta Q = C \times \Delta V$ is one of the factors that determines how frequently the charge pump will be turned on in operation. The larger the amount of charge transferred in a fixed period, the longer the pump can stay idle and save power. However, the goal of pump design is to have a smaller layout area and run the circuit at a higher clock frequency. At regulation level, ΔQ dumped to the output by this design approach would be smaller compared with the other approaches. If I is unchanged, the designer might not be able to achieve the smallest die size while minimizing the I_{cc} of the charge pump at the same time. The tradeoff between the size of the charge pump and the power consumption of the charge pump must be decided.

Now on to the second factor concerning I. On the output of the charge pump, there are three current components: active current I_{active}, regulation current $I_{regulation}$, and leakage current $I_{leakage}$. The active current and leakage current are components that probably cannot be changed. Once the design architecture is fixed, I_{active} and $I_{leakage}$ are determined fixed by the operation. The leakage current discussed here is mainly due to reverse-biased junction leakage. The only component left is $I_{regulation}$. We discussed the regulation schemes earlier in this book. The benefit of the capacitive divider feedback is significant here because there is no DC current through the regulation path. This scheme is the best for reducing the extra component $I_{regulation}$.

How about resistor divider regulation? The resistor divider has many benefits: ease of implementation, dependence of the output regulation only on the resistor ratio, and less temperature variation or process variation impacts. One of the key drawbacks is the DC current through the regulation path. Is it possible for a design to have the benefit of using the resistive divider feedback scheme but not at the expense of the power consumption issue? The answer is yes.

Figure 6-15 shows a generic charge pump with resistive divider feedback regulation. The sampled voltage is compared with the on-chip band gap reference voltage V_{ref}. The comparison result is then used to enable

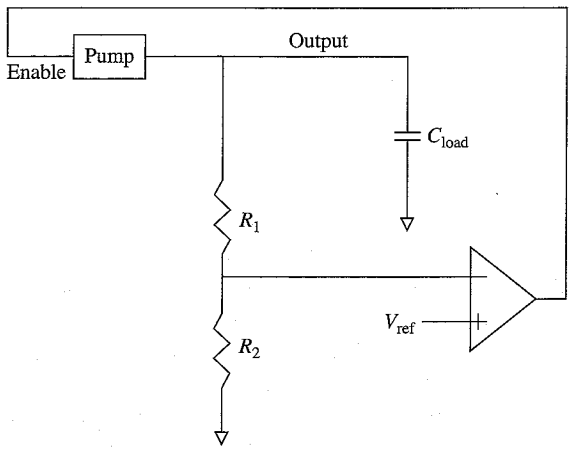


Figure 6-15 Resistive divider feedback regulations.

for regulation control. Figure 6-16 shows the output waveform of the charge pump in regulation. The frequency of the pump being enabled is a function of the load current, as shown in Equation 6-21. Because the active load current and the leakage current cannot be reduced via design technique after architecture is fixed, they are not of concern in this discussion. The only current of interest here is the DC current through the resistor chain:

$$f_{osc} = \frac{1}{t_1}$$
$$t_1 = \frac{Q}{I_{total}} = \frac{C_{load}\Delta V}{I_{regulation} + I_{active}} \quad (6\text{-}21)$$

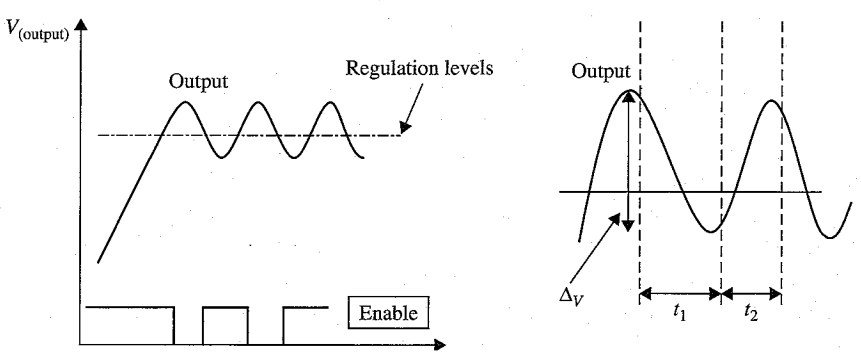


Figure 6-16 Pumping frequency in regulation.

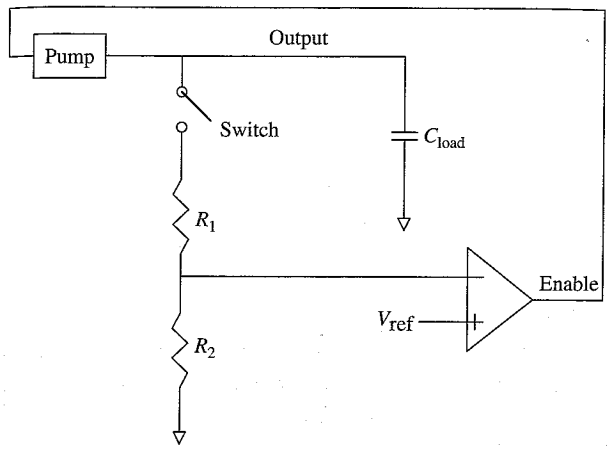


Figure 6-17 Modified resistor divider feedback scheme.

Figure 6-17 shows the modified resistor divider feedback scheme. As shown in the diagram, an additional high-voltage switch connects the output node to the resistor divider chain. In operation, as long as no current load is connected to the pump, or if the pump output reaches regulation level, the switch shown can be turned off, as well as the charge pump. The DC current path between the output node and the ground node is disconnected. With the output node left floating, two scenarios can occur: First, the active current is negligible. Second, the active current is larger. If the active current (plus the leakage current) is negligible, the output of the charge pump may be left floating for the duration of the operation as long as the error on output voltage can be tolerated. If the active current is large, the regulation and the pump should be turned on at a predetermined frequency. Enabling the charge pump and regulation path allows the regulation level to be maintained on the pump output again. By reducing the DC current through the resistive divider, the frequency of the pump being enabled can be minimized, as well as the overall active pump power consumption. The switch in the diagram can be implemented by either NMOS or PMOS. The insertion point can vary depending on the operation requirement.

6.5 Charge Pump Area Efficiency

Sometimes designing a smaller charge pump does not necessarily mean the pump boosting capacitance per stage needs to be the smallest in size overall. The total area of the charge pump includes pump capacitance per stage, diode-connecting successive stages, and the peripheral supporting circuits. To make the pump capacitance small, one of the approaches is to speed up the charge pump clock frequency. While the

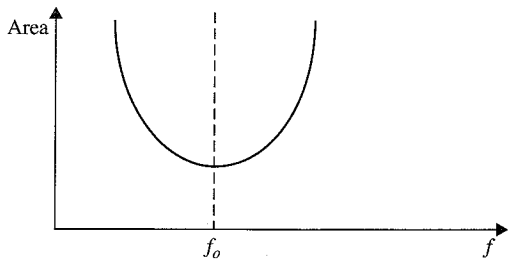


Figure 6-18 Pump layout area versus pump frequency.

clock frequency is increased, many supporting circuits have to be sized up to meet the high-speed operations. With the increased device sizing, the area consumed by transistor spacing, gate spacing, and routing area and spacing can increase proportionally. The area efficiency versus pump clock frequency is described in Figure 6-18.

Initially, as the pump frequency is increased, the total area of the pump layout should decrease. As the percentage of the overall capacitance decreases, the gain of area reduction is reduced while the frequency is increased. The most optimum frequency should be f_0, as shown in Figure 6-18. At this operating point, the circuit gives the best overall layout area. If the pump frequency is increased further, the extra layout area consumed by the larger device size, spacing, and routing would erode the gain from the reduction of the capacitive area. The overall layout area would increase as the frequency increases for $f > f_0$. Determining the optimum clock frequency, f_0, is difficult. Trial and error helps determine the optimum combination of the frequency and area.

6.6 Layout Requirements for Pump Design

Like many other analog circuits, charge pump performance is strongly dependent on layout. A designer can simulate the charge pump circuits drawn on the schematic with all the simulations shown to 100% satisfaction, but the real silicon may not function at all. Charge pump designers not only have to think about the circuits and simulations, but they also have to estimate the impact from the physical placement of capacitance, interconnections, and signal routings. Usually the physical layout of a charge pump requires a joint effort from both the design and layout departments concerning placement and routings.

Ideally, modeling every aspect of circuit components on the real silicon is desirable. However, in reality too much detailed modeling can distract the attention of the designers and lead to a very long simulation time. It is recommended that designers model only the necessary components. Designers should simplify the modeling as much as possible and always understand the tradeoffs, while using a keen sense of judgment to optimize the key parameters. Let us look at what layout factors should be

6.6.1 Parasitic capacitance

Parasitic capacitance is the unwanted capacitance associated with a physical design. It is "unwanted" because it may not be foreseen in the initial design phase. Normally the size of the circuit blocks, the distance the signals will travel, the layers of metals used, and the signals that will be routed nearby are not obvious at the beginning. The exact information cannot be extracted until the final physical layout is completed. For example, suppose a 5-stage charge pump is being designed, and each stage has an identical boosting capacitance of 10 pF. Unless the designer has estimated the parasitic loadings and put them into the schematic, the schematic view of the pump will have zero parasitic capacitance. On the physical layout, 10 pF may be drawn in the size of 50 μm × 50 μm. For the interconnection between stages, the wire distance could be 50 μm long or it could be 100 μm long. The exact information is not available in the early stages of the design.

Some parasitic capacitances have a greater impact on design. For example, the interconnection between stages should be as small as possible. This parasitic capacitance acts as extra loading in each pump stage and takes the charge away. If it became a larger percentage of the total boosting capacitance than expected, the pump efficiency will suffer. Another example of bad parasitic capacitance is associated with the boosting gate in the 4-phase charge pump design. The conduction of the transfer device is strongly dependent on how high the gate voltage can be coupled up. The size of the boosting capacitance for the gate is typically small. Any additional capacitance added to the gate will reduce the coupling efficiency and impact the circuit performance.

Figure 6-19 shows a cross-section view of a two-layer metal process and the associated parasitic capacitance. M_2 is the top metal layer, and M_1 is the bottom metal layer. The middle M_1 signal is the focus of the discussion. M_1 is normally used for local connections between poly, substrate, and M_2. On most occasions, M_1 passes over the P-substrate separated by field oxide.

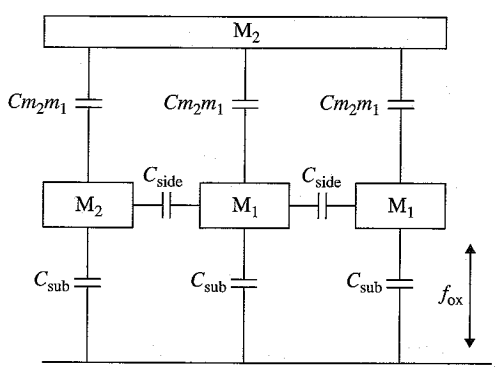


Figure 6-19 Parasitic capacitance of signals.

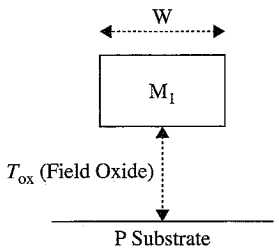


Figure 6-20 Parasitic capacitance of signals.

The first component, C_{sub}, is described in Equation 6-22 as the parallel plate capacitance per unit length between M_1 and the P substrate. The cross-section view of this structure is shown in Figure 6-20:

$$C_{sub} = \varepsilon(f_{ox}) \times W(m_1) / T_{ox}(f_{ox}) \tag{6-22}$$

The second component is C_{side}, as shown in Figure 6-21. It is actually made of three components, as described by Equation 6-23. The final size of C_{side} is given in Equation 6-24:

$$C_{m_1 m_1} = \varepsilon \times \text{Height}(m_1) / \text{Space}(m_1) \tag{6-23}$$

$C_{fg}(m_{1,sub})$ fringing capacitance from M_1 to P substrate

$C_{fg}(m_1 m_2)$ fringing capacitance from M_1 to M_2

$$C_{side} = 2\left[C_{m_1 m_1} + C_{fg}(\text{m}_{1,sub}) + C_{fg}(m_1 m_2)\right] \tag{6-24}$$

Modeling of the exact fringing capacitance is very difficult. There are various approximation formulas to calculate fringing capacitance based

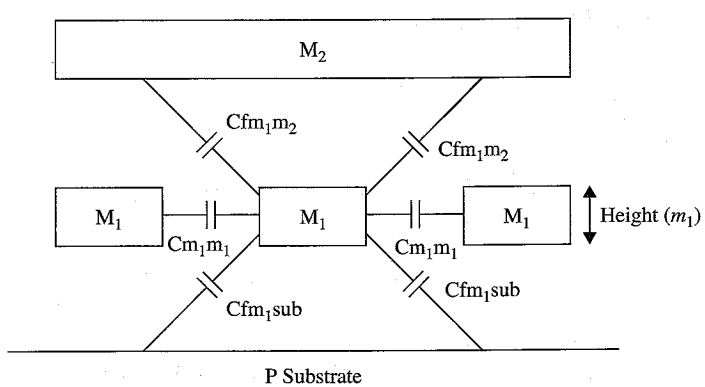


Figure 6-21 Side wall and fringing capacitance of M_1 signals.

on different simplifications. However, for real layout with various patterns, it is a difficult job to look at every layout pattern in detail. It is recommended that the designers find a method they feel most comfortable with in practice, try it, and refine the estimation with real silicon results. 3-D extraction software is available for sub-micro post-layout analysis. It is a good practice to do a post-layout analysis and compare the derived result with the one with initial parasitic estimates.

6.6.2 Miller effect (parasitic capacitance)

Bad parasitic capacitance would degrade the pump performance from the initial design target. Sometimes this effect can be amplified due to the nearby signals and layout patterns.

On real silicon, there are many occasions of group of signals where one signal is switching in one direction and the neighboring signals switch in the opposite direction, as shown in Figure 6-22(a). Let us analyze the effect of capacitance as seen from the middle node, signal b. It is better to analyze the circuit for the charge point of view. As shown in Figure 6-22(b), there is a single capacitance of C, and two nodes of the capacitor are connected to n_1 and n_2. Assume n_1 is switching from 0 to V_{delta}, and n_2 is switching in the opposite direction, from V_{delta} to 0, simultaneously.

As shown in Figure 6-22, the static capacitance between node n_1 and n_2 is C. Dynamically, if two nodes of the capacitance switch in the opposite direction, the equivalent capacitance seen from either node would be doubled. The derivations are given in Equation 6-25 and Equation 6-26. This phenomenon is called the Miller effect. From the charge point of view, the total charge needed to charge up the capacitance on the middle node is twice the amount compared with the case where one terminal

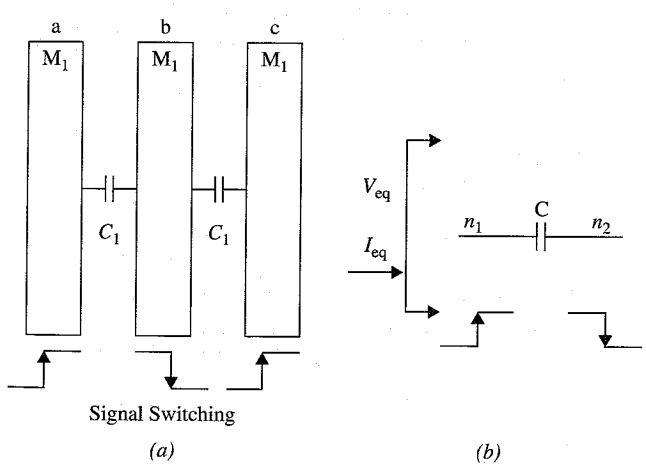


Figure 6-22 Miller effect of capacitance.

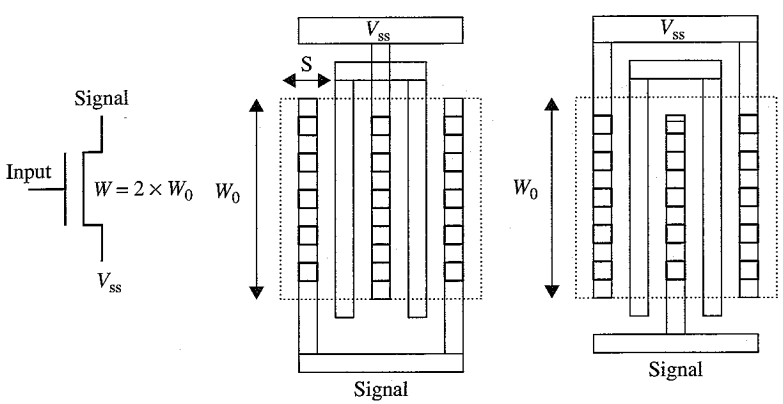
is held at small signal ground. To express this another way, if viewed from node n_1, as node n_2 swings in the opposite direction, there would an equivalent capacitance of $2C$ connected between n_1 to the small signal ground. In pump design, a lot of clocking signals are used. Many of them have to be routed to many different blocks. Without careful planning, the equivalent parasitic capacitance seen by any clock driver could be easily doubled in real silicon. This would cause a performance downgrade.

$$\begin{aligned} I_{eq} &= C\frac{\partial(V_{n1}-V_{n2})}{\partial t} \\ &= C\frac{\partial(V_{delta}-(-V_{delta}))}{\partial t} \\ &= C\frac{\partial(2V_{delta})}{\partial t} \\ &= (2C)\frac{\partial(V_{delta})}{\partial t} \end{aligned} \tag{6-25}$$

$$C_{eq} = 2C \tag{6-26}$$

6.6.3 Junction capacitance

Another simple but not very obvious layout technique is shown in Figure 6-23. Figure 6-23(a) is a schematic view of an NMOS device with a width equal to $2W_0$. The source of the NMOS is connected to V_{ss}, and the drain is connected to the output "Signal." Two different layout views are compared in Figure 6-23(b) and Figure 6-23(c). Both of these would pass LVS (layout versus schematic comparison during tape-out process) because both of them match the device dimensions between the layout



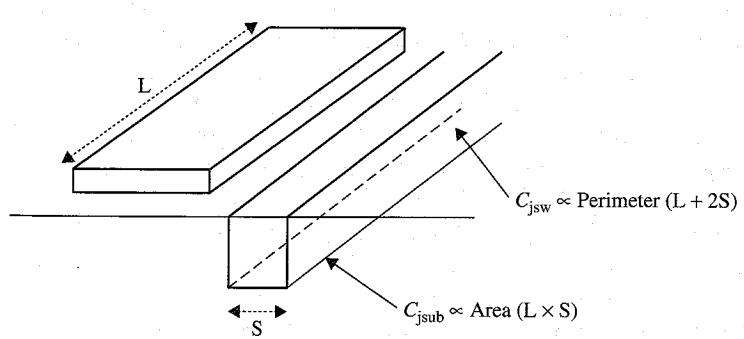
view and the schematic view. However, one layout is definitely better than the other. In Figure 6-23(b) and 6-23(c), the gate of NMOS transistor is drawn with two legs of ploy. Each gate has a width of W_0. The difference between the two layout views is which junction in the layout is connected to the output "Signal" and which junction is connected to the power supply. In Figure 6-23(b), the junction located in the center is connected to the power supply V_{ss}. The output signal is connected to both left and right junctions of the device. In terms of junction capacitance, this can be divided into two components: C_{jsw} and C_{jsub}. C_{jsw} is the sidewall capacitance of the junction toward the well (or substrate). It is proportional to the perimeter of the junction. C_{jsub} is the area capacitance of the junction toward the well (or substrate). It is proportional to the total area of the junction. The cross-section diagram for the junction capacitance is shown in Figure 6-24. C_{jsw} and C_{jsub} are differentiated in this figure.

In Figure 6-23(b), the total junction capacitance connected to V_{ss} is given by Equation 6-27, and the total junction capacitance associated with the signal is given in Equation 6-28. The sidewall capacitance between the junction and the gate is not counted in the calculation. Because a channel is formed in between, L is not counted toward the calculation in terms of sidewall capacitance. Comparing the difference, the signal side junction capacitance is more than double the total capacitance associated with the power signal.

$$\begin{aligned} C_{b1} &= C(\text{side wall}) + C(\text{substrate}) \\ &= C_{jsw} \times 2 \times s + C_{jsub} \times s \times L \end{aligned} \tag{6-27}$$

$$\begin{aligned} C_{b2} &= C(\text{side wall}) + C(\text{substrate}) \\ &= 2 \times [C_{jsw} \times (2 \times s + L) + C_{jsub} \times s \times L)] \end{aligned} \tag{6-28}$$

Let us examine the junction capacitances in Figure 6-23(c). The total junction capacitance on V_{ss} is given by Equation 6-29, and the total



junction capacitance on the output node Signal is given by Equation 6-30. In general, the parasitic capacitance on power net is good capacitance. Capacitance associated with power supply lines is for decoupling purpose. It can filter out the power supply noise. On the other hand, the parasitic capacitance associated with signal nodes is bad. It introduces extra delay in signal transition as well as higher power consumption for the charging and discharging of unwanted capacitance. If possible, the mask designer should always choose the approach in Figure 6-23(c) over that in Figure 6-23(b). The designer should also keep checking the layout to minimize any unwanted parasitic capacitance on all internal nodes of the pump.

$$\begin{aligned} C_{c1} &= C(\text{side wall}) + C(\text{substrate}) \\ &= 2 \times [C_{\text{jsw}} \times (2 \times s + L) + C_{\text{jsub}} \times s \times L)] \end{aligned} \tag{6-29}$$

$$\begin{aligned} C_{c2} &= C(\text{side wall}) + C(\text{substrate}) \\ &= C_{\text{jsw}} \times 2 \times s + C_{\text{jsub}} \times s \times L \end{aligned} \tag{6-30}$$

6.6.4 Improving layout efficiency per unit area

To minimize the total layout area occupied by the capacitance and circuits, design techniques can be used to improve pump efficiency. On the other hand, careful placement and routing of signals can also help reduce the unwanted parasitic capacitance. Smaller parasitic capacitance allows for a smaller charge pump. In addition to the two methods just mentioned, an additional method may be used at no extra cost.

Normally capacitance is built based on the gate capacitance. Gate oxide thickness on silicon is usually 1/10 ~ 1/20 or even less than the thickness of the field oxide (or field dielectric). That is why the boosting capacitor in the charge pump is usually built over thin gate oxide or thick gate oxide devices. One of the great layout techniques can be used for free. If gate capacitance can be used, why not utilize the dielectric material vertically above the gate to achieve better utilization of the silicon area?

Figure 6-25 shows one example of how to build extra capacitance vertically above the gate area. In this example, two metal layers are available in the process. The main component of the capacitance is the gate capacitance between the poly silicon and active area. The source and drain are connected together by M_1 in the diagram. As shown, M_1 lines connecting source/drain overlap the poly silicon gate. In addition, the poly gate is connected to M_2, which overlaps the source/drain M_1 lines underneath. If we consider the dielectric material between M_1/polysilicon

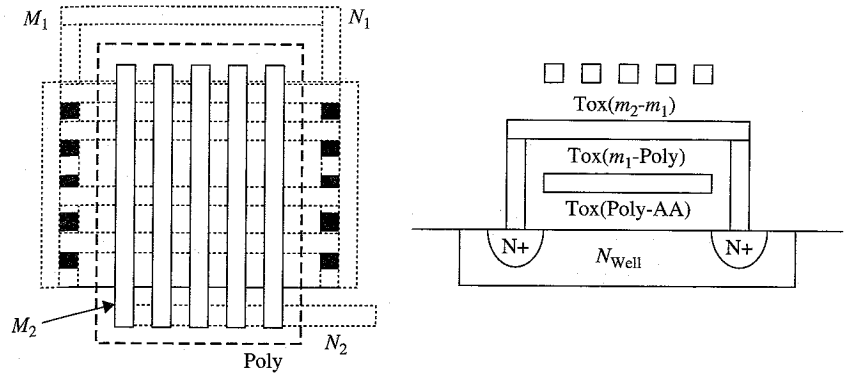


Figure 6-25 Improved capacitance layout.

and M_2/M_1, the additional capacitance added over the same active area overlapped by polysilicon can be roughly estimated as follows:

$$\text{ratio} = \frac{T_{\text{ox}}(m_1/\text{poly}) + T_{\text{ox}}(m_2/m_1)}{T_{\text{ox}}(\text{poly}/AA)} \times 100\% \tag{6-31}$$

If the gate oxide thickness is 100 A, the thickness between M_1 and the poly silicon is 1000 A, and the dielectric thickness between M_2 and M_1 is 1000 A, using Equation 6-31 we can estimate that nearly 20% more capacitance can be achieved over the same gate area. If the process allows more metal layers, then the ratio in Equation 6-31 could be further increased. The previous assumption is based on pure parallel plate capacitance between nodes. If fringing capacitance is more dominant than the parallel plate capacitance, the layout can be changed in a little bit different way to achieve an even greater benefit.

In Figure 6-26, the layout pattern is changed a little bit to allow a larger fringing field component to be realized. N_1 and N_2 nodes are

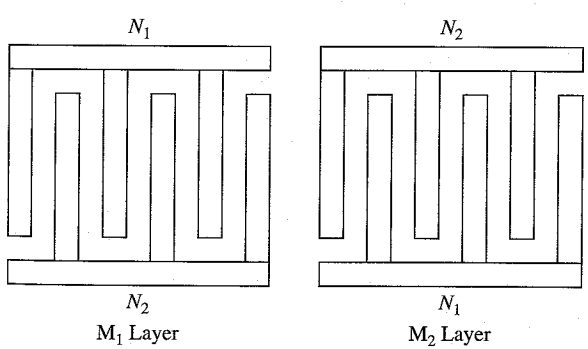


Figure 6-26 Serpentine layout technique.

both laid out in M_1 and M_2 layers. Unlike in Figure 6-22, N_1 takes only the metal 1 layer and N_2 takes both the poly and metal 2 layers. In Figure 6-26, both signals are laid out using identical layers. The forking patterns are used between these two signals at each one layer. When moving to the upper metal layer, the pattern is swapped between these two signals. For each segment of the signal, the neighbor segment is always associated with the other node of the capacitance. In practice, this layout pattern could yield more capacitance per unit area than a simple parallel plate capacitance.

6.6.5 Well resistance

In pump design, large size capacitors are often encountered. When the capacitance is drawn on the layout, the effective capacitance may not be the same as what is expected in simulation. For example, assume that 2000 μm^2 NWCAP (n-type well to poly capacitance) capacitance needs to be drawn. If the height of the poly drawing is limited to 20 μm, then the width of the poly must be drawn 100 μm wide. This layout drawing is shown in Figure 6-27.

If the layout is translating a 2000 μm^2 NWCAP into a 20 $\mu m \times 100$ μm poly over the active area, as shown in Figure 6-27, there is a problem with the device on silicon matching the original design expectations. If a 2000 μm^2 NWCAP is instantiated in the schematic, the SPICE model would probably treat it as a normal transistor with $L = 20$ μm and $W =$ 100 μm in an ideal situation. In reality, the mismatch between simulation and silicon are due to poly resistance and WELL resistance.

Figure 6-28 is a cross-section view of NWELL capacitance. Because NWELL capacitance is operating in the accumulation region, the charge needs to be supplied from NWELL TAP to the substrate underneath the gate. All NWELL TAPs are connected together and are of equal potential. However, the surface potential of the NWELL underneath is not equal. There is potential difference between nodes near and far away from the NWELL TAP due to the NWELL resistance.

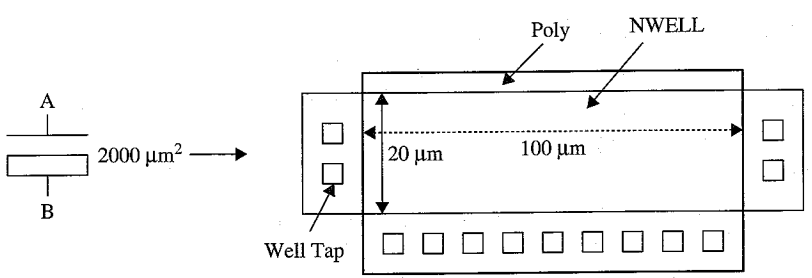


Figure 6-27 Large capacitance layout.

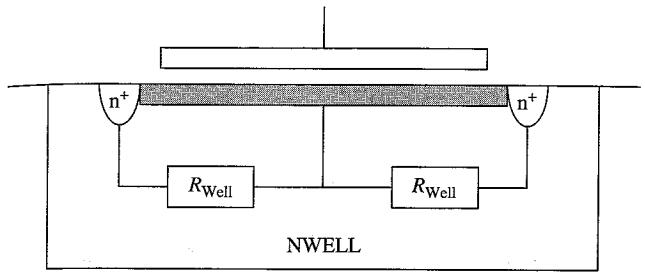


Figure 6-28 Cross-section of NWELL capacitance.

On the other side of the NWELL capacitance, the poly gate is contacted at one side. If the poly can be dissected into a group of parallel polygons, then the gate resistance can be easily visualized, as shown in Figure 6-29. The potential voltages near the gate contacts and far away from the gate contacts are not at the same potential. The capacitance is a simplified view of two individual nodes being separated by dielectric materials. Those two nodes are reduced from any shape of the surface if equal potential happens. If in reality the condition for equal potential does not hold, the actual effective capacitance will be smaller.

As shown in Figure 6-30, a simplified capacitance (a) in a real layout should be represented by the capacitance with two resistors in serial, as shown in (b). Depending on the clock frequency on n_a/n_b and the RC time constant, the equivalent capacitance could be very different. Let us review two extreme cases to see how resistance could impact the equivalent capacitance on real silicon.

In the first case, if $RC < T_{cycle}$, then within one clock cycle almost 100% of charge can be propagated and accumulated over the two plates of the capacitance. There is no degradation of efficiency for charging and

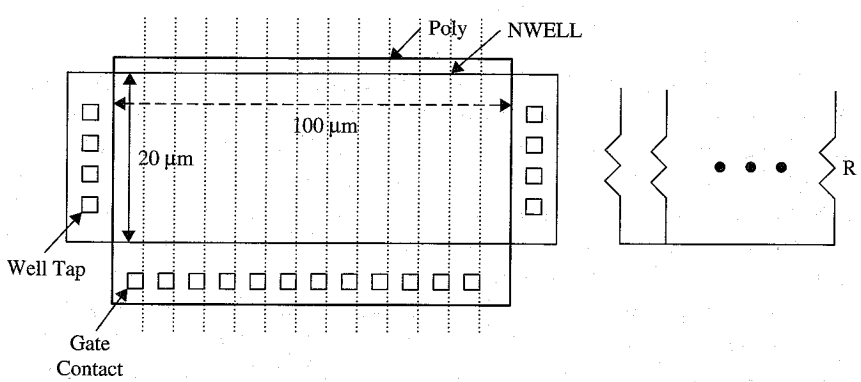


Figure 6-29 Poly gate resistance.

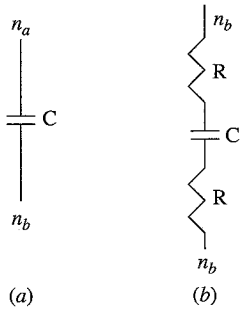


Figure 6-30 Capacitance with resistance in play.

discharging of capacitance. That is why the value of the effective capacitance is close to the given capacitance, as shown in Equation 6-32.

$$C_{eq} = C$$
$$Q_{total} = C_{eq} V = CV \tag{6-32}$$

In the second case, if $RC = T_{cycle}/2$, then the effective capacitance across the nodes would be different from the capacitance specified. As shown in Equation 6-33, if RC is equal to the half clock cycle period, then only 63% of the charge can be transferred at end of the clock cycle. The equivalent capacitance in this case is only 63% of the given value.

$$V(t) = V \times (1 - e^{-t/RC})$$
$$V(t_{cycle}/2) = V \times (1 - e^{-1}) = 0.63 \times V$$
$$Q_{total} = V(t_{cycle}/2) \times C = (0.63 \times C) \times V \tag{6-33}$$
$$C_{eq} = 0.63 \times C$$

This is a very critical point in circuit design and layout planning. Choosing the correct clock frequency and determining the minimum repeating unit of drawn capacitance play critical roles in the real silicon performance. In pump layout, two approaches can be used. The first approach is to determine the minimum layout unit of the capacitance. Any large size capacitance would be broken down into groups of minimum units to avoid large resistance in gate or in substrate. The second method involves having open holes in the middle of the drawing layer to allow metal contacts to be dropped over the layer of interest to reduce the serial resistance. Both methods should work effectively on real silicon.

6.6.6 Critical signal width

Charge pump performance can be impacted by wire RC delay. In the designing phase, it is important to estimate the routing distances of these critical signals. Based on the estimation, it is easy to plan the interconnection layer and the dimensions for the signals.

So how does one pick the right interconnect layer for the signals? How does one determine whether the width of the signal can meet the design requirements? With a given process, the sheet resistance and unit capacitance for different materials and different dimensions can be obtained. It is common that the top metal layer processed at a later stage has smaller sheet resistance compared with one from earlier process stages. The unit capacitance is usually provided for many conditions.

It is common to choose two lower metal layers for local interconnections within the layout blocks, such as using metal 1 and metal 2. Then upper metal layers can be chosen for the global connection, such as metal 3. The pump layout can be drawn as chains of repeated pump units. The size of each unit is relatively small in terms of block width and block height. The connection within the pump unit, that is the local interconnects, can use metal layers with relatively high sheet resistance and unit capacitance. Signals connecting globally should use a metal layer with low sheet resistances. The clock signal's drive pumps should connect to all pump units, and the logic effort of the signal is very high. Using top metal layers with low sheet resistance can reduce the signal delay.

After the metal layers for the signals have been picked, it is important to determine the width of the critical signal lines. How wide should a signal be designed? Can we make it a minimum width to save some layout area? For logic control signals that are not speed critical, it is okay to adopt the minimum-width approach. This helps to reduce the power consumption and total layout area. For analog signals that consume AC/DC power, or for those that are speed critical, the widths of the signals need accurate calculation. For analog signals that burn DC power, we need to consider the current density of the interconnection and the IR drop of the signal line.

For analog signal lines that carry a lot of current, the width of the wire needs to be greater to counter the electron-migration problem. For wires conducting high current, the electron flow will push around the metal grains. Over a longer period of time, the metal wire will become thinner. If the width of the metal wire is not large enough, the wire will eventually be broken. Normally design rules provided by the foundry will specify the maximum current density allowed for each metal layer of a specified width. The designer needs to estimate the current density for these critical paths conducting a high current and determine the width of the signal line.

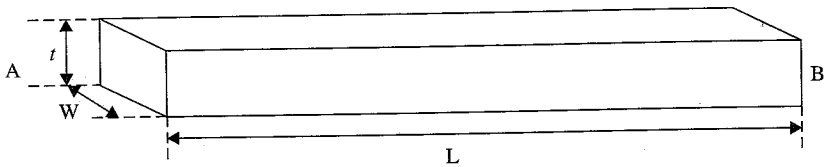


Figure 6-31 Wire delay between node A and node B.

RC delay is another critical concern for sizing up the width of critical metal lines. Figure 6-31 shows a metal interconnection between node A and node B in metal. The resistance of the wire from node A to node B is given in Equation 6-34.

$$R_{total} = \frac{\rho L}{tW} = \left(\frac{\rho}{tW}\right) L = R_{sheet} L \tag{6-34}$$

The thickness and the resistivity of the metal are fixed by process. It can be seen that the resistance of the metal line is directly proportional to L and inversely proportional to W.

In Figure 6-32, the cross-section view along the width of the metal wire shows how the electric field lines terminate on the substrate. In Figure 6-33, the wire capacitance can be simplified into two components: the parallel plate capacitance, C_1, and the fringing capacitance, C_2, on both sides.

As shown in Equation 6-35, C_{total} is proportional to the length of the metal wire. In terms of wiring delay, it was calculated by τ in Equation 6-36.

$$\begin{aligned} C_{total} &= C_1 + 2C_2 \\ C_{total} &= C_{1unit\_length} L + C_{2unit\_length} L \\ C_{total} &= C_{unit\_length} L \end{aligned} \tag{6-35}$$

$$\begin{aligned} \tau &= R_{total} \times C_{total} \\ \tau &= (R_{sheet} L)(C_{unit\_length} L) \\ \tau &= [R_{sheet}(C_{1unit\_length} + C_{2unit\_length})]L^2 \end{aligned} \tag{6-36}$$

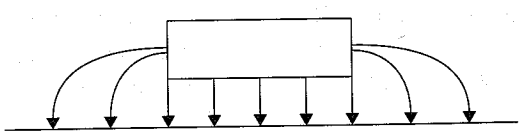
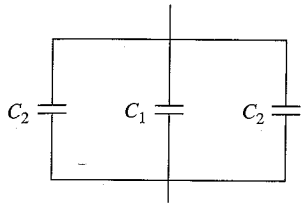


Figure 6-32 Wire parasitic capacitance.



C_1: Parallel Plate Capacitance
C_2: Fringing Capacitance

Figure 6-33 Simplified model of wire parasitic capacitance.

If the distance (L) between node A and node B is fixed by placement, can the delay through the metal wire be reduced by increasing the width of the wire? Let us assume there are two wires of the same length (L). One has width W and the other has width $2W$. The time constant of the first wire with width W is given in Equation 6-37. The time constant of the second wire with width $2W$ is given in Equation 6-38.

$$\tau_1 = [R_{sheet}(C_{1unit\_length} + C_{2unit\_length})]L^2 \tag{6-37}$$

$$\begin{aligned} \tau_2 &= \left[\frac{R_{sheet}}{2}\left(2C_{1unit\_length} + C_{2unit\_length}\right)\right]L^2 \\ \tau_2 &= \left[R_{sheet}\left(C_{1unit\_length} + \frac{1}{2}C_{2unit\_length}\right)\right]L^2 \end{aligned} \tag{6-38}$$

Comparing τ_1 and τ_2, τ_2 is less than τ_1. From a physics point of view, on the one hand, if the width of the wire is increased, the associated resistance would decrease proportionally with the width. On the other hand, increasing the width of the wire increases the parallel plate capacitance C_1 proportionally. C_2 is the component that is not changed with the width change. The total capacitance of the wire is increased at a slower rate compared with the rate of wire resistance being reduced. As shown, the delay through the wire can be reduced by increasing the width of the metal line.

A pump requires coupling capacitance to be driven by clock signals and transfer the charge within a given clock cycle. Because the equivalent capacitance seen by pump clocks is very high, it is important to design the clock signal with a proper layer and appropriate width to meet the frequency demands of the pump design.

6.6.7 Clock buffering

Clock signals need to be studied. As the chip size becomes larger, and power supplies trend down, the clocks used by the charge pump

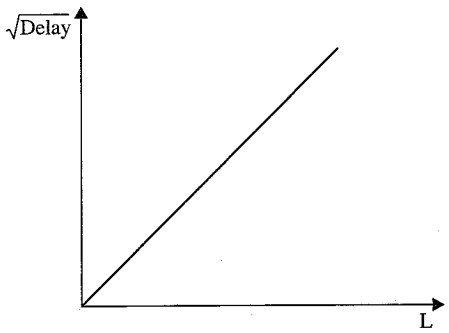


Figure 6-34 RC delay of a wire.

need to be buffered. As seen in the *RC* delay equation of the wire, τ is proportional to the second power of the wire length. If the length of the wire is doubled without any other parameters being changed, the propagation delay through the wire from one end to the other will be quadrupled. This relation is plotted in Figure 6-34.

This is alarming for pump design because it has tended to run at higher and higher clock frequencies. As shown in previous section, Critical Signal Width, widening the width of the signal line would help reduce the point to point delay. However, the improvement is limited. To conquer this delay issue, clock signals need to be buffered. Serials of buffers must be inserted along the signal path and gradually sized based on the load capacitance. There are two benefits of clock buffering. The first one is a reduction in wire *RC* delay. Instead of the square relationship between delay and wire length, the delay would grow linearly with the length of the wire. The second benefit is a reduction in the crossbar current in signal transition.

Figure 6-35 shows a model of the delay from node A to node B. The first stage driver is an inverter of size 1*x* that has an intrinsic delay of t_d. It is driving a load inverter of size 100*x*. The interconnection is a wire of length *L*. Figure 6-36 shows the simplified *RC* delay model. The driving transistor can be modeled as R_{wire}, and the metal wire is modeled by π networking. C_{load} is the input gate capacitance of the 100 × inverter driver. Delay from node A to node B can be calculated using Equation 6-39. As shown, π_1 is composed of three products. What if a

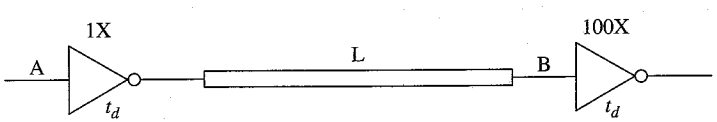


Figure 6-35 Delay in clocking.

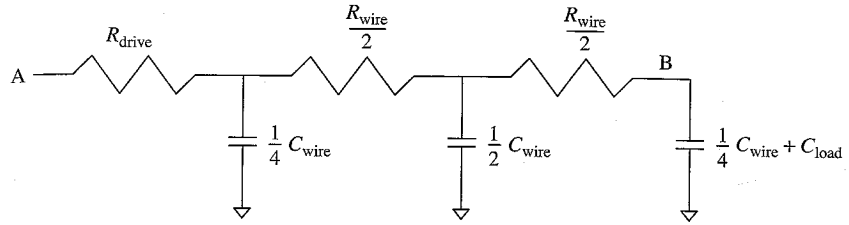


Figure 6-36 Simplified delay model.

driver of 10X strength is inserted in the middle of the wire, as shown in Figure 6-37?

$$\tau_1 = R_{drive}\left(\frac{1}{4}C_{wire}\right)+\left(R_{drive}+\frac{1}{2}R_{wire}\right)\left(\frac{1}{2}C_{wire}\right)+(R_{drive}+R_{wire})\left(\frac{1}{4}C_{wire}+C_{load}\right)+t_d$$

$$\tau_1 = R_{drive}C_{wire}+\frac{1}{2}R_{wire}C_{wire}+(R_{drive}+R_{wire})C_{load}+t_d \qquad (6\text{-}39)$$

$$\tau_1 = R_{drive}C_{wire}+0.5R_{wire}C_{wire}+100R_{drive}C_{in}+100R_{wire}C_{in}+t_d$$

The delay from node A to node B with a buffer of 10X strength inserted in the middle can be calculated using Equation 6-40.

$$\tau_{21} = \frac{1}{2}R_{drive}C_{wire}+\frac{1}{8}R_{wire}C_{wire}+\left(R_{drive}+\frac{1}{2}R_{wire}\right)(10C_{in})$$

$$\tau_{22} = \left(\frac{R_{drive}}{10}\right)\left(\frac{1}{2}C_{wire}\right)+\frac{1}{8}R_{wire}C_{wire}+\left(\frac{R_{drive}}{10}+\frac{1}{2}R_{wire}\right)(100C_{in}) \qquad (6\text{-}40)$$

$$\tau_2 = \tau_{21}+\tau_{22}+2t_d$$

$$\tau_2 = 0.55R_{drive}C_{wire}+0.25R_{wire}C_{wire}+20R_{drive}C_{in}+55R_{wire}C_{in}+2t_d$$

Comparing the difference between τ_1 and τ_2 in Equation 6-41, T_d is the intrinsic delay through the inverter. As shown, with the insertion of a

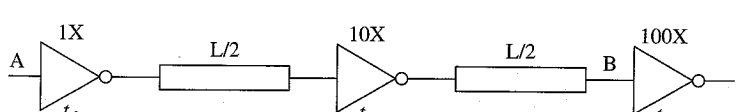


Figure 6-37 Simplified wire delay model after buffer insertion.

10X buffer in the middle of the wire, not only is the wire RC delay reduced by half, but the delay to drive 100X load is also reduced.

$$\begin{aligned}\Delta\text{delay} &= \tau_1 - \tau_2 \\ &= 0.45R_{\text{drive}}C_{\text{wire}} + 0.25R_{\text{wire}}C_{\text{wire}} + 80R_{\text{drive}}C_{\text{in}} + 45R_{\text{wire}}C_{\text{in}} - 2t_d\end{aligned} \tag{6-41}$$

Even though the purpose of this subtopic is to show the importance of clock and signal buffering, the actual process of buffer sizing and placement process can be followed from the previous chapter on logical effort.

6.6.8 Power bus and decoupling capacitance

Power bus and decoupling capacitance are probably the two things designers pay the least attention to. For example, the chip specification may state power supply V_{cc} to be 3 V ± 10% variation. For the worst-case design, the designer would probably choose V_{cc} = 2.7 V. Now, past experience from different sources may tell the designer to have an additional 0.2 V margin on the power supply for simulation. Eventually the simulation condition would be set to V_{cc} = 2.5 V for all worst-case simulations. For ground supply, it is common to set V_{cc} = 0 V.

In reality, how much supply noise do designers need to anticipate for a safe design? How does one guarantee that the margin set in simulation holds true in the real product? In order to answer these questions, it is important to understand where the power supply noise comes from. Figure 6-38 shows a simplified view of a chip. On the left side of the diagram is the package representation. The right side shows the internal circuits connected to the power buses. On the package

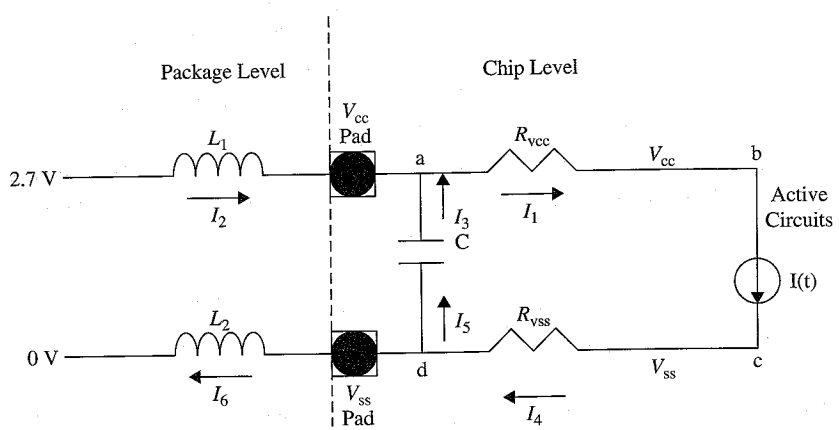


Figure 6-38 Simplified power dissipation diagram.

level, a bonding wire connects to the pad. Bonding wire has self-inductance, mutual inductance, capacitance, and resistance associated with it. In this discussion, only the inductance of the pin is a first-order parameter that concerns us. The other two parameters are secondary. On the chip level, there are circuits to perform certain operations and consume power. These have been lumped into $I(t)$, which is a function over time. From the bonding pads to the supplies of the internal active circuits, the power lines go through different metals, vias, contacts, and so on. All these interconnections can be lumped into R_{vcc} and R_{vss}, as shown in Figure 6-38. A lumped decoupling capacitance, C, is located between the internal V_{cc} and V_{ss} power lines.

Assume V_{cc} = 2.7 V and V_{ss} = 0 V are supplied from the system. If there is no operation internal on the chip, there will be no large active current flow on the V_{cc} or V_{ss} power buses. Therefore, the internal power supplies should be at the same level as the external supplies on the package level. The decoupling capacitance C should be fully charged. The total charge stored on this capacitance is $Q = 2.7\ C$.

What if the chip is switching from standby into active mode? Because circuits need to do their work and burn power, the charge would flow from V_{cc} through the devices, and eventually be discharged to V_{ss}. To quantify the power consumption, we use the $I(t)$ function to represent the current profile of all internal active circuits at any given time. Let us analyze the voltage drop on the V_{cc} power bus during active operation. The V_{ss} power bus analysis can be calculated in a similar fashion. Although the power bus is a distributed network of resistance and capacitance, this discussion uses a simplified view to lump all the power capacitance between node a and node d into C. The resistance is lumped together into R_{vcc} and R_{vss}. Kirchhoff's current law applies to both node a and node b. They are summarized in Equation 6-42.

$$\begin{aligned}&\text{Node } a\text{: } I_1 = I(t) \\ &\text{Node } b\text{: } I_1 = I_2 + I_3\end{aligned} \tag{6-42}$$

Two components contribute to the voltage drop from the external supply (2.7 V) to the power supply node near the active circuits. The first component is associated with the IR drop of the power bus. The second component is due to the bonding wire inductance.

$$\begin{aligned}V_{cc} &= \text{V}_{\text{external}} - L\frac{\partial(I_2)}{\partial(t)} - I_1R_{vcc} \\ V_{cc} &= \text{V}_{\text{external}} - L\frac{\partial(I_1 - I_3)}{\partial(t)} - I_1R_{vcc}\end{aligned} \tag{6-43}$$

As shown in Equation 6-43, the *IR* drop on the power bus due to I_1R_{vcc} is unavoidable. The charge has to move from the V_{cc} pad through the interconnect layers to the power bus near the circuits that consume the current. The larger the chip, the longer the interconnect layers are routed. The second component of the voltage drop is due to bonding wire inductance. Faraday's Law states that any change in the magnetic environment of a coil of wire will cause a voltage (EMF) to be "induced in the coil." No matter how the current change is produced, the voltage is generated in the direction to oppose the change of current. Because all charge has to be transferred in from the bonding wire, the average current consumed by circuits has to be same as the average current flow into the chip supply, as shown in Equation 6-44.

$$I_{2\text{average}} = I(t)_{\text{average}} \tag{6-44}$$

However, the key to Faraday's Law in this case is $\frac{\partial(I)}{\partial(t)}$, which is the gradient of current change through the inductor. The decoupling capacitance C added between node a and node d is acting as a low pass filter to smooth out the current change through the power bus's bonding wire inductance. The amount of charge needed by the active circuit still passing through the bonding wire is 100%, but the change of current amplitude over time is smoothed by the decoupling capacitance C.

Now let us revisit the topic of simulation margin for the power supply. As you can see, the actual supply voltage shown near the active circuit is different from the external supply voltage. The difference is given in Equation 6-45.

$$\Delta V_{cc} = L\frac{\partial(I_1 - I_3)}{\partial(t)} + I_1R_{vcc} \tag{6-45}$$

This difference determines how much of a power supply simulation margin a designer should adopt in circuit simulations. In order to guarantee that on the silicon the power supply variation is not worse than the value given in Equation 6-45, the layout of the power buses on the chip have to meet two requirements. The first requirement is the power bus width and length; the second requirement is the size of the total decoupling capacitance allocated between the power buses on the chip. Both of these consume die size. The best approach is to balance between *IR* drop and the inductive effect of the power supply pins to minimize the overall die size. As long as layout engineers can follow the power bus requirement and the decoupling capacitance requirement provided by the designers, the pump design can assume the simulations' conditions with the specified noise margins. Without this procedure,

the simulations may seem reasonable, and the actual charge pump will underperform on the silicon.

6.7 Conclusion

With the knowledge built over the previous five chapters, this chapter provided details on how to design a better charge pump. Various parameters are revisited with new perspectives, and the optimization schemes for parameters are shown in intuitive fashion. Performance improvement was discussed in detail with discussion in terms of design architecture, device physics, circuit design, physical design and layout, and system-level point of view. All the techniques introduced are practical solutions used by designers in the chip industry. This chapter should have reinforced charge pump design skills for readers, and provide the basis for more advanced discussion in later chapters.

References

1. Dickson, J.K. "On-chip high voltage generation in NMOS integrated circuits using an improved voltage multiplier technique." *IEEE Journal of Solid-State Circuits*, Vol. SC-11, pp. 374–378, June 1976
2. Pan, et al. "Four phase charge pump operable without phase overlap with improved efficiency" U.S. patent 7,030,683
3. Silva-Martinez, J. "A switched Capacitor Double Voltage Generator," *IEEE Proceedings of Mid-West Symposium. Circuits and Systems*, Vol. 1, pp. 177–180, 1994.
4. Favrat, P., et al., "High-Efficiency CMOS Voltage Doubler,"
5. San, H., et al. "Highly-Efficient Low-Voltage-Operation Charge Pump Circuits Using Bootstrapped Gate Transfer Switches." *IEEE Journal of Transactions Electrical Impedance Spectroscopy*, Vol. 120-C, October 2000
6. Wu, J.T. and L.K. Chang. "MOS Charge Pumps for Low-Voltage Operation." *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 4, April 1998.
7. Wu, Jieh-Tsorng and Chang, Kuen-Long. "Low Supply Voltage CMOS Charge Pumps." www.ics.ee.nctu.edu.tw/ ~jtwu/publications/pdf/97vlsi-cp.pdf
8. Gray, R.P., J.P. Hurst, H.S. Lewis, and G.R. Meyer, et al. *Analysis and Design of Analog Integrated Circuits*, Fourth Edition. John Wiley & Sons New York, 2001.
9. Pan, "High voltage ripple reduction." U.S. patent 6,734,718.

Chapter

7

Different Charge Pump Architectures

In previous chapters, we discussed the basics of charge pump design and the optimization of parameters to improve charge pump performance. With all the information you have been given so far, it would be rewarding to analyze some of the practical charge pumps used in the industry. By studying these designs, you can learn various practical techniques for optimizing performance and the tradeoff between different parameters. Hopefully, an understanding of these architectures will trigger new ideas and new ways of thinking.

7.1 The 2-Phase Positive Charge Pump—Revisited

The 2-phase charge pump is one of the most commonly used charge pumps in the industry. Due to its simple clocking scheme and circuit implementation, the design is less prone to failure on silicon. It has been widely adopted by designers in practice.

The best example of a 2-phase charge pump is the Dickson charge pump.[1] It is one of the fundamental architectures for explaining pump design in general. The Dickson charge pump can be divided into stages. The structure of each stage is identical. Each stage operates in a similar way. Figure 7-1 shows a schematic view of an N-stage 2-phase charge pump. In each stage, there are only two elements: the boosting capacitance (C_{boost}) and the isolation diode MOSFET transistor. Because it is designed for positive high-voltage generation, NMOS is used. The input to the first stage is directly connected to the chip power supply, V_{cc}. For consecutive stages, the input node is always connected to the output node of the preceding stages.

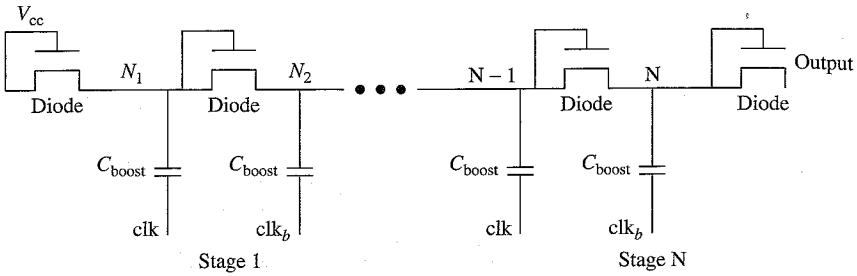


Figure 7-1 The 2-phase charge pump with N pump stages.

Figure 7-2 shows the clock phase for clk and clk_b used by the 2-phase charge pump in Figure 7-1. It is common for clk and clk_b to be non-overlapping. The high phases of clk and clk_b would never overlap each other. The non-overlapping clocking scheme claims to allow higher efficiency for charge transferring in each clock period. The boosting capacitance in the charge receiving stage needs to be coupled low first before the boosting capacitance in the charge given stage to be boosted up to transfer the charge. In reality, clk_b can simply be the inverted output of clk. Using this approach, the circuit design is simple, and there should not be any performance degradation.

Figure 7-3 shows the internal operations of nodes in the first pump stage.

The characteristics of the charge pump are represented by two parameters given in Equation 7-1. As described in earlier chapters, the limitations to the performance of the Dickson charge pump are the threshold

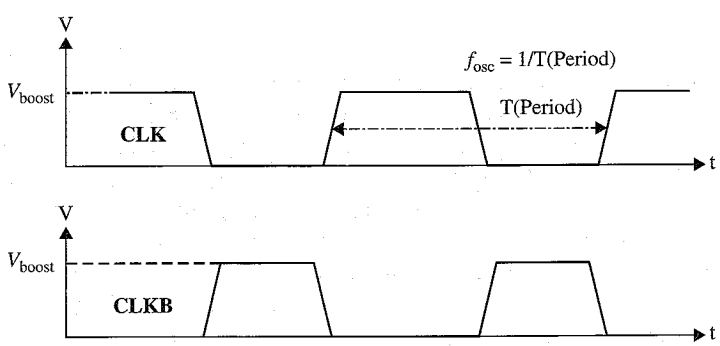


Figure 7-2 The 2-phase clocking scheme.

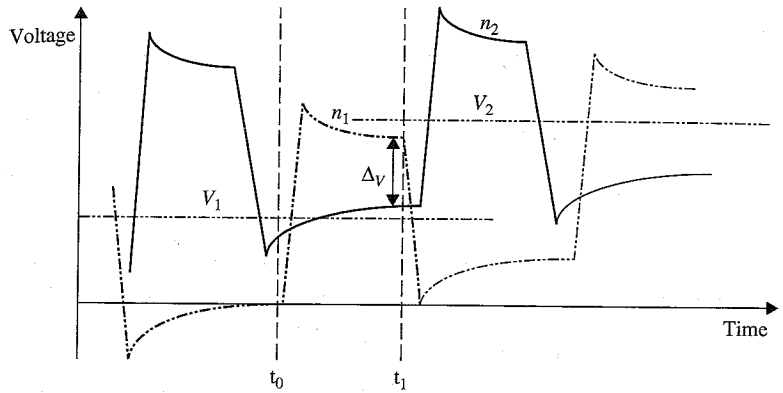


Figure 7-3 Internal waveforms of N_1 and N_2.

voltages of the diode transistor (NMOS) and the frequency limitation on the pump clock that could be applied.

$$R_s = \frac{n}{(C+C_s)f_{osc}}$$

$$V_{out} = V_{in} + n\left[\left(\frac{C}{C+C_s}\right)V_{clock} - V_{tn} - \frac{I_{out}}{(C+C_s)f_{osc}}\right] - V_{tn} \qquad (7\text{-}1)$$

The threshold of an NMOS device reduces the amount of charge that can be transferred to the next stage. The efficiency of the charge pump is immediately reduced due to the limitation. As the number of stages increases, or the voltage on the internal nodes increases, the body bias effect could further increase the absolute value of the threshold voltage, V_t. This pump architecture allows fewer and fewer charges to be transferred in the stages closer to the output. Due to the diode configuration of NMOS and V_t, the allowed pump clock frequency for operation is limited by the effective RC delay.

Due to the limitations of the 2-phase charge pump, many new designs have followed that target canceling (or reducing) the effect of V_t, reducing the effect of body bias of threshold voltages, or reducing the equivalent resistance per stage. Those techniques could greatly improve the pump's efficiency. With these improvements, either the pumps can deliver more current at the same-targeted regulation level with the same silicon area, or they can deliver the same performance with less silicon area.

7.2 The 4-Phase Positive Charge Pump

In practice, the positive 4-phase charge pump[2,3] is one of most commonly used charge pumps by designers. It is one of the best examples to represent the V_t cancellation scheme discussed earlier. The limitation of the 2-phase charge pump involves the V_t of NMOS diode connected transistors. This design targets canceling the effect of V_t to allow a full transfer of charge between stages.

Through the introduction of additional bootstrapping circuits on the gates of NMOS pass transistors, the gate voltage of NMOS pass transistors can be boosted higher than the drain voltage. Under this condition, an NMOS pass transistor is fully turned on and could allow drain and source nodes to be equalized to the same potential during the charge-transferring phase. The penalty of this design is the extra bootstrapping circuits needed per pump stage and the additional routing area for two new clocks and other wirings. The gain is the significant improvement in charge transfer efficiency over the same silicon area.

Figure 7-4 shows the configuration of a 4-phase positive N-stage charge pump. In the first stage, beside the original boosting capacitance C_1 and M_1, C_{b1} and M_2 are introduced. In the 2-phase clocking scheme, only Clk_1 and Clk_3 are used. In this scheme, two additional clocks, Clk_2 and Clk_4, with different phases are added. Figure 7-5 shows the 4-phase clocking scheme. Four different pump clocks with different clock phases are used to drive the main boosting capacitors and the bootstrapping capacitors in different stages. Clk_1 and Clk_3 are the main boosting clocks used to transfer charge and elevate the potential energy of the charge. Clk_2 and Clk_4 are the supplemental clocks that are used to facilitate the charge transfer for Clk_1 and Clk_3.

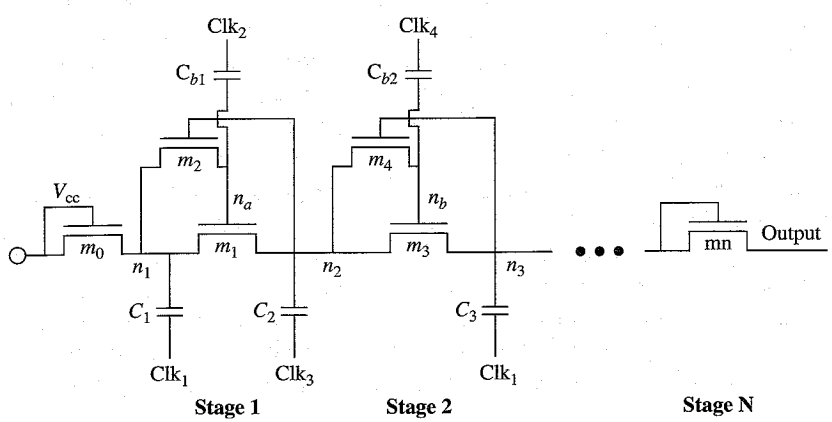


Figure 7-4 The 4-phase positive N-stage charge pump.

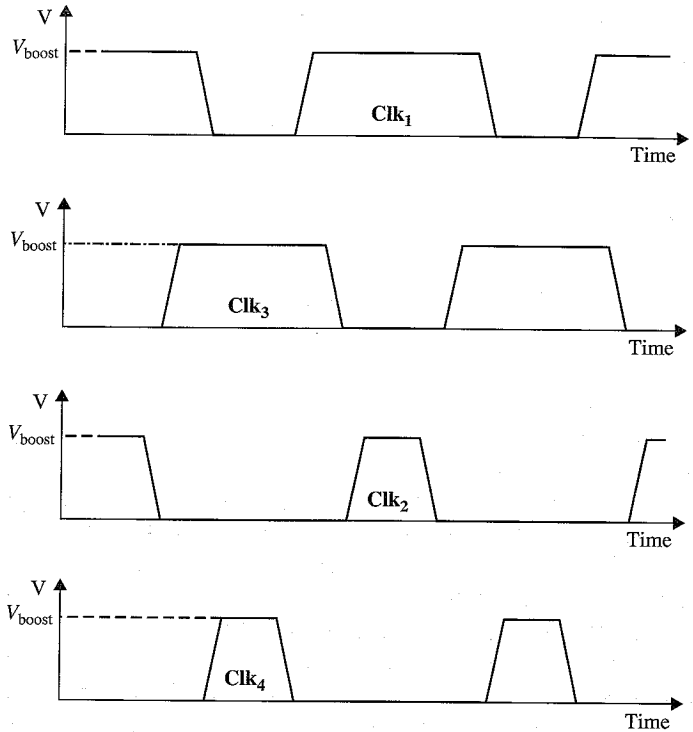


Figure 7-5 The 4-phase clocking scheme.

Because all pump stages operate in a similar fashion, we will look at only the first stage of the 4-phase charge pump to describe the basic operations of the 4-phase charge pump. The emphasis is on how charge can be fully transferred between stages via bootstrapping techniques. This also shows how the gates of NMOS pass transistors are shut off properly in this scheme to act as diodes in the off phase to prevent any potential reverse leakage.

In Figure 7-4, the first stage contains all the circuits connected in between node n_1 and n_2, as shown in Figure 7-6. M_0 is part of the initialization circuit, and C_2 is part of the second stage. They are included in Figure 7-6 for the discussion of first-stage operation only. Let us examine the devices in Figure 7-6 to understand their corresponding functions.

First, n_1 is connected to the source of NMOS M_0. M_0 acts as a clamping device or precharging device to n_1. If at any given time the potential of node n_1 falls below $V_{cc} - V_t$, M_0 would be turned on. M_0 would pull up the potential of n_1 until $V_{gs} - V_t = 0$ for transistor M_0. Due to this clamping device, the voltage at n_1 should be at least greater than or equal to $V_{cc} - V_t$ at any time.

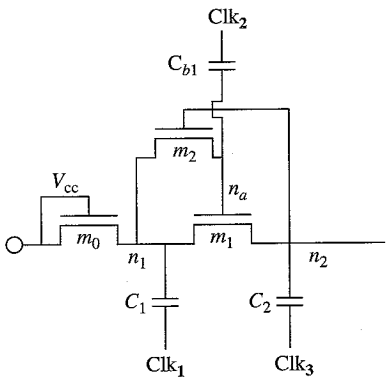


Figure 7-6 First pump stage of the 4-phase charge pump.

Second, the main boosting capacitance, C_1, is connected between Clk_1 and n_1. This device is similar in function to the one in the 2-phase charge pump. It has two purposes: The size of C_1 determines how much charge can be transferred from stage to stage per half clock cycle. Ignoring any secondary effects, the size of this capacitance would partially determine the maximum pump output power. Also, C_1 acts as an isolation device between Clk_1 and n_1. It allows the potential of the charge at n_1 to be elevated if Clk_1 goes from low to high. As the charge moves from stage to stage, the potential of the charge is moved higher and higher, which is similar as the operation of the 2-phase charge pump. If Clk_1 goes from high to low, it allows C_1 to receive the charge from the preceding stage. This is how the charge is transferred from stage to stage.

Third, NMOS transistor M_1 is connected between nodes n_1 and n_2. M_1 has a function similar to the diode-connected transistor used in the 2-phase charge pump. However, the connection for the gate of M_1 is different from the one in the 2-phase charge pump design. Node n_a, the gate of M_1, is not connected directly to the drain of M_1 as a diode. Instead, the gate is connected to n_1 through a second NMOS M_2 transistor. The purpose of this special connection will become apparent as this chapter unfolds.

Fourth, NMOS M_2 is connected between n_1 and n_a. The M_2 gate is connected to n_2, which is the output of the first stage and input to the next stage. M_2 is used to precharge n_a in one half clock cycle and then to discharge n_a in another half clock cycle.

Fifth, an additional boosting capacitance, C_{b1}, is connected to n_a. The other end of the capacitance is connected to Clk_2, which is different in phase compared with Clk_1. This capacitance allows node n_a to be boosted higher in the charge-transferring phase. This is the key to the V_t cancellation scheme.

The addition of M_2 and C_{b1} allows n_a, the gate of transferring device M_1, to be bootstrapped higher than its drain voltage during charge transfer.

Higher gate voltage could fully turn on transistor M_1. Once M_1 is fully on to conduct, the effect is the threshold voltage of M_1 being cancelled. The charge can be completely passed from n_1 to n_2. This is the key that makes this design unique from other approaches.

We will analyze step by step the internal node operations to understand the mechanism of the 4-phase charge pump. As its name states, the 4-phase charge pump uses four pump clocks with unique phases to drive capacitance. As shown in Figure 7-6, Clk_1 and Clk_3 are the main pump clocks used to drive the boosting capacitors C_1 and C_1 in two alternating stages. Figure 7-5 shows the phases of Clk_1 and Clk_3 relative to each other.

In general, Clk_1 and Clk_3 are opposite in clock phases. When one clock is being boosted up to transfer the charge to the next stage, the other stage is being coupled down to receive the charge from the preceding stage. Not only are they opposite in clock phases, Clk_1 and Clk_3 must be overlapped in high clock phases. As shown in Figure 7-5, whenever Clk_3 switches from 0 V to V_{boost}, Clk_1 would stay at V_{boost} for a short duration before switching to 0 V. Similar operations occur during the rising edge of Clk_1 and falling edge of Clk_3. This overlapping characteristic of pumping clocks is one of the most fundamental requirements of the 4-phase charge pump. Please note that Figure 7.5 exaggerates the amount of clock overlap; in general, the clock overlap should be less than ten percent of the total clock period.

The overlapping of the high phases of Clk_1 and Clk_3 has two functions: One, it allows precharging of the gate of transfer device to a certain potential before bootstrapping happens; two, it allows for the discharging of the gate of transfer devices to prevent charge leaking backward. In each pump stage, those two jobs are completed at different half cycles of pump clock. Within the same half clock cycle, those two jobs are completed at alternating stages of the pump. For example, if stage 2 is precharging to n_b, the gate of NMOS M_3, then stage 1 must be discharging n_a at the same moment. The other two clocks, Clk_2 and Clk_4, are associated with Clk_1 and Clk_3, respectively. Clk_2 is paired with Clk_1, and Clk_4 is paired with Clk_3. In Figure 7-5, Clk_2 is pulsed within the high phase of Clk_1. Clk_4 is pulsed in the high phase of Clk_3. Clk_2 should not overlap with Clk_3 in high phase, and Clk_4 should not overlap with Clk_1 in high phase either. Clk_2 is used to boost up the potential of node n_a, which is the gate of NMOS M_1. Higher potential on the gate of the transfer transistor allows partial or full V_t cancellation during charge transferring. Clk_4 serves a similar purpose for transistor M_3.

Figure 7-7 shows the waveforms of internal nodes n_1, n_a, and n_2 in the first pump stage over two completed pump clock cycles. At time t_0, Clk_3 went from 0 V to V_{boost}. Node n_2 was coupled higher by capacitor C_2. At time t_0, Clk_2 was at 0 V. Node n_a tried to equalize with node n_1

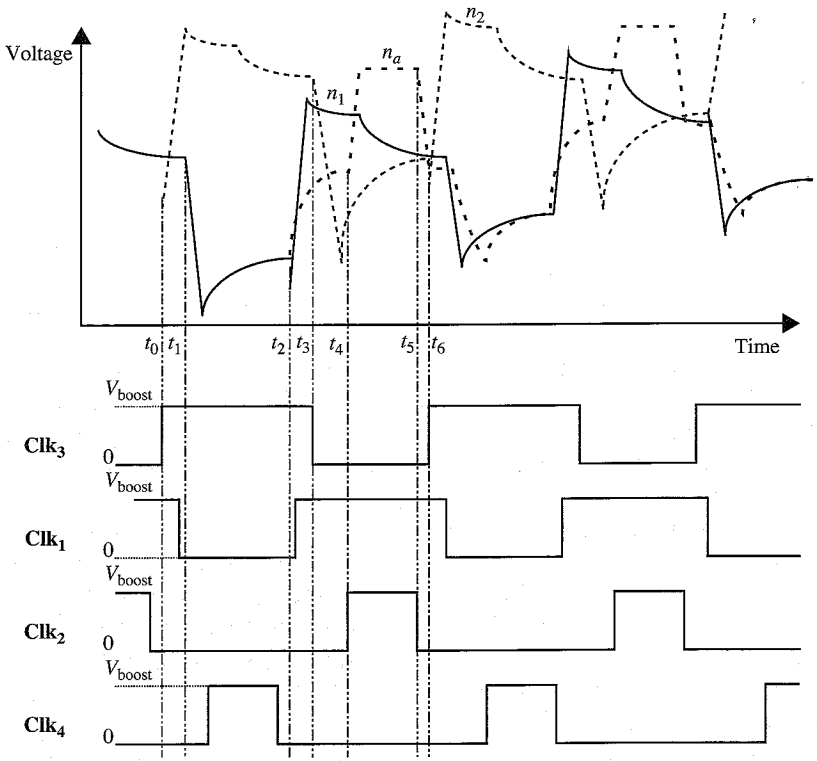


Figure 7-7 Internal node waveforms for a 4-phase charge pump's first stage.

through transistor M_2. Because both the drain and gate of M_1 were at the same potential, this formed a diode-connected device. Even though the voltage at node n_2 is being boosted higher than that of n_1, the charge could not flow backward from n_2 to n_1. This characteristic is similar to the diode-connected NMOS used in a 2-phase charge pump. At time t_1, Clk_1 went from V_{boost} to 0 V. Node n_1 was coupled down by Clk_1. The amplitude of voltage change on n_1 depends on the coupling ratio, which can be calculated using Equation 7-2. C_1 is the size of the main boosting capacitance. C_{n1} represents all the capacitance reference to ground potential on node n_1. It includes all capacitances—the source/drain-to-gate overlap capacitance, source/drain junction capacitance, and parasitic capacitance associated with any interconnection on this node.

$$\Delta V = \frac{C_1}{C_1 + C_{n1}} V_{boost} \tag{7-2}$$

At the same time, M_2 is still in conduction state. Node n_2 stayed at a higher potential than node n_1. Node n_a followed node n_1 as it was coupled lower. n_a would be at a potential very close to that of n_1.

Equation 7-3 proves that NMOS M_1 was still in the off state at time t_1. No current was able to flow backward from n_2 to n_1. This state of operation is similar to the diode in the reverse-biased state. No conduction will happen. In between time t_1 and t_2, node n_1 could not stay low after being coupled down. n_1 has to take charge from the preceding stage. In the first stage, the preceding stage is precharging transistor M_0. It acts as a diode-connected NMOS between V_{cc} and n_1. M_0 tries to pull up n_1 as long as it is conducting. $V_{cc} - V_t$ should be the upper bound that n_1 could reach during this period, as shown in Figure 7-8. For stages other than the first stage, the current stage takes charge from the preceding stage between time t_1 and t_2.

$$V_{gs} - V_t = V_{na} - V_1 - V_t = 0 - V_t = -V_t$$
$$V_{gs} - V_t < 0 \tag{7-3}$$

At time t_2, n_1 should reach approximately the level near $V_{cc} - V_t$. V_t is the threshold voltage of the NMOS transistor M_0 with back-bias applied. In between t_1 and t_2, after Clk_1 goes low, Clk_4 would be pulsed low to high while Clk_3 is still in high phase. There should be a delay between the falling edge of Clk_4 to the rising edge of Clk_1. The pulsing of Clk_1 allows the charges stored on n_1 to be transferred to n_2. At time $t2^+$, Clk_1 switches from low to high. Node n_1 would be boosted up by C_1. The change of voltage potential (ΔV) is given by Equation 7-2. At the time $t2^+$, Clk_3 is still at V_{boost}, which means n_2 will stay at a higher potential than that of n_1. NMOS M_2 will precharge node n_a, the gate of M_1. At this moment, Clk_2 is kept at a low phase. The maximum voltage (n_a) could be precharged as shown in Equation 7-4.

$$V_{precharge} = V_{n1,boosted} - Vt_{m2} \tag{7-4}$$

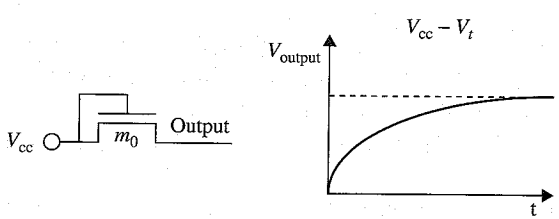


Figure 7-8 Clamping device used in first stage.

The precharging of n_a is only the first step for the V_t cancellation scheme. The overlapping of the high phase of Clk_1 and Clk_3 between t_2 and t_3 is used for the precharging operation in the first stage. At t_3, Clk_3 switched from V_{boost} back to 0 V. The voltage at n_2 was coupled lower than the voltage at n_1. The operation of NMOS M_2 changed from the saturation region into the cut-off region. The charge stored on n_a held at the previously precharged level. There is no conducting path between n_a to n_1. Because n_a was precharged to $V_1 - V_t$, NMOS M_1 is probably being weakly turned on for conduction. At this moment, some of the charge could be transferred from n_1 to n_2 through M_1.

At time t_4, Clk_2 switched from 0 V to V_{boost}. Node n_a was coupled further higher by C_{b1}. The only device that allowed node n_a to be discharged was M_2. At t_4, it was in cut-off state. The conservation of charge on node n_a should hold because the leakage current was smaller for this operation. Because $C_{b1} > C_{gate,m1} + C_{J,m2}$, node n_a should be boosted up by almost 100% of V_{boost}, which is the clock amplitude of Clk_2. In practice, the design should allow at least 80~90% coupling ratio in order to make the V_t cancellation scheme more efficient. The higher the ratio, the better the transfer efficiency. However, as C_{b1} becomes larger and larger, the gain of efficiency is not proportional.

For 3 V and 5 V designs, it is common to use chip power supply to provide power directly for the pump clock drivers. If $V_{cc} = 3.0$ V and $V_t = 1$ V for calculation purposes, M_1 should be in conduction, as shown by the derivation in Equation 7-6:

$$V_{na,boosted} = V_{precharge} + V_{boost} = 3V_{cc} - Vt_{m0} - Vt_{m2} \tag{7-5}$$

$$V_{gs}(m_1) - Vt_{m1} = V_{na,boosted} - V_1 - Vt_{m1} = V_{cc} - Vt_{m2} - Vt_{m1} = 1\text{V} > 0 \tag{7-6}$$

In between t_4 and t_5, the bootstrapping mechanism allows node n_a to reach a much higher potential, as given in Equation 7-5. Transistor M_1 was in strong conduction with this higher gate bias, as given in Equation 7-6. With these biased voltages applied, the charge can be quickly moved from n_1 to n_2 until they are of equivalent potential. As charge is being transferred, V_{ds} across M_1 is reduced over time. Eventually the condition of $V_{ds} < V_{gs} - V_t$ hits. Crossing over this boundary, M_1 switches from the saturation region into the linear region. Because the gate voltage of M_1 was at n_a, which is at a much higher voltage than the drain and source voltages of M_1, the equivalent impedance of M_1 is still much smaller than that of a pure diode-connected NMOS used in 2-phase charge pump design.

Figure 7-9 shows the equivalent resistances of transfer devices used in each pump stage for a 4-phase charge pump and a 2-phase charge pump.

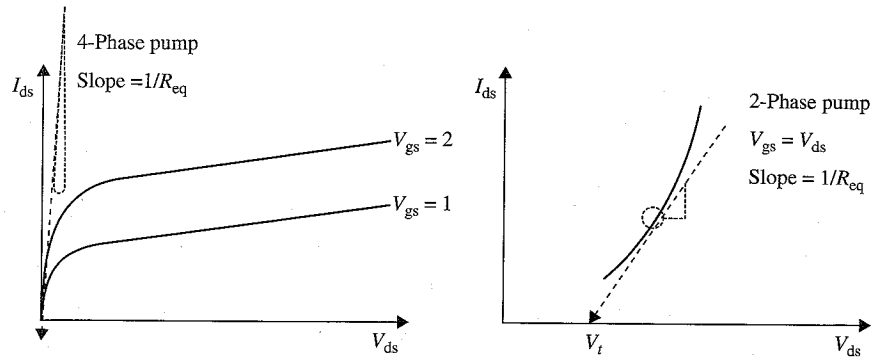


Figure 7-9 Equivalent resistance of a 4-phase charge pump and a 2-phase charge pump.

As charge is being transferred, V_{ds} is reduced. In a 2-phase charge pump, the slope of the arrow in Figure 7-9 shows the trajectory of how I_{ds} would vary with V_{ds} being decreased. In this plot $V_{ds} = V_{gs}$ under all conditions for 2-phase charge pump. It finally stops at the point where $V_{ds} = V_t$. If the plot I_{ds} versus V_{ds} is used, curves associated with different V_{gs} are plotted. As the source voltage changes, the operating points jump over different V_{gs} curves as V_{ds} drops. Connecting all the operating points together would be exactly as shown in the right portion of Figure 7-9. The stopping point of the IV curve is at $V_{ds} = V_{gs} = V_t$.

The slope of the curve at each operating point on the curve represents the inverse of 2-phase pump equivalent impendance. As for a 4-phase charge pump, the transition eventually goes all the way from the saturation region into the cut-off region. Comparing the slopes of trajectory in this region, it is obvious that with the bootstrapping gate voltage, the equivalent impedance of the 4-phase charge pump is much smaller than that of the 2-phase charge pump. Lower equivalent impedance per pump stage indirectly proves that charge can be transferred faster and more efficiently.

At t_5, in Figure 7-7, Clk_2 switches from high to low. Due to the conservation of charge at node n_a, the voltage of n_a is coupled lower near its starting level at t_4. Between t_4 and t_5, it is ideal for nodes n_1 and n_2 to be charge-shared and equalized to the same potential. This phenomenon is referred to as V_t cancellation. The direct result is better charge pump transfer efficiency per stage.

At t_6, Clk_3 switches from low to high. This operation is similar to the operation at time t_0. Between t_4 and t_5, n_2 has recovered fully by taking the charge transferred from n_1. At this moment, $t2^+$, n_2 is boosted to an even higher potential by capacitance C_2. The charge stored on n_2 is

elevated in potential and then is transferred to n_3 at a later time. Unlike 2-phase charge pump design, the charge here can be fully transferred from one stage to the other in a half clock cycle with a little penalty. It allows better charge pump efficiency to be realized. Better charge pump efficiency allows for smaller boosting capacitance and possibly fewer number of stages needed.

For 4-phase charge pump design, the equivalent resistance and the magnitude of output voltage that can be reached can be calculated by Equation 7-7. This formula is similar to the one used by the 2-phase charge pump in Equation 7-1. However, there are some differences regarding certain terms in Equation 7-7.

$$R_s = \frac{n}{(C+C_s)f_{\text{osc}}}$$
$$V_{\text{out}} = V_{\text{in}} + n\left[\left(\frac{C}{C+C_s}\right)V_{\text{clock}} - \frac{I_{\text{out}}}{(C+C_s)f_{\text{osc}}}\right] \quad (7\text{-}7)$$

First, no V_t term appears in Equation 7-7. With the bootstrapping, the V_t of the NMOS transistor is fully cancelled ideally. The direct impact is that there is no more V_t drop in charge-transferring phases. For example, if V_t of the NMOS is 1.0 V (ignoring body effect), a 5-stage charge pump could have total 5 V drop along five pump stages. A 4-phase charge pump allows the barrier of inhibiting charge transfer to be removed.

Second, the parasitic capacitance, C_s, is larger in 4-phase charge pump design than that of 2-phase charge pump design. C_s represents the total parasitic capacitance on n_1 in stage 1. In a 2-phase charge pump, C_s is composed of wire capacitance on n_1, one side junction and the gate-overlapped capacitance of transistor M_0, and one side junction capacitance and the gate capacitance of M_1. In 4-phase charge pump design, in addition to the terms given in a 2-phase charge pump, there is parasitic capacitance due to C_{b1} layout, wiring capacitance of n_a, two side junction capacitance, and the channel capacitance of M_2. This is the extra penalty 4-phase charge pump design has to pay.

Third, pump clock frequency f_{osc}, could be operated faster in 4-phase charge pump design than 2-phase pump design. The equivalent RC delay of the pump stage is much smaller for 4-phase charge pump design. This factor is one of the main reasons that limit how fast the charge pump can be operated.

Fourth, given the same clock frequency, because of less RC delay per pump stage, a 4-phase charge pump can transfer charge much more efficiently per pump clock cycle than a 2-phase charge pump.

Ideally, the effect of the gain from V_t cancellation, the gain from faster clock frequency, and the gain from better charge transferring efficiency will outweigh the penalty of extra area due to new devices and extra routings. Several design concerns need to be watched carefully for a practical 4-phase charge pump design.

The first concern involves the overlapping period of the high phases of Clk_1 and Clk_3. As discussed earlier, the overlapping clocks serve two completely different functions. Both operations happen in alternating stages at the same time or in the same stage at different clock cycles. The first purpose of overlapping the high phase of clocks allows for the precharging of the gate of the pass transistor for bootstrapping purposes. The second purpose is to properly discharge the gate of the pass transistor between stages to make it act like a diode. Without one or the other, the 4-phase charge pump would fail. The minimum amount of time needed for overlapping is based on precharging time, discharging time, the RC delay of those gates, clock skews, and the RC delay of clocks. A minimum time margin is needed to guarantee a working design on silicon in the worst process corner with the worst conditions.

The second concern involves the capacitance associated with the gates of each diode-connected device, such as M_1 and M_3. Figure 7-10 shows the devices connected to node n_a in the first pump stage. There are two transistors, M_1 and M_2, and one capacitor, C_{b1}, connected to n_a.

The key to a 4-phase charge pump is to be able to bootstrap the gate of the passing transistor high enough to fully turn on the transistor for conduction. Effectively, the threshold voltage V_t of the pass transistor is cancelled. C_{b1} is the main boosting capacitor. From its point of view, all other capacitance associated on node n_a to ground are considered loading capacitances to C_{b1}. Those capacitors include the gate capacitance of M_1, the junction capacitance of M_2, and the single side gate/drain overlap capacitance of M_2. There are some other capacitances not shown in the diagram, such as the parasitic capacitance of the wiring of the n_a node.

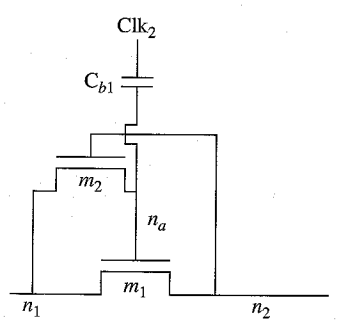


Figure 7-10 Capacitance associated with node n_a.

After the precharging of n_a, the additional boosting from Clk_2 to node n_a can be formulated using Equations 7-8 and 7-9.

$$\text{Coupling ratio} = \frac{C_{b1}}{(C_{\text{load}} + C_{b1})} \tag{7-8}$$

$$\begin{aligned} V(\text{delta}) &= V_{\text{boost}} \times \text{Coupling ratio} \\ &= V_{\text{boost}} \times \frac{C_{b1}}{(C_{\text{load}} + C_{b1})} \end{aligned} \tag{7-9}$$

The coupling ratio from Clk_2 to n_a is defined by Equation 7-8, which describes the percentage of voltage swing from Clk_2 that can be translated to change of voltage on n_a. Equation 7-9 characterizes all the loading capacitances to ground seen by C_{b1}. To achieve a good V_t cancellation effect, the boosting ratio defined by Equation 7-8 should be as large as possible to couple almost the full amplitude of Clk_2 onto n_a to cancel the V_t. In practice, this ratio should be about 85%~90%. A higher ratio ensures that up to 90%~95% of the charge can be transferred from n_1 to n_2 within the given half clock cycle.

One way to increase the coupling ratio is to make $C_{b1} > C_{\text{load}}$. A larger-sized C_{b1} translates to a larger layout area. So, how do you determine an optimum operating point with a reasonable size of capacitance while maintaining the coupling ratio in a good range? In Figure 7-11, the coupling ratio from Clk_2 to n_a is plotted based on the relationship of C_{b1} and C_{load}. C_{load} is assumed to be $1x$ of the base unit capacitance. C_{b1} varies from $1x$ to $50x$ of the base unit capacitance. The coupling ratio is plotted against C_{b1}. As shown, if C_{b1} varies from $1x$ to $5x$ of C_{load}, the coupling ratio would change from 50% to 85% in Figure 7-11.

In Figure 7-12, the change of coupling efficiency versus the size of C_{b1} is plotted. As C_{b1} increases, the rate of increment of the coupling ratio drops. The coupling ratio would gain 16% if C_{b1} changes from $1x$ to $2x$. As the size of C_{b1} increases further, the gain is reduced further. The coupling ratio only increases 3.33% when C_{b1} changes from $4x$ to $5x$. As the size of boosting capacitance continues to increase, the coupling ratio levels off in Figure 7-12. As C_{b1} changes from $9x$ to $10x$, the size of the capacitance increases by 11%, and the coupling ratio increases only 0.9%. It is obvious from Figures 7-11 and 7-12 that a coupling ratio near 85% would be an optimum choice for coupling efficiency and layout area considerations.

$$C_{\text{load}} = C_{\text{gate}}(m_1) + C_j(m_2) + C_{\text{gdov}}(m_2) + C_{\text{wire}}(n_a) \tag{7-10}$$

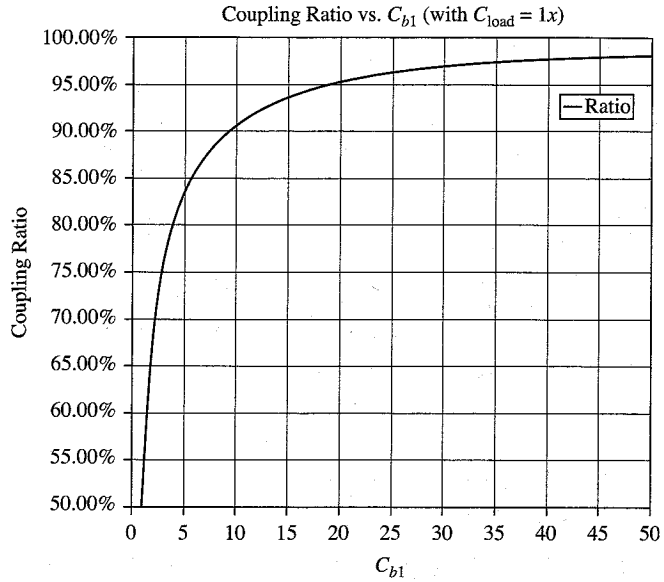


Figure 7-11 Coupling ratio vs. boosting capacitance size.

Increasing the sizing of the boosting capacitance is one way to improve coupling efficiency. Another way is to make C_{load} as small as possible. The components in C_{load} are described in Equation 7-10. Each capacitance should be individually optimized. For example, reducing the width of M_2 would reduce the overall $C_j(M_2)$ and $C_{\text{gdov}}(M_2)$. Those two components are proportional to the width of NMOS transistor M_2. A layout technique

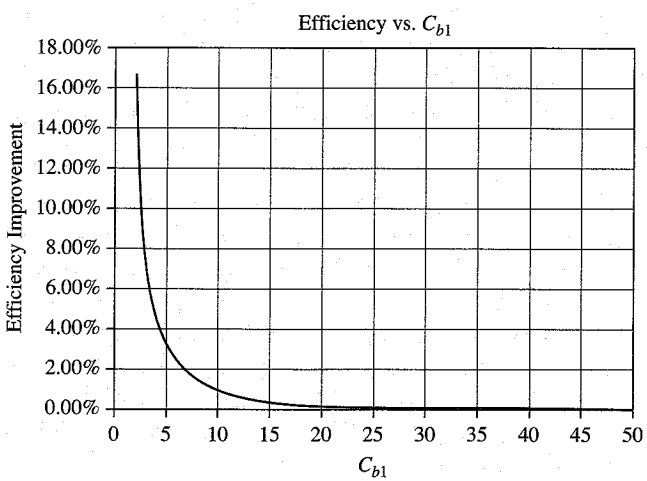


Figure 7-12 Coupling ratio efficiency change vs. boosting capacitance size.

should be used to share the drain or source to reduce the total junction area. On the one hand, the width of M_1 has to be properly sized to allow the delivery of charge within the half clock cycle; on the other hand, the gate area of M_1 cannot be too large to cause an unnecessary increase of C_{b1}. Overdesigning of M_1 would erode the die-size savings from applying the 4-phase charge pump scheme. Special layout techniques could be applied to change the wiring parasitic capacitance from loading capacitance to a good coupling capacitance in parallel with C_{b1}.

Compared with a 2-phase charge pump, a 4-phase charge pump has much better pump efficiency and better area efficiency. There is a small penalty due to the extra circuits needed per pump stage. Both design and layout need special attention in terms of the overlapping of clock signals and the parasitic load capacitance in the internal nodes.

7.3 The Modified 2-Phase Positive Charge Pump with Doubled Pump Clock Amplitude

V_t cancellation is one of the schemes that could improve the charge pump's efficiency. One of the best examples to explain this scheme is the 4-phase charge pump architecture discussed earlier. Another method to improve the charge pump's efficiency involves trying to reduce or minimize the effect of NMOS V_t on the charge transfer. Making V_t a smaller percentage of the amplitude of pump clock would serve this purpose.

$$\text{Loss} = \frac{V_t}{V_{\text{clock}}} \tag{7-11}$$

You will recall Equation 7-11 from Chapter 6 (Equation 6-12). The assumption is that boosting capacitance is significantly larger and that the coupling efficiency is 100%. Full amplitude of the pump clock can be coupled to the internal node in the pump stage to elevate and transfer the charge.

In a 2-phase charge pump, the loss of efficiency per stage due to the threshold voltage of the NMOS diode-connected device could be calculated by Equation 7-11. In a 2-phase charge pump, as charge is being transferred from stage to stage, one of the major losses of charge is due to the inability to fully transfer the charge across the diode-connected device. There is always an amount of $\Delta Q = V_t \times C_{\text{load}}$ trapped in the current stage that is not able to pass the barrier of V_t. For example, if pump clock amplitude V_{clock} is 3.0 V, the V_t of the NMOS transistor is 1.0 V, ignoring the body bias effect. The charge lost per pump stage would be 33.3% from the calculation in Equation 7-11.

In Figure 7-13, an experiment is done to illustrate how much the threshold voltage V_t per pump stage could affect the charge transfer efficiency.

Stage	Efficiency (Clock = 3v)
0	100.00%
1	67.00%
2	44.89%
3	30.08%
4	20.15%
5	13.50%
6	9.05%
7	6.06%
8	4.06%
9	2.72%

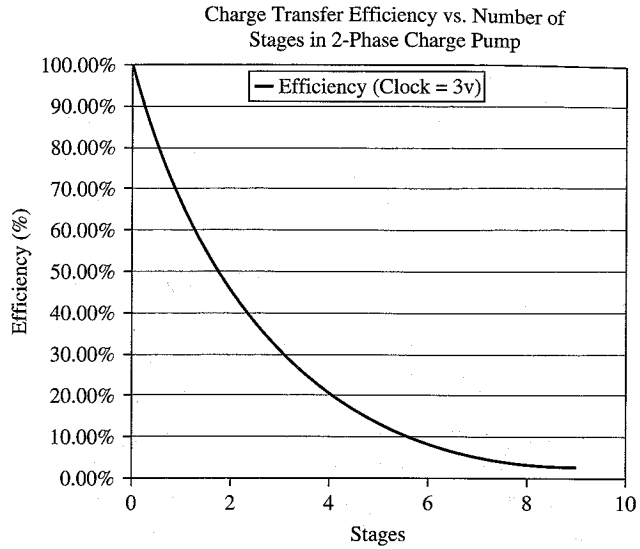


Figure 7-13 Charge transferring efficiency vs. stages.

The body bias effect is ignored for the sake of simplicity. If we assume the drop per stage is 33.3% from the previous case, at stage 0 the efficiency is 100%. As the number of pump stage increases, the charge is lost along the way. After passing stage 1, only 67% of the charge is left. After passing stage 5, only 13.5% of the original charge from stage 1 is left. After stage 9, there is only 2.72% of the original charge left.

What does this phenomenon reveal to us? If the output of stage 9 is the same as the charge pump output, and if the charge delivered to the output within the given clock cycle is able to meet the summation of all loading currents, then the input stage and successive stages have to be designed to take at least $37x$ (assuming the charge transfer efficiency is 2.72% of the first stage) the final delivered charge within the same clock cycle. The boosting capacitance has to be sized up to enable this capability, and so do the clock drivers and all supporting circuits. After internal nodes are charged up to the equilibrium levels, pump clocks still needed to couple up or couple down those capacitances. A lot of power is wasted and does not do any real useful work.

It is alarming to see the efficiency of a 2-phase charge pump decrease at such a high rate with increased pump stages and even with a fixed V_t assumption. Accounting for the back-bias effect on V_t, the efficiency curve in Figure 7-13 would move even lower than the one currently shown.

With the challenge of threshold voltage per stage, what could be the solution to improve efficiency? In Equation 7-11, a higher amplitude

Stage	Efficiency (Clock = 3v)	Efficiency (Clock = 6v)	Difference
0	100.00%	100.00%	0.00%
1	67.00%	83.33%	17.33%
2	44.89%	69.44%	25.88%
3	30.08%	57.87%	29.12%
4	20.15%	48.23%	29.25%
5	13.50%	40.19%	27.66%
6	9.05%	33.49%	25.22%
7	6.06%	27.91%	22.45%
8	4.06%	23.26%	19.66%
9	2.72%	19.38%	17.00%

Figure 7-14 Calculation of charge-transferring efficiency.

clock could reduce the charge loss during charge transfer. With this understanding, what if the pump clock amplitude is doubled? What would happen to the efficiency of the charge pump? If V_{clock} could be boosted up to 6 V, and the V_t of NMOS is still assumed to be 1 V for ease of calculation, then the charge loss per stage would be 16.7%. This approach literally cuts down the charge loss per stage by 50%.

Figure 7-14 shows the comparison of charge transfer efficiencies for clock amplitudes of 3 V and 6 V for a 10-stage charge pump design. The threshold voltage V_t of NMOS is assumed to be 1 V (ignoring the secondary effect of body bias). Figure 7-15 plots the charge transfer efficiency curves for both designs. The efficiency difference is also plotted in the same figure. Comparing the data points clearly demonstrates which design approach would achieve better charge transfer efficiency. With a 3-stage design, the efficiency of a 6 V pump is already double the performance

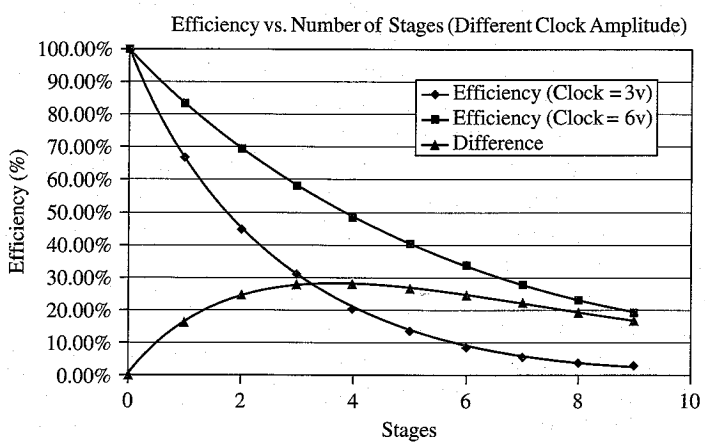


Figure 7-15 Charge-transferring efficiency vs. stages (different clock amplitude).

of a 3 V design. As the number of stages increases, the efficiency of the 6 V clock is significantly higher than that of the 3 V clock design. At stage 9, the 6 V design has a charge-transferring efficiency of 19.66%, while the 3 V design has only 2.72%. The ratio is 7*x*.

Given the example shown in Figure 7-14 and Figure 7-15, can a 2-phase charge pump be modified to double its pump clock amplitude and achieve the predicted charge transfer efficiency given in Figure 7-15? Yes, all the existing circuits in the pump stages do not need to be modified. The focus is on how to make the pump clock amplitude higher than the given system power supply. This design involves the design of a smaller scale charge pump.

Figure 7-16 is the block diagram of a generic modified 2-phase positive charge pump with doubled pump amplitude.[4,5] As shown, the diagram was composed of two parts: The first part is the conventional charge pump, located on the right side. This could be assumed to be a typical 2-phase charge pump design. The second part is the CLKGEN circuit located on the left side. It takes the clock inputs from Clk/Clk_b at V_{cc} level and then generates two $2 \times V_{cc}$ clocks: V_a/V_b. V_a and V_b are used to drive the remaining conventional charge pump stages. The challenging part of this design is how to design the clock generation circuit that can double the amplitude of the system supply. In reality, due to capacitive loading on the output and parasitic capacitance, the derived clock amplitudes are higher than system supply voltages, but they can never fully reach $2 \times V_{cc}$.

Figure 7-17 is a generic representation of a $2 \times V_{cc}$ clock generation circuit. M_0 and M_1 are NMOS transistors that are connected from source V_{cc} to V_{ab} and V_{ba}. The gates and sources are cross-coupled to each other. This configuration allows sufficient precharging on nodes V_{ab} and V_{ba} in opposite clock phases. Two boosting capacitances are driven separately by Clk and Clk_b. The boosted output V_{ab} and V_{ba} are the supplies to two inverters that are used to generate the final clocks. The final generated signals, V_a and V_b, switch between 0 V and $2 \times V_{cc}$. The amplitudes of the clocks are higher than those of the system power supplies.

There are several important design requirements. First, the wells of PMOS M_2 and M_3 should be connected to the highest potential voltages

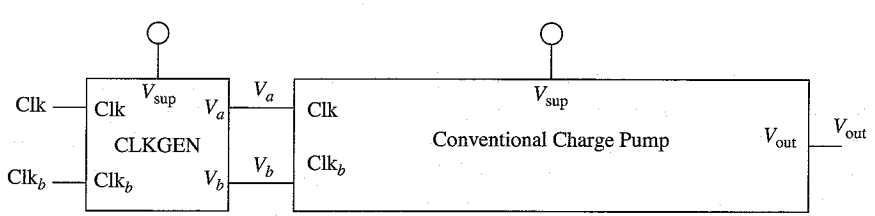


Figure 7-16 Modified 2-phase positive charge pump with doubled pump clock amplitude.

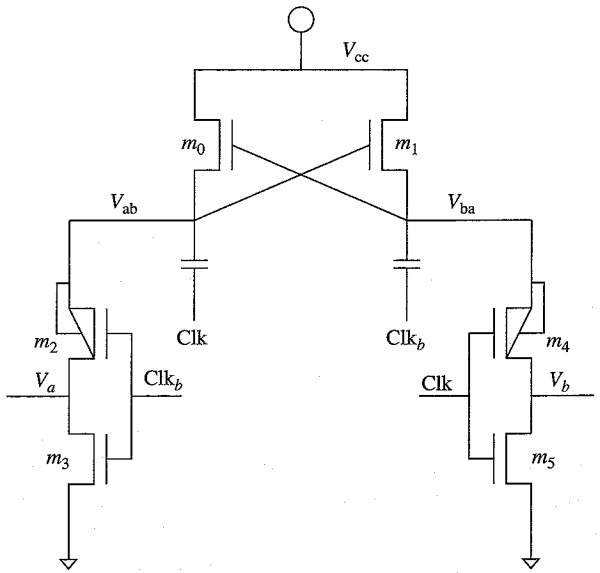


Figure 7-17 $2 \times V_{cc}$ clock generation.

seen by those PMOS transistors. They are V_{ab} or V_{ba}, depending on the phases of the clocks. Second, the punch-through effect or the breakdown characteristic of all transistors used in this design must be characterized. For example, in Figure 7-17, because the internal voltages were not at normal V_{cc} level, high-voltage design rules should be considered when working on those devices.

Another way to understand why doubling clock amplitude could significantly improve the pump efficiency is to compare the same pump performance at 5 V and at 3 V. As described in 2-phase charge operation, the V_t of a diode-connected transistor is one of the major factors reducing charge pump efficiency and limiting the clock frequency improvement. In 5 V design, the pump clock is typically powered by the chip supply V_{cc}. The efficiency loss per stage due to the threshold voltage V_t of diode-connected NMOS can be calculated using Equation 7-12.

$$\text{Efficiency Loss} = \frac{V_t}{K \times V_{cc}} \tag{7-12}$$

K is the coupling ratio based on the design and is always less than 1. For example, if V_t of NMOS is 0.8 V, the coupling ratio is 0.9, and V_{cc} is 5 V (±10%), the efficiency loss per stage is 19.7% per clock cycle per stage for the worst case, V_{cc} = 4.5 V. If the system supply drops from 5 V to 3 V, the efficiency loss per stage would be 33% for the worst case, V_{cc} = 2.7 V.

As shown, if the pump clock drivers are powered by system supply V_{cc}, while the power supply change from 5 V to 3 V represents a 40% supply voltage drop, the charge-transfer efficiency loss per stage represents an almost 67% increase based on previous assumptions. So it would be ideal to increase the amplitudes of pump clocks higher than those of system supplies to improve charge transferring efficiency. The gain of area savings due to improved charge-transferring efficiency is larger than the extra area needed to generate those higher amplitude clocks. As a consequence, the overall power consumption of the new charge pump design will be reduced due to its better circuit efficiency.

In general, for designs that use a doubling in pump clock amplitude, the conventional charge pump in Figure 7-16 could be almost any existing charge pump. The only additional change would be the extra clock-generation circuit introduced. This architecture could be easily applied to any low-voltage charge pump design required. If old-generation charge pumps are still working, by keeping the existing pump architecture and only modifying the clock-generation circuits, the design effort can be minimized. For the technology transition of chip design, such as requiring the chip power supply to be scaled from 5 V to 3 V, or from 3 V to 1.8 V or even lower voltages, the existing pump architecture used in older generation designs does not need to be modified. With the addition of a new clock generator, the existing charge pump could do the same job as before.

7.4 The Static CTS Charge Pump

Static CTS charge pump[6,7] is one type of charge pump that uses dynamic feedback to improve the charge-transfer efficiency. This architecture follows the design approach of the V_t cancellation scheme.

Figure 7-18 is a schematic view of a static CTS charge pump. Looking at the circuits below the divider line, the bottom portion of the static

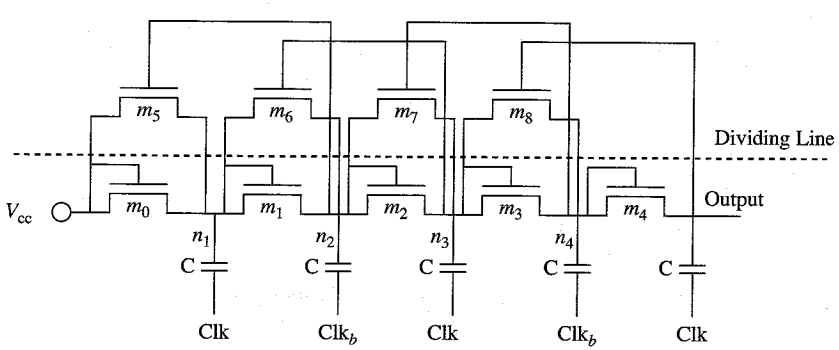


Figure 7-18 Static CTS charge pump.

CTS pump is identical to a generic 2-phase charge pump. The top portion is the new circuit being added to cancel the V_t of NMOS dynamically. Let us use the first pump stage as an example. Assuming $C > C_s$, to cancel the threshold of M_5 at the high phase of Clk_b, Equation 7-13 needs to be satisfied.

$$\begin{aligned} V_{n1} &= V_{cc} - V_t \\ V_{n2} &= V_{n1} + V_{cc} - V_t + V_{cc} = 3V_{cc} - 2V_t \\ V_{n2} - V_t &>= 0 => 3V_{cc} - 3V_t >= 0 \end{aligned} \tag{7-13}$$

At the low phase of Clk_b, M_5 needs to be in cut-off state after Clk_b goes low. Now Equation 7-14 needs to be satisfied.

$$\begin{aligned} V_{n2} - V_t &\le 0 \\ V_{cc} - V_t + V_{cc} - V_t &\le 0 \\ 2V_{cc} - 2V_t &\le 0 \end{aligned} \tag{7-14}$$

Equation 7-13 can be satisfied as long as the chip supply is higher than the V_t of NMOS. If Equation 7-13 is satisfied, Equation 7-14 cannot be satisfied at the same time. Because Equation 7-14 cannot be satisfied, reverse leakage current occurs while the node is pumping.

In order to fix the issue shown in Equation 7-14, a modified CTS charge pump stage[7,8] is presented in Figure 7-19. An NMOS M_1 transistor and a PMOS M_2 transistor were added to control the gate voltage of M_5. PMOS M_2 still allows the higher voltage potential from the next

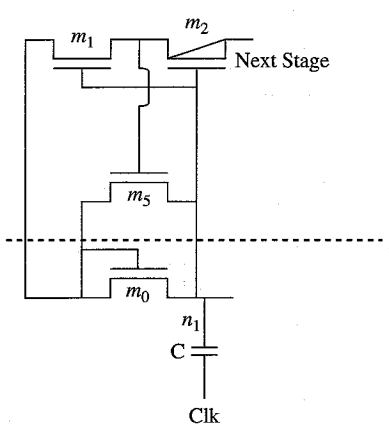


Figure 7-19 Segment of a modified static CTS charge pump.

stage to be applied to the gate of M_5 during the charge-transferring phase. NMOS M_1 in this case would shut down gate M_5. This is to prevent Equation 7-14 from happening when Clk_b goes low.

Although the simulation shows improved performance over the original CTS charge pump, there are other limitations to this design. First, a total of three extra transistors need to be added, compared with the 2-phase charge pump. One of them is PMOS with NWELL. The 4-phase charge pump only adds one extra transistor and one capacitance. Considering the routing area and extra spacing needed between the devices and NWELL, the modified static CTS charge pump should occupy a larger die size. Second, a PMOS device is added to the high-voltage path. The technology should be available to process high-voltage PMOS. Third, considering the performance of the modified static CTS charge pump, because of the additional three devices, the extra parasitic capacitance introduced is quite significant. Therefore, this design may not have an advantage over the 4-phase charge pump design.

It should be noted that using dynamic feedback to improve the charge pump efficiency is one common technique. It is also known that the latter stage has higher potential than the current stage. The latter stage voltage can always be utilized to improve the charge transfer of the early stage somehow. The modified CTS charge pump is a good example of this technique.

7.5 The Positive Charge Pump with Very High Amplitude Pump Clocks

There is another type of charge pump that is based on the concept of very high clock amplitude design. As discussed earlier, by just doubling the clock amplitude for a 2-phase charge pump design, the efficiency can be significantly improved. Because higher amplitude clocks can reduce the inhibiting effect of V_t on charge transfer (refer to Equation 7-11), the clock amplitude can be further doubled using clock adders or doublers, in theory to improve the charge pump efficiency.

Figure 7-20 is a generic representation of how to generate very high amplitude pump clocks. The $2 \times V_{cc}$ CLKGEN was shown in Figure 7-17.

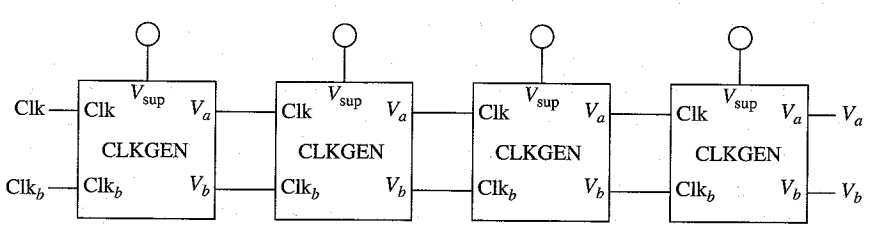


Figure 7-20 Generation of very high amplitude pump clocks.

With four stages CLKGEN cascaded in serial, V_a and V_b should get closer to $5 \times V_{cc}$ level in clock amplitude. As the clock amplitude increases, the number of stages can be reduced proportionally. The impedance of the charge pump is proportional to the number of pump stages in serial. The greater the number of stages in serial, the higher the pump impedance will be. With higher impedance, to deliver the same amount of power on output, the area of the charge pump needs to be larger. Equation 7-12 could clearly support this design approach to get better efficiency.

$$R_s = \frac{n}{(C + C_s')f_{osc}}$$
$$V_{out} = V_{in} + n\left[\left(\frac{C}{C + C_s}\right)V_{clock} - \frac{I_{out}}{(C + C_s)f_{osc}}\right] \quad (7\text{-}15)$$

Very high amplitude clock design also has its own drawbacks. Because the output is a clocking signal, it requires the charge to be elevated to generate this clock. Theoretically, capacitive coupling is needed. To generate a $2 \times V_{cc}$ clock, one stage of capacitance is needed. A $3 \times V_{cc}$ clock requires two stages of capacitance in serial. As more capacitance is stacked in serial, the equivalent capacitance is reduced. To compensate for this reduction, the capacitance at each stage needs to be increased up proportionally. A practical design needs to balance the size, power consumption, and efficiency of the pump as a whole system. Another issue with very high amplitude clock generation is the devices issues. Since the clock amplitude is switching from 0 V to very high voltage potential, the design is prone to snap-back, punch-through, etc. Extra attention is needed when using this pump architecture in practice.

7.6 The 2-Phase Negative Charge Pump

The charge pumps discussed so far are all positive charge pumps. The source of the charge is from the system power supply, V_{cc}. Through every stage the potential energy of the charge is elevated by the pump clock. On the output, a relatively high potential exceed system supplies is achieved.

If the potential of the charge can be raised, then the potential of the charge can be lowered too. In modern chip design there are occasions when a negative voltage lower than the ground supply is needed. For example, on a DRAM chip, in order to reduce the background leakage current, a negative voltage is generated to bias the P substrate instead using the ground supply V_{ss}.

Figure 7-21 shows a generic representation of a K stage 2-phase negative charge pump.[9–11] Compared with its counterpart, the 2-phase

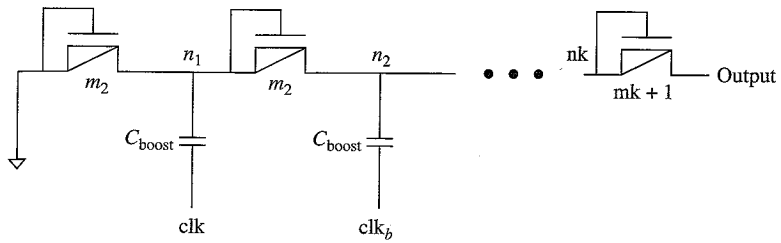


Figure 7-21 Generic K stage 2-phase negative charge pump.

positive charge pump, there are two major changes: First, all NMOS transistors are replaced with PMOS transistors, with the NWELL connecting to the most positive voltage per stage. Second, the source for the first stage is not V_{cc}, but V_{ss}.

Figure 7-22 shows the 2-phase clocking scheme used by a 2-phase negative charge pump. This diagram looks similar to its positive counterpart. In a positive charge pump, the high phase of the clock is used to transfer the charge. For the negative charge pump, the low phase of the clock is designed to transfer the charge.

In Figure 7-21, the PMOS transistor is connected in diode fashion. The negative charge pump removes the positive charge from the pump output node and transfers it to ground node, at the beginning of the pump. As charge is carried away from the pump output, the voltage reduces from the initial level (say positive V_{DD} level) to a negative voltage level. The direction of charge movement is exactly the opposite compared with that of the positive charge pump.

Let us examine the basic operations of the first stage of the negative pump. As shown in Figure 7-23, when n_1 is being boosted up by Clk, its final level cannot exceed $|V_{tp}|$ of PMOS M_1. M_1 clamps the positive voltage on n_1. The positive charge cannot propagate to n_2 through M_2.

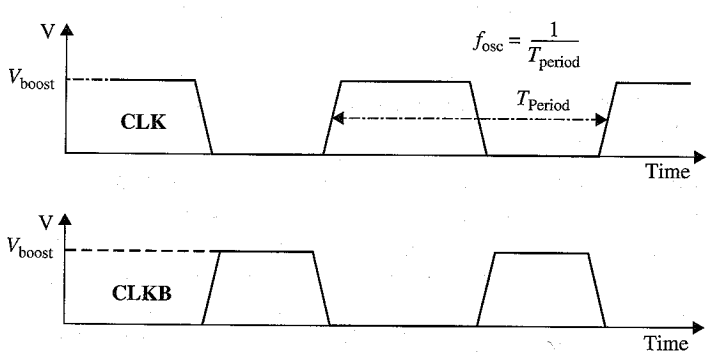


Figure 7-22 The 2-phase clocking scheme for a negative charge pump.

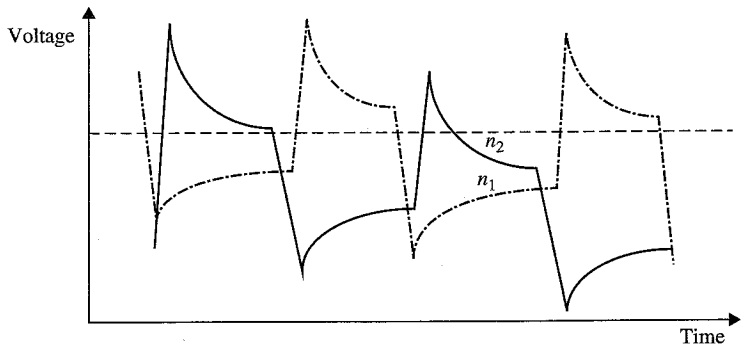


Figure 7-23 Internal waveforms of n_1 and n_2 in a negative charge pump.

When n_1 is being boosted down by Clk, the node is coupled into the negative range. M_1 is in the cut-off region because of the diode connection of PMOS. At the same time, node n_2 is being boosted higher by Clk_b. Due to the diode-connected device M_2, the positive charge on n_2 is removed and transferred to n_1. In the next clock phase, n_1 will be boosted higher to dump a positive charge to V_{ss}. Node n_2 is boosted even lower to transfer the negative charge down to n_3.

The characteristics of a 2-phase negative charge pump can be modified from Equation 7-1 to give the formulas in Equation 7-16. All the design constraints of the 2-phase clocking scheme and the design concerns of the 2-phase positive charge can be applied to the 2-phase negative charge pump design.

$$R_s = \frac{n}{(C+C_s)f_{osc}}$$

$$V_{out} = -n\left[\left(\frac{C}{C+C_s}\right)V_{clock} - |V_{tp}| - \frac{I_{out}}{(C+C_s)f_{osc}}\right] + V_{tp} \qquad (7\text{-}16)$$

7.7 The 2-Phase Negative Charge Pump with Triple Well Technology

Because the body bias effect of the PMOS transistor is usually larger than that of the NMOS transistor, the 2-phase negative charge[9–12] using the preceding implementation is only suitable for low-amplitude negative charge pump design. To improve the negative charge pump's efficiency (or for higher amplitude negative charge pump design), a certain technique can be applied. The diode transfer PMOS is in NWELL. NWELL is typically biased to V_{cc}, the highest potential on chip.

If the NWELL of PMOS is biased to V_{cc}, then V_{sb} would increase with the increasing number of stages due difference between the negative potential on the internal nodes and NWELL bias V_{cc}. As shown in Equation 7-17, the absolute value of V_t will go higher and higher.

$$\Delta V_T = \frac{\sqrt{2\varepsilon_s q N_a}}{C_{ox}}\left(\sqrt{(2\phi_F + V_{SB})} - \sqrt{2\phi_F}\right) \qquad (7\text{-}17)$$

One approach to slow down the trend of V_{tp} increment with V_{sb} is to switch NWELL potential from V_{cc} down to 0 V when the output hits some predetermined negative voltage level. For example, –3 V may be a good candidate. This approach will temporarily reduce the absolute value of V_{sb} by V_{cc}, which in turn reduces the absolute value of V_{tp} for PMOS. With –3 V on the output, the voltage on the internal stage is lower than 0 V. Switching the potential of NWELL from V_{cc} to 0 V will not cause the forward biasing on the source/drain junctions in internal nodes to occur.

This technique only temporarily reduces the absolute value of V_{tp}. As stages increase, or as the voltage goes more negative, Equation 7-14 will still hold. With the help of process technology, a new type of transistor was created. It allows the negative charge pump design not to suffer from the body bias of diode-connected PMOS anymore.

In a normal CMOS process, the substrate of silicon is p-type. NMOS devices can be processed directly on top of p-substrate. To create PMOS devices, NWELL is created in p-substrate to allow the definition of PMOS devices.

In Figure 7-24, with the addition of an extra PWELL created in the NWELL, a new type of N-channel MOSFET transistor is created. If the PWELL and p-substrate are both biased to V_{ss}, and the NWELL is biased to V_{cc}, this device is not much different from a normal NMOS device defined on top of p-substrate. However, if the PWELL is tied to

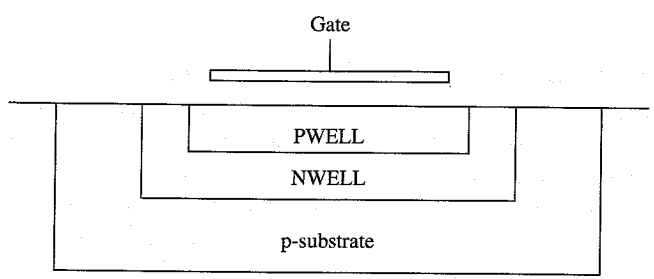


Figure 7-24 Deep NWELL n-type transistor.

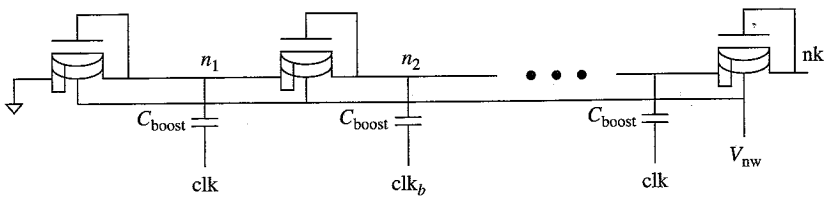


Figure 7-25 A 2-phase negative charge pump with a deep NWELL transistor.

the source of the transistor and the NWELL is biased to V_{cc}, the source and the local PWELL can go down to negative potential. If a normal NMOS source is going down to negative voltage level, the junction will become forward biased. Therefore, it is possible that latch-up issues could arise.

With the aid of a deep NWELL NMOS device, the negative charge pump can be configured like the one in Figure 7-25. This negative charge pump with triple well technology is very similar to the 2-phase positive charge pump. The exception is that the first stage is connected to V_{ss} rather than to V_{cc}. The gates of all NMOS diodes are switched to the other side.

All diode-connected deep NWELL NMOS devices have their local PWELL connected to the source. This connection allows $V_{ss} = 0$, and there is no body bias effect on diode-connected PMOS. Deep NWELL can be tied to V_{cc}, or it can switch like it does in the previous design to reduce well-to-well potential difference. The operation of this negative charge pump is simple.

In Figure 7-25, at each pump stage, if the node is being coupled higher, it will dump the positive charge toward the left side to the ground; if the node is being coupled lower, it transfers this negative potential toward the output. The detailed derivation of the internal node operations and corresponding waveforms are not covered here. It would be a good exercise for the reader to analyze this pump architecture using the techniques covered in this chapter.

7.8 Conclusion

After introducing the fundamentals of charge pump designs and optimization process to achieve better charge pump efficiencies, this chapter has sought to reinforce the reader's understandings with real life charge pump implementations. All the architectures presented in this chapter are broadly used by many chips in different products. The continuous life spans of the architectures in real products are the best supports to prove their merits. Each architecture was analyzed thoroughly to highlight the unique approach behind it. These architectures are refined

by generation after generation of product refinements. One of the goals of this chapter was to help provide the reader with a thorough understanding of different charge pump designs. Another goal was to spark new ideas and new implementations for charge pumps.

References

1. Dickson, J.K. "On-chip high voltage generation in NMOS integrated circuits using an improved voltage multiplier technique," *IEEE Journal of Solid-State Circuits*, Vol. SC-11, pp. 374–378, June 1976.
2. Lin, H., N. Chen, and J. Lu. "Design of Modified Four-Phase CMOS Charge Pumps for Low-voltage Flash Memories." *Journal of Circuits, Systems, and Computers*, Vol. 11, No. 4, pp. 393–403, 2002.
3. Pan, et al. "Four phase charge pump operable without phase overlap with improved efficiency," U.S. patent 7,030,683.
4. Di Cataldo, G and G. Palumbo, "Double and triple charge pump for power IC: Dynamic models which take parasitic effects into account." *IEEE Transactions on Circuits and Systems*. I, Vol. 40, pp. 92–101, February. 1993.
5. Starzyk et al. "A DC–DC Charge Pump Design Based on Voltage Doublers." *IEEE Transactions On Circuits And Systems—I: Fundamental Theory and Applications*, Vol. 48, No. 3, March 2001.
6. Wu, J.T. and L.K. Chang. "MOS Charge Pumps for Low-Voltage Operation." *IEEE Journal of Solid-State Circuits*, Vol. 33, No. 4, April 1998.
7. Tsang, B. and E. Ng. "Switched Capacitor DC–DC Converters: Topologies and Applications," www.ocf.berkeley.edu/~eng/classes/EE290cPresentation.ppt.
8. Wu, J.-T. and Chang, K.-L. "Low Supply Voltage CMOS Charge Pumps," www.ics.ee.nctu.edu.tw/ ~jtwu/publications/pdf/97vlsi-cp.pdf.
9. Lin, H., H.K. Chang, and C.S. Wong. "Novel High Positive and Negative Pumping Circuits for Low Supply Voltage." *IEEE International Symposium on Circuits and Systems*, Vol. 1, pp. 238–241, May 30–June 2, 1999.
10. Naso, et al. "Negative-voltage charge pump with feedback control." U.S. Patent 5,168,174.
11. Jinbo, T., et al. "A 5-V-only 16-Mb flash memory with sector erase mode." *IEEE Journal of Solid-State Circuits*, Vol. 27, pp.1547–1553, 1992.
12. Atsumi, S., et al. "A 16-Mb Flash EEPROM with a New Self-Data-Refresh Scheme for a Sector Erase Operation." *IEEE Jounal of Solid-State Circuits*, Vol. 29, pp. 461–469, 1994.

Chapter

8

Future Design References

After discussing the basic operations of charge pumps and how to optimize various parameters to achieve the design targets, we reach a point where we must think about the trend of future pump designs. The goals of a better charge pump design include targeting a smaller die size, higher power efficiency, and less output noise at regulation level. The trend of charge pump design is still to try to meet all these requirements.

8.1 Area Versus Performance

In general, the area of a charge pump is inversely proportional to the pump clock frequency, and the area of a charge pump is inversely proportional to the threshold of the diode-connected transistors. To reduce overall charge pump area, the performance and the efficiency of the charge pump have to be increased compared with old generations. If the overall optimizations of architectures, device characteristics, design parameters, and circuit techniques allow the final design to have better output performance than previous designs, the area of the charge pump would be minimized by default.

8.1.1 Output performance versus pump clock frequency

First let us plot I_{output} versus pump clock frequency for different charge pump designs. A design that can achieve the highest possible frequency has the capability of reducing pump boosting capacitor size while still meeting the same performance target. Figure 8-1 is the plot of pump output current versus clock frequency for a 2-phase charge pump design such as

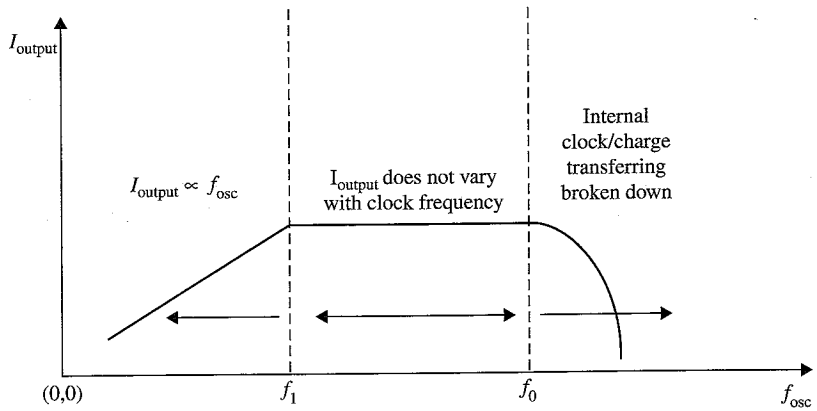


Figure 8-1 I_{output} vs. pump clock frequency for a 2-phase charge pump.

the Dickson Charge pump.[1] The curve can be divided into three regions according to the characteristics described in Equations 8-1 to 8-3.

$$f_{osc} < f_1 \qquad I_{output} \propto f_{osc} \tag{8-1}$$

$$f_1 < f_{osc} < f_0 \qquad I_{output} \text{ is approximately constant.} \tag{8-2}$$

$$f_{osc} > f_0 \qquad I_{output} \text{ decreases as the pump is near breakdown.} \tag{8-3}$$

In Equation 8-1, for a clock frequency that is lower than f_1, the charge in the pump could be fully transferred between successive stages within each clock cycle. The amount of time it takes to transfer the charge is less than the clock period. As f_{osc} increases, more charge is dumped to the output within the same given period. This is why the output current increases proportionally as the pump frequency increases in Figure 8-1.

In Equation 8-2, if the pump clock frequency operates between f_1 and f_0, the charge cannot be fully transferred between stages within the given clock period. As frequency increases, less charge would be transferred in a given clock period. The decrement in the total amount of charge being transferred per cycle is compensated for by the increment of the pump clock frequency. As shown in Equation 8-4, I_{output} is almost unchanged as the frequency varies. In Figure 8-1, a flat curve is shown for I_{output} in this region.

$$\begin{aligned} Q_2 &= Q_1 - \Delta Q \\ f_2 &= f_1 + \Delta f \\ I_2 &= \frac{Q_2}{t_{cycle}} = Q_2 \times f_2 = (Q_1 - \Delta Q) \times (f_1 + \Delta f) \end{aligned} \tag{8-4}$$

In Equation 8-3, as the clock frequency eventually goes beyond the limit (f_0), either the pump clock generation fails, or there is not enough time for the charge to be transferred between stages. As a consequence, I_{output} will fall quickly at an exponential rate. The charge pump efficiency gets worse with increasing clock frequency. Eventually, the pump will not function anymore. As shown in Figure 8-1, the curve exponentially decreases for a clock frequency that is higher than f_0.

Figure 8-2 is the plot of the pump output current versus pump clock frequency for a V_t cancellation charge pump. The critical frequency f_0 divides the operating range into two regions as shown in Equation 8-5 and 8-6.

$$f_{osc} < f_0 \qquad f_{output} \propto f_{osc} \tag{8-5}$$

$$f_{osc} > f_0 \qquad f_{output} \text{ decreases as the pump is near breakdown.} \tag{8-6}$$

For Equation 8-5, if the pump operates at a frequency lower than f_0, the charge can be fully transferred between successive stages in each clock cycle. As f_{osc} increases, the output current I_{output} will increase proportionally with the increased pump clock frequency f_{osc}. For designs using very high amplitude pump clocks, the lost charge due to V_t of diode-connected transistors is either very small or is negligible, and Equation 8-5 can still be applied. For Equation 8-6, as the clock frequency increases further over the limit (f_0), either the clock generation circuits start to fail or the charge has too little time to be transferred. As a consequence, I_{output} will fall almost at an exponential rate until the pump fails completely.

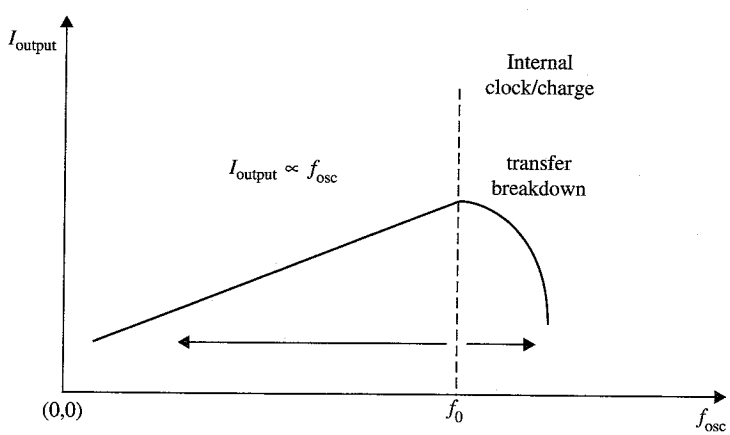


Figure 8-2 I_{output} vs. pump clock frequency for V_t cancellation charge pump.

8.1.2 Output performance versus pump clock amplitude

Figure 8-3 is the plot of pump output current versus pump clock amplitude. Equation 8-7 is for the pump clock amplitude lower than V_0. Due to the threshold of MOSFET transistors, if V_{clock} is below the threshold voltage, the pump cannot transfer charge at all. Only if V_{clock} crosses over this threshold voltage will the performance of the charge pump improve. For Equation 8-8, if the clock amplitude is larger than V_0, I_{output} should increase linearly with the increased pump clock frequency.

$$V_{clock} < V_0 \quad \text{Internal stage is inefficient.} \tag{8-7}$$

$$V_{clock} > V_0 \quad I_{output} \propto f_{osc} \tag{8-8}$$

8.2.3 Output performance versus number of pump stages

Pump performance has a strong relationship with the number of pump stages in series.[3] Figure 8-4 is a plot of pump efficiency versus regulated pump output voltage. The efficiencies for two similar designs with different numbers of pump stages were plotted. The efficiency can be defined in two ways. The first way is to use the ratio I_{output}/I_{cc} to judge the efficiency of circuit. The design that consumes the lowest system current I_{cc} while delivering the required output power has better circuit efficiency.

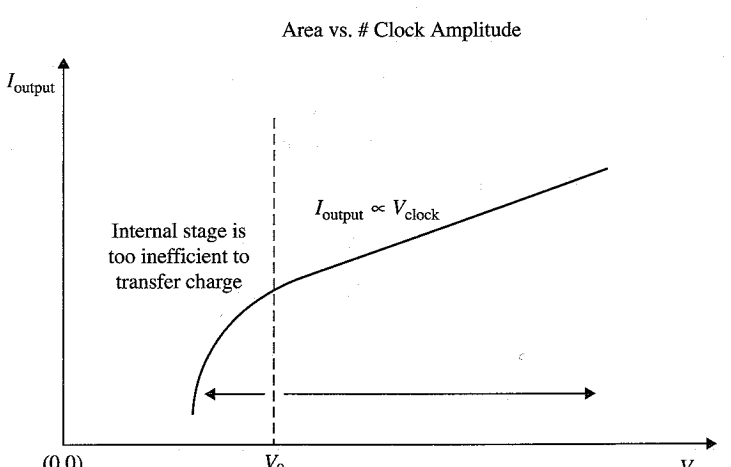


Figure 8-3 I_{output} vs. pump clock amplitude.

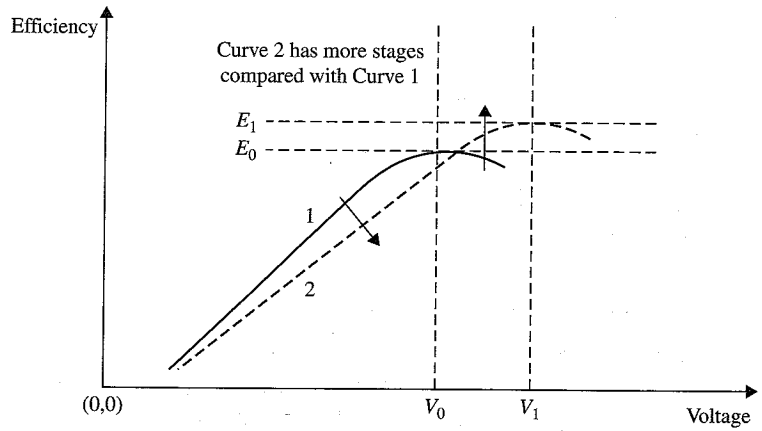


Figure 8-4 Pump efficiency vs. number of pump stages.

The second way to define the efficiency of charge pump is based on the following ratio: I_{output} / *Area*. Here, *Area* is the total layout area of the charge pump and its supporting circuit. If the charge pump is efficient, less area is occupied to generate the required output power. The higher the ratio, the more efficient the design has to be.

In Figure 8-4, the performances of two pump designs with different numbers of pump stages are compared. The individual pump stages between two designs are identical for comparison purpose. Curve 1 has a fewer number of stages compared with Curve 2. Pump efficiencies are plotted against the output regulation voltages for both designs.

In Equation 8-9, as long as the regulated output voltage is lower than V_0, Curve 1 has better efficiency than Curve 2, as shown in Figure 8-4. In this operating range, the serial impedance of the pump with fewer of pump stages is low compared with the other one. In Equation 8-10, for the regulation level that is in between V_0 and V_1, the efficiency of Curve 1 drops off from its peak. Curve 2 overtakes Curve 1 in terms of output efficiency. In Equation 8-11, for a regulation level that is higher than V_1, the efficiencies for both Curve 2 and Curve 1 fall. Curve 2 still has better efficiency than Curve 1. If output is designed to operate in this range, more pump stages should be added in serial to allow better pump efficiency and maintain the slope of the curve.

$$V < V_0 \tag{8-9}$$

$$V_0 < V < V_1 \tag{8-10}$$

$$V > V_1 \tag{8-11}$$

8.2 Power Consumption

Power consumption of the charge pump is one of critical design specifications. With the scaling of technology and power supply voltages, the demand for charge pumps with low power consumption is strong. The power consumption of the charge pump is affected by many charge pump design parameters. The next few subtopics will address some of the major parameters.

8.2.1 Power consumption versus pump clock frequency

Power consumption for charge can be divided into two components: the first component is the charge transferred from stages to stages in serial, and eventually to the output of the charge pump; the second component is the power consumed by peripheral supporting circuits.

$$f < f_0 \quad \text{Power} \propto f_{osc} \tag{8-12}$$

$$f > f_0 \quad \text{Pump malfunction} \tag{8-13}$$

For pump clock frequency that is less than the critical frequency, f_0, as shown in Equation 8-12, power consumption of the pump is proportional to f_{osc}. If the clock period is relatively long, the charge transferred between stages could be a near constant term. As frequency increases, more charge can be transferred in a fixed period, so the associated power consumption will increase proportionally. If the clock period is too short, the amount of charge being transferred is cut short. However, within the given period of time, there are more clock pulses. The total amount of charge being transferred is relatively near a constant value. The power does not change with clock frequency. For supporting circuits, such as clock generators and clock buffers, the total power consumed is linearly proportional to clock frequency, f_{osc}. For frequency that is lower than critical frequency, f_0, the combined power from those two terms results in the total power consumption being a linear function of the pump clock frequency, f_{osc}, as shown in Figure 8-5.

For frequency that is greater than the critical frequency, f_0, as shown in Equation 8-13, the internal circuits of the pump start to fail and eventually cause the pump to malfunction. Although the current associated with supporting logic increases, the charge cannot be transferred between stages. The second term is a larger component in terms of power. Overall the power consumption should decrease.

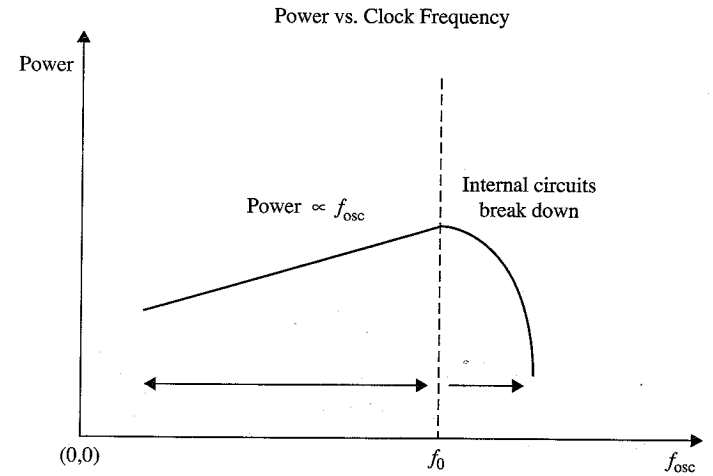


Figure 8-5 Power consumption vs. pump clock frequency.

8.2.2 Power consumption versus number of pump stages

The pump design itself is flexible regarding the number of pump stages to be used in implemented. Although the specified design targets can be achieved by both designs with different number of pump stages, the power consumption in two designs could be different.

Figure 8-6 is the plot of power consumption versus output voltage. There are two designs shown in the plot. Design 2 has more number

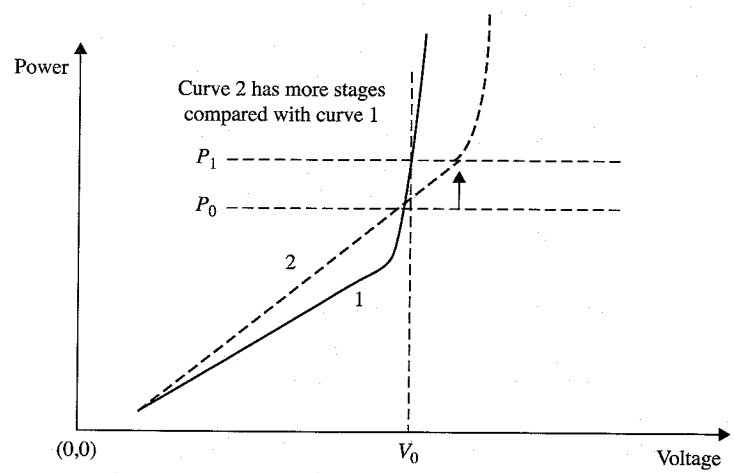


Figure 8-6 Power consumption vs. number of pump stages.

of pump stages than Design 1. The pumping capacitance at each stage would vary to meet the output power requirements. For this discussion, it is assumed that both pump stages are identical.

$V < V_0$ Design 2 consumes more power. (8-14)

$V > V_0$ Design 1 consumes more and fails first. (8-15)

The critical output regulation level V_0 is associated with the design. Charge pumps perform differently across this boundary.

For an output voltage level less than V_0, as shown in Equation 8-14, to reach the same pump output regulation level, Design 2 has to consume more power than Design 1. The operation is shown in Figure 8-6. The charge lost due to the internal parasitic capacitance and V_t drop (if any) is more for Design 2 because it has more pump stages in serial. For the supporting circuit to the charge pump, more power has to be consumed to charge or discharge larger capacitance and extra stages in Design 2. Overall, Design 2 should consume more power than Design 1 in order to meet the same output regulation level.

For an output voltage level greater than V_0, as shown in Equation 8-15, Design 1 overtakes Design 2 in terms of power consumption in Figure 8-6 because the charge pump output voltage level is proportional to the number of stages in serial. For an output voltage that exceeds V_0, Design 1 becomes less efficient for delivering the charge at a high potential voltage. The number of stages becomes a limiting factor for Design 1. As a consequence, more boosting capacitance needs to be added in each stage to compensate for the reduction of circuit efficiency due to the number of pump stages. If the regulation level is further raised, Design 1 eventually fails to operate because it just cannot reach the regulation level with its existing number of pump stages.

It is important to study the design specification and build extra margin in the design. Variation of supply voltage, parasitic RC, threshold voltage of transistors, and so on, on real silicon could all cause the pump to fail if not enough design margin is built in at the beginning.

8.2.3 Power consumption versus pump clock amplitude

As discussed in Chapter 6, a higher amplitude pump clock allows the pump to have better output efficiency compared with one using a lower amplitude pump clock.

Figure 8-7 has been taken from Chapter 7 to show a charge pump with a 3 V pump clock and a charge pump with a 6 V pump clock. This plot shows the power transfer efficiency through the pump stages. Higher pump clock amplitude allows better charge pump transfer efficiency overall.

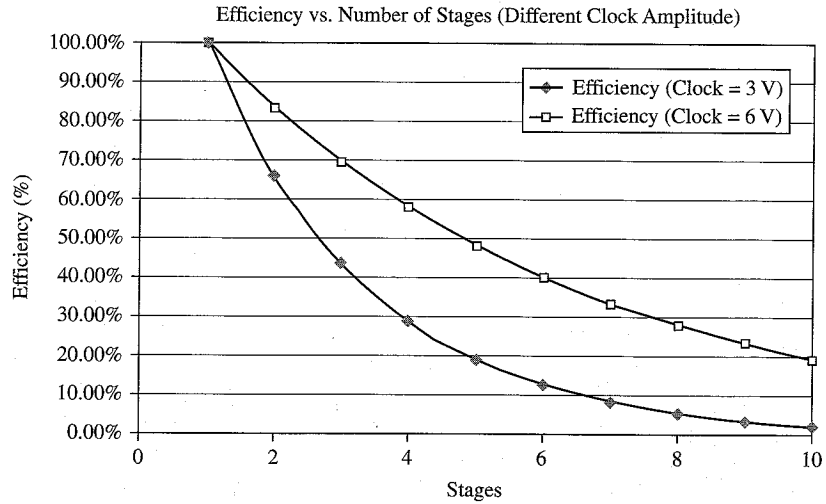


Figure 8-7 Charge pump efficiency vs. pump clock amplitude for 3 V and 6 V design.

This plot does not include the power consumption component from the peripheral supporting circuits. Normally the power consumption of the peripheral supporting circuits is low, and Figure 8-7 should not be affected too much if this extra component is included.

Figure 8-8 is a plot of the power consumption of the charge pump versus pump clock amplitude. The X axis is the output voltage of the charge pump. The Y axis is the total power consumed by the pump itself.

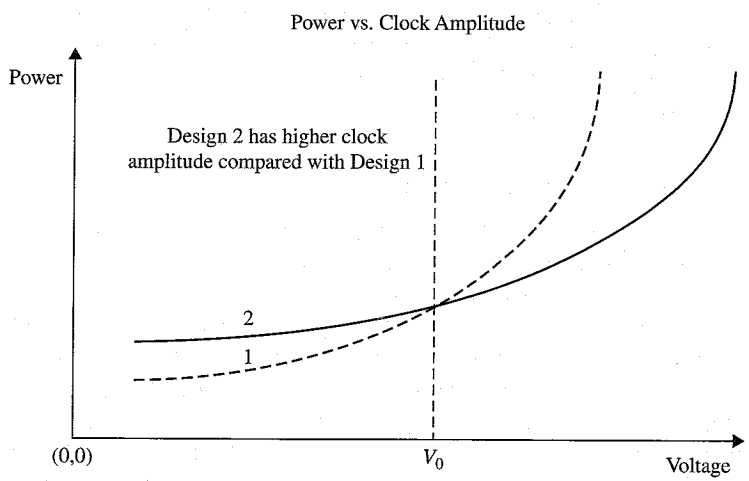


Figure 8-8 Power consumption vs. pump clock amplitude.

Two designs are compared in Figure 8-8. Design 1 is the charge pump with pump clock has regular clock amplitude. Design 2 has higher pump clock amplitude than Design 1. Additional on-chip circuits using a normal chip power supply generate the higher clock amplitude.

$V < V_0$ Design 2 consumes more power. (8-16)

$V > V_0$ Design 1 consumes more power. (8-17)

On the X axis, there is a critical output voltage level, V_0, that divides the pump operating spectrum. For an output voltage level that is less than V_0, although the higher amplitude clock in Design 2 can transfer the charge more efficiently, the additional power consumed by the clock generator will offset the gain. As a consequence, Design 2 consumes more power than Design 1. This difference should be small, as in Figure 8-8. For an output voltage level that is greater than V_0, the gain of the higher amplitude clock will exceed the extra power consumption in the clock generator. The benefit is obvious based on the calculation and is also shown in Figure 8-8. As a consequence, the overall power efficiency for Design 2 exceeds that of Design 1. Understanding the design specification and choosing the right approach can save on the total power consumption of the charge pump.

8.3 Noise Controls for the Charge Pump

Output noise of the charge pump can have detrimental effects on chip operations if the magnitude of noise exceeds the limit. With the pump output being regulated, the output noise will always exist in operations. Since the output load circuits vary and pumps are designed to work in different applications, minimizing the charge pump output noise becomes a challenging task. There are some common rules that can always be applied in real practice.

8.3.1 Noise versus filtering capacitance

The noise on the output of a charge pump should be contained within the limits given by the design specifications. What are some of the methods that can be easily applied in practice?

A very practical approach that can reduce the noise on the output node of the charge pump is to increase the decoupling capacitance connecting to the output of the charge pump. Figure 8-9 shows a typical representation of the pump output loads. First, regulation impedance (Z_{reg}) connects to the output node. This component is used to maintain the regulation level. This could be a passive element such as capacitance or resistance, or an active feedback control circuit. Second, a filtering capacitance (C_{filter})

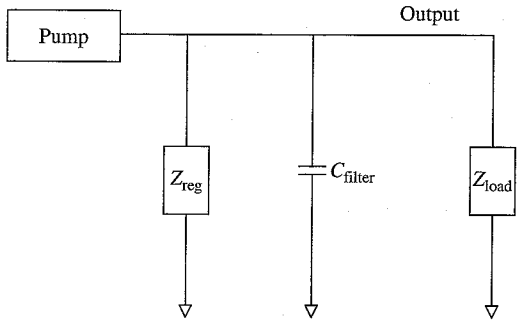


Figure 8-9 Pump output loadings.

is used to filter out some amount of noise on the pump output node. Third, the loading circuit (Z_{load}) uses high-voltage output to do its work.

How does C_{filter} come into play in the charge pump design? Because the gain of the amplifier is finite, and there is always delay from the regulation feedback to the active control of the charge pump, the output voltage of the charge pump always has some amount of output noise near the regulation levels.

If the ΔQ is the peak error near the regulation level, then the effect on the output voltage would be ΔV_{output}, as given in Equation 8-18. ΔV_{output} represents the noise on the pump output node near the regulation level. It has a linear relation with C_{filter}, as shown by the calculation.

$$\Delta V_{output} = \frac{\Delta Q}{C_{filter}} \tag{8-18}$$

As shown in Figure 8-10, as C_{filter} increases, ΔV_{output} goes down inversely proportional to the size of C_{filter}. The larger the filter capacitance, the less noise there is. Larger filter capacitance has a negative impact in terms of pump output voltage setup speed. Because the filter

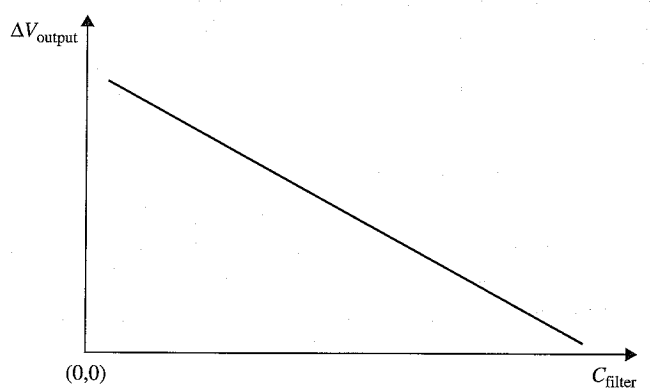


Figure 8-10 Pump output noise vs. decoupling capacitance.

capacitance needs to be charged up to the regulation level, more charge is needed to fill it up. If the output setup speed is not a concern, increasing C_{filter} would be a very easy approach. If output setup speed is critical in the design, the filter capacitance needs to be optimized for both noise and the setup speed of the output.

Another negative impact would be the layout area concern. In particular, the capacitance needs to use thick oxide devices to prevent any oxide breakdown issues. Therefore, the layout area will be larger than the one using thin oxide dielectric materials. However, the chip design usually has overdesigned the decoupling capacitance to filter out noise on power lines. Balancing and converting some of those decoupling capacitances into pump filtering capacitance may not cause serious power glitches.

8.3.2 Noise versus balance of pump power

The noise on the charge pump output shows a strong relation between the pump output power that can be delivered and the output loading circuit power consumption.[2] Refer back to Figure 8-9. According to Kirchoff's current law, at any moment the summation of all currents going into the output node of the pump should be equal to 0.

Equation 8-19 shows that, at any given time, the output current delivered by the pump is consumed by all the circuits connected to the output node. Three components are described in Equation 8-19. The first one is the regulation current, $I_{reg}(t)$. It can be either passive current consumed or active current shunt away. The second component is the loading circuit current, $I_{load}(t)$. The third component is $I_{filter}(t)$, which is the AC current going into or out of the decoupling capacitance.

$$I_{output}(t) = I_{reg}(t) + I_{load}(t) + I_{filter}(t) \tag{8-19}$$

If we rewrite Equation 8-19 into Equation 8-20, the difference between $I_{output}(t)$ and the sum of $I_{reg}(t)$ and $I_{load}(t)$ should be balanced by $I_{filter}(t)$ at any given time. If there is an introduction of noise, suddenly by $I_{output}(t)$, or a variation of $I_{reg}(t)$ or $I_{load}(t)$, the noise has to be filtered out by decoupling capacitance.

$$I_{filter}(t) = I_{output}(t) - \left[I_{reg}(t) + I_{load}(t)\right] \tag{8-20}$$

$I_{filter}(t)$ is formulated in Equation 8-21. It is proportional to the product of C_{filter} and ΔV_{output}. It is inversely proportional to the duration of this change, Δt. Rewriting Equation 8-21 into Equation 8-22, if the magnitude of $I_{filter}(t)$ is determined and Δt is unchanged, increasing filter capacitance C_{filter} allows ΔV_{output} to be decreased.

$$I_{filter}(t) = C_{filter}\frac{\Delta V_{output}}{\Delta t} \tag{8-21}$$

What if $I_{filter}(t)$ is reduced in magnitude? Then without a change in C_{filter}, ΔV_{output} would decrease by default, as in Equation 8-22. From Equation 8-20, at a given regulation level, if the power delivered to the output by the charge pump is closer to the power consumed by load and regulation circuits, then $I_{filter}(t)$ will be reduced, and so will the noise on the output of the charge pump.

$$\Delta V_{output} = \frac{I_{filter}(t)\Delta t}{C_{filter}} \tag{8-22}$$

Let us look at a simple example to explain the concept of the balance of different powers. In Figure 8-11 for charge pump output voltage, there are two regions distinguished in this figure. The first region is the ramp-up region. In this region, the charge pump has to charge up all the parasitic capacitance. In addition, it has to supply all the other load current and regulation current at the same time. The second region is the regulation region. The charge pump should supply only the load current and regulation current. Because capacitance is being charged up in the ramp-up phase, no DC current is associated with the capacitance in this phase. If there is a ramp-up speed requirement for the charge pump, the total current should be delivered by the charge pump to its output, is described in Equation 8-23.

$$I_{output}(t) = C_{cap}\frac{V_{reg}}{t_{rampup}} + I_{reg}(t) + I_{load}(t) \quad \text{for} \quad t < t_{rampup} \tag{8-23}$$

In the regulation region, the charge pump has to meet the output current requirement, as shown in Equation 8-24. The difference between Equation 8-23 and Equation 8-24 at $t = t_{rampup}$ is the output current difference between the two regions. This is summarized in Equation 8-25.

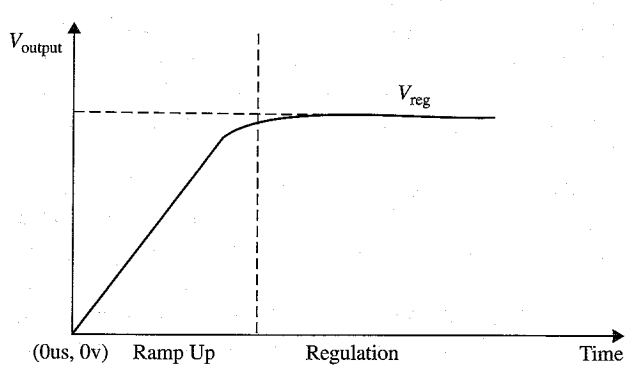


Figure 8-11 Pump output voltage vs. time

This difference involves the additional current needed to charge up the capacitive load in the ramp-up time required.

$$I_{output}(t) = I_{reg}(t) + I_{load}(t) \quad \text{for} \quad t > t_{rampup} \tag{8-24}$$

$$I_{output}\left(t_{rampup}\right) = C_{cap}\frac{V_{reg}}{t_{rampup}} \quad \text{for} \quad t = t_{rampup} \tag{8-25}$$

As a consequence, the charge pump needs to meet the maximum power requirement of the ramp-up region and the regulation region. The difference from Equation 8-25 becomes one of the main sources of noise on the output voltage of charge pumps. It can only be absorbed by the decoupling capacitor. Even the active control methods may not have enough time to respond in time to remove this noise. If this is known to be a noise source, why not deal with this noise from its root cause? The pump has to meet the power requirement of the worst-case scenarios. For the operation region that requires less pump output power, can the pump be adjusted accordingly to meet only the need of that region? The answer is yes.

To modify the output power of the charge pump in operation, we can apply several methods. You already know that the output power of the charge pump has the relationships shown in Equation 8-26.

$$\begin{aligned} P &\propto f \\ P &\propto C \\ P &\propto V_{clock} \end{aligned} \tag{8-26}$$

The charge pump output power is proportional to the pump clock frequency, f. Based on operation need, the clock frequency of the charge pump can be adjusted. If lower power consumption is expected, the clock frequency can be slowed down. The output power of the charge pump is reduced proportionally to frequency, too. If higher output power is needed on the output of the charge pump, clock frequency can speed up as well. The design needs to be checked out because not all charge pumps can adjust output power by varying clock frequency. This is architecture dependent.

The charge pump output power is proportional to the coupling capacitance, C, in each pump stage. If somehow the equivalent capacitance can be adjusted smaller or larger, the output power of the pump can be changed, too. Adjusting the equivalent capacitance means either switching in or out of pump capacitance, or several pump branches can be enabled or disabled based on operation. The equivalent is the same as

modifying the size of the coupling capacitance. The charge pump output power is proportional to the pump clock amplitude, V_{clock}. If the pump clock amplitude can be adjusted based on operation, the pump output power will vary as well.

Besides the methods given in Equation 8-26, many other approaches can be used to change the pump's performance. Switching between two operating regions can be easily detected by some analog circuits or based on state machine control signals. If analog detection is used, the feedback control signals can be used to control the charge pump internal circuits to vary the power consumption. In the worst case, a simple timing technique can be used to time out the different operations and then adjust the charge pump output power accordingly.

Balancing the charge pump output power is one of the most straightforward methods for minimizing the noise on the output of the charge pump. It is also one of the most powerful ways. The other methods that work on the outcome always lag behind in performance.

8.4 Off-Chip Charge Pump

For all the chapters presented so far, the focus has been on on-chip charge pump designs. As the power supply voltages of the chip are being continuously scaled down, and lower power consumption is demanded in the chip, the charge pump circuit on silicon requires a larger and larger layout area. The efficiency of pump design is also reduced as technology advances. Larger silicon area, increased parasitic components, larger I_{cc}, and larger IR drop are all factors that work against better designs. Could there be any alternative approach to the charge pump design that can alleviate these pressures?

In the early days, high-voltage generation was created off chip for devices operations such as using EPROM/EEPROM programmers. As technology advanced, charge pumps were moved to chips for convenience of design. As the power supply scaled down and the power demand increased on chip, the penalties for the design increased. The die size needed to be increased to meet the power demand. As die size is increased, larger IR drop occurs on the internal power buses. Because a larger current is needed at a lower power supply voltage, bonding wire causes more inductive noise on the internal power supply in transit. Larger decoupling capacitance was needed to keep the power supply noises low. All these effects would have a negative impact on the chip designs.

Taking these drawbacks into account, would it be better to move some parts of the high-voltage generation circuits off the chips? This indeed goes against the trend of integrating all components onto a single chip. The penalty associated with this approach is the extra discrete components

required by the system-level design. However, the penalty could be recouped by the gain from smaller chip size and better charge pump efficiency. In this chapter, two possible approaches are presented for discussion. One approach is the use of off-chip capacitive charge pumps, and the other approach is the use of off-chip inductive charge pump designs.

8.4.1 Off-chip capacitive charge pump

By moving some of pump capacitance external to the chip, using an off-chip capacitive charge pump could be one of the approaches to improving charge pump efficiency. Figure 8-12 shows a simple configuration of off-chip capacitive charge pump design. The chip has two extra pins: the HV pin receives high-voltage input potential from the external supply, and a pin for the CLK signal is used to drive the off-chip capacitance, C_1. There are two extra discrete components at the system level: One is boosting capacitance, C_1, connected between signals HV and CLK. The other is the diode D_1 connected between the off-chip power supply V_{cc} and the HV pin. If we look at the configuration of the diode and boosting capacitance as a whole, this configuration looks very similar to one of the pump stages used in 2-phase charge pump design.

As shown in Figure 8-13, clock CLK is pulsing between 0 V and V_{cc} over time. The output voltage HV is responding to the clocking signal through capacitance C_1. This generated high-voltage potential is the supply to the internal of the chip.

Between time 0 and t_1, CLK is 0 V. HV is being precharged to at least $V_{cc} - V_d$ by V_{cc} through diode D_1. In between time t_1 and t_2, CLK switches from 0 V to V_{cc}. If boosting capacitance C_1 dominates over the

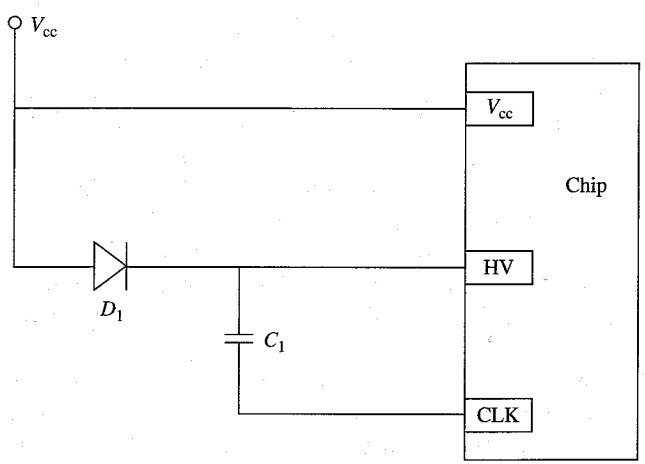


Figure 8-12 Off-chip capacitive charge pump.

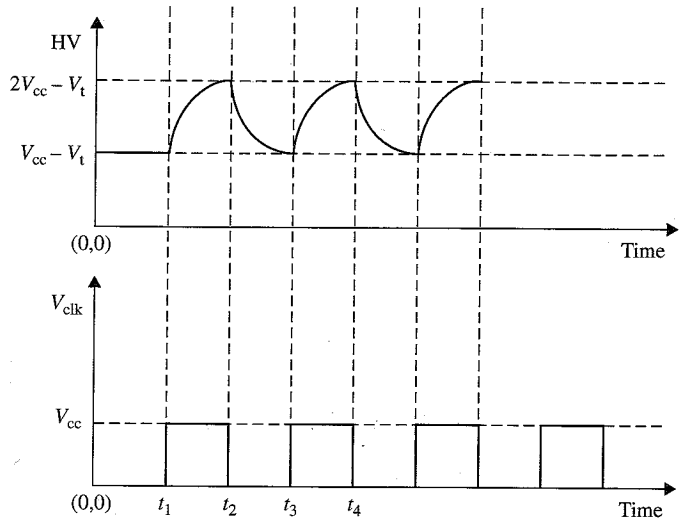


Figure 8-13 Waveforms of the off-chip capacitive charge pump.

loading capacitance seen on the HV node, capacitive coupling through C_1 will bring the HV node to $2V_{cc} - V_d$. At t_2, CLK switches from V_{cc} to 0 V. Reverse coupling would couple down HV. However, because diode D_1 is presented, HV would be clamped to a voltage level at least near $V_{cc} - V_d$. As a consequence, the average potential of HV exceeds V_{cc}.

Because the precharge supply is off-chip, the IR drop due to routing on chip is removed. The clock driver is near the pad, so the IR drop to the driver will be significantly less than that of the on-chip charge pump. On the chip itself, this elevated potential voltage (HV) could be used to drive the other part of pump circuits. The advantage of using a high-amplitude clock design was presented in Chapter 6. The benefit is obvious from this discussion.

In Figure 8-12, the design is only targeting for the supply power within the half clock cycle. In the other half clock cycle, the potential on HV drops because there is no driver. To supply the charge in the complement phases of the clock, another branch of diode and capacitance can be added. This capacitance is driven by the complement of the clock signal. The revised design is shown in Figure 8-14. To isolate two branches from interfering with each other, two extra diodes are added to the output of each branch. These additions prevent the charge on HV from leaking backward.

Figure 8-13 and Figure 8-14 show a design using off-chip capacitive charge pump designs capable of delivering a voltage level closer to $2V_{cc}$. If voltage higher than $2V_{cc}$ is needed, the design can be revised to add extra stages.

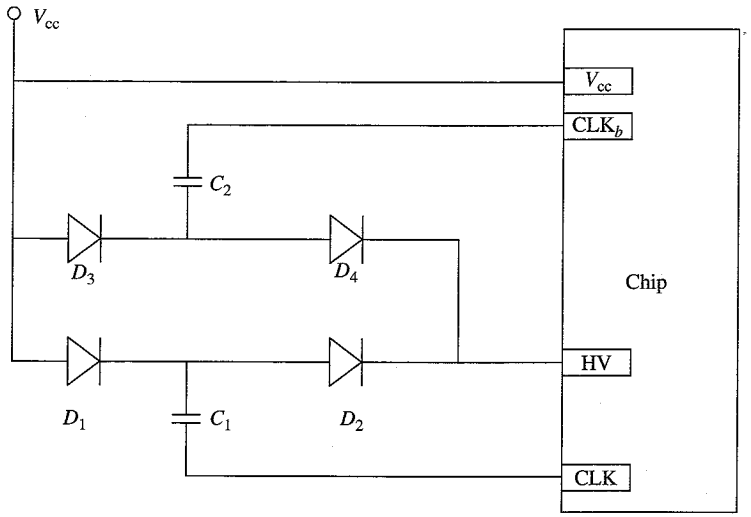


Figure 8-14 Revised off-chip capacitive charge pump with complementary clocking scheme.

Figure 8-15 is an off-chip capacitive charge pump design using an approach similar to the normal 2-phase charge pump. As more diodes and capacitance are being cascaded in serial, higher and higher potential can be achieved on node HV. Calculations must definitely be done to show this approach can be more beneficial in reducing the overall design cost.

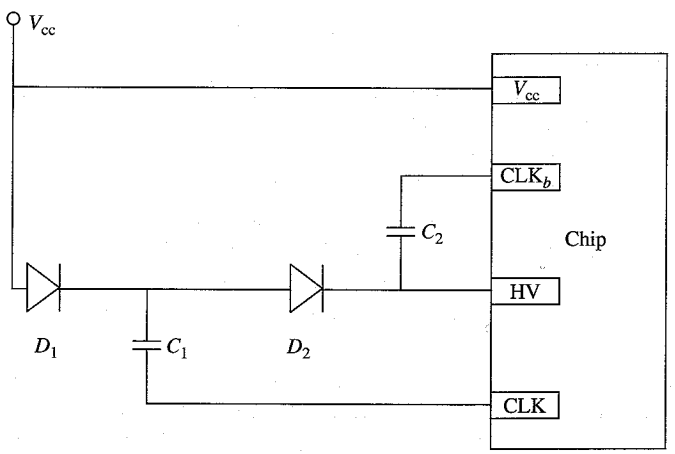


Figure 8-15 Off-chip 2-phase capacitive charge pump.

8.4.2 Off-chip inductive charge pump

Most of the charge pumps are constructed by some means of capacitive coupling. As seen in Figure 8-14 and Figure 8-15, more discrete components are needed to get higher and higher voltages. The cost adds up with each discrete component being introduced. How about using capacitance instead? Can inductive components be used in pump design? The answer is yes. In earlier chapters, we mentioned that high voltage potential could be generated by transformers and rectifiers. The design works for an AC power supply but not a DC supply. Inductance requires a change of current over time to generate high voltage potential electromagnetically.

Figure 8-16 shows a generic design using off-chip inductance to generate high voltage potential. The chip has one port, HV, connected to the external supply. There is one inductor, L_1, connected between V_{cc} and HV. Decoupling capacitance C_1 is used to stabilized the off-chip power supply, V_{cc}. On the chip itself, there is one NMOS transistor, N_1, with a gate connected to CLK. The drain is connected to the HV pin, and the source is connected to the ground power supply. There is one diode, D_1, that is connected between the HV pin and the internal high-voltage supply V_{sup}.

How does the inductive charge pump work? As shown in Figure 8-16, the NMOS transistor N_1 is being clocked. Between time t_1 and t_2, CLK is high and N_1 is conducting a DC current of t_{N1}. The HV pin is being pulled down all the way to ground. As the current flows through inductance L_1,

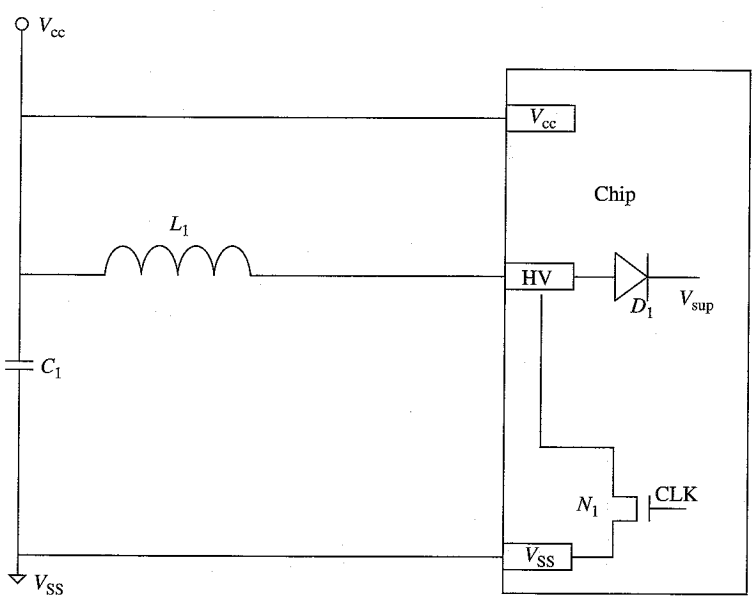


Figure 8-16 Off-chip inductive charge pump.

the electromagnetic energy is gradually stored in L_1 during this period. Between t_2 and t_3, CLK goes from high to low, and N_1 switches from conducting a DC current of t_{N1} to 0. For L_1, when the current changes in amplitude abruptly, the electromagnetic energy stored in the inductor will be released to impede the change of current. As a consequence, HV increases sharply according to Equation 8-27 to maintain the current flowing in the same direction.

$$\Delta V = L\frac{\partial I(t)}{\partial t} \tag{8-27}$$

At the peak of each pulse in Figure 8-17, voltage will be passed to V_{sup} through the diode D_1. Resonance occurs as the energy is transferred between the capacitance and inductance. At every peak of positive voltage, charge will be passed into the internal node.

As shown in Figure 8-18, V_{sup} rises in staircases as the charge is transferred to it. The diode prevents the charge from leaking backward once it goes up in amplitude. Over a few pulses, internal supply V_{sup} can reach very high potential. The highest voltage the output can reach is calculated based on Equation 8-27. Inductive charge pump design is simple in principle, and much fewer components are needed to generate a high voltage. It could be a good candidate for improving layout area efficiency.

Off-chip inductive charge pumps have been used in many applications and the method was proposed recently by a leading company producing non-volatile memories as a means to reduce silicon area and the chip's power consumption.

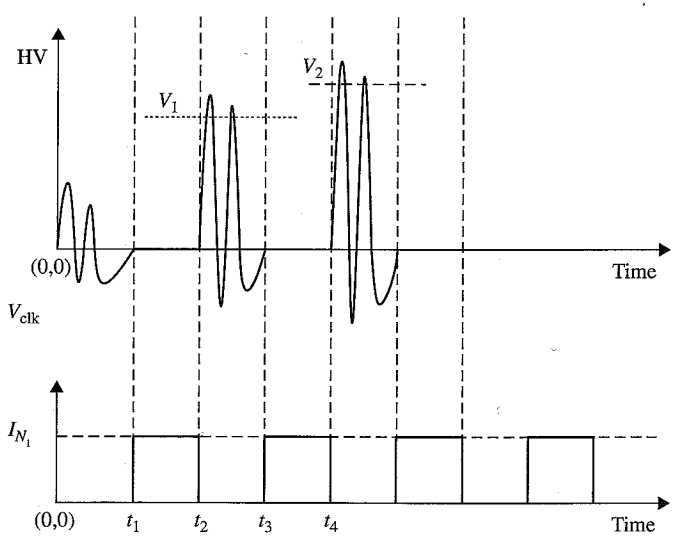


Figure 8-17 Off-chip inductive charge pump waveform.

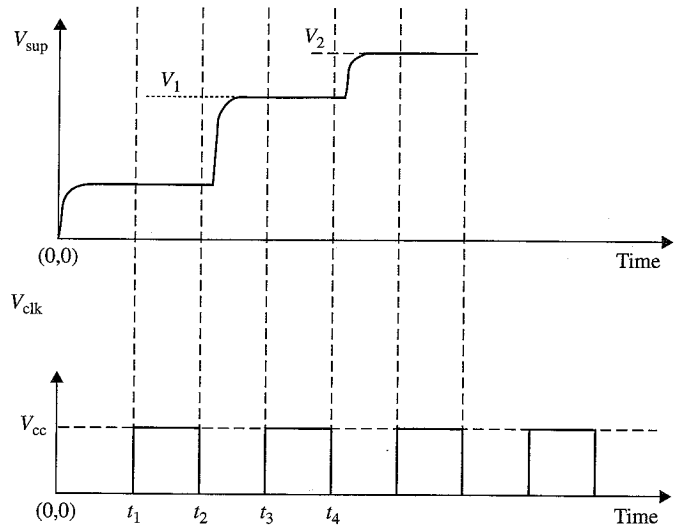


Figure 8-18 Off-chip inductive charge pump output voltage.

8.5 Conclusion

Charge pump design is challenging. The attempt to constantly improve the existing approaches and find a better solution has never stopped. Although we cannot foresee the future, the problems pump designers face today should be the targets that new ideas try to conquer. The future pump has to be small, power efficiency and well regulated on its output. In this chapter, we have discussed the design approaches to improve each aspect. We hope these starting points could lead others to develop better approaches.

References

1. Dickson, J.K. "On-chip high voltage generation in NMOS integrated circuits using an improved voltage multiplier technique." *IEEE Journal of Solid-State Circuits*, Vol. SC-11, pp. 374–378, June 1976.
2. Pan. "High voltage ripple reduction." U.S. patent 6,734,718.
3. Di Cataldo, G. and G. Palumbo. "Design of an Nth order Dickson voltage multiplier." *IEEE Transactions on Circuits and Systems*, I, Vol. 43, pp. 414–418, May 1996.

Chapter

9

A Practical Charge Pump Design Example and Analysis

To complete our studies of the charge pump, we will build a simple charge pump circuit in this chapter. We will simulate, study, and analyze its operation while performing a series of experiments involving the pump's output voltage variation versus clock frequency and clock frequency amplitude, number of stages, and parasitics variation. We will also derive the pump efficiency and pump I-V characteristics. The simulations and the explanations in this chapter will help fine-tune different aspects of the charge pump's design variables, such as the actual number of stages, the optimum clock frequency, the optimum output operating voltage, the MOSFET diode sizes, the pump clock amplitude, and so on.[3] Even though at the time of the writing of this book, almost all leading companies are using a 0.18 μ technology process or beyond, most of the circuit's characteristics remain unchanged throughout the various generations of the technology.

For this chapter, simulations were performed using a 0.8 μ CMOS AMI process available through MOSIS technologies. This CMOS process is identical to a standard 0.8-μm process, with the addition of a second poly layer to accommodate linear capacitors. This provides dense high-performance capable of integrating complex analog functions. The process features two metal layers and one poly layer, and has a linear cap option. It is important to note that, even though these models may not be up to date, all simulations were done using them. The main intention of all the following simulations is to show the "trend" and different characteristics of the pump.[1] Following Figure 9-1 are some details about the process.

Run: N85F			Vendor: AMI		
Technology: SCN10			Feature Size: 1.0 microns		
Transistor Parameters	W/L	N-Channel	P-Channel		Units
Minimum Dimension	1.5/1				Microns
V_{th}		0.72	–0.97		Volts
I_{dss}		387	–187		μa/μm
Delta length		0.38	0.3		Microns
(L_eff = L_drawn-DL)					
Delta width		0.33	0.55		Microns
(W_eff = W_drawn-DW)					
K' ($U_o \times C_{ox}/2$)		55	–17.2		μa/V^2

Figure 9-1 0.8 μ CMOS AMI process.

The BSIM3v3 level 49, NMOS and PMOS models used for the simulations are shown here:

```
* N85F SPICE BSIM3 VERSION 3.1 (HSPICE Level 49) PARAMETERS

* DATE: 98 Jul  7
* LOT: n85f                    WAF: 07

.MODEL CMOSN NMOS                                    LEVEL   = 49
+VERSION = 3.1          TNOM    = 27                 TOX     = 1.54E-8
+XJ      = 1.5E-7       NCH     = 1.7E17             VTH0    = 0.6910337
+K1      = 0.8700926    K2      = -0.0394636         K3      = 11.3177837
+K3B     = -1.2220445   W0      = 4.359393E-6        NLX     = 9.229978E-9
+DVT0W   = 0            DVT1W   = 5.3E6              DVT2W   = -0.032
+DVT0    = 2.2790092    DVT1    = 0.42292            DVT2    = -0.1302476
+U0      = 560.1408795  UA      = 1.158328E-9        UB      = 9.045045E-19
+UC      = 4.943224E-11 VSAT    = 1.055535E5         A0      = 0.7340357
+AGS     = 0.1064104    B0      = 2.384737E-7        B1      = 6.630751E-7
+KETA    = -9.0161E-3   A1      = 0                  A2      = 1
+RDSW    = 1.859947E3   PRWG    = -1E-3              PRWB    = 0
+WR      = 1            WINT    = 1.52595E-7         LINT    = 1.84968E-7
+DWG     = -1.522374E-8 DWB     = 2.371351E-8        VOFF    = -0.0959822
+NFACTOR = 1.0215504    CIT     = 0                  CDSC    = 4.654432E-4
+CDSCD   = 0            CDSCB   = 0                  ETA0    = 1.937282E-3
+ETAB    = 0            DSUB    = 0.03403            PCLM    = 0.7224211
+PDIBLC1 = 0.019077     PDIBLC2 = 9.376978E-4        PDIBLCB = -1E-3
+DROUT   = 0.3136688    PSCBE1  = 1.732251E9         PSCBE2  = 5E-9
+PVAG    = 0.0245002    DELTA   = 0.01               MOBMOD  = 1
+PRT     = 0            UTE     = -1.5               KT1     = -0.11
+KT1L    = 0            KT2     = 0.022              UA1     = 4.31E-9
+UB1     = -7.61E-18    UC1     = -5.6E-11           AT      = 3.3E4
+WL      = 0            WLN     = 1                  WW      = 0
+WWN     = 1            WWL     = 0                  LL      = 0
+LLN     = 1            LW      = 0                  LWN     = 1
+LWL     = 0            CAPMOD  = 2                  CGDO    = 3.5E-10
+CGSO    = 3.5E-10      CGBO    = 0                  CJ      = 4.482658E-4
+PB      = 0.8907125    MJ      = 0.4317331          CJSW    = 3.002331E-10
+PBSW    = 0.4673283    MJSW    = 0.195429           PVTH0   = -4.584654E-3
+PRDSW   = -490         PK2     = 0.0147474          WKETA   = 3.122681E-3
+LKETA   = -1.780912E-3
*
```



```
.MODEL CMOSP PMOS                                    LEVEL   = 49
+VERSION = 3.1          TNOM    = 27                 TOX     = 1.54E-8
+XJ      = 1.5E-7       NCH     = 1.7E17             VTH0    = -0.9307976
+K1      = 0.4138836    K2      = 0.0229645          K3      = 19.3590002
+K3B     = -0.2309329   W0      = 1.119804E-6        NLX     = 1.597682E-7
+DVT0W   = 0            DVT1W   = 5.3E6              DVT2W   = -0.032
+DVT0    = 2.8021132    DVT1    = 0.408755           DVT2    = -0.0673098
+U0      = 209.7071291  UA      = 2.350293E-9        UB      = 1.022986E-18
+UC      = -4.43816E-11 VSAT    = 1.693469E5         A0      = 0.9126827
+AGS     = 0.1608997    B0      = 9.768108E-7        B1      = 2E-6
+KETA    = -6.464748E-3 A1      = 0                  A2      = 1
+RDSW    = 2.641364E3   PRWG    = -8.079845E-6       PRWB    = -1E-3
+WR      = 1            WINT    = 2.028229E-7        LINT    = 1.370548E-7
+DWG     = -2.468532E-8 DWB     = 2.892704E-8        VOFF    = -0.0751808
+NFACTOR = 0.410862     CIT     = 0                  CDSC    = 4.363744E-4
+CDSCD   = 0            CDSCB   = 0                  ETA0    = 0.0132505
+ETAB    = -1.09027E-3  DSUB    = 0.1228106          PCLM    = 3.4356635
+PDIBLC1 = 0.9994144    PDIBLC2 = 2.090274E-3        PDIBLCB = 0
+DROUT   = 0.7969462    PSCBE1  = 3.519422E9         PSCBE2  = 5E-9
+PVAG    = 0.8461572    DELTA   = 0.01               MOBMOD  = 1
+PRT     = 0            UTE     = -1.5               KT1     = -0.11
+KT1L    = 0            KT2     = 0.022              UA1     = 4.31E-9
+UB1     = -7.61E-18    UC1     = -5.6E-11           AT      = 3.3E4
+WL      = 0            WLN     = 1                  WW      = 0
+WWN     = 1            WWL     = 0                  LL      = 0
+LLN     = 1            LW      = 0                  LWN     = 1
+LWL     = 0            CAPMOD  = 2                  CGDO    = 3.5E-10
+CGSO    = 3.5E-10      CGBO    = 0                  CJ      = 6.292692E-4
+PB      = 0.9263477    MJ      = 0.4891402          CJSW    = 4.460985E-10
+PBSW    = 0.95         MJSW    = 0.4270889          PVTH0   = 0.0462392
+PRDSW   = -680         PK2     = 8.816106E-3        WKETA   = 7.199117E-3
+LKETA   = -4.523899E-3
*
```

Source: These MOSFET models have been printed for use in this book only, with kind permission from Tamera Drake, the WW Marcom Manager for AMI Semiconductor (http://www.amis.com).

Simulations were performed using WinSpice, a HSPICE-based circuit simulator, on a PC running Windows XP. WinSpice is ported to run in a window as a native 32-bit application. It can generate waveform plots in individual floating windows and contains a powerful scripting language. It is available for free for individual use and for a nominal license fee for commercial use and can be downloaded from http://www.winspice.com. WinSpice supports BSIM3, JFET2, and BSIM4 device models. It has zoomable plot windows and also supports parameterized sub-circuits. It also has an excellent manual and offers a built-in tutorial. WinSpice itself does not contain a schematic editor. However, third-party schematic editors can be linked to it. A schematic editor that uses WinSpice for its simulation engine is available at http://www.5spice.com.

9.1 Design Specification

Referring back to Chapter 4, a simple 2-phase charge pump was designed as shown in Figure 9-2. The initial design specifications are shown in Table 9-1.

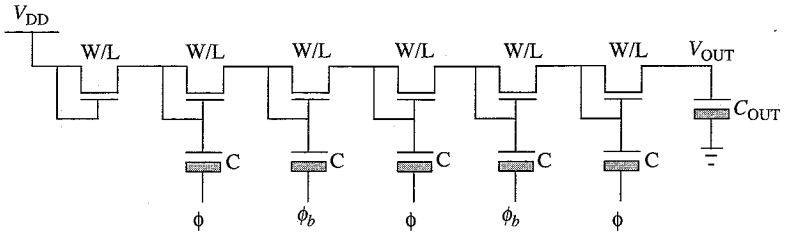


Figure 9-2 A 6-stage Dickson charge pump

9.2 Design Steps

Next, the following steps will take you through a detailed series of four steps needed to define the basic parameters of the charge pump, such as how to accurately determine the number of pump stages, the initial pump operating frequency, the pump boosting capacitor size and the size of the diode-connected MOSFETs, all geared toward creating a more efficient charge pump.[2]

9.2.1 Step 1: Determine *N*, the initial number of stages

Because the target V_{out} is 15 V, as a rule of thumb, we should design a charge pump such that the required regulated output voltage is about 75% of the maximum charge pump's output voltage. More details on this assumption are provided later in this chapter when we extract the pump's I-V characteristics. Because our target charge pump output voltage requirement is 15 V, we should design a charge pump whose maximum output voltage is close to or higher than 20 V. The assumption is that because this is a high-voltage-generating pump, the output needs to be regulated precisely at a fixed amplitude. Hence, in a real implementation, we will need to use a voltage regulator to regulate the output voltage at 15 V.

Using the original charge pump's formula $V_{out} = NV_{\phi} + V_{in} - (N+1)V_D$, the output voltages for three different conditions are derived and shown in Table 9-2. Note V_{in} is actually V_{DD} in this case.

TABLE 9.1 Charge Pump Specifications

Specification	Specification Value
Steady output voltage	15 V
Output load capacitance	20 pF
Output voltage ramp-up time	10 μS
Pump power supply voltage	5 V
Average pump current consumption	1 mA

TABLE 9.2 Dickson Charge Pump's Output Voltage vs. Input Conditions

Iteration	N	V_{ϕ} (V)	V_{in} (V)	V_d (V)	V_{out} (V)
1	6	4	4	1.5	17.5
2	6	5	5	2.0	21.0
3	7	5	6	2.1	23.2
4	6	6	6	2.2	26.6

A few points need to be clarified here. V_D, the MOSFET's threshold voltage, is actually a function of the source voltage and will gradually increase as the charge pump is ramping up. Even though the zero bias V_t is close to 0.69 V, the actual V_t in the last few stages of the pump will rise close to 1.8 V or higher (note that the V_t dependence on V_{SB} can be found in a particular technology's process manual document or can be derived through a few sets of simulations). Here in Table 9-2, a simple average V_t has been assumed, for preliminary simulation purposes. For now, we can start our simulations with the values for iteration 2 in Table 9-2.

9.2.2 Step 2: Determine *f*, the initial pump operating frequency

In general, the pump's operating frequency does not need to touch the stratospheric frequency limits of today's semiconductor processes. A modest frequency in the range of about 2 MHz to 50 MHz is generally used. The operating frequency is dictated more by the pump capacitor size, diode size, and parasitics than by the process limits. For the present case, we will assume a frequency of 20 MHz, which translates to a 50-ns clock period.

9.2.3 Step 3: Determine *C*, the initial size of the pump capacitor

Using Equation 5-4 in Chapter 5 we can roughly determine the value of the average linear pump current during ramp up. For our case, C_{load} = 20 pf, V_{out} = 15 V, and T_{rampup} = 10 μS.

$$i_{linear} = C_{load}\frac{V_{out}}{T_{rampup}} = 20\times10^{-12}\frac{15}{10\times10^{-6}} = 30\,\mu A$$

Now use i_{linear} to find the value of C using Equation 5-5 from Chapter 5. In this case, we will assume dV = 1 V and the frequency of oscillation is 20 MHz.

$$I_{out} = 30\,\mu A = C\frac{dV}{dT} = C\frac{1}{50\times10^{-9}}$$

$$\Rightarrow C = 1.5\,pF$$

We will use this capacitor size for our initial simulation and then optimize it along with other parameters.[4]

9.2.4 Step 4: Determine the diode *W/L*, the initial size of the diode-connected MOSFET

The MOSFET width should be at least 10 to 15 times higher than the average output current it is delivering at the output. The reason is that, in general, the pump will be driving a high current at the output at the initial stages of ramp up, and this current will gradually taper off exponentially as the output voltage reaches the target value. The initial current can be close to 10 × higher than the average current the pump is expected to deliver, and hence the channel width should be high enough to accommodate this I_{ds}. The appropriate value of the channel width for a particular I_{ds} can be determined from the MOSFET I_{ds}–versus–V_{ds} plots. Next, because the defined minimum channel length is 1 μ, we will use MOSFETs with a 1-μ channel length to produce the minimum resistance possible. For this particular process, we will use a gate width of 100 μ and a gate length of 1 μ.

9.3 Initial Simulation and Analysis

With the initial values of the preceding basic parameters already determined, we can start our simulations with the circuit shown in Figure 9-3. Apart from the pump circuit, we also need to create a non-overlapping clock generator and clock booster. For simplicity, those circuits will not be shown here, but it is assumed that the clock sources clk and clk_b are derived from a dedicated clock generator. We will simulate the iteration 2 case from Table 9-1, which will produce a maximum output voltage of about 21 V, because we assumed that we should target our pump's maximum voltage such that the actual required voltage, 15 V in this case, is about 75% of the maximum voltage attained.

Figure 9-4 shows the charge pump output voltage over a time period of 25 μs. As you can see, the output voltage has not saturated yet, and

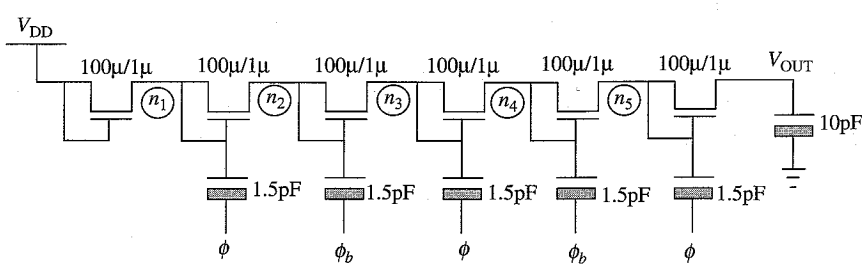


Figure 9-3 A 6-stage Dickson charge pump with initial values of device sizes.

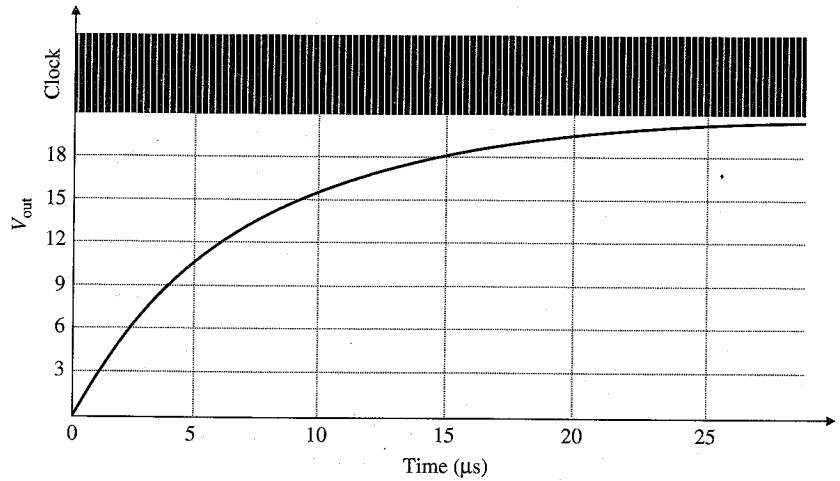


Figure 9-4 Pump output voltage vs. time.

if left simulating it will rise up to about 21 V. The output shows an almost exponential characteristic. As the output and internal node voltages rise, the transistor's V_t also increases, thereby limiting the charge transfer per stage and thus slowing down the output charge-up time. Further, this simulation was done with no resistive load at the output, so the output voltage characteristics will be lower than this simulation's results, if such resistive load exists.

Figure 9-5 shows the internal charge pump voltages during a certain phase of the charge pump's operation. The internal nodes n_1–n_5 have been marked during the phase when cllk is high and clkb is low. As can be seen from the waveforms, the charge transfers between each node are almost complete, and the nodes (such as n_1–n_2, n_3–$_n4$, n_5–V_{out}) are almost stable before the start of the next clock phase. In general, the slope of these node voltages will determine the suitability of the choice of the clock frequency and the overall pump efficiency. As discussed earlier, the difference between the node voltages, such as n_1 and n_2, is a combination of the V_t of the corresponding MOSFET connecting these two nodes and the resistive voltage drop if the pump is supplying an average DC current. But in general the MOSFET V_t is the dominant factor.

The preceding charge pump simulation assumes an ideal scenario (i.e., no parasitic resistances or capacitances are present). But as mentioned in the preceding chapters, parasitics play a major role in defining the charge pump's output voltage and output current. As a rule of thumb, we can assume parasitic capacitances at the internal pump nodes to be about 10% of the charge pump's capacitors. Because we are assuming 1.5 pF pump capacitance, let us assume 0.15 pF parasitic capacitance.

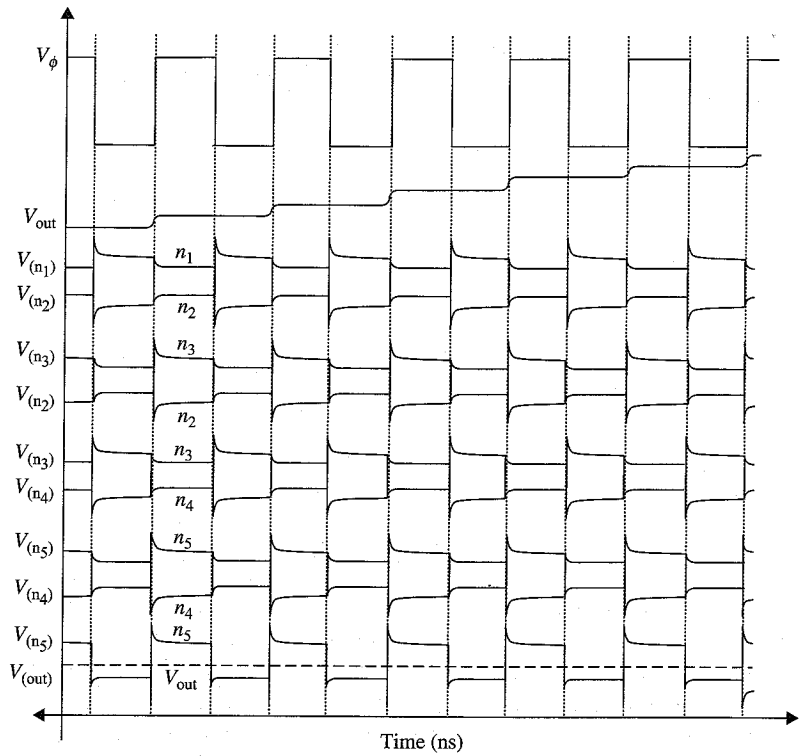


Figure 9-5 Pump internal waveforms/stage.

Depending on the layout and the pump size, this assumption should be modified. We can ignore the circuit resistances by assuming we will use wide metals for the interconnection and a generous number of contacts and vias to keep the connecting resistances down. Our modified simulation setup with the parasitic capacitances is shown in Figure 9-6.

You need to understand that the pump simulations must be run again after the complete layout is done. Also, the parasitics, including

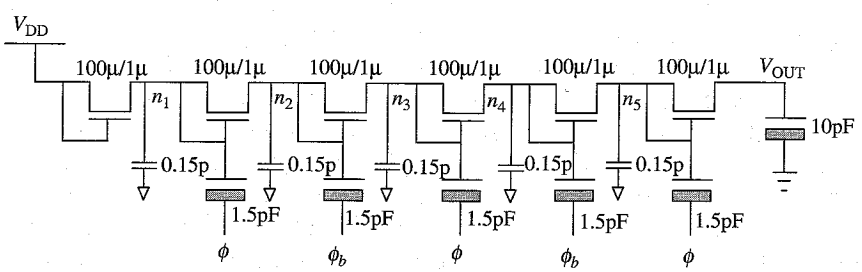


Figure 9-6 Charge pump schematic with parasitic capacitances.

resistances and capacitances, are extracted from the layout and back-annotated to the SPICE simulation. It is almost always the case that the pump will exhibit some more degradation, due to unseen parasitic capacitances, mostly from cross-coupling, extra routing, or from signal lines running over the pump in higher metal layers. Further, finite resistances at different nodes will also contribute to some degradation. It is the task of the circuit designer to carefully monitor each parasitic capacitance (or group of capacitances) and improve the layout wherever possible by tweaking the metal routing and connections. This may create a few iterations, but a rigorous attempt must be made to make sure that the parasitics are kept to a minimum.

A final back-annotated simulation must be made to verify pump performance. Figure 9-7 shows such a pump simulation result. The top waveform corresponds to the pump characteristics with no parasitics. The middle one corresponds to the pump characteristics with the estimated capacitance, and the bottom waveform corresponds to the waveform with back-annotated parasitics. Also, owing to the availability of only one type of model (typical), all the simulations performed here were at a typical corner (i.e., typical process, room temperature, and typical supply voltage). In reality, because all the three major parameters are independent variables, simulations should be performed across different conditions, and the worst case should be chosen. In general, the worst-case process corner (high V_t), high temperature (higher resistances), and lowest possible in-circuit supply voltage will exhibit the worst case scenario.

As mentioned in Chapter 5, the charge pump may be modeled as a voltage source with a finite input impedance. Hence, when a load capacitance is connected at the output, the voltage across the capacitor

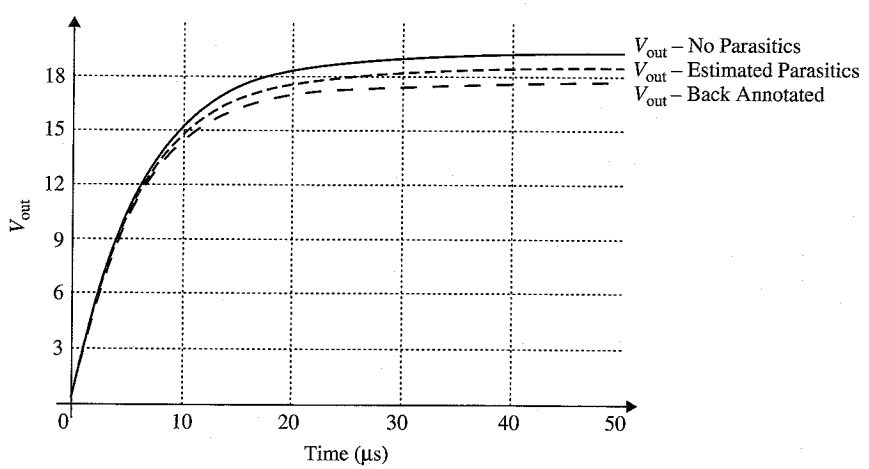


Figure 9-7 Pump output vs. pump parasitics.

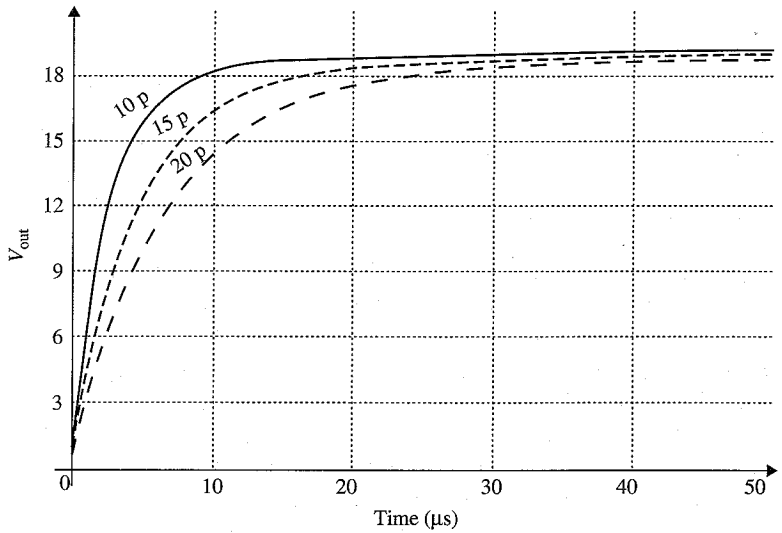


Figure 9-8 Pump output vs. output load capacitance.

will increase exponentially to attain the final output voltage. Again, if the output capacitance is increased/decreased, the output voltage will change accordingly. Figure 9-8 shows such a condition. The bottom waveform shows the output ramp with the default capacitive load, and the next two correspond to incremental loads.

Figure 9-9 shows the output voltage variation with clock frequency. The present simulations were done using a 50-ns clock, and then the

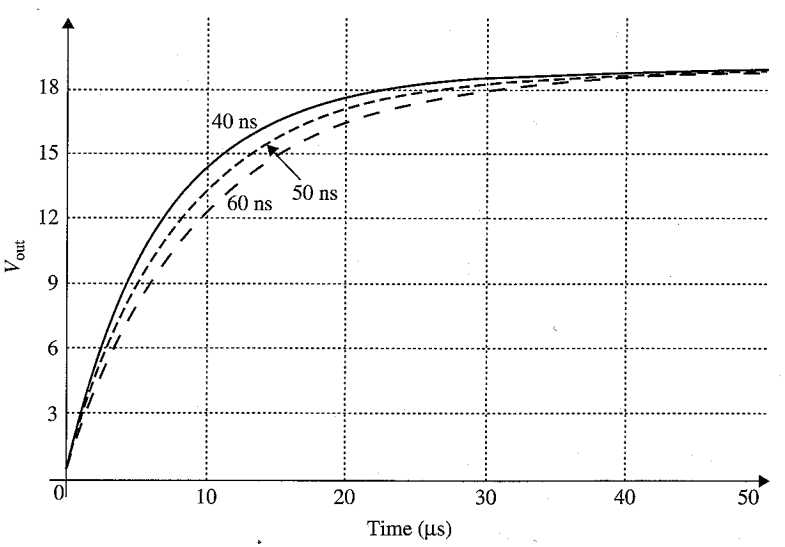
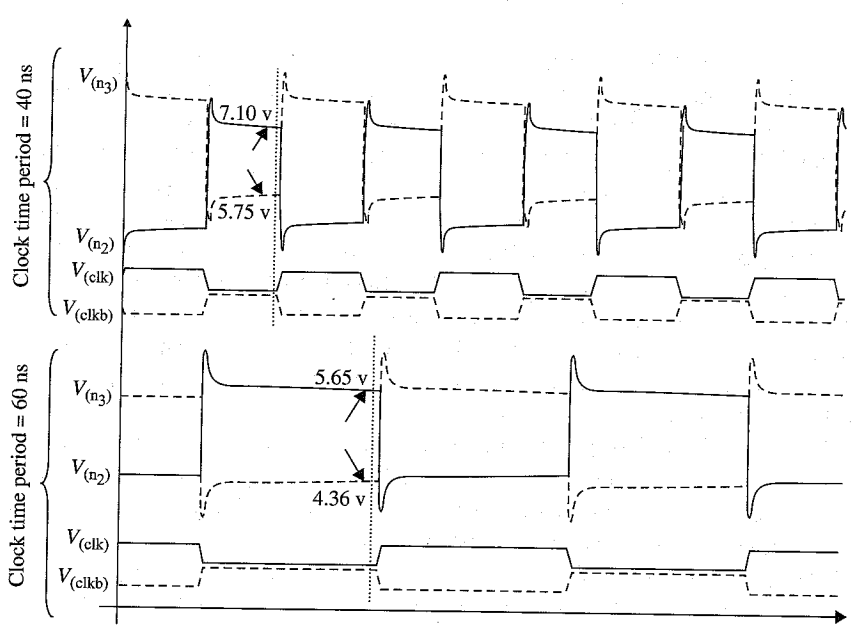


Figure 9-9 Pump output vs. clock frequency

clock was varied 20% higher and lower to observe the pump performance. As can be seen, increasing the clock frequency will increase the amount of charge transferred over a certain fixed time interval, which will increase the output voltage faster. But this process has its limits—increasing the clock frequency too high will result in incomplete charge transfer, resulting is reduced pump performance and efficiency.

Figure 9-10 shows the pump's internal voltages, almost at the similar time during initial ramp up. As can be seen, for the clock with the 40-ns time period, the voltages at nodes n_2 and n_3 increase faster than when the clock time period is 60 ns. The voltage difference at the end of the charge transfer for the case with clock period of 60 ns is about 1.29 V, whereas the voltage difference for the case with the clock period of 40 ns exhibits 1.35 V. Because there is no DC current load component, we can assume that these voltages are closer to the MOSFET's V_t, which is true at least for the 60-ns clock case. Because it is almost always desirable for a charge pump to ramp up as fast as possible, increasing the clock frequency is a quick-and-easy way to accomplish this. Studying the internal node voltages will be crucial for determining the actual operating clock frequency.

Looking back at nodes n_2 and n_3 for the 60-ns clock period case in Figure 9-10, we can see that the node voltages are almost saturated and there is very little charge sharing after about halfway when the



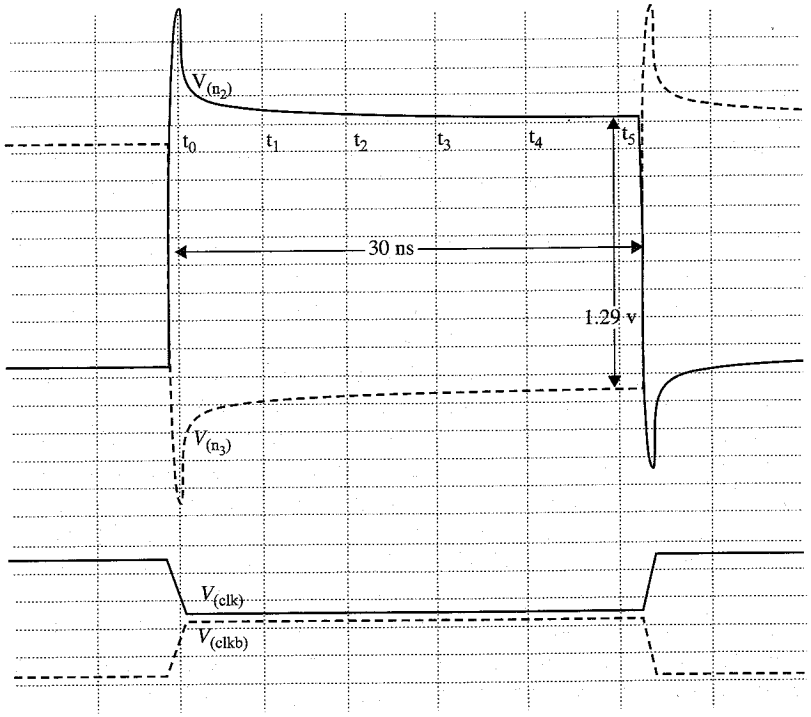


Figure 9-11 Detailed pump internal waveforms vs. clock frequency.

clock clk_b went high. Figure 9-11 zooms in on this region and shows it in greater detail. The whole 30-ns clk_b pulse high period has been divided into small segments and marked as $t_0, t_1, \ldots t_5$. We can see that most of the charge sharing occurs during the time t_0 to t_3. After that, the rate of charge transfer slows down and is predominated by the V_t of the MOSFET connecting nodes n_2 and n_3. Therefore, in this scenario, for this type of charge pump, the clock period can be reduced to at least t_4 (see Figure 9-11) and to at most t_3. More simulations with all parasitics need to be run at different simulation conditions to determine the suitable operating frequency. Further, the circuit designer needs to pay attention to the pump's increased current consumption due to the higher frequency.

Next, we should discuss the effects of the pump clock amplitude. Referring back to Table 9-1 at the start of this chapter, it was calculated that if we use the clock amplitude of 4 V, we should expect about 17 V at the output. A 6-V clock will allow us to go higher than 25 V, and a 5 V clock, the one we are using as the default, will allow us to go as high as 21 V. A simulation was run with clock amplitudes of 4 V and 5 V for comparison, and the results are plotted in Figure 9-12. Again, if the

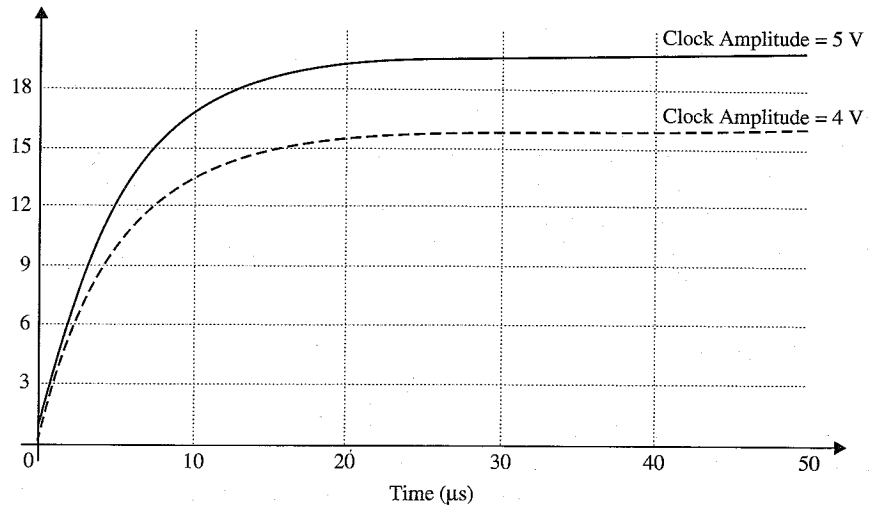


Figure 9-12 Pump output voltage over a time range for different clock amplitude.

simulations were run for a longer time, and as they can be interpreted from the figure, the voltages will go up by about 1 V, although this will take a long time.

Figure 9-13 shows the pump's internal waveform at the same time for the two different clock amplitudes. The top one corresponds to the clock amplitude of 4 V, and the bottom one corresponds to the clock amplitude

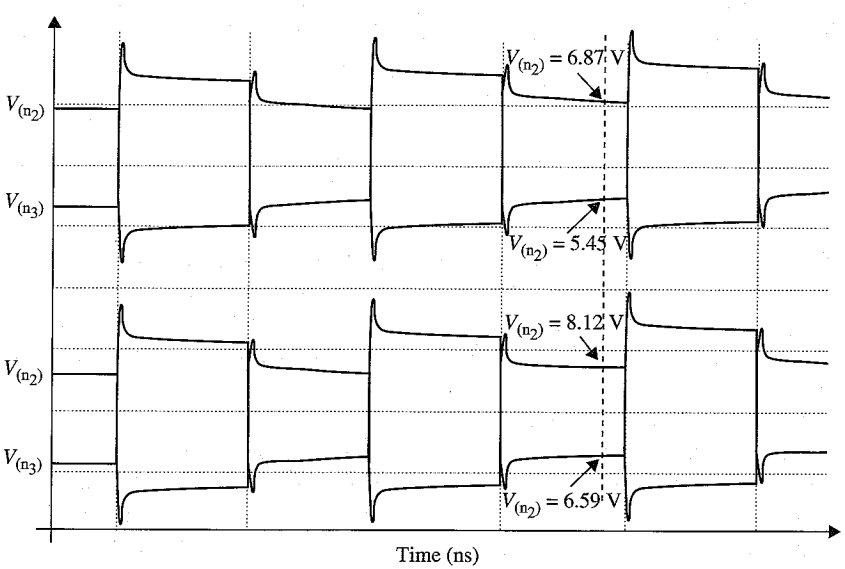


Figure 9-13 Pump internal waveforms vs. clock amplitude.

of 5 V. Increased clock amplitude will allow us to mitigate the effects of MOSFET V_t and hence allow more charge transfer, resulting in a faster output ramp up and higher output voltage. The purpose of using two different clock amplitudes differing by a volt is to show the effects of the clock amplitude change to the pump's output voltage and its internal waveforms. In real circuit implementation, the clock amplitudes will be determined by the supply voltage and a multiple of it, mostly in the case when we are using a clock amplitude doubler or tripler, as discussed in earlier chapters.

9.4 Pump Performance Characterization

In the previous section we designed a simple charge pump and saw the pump's operation and pump output response versus various design and external parameters, such as the pump output response versus time, number of stages, clock frequency, clock amplitude, and output load capacitance. Next in this section we will characterize the inherent capabilities of the charge pump, such as calculate the pump's efficiency, plot its I-V curves, output current versus frequency plots, and a few other parameters crucial for understanding the charge pump's capabilities.

9.4.1 Pump efficiency calculation

Next we will focus our attention on generating the pump's efficiency, I-V, and other essential characteristics, which are important when designing charge pumps for chips requiring a low power budget, such as those in handheld devices and other portable operation. We need to make sure that the pump is operated as close to its peak efficiency level as possible. The charge pump's efficiency can be calculated according to the formula

$$\text{Efficiency} = \frac{i_{\text{pump\_out}} V_{\text{out}}}{\left[\sum i_{\text{subcircuit}}\right] V_{\text{DD}}} \qquad (9\text{-}1)$$

where $i_{\text{subcircuit}}$ is the V_{DD} power consumption of each subcircuit component of the whole charge pump. It needs to be pointed out that the $i_{\text{subcircuit}}$ current is the average DC current over a clock period. For better results, more than one clock period may be considered to determine the average DC current.

In general, some of the major components of power consumption of the whole charge pump circuit have been categorized in Figure 9-14. As you can see, the circuits for generating the boosted clocks and the pump's regulator are the major components of power consumption. The circuits that are not directly related to the actual charge pump's operation need to be partitioned out from the pump efficiency calculation, such as the

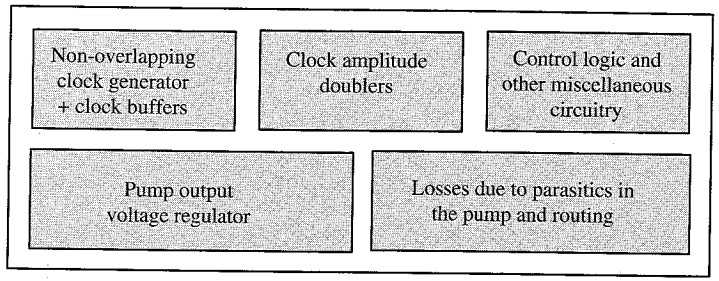


Figure 9-14 Charge pump components of power consumption.

pump clock oscillator, and a case can be made to leave out the control logic and miscellaneous components. Hence, the denominator of the preceding equation can be expressed as

$$\sum i_{\text{subcircuit}} = i_{\text{clock\_generator}} + i_{\text{clock\_buffer}} + i_{\text{control\_logic}} + i_{\text{regulator}} + i_{\text{misc}} \qquad (9\text{-}2)$$

One of the easiest ways to calculate the efficiency is to observe the following SPICE simulation setup shown in Figure 9-15. Assuming the charge pump is being simulated along with different other circuits on the chip, a 0 V DC source has been put between the real V_{DD} power supply and the V_{DD} nodes of the pump's subcircuits. This setup will allow us to monitor the average DC current, I_{in}, flowing from the power supply into the pump. The second voltage source, V_0, needs some clarification. We know that the charge pump delivers a particular output current for a corresponding output voltage. By connecting the output to a steady DC voltage source, the pump's output is held constant to determine the amount of current the pump can deliver at

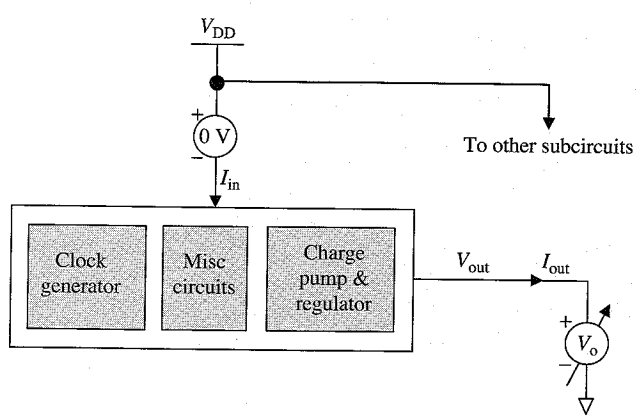


Figure 9-15 Simulation setup for calculating efficiency.

that output voltage per clock cycle. Next, the output voltage source's DC value may be varied to determine the pump output current for a new voltage setting. This operation may be done in steps to determine the pump's I-V characteristics. This will be shown and discussed later in the chapter.

Hence, with the simulation setup shown in Figure 9-15, the efficiency can easily be calculated as follows:

$$\text{Efficiency} = \frac{V_o \times I_{\text{out}}}{V_{\text{DD}} \times I_{\text{in}}} \tag{9-3}$$

Figure 9-16 shows the pump efficiency plot, over a range of output voltages. It is easy to observe that the peak efficiency occurs when the pump is delivering an output of about 15 V, which is about 75% of the peak output voltage of 20 V. This is one of the reasons why we targeted a pump delivering 20 V, when the requirement was a 15 V output, as assumed early in this chapter. Operating the pump close to its peak output voltage will mean less efficiency and further lower output drive current.

Following the original equation $V_{\text{out}} = NV_{\phi} + V_{\text{in}} - (N + 1)V_D$ and Table 9-1, it can be shown that if the number of stages, N, is increased from 6 to 7, the output voltage will jump up above 25 V. A new set of simulations was run with $N = 7$, and the new efficiency was calculated and plotted along with the efficiency plots for $N = 6$ in Figure 9-17. The two plots show an interesting trend. After we increase the number of stages, the peak efficiency figure is reduced, the curve goes flatter on

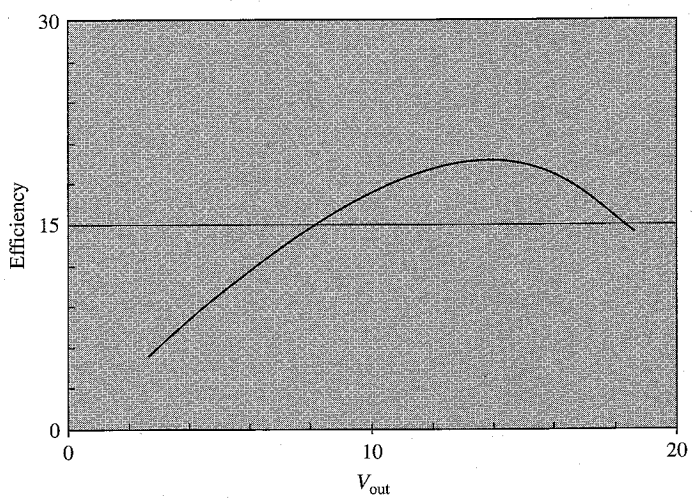


Figure 9-16 Charge pump output voltage vs. pump efficiency.

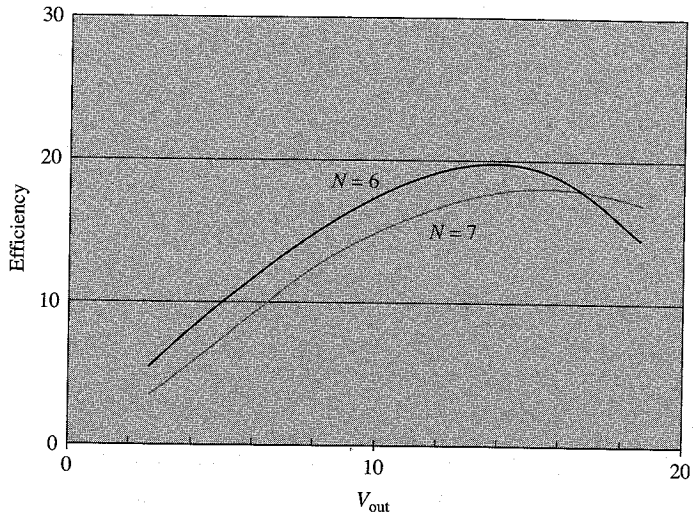


Figure 9-17 Charge pump output voltage vs. pump efficiency for a different number of stages.

the top and is shifted to the right, depicting a new peak efficiency at 16 V–17 V. In the same way, as the number of stages is increased, the curves will get lower in amplitude and gradually get flatter and shift right. Hence, from Figure 9-17, we can see that our initial assumption for using $N = 6$ is valid and produces the highest efficiency at the target output voltage level. In case the efficiency maxima lie on either side of 15 V, we need to change N or other circuit parameters to bring the peak pump efficiency closer to 15 V.

9.4.2 Pump I-V characteristics

The charge pump's I-V characteristics are plotted in Figure 9-18 for two cases, $N = 6$ and $N = 7$. The current output shows an almost linear response to the rising output voltage, confirming the simple voltage source model of the charge pump. The slope of the curve can be used to determine the pump's input impedance. It is interesting to note that even though the current output of the pump with seven stages ($N = 7$) is lower than the current output of the pump with six stages, the 7-stage charge pump will sustain a little higher current at high voltage than its 6-stage counterpart. Also, you can see that for the 6-stage pump, the available output current at around 20 V will be much less than that available at about 15 V. Because, in general, a charge pump is always expected to be able to deliver some appreciable output drive current, you should always design it such that the actual required pump output voltage is lower than the maximum pump output voltage.

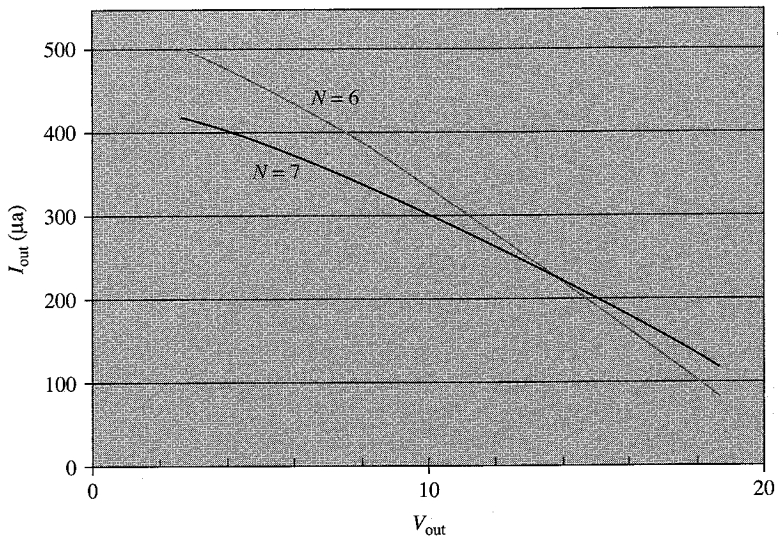


Figure 9-18 Charge pump I-V characteristics for a different number of stages.

9.4.3 Pump output current versus pump clock frequency

Next, the pump output current dependency versus pump clock frequency was determined at different output voltages, as shown in Figure 9-19. To generate this graph, the simulation setup shown in Figure 9-15 was used.

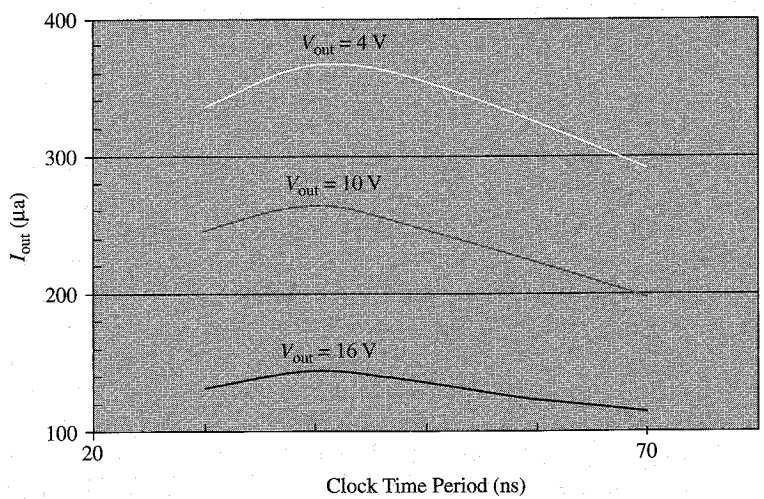


Figure 9-19 Charge pump I-V characteristics for a different number of stages.

By holding the output voltage constant at 4 V, 10 V, and 16 V, the pump oscillator frequency was varied over a range of 25 ns to 75 ns and the corresponding I_{out} was recorded. This graph also shows a resemblance to the graph in Figure 9-15—the pump produces the maximum current at a particular clock frequency, and it rolls off at frequencies higher and lower than the optimum one. As shown earlier, when the clock period is long, complete charge transfer takes place and then the pump is at a steady state for the rest of the period. Hence, by closely monitoring the internal nodes, this steady-state region can be removed completely (i.e., by reducing the clock time period). As shown earlier in Figures 9-10 and 9-11, the clock period can be reduced from 60 ns to close to 40 ns. A frequency sweep of the charge pump shows that the optimum operating frequency is indeed around 40 ns and falls off below this number. As explained earlier, when the time period is reduced below the optimum number, incomplete charge transfer takes place per clock cycle, and this reduces the effective output current. This type of simulation is extremely crucial in the circuit design phase to find the optimum pump operating frequency.

9.4.4 Pump output current versus MOSFET sizes

Figure 9-3, earlier in the chapter, shows the diode-connected MOSFET sizes as about 100 μ in width. This was simply an assumption, and we need to tune the MOSFET size to find the one that gives the best performance. One of the best ways to do this is to generate an I-V curve for a few different MOSFET sizes, both higher and lower than the current one, and then find the size that produces the optimum results. As was discussed in the preceding chapters, increasing the MOSFET size will reduce the channel resistance and hence allow for faster charge transfer, but this does have a limitation. As the MOSFET size is increased, the layout geometry also increases, as well as the source, drain parasitic capacitances, and other associated routing capacitance that will act to reduce the advantage as the pump frequency is increased or as the pump output gets higher. Figure 9-20 shows a similar case in which the channel width was varied from 80 μ to 120 μ, in 20 μ steps. As you can see from the figure, the output current supplied is higher at any particular voltage as long as it is in the lower range (i.e., below 12 V). But as the output voltage rises, the difference becomes very small. Parasitics play an important role as the voltage gets higher along with the rising V_t. Hence, in this scenario, an 80 μ or less device width would suffice if the layout is not extremely space limited, but again it would not hurt to keep the width at 100 μ.

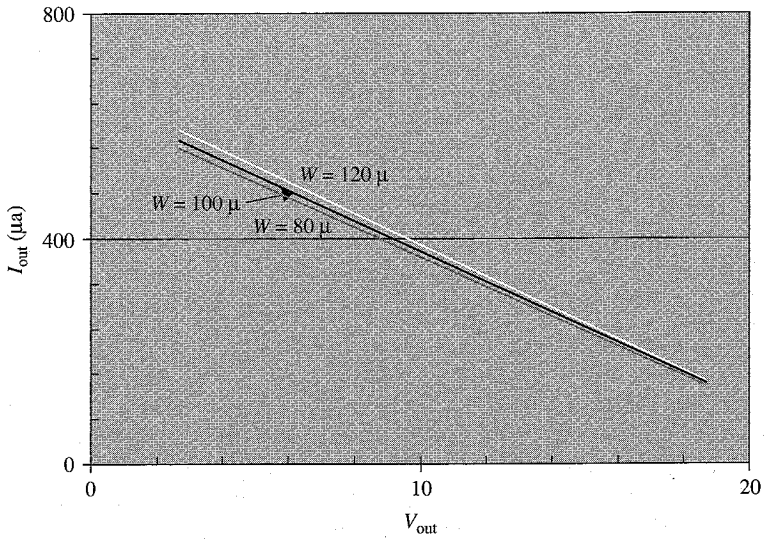


Figure 9-20 Charge pump I-V characteristics for different MOSFET widths.

9.4.5 Pump output current versus clock driver size

Last but not the least, we need to pay attention to the clock driver sizes. In general, buffer driving loads of 4 pF–5 pF need to be really huge, and they place a heavy toll on the circuit power consumption, consuming a huge layout area. However, we also need to keep the driver sizes as big as possible, because a slow rising/falling clock will not allow an effective charge transfer if the clock time period is not long enough. Continuing with the findings from the preceding pages, consider the case when the clock time period is 40 ns (i.e., we have about 20 ns for charge transfer including clock rise and fall time). If the clock rise and fall time is not sharp enough—say, it becomes 10 ns each side—then we have already shaved off about 10 ns from the 20 ns available for charge transfer, which will be grossly inadequate.

Figure 9-21 shows a simulation performed with different clock buffer sizings. This simulation was performed by including a multiplier factor in the clock driver sizes and then altering the multiplier parameter during the simulations to generate the I-V characteristics. As shown in Figure 9-21, as the driver size is increased from 0.8x to 1x to 1.2x, the pump delivers higher current, because it has more time for charge transfer and because it allows more charge to be transferred across the pump capacitor. But when the rise/fall times are sharp enough, there will be no real advantage in increasing the clock driver sizes, which will result in higher power dissipation. Judging from Figure 9-21, it can be

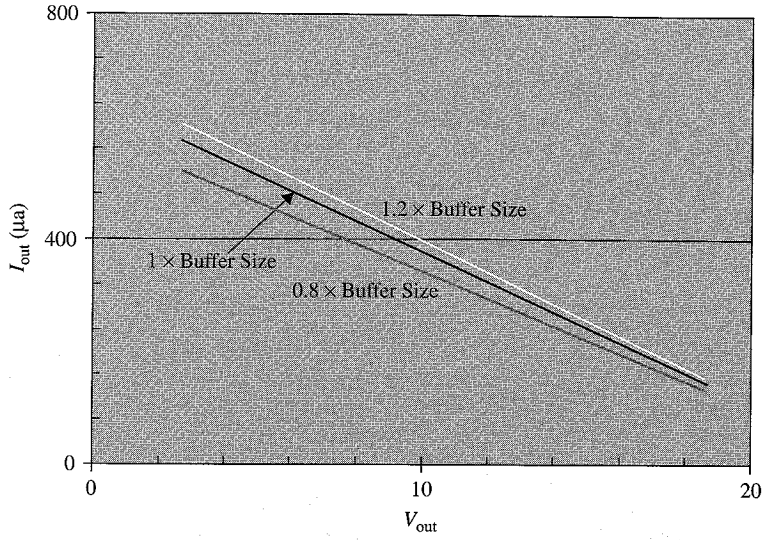


Figure 9-21 Charge pump I-V characteristics for different clock buffer sizes.

said that a 1.4 × buffer size will produce no real advantage in terms of drive current, and the 1 × buffer size to 1.2 × buffer size is appropriate in this case.

9.4.6 Design summary

Referring back to the original design specification in Table 9-1, we started our simulations by making a few calculated assumptions, and we performed many simulations to determine the pump's characteristics, test our assumptions, and tune the parameters for better performance. Table 9-3 attempts to summarize our analysis findings and substantiate the best design specifications for this particular case.

For the present design requirements, the parameters shown in Table 9-3 should be sufficient to take it to the next stage (i.e., layout implementation). Yet, the design is not complete unless the layout, after it is done, is carefully scrutinized and a parasitic extraction is performed. The circuit designer should have an idea of the estimated parasitics at all crucial nodes and should carefully analyze parasitics that are higher than the estimate. The layout may need to be modified and the whole process may need to be repeated again. In parallel, an RC back-annotated simulation should be performed under all possible worst-case conditions to make sure that all the design parameters are sufficient to meet performance under worst-case conditions. Generally, after back-annotated simulation, certain parameters such as pump capacitance (C),

TABLE 9.3 Design Summary for the 6-Stage Dickson Charge Pump

Parameter	Description	Final value
N	An initial value of $N = 6$ was chosen so as to produce a maximum V_{out} of about 20 V, such that the target V_{out} is about 75% of the maximum, when the input clock amplitude is 5 V. Simulations done with parasitics, clock frequency, and amplitude variations show that $N = 6$ produces optimum results. This was further validated with the pump efficiency simulation, which showed that at the desired output voltage, the pump is operating at its maximum efficiency.	$N = 6$
f	As shown in Figure 9-9, the initial frequency estimate of 20 MHz (i.e., 50 ns) was a good start. However, thorough analysis in Figures 9-9, 9-10, and 9-11 shows that the clock time period could be reduced to 40 ns, by which time a complete charge transfer can be guaranteed and the next cycle can be started. Hence, the new frequency of operation can be set at 25 MHz. However, before finalizing this value, a thorough simulation should be made at all different process, temperature, and voltage skews to ensure this 40 ns time period is sufficient to guarantee complete charge transfer.	$f = 25$ MHz
C	The initial estimate of 1.5 pF for the pump capacitor size was a good estimate because it allowed us to meet all the required design specifications. More analysis should be performed to see whether reducing the size still suffices for the pump's specifications.	$C = 1.5$ pF
W	The initial width of the MOSFET (100 μ) was a good estimate. As shown in Figure 9-20, using a MOSFET of higher width is no advantage. But if needed, the width can be reduced to 80 μ when layout size is at a premium.	$W = 100$

diode-connected MOSFET width, and frequency of operation need to be trimmed for optimum performance.[5]

9.5 Conclusion

We saw that the design of the charge pump starts by back-calculating the values of N, V_{clock}, and V_{in} from an assumed output voltage, which is actually higher than the required output voltage. The next stage is to effectively determine the pump parasitics, clock booster, pump capacitor size, diode-connected MOSFET size, and the clock oscillator frequency. Details of all these crucial parameters and their calculations have been explained in the previous chapters. Once a preliminary set of parameters is determined, we must run a preliminary set of simulations to observe the pump's output voltage ramp up, V_{DD} power consumption, and the amount of average current delivered at the output.

These numbers should give us a feel for the pump's performance and can effectively point to the deficiencies in any of the different parameter assumptions. A few iterations of the design modification may be made and simulations performed to make sure the basic design targets are met.

Next, most of these performance-tuning simulations must be performed to determine the optimum numbers, such as the best clock frequency, number of stages, clock booster sizes, pump diode and pump capacitor sizes, and a host of other parameters. In general, the simulation setup may be made somewhat automatic through scripting to allow easier re-simulation and re-extraction of parameters. Once most of the numbers are in a satisfactory range, layout should be completed and a re-simulation performed with back-annotated parasitics. Also, the pump performance must be checked against previously extracted data. In the case of large discrepancies, the circuit designer must take steps to modify and enhance the layout and repeat the back-annotated re-simulation again to get better results. Granted, most of the steps explained here may not be adequate or, in some cases, may be too time-consuming, but this depends on the circuit designer's own decision to create a rugged design, by testing all variables and making sure that the pump performs reliably, even under unseen extreme conditions, because in real applications such conditions do happen.

References

1. Choi, K.-H. et al. "Floating-well Charge Pump Circuits for Sub-2.0 V Single Power Supply Flash Memories," Digest of Technical Papers, Symposium on VLSI Circuits, pp. 61–62, June 12–14, 1997.
2. Papaix, C. and J.-M. Daga. "High Voltage Generation for Low Power Large V_{DD} Range Non Volatile Memories," PATMOS 2001. http://patmos2001.eivd.ch.
3. Witters, J.S., G. Groeseneken., and H.E. Maes. "Analysis and Modeling of On-Chip High Voltage Generator Circuits for Use in EEPROM Circuits." *IEEE Journal of Solid-State Circuits*, Vol. 24, No. 5, pp. 1327–1380, October 1989.
4. Tanzawa, T. and S. Atsumi. "Optimization of Word-Line Booster Circuits for Low-Voltage Flash Memories." *IEEE Journal of Solid-State Circuits*, Vol. 34, No. 8, August 1999.
5. Brugler, J.S. "Theoretical Performance of Voltage Multiplier Circuits." *IEEE Journal of Solid-State Circuits*, June 1971, pp. 132–135.

Index