Semiconductor-technology exploration : getting the most out of the MOST

DOI:
10.6100/IR556906

Document status and date:
Published: 01/01/2002

Document Version:
Publisher’s PDF, also known as Version of Record (includes final page, issue and volume numbers)

Semiconductor-Technology Exploration
getting the most out of the MOST

Harry Veendrick

Dit proefschrift is goedgekeurd door de promotoren:

prof.dr.ir. A.H.M. van Roermund
en
prof.dr.ir. R.H.J.M. Otten

Colophon:
Cover design: Hennie Alblas
Figures: Hennie Alblas
Type setting: Ivonne Hermens
Printed by: University Press, Eindhoven

Explanation of the cover:
The cover shows the limits of the Semiconductor Road, symbolically represented by the icebergs appearing on the horizon.

The work described in this thesis has been carried out at Philips Research Laboratories Eindhoven.

CIP-DATA LIBRARY TECHNISCHE UNIVERSITEIT EINDHOVEN

Veendrick, Harry

Semiconductor-Technology Exploration: getting the most out of the MOST
by Harry Veendrick. - Eindhoven : Technische Universiteit Eindhoven, 2002.
Proefschrift. - ISBN 90-74445-56-X
NUGI 832
Trefw.: CMOS circuits / performance / density / signal integrity / speed / power / noise / robustness / semiconductor technology scaling

Semiconductor-Technology Exploration
getting the most out of the MOST

Proefschrift

ter verkrijging van de graad van doctor aan de
Technische Universiteit Eindhoven, op gezag van de
Rector Magnificus, prof.dr. R.A. van Santen, voor een
commissie aangewezen door het College voor
Promoties in het openbaar te verdedigen
op vrijdag 28 juni 2002 om 16.00 uur

door

Harry Veendrick

geboren te Hummelo en Keppel

Preface

This thesis describes part of the research that has been done at Philips Research Laboratories in Eindhoven, The Netherlands, during a period of 25 years from 1977 to 2002.

This research was particularly focussed on three different aspects of circuit design: optimising performance (speed, power), maximising density (area) and improving robustness (signal integrity and noise). Although the selected subjects for this thesis share the same motivation: *getting the most out of the MOS transistor (MOST)*, they are very diverse regarding the fundamentally different requirements dictated by the various application areas. The level of detail in the underlying scientific publications is quite different, mainly because of the available publication space, e.g. as a short paper in a conference digest or as a long and detailed article in a journal. Therefore this thesis is divided into two parts.

Part I presents an anthology of the work, in which the most important research topics and results of each subject are discussed at an equal level of detail.

Part II, on the contrary, includes the related detailed scientific papers as they were published in the conference digests, magazines or journals.

I leave it to the readers' interest in the specific subject, which level of detail he prefers.

This work could not have been done without the inspiring environments of both the Philips Research Labs and the large number of high-talented colleagues. I want to thank all of them, and particularly those with whom I have closely worked together in different IC design and research projects and whose co-operation has resulted in co-authorship of several of the papers and patents during this 25 years period of time.

I want to acknowledge Philips Research management for allowing me a great deal of freedom in selecting relevant research subjects and for the opportunity to publish them all. I also like to thank them for supporting this work.

Finally, I want to thank my family for their understanding and allowing me to spend again many private hours on professionally related activities, such as the writing of this thesis.

Eindhoven, June 2002
Harry Veendrick

Contents

Part II Detailed scientific work

Introduction

Background and motivation

The complexity of a VLSI chip has increased from just a few components in the early sixties, to several tens of millions of devices today. It was Moore, already in 1964, who predicted this enormous increase. There are several factors that drove this complexity to the level that has been reached today. First of all, the rapid scaling of the minimum feature sizes allowed us to almost double the number of devices on the same silicon area every new technology generation. Particularly in the first three decades, also the speed of the circuits doubled. As a result, we could have about four times more functionality (or computing power) on the same die area, every 18 to 24 months. However, the complexity of an IC not only depends on what is offered by the technology. A large part of it is also dominated by the requirements of the application area for which they are designed. Important parameters in this respect are:

1- functionality/features (density)
2- performance (speed and power)
3- product volume (density)
4- system size (density)
5- time-to-market (turn-around time)
6- robustness (signal integrity and reliability)

The order of priority in these parameters depends on the application area.
In personal computers, for example, speed is the most important requirement because it determines how fast they can run the complex software programs. A state-of-the-art PC system is built around very high-speed microprocessors, such as the Pentium 4 (Intel), the PowerPC (IBM) or the Thunderbird (AMD) processor.
Consumer electronic applications like audio and video, however, are less driven by speed and more by an increased functionality (features: stereo sound, Dolby surround, teletext, noise reduction, 100Hz, wide and dual screen, picture-in-picture, and MPEG applications such as set-top boxes, DVD, etc.
In portable electronics, both the advances in semiconductor technology and design have been used to reduce the physical sizes and simultaneously increase the system features while keeping the power budget constant. Mobile phones show the strongest improvements in this category of electronic products. They may include lots of different features: calculator, radio, games, MP-3 player, remote control, GPS, Internet access, email, digital camera, etc.

Because of the high innovation rate, the lifetime of a new generation of most of these products is only six to twelve months. Time-to-market has become one of the most critical requirements for many of today’s products.
Finally, the increase in complexity of current ICs has led to a rapid increase of signal activity on a chip: more signals switch at the same time and at a higher rate. Moreover, they propagate at smaller mutual distances. The resulting physical effects, such as supply noise, voltage drop, cross-talk and electromigration have a negative influence on circuit behaviour and require special design solutions. The continuous drive to improve the performance of integrated circuits, while reducing the physical dimensions of both the transistors and the interconnect, causes an increased manifestation of these so-called deep-sub-micron effects. This will put a burden on maintaining circuit performance and robustness and will have severe consequences for the design of complex VLSI chips.

This work is the result of this continuous drive for more speed, less power and/or an increased density, to get the most out of the MOST (MOS transistor). At the same time it is tried to keep the robustness of circuit operation at a sufficiently high level.
Although the selected subjects for this thesis share the same motivation, they are very diverse regarding the fundamentally different requirements dictated by the various application areas.
Several research topics are more or less dated, however, the basic solutions or implementations are still valid and used today and the results have been placed in a today’s perspective.

Outline of the thesis

This thesis describes the research that has been done regarding three important aspects of IC design: performance (speed, power), density (area) and robustness (signal integrity and noise). Although the research covered different levels of IC design, the focus here will be particularly on circuits and circuit design techniques. Also examples of ICs will be given in which the particular techniques have been applied. Most of the underlying work has been published, e.g. as a short paper in a conference digest or as a long and detailed article in a journal. The level of detail of these publications is quite different. Therefore this thesis is divided into two parts.
Part I highlights the main research topics and results with an equal level of detail. In fact, it presents an anthology of the work. The different subjects are presented in an order, which almost reflects the order in which the research was performed.

Part II contains the detailed papers as they were published in the scientific magazines and conference digests. The order in which these publications are presented here, is synchronised with Part I.
Therefore, the outline is as follows:

Part I contains four chapters.
Chapter 1 describes the work that has been done with respect to performance optimisation. The first circuits presented here, were designed in nMOS technology and used in digital video applications. A special circuit technique was developed to achieve high-speed operation in a digital chrominance filter and is also used in the design of a digital potentiometer. The second topic in this chapter discusses the design of a programmable high-speed video signal processor. In this design, various circuit techniques are applied to accommodate the high-speed requirements and support the signal integrity. The discussions, however, are mainly focussed on the high-bandwidth switch matrix that enables the high-speed communication between the different video processing elements on that chip.
Power consumption is an important parameter that also reflects the performance of a chip. There are several different factors that contribute to the total power consumption of a CMOS chip. One of them is called short-circuit dissipation. The third topic discusses the cause of this short-circuit power dissipation. It also presents an expression to estimate this power and ways how to limit it.
Chapter 2 is focussed on the optimisation of the density of different types of integrated circuits. Due to coding of the large number of pixels in a TV screen, the data storage of one complete TV picture requires relatively large memories. The first topic in this chapter is the design of two digital video memories, implemented in charge-coupled device (CCD) MOS technologies. The operation of a CCD perfectly matches the serial character of video samples transmission.
Since the product life cycle for many application areas is reduced, time-to-market has become an important aspect for a lot of different ASIC products. Fast turn-around time in the waferfab is one way to support a short time-to-market. This can be achieved by using prefab wafers (off-the-shelf available), which already contain the processed transistors. A few remaining contact and metal interconnect masks can then ‘simply’ complete its functionality. In this way the turn-around time in the waferfab is reduced from an average 10 to 13 weeks for a custom IC, to about two to four weeks for these so-called gate-arrays. Because these gate-arrays are less customised, they show a lower gate density than other IC implementations like standard-cell or bit-slice layout. Therefore, as a second topic in this chapter, an efficient and flexible gate-array architecture is described, which supports a relatively high-density realisation of both logic and memory circuits.
Due to the continuous process of technology scaling, the feature sizes of the transistors and their interconnections reduce while their numbers increase. At the same

time the clock frequencies increase. This poses a burden on the robustness of circuit operation: their reliability and signal integrity.

Chapter 3 therefore discusses several of these aspects of IC design. The discussion starts with the design of a synchroniser, which is based on the basic operation of a flipflop. Due to the sampling of asynchronous signals, a flipflop may reach a meta-stable state, in which its output levels are logically undefined, causing a glitch. The research was focused to optimise the reliability of synchronisers such that the chance of occurrence of a glitch was minimised. This could not be done without a detailed study of the influence of noise on the meta-stable state behaviour.

The continuous process of scaling also causes an increase of the supply currents every new technology. As a result, the current density in the on-chip supply lines can easily reach the maximum allowed levels defined by electromigration requirements. These requirements are strongly related with the temperature of these supply lines. The power consumed in these supply lines will increase their temperature. A quantitative discussion on this so-called wire self-heating is presented in the second subject in this chapter.

The third topic of this chapter is related to signal integrity. More devices are produced at smaller distances and switch at higher frequencies, causing much more on-chip noise and interference. Particularly the cross-talk between neighbouring signal lines and the noise on the supply lines may have a dramatic impact on the performance and signal integrity of deep-sub-micron ICs. For compensation of the huge peak currents that occur during heavy simultaneous switching and which are the cause of the supply noise, de-coupling capacitors are implemented on-chip.

Finally, *chapter 4* discusses the effects and challenges of scaling with respect to the three previous chapters: performance, density and robustness of future ICs.

Part II contains the detailed scientific papers, which are presented in three different chapters. The subjects and numbering of these chapters and their sections correspond with those of Part I.

Part I is directly following by its reference list. The detailed papers in Part II all have their own reference lists.

The overall conclusions have been placed at the end of the thesis, directly after the final detailed paper.

The thesis ends with a list of patents and recent publications of the author.

Part I

Anthology of selected work

Chapter 1

Design for performance improvement

The consumer market knows a large variety of different integrated circuits applications. Some of the applications are performance (power, speed) driven, while others are features driven: more functionality. The subjects in this chapter are focussed on performance and particularly on performance improvement. The first topic that will be discussed here is related to circuit design optimisation for high-speed video applications.

High performance does not only mean high speed. It may as well reflect the efficiency of a circuit in terms of power usage. So, low power is an equally important performance indicator as high speed. The second topic is therefore a power related subject that describes how the short-circuit component in the total power consumption can be reduced to negligible values.

1.1 High-speed circuit design

1.1.1 Introduction

There are many ways in which the speed of integrated circuits can be improved. This can be done at architecture level, at logic implementation level, at circuit level and at device (technology) level. This section first discusses an example of achieving high speed circuits by the development of a new style of nMOS logic gates, which could approach the speed of CMOS circuits at that time. The second example, the design of a high-bandwidth communication bus, shows that both the architectural and circuit levels are used to achieve high speed.

1.1.2 High-speed logic gates: race-compensated MOS logic

Before the move to CMOS in the early eighties, most digital MOS circuits were made in nMOS technology [Mavor, 1983]. Initially nMOS technology only contained one type of transistor: the enhancement transistor. Fig. 1 shows an (E/E) inverter that is built up with an enhancement (E) driver and an enhancement (E) load transistor (positive threshold voltage (V_t > 0V).

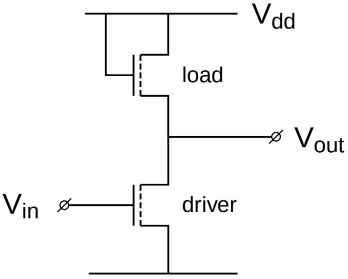


Fig. 1 Inverter in Enhancement/Enhancement (E/E) nMOS logic

A major disadvantage of this type of E/E logic is that, when the output is rising, the gate-source voltage of the load is reducing with as final result that the output level can not get higher than a voltage level equal to $V_{dd} - V_t$, in which V_{dd} represents the supply voltage. In a 5V technology this threshold voltage loss can be as high as 1.5V, such that the output high level of an E/E logic gate may only reach 3.5V. This reduced high level is received by the connected logic gates, which on their turn become more than a factor of three slower than when they would receive a high level as high as V_{dd} (so without the threshold voltage loss).
Arithmetic type of circuits, such as adders, multipliers and filters usually use half and full-adder type of cells to perform their functionality. Fig. 2 shows the sum function of an E/E full-adder cell. Including the generation of the inverse inputs, this cell consists of 21 transistors. Normally, when a logic function uses the direct inputs and their inverse ones, it takes two gate delays to perform this function. Using race-compensated MOS logic allows both creating and using the inverse data in one single logic gate. It results in a hardware reduction (smaller chip size) as well as in a relatively large increase in circuit speed. Fig. 3 shows an alternative realisation for this cell. This race-compensated MOS logic cell contains only 18 transistors and is much faster (about a factor of two).

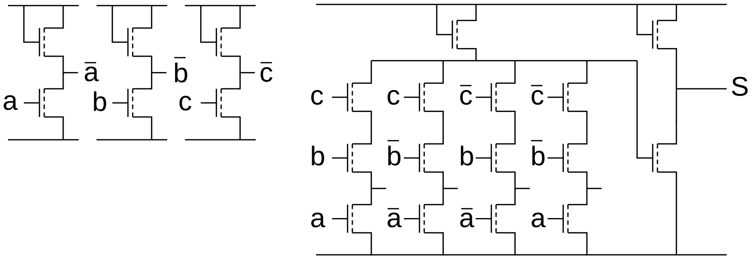


Fig. 2 Full adder Sum function in E/E nMOS logic

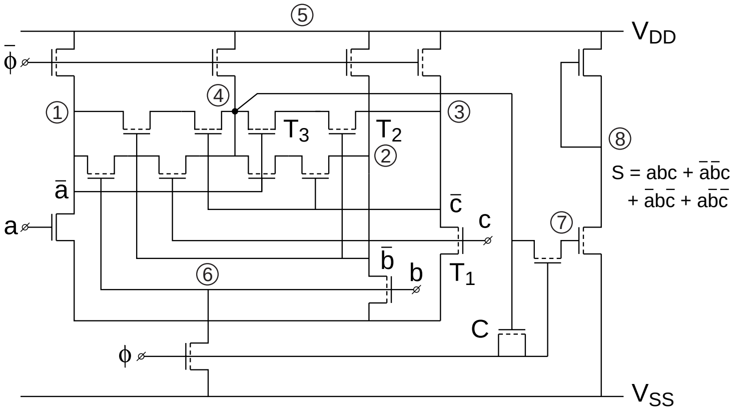


Fig. 3 The same Sum function in race-compensated MOS logic

The operation of this cell is as follows. When the clock ϕ is low, nodes 1, 2, 3 and 4 are pre-charged high. During ϕ, the data at the input nodes *a*, *b* and *c* are sampled and at the same moment the inverse data a_{not}, b_{not} and c_{not} are generated and used to control other gates. Under certain input conditions this introduces a race, by causing a voltage drop at node 4, right after a rising edge of clock ϕ, at a moment that it is not allowed. For instance, when the inputs *a* and *c* are at high level and input *b* is low, then transistors T_1 and T_2 must conduct, while T_3 should not. However, when the sample clock ϕ is switched on, it will take some time before node a_{not} has been discharged, meaning that transistor T_3 is conducting during a very short period of time, which causes a temporary conducting path from node 4 through transistors T_3, T_2 and T_1 to ground. This results in a voltage drop at node 4. The bootstrap capacitor *C* is charged during the pre-charge period (clock ϕ is low) and right after the clock switches to high, its charge is used to compensate any voltage drop at node 4 caused by either charge sharing or by a race.

This circuit design technique has been used in two different video signal processing ICs made in nMOS technology to increase the circuit speed by more than a factor of two: a digital low-pass filter for the separation of the luminance and chrominance (see Part II, Chapter 1.1A) and a digital potentiometer for high-speed video applications (see Part II, Chapter 1.1B and US patent US04513388).

Conclusions

In nMOS circuits that consist of enhancement transistors only, the output high level of a logic gate suffers from a threshold-voltage loss, causing a reduced high level at the inputs of the connected gates. These gates therefore show a relatively slow switching operation. A new dynamic logic-gate concept has been introduced, which allows generating both the logic function and its inverting inputs in one single gate. Without any further design measure, this would lead to races causing additional voltage drops at the gate's output. By using a bootstrap capacitor, charge can be pumped from the clock node to the logic gates' output nodes to compensate for both types of voltage drop. The circuits that were built with this type of logic show a 15% higher density and a two times improvement in speed.

In today's perspective

Since we have CMOS technologies today, it is not needed to use charge-pumping (bootstrap) techniques to speed up the basic logic circuits. The subject described here shows that there are on-chip electronic solutions to improve the performance even more then what is intrinsically available from a technology. Today, bootstrap concepts are still used, e.g. in such non-volatile memories as flashes and E(E)PROMs, to generate the higher voltages needed during the programming and/ or erasure mode.

1.1.3 A switch-matrix for high-bandwidth communication

One of the most important characteristics of real-time video signal processing is its relatively high sample rate. This may range from 3MHz for a single chrominance signal in a conventional TV receiver, up to 108MHz or higher, for advanced high-definition TV signals. Accordingly, the required signal-processing power and communication bandwidth can be very high [Roizen, 1986; Chen, 1992]. For this purpose, a programmable general-purpose video signal processor (VSP) has been developed. The amount of parallelism, combined with the speed of operation resulted in a total processing power of 1.5GIPS (Giga Instructions per Second). Fig. 4 shows the modular architecture of this VSP.

The chip contains 28 pipelined processing elements (PE): twelve Arithmetic Logic Elements (ALE), four Memory Elements (ME), six Buffer Elements (BE) and six Output Elements (OE). Every input sample of a PE can temporarily be stored in a small buffer memory (silo). The number of each category of PEs is chosen to cover most of the intended real-time video processing algorithms. Such an algorithm can be regarded as a combination of several operations (add, subtract, multiply, store, read, etc.). According to Fig. 4, there are 28 PEs, with a total of 60 PE inputs. Each operation must be assigned to one of the PEs in the proper time slot, such that these PEs are optimally used. As a result of this requirement, each PE input must be able to select either an output of any of the PEs, or of any of the six external inputs. In fact, the efficiency of such an architecture is determined by the flexibility and throughput of its communication structure. Therefore, a switch matrix is used to perform the complete communication interface between the blocks on the chip as well as to the chip inputs and outputs.

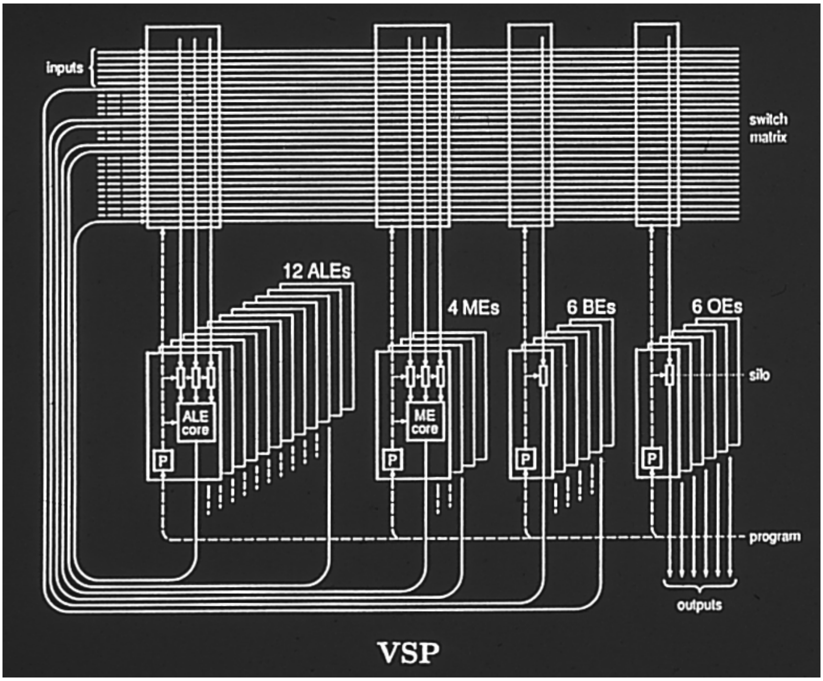


Fig. 4 Architecture of the VSP chip

Fig. 5 shows the schematic of the switch-matrix.

Since the word width is twelve bits, the switch matrix consists of a 28 × 12 bits = 336 bits wide bus, which crosses the complete chip, from left to right (see Part II, Chapter 1.1C, Fig. 6). The biggest challenge in the design of this switch-matrix was its complexity (28 inputs and 60 outputs, each of 12 bits) and its minimum required switching speed of 54MHz. Every clock cycle it must be able to connect each PE input to any of the 28 12-bits busses of the switch-matrix. This also allows one input to be connected to multiple outputs. The connections are done by pass-gates, which are located on every crossing of a PE input line with a switch-matrix bit. These pass-gates are selected by a decoder, which is located in the switch-matrix, below the metal busses. Since each PE input requires its own decoder, 60 of these decoders had to fit in this switch-matrix. Each decoder is controlled by the program memory P of the related PE (Fig. 4). So, within one clock-cycle it is required to read out the program memory P, decode its data to select the switches in the switch-matrix, and then store the selected data into the PEs input buffers (silo in Fig. 4). The combination of the necessary speed and density required a full-custom design of this switch-matrix. Fig. 6 shows the layout of a part of the switch-matrix.

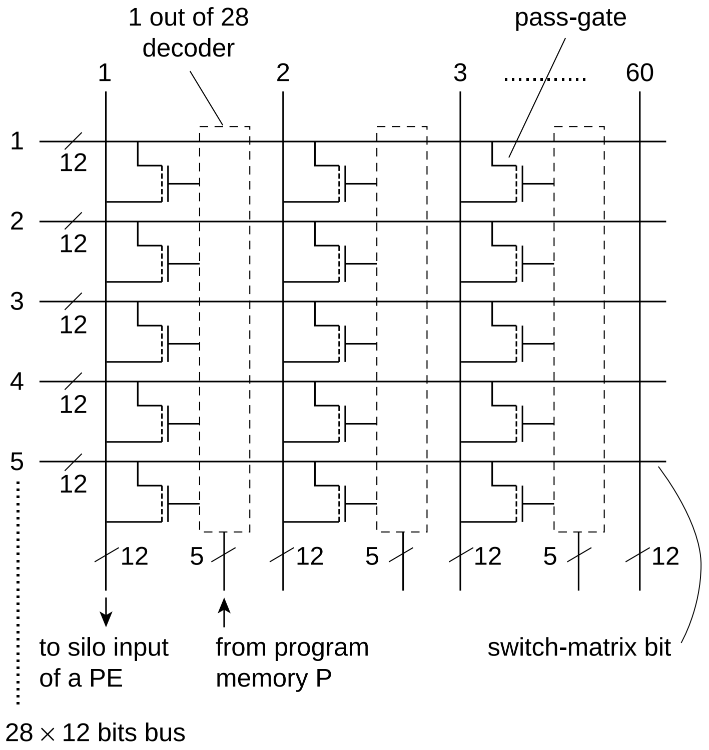


Fig. 5 Schematic diagram of the switch-matrix

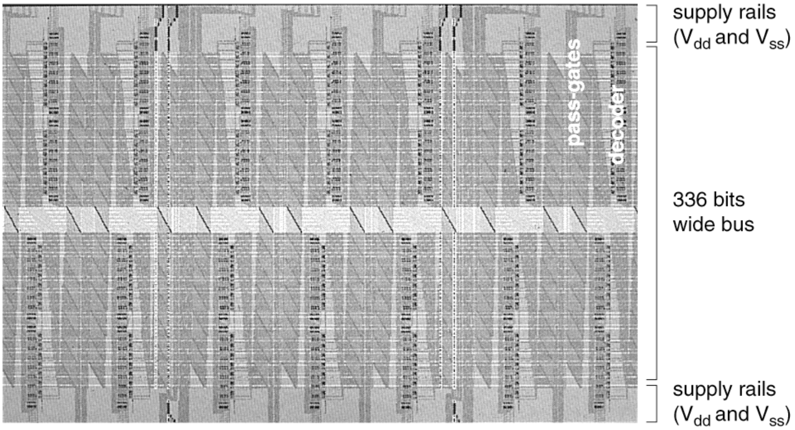


Fig. 6 Layout of a part of the switch-matrix

In the worst-case switching situation, all switch-matrix busses could switch simultaneously in the same direction. This would have caused huge peak currents and related supply noise across the supply network. For this reason the switch-matrix has been encapsulated within a large supply rail network, which, in turn, is connected to many supply pins in the chip periphery. Next to this, 20 nF of decoupling capacitance has been implemented on the chip to reduce the supply noise. This capacitance is charged during steady-state and its charge is used during peak activity.

By means of extensive area optimisation, this custom-designed switch-matrix is very compact and offers a flexible and programmable high-speed communication structure between the PEs. As discussed before, this switch-matrix has been used in the design of a high-speed video-signal processor (Part II, Chapter 1.1C).

Conclusions

General-purpose video-signal processors require many different processing elements operating at a relatively high frequency. In many cases the intermediate or final results of these operations need to be stored as well. This requires a very flexible and high-speed interface. The developed switch-matrix with 28 inputs and 60 outputs, each of 12 bits, offers a minimum bandwidth of 18 Gbit/s. It allowed the developed video signal processor to run even at 100MHz on some dies.

In today’s perspective

Because this switch-matrix architecture allows each input of a processing element to be connected to any of the others' or its own output, it is still used today as one of the most flexible communication interfaces in modern complex signal processors.

1.2 Low-power circuit design

1.2.1 Introduction

As long as the existence of the integrated circuit, its power consumption, limitation and reduction have been a major subject for research. During the seventies and early eighties, the most dominant MOS technology was the enhancement/depletion nMOS technology [Veendrick, 2000]. Through the eighties, however, due to the increase in both IC complexity and speed (clock frequency), the power consumption of an average nMOS ASIC chip reached 1W, which is the maximum power consumption of a cheap plastic package. This was one of the main driving forces for moving from nMOS to CMOS technologies in the first place. Now, after about two decades, the average CMOS ASIC chip has reached this 1W power limitation again, however with the difference that we don't have an alternative technology this time. Next to this average ASIC category of ICs, also the two other categories, namely the ICs in hand-held devices (battery-operated products like cordless and cellular phones, PDAs, palmtops, etc.) and those in high-speed microprocessors (Pentium (Intel), PowerPC (IBM), Thunderbird (AMD)) face a strong pressure on power limitation/reduction. Therefore, power limitation (reduction) has become one of the most important requirements for IC design in this new millennium.

To describe the work that has been done with respect to the power reduction of CMOS circuits, it is good to present an overview of the different sources that contribute to the total power consumption of a CMOS circuit.

Consider in this respect the two CMOS inverters presented in Fig. 7.

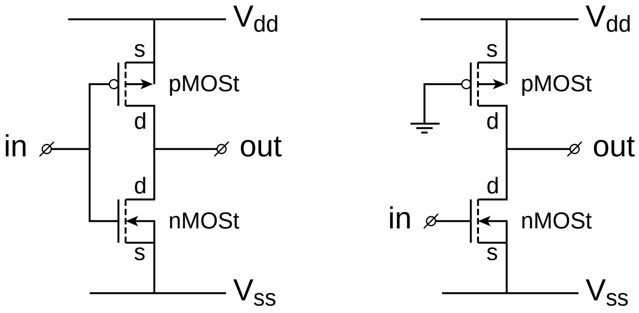


Fig. 7 (a) Basic CMOS inverter; (b) Pseudo nMOS inverter

During the operation of CMOS circuits their total power consumption consists of four different components:

$$P_{total} = P_{dyn} + P_{stat} + P_{short} + P_{leak} \tag{1}$$

P_{dyn} is the power consumed during charging and discharging (switching) of the output (see Fig. 7):

$$P_{dyn} = C \cdot V^2 \cdot a \cdot f \tag{2}$$

where C is the total capacitance at the output node (load + parasitic capacitance), V is the voltage swing, f is the clock frequency. The activity a factor of a logic gate represents the number of its switching transients per clock period, which can vary from below 0.1 (low activity) to 1 (high activity). Sometimes a gate may even switch more than once in a clock period, due to the occurrence of glitches, causing the activity factor to be higher than 1.

P_{stat} represents the static dissipation, which is the power consumed as a result of static current. This current can only flow when a circuit, in steady-state, has a DC current path from supply to ground, when its output is low. Although the circuit in Fig. 7b is a CMOS circuit, it is called *pseudo nMOS* because it operates similar to an nMOS gate (see Fig. 1). In such a gate the logic function is implemented in the nMOS transistors only, while all pMOS transistors are replaced by only one pMOST with its gate connected to the V_{ss}. The static power of such a logic gate is expressed as:

$$P_{stat} = I_{average} \cdot V \tag{3}$$

where $I_{average}$ represents the average DC current. Due to this DC current, pseudo nMOS logic gates consume 10 to 20 times as much as their *full CMOS* counterparts. Particularly in low-power applications, this type of logic is not used, thereby eliminating the static power component.

P_{leak} is the power dissipated as a result of substrate leakage, sub-threshold leakage and gate leakage currents. In current technologies the sub-threshold leakage is by far the largest contribution to this power component. Due to scaling of the technology, also the supply voltage is reduced. Because of the speed requirement of a new technology, also the threshold voltage is reduced. However, a reduction of 100 mV in the threshold voltage V_t leads to an increase of the leakage current (at V_{gs} = 0V) of a factor between 10 to 16 [Veendrick, 2000]. So, on the one hand, the high-speed requirement demands a low V_t while, on the other hand, the low-power requirement demands a high V_t. Because of these contradictory requirements, most advanced CMOS technologies offer both a low and high V_t.

P_{short} is the power consumption in an inverter (or in a logic gate), whenever transients on the inputs cause a temporary current to flow directly from supply to ground. Let us assume that the input of the CMOS inverter without a load (Fig. 8a) is at low level and its output is at high level. In this case the pMOS transistor is on, while the nMOST is off.

Next, let us assume that the input switches slowly to high level. When its level passes the V_{t_n} of the nMOST, this transistor switches on, while the pMOST is still on. This causes a short-circuit current from supply to ground. This current flows as long as the input voltage is higher than V_{t_n} above V_{ss} and more than $|V_{t_p}|$ below V_{dd}. It creates a temporary short between V_{dd} and V_{ss} and is responsible for the short-circuit power component.

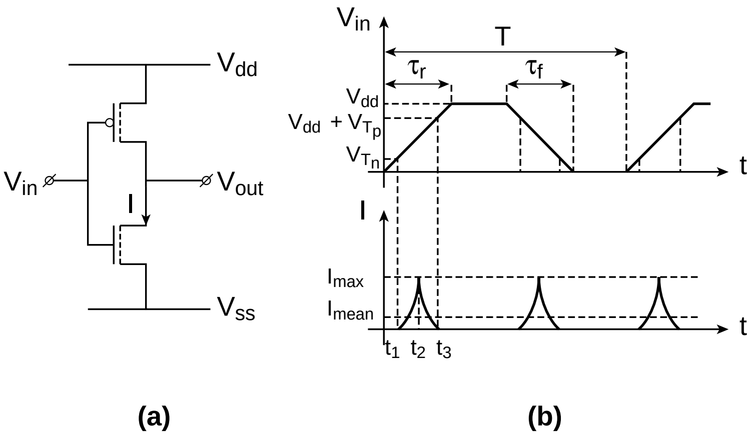


Fig. 8 (a) CMOS inverter without a load
(b) Current behaviour of an inverter without load

The following paragraph focuses on circuit design techniques to reduce this short-circuit component.

1.2.2 Design concepts to reduce the short-circuit power dissipation

Fig. 8b shows the input voltage waveform, when it switches from low to high level and back, and the corresponding short-circuit current.
The short-circuit dissipation can then be described as the product of the average current I_{mean} and the voltage V:

$$P_{short} = I_{mean} \cdot V \tag{4}$$

For simplicity we assume that the inverter has a symmetrical behaviour, which means that:

$$\beta_n = \beta_p = \beta \quad \text{and} \quad V_{t_n} = -V_{t_p} = V_t \tag{5}$$

where β_n and β_p are the gain factors of the nMOST and pMOST, respectively. We also assume that the rise and fall times of the input signal, τ_r and τ_f, respectively, are equal:

$$\tau_r = \tau_f = \tau \tag{6}$$

In Part II, Chapter 1.2 it is derived that for this inverter the short-circuit dissipation equals

$$P_{short} = \frac{\beta}{12} \cdot \left(V_{dd} - 2V_t\right)^3 \cdot \frac{\tau}{T} \qquad (7)$$

where T represents a full period of the input signal.
From this expression we can see that the short-circuit dissipation not only depends on the switching frequency ($f = 1/T$) and rise and fall times (τ) of the input signal, but also on the technology (V_t and β) and on the design of the inverter (β).
From expression (7) it is clear that the short-circuit power consumption is largest in circuits that contain transistors with large βs. There is a linear relation between β and the width W and length L of the transistor channel:

$$\beta = \beta_{\square} \cdot \frac{W}{L} \qquad (8)$$

where $\beta_{\square}$ is the gain factor (β) of a square transistor (with $W = L$).
In other words, circuits that contain large W/L ratios will generate the largest short-circuit dissipation. In digital CMOS ICs, the on-chip driver (buffer) circuits, such as bus drivers, clock buffers and output buffers may contain transistors that have W/L ratios between 20 and 500 to drive large load capacitances. On the other hand, the W/L ratios used in a typical logic gate usually varies from 1 to 10.

CMOS buffer design

Suppose the signal on a bus line (or bonding pad) with capacitance C_N must follow a signal at the output node A of a logic gate, which is capable of (dis)charging a capacitance C_0 in τ ns. An inverter chain such as illustrated in Fig. 9 can be used as a buffer circuit between node A and the bus line (or bonding pad).

From formula (7), it is clear that the rise and fall times on each input of the inverters in the above chain should be short. Moreover, it has been shown in Part II, Chapter 1.2 that minimum dissipation can be achieved when the rise and fall times on each of these inputs are equal to the rise and fall times (τ) at the buffer output. The inverter chain must therefore be designed such, that the rise and fall times on the inputs of each of its inverters are also equal to τ. According to literature [Mead, 1980], a minimum propagation delay time across the buffer is obtained when the ‘tapering factor’ r between the β’s of successive inverters is e, the base of natural logarithm. In terms of dissipation and silicon area, however, this tapering factor will not lead to an optimum design. Design for minimum dissipation and silicon area requires a different approach.

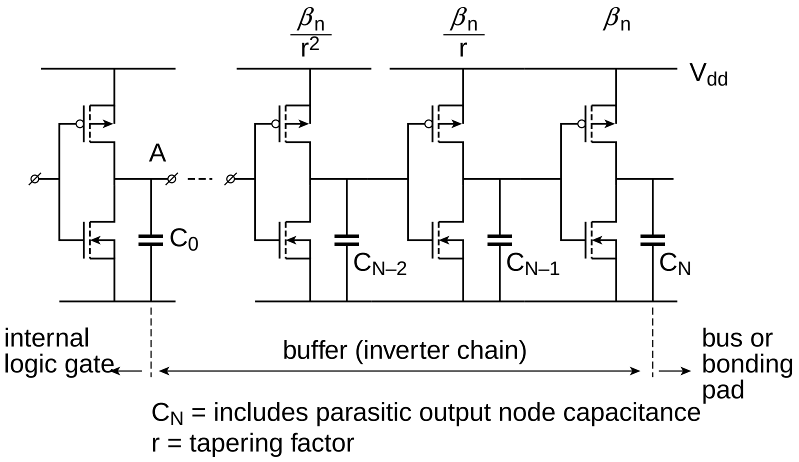


Fig. 9 A buffer circuit comprising an inverter chain

Example

A signal is produced by a logic gate and must be buffered to drive a capacitive load C_L = 10pF with a rise and fall time τ equal to 1ns. The channel length of both the pMOS and nMOS transistor is 0.25μm, while $\beta_{\square n}$ = 240μA/V^2 and $\beta_{\square p}$ = 60μA/V^2. To determine the right tapering factor for minimum area and power consumption, three different inverter chains are examined (Fig. 10).

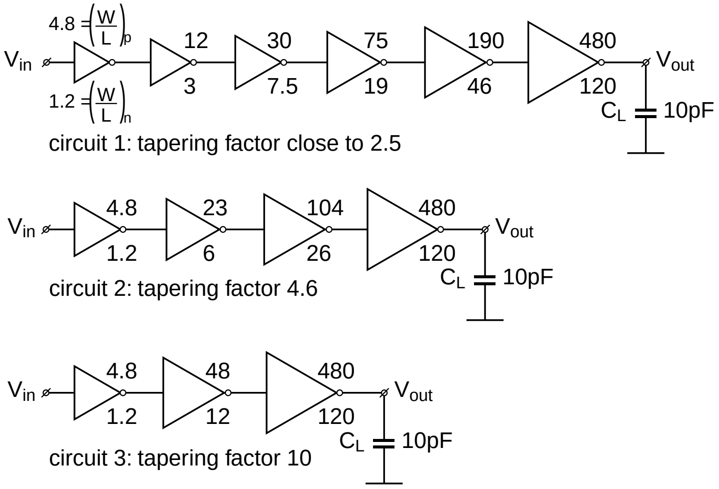

$4.8 = \left(\frac{W}{L}\right)_p$
$1.2 = \left(\frac{W}{L}\right)_n$
V_{in}
12
3
30
7.5
75
19
190
46
480
120
V_{out}
C_L
10pF
circuit 1: tapering factor close to 2.5
V_{in}
4.8
1.2
23
6
104
26
480
120
V_{out}
C_L
10pF
circuit 2: tapering factor 4.6
V_{in}
4.8
1.2
48
12
480
120
V_{out}
C_L
10pF
circuit 3: tapering factor 10

Fig. 10 Inverter chains with different tapering factors, all driving the same load with almost equal rise times

The characteristics of these inverter chains can be expressed with the variables: power dissipation, propagation delay, maximum current change d*I*/d*t* and area. The importance of a low value for d*I*/d*t* will be explained in section 4.2. Fig. 11 shows the simulation results of these inverter chains.

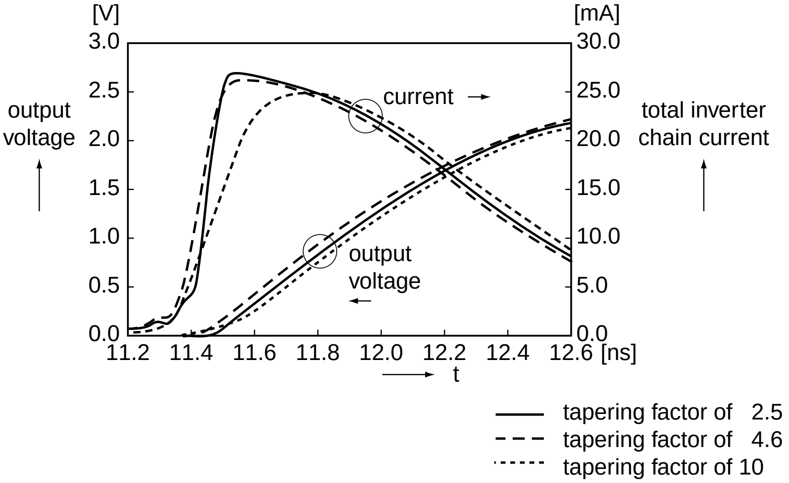


Fig. 11 Simulation results for inverter chains of Fig. 10

The input signal, V_{in}, to the three different inverter chains, is connected to identical first inverters with the same effective *W*/*L* ratio which mimic an equivalent logic. The diagram shows the total inverter chain currents and the output signals. Detailed overall results for these circuits are given in Table 1.

Table 1 Comparison of inverter chain buffers with different tapering factors

Circuit number	1	2	3	Dimension
Tapering factor	2.5	4.6	10	
Number of inverters	6	4	3	
Total power (relative)	1.14	1.11	1	
Total area (relative)	1.55	1.21	1	
Max d*I*/d*t* (relative)	4.6	3	1	
Max d*I*/d*t* (absolute)	$2.8{\cdot}10^8$	$1.8{\cdot}10^8$	$0.6{\cdot}10^8$	[A/s]
Propagation delay (relative)	0.98	0.94	1	
Propagation delay (absolute)	0.92	0.88	0.94	[ns]

The tapering factor *e* (close to 2.5), which is derived in literature [Mead, 1980] to achieve minimum propagation delay, scores very badly with respect to the maximum d*I*/d*t* and to the area. Since the noise margins of a CMOS IC reduce with every new technology, due to the voltage reduction, the d*I*/d*t* should be as low as possible, but without deteriorating the performance too much. The table shows that the inverter chain with a tapering factor of 10, which was derived to achieve minimum power, also yields optimum overall performance (power, delay, area and noise). Research from previous CMOS technology generations (2.5μm CMOS and 1μm CMOS) had also resulted in an optimum tapering factor of around 10 (Part II, Chapter 1.2). Generally we can conclude that a tapering factor close to 10 will still result in optimum buffer design.

1.2.3 Conclusions

An important contribution to the total power consumption in CMOS circuits is the short-circuit power, which occurs during signal transients on the input(s) of a logic gate. An expression for this short-circuit power is derived and it shows that it could be a relatively large part of the total power, if it is not given proper attention during the design phase. Particularly in circuits that have large transistors, such as in clock drivers, bus drivers and output buffers, this power consumption could be large. These circuits usually consist of a chain of inverters, of which the sizes of the successive inverter transistors are tapered. It turns out that a tapering factor equal to about 10 shows the best circuit characteristics in terms of power consumption, area and noise.

In today's perspective

Over several generations of CMOS technologies, the transistor channel length has reduced while keeping the supply voltage constant. The resulting high electrical field across the channel causes so-called velocity saturation of the charge carriers in the channel. The transistor saturation current then changes from a quadratic relation with the voltage to a more linear one. This will also have its impact on expression (7) for the short-circuit power consumption, which will now change to:

$$P_{short} = \frac{\beta}{12} \cdot (V_{dd} - 2V_t)^j \cdot \frac{\tau}{T}$$

with 2< j< 3.

However, the reduced saturation current would increase the rise and fall time τ to:

$$\tau = \frac{C \cdot V}{I} = \frac{C \cdot V_{dd}}{\frac{\beta}{2} \cdot (V_{gs} - V_t)^{j-1}}$$

To maintain the same rise and fall time τ, the β of the transistors, and so their *W/L*-ratios have to be increased. As a result of these considerations, relatively, the short-circuit contribution to the total power consumption will hardly change.

Regarding the optimum tapering factor (see Table 1) it can be stated that this hardly changes with technology scaling, since almost all capacitances scale with about the same factor. In a 0.12μm CMOS, a tapering factor of 10 still shows the best numbers in terms of area, max d*I*/d*t* and power consumption (from simulations).

Chapter 2

Design for high density

The smaller the chip area, the more dies will fit on a wafer. Moreover, the production yield of integrated circuits is exponentially proportional with the chip area. So, the size of a chip has a great influence on the eventual selling price. Since the price erosion of consumer products is relatively high compared to other goods, the main focus in the design of a consumer IC is on its chip area. This generally holds for video signal processing functions, but even more particular for video memories, as they require relatively large capacities for storing complete video frames. The charge-coupled device (CCD) concept was known to offer two to three times higher bit density compared to dynamic random-access memories (DRAMs) in the same technology node. The first topic to be discussed in this chapter describes two generations of low-cost video memories, implemented in different CCD-MOS technologies. With only one or two additional masks, these technologies offered the combination of high-density memory with specific video processing on the same chip. Because most of the vendors of high-density memories focused their technologies on the DRAM concept, the learning curve of these devices eventually surpassed that of the CCD memories.

Introducing new features into a TV or VCR, requires a fast turn-around time and short-time-to-market of the different ICs from which these features are built. In many cases new-feature TV sets are put onto the market as prototypes, to allow fast market penetration and market survey. In many cases such prototype systems are implemented as gate arrays, which are pre-fabricated unfinished ICs, containing large arrays of transistors or logic gates. These gate arrays are available off the shelf in different categories and only need one or a few interconnect and contact layers to complete their functionality. This allows turn-around times of just a couple of weeks (2-4), compared to the relatively long throughput times of a complete CMOS process (10-15 weeks). Since the volumes for prototyping were usually not so large (a couple of thousands per design), the cost of silicon was only a fraction of the design, test and packaging costs. It was thus not so much of a problem that these gate arrays did not offer the density that could be achieved by implementing the same function with a standard-cells. However, the volumes for prototyping steadily increased from a couple of thousands to several ten thousands per design, thereby relatively increasing the silicon costs per chip. This was a drive for research into the density improvement of gate arrays. The second part of this chapter discusses, as an outcome of this research, an efficient and flexible architecture for high-density gate arrays.

2.1 High-density CCD video memories

2.1.1 Introduction

The introduction of digital memories in TV and VCR equipment has made it possible to enhance TV pictures with additional features, such as 100Hz, noise reduction, still picture, fast teletext page access, etc. [Berkhoff, 1983; Fisher, 1982]. Storage of complete TV pictures requires relatively large memories due to the number of pixels from which the TV field is built. By the time this research was executed, the charge coupled device (CCD) concept offered two to three times higher bit-density than DRAMs in comparable technologies. The next sections discuss the implementation of two CCD video memories: a 308Kb and an 835Kb respectively. The contributions made to this subject were mainly pointed at the integration level, rather than at the CCD device level.

2.1.2 A 308Kb CCD video memory

The basic CCD cell used in this memory is built from two transistors, one with an aluminium transfer gate and one with a polysilicon storage gate. Fig. 12 shows the CCD cell concept.
Both transistors, which are controlled by the same clock signal, are fabricated with a different gate oxide thickness, resulting in different threshold voltages. Data transfer in a CCD memory can be achieved by switching neighbouring memory cells by different clock signals. A short description of the basic data (charge) transfer will be given first.
In many cases a 2-phase clock is used for this shift operation. Fig. 13 shows the basic shift operation of a 2-phase CCD.

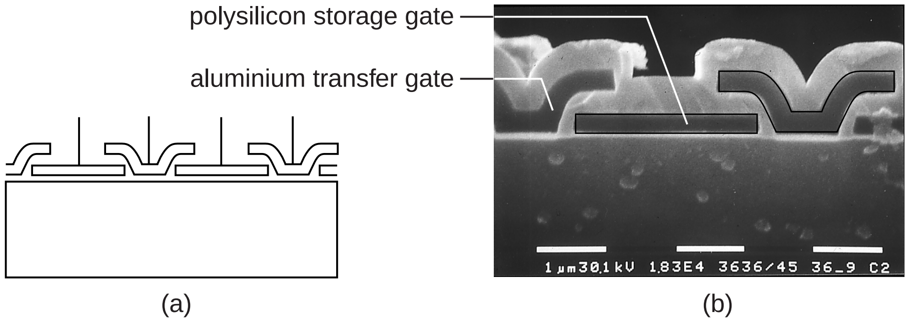


Fig. 12 CCD cell concept (a) and technology cross section of the cell (b)

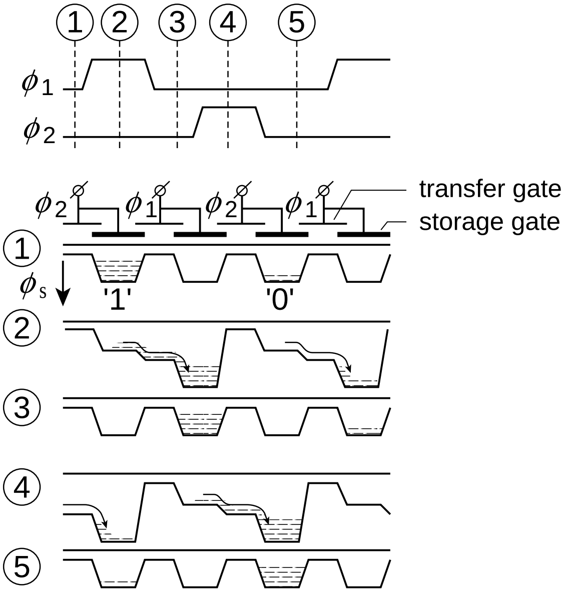


Fig. 13 The shift operation in a basic 2-phase CCD

According to the figure, this shift operation is similar to a repetitive operation of filling buckets with water and then empty them again. The *depth of the buckets*, in this CCD, is determined by the difference between the threshold voltages of the storage and transfer gate.

Suppose the first and third storage gates contain a full and an empty charge packet, representing the logic levels ‘1’ and ‘0’, respectively. The charge packet stored in the first cell is then full of electrons. This is represented by a full *charge packet* under its storage gate. The charge packet stored in the third cell, however, is almost empty, i.e. it is practically devoid of electrons. At time point 1, both ϕ_1 and ϕ_2 are ‘low’ and the storage gates are separated from each other. At time point 2, ϕ_1 has switched from a low to a high level and the charge is transferred from the ϕ_2 storage gates to the ϕ_1 storage gates. At time point 3, both ϕ_1 and ϕ_2 are ‘low’ again and the charge is now stored under the next ϕ_1 storage gates. The description of the shift behaviour at time points 4 and 5 is obtained by replacing ϕ_1 by ϕ_2 in the above descriptions for time points 1 and 2 respectively. A comparison of the time points 1 and 5 shows that the charge has been transferred from the first to the third bucket in one complete clock period. In fact, the charge is transferred from one CCD ‘cell’ to another in one single clock period. So, each CCD memory cell, here, clearly requires two storage elements, which are analogous to the master and slave latches in a *D-type flipflop*.

The potential of a CCD cell to collect charges also holds for the leakage charge from thermal generation of minorities [Slotboom, 1981]. This charge is able to slowly fill the buckets of a CCD. The CCD that is described here is a surface-channel CCD, meaning that the charge transfer occurs right at the silicon surface under the gates. Unfortunately, the surface is somewhat inhomogeneous and plagued by surface states able to trap electrons. Usually the time that charge is being trapped by such surface states is relatively longer than one shift period of a packet transfer. So, the charge, which is ‘stolen’ from one packet, can be released some time later, thereby joining another packet. If the charge is conducted through a very long chain of CCD cells, the full buckets loose charge, while the empty packets will get filled. The amount, to which packet charge is lost, is called charge-transfer efficiency. The leakage and the transfer efficiency are two important effects in a CCD, which influence the architecture and design of a CCD memory.

Storage of one bit of a digitised TV field requires a memory capacity of 308 lines of 1024 bits, with a storage time of 10 or 20ms, depending on the application. So, coding 8-bit video samples requires eight times this memory capacity and so eight of these 308Kb memory chips. If the 308Kb was implemented in one large SPS CCD block, the density would be high, but each individual sample would be subjected to many transfers, thereby gradually loosing its charge due to the transfer inefficiency. A good compromise between area, speed and power and the number of transfers, lead to the choice to realise this 308Kb memory with eight 39Kb serial-parallel-serial (SPS) CCD structures (Fig. 14).

The serial stream of video samples at the input (DI) is de-multiplexed over these eight SPS blocks and the data at the outputs of these SPS blocks is multiplexed again to regenerate the serial bit stream. Each SPS block contains two serial CCD registers, one for the input and one for the output, each implemented as a 2-phase 128b register. The 128 parallel registers in the SPS contain 170 storage gates each. First the serial register is filled with a high-speed clock and then a parallel transfer empties this serial register into the parallel registers. So, the frequency of the parallel clock registers is only 128-th of the serial register clock frequency, leading to much less power consumption compared to the case that all samples would shift through one large 39Kb serial register. Each charge packet thus faces a limited number of total transfers (128 transfers in the serial register and another 170 in the parallel register) making it less sensitive to charge-transfer loss. A 10-phase ripple-clock is used in these parallel registers, which results in one empty bit followed by nine data bits. First the ninth data bit moves its charge to the neighbouring empty bit position, thereby moving the empty bit to the ninth position. Next the eighth data bit moves its charge to the empty bit, thereby moving the empty bit to the eighth bit position, and so on. In this way, a storage density close to one-electrode-per-bit is achieved. As a consequence of this architecture of an SPS, it can store a data bit up to 20ms without refresh. The serial and parallel clock signals needed to operate an SPS block, are generated by the combination of the Gray-code counter, the clock decoder and the ripple-clock generator.

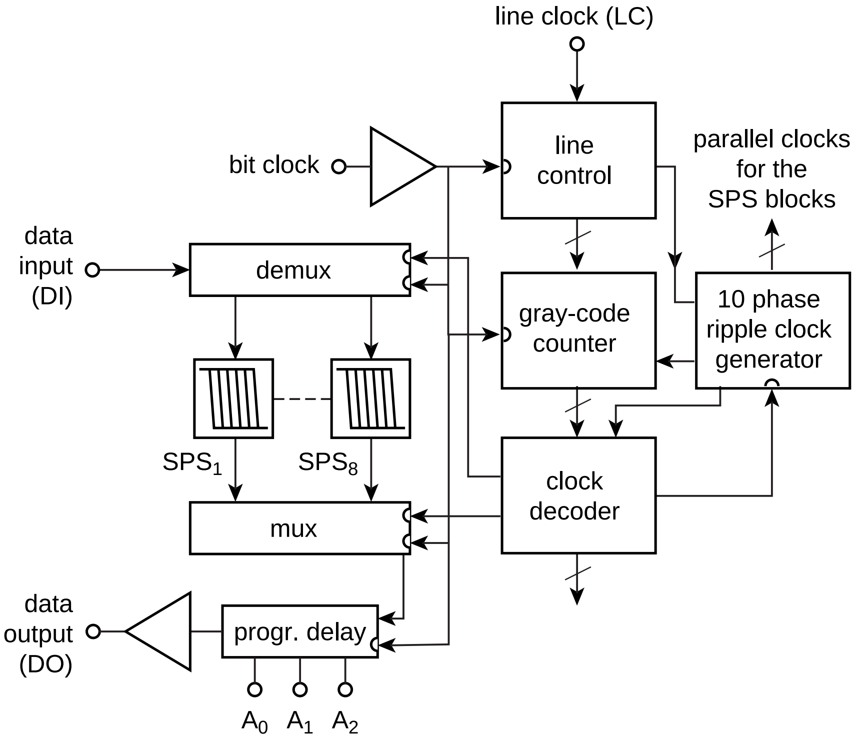


Fig. 14 Block diagram of the 308Kbit memory

Storing a 'still picture' in the memory requires the data to re-circulate within the CCD memory, meaning that the data output (DO) samples have to be fed back to the data input (DI). The programmable 7b-delay has to take care of a correct data synchronisation during re-circulation. The line clock control block allows this memory to be easily locked to the TV picture.

An optimised technology (2 μm CCD-nMOS) resulted in a high transfer efficiency ($\varepsilon \cong 5.10^{-4}$) and a very low leakage current (0.2μA/cm^2 at 90°C). This technology, combined with a dedicated chip architecture and an optimised SPS structure, allowed a very dense integration of this video memory, whose area was less that 35mm^2 (somewhat more than half the size of a 256Kb DRAM at that time). Bootstrap techniques are used to speed up the driver circuits and clock buffers (US patent 04697111). More, design considerations, performance parameters and a chip photograph of this memory can be found in Part II, Chapter 2.1A.

2.1.3 An 835Kb video serial memory

The major differences between this CCD memory and the previously discussed one are the technology, the architecture and the bit density. The discussions in this section will therefore mainly be focused on these topics only. The elementary CCD cell in this memory concept is built from a first polysilicon storage gate (on 25 nm thick oxide) and a second polysilicon transfer gate (on 40 nm thick oxide). Due to the 4-phase clock used in this device, the basic shift operation is somewhat different from the previous one. Fig. 3 in Part II, Chapter 2.1B shows this shift operation in detail.
According to the CCIR standard, in the PAL system each field contains 288.5 lines of 720 active samples. In the 4b wide memory (Fig. 15), each bit plane is thus implemented as a memory block of 290 lines of 720 bits (208,800 bits).

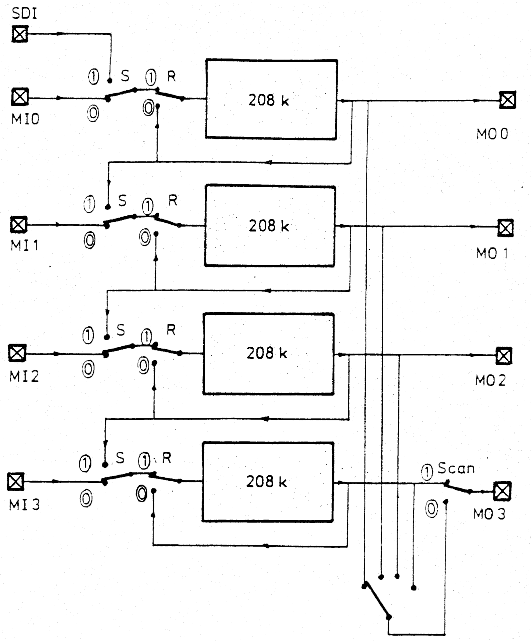


Fig. 15 Block diagram of the 835Kb memory

To avoid problems with speed, power and transfer efficiency, the data-flow within a bit-plane is de-multiplexed over eight SPS memory arrays of 26Kb each. The operation of an SPS is analogous to the one discussed in the previous section. The major topic of this chip is the minor technology adaptation (only two more masks) of a baseline 1.2μm CMOS process to a CCD-CMOS technology. This allows an effective combination of dense memory with logic and enabled the inclusion of features into the chip that support several operation modes which facilitate its use in digital video systems:

1- the 258 lines mode to support the NTSC system (240 active lines per field).

2- the normal 208Kb by 4 mode (switches S and R in positions 0 and 1 respectively).

3- the multiplex mode, in which the four inputs and outputs are multiplexed over two pins each, saving four I/O pins per memory.

4- the serial mode allows an 835Kb by 1 operation of the chip (both switches S and R in position 1).

5- The re-circulate mode (switches R in position 0) for still picture applications.

The memory requires no addressing and can be controlled with only two external clock signals, identical to the memory bit clock and the memory line clock in Fig. 14, through which it can be locked to the TV picture. This application-specific design combined with the high bit rate of this memory resulted in a reduction of system overhead at the printed-circuit board level. The dense integration of the CCD memory cell, together with an optimised architecture resulted in a chip area of 29mm^2, which was about 30% smaller than a comparable DRAM implementation. The power consumption of only 250mW, was more than a factor of three less than video memories implemented in DRAM technology.

More details on this chip and a micro-photograph can be found in Part II, Chapter 2.1B.

2.1.4 Conclusions

Basically, a CCD is a device that shifts charge packets through a serial chain of cells. Each packet contains an amount of charge, which can represent both analogue and digital values. The serial character of operation and the fact that a logic zero and a logic one can easily be represented by an empty and a full charge packet, makes the CCD concept particularly suited for the storage of video pictures. No random access is required, which means that we don’t need bit lines and word lines to be pre-charged every clock cycle. This saves a lot of area overhead and improves both the power consumption and density of video memories by almost a factor of two.

In today's perspective
CCD memories lost attention in the late eighties, not because they lost their density and power advantages, but due to the fact that there were many DRAM competitors sharing the development costs of these memories resulting in a faster DRAM learning curve. Today, the CCD devices are still used in digital and video cameras as sensors to capture the picture. These CCDs also contain a storage part, which is used during read out. The parallel data is then transferred into a serial bitstream. Charge transfer efficiency is currently less of an issue, because of two reasons. First, the use of polysilicon gates allows optimised annealing steps to reduce the number of surface states that could trap charge. Second, in stead of surface-channel devices, buried-channel devices are now used, which are much less sensitive to surface states since the charge transport is now below rather than at the silicon surface.

2.2 High-density gate arrays

2.2.1 Introduction

Gate arrays already exist for several decades to support fast prototyping of new system concepts. Since these gate arrays must support a wide range of potential applications, their architecture has to be flexible. However, the flexibility of general-purpose architectures is usually at the cost of additional area. Therefore, several high-density gate array (HDGA) architectures have been proposed in literature, to accommodate a dense integration of prototype circuits. The goal of the research performed in this area was to increase the efficiency of the functions mapped onto the gate array, without loosing any of the gate arrays' flexibility.

2.2.2 An efficient and flexible architecture for high-density gate arrays

In gate arrays, the basic devices or cells can be isolated by means of oxide isolation [Wong, 1986; Takahashi, 1985] (Fig. 16a), often referred to as 'sea-of-gates', or by means of the gate-isolation technique [Ohkura, 1982] (Fig. 16b), which is often associated with the 'sea-of-transistors' architecture. This architecture consists of a continuous array of transistors, in which the logic gates are separated from one another by transistors that are switched off. This can be done by including one couple of nMOS and pMOS isolation transistors in every logic cell. Their gates are then connected to the ground and supply lines, respectively.

Either of these gate arrays, however, offer nMOS and pMOS transistors of only one size each. Particularly, circuits such as transmission-gate flipflops, ROMs, RAMs and PLAs, as well as dynamic CMOS circuits, require transistors of different sizes. In the sea-of-gates architecture as proposed by [Duchene, 1989] and shown in Fig. 16c, the nMOS transistor is split up into two smaller ones in parallel, each having its own connection(s). However, in memory arrays, many transistor gates usually share a common word line and thus do not require individual connections. These considerations have been used in the development of a new HDGA architecture (US patent US05250823) (Fig. 17).

Each basic cell in this architecture provides three nMOS and three pMOS transistors. Every wide transistor lies in between two narrow ones. Both the nMOS and pMOS transistors each share a common gate. The contact positions in both horizontal and vertical directions are on the same grid, defined by the routing pitch. By the time this research was done, triple-metal CMOS processes were used to implement HDGAs. However, due to additional complex planarisation and contact-etching steps, the third metal layer in this technology would increase the silicon cost by up to 25% and turnaround time by 35%. Replacing this third metal layer by a titanium-silicide ($TiSi_2$) layer, increases the silicon cost and processing time by no more than 5 to 10%, because connections in this layer, called *straps*, enable direct contact between polysilicon gates and source or drain areas of transistors, without the use of metal layers and contact holes. These straps are used to bridge only short distances, such as those within library cells. A detailed description of using these straps to efficiently implement ROM cells (US patent US05053648), RAM cells, D-type flipflops and basic logic gates on the developed architecture can be found in Part II, Chapter 2.2. It is clear that the small (narrow) transistors offered by the architecture allow smaller memory cells. In this section, however, the focus will be on the use of the straps in combination with the architecture, to show that it also allows a dense integration of logic gates. The Exclusive-NOR (EXNOR) and multiplexer gates of Fig. 18 are realised with only one metal layer in a three metal layer technology. The other two metal layers are then completely available for routing purposes.

The narrow transistors connected as inverters to create the inverse functions $\overline{x}$ and $\overline{y}$. The wide transistors create the EXNOR-gate. The polysilicon tracks that carry the inverted signals $\overline{x}$ and $\overline{y}$ not only act as inputs to the wide transistor gates, but also serve to connect the drains of the widely spaced narrow nMOS and pMOS transistors that form the inverters. So, creating logic gates that have inverting input signals, such as in the above examples in Fig. 18, these inputs need no separate large inverters, because these are implemented in the small transistors, parallel to the large ones, which create the logic gate. This architecture is therefore particularly suited to realise multiply and add functions with two times higher densities than previously presented gate arrays, in a three metal layer technology.

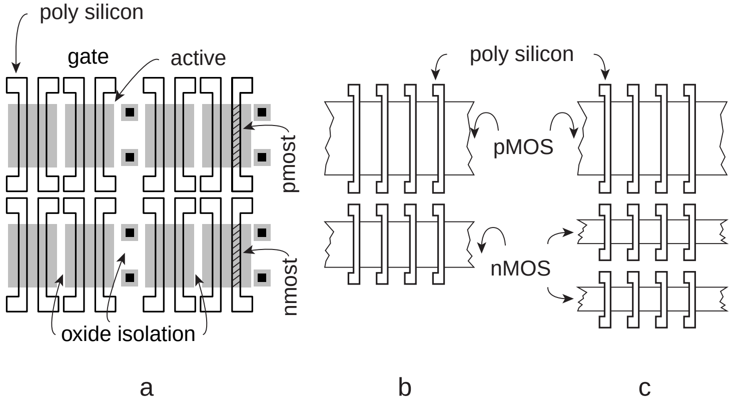


Fig. 16 (a) Typical example of a sea-of-gates architecture
(b) Typical example of a sea-of transistors architecture
(c) A sea-of-gates example that supports memory implementation

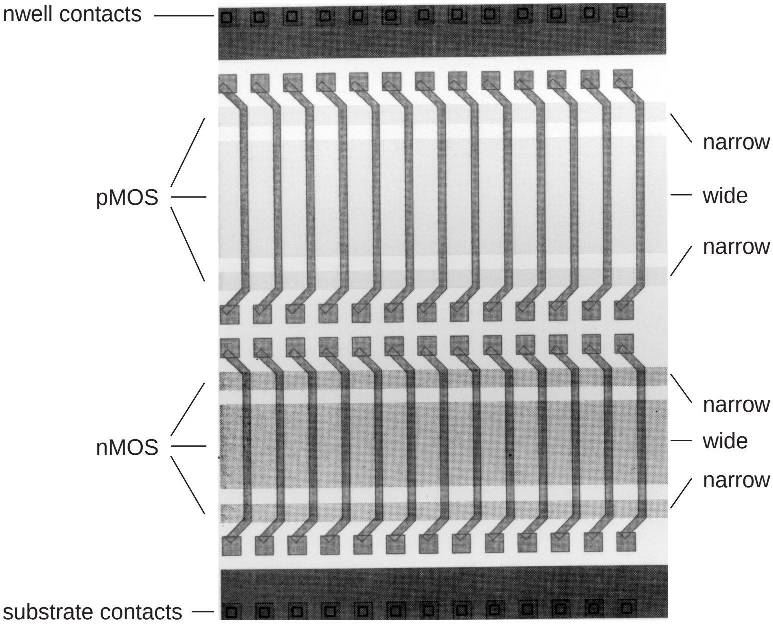


Fig. 17 The new common-gate HDGA architecture

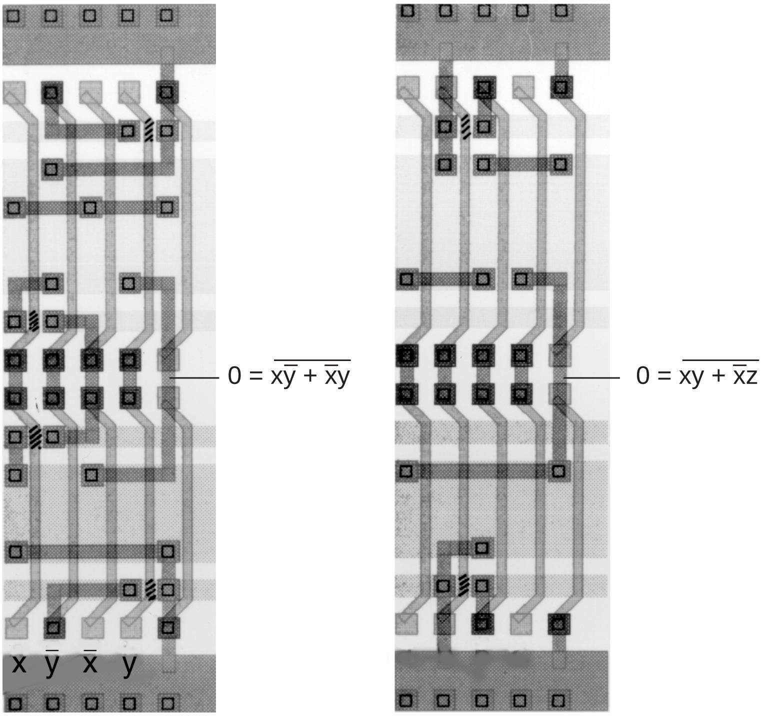

$0 = \overline{x\bar{y} + \bar{x}y}$
$0 = \overline{xy + \bar{x}z}$
x $\bar{y}$ $\bar{x}$ y

Fig. 18 The use of small transistors in creating logic gates: an EXNOR (a) and a multiplexer (b)

2.2.3 Evaluation and results

A test chip has been designed, which includes several different implementations of a 10 by 10 bits multiplier, a 24Kb ROM and some performance and technology evaluation modules. However, the regularity of a multiplier is probably not representative for a general logic circuit. Therefore a further evaluation of the HDGA architecture has been performed. The results of this evaluation are presented in Table 2. Next to two different multipliers, also a complex logic block of a compact disc servo-control chip (CD-BLOCK) and two different fast-fourier transform (FFT) designs (FFT A and FFT B) have been mapped onto the HDGA architecture. FFT B is also realised with three different layout aspect ratios (*W*/*L* ≈ l, 4, ¼, respectively), to investigate the relation between this ratio and the eventual layout area. Their densities are compared with a standard-cell implementation in a two-metal layer and straps CMOS technology.

Table 2 Comparison of different logic functions in standard-cell and HDGA design

		Standard cell (SC)	**Common-gate HDGA**		**HDGA/SC**
Design	# of gates (2-NAND)	Aspect ratio	Aspect ratio	Transistor utilisation	Area ratio HDGA/SC
MPY (20×20)	20800	1.5	1.50	96%	0.97
MPY (10×10)	5000	1.51	1.51	95%	1.00
CD-BLOCK	1350	2.06	2.34	94%	1.00
FFT A	2637	1.02	1.07	87%	0.68
FFT B (1)	1980	1.06	1.03	86%	0.77
FFT B (4)	1980	3.92	3.92	74%	0.82
FFT B (¼)	1980	0.33	0.27	81%	0.82

The 10 by 10 bits multiplier and the more complex 20 by 20 bits multiplier occupy about the same area as their standard-cell counterparts. The HDGA implementation of the CD-BLOCK (a logic block of a compact disc servo control chip) occupies exactly the same area as the standard-cell version. All HDGA versions of the FFT designs score even better than the standard-cell versions. Because their net lengths and transistor sizes are also about equal, the HDGA circuits show the same performance as the standard-cell versions. The overall conclusion from this table is that the HDGA implementation showed equal performance at comparable or even smaller chip areas. This is quite a good result, since usually there was an area difference of a factor of 1.5 to 2, in favour of the standard-cell versions. There are two reasons for this increased HDGA efficiency. First, the use of straps in HDGAs turns out to be very efficient compared to the metal interconnections in the standard-cell designs. Second, the use of the narrow transistors in the HDGA architecture enables a very dense implementation of a large variety of cells. More details on this subject can be found in Part II, Chapter 2.2.

2.2.4 Conclusions

A flexible high-density gate-array architecture is presented. The inclusion of narrow transistors in parallel with wide transistors in the basic cell shows a lot of advantages. First they are used to generate local inversions within logic gates, without any area penalty. Next, it supports the mapping of flipflops that require small feedback transistors. Finally, memory cells like SRAM and ROM cells are easily mapped onto this architecture. A test chip shows that even with a regular array of transistors logic gates and memories could be implemented with a density improvement of about a factor of two compared to existing gate arrays.

In today's perspective

High-density gate arrays are still used for fast prototyping in many different applications. However, due to the increased density of the transistors over the years, the number of metal layers that these gate arrays currently use, has gone up to about four or five. This has lengthen the turn-around time to four to five weeks, compared to about two weeks that it took more than a decade ago. Also the NRE costs (non-recurring engineering costs, which includes almost all costs related to design, test and packaging, except for the silicon costs) have increased to above US$ 100,000. Because of the expected increase in the costs of a mask set, more flexibility in terms of design and redesign is required. Over the last couple of years, programmable logic, such as FPGAs (field-programmable gate arrays) gained much more attention and, in some cases, have overtaken applications which were formerly implemented with gate arrays.

Regarding future VLSI design, hundreds of millions of transistors are expected to be integrated on a digital chip. A higher degree of regularity and a lower variety of differently shaped transistors could help to improve yield. The concepts of a regular array of transistors, such as in the HDGA discussed in this paragraph, could very well be applied to the development of a standard-cell library, to allow technologists to only focus on a limited number of transistor topologies to increase yield.

Chapter 3

Design for robustness

The robustness of an integrated circuit has been a major topic already since the early days of its existence. Two important issues regarding this robustness are reliability and signal integrity. Reliable circuit operation includes the design tolerance with respect to the requirements of the application (specification) as well as to those of the process technology (design rules). Signal integrity reflects the degree to which the shape of a signal is affected by cross-talk, noise or propagation delay. Due to the continuous process of scaling of the minimum features sizes, the supply voltage and the clock periods, the design margins have dramatically reduced, thereby threatening the reliability and signal integrity of current and future ICs.

Two subjects relate to circuit reliability. The first one is an example in which the reliability requirements are defined by the application. It describes the problems that arise from communication between asynchronous (sub)systems and the solutions that have been developed to guarantee a reliable communication between such systems. As such, the design of a synchroniser will be presented together with a prediction of its failure rate. The second subject is an example in which the reliability requirements are defined by the technology design rules. Due to the increased current densities in the on-chip metal supply and interconnect lines, electromigration is becoming increasingly important. Particular its exponential dependence on the temperature has intensified the research in such effects as wire self-heating due to the expected increase of the power consumption in these resistive metal lines. Both a qualitative and quantitative discussion of wire self-heating is presented.

An increased number of devices on a chip, combined with reduced feature sizes and reduced physical spacing, causes an increase of on-chip generated switching noise and interference. At the same time the supply and transistor threshold voltages decrease, thereby reducing the noise margins. In other words, due to scaling the noise increases, while the margins reduce. This puts a burden on the signal integrity within an IC and on the reliability of its operation. Maintaining signal integrity and design robustness at a sufficiently high level will increasingly require specific design measures. These signal integrity issues will be discussed in the third subject in this chapter.

3.1 Reliable communication between asynchronous systems

3.1.1 Introduction

Most signals that are input to a synchronous chip, are sampled by flipflops, which are controlled by a clock signal. In the communication between digital (sub)systems that do *not* share a common time reference, signals may occur which are not logically defined [Chaney, 1973]. This means that signals are generated randomly in time (asynchronous), with respect to each other. It is particularly possible that an input signal will change during an edge of the sample-clock. There is a large possibility that such a situation will cause a system failure. This can only be prevented by using a synchroniser circuit, most commonly a flipflop, which must support reliable communication between such asynchronous systems. A flipflop can adopt three different states: a logic ‘one’, a logic ‘zero’ and a meta-stable state (MSS). In a (differential) flipflop, the exact meta-stable point is defined to be the point in which both flipflop nodes show equal voltages. Sampling asynchronous (random) data may put the synchronising flipflop into an undefined MSS. Due to the feed-back loop in the flipflop, it will recover from such an MSS. The closer to the meta-stable point, the more time it takes to recover from it, and the longer an MSS lasts. If such an MSS lasts almost a full clock period, it will cause anomalous response times [Pechoucek, 1976]. Part II, Chapter 3.1 describes the design, operation and performance of a synchroniser in detail, as well as the influence of noise on its MSS behaviour. The next subsection discusses the most important topics and results of this research.

3.1.2 Synchroniser behaviour

Sampling asynchronous data signals at an input of a chip, may result in situations in which the input signal is changing (e.g. from a ‘0’ to a ‘1’) during the sample moment of the clock signal ϕ (Fig. 19a). For a better understanding, slow rising and falling edges of the input signal are drawn in this figure.

The sample transistor is drawn as a MOS pass transistor, whose output is connected to a flipflop, which consists of two cross-coupled inverters only. Let us assume that the feed-back loop in the flipflop is very weak compared to the sample transistor. Now, this flipflop will reach a MSS (Fig. 19b), when the falling edge of the sample clock appears within a small time interval $\delta(t)$ during a rising or falling edge of the input signal.

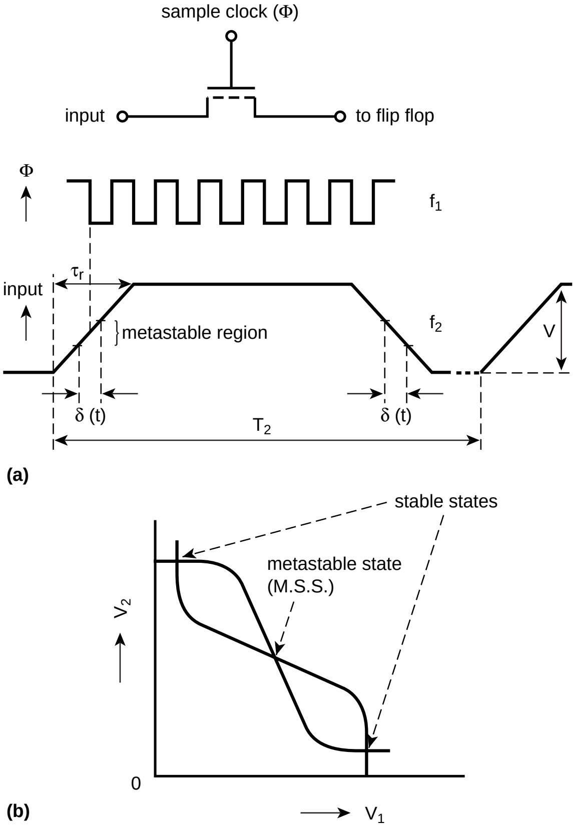


Fig. 19 (a) Possible signal configuration at asynchronous communication
(b) Possible states of a flipflop, consisting of two inverters of which the transfer characteristics are given

Because the input signal is asynchronous, its edges appear randomly in time. The sampled signal values during these edges are therefore uniformly distributed. Fig. 20a shows the uniform distribution of sampled voltage values at *t* = 0 around the meta-stable point (0) and the path that each individual sampled state follows over the time. It is mathematically proven in Part II, Chapter 3.1, but also clear from the figure, that this distribution remains uniform within a certain region around the meta-stable point. When the signal values approach one of the logic levels, this uniform distribution is no longer maintained, but then the signal will soon become a logic one or zero itself and can no longer return to the meta-stable state by a noise pulse.

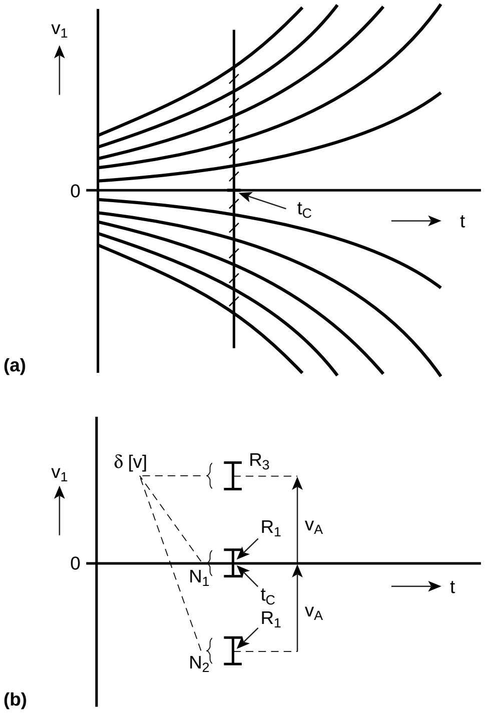


Fig. 20 (a) Behaviour of the nodal voltages of a flipflop starting from uniformly distributed sample values at $t = 0$ (only a limited number is drawn)
(b) The number of states N_1 that is forced out of the MSS region by a noise voltage V_A at a time $t = t_c$, is replaced by the number N_2 of states that is forced back into the MSS region by the same noise voltage

Let us assume that at a randomly chosen time, $t = t_c$ (Fig. 20), the flipflop is ‘hit’ by a positive noise voltage peak V_A. Any meta-stable state, whose voltage value at $t = t_c$, is still within region R_1 in Fig. 20b, would be moved up to region R_3, thereby soon leaving the meta-stable region to become a logical ‘1’. However, any meta-stable state, whose voltage value at $t = t_c$, is within region R_2 and almost becoming a logical ‘0’, is put back into region R_1, close to the meta-stable state again, by the same noise voltage peak V_A. In other words, over the time, the number of meta-stable states that will be forced out of the meta-stable region by a noise pulse, is identical to the number that is forced back into it. This is one of the most important conclusions of this research: ‘*although each individual MSS is affected by random circuit noise, the average number of MSSs during a constant time period is independent of the noise’.*

This result has dramatically reduced the complexity of handling flipflop failures and estimating failure rates.

3.1.3 Synchroniser design

In Part II, Chapter 3.1 the relation between the probability of occurrence of a MSS and the design of a latch, which is the basic circuit behind a synchroniser (Fig. 21), is derived.

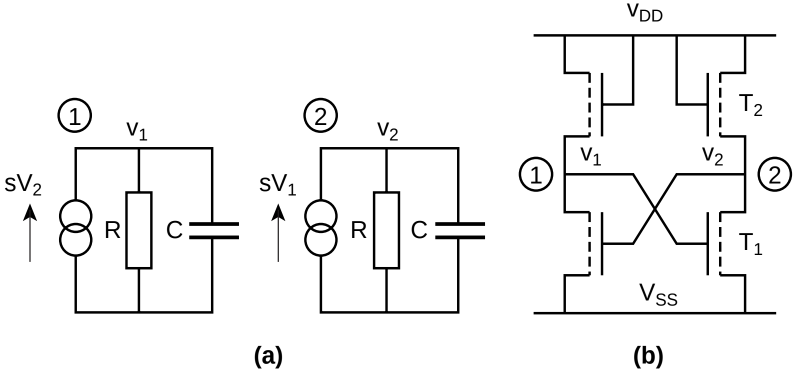


Fig. 21 (a) Small-signal model of an nMOS latch
(b) A latch in nMOS technology

According to this, the probability of occurrence of a MSS, whose duration *t* is longer than a time t_n is given by:

$$P(t > t_n) = \exp\left\{\left(-\frac{A-1}{\tau}\right)\cdot t_n\right\} = \exp\left\{\left(\frac{1}{RC} - \frac{s}{C}\right)\cdot t_n\right\} \qquad (9)$$

where A represents the amplification of the latch, while τ is the *RC*-product, *s* is the trans-conductance of the driver transistors (T_1), *R* the differential resistance of the loads (T_2) and *C* the capacitance at nodes 1 or 2.

To minimise this probability, the factor (*A*-1)/τ must be maximised. Variables *R*, *C* and *s* are all dependent on the sizes of the transistors in the latch. It is derived in Part II, Chapter 3.1, that this factor is a function of the ratio between the sizes of the driver and load transistors, and, in different technologies, it has a maximum at:

$$\frac{(W/L)_{T_1}}{(W/L)_{T_2}} \approx 8 \qquad (10)$$

Proceeding from this result, a synchroniser has been designed according to Fig. 22. Next to the above derived W/L ratio, also much attention has been given to reduce the parasitic capacitances at the internal flipflop nodes.

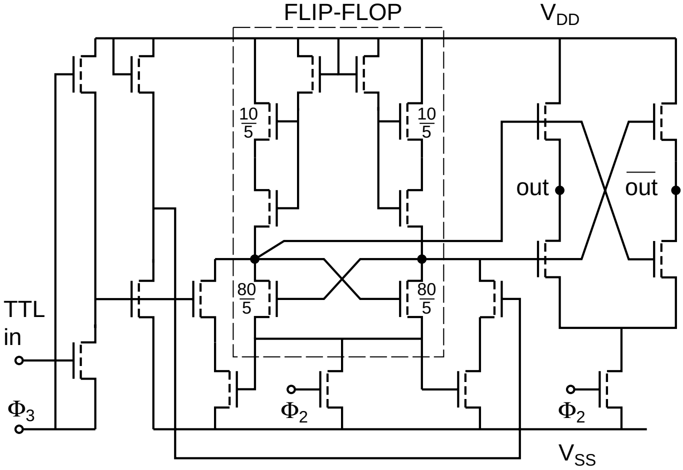


Fig. 22 A synchroniser, designed according to the theory developed in this paper

The next subsection presents a qualitative discussion of this synchroniser.

3.1.4 Reliability of this synchroniser and prediction of its failure rate

It has already previously been described, that an MSS will only occur, when the falling edge of the sample clock will appear within a small time interval $\delta(t)$ during a rising or falling edge of the input signal (see Fig. 19). The probability that this happens can be described with:

$$P = \frac{2\delta(t)}{T_2} \cdot \frac{f_1}{f_2} = 2 \cdot \frac{\Delta v}{V} \cdot \frac{\tau_r}{T_2} \cdot \frac{f_1}{f_2} = 2 \cdot \frac{\Delta v}{V} \cdot \tau_r \cdot f_1 \qquad (11)$$

where the variables are depicted in Fig. 19. f_1 and f_2 are the sample clock frequency and the frequency of the asynchronous input signal, respectively.
With n seconds a year ($n = 31.5 \cdot 10^6$), the average number N of occurrences of such MSSs during one year is given by:

$$N = P \cdot f_2 \cdot n = 2 \cdot \frac{\Delta v}{V} \cdot \tau_r \cdot f_1 \cdot f_2 \cdot n \qquad (12)$$

All variables in equation (12) are known, once we know the application. The following values, which were typical for an application at the time of this research, are assumed as an example:

data input rise/fall time	t_r = 10ns;
meta-stable region	Δv (to be determined later);
clock frequency	f_1 = 4MHz;
average data-in frequency	f_2 = 400kHz;
data-in amplitude	V = 10V;
allowed MSS duration	40ns (depends on f_1).

Only the meta-stable region Δv depends on the design of the synchroniser. If we assume that the maximum duration of an MSS is limited to 40ns (depending on the application), then, by using a circuit simulator, we can determine the meta-stable region. For the designed synchroniser it is simulated that this meta-stable region, under the above condition, is equal to: $\Delta v = 6{\cdot}10^{-15}$V. All sampled values, which are within this meta-stable region Δv, cause MSSs that last longer than the allowed duration of 40ns and cause a failure. With equation (12) we can calculate that the failure rate of this synchroniser in the above application is $6{\cdot}10^{-4}$ times/year.
Further details on this research can be found in Part II, Chapter 3.1.

3.1.5 Conclusions

It is shown that random noise has no influence on the average number of meta-stable states in a synchronising flipflop and therefore does not affect its average failure rate. An optimised synchroniser design is achieved by adopting an aspect ratio of eight between the driver and load transistors and by careful layout, such that its parasitic nodal capacitances are minimised. One synchroniser per bit is needed in asynchronous communication, meaning that in a 16-bit wide communication bus, only 16 of these synchronisers are needed, which hardly causes any area overhead.
With an optimised design, the failure rate can be reduced, but not to zero. To solve the problems associated with asynchronous communication fundamentally, synchronisers have to be developed that have an infinite factor: $(1/RC - s/C)$. Such synchronisers, however, do not exist.

In today’s perspective

The concepts of this research are still being used in the design of synchronisers and to the bit-error rate specification of analogue to digital converters in advanced digital transmission systems.

While supply voltages have reduced over the last couple of technology generations, the noise levels did not. So, it is important to have a feeling about the levels of thermal noise. Thermal noise has the flat frequency spectrum of white noise with an rms level V_n of [Gray, 1996]:

$$V_n = \left(\sqrt{4KTRB}\right) = \left(\sqrt{4 \cdot \frac{KT}{q} \cdot qRB}\right)$$

where *KT/q* represents the thermal voltage (about 25mV at room temperature), q the elementary charge of an electron, *R* the differential resistance of the MOST and *B* the bandwidth of the circuit. Let us assume the synchroniser will be made in a 0.12μm CMOS technology. In the meta-stable region both transistors of the latch are in the amplification mode. Their *W/L*-ratio may be as high as 4 and for a typical transistor in this technology it holds:

$$R = \frac{\partial V_{gs}}{\partial I_{ds}} = \frac{L}{\mu \cdot W \cdot C_{\text{ox}} \cdot \left(V_{gs} - V_T\right)} \approx 2k\Omega$$

A typical value for the parasitic capacitance of a latch node is: $C \approx$ 5fF.
This results in a bandwidth of

$$B = \tfrac{1}{2\pi RC} \approx 16GHz.$$

With the above expression for V_n this results in an rms noise level of:

$$V_n \approx 700\mu V.$$

This noise level is much less than the linear region of the transistors when operating in the meta-stable region. So, today, the conclusion that noise does not affect the average number of failures, is still valid.

3.2 Wire self-heating in supply lines on bulk-CMOS ICs

3.2.1 Introduction

Electromigration (EM) is the effect by which metal ions in interconnect lines are physically moved in the same direction as the electrons, due to a very high current density which exceeds a certain maximum allowed level. Every technology design manual includes design rules regarding the minimum metal width to fulfil EM requirements at specific die temperatures.

As a result of the continuous process of technology and voltage scaling, current densities are expected to increase. This leads to an increased chance of EM and to more power dissipation in the metal wires and, more specifically in the supply lines. Particularly due to the increasing distance between top-level metal wires and substrate and the use of low-k, high-thermal resistive dielectrics, the temperature of these wires is expected to increase dramatically due to so-called wire self-heating [Banerjee, 1999; Banerjee, 2000; Streiter, 2000]. EM effects increase exponentially with temperature, causing a relatively large reduction of the maximum allowed current densities in on-chip interconnect

This section evaluates the effects of wire self-heating in supply lines on bulk-CMOS from a design perspective. Due to sufficient on-chip de-coupling capacitance, the supply currents are assumed to be almost constant. Therefore the dynamic aspects of thermal conductance are not taken into account.

3.2.2 Metal width requirements

The maximum current density J_{max} in a metal wire is limited by EM. This is usually specified in the technology design manual. Since most IC data sheets show a maximum ambient temperature around 70°C or higher, the real worst-case junction temperature of the silicon itself may exceed 100°C in many applications. Therefore it is common design practice to use the value for J_{max} at 125°C.

All further calculations and diagrams in the next figures are based on the technology characteristics presented in Table 3.

Table 3 Metal characteristics for 0.12μm and 0.18μm bulk-CMOS technologies

Technology and metal layer	R_{sheet}	H	J_{max} @ 125°C
0.18μm CMOS second metal	72mΩ/□	550nm	2.3mA/μm^2
0.18μm CMOS upper metal	35mΩ/□	900nm	2.3mA/μm^2
0.12μm CMOS second metal	85mΩ/□	350nm	3.5mA/μm^2
0.12μm CMOS upper metal	26mΩ/□	900nm	3.5mA/μm^2

The minimum allowed width W_{em} of a metal wire with height H, to carry a current I, according to this EM requirement, is then equal to:

$$W_{em} = \frac{I}{\left(J_{\max} \cdot H\right)} \tag{13}$$

Usually a circuit is designed such that it operates completely according to the specification, allowing external supply voltage margins of +10% or –10%. Therefore an additional average voltage drop across the on-chip supply lines must be strictly limited. The maximum allowed voltage drop ΔV_{max} across the supply lines is only a fraction r of the total supply voltage V_{dd}, or: $\Delta V_{max} = r \cdot V_{dd}$. The requirement on ΔV_{max} $(= I \cdot R_{wire})$ determines the width $W_{\Delta V_{\max}}$ of each individual supply wire according to:

$$W_{\Delta V_{\max}} = \frac{I \cdot L \cdot R_{sheet}}{\left(r \cdot V_{dd}\right)} \tag{14}$$

where L represents the length of the supply line and R_{sheet} its sheet resistance (resistance of a metal line with identical length and width). There is a wire length at which the requirements on EM and those on voltage drop result in the same metal width: $W_{em} = W_{\Delta V_{\max}}$. This cross-over wire length L_{co} at which this occurs is:

$$L_{co} = \frac{r \cdot V_{dd}}{\left(R_{sheet} \cdot J_{\max} \cdot H\right)} \tag{15}$$

Below this length, the minimum width of the metal line is constant and determined by electromigration. Above this length, the minimum width of the metal line is determined by the maximum allowed voltage drop and increases proportionally with its length. L_{co} is fixed for a given metal layer (and J_{max} at a given temperature) in a given technology (Fig. 23).

Let us assume a core on a chip, supplied with metal lines of a certain length and carrying a current of 100mA. For this supply current, the above figure shows the minimum required metal widths, by the combined requirements for electromigration (@125°C) and voltage drop, for second and top layer wires in both 0.12μm and 0.18μm CMOS technologies with copper and aluminium back-end, respectively. The widths of wires with lengths above L_{co} increase proportionally with the length to maintain a constant resistance to fulfil the voltage drop requirement. We now know the width (W) of the supply lines to this 100mA core, depending on its distance (*length* in Fig. 23) from the supply pad(s). The amount of wire self-heating depends on the total cooling area (*W*x*L*) of the supply line, and its power dissipation.

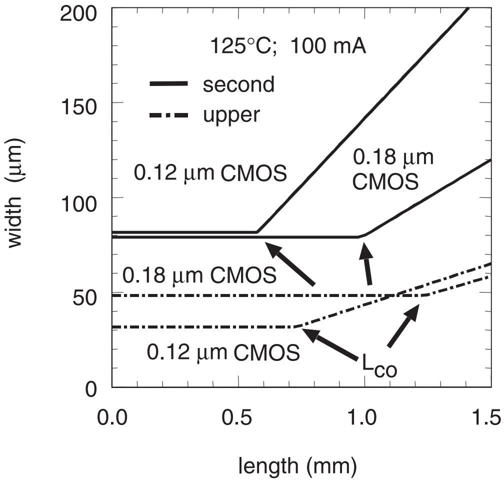


Fig. 23 Minimum wire width required by EM and by 5% voltage drop as a function of the wire length, for the second and upper metal layers in 0.12μm and 0.18μm bulk-CMOS technologies

3.2.3 The maximum power dissipation in a metal track

As shown in the previous subsection, the width of a wire with a length less than L_{co} is only determined by EM. In such wires, the resistance is less than in wires with lengths larger than L_{co}, whose resistance is constant, due to the maximum allowed voltage drop across them. In other words: the power dissipated in wires, longer than L_{co}, is identical to the power dissipated by a wire whose length exactly matches L_{co}. Fig. 24 shows this power dissipation as a function of the length of the wires for different metal layers in different technologies.

In this diagram and in all further discussions, the current through the supply lines is equal to 100mA. However, this value is only taken as an example and the conclusions on temperature rise and wire self-heating will be independent of the value of this current.

The maximum power dissipation in the supply lines is:

$$P_{wire} = I \cdot r \cdot V_{dd} \qquad (16)$$

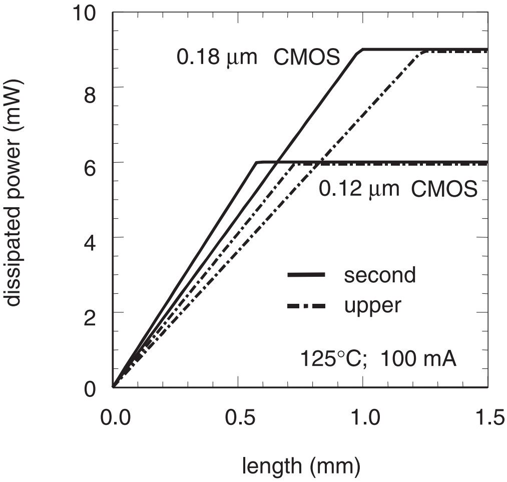


Fig. 24 Power consumption in a wire, for the second and upper metal layers in 0.12μm and 0.18μm bulk-CMOS technologies, as a function of its length

It is common design practice to limit the voltage drop across the supply lines to less than 5% of V_{dd}, so r = 0.05. At a supply current of 100mA, this leads to a maximum power consumption in the wire of 9mW and 6mW for a 0.18μm and 0.12μm bulk-CMOS technology, respectively.

For wire self-heating not only the power dissipated in the wire is of importance, but also its thermal resistance to the substrate.

3.2.4 The temperature increase of a wire by self-heating

As discussed before, the maximum dissipation in a metal wire occurs when its length reaches the previously defined cross-over length L_{co}. However, the temperature rise in a wire also depends on its thermal resistance to the substrate. The effective thermal resistance of a metal line to the chip's substrate is defined as [Banerjee, 1999):

$$R_\Theta = \frac{t_{ins}}{\left(K_{ins} \cdot L \cdot W_{eff}\right)} \tag{17}$$

where

$$W_{eff} = W + \phi \cdot t_{ins} \tag{18}$$

K_{ins} is the thermal conductivity normal to the plane of the dielectric (K_{ins} = 0.6 [W/m °K] for low-K dieletrics (HSQ)) and t_{ins} the thickness of the insulator below the wire.

The width correction in equation (18) is necessary because equation (17) is based on a quasi-1-D heat conduction model. ϕ represents a heat-spreading parameter, which equals 0.88 for $W/t_{ins} \geq 0.4$ (common for top level metal (worst-case) lines in deep-submicron technologies [Banerjee, 1999]). Combining Fig. 23 with equation (17), results in the diagram of Fig. 25, showing the effective thermal resistance as a function of the wire length, only for the top layers in 0.18μm and 0.12μm CMOS technologies.

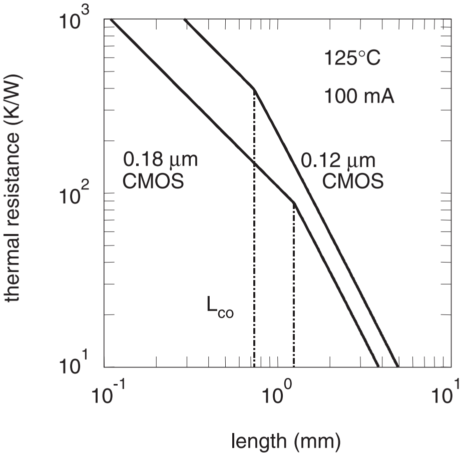


Fig. 25 Thermal resistance as a function of the wire length, for the upper metal layer in 0.12μm and 0.18μm bulk-CMOS technologies

For small lengths, the width is constant, so R_Θ is proportional to $1/L$. For lengths above L_{co}, the width increases with L causing R_Θ to be proportional to $1/L^2$.

We can now calculate the temperature rise [Banerjee, 1999] (since supply currents are almost constant, dynamic effects of thermal conductance are neglected here):

$$\Delta T_{wire} = P_{wire} \cdot R_\Theta = I^2 \cdot R_{sheet} \cdot \left(\frac{L}{W_{em}}\right) \cdot R_\Theta \qquad (19)$$

Combining equation (17) and (19) leads to:

$$\Delta T_{wire} = P_{wire} \cdot R_\Theta = I^2 \cdot R_{sheet} \cdot \left(\frac{L}{W}\right) \cdot \frac{t_{ins}}{(K_{ins} \cdot L \cdot W_{eff})}$$

The relatively wide supply lines allow us to assume W_{eff} and W to be (almost) equal, leading to:

$$\Delta T_{wire} = I^2 \cdot R_{sheet} \cdot \frac{t_{ins}}{\left(K_{ins} \cdot W^2\right)} \tag{20}$$

ΔT_{wire} only depends on the width W of the wire. For wires shorter than L_{co} the width is constant and equal to W_{em} resulting in a constant ΔT_{wire}. However, the width of a wire longer than L_{co} is proportional to the length, which means that above L_{co}, ΔT_{wire} is inversely proportional with L^2. Fig. 26 shows the temperature rise for the upper metal layer in a 0.18 and a 0.12μm technology.

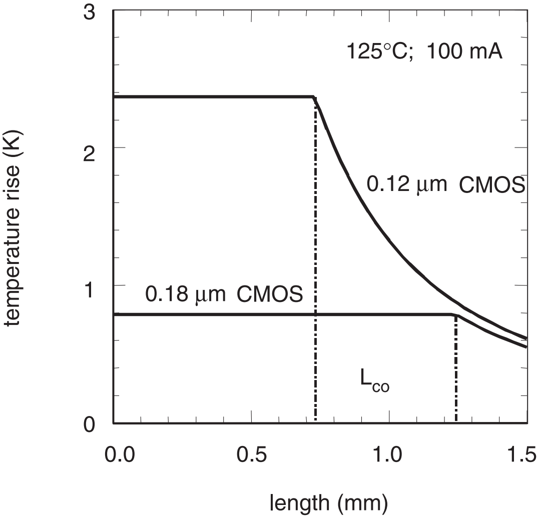


Fig. 26 Temperature rise as a function of the wire length for the upper metal layer in 0.12μm and 0.18μm bulk-CMOS technologies

The maximum ΔT_{wire} depends on the temperature at which the width W_{em} is specified by EM requirements (through J_{max}). The figure shows only a limited temperature rise when W_{em} is specified at 125°C.

The maximum EM-allowed current through a metal wire roughly halves for every 25°C increase in temperature. This means that if W_{em} is determined using a value for J_{max} at a lower temperature, this W_{em} would be much smaller than the one determined using the value for J_{max} at 125°C (e.g. ≈ 3.5mA/μm^2 for upper metal line in 0.12μm CMOS. Since the cross-over length L_{co} of the metal line depends on J_{max} it is also dependent on the temperature at which J_{max} is taken.

Consequently, the wire self-heating will be larger when we specify W_{em} at lower temperatures. Fig. 27 shows the temperature rise due to wire self-heating of a top-metal track, again in both technologies, as a function of the temperature at which W_{em} is specified through J_{max}.

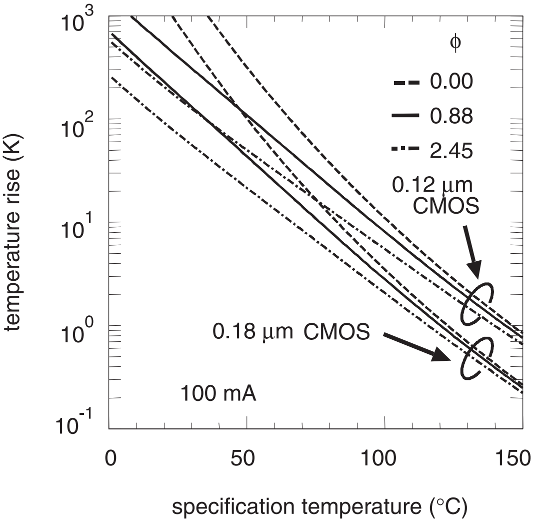


Fig. 27 Temperature rise as a function of the specification temperature for different values of the heat-spreading parameter ϕ for the upper metal layer in 0.12μm and 0.18μm bulk-CMOS technologies

So, the wire self-heating strongly depends on the temperature at which the wire width was specified by EM requirements. As stated before, it is common design practice to take 125°C or higher (150°C) as the specification temperature for the required wire width. Because of this specification requirement, we can read from Fig. 27 (and also from Fig.26) that the wire self-heating in the supply lines of any 0.12μm bulk-CMOS product will be limited to only 2.5°C. This value is independent of the supply current that we have used in the examples, because more current means more dissipation in a wider wire, which therefore has a proportionally larger cooling area.

In other words, when a chip has been designed according to appropriate requirements on EM and voltage drop, the effect of wire self-heating in the supply lines can be neglected. If these requirements were not taken into account during the design phase, then wire self-heating may lead to a temperature rise in the order of 10K to 100K, causing reliability problems.

3.2.5 Conclusions

In a real VLSI design there are two requirements that determine the width of a wire and limit its temperature rise due to the self-heating mechanism: EM and maximum allowed voltage drop across the wire. It is shown that for a given technology there exists a worst-case wire length for self-heating in every metal layer at a given maximum temperature, most commonly 125°C or more. At this condition, it has been shown that for different technologies (0.18μm bulk-CMOS with aluminium back-end and 0.12μm bulk-CMOS with copper and low-*k* back-end), wire self-heating causes only a limited temperature rise of the wire of just a few degrees. This temperature rise is by far negligible compared to the temperature rise due to the power consumption of the silicon part of the chip. From this result it can be concluded that wire self-heating in the supply lines of current (and near future) properly designed CMOS VLSI chips should not be a real issue.

3.3 Signal integrity in deep-submicron CMOS designs

3.3.1 Introduction

The increase in complexity of ICs over the last couple of decades has enabled the integration of millions of transistors on one single die. This has resulted in complete systems on a chip (SOC). Not only the functionality of a board, but also its problems concerning clock skew, supply network and supply de-coupling, interference and EMC will be integrated on the chip as well. The increased manifestation of these effects on a chip is threatening the signal integrity of deep-submicron ICs in different ways. Many of the problems are related with the current, the current peaks and with the back-end of the manufacturing process: the metal layers. The supply lines and interconnections in these metal layers are already dominating the ICs performance and signal integrity today. This section will focus on three important signal integrity topics. At frequencies below 1GHz on-chip metal lines predominantly behave like resistors. When large currents flow through these metal lines, like in supply lines, these currents will cause a voltage drop in these supply lines. Moreover, the many simultaneous switching nodes on a chip, today, cause large current peaks to be supplied by the bond wires and the board wiring. Due to their self inductance, these current transients also cause voltage drop. The total voltage drop across supply lines is often also regarded as supply noise. This supply noise severely affects the behaviour of the connected circuits. Finally, due to the reduced spacing between neighbouring signal lines, they influence each others signal shape and propagation. This so-called cross-talk also influences the overall IC performance and signal integrity.

3.3.2 Voltage drop (supply noise)

Over generations of ICs, the bus widths, the number of outputs, the clock load, etc. have increased dramatically. For example, microprocessor bus widths have increased from four to eight bits in the late seventies, to 64 or even 128 bits at the present time. Consequently, the total load capacitance that has to be simultaneously charged and discharged has increased proportionally. As a result of the higher resistance of the power supply network (larger chip and block sizes) and the increased current slew rates (d*I*/d*t*), the peak currents can no longer be completely supplied in time by the power network. This causes large voltage bounces on its supply lines (supply noise). The associated voltage drop (Δ*V*) is a result of the resistance (*R*) of the on-chip supply line and the self-inductance (*L*) of the board wiring, the bond wires, the package leads and the on chip supply lines themselves:

$$\Delta V = I\,R + L \cdot \frac{\mathrm{d}I}{\mathrm{d}t} \tag{21}$$

The self-inductance is mostly dominated by the package and board wires. The self-inductance of the on-chip wires only becomes important at frequencies close to 1GHz (Fig. 28). At this frequency the inductance value is about ten percent of the resistance value and can no longer be neglected. Above this frequency also transmission line effects must be take into account.

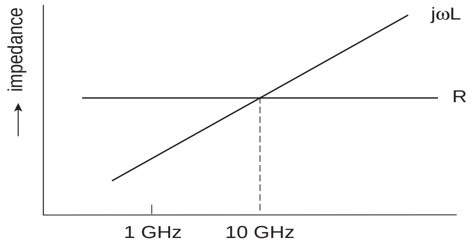


Fig. 28 Resistance and inductance behaviour of on-chip metal lines as a function of the frequency

For each package, the value of *L* can vary per pin due to different lengths of the bond wires (about 10nH/cm) and different bonding structures. A dual-in-line IC package has the largest inductance (2–50nH) per pin, because of the absence of a ground plane (usually) and because of its longer lead lengths. TAB (Taped Automated Bonding) usually results in smaller inductance values (0.5–10nH), because of the presence of a ground plane and shorter and thicker leads. An example of low inductance values (0.15–1nH) is flip-chip technology, in which the chip is mounted upside down with its solder bumps directly attached to the substrate.

The increase in total chip capacitance, combined with faster logic and driving circuits, causes very high values of dI/dt. Currently, supply current changes of a hundred milli-amperes to even amperes per nanosecond are no exceptions. These huge current peaks may cause different problems. The first problem may be the contribution to the supply noise. If a dI/dt of 100mA/ns is to be supplied by one bonding wire (with a length of 1cm and a self inductance L = 10nH/cm), then the associated supply or ground bounce is equal to:

$$\Delta V = L \cdot \frac{\mathrm{d}I}{\mathrm{d}t} = 1\,\mathrm{V} \tag{22}$$

This effective voltage drop is determined by the number of nodes that switch simultaneously. It temporarily reduces the effective supply voltage and it introduces additional delays in the connected logic circuits, leading to slower circuit operation. Secondly, the current peaks can cause glitches. In circuits which are connected to different supply domains, supply noise in one domain may lead to glitches in the other, because this supply noise is super imposed on the communication signals between these circuits [Bakoglu, 1990]. The third problem that may occur from large current peaks, particularly when they flow through the bond wires and the package and board leads, is electromagnetic interference (EMI). Because these wires all act as antennae, they are susceptible to, or can generate EMI, which can dramatically effect neighbouring electronic circuits and systems.

There exist several design measures to reduce the LdI/dt. In this particular section, however, the discussions will be limited to the use of on-chip de-coupling capacitors, and more specifically on the implementation of the capacitor itself.

De-coupling capacitors are charged during steady state while during periods of increased activity, they act as a temporary power supply. Board de-coupling capacitors are already in use for a couple of decades. They are placed very close to the supply pads of the power-intensive chips to reduce the voltage fluctuations that result from the on-board self-inductance of the board wiring. However, on-chip generated current transients, as described above, cannot be reduced by putting de-coupling capacitors next to the IC package on the board. Since they appear only for a very short time, these on-chip current peaks need the charge instantly and the necessary charge will be collected locally, if available. If not, they will cause the described voltage fluctuations. As a result of this, the additional de-coupling capacitors are currently integrated within the standard-cell blocks on a chip. The amount of the required additional de-coupling capacitance depends on the technology (switching speed), the application (frequency) and the functionality of the specific logic block (switching activity). Today's high performance microprocessors, for example, may have several hundreds of nano-farads of on-chip de-coupling capacitance, occupying several tens of square millimetres of chip area [Anderson, 2001; Jain, 2001]. The amount of de-coupling capacitance in an average ASIC standard-cell block varies from 5 to 15% of its area.

Currently, different de-coupling capacitors are used on a chip. Gate-oxide is the thinnest dielectric used in a CMOS production process. Its thickness is only a few nano-metres and, compared to the dielectrics used in between successive metal layers, its capacitance value per unit of area is about 50 to 100 times larger. For this reason, on-chip de-coupling capacitors are implemented, based on this thin gate oxide. Fig. 29 shows the two different implementations: the transistor gate-capacitor cell and the tie-off cell. The most important restriction on both cells is that they have to be placed in between library cells within a standard-cell block and must thus be standard cells themselves.

Both cells show different properties with respect to their capacitance value per unit of area, their efficiency with respect to the capability to deal with fast voltage spikes and their reliability with respect to ESD.

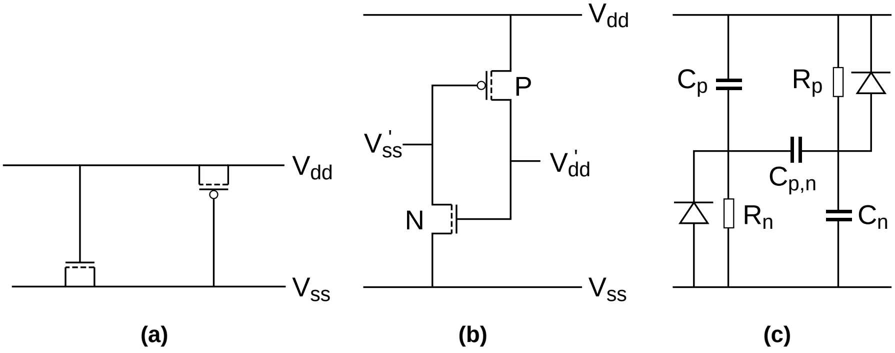


Fig. 29 Different de-coupling capacitors
(a) Transistor de-coupling cell
(b) The tie-off cell
(c) The equivalent circuit diagram of the tie-off cell

The first cell (Fig. 29a) consists of both an n-type and p-type transistor. The nMOS transistor has its gate connected to the V_{dd}, while its source and drain are both connected to ground. The connections to the pMOS transistor are complementary to the nMOST, in that its gate is connected to ground and its source and drain to the V_{dd}. In this cell, the full supply voltage is right across the thin gate oxides of both transistors, having only a very low resistance of the connection in series. Whenever this cell is placed within a logic block that is positioned close to the supply pads of a chip, it is susceptible to damage by electrostatic discharge (ESD). During ESD tests, pulses up to 2000V are also being applied between the supply pads (V_{dd} and V_{ss}). Although all supply pads include ESD protection, it can not always be prevented that part of the ESD pulse might reach such a de-coupling capacitor and damage it by a gate-oxide breakdown.

Actually, for reliability reasons, it is usually not allowed to connect any transistor gate directly to the V_{dd} or V_{ss} supply lines. The cell in Fig. 29b [veendrick, 2002] shows better capabilities in this respect. After power-up of the chip, this 'tie-off' cell generates a dummy V_{dd} and V_{ss} on its internal nodes. For certain reasons it might be required that a transistor gate must be connected to V_{dd} or V_{ss} to permanently switch-off a pMOST or nMOST, respectively. If this is the case, it is not allowed, as described above, to directly connect them to the supply lines. Instead, these transistors have to be connected to a dummy V'_{dd} or V'_{ss}, such as generated by the tie-off cell in Fig. 29b. This is the reason for its name. However, it turned out that this cell can be used as a very good de-coupling capacitor itself. Since the gate of the nMOST is at V_{dd} level, while its source and drain are both at V_{ss} level, the gate capacitor is completely charged. The same holds for the pMOST, since it terminals are at the complementary voltage levels. In other words, both transistors in this cell are completely on and we can make full use of both of their capacitances.

From the equivalent circuit diagram of Fig. 29c, it can also be seen, that the gate capacitance C_n of the nMOST is charged (and discharged during a power dip) through the resistor R_p, which represents the channel resistance of the pMOST. Complementary to this, the gate capacitance C_p of the pMOST is charged (and discharged during a power dip) through the channel resistor R_n of the nMOST. In this way this cell has much better ESD reliability properties in that the transistor gates are only indirectly connected to the supply lines. A disadvantage of this cell is the timing requirements on both *RC*-products. Since a potential power dip in an average today's IC lasts about 1 to 3ns, both R_pC_n and R_nC_p products may not be larger than about one tenth of the supply dip duration, in order to deliver the cell's de-coupling charge in time. This restriction puts some serious demands on the implementation and the layout of the transistors. The sizes of both transistors are expected to be large, so as to create a large de-coupling capacitance for the cell. However, each transistor also acts as a resistor to (dis)charge the capacitance of the other and since this resistance must be small, its channel length can not be too large. The cell of Fig. 29a does not have this restriction, since it has only a marginal resistance of the connections in series with its de-coupling capacitors. Fig. 30 shows a layout example of the tie-off de-coupling capacitor cell. Because the mobility of holes in a pMOST is less than that of an nMOST, the pMOST consists of two parallel transistor in order to reduce its channel length to achieve the required small R_pC_n-product as discussed above.

Since it is expected that Moore's law is still valid for at least a couple of generations, more capacitance needs to be (dis)charged in a shorter time. This requires an increase of additional on-chip de-coupling capacitance and drives the search for a continuous improve in the efficiency of capacitor cells, both in area as well as in its capacitance and its *RC*-time constant, while maintaining its reliability at a sufficiently high level. Not only the *RC*-time constant of the cell is important, also the increase in on-chip capacitance may cause ringing in the supply network. Therefore, for the supply network must have sufficient resistance to damp potential ringing effects.

Other solutions to reduce peak currents may include a more evenly distributed switching activity over the clock period.

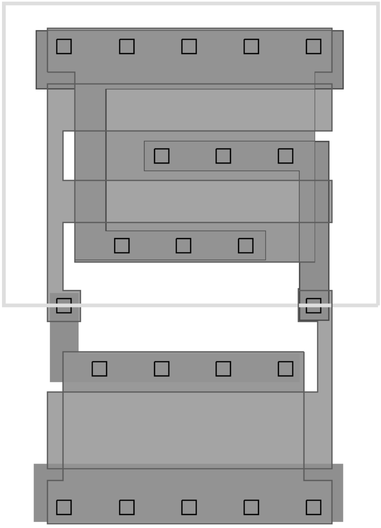


Fig. 30 Possible layout implementation of the tie-off de-coupling capacitor cell

3.3.3 Cross-talk noise and signal propagation

There are two important topics that affect the signal behaviour across the interconnect: cross-talk noise and signal propagation. Both effects are dominated by the mutual capacitance between neighbouring signal lines.

Cross-talk

Cross-talk is an effect that causes one signal to influence another one by means of capacitive, resistive or inductive coupling. On-chip inductive cross-talk, other than previously discussed, can still be neglected. There are several forms of cross-talk.

Particularly in dynamic CMOS circuits, some nodes may float (no conducting path to either supply or ground) during part of the clock period. These floating (high impedance) nodes are particularly sensitive to interference (cross-talk) from neighbouring signal tracks. This section, however, focuses on the cross-talk between parallel signal lines in general and how it affects the signal propagation across long interconnects.

As stated before (Fig. 28), the self-inductance of on-chip wires becomes important at frequencies above 1GHz. Also transmission-line phenomena such as reflections and characteristic impedance will then become increasingly important [Bakoglu, 1990]. Because this is not yet the case for most of the ICs, we will restrict ourselves to noise (cross-talk) generated by the capacitive coupling between the signal lines. The combined effect of larger chip areas and higher complexity causes both an increase of the current densities and of the length of signal wires. This prevents interconnection layer thickness from being scaled with the same ratio as the minimum feature sizes do in every new process generation. The side wall capacitance (and thus the coupling between two signal lines at minimum spacing) increases rapidly over the process generations. Moreover, the increase of the number of interconnection layers makes the cross-talk problem much more complex.

Fig. 31 shows a comparison of a conventional 1μm CMOS double metal process and the lower three metal levels of a 0.25μm six-metal layer CMOS process. This figure clearly demonstrates the pace of scaling of the interconnections within a time frame of less than 10 years.

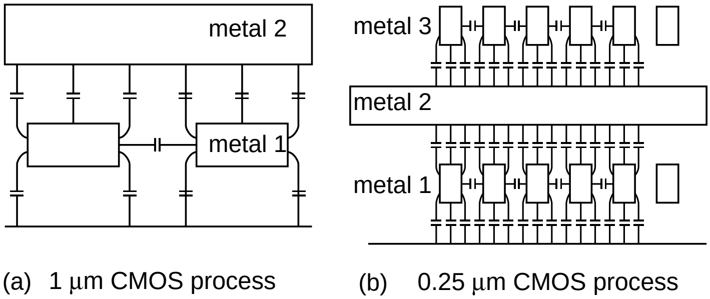


Fig. 31 (a) Capacitances in a double level metal 1μm CMOS process
(b) A 0.25μm CMOS process (only the lower three layers are shown)

Fig. 32a contains a very simple, but representative model, which shows the cross section of two metal tracks in the same layer (metal 2). Capacitance C_m represents the mutual capacitance between the two tracks, while capacitance C_{ground} represents the total capacitance of track M_2 to ground. The equation in the left upper part of the figure expresses the level of cross-talk, in case the signal on M_1 is changing, while M_2 is left floating.
Fig. 32b shows second metal track capacitance values for different technology generations.

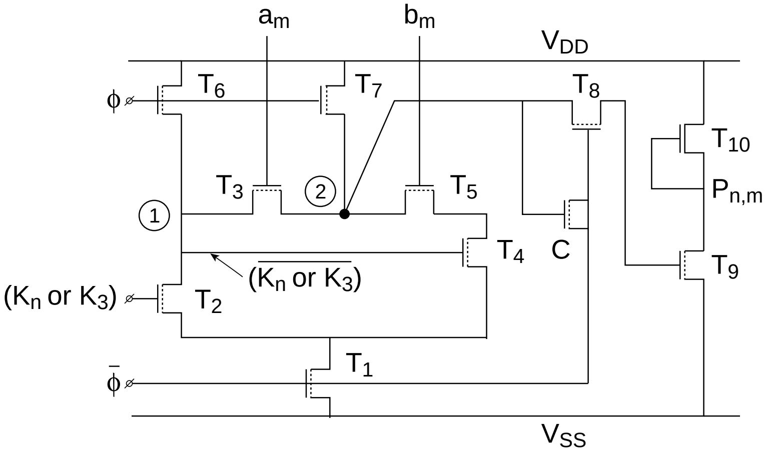


Fig. 32 (a) Simple capacitance model of two metal wires in the same layer (2nd metal)
(b) Capacitance values for 2nd metal in different technologies

The following example shows the severeness of the cross-talk phenomenon in a 0.25μm process. Representative values for these capacitances per mm length are: C_m = 80fF/mm and C_{ground} = 40fF/mm. If track M_2 is floating, then a voltage swing ΔV_1 on track M_1 would generate a noise pulse ΔV_2 on M_2 equal to:

$$\Delta V_2 = \frac{80}{120} \cdot \Delta V_1$$

With a nominal supply voltage of 2.5V and a maximum voltage swing ΔV_1 of 2.5V on M1, the noise pulse on M2 (the victim wire) will be 1.6V. If this victim wire would have two simultaneously switching neighbours in stead of one, the noise level would almost be as high as 2V.
Because this is unacceptable, we can conclude that tri-state buses in (signal) processor ICs or pre-charged bit lines (in memories) are extremely susceptible to cross-

talk. Attention should be paid to ‘floating lines’, pre-charged lines and tri-state buses, etc., not only with respect to their logic levels (which will be corrupted by the high level of cross-talk) but also with respect to reliability.

If, for instance, one line (a in Fig. 33) is (pre-)charged high and if after some time its neighbour also switches to high level, the first line may reach a voltage level (in a 0.25μm CMOS process) of 2.5 + 1.6 (cross-talk) = 4.1V. If this voltage comes across a channel of a single transistor, hot-carrier effects may result in a change of the threshold voltage and in degradation of transistor reliability.

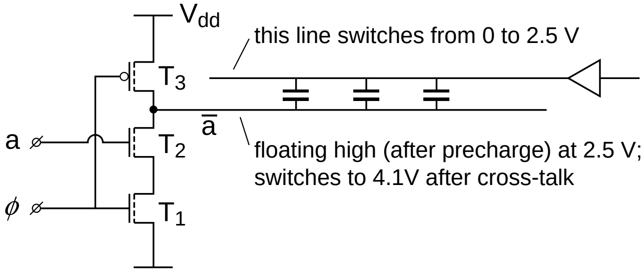

V_{dd}
this line switches from 0 to 2.5 V
T_3
$\bar{a}$
a
T_2
floating high (after precharge) at 2.5 V; switches to 4.1V after cross-talk
ϕ
T_1

Fig. 33 Dynamic circuit with temporary floating output

This voltage will, however, be reduced rapidly by the pMOS transistor and by the p^+n forward biased junction of the pMOS transistor.

Although many circuits do not have floating signal lines, cross-talk still might corrupt signal levels, particularly when long tracks are involved.

Fig. 34 shows a model for the simulation of the cross-talk between two 20mm long second-metal wires at a minimum spacing in a 0.25μm CMOS process. These wires are modelled by 20 stages, each consisting of two resistors *R*, one coupling capacitor C_c and two track capacitors C_p.

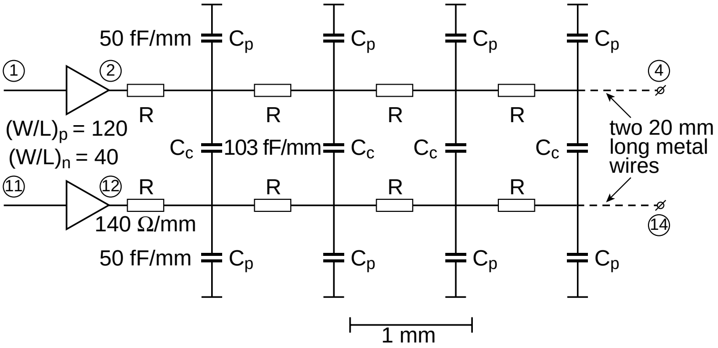


Fig. 34 Model for cross-talk simulation between two parallel wires

The results of a circuit simulation of the above model, in which 1mm wire section is represented by one resistor-capacitor stage, are shown in Fig. 35.

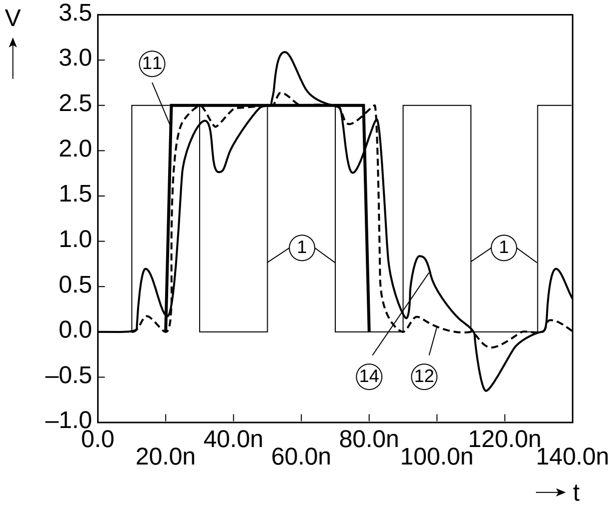


Fig. 35 Mutual cross-talk between two long metal lines in a 0.25μm process

The figure shows cross-talk noise pulses with amplitudes at wire ends (node 14) of about 0.65 to 0.85V. These values exceed the threshold voltage levels and the reader can imagine the effects when everything is even somewhat worse or when two neighbouring lines switch simultaneously. In this case, the in-between line may show cross-talk noise at the end of the wire that exceeds 1.3V!

As signal wire lengths on VLSI chips of 10 to 20mm are no exception, the cross-talk may have dramatic effects on the proper behaviour of the circuit.

In conclusion, long signal lines behave almost like tri-state buses at wire ends and are therefore very susceptible to cross-talk.

It is clear that deep-sub-micron designs will exhibit severe cross-talk problems if its prevention is not part of technology development, design style and design flow. Low-voltage swing signalling or increasing the space between critical interconnects are among the solutions that can reduce the effects of cross-talk. Large designs in deep-submicron technologies therefore require tools to manage the cross-talk problem. Some tools have a cross-talk feature: they need characteristic technology input. This allows them to detect nodes inside logic blocks that show too much cross-talk. Another feature in this tool is that it can re-route the critical nodes, either by placing the tracks further apart, or by placing them in different layers, to reduce the mutual capacitance [Choudhury, 1995]. Beyond 0.25μm feature sizes, metal heights, oxide thickness and dielectric constants are expected to scale according to the SIA roadmap, see Table 4.

Table 4 SIA interconnection

Technology	250nm	180nm	150nm	130nm	100nm	70nm
Metal height (nm)	450	324	300	273	240	189
Minimal metal spacing (nm)	340	240	210	170	140	100
Oxide thickness (nm)	792	572	504	450	378	290
Relative dielectric constant	3.0–4.1	2.5–3.0	2.0–2.5	1.5–2.0	1.5–2.0	<1.5

If technology developments keep pace with this roadmap, the cross-talk will almost remain constant, as shown in Fig. 36.

$$\Delta V_{victim} = \frac{2\cdot C_{lateral}}{C_{vertical} + 2\cdot C_{lateral}}\cdot\Delta V_{source}$$

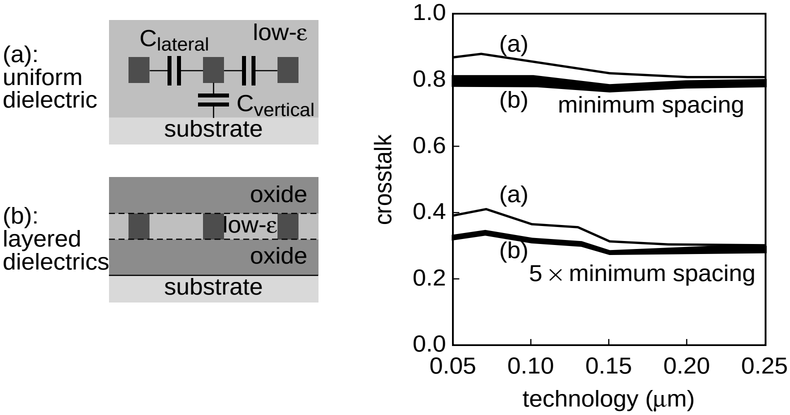


Fig. 36 Cross-talk trends according to capacitance values predicted in SIA roadmap

The figure shows the cross-talk, as a fraction of the voltage swing, for a uniform, low-ε dielectric and for layered dielectrics. It also shows the reduction in cross-talk if the minimum spacing increases by a factor of five, for example. However, there are signs that the low-ε dielectric materials can be integrated into the CMOS processes at a much slower pace than expected.

Signal propagation

Signal propagation was thought to be a problem in conservative one-metal layer processes. In these processes, the crossing of data lines and clock lines was often

implemented via a polysilicon bridge with a resistance of $R_{ps} = 40\Omega/\square$. Whenever the data lines were long, this could lead to relatively large delays (tens of nanoseconds) and a dramatic reduction of the performance or the malfunctioning of the chip.
Current processes offer several metal layers (Al, Cu) which have resistances of 30–70m$\Omega/\square$, depending on the layer thickness. In these processes, the *RC*-delay is not a problem for most of the signals, because they are routed over average distances and have an average load (number of inputs to which they are connected). However, there are signal lines such as clock lines, scan control lines and data buses which may run all over the chip to provide global control or communication. As discussed, such wire lengths may exceed some tens of millimetres or even hundreds in the case of clock signals.

Particularly in the case of buses, signal lines are completely embedded. Next to the fact that their behaviour interferes with each other, leading to cross-talk as discussed in the previous subsection, they also affect each others signal propagation delay. Fig. 37a shows a victim line, which is embedded within two aggressors. The design has been made in a 0.18μm technology, with minimum line widths and minimum spacings between the victim line and its aggressors.
In the following we distinguish three cases of operation:

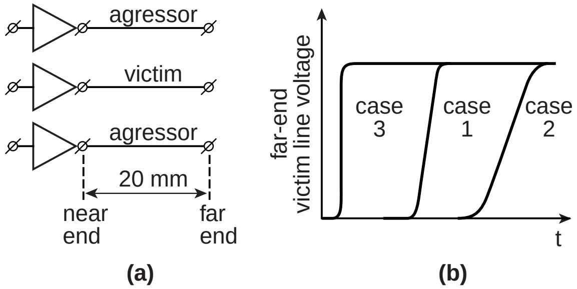


Fig. 37 (a) Model for signal propagation in buses
(b) Far-end victim line signal for different cases

Case 1: Victim switches from low to high while aggressors remain low

In this case, the victim line shows a capacitance to ground and to its two neighbours, as modelled in Fig. 34 by capacitors C_p and C_c, respectively. The far-end victim line signal is represented by case 1 in Fig. 37b.

Case 2: Victim switches from low to high while aggressors switch opposite

Now the victim line shows about the same bottom capacitance but it looks like the mutual capacitance to both of its neighbours has doubled. Due to the minimum spacing between the signal lines, these mutual capacitances are much larger than the bottom capacitance. Case 2 in Fig. 37b represents the far-end victim line signal, which shows about twice the propagation delay compared to case 1.

Case 3: Both victim and aggressors switch from low to high

Since the neighbours switch in the same direction, the victim line almost only shows a bottom capacitance and therefore the total propagation delay to its far end has dramatically reduced with respect to both previous cases.

Fig. 38 shows a photograph of the signals from Fig. 37b, measured on a 0.18μm CMOS chip.

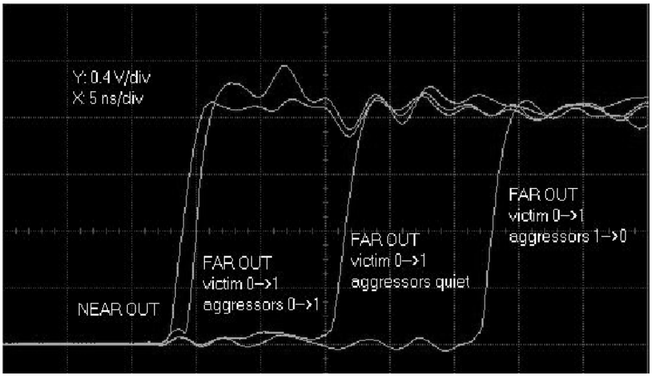


Fig. 38 Photograph of cross-talk measurement in 0.18μm CMOS technology (corresponding to Fig. 37

The difference in signal propagation between the worst case (opposite switching) and best case (same switching) is about a factor of ten! Since this effect is not yet included in all the design tools, it can lead to severe timing problems and it is the designers responsibility to include accurate wire models and signal propagation in the overall design simulation and verification phase. Further information on future signal propagation trends is also presented in the next chapter on scaling trends.

Even with the use of other metals for interconnection (copper) and the use of low-ε dielectrics from 0.18μm technologies onwards, the purely interconnect *RC*-delay is still expected to increase according to the ITRS roadmap, as shown in Fig. 39. In summary, it is advisable that signal lines, which exceed a certain critical length and load, are automatically detected at the end of the design cycle. The designer is then able to check whether the corresponding *RC*-delay is critical. At frequencies below 1GHz, the self-inductance of on-chip wires hardly influences the propagation delay (Fig. 28) and is therefore neglected here.

Due to the increasing interconnect propagation delay, design styles have to be changed as will also be discussed in the next chapter.

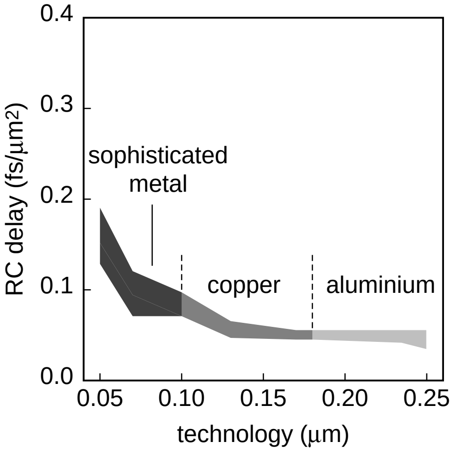


Fig. 39 RC-delay scaling of interconnection layers according to the ITRS roadmap

3.3.4 Conclusions

The continuous decrease of the physical sizes of transistors and interconnect have introduced a lot of so-called deep-submicron effects . Particularly supply noise and cross-talk may dramatically reduce the overall signal integrity in a design.

Currently CAD vendors introduce signal integrity tools for cross-talk, power-grid analysis and even substrate noise. These will have to become part of the design flow to maintain design robustness at a sufficiently high level.

Chapter 4

Effects of scaling on MOS IC design and consequences for performance and robustness

If we continue to increase the complexity of ICs at the same pace as we have done since 1960, we will have reached a level of a billion transistors per chip within this decade. Moreover, the clock period is ‘expected’ to be well below a hundred picoseconds. Even if we do not believe these excessive numbers, design styles and methods have to be changed to fully exploit the potentials of IC complexity and speed, as predicted by the Semiconductor Industrial Association (SIA) roadmap (or International Technology Roadmap for Semiconductors [SIA, 2001].

This chapter discusses the consequences of the scaling process for deep-submicron IC design, with the focus on future trends of power, speed and robustness. The increasing dominance of physical effects that create interference and noise in VLSI designs requires more and more analogue measures to limit their influence. In the race towards a 1 Giga transistor (1Gtor) heterogeneous System On a Chip (SOC), see Fig. 40, design styles have to be changed to make the design manageable (system design aspects) and to make a functional design (physical design aspects).

The complexity of such a SOC cannot be managed with traditional design concepts and requires:

- a platform with integrated hardware/software architecture and application development tools
- system level synthesis to improve design efficiency
- design reuse
- increased design resources per chip.

The first three items deal with system level design aspects. The increased design resources, however, are not only required to manage the SOC design complexity, they are also needed to:

- develop testing, debugging and diagnosing concepts to enhance testability and observability when using IP cores
- manage clock synchronisation in different clock domains
- limit on-/off-chip noise and interference, and support EMC.

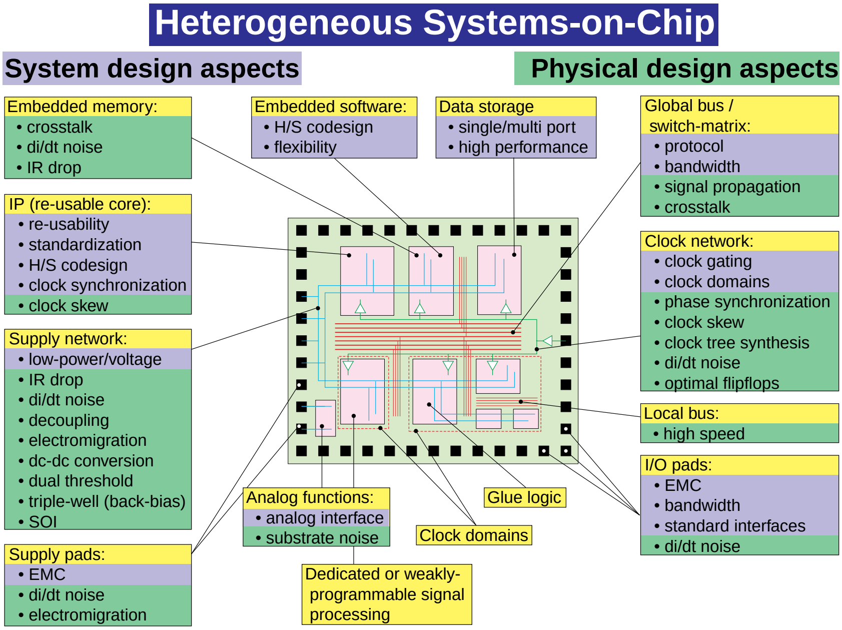


Fig. 40 Important aspects of a (heterogeneous) System On a Chip

Particularly these last two items are heavily influenced by the backend part of the process (the metal layers). Previously, only analogue circuits were susceptible to these physical effects. In future process generations, these effects will dominate the SOCs performance and robustness, while some of these effects are already threatening the performance of complex VLSI chips. Future VLSI design therefore requires a more analogue approach. Design is no longer about switches and ones and zeros only, but also about resistors, capacitors, inductors, noise, interference and radiation. This requires an increased level of details during simulation and verification, which will take an increasing part of the total design time.

Basically, a VLSI chip is just a bunch of transistors that perform a certain function by the way they are interconnected. The next sections focus on the influence of scaling on the basic elements. First the effects of scaling on transistor performance and reliability will be presented. The interconnect has become an important factor in the overall performance and signal integrity of an IC. This dominance will even further increase as a result of further scaling. Finally, the scaling consequences for overall performance, density and robustness of deep-submicron IC designs are discussed. The results are summarised in a table, which represents the overall constant-voltage scaling effects.

This chapter is the result of a lot of research and is included as a preview of future challenges related to '*getting the most out of the MOST*'.

4.1 Transistor scaling effects

When scaling transistor sizes and bias voltages by a factor of s ($s \approx 0.7$), the transistor current scales by the same factor. To maintain performance, the threshold voltage is also required to scale with s. The threshold-dependent leakage current is an important factor that limits the pace of scaling. This sub-threshold leakage current can be estimated by the following, which means that this current increases by about a factor of 12 for every 100mV decrease in V_T:

$$I_{\text{sub-threshold}}(\text{scaled}) = 12^{10(1-s)V_T} \cdot I_{\text{sub-threshold}}$$

For large SOCs, this background leakage current will be higher than the current from a gate oxide short, for example, which will dramatically limit the potentials of I_{ddq} testing. The eventual reduction of the threshold voltage requires alternative techniques, such as the Dual-V_T concept [Izumikawa, 1997] and the Triple-well concept [Kuroda, 1996], to limit the sub-threshold power consumption during standby and test modes.

To maintain a high on-current in the transistors, the gate oxide thickness (t_{ox}) is also required to be scaled with s. Below a t_{ox} of about 2.5nm [Lo, 1997], quantum-mechanical tunnelling of charge through the gate oxide may occur, resulting in additional standby currents and possibly a reliability problem. It is expected that this gate leakage current surpasses the sub-threshold leakage in the 70nm technology node. As long as the semiconductor industry is incapable to integrate high-ε gate dielectrics into future CMOS technologies, the gate leakage problem may become one of the biggest problems of this decade, for which there exist no design or technology solutions yet.

Finally, the continuous scaling of the channel length will increase mismatching of ‘equal’ transistors, as a result of the increased spread of the number of dopants in the transistor channel. Matching of transistors means the extend to which two identical transistors, both in type, size and layout topology show equal device parameters, such as β and V_t. Particularly in analogue circuits (a memory is also an analogue circuit) where transistors are required to have a very high level of matching [Pelgrom, 1989; Vertregt, 1999], the spread in V_t due to doping statistics in the channel of the MOS transistors results in inaccurate or even anomalous circuit behaviour. For minimum transistor sizes (area), this effect increases every new IC process generation, such that both the scaling of the physical size and the operating voltage of analogue CMOS circuits lag one or two generations behind the digital CMOS circuits. Also for digital (logic) CMOS circuits, matching of transistors is becoming an important issue, resulting in different propagation delays of identical logic circuits. Fig. 41 shows two identical inverter chains (e.g. in a clock tree), but due to the spread in V_t, they show different arrival times of the signals at their output nodes. For circuits in a 0.1μm technology this time difference is in the order of several gate delays, depending on the depth of the logic chain. Particularly for high-speed circuits, for which timing is a critical issue, transistor matching and its modelling is of extreme importance to maintain design robustness at a sufficiently high level.

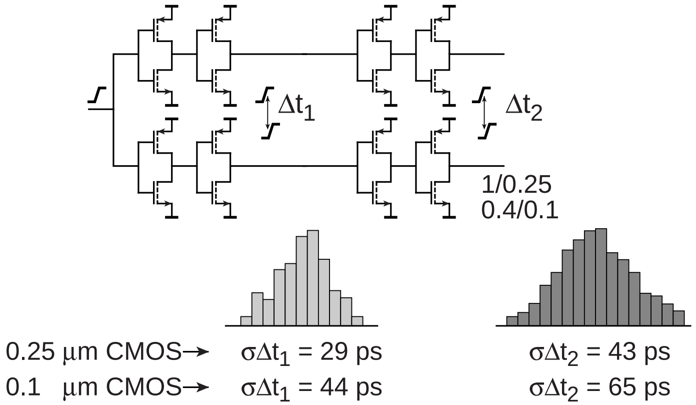


Fig. 41 Spread in signal arrival times in a clock tree, due to transistor mismatch

To reduce the timing effects of transistor mismatch in drivers in a clock tree, the transistor sizes in these circuits can be increased analogous to making analogue circuits less sensitive to process parameter spread.

4.2 Interconnect scaling effects

Scaling of widths and spacings has caused the metal interconnections to start dominating the ICs performance, reliability and signal integrity. The output load of a logic gate is equal to the total of the fan-in capacitances of its connecting gates and the total wire load of the interconnections. Table 5 shows the increase in the average ratio between wire load and fan-in, for standard cell blocks, caused by scaling. These numbers represent average values; for each individual chip, this ratio may be different from the table.

Table 5 Increasing interconnect dominance

Technology	Ratio: wire load/fan-in
0.35μm	50/50
0.25μm	58/42
0.18μm	66/34
0.13μm	75/25
0.1μm	80/20

Higher interconnection resistance leads to larger voltage drops. Larger mutual capacitances lead to increased cross-talk, while the combination leads to larger signal propagation delays. At 1GHz, the required rise and fall times should be less than 50ps to perform some computational tasks within the available 1ns time frame. Even on-chip wires then cause interference with other modules. For such signal edges, line lengths of 3mm and above become critical and require transmission line modelling. Fig. 42 shows the signal propagation delay across an embedded metal track (between two other metal tracks at minimum spacing) in different technologies. This has been simulated using the model from Fig. 3, but then extended with another neighbour.

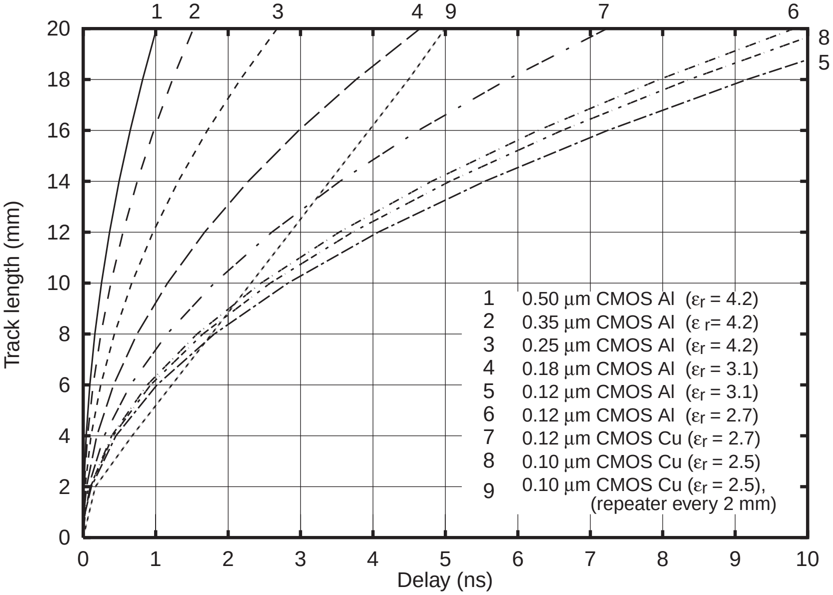


Fig. 42 Propagation delay of an embedded track in different technologies

There are several approaches to reduce the negative effects of scaled interconnections. One is to reduce the capacitance, which is expressed as $C = \varepsilon_0\varepsilon_r A/t_{\mathrm{ox}}$. Current values for the dielectric coefficient ε_r are between 3 and 4. In the SIA roadmap for the year 2007, values around $\varepsilon_r \approx 2$ are expected, leading to a capacitance reduction of a factor of 2. The second improvement we can make is to reduce the resistance. The sheet resistance of conventional aluminium alloys is around 3$\mu\Omega$cm, while that of copper is about 1.8$\mu\Omega$cm. However, the potentials of the reduced copper resistance cannot fully be exploited. Because copper diffuses through oxides, it cannot be deposited and etched like aluminium. By applying a damascene back-end flow, copper can be completely encapsulated within a barrier material, as shown in Fig. 43.

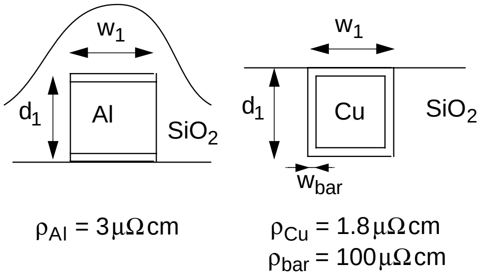
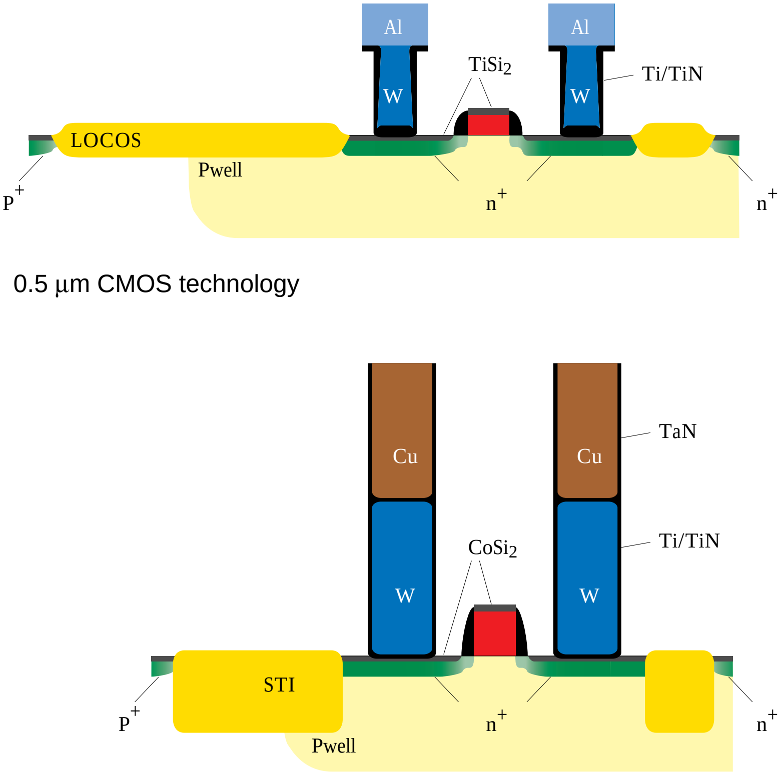


Fig. 43 Basic differences between the formation of aluminium and copper interconnections

The effective sheet resistance of copper wiring depends on the barrier material and will reach values close to 2.2$\mu\Omega$cm. The advantage of this lower resistance will be used to reduce the track thickness, because a reduction of the mutual track capacitance is preferred to a reduction of the resistance. The total track capacitance is then reduced by a factor of about 1.25. Thus, the combination of copper with low-ε dielectrics may reduce the total track capacitance and propagation delay by a factor of about 2.5 at the maximum. Since the power consumed by the charging and discharging of a metal interconnection is proportional to the capacitance, this power will reduce by the same factor as well. The use of copper and low-ε dielectrics will help only for about two generations. Fig. 42 also shows the individual influence of copper and low-ε dielectrics. The signal propagation delay over a metal wire is proportional to the square of its length. The use of repeaters, however, reduces the propagation delay to a linear dependency on length. Particularly for longer wires, this may improve the signal propagation by more than a factor of two (curves 8 and 9). Yet, another problem is dawning on the horizon. In the 65nm technology node, the local copper interconnect wires are getting so narrow, that the electrons suffer from surface scattering, which is expected to almost double the effective wire resistance.

The increasing clock skew and signal propagation delay for across global wires are in direct contrast to the reducing clock period. Therefore, there will be an increased drive to limit the size of the standard cells blocks to about 100K gates, which will also limit local interconnect lengths and clock skew. Designs will therefore become globally asynchronous and locally (within blocks) synchronous (GALS). To further relief the propagation delay problems, pipelines could be built into the global interconnects, but the bus latency will then become an important design parameter.

Fig. 44 shows cross-sections of a 0.5μm and a 0.1μm transistor. Both transistors are drawn to different scales, so that the channel lengths in the drawing are equal. The figure clearly demonstrates the increased dominance of the interconnect in future deep-sub-micron processes.



0.5 µm CMOS technology

0.1 µm CMOS technology

Fig. 44 Cross-sections of a 0.5μm and a 0.1μm transistor with normalised channel lengths

4.3 Scaling consequences for overall IC design

If we integrate complete boards on one single chip, not only the functionality of the board is included, but also common board problems, such as supply noise, interference and EMI. In this section, we discuss the scaling consequences for overall chip performance, design reliability and signal integrity. As the transistor bias voltages scale with the same factor as the transistor dimensions, the scaling from 0.35μm processes onwards is called constant-field scaling. Table 6 shows the effects of constant-field scaling on the overall chip properties. These depend on the scale factor (s) between successive process generations, being about equal to $s \approx 0.7$.

Table 6 Effects of scaling on transistor and chip characteristics

Property	Effect	Scale factor dependence
basic parameters		
electrical field strength		1
device demensions ($W, L, t_{ox},$ junction depth)		s
transistor area (logic gate area) ($W \cdot L$)		s^2
capacitances per unit area (C_u)		$1/s$
capacitances ($C = W \cdot L \cdot C_u$)		s
bias voltage (V) and V_T		s
body effect (K- factor)		$\sqrt{s}$
bias currents ($I = \mu C_{ox} \frac{W}{L} (V_{gs} - V_T)^2$)		s
performance		
packing density (logic gates per unit area)	complexity	$1/s^2$
number of pins	bandwidth	$1/s$
gate delay ($T_{min} = C \cdot V / I$)	speed	s
power dissipation/gate ($C \cdot V^2 \cdot f_{max}$)	power	s^2
power delay product ($I \cdot V \cdot T_{min}$)	efficiency	s^3
reliability		
ESD susceptibility ($1/ t_{ox}$)	ESD	$1/s$
current density ($I / area$)	electromigration	$1/s$
power density ($I \cdot V / area$)	heating	1
latch-up sensitivity ($V > 1.2V$)	latch-up	s
hot-carrier degradation ($1.5V < V < 3.0V$)	hot carrier	~s
integrity		
sub-threshold current	leakage	$12^{10(1\text{-}s)V_T}$
noise margins (V_{dd}, V_T)	signal integrity	s
signal interference	mutual cross-talk	1/s
current density slew rate ($\frac{1}{area} \cdot \frac{dI}{dt}$)	EMC	$1/s^2$
$L \cdot \frac{dI}{dt}$ noise density ($\frac{dI}{dt} / \# pins$)	supply bounce	$1/s$
ΔV per unit wire length ($I \cdot R / V$)	voltage drop	$1/s$
α-particle sensitivity	radiation	$1/s^2$

The relation between a property and such parameters as *V*, *I*, *C*, *t*, *W*, *L* and t_{ox} are also given in the first column of the table. The scale factor dependence for each of the properties is given in the third column and is the result of calculating these relations. The following subsections discuss these relations and are particularly focussed on performance, density and robustness.

4.3.1 Scaling consequences for overall chip performance and density

Until the late nineties (1997/1998 for volume production), transistor bias voltages were not scaled with each process generation. This was called constant-voltage scaling. During this conventional scaling era, speed and power efficiency scaled quite differently to the current constant-field scaling process. Fig. 45 shows the improvements in speed and power efficiency, based on the evolution during the past two decades and the expected evolution in the next decade.

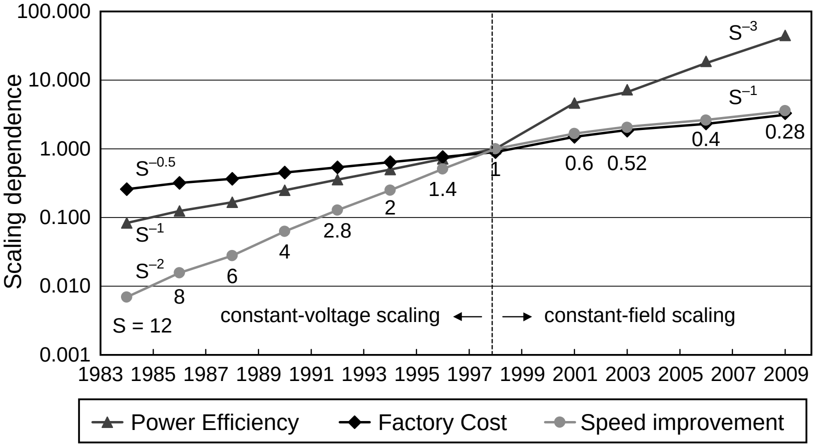


Fig. 45 Relative change in scaling dependencies of power efficiency, speed and factory costs, with respect to the year 1997/1998

Until recently, the costs of building and equipping a factory increased by a factor of 1.5 every three years [Schaller, 1997]. However, because of the rapidly increasing demands on the accuracy of the equipment and the class of the clean room, the factory costs are expected to double every factory generation. (About a three-year period, twice the average period between successive process generations.)

As can be concluded from this figure, we might state that, until recently, chip size reduction and speed improvement were the drivers behind scaling. However, now, with constant-field scaling, power efficiency (i.e. computing power per watt) has become the main driver behind the scaling process. Finally, constant-field scaling is threatening mixed analogue/digital designs, which are becoming increasingly challenging, and become questionable in performance and cost efficiency below 1.5V.

4.3.2 Scaling consequences for overall design robustness

Robustness of an integrated circuit deals on the one hand with the electrical design tolerance and on the other hand with the physical design tolerance. For associated scaling dependencies, please refer to Table 6. The effect of scaling on signal integrity has already been discussed in Chapter 3. This section therefore focuses only on the new challenges that dawn on the horizon.

Electrostatic discharge (ESD)

As a result of the scaling of the gate-oxide thickness, the input transistor sensitivity for ESD will increase by a factor of 1/*s*. Because the protection must sink current to limit or clamp voltages to a certain level, power will be consumed inside the protection circuits. To limit power density, protection circuits cannot be scaled similarly to functional circuits. They will either occupy relatively larger areas or they will require novel structures.

Soft-errors

Soft-errors may be caused by charge particles (α-particles or cosmic particles), which may hit an IC in one or more electronic circuits, such as a logic or memory cell. Radio-active particles (α-particles) are high-mass particles which origin from radio-active decay from material (e.g. chip, package, solder). They can create a lot of electron-hole pairs along their track. Cosmic particles are high-energy particles from outer space. They can fracture silicon nucleus, causing liberation of a large number of electron-hole pairs. Both charges can be captured by capacitors and may flip states in memory cells and flipflops. Since the cosmic particles may easily propagate through more than 1.5m of concrete [Cataldo, 2001], it is very difficult, if not impossible, to protect ICs from this type of radiation. However, α-particles occur at a much higher rate and are ‘easier’ to control. On the one hand, the selection of proper materials in IC packages, in IC processes and during bonding will reduce α-particle radiation. On the other hand, measures can be taken in the design to lower its sensitivity to α-particles. Also silicon can be coated with a blocking layer, most commonly 30μm polyimide, which is used as a shield to effectively reduce the number of α-particle hits.

If an output node of a logic gate in a digital logic block is hit by an α-particle, the logic value at that node may only temporarily be destroyed. This is because the input signals to that logic gate are maintained, thereby regenerating the original output state again after a very short time. This will hardly result in an operating failure. However, if the diffusion area of an output node of a latch (or flipflop) is hit by such a particle, this might cause a permanent change of state of that latch. The stored charge on an output node of a minimum sized cross-coupled latch in a 0.25μm CMOS technology is roughly 2 to 10fC.

This amount of critical charge to disturb proper circuit operation is much less than the maximum charge (130fC), which an α-particle can deposit on a single node of a present-day device [Dennard, 1997]. Therefore, it is not a question of whether the latch will change its state when it is hit by an α-particle, but what is the chance of being hit and causing a so-called soft-error. This chance is relatively low, because the critical latch area (output diffusion area) is much smaller than the total latch area. Also, the number of latches in a standard cell block is relatively low, compared to the many memory cells in a large memory chip. α-particle induced soft-errors are therefore becoming a growing thread, particularly to the robustness of memory intensive systems. However, with scaling, the soft-error rate increases by a factor $1/s^2$, as the number of memory cells and flipflops per mm^2 increases with that factor. Also in logic circuits, the soft-error rate is expected to increase rapidly, since.

Fig. 46 summarises the trends in design robustness based on further scaling.

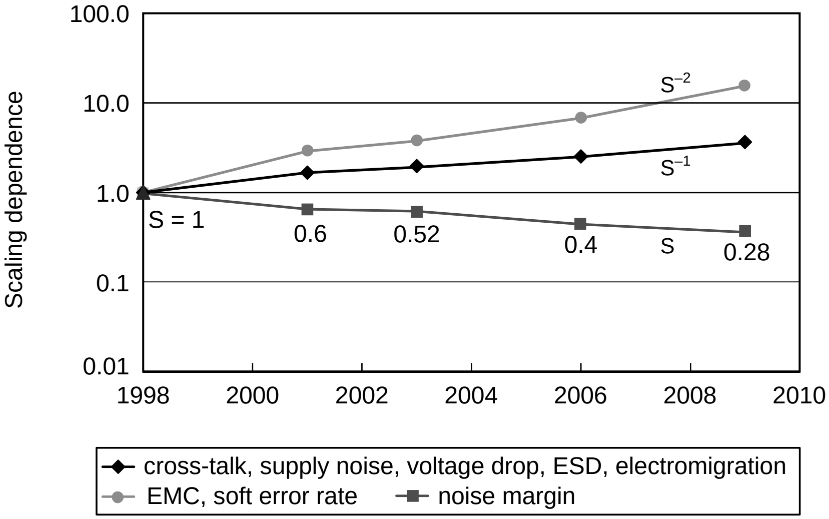


Fig. 46 Scaling dependencies of robustness parameters

4.4 Potential future design challenges

The observation of Gordon Moore's law (a quadrupling of IC complexity every three years) has proven its validity from the invention of the chip until now. It is sometimes called a self-fulfilling prophecy, and is viewed as a measure for future trends and sets the pace of innovation. Almost according to this law, the Semiconductor Industrial Association has set up its roadmap for the next couple of years. Table 7 shows the most important design parameters of this roadmap [SIA, 1999].

The previously-discussed scaling trends show that there are many key factors that may limit the pace of scaling. The first one is the design complexity of CMOS ICs, which increases exponentially with time, as can be seen from the table. However, the complexity of the design and test tasks is accelerated, and forms a potential barrier to obtaining full exploitation of the available manufacture potentials. The overall success of the semiconductor industry will be increasingly dominated by how the complex design, engineering and test challenges will be addressed [SIA, 1999]:

- managing design databases
- complexity of design flows
- hierarchy of synthesis
- integration of reusable cores
- on-chip communication and interfaces
- complexity of timing modelling (clocks, interconnections)
- design effort (design teams)
- noise and reliability
- number of redesigns
- complexity of failure analysis
- costs of testing (test time, hardware and software).

Reuse of IP cores is a must to improve design efficiency and a challenge to integration. By the end of this decade we will have arrived at the one billion transistor System on Chip. Even if we assume that 80% of all transistors is in a memory, we still have 200 million transistors in the logic blocks. With a maximum number of about 100K gate equivalents per logic block, we still need to integrate 50 different blocks. A possible solution to the complexity problem could be to introduce a higher degree of regularity at all levels of design. It may therefore be attractive to design the standard-cell blocks with a kind of high-level standard-block approach, e.g. in which all blocks have the same height. This also allows a more regular and simpler supply grid across the chip. Further standardisation of these blocks regarding design, test, timing and interfacing is a prerequisite to a successful concept of reuse.

The ability to completely verify, test, debug and diagnose future complex designs will reduce dramatically. The increased costs of a mask set, which is expected to double each new technology node, no longer allows design iterations or redesigns. It is therefore likely that current design styles with fixed and dedicated logic will (partly) be replaced by design styles that allow flexibility and (re)configurability. This flexibility can be enhanced by software solutions (programmability) as well as hardware solutions (*configurable computation* such as embedded *FPGA* and/or sea-of-gates architectures). Remaining bugs can then be bypassed by changing the program or by re-mapping the function, respectively.

As discussed in Chapter 3, the increased manifestation of physical and electrical effects in deep-sub-micron technologies will bring the circuit noise to unacceptable levels. In addition to this, the noise margins of future processes will further decrease due to the reduction of the supply and threshold voltages (Fig. 47) [Veendrick, 2000]. Every new technology requires additional design and/or technology measures to reduce the noise and increase the gap between the noise and the noise margin. However after scaling to the next technology, the problem is the same again and new measures are required. Relatively large additional chip areas must be devoted to on-chip measures like de-coupling capacitances and to more widely-spaced buses and other global signal interconnections etc.

The increasing amount of devices on a chip combined with decreasing feature sizes leads to a growing sensitivity for certain defect mechanisms as shorts, opens, stuck-at-one, stuck-at-zero and bridging and cause permanent faults, which reduce the yield. Due to the decreasing noise margins, circuits become increasingly sensitive to such deep-submicron effects as cross-talk, supply noise, voltage drop and soft-errors, which cause temporary faults. Future design approaches must therefore include fault tolerance, redundancy and/or self-repair techniques as a way to maintain robustness and yield in order to keep scaling commercially attractive.

All design measures that need to be taken to cope with the above effects, reduce the chance of fully exploiting the potentials of the new process generations. The level to which these effects will limit the efficient use of chip area cannot be predicted because it also depends on the creative design alternatives that will be developed in the near future.

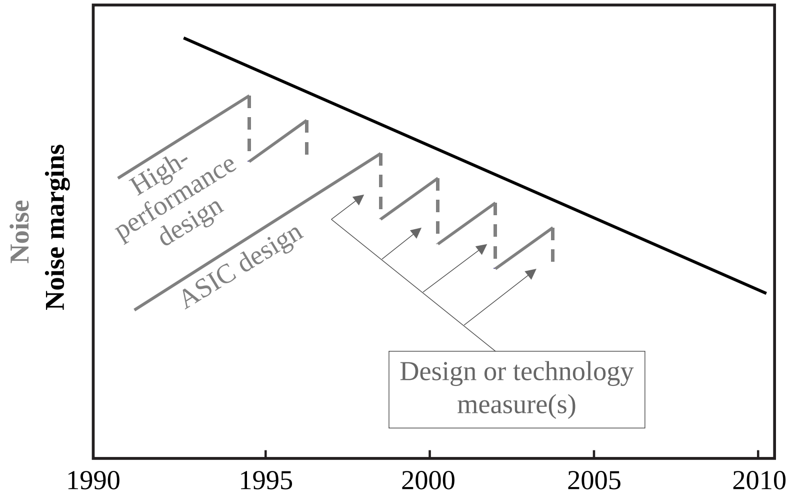
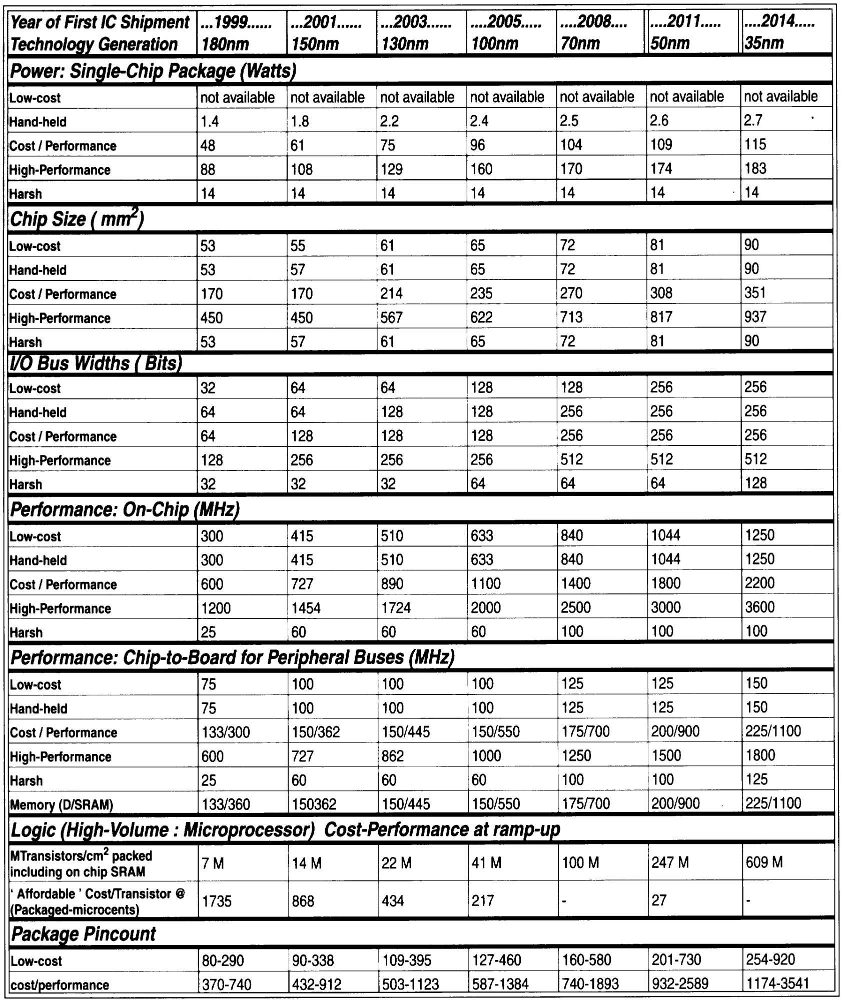


Fig. 47 The reducing gap between noise and noise margins

Table 7 Most important IC characteristics and their change according to SIA (ITRS) roadmap



Year of First IC Shipment	*...1999......*	*...2001......*	*...2003......*	*....2005.....*	*....2008....*	*....2011.....*	*....2014.....*
Technology Generation	*180nm*	*150nm*	*130nm*	*100nm*	*70nm*	*50nm*	*35nm*
Power: Single-Chip Package (Watts)							
Low-cost	not available	not available	not available	not available	not available	not available	not available
Hand-held	1.4	1.8	2.2	2.4	2.5	2.6	2.7
Cost / Performance	48	61	75	96	104	109	115
High-Performance	88	108	129	160	170	174	183
Harsh	14	14	14	14	14	14	14
Chip Size (mm^2)							
Low-cost	53	55	61	65	72	81	90
Hand-held	53	57	61	65	72	81	90
Cost / Performance	170	170	214	235	270	308	351
High-Performance	450	450	567	622	713	817	937
Harsh	53	57	61	65	72	81	90
I/O Bus Widths (Bits)							
Low-cost	32	64	64	128	128	256	256
Hand-held	64	64	128	128	256	256	256
Cost / Performance	64	128	128	128	256	256	256
High-Performance	128	256	256	256	512	512	512
Harsh	32	32	32	64	64	64	128
Performance: On-Chip (MHz)							
Low-cost	300	415	510	633	840	1044	1250
Hand-held	300	415	510	633	840	1044	1250
Cost / Performance	600	727	890	1100	1400	1800	2200
High-Performance	1200	1454	1724	2000	2500	3000	3600
Harsh	25	60	60	60	100	100	100
Performance: Chip-to-Board for Peripheral Buses (MHz)							
Low-cost	75	100	100	100	125	125	150
Hand-held	75	100	100	100	125	125	150
Cost / Performance	133/300	150/362	150/445	150/550	175/700	200/900	225/1100
High-Performance	600	727	862	1000	1250	1500	1800
Harsh	25	60	60	60	100	100	125
Memory (D/SRAM)	133/360	150362	150/445	150/550	175/700	200/900	225/1100
Logic (High-Volume : Microprocessor) Cost-Performance at ramp-up							
MTransistors/cm^2 packed including on chip SRAM	7 M	14 M	22 M	41 M	100 M	247 M	609 M
' Affordable ' Cost/Transistor @ (Packaged-microcents)	1735	868	434	217	-	27	-
Package Pincount							
Low-cost	80-290	90-338	109-395	127-460	160-580	201-730	254-920
cost/performance	370-740	432-912	503-1123	587-1384	740-1893	932-2589	1174-3541

In this SIA roadmap, the following definitions are used for the different IC categories:

- Low-cost < $300 consumer products, micro-controllers, disk drives, displays
- Hand-held < $1000 battery-powered products; mobile products, hand-held cellular and other hand-held devices
- Cost-performance < $3000 notebooks, desktop personal computers, telecommunications
- High-performance > $3000 high-end workstations, servers, avionics, supercomputers, most demanding requirements
- Harsh under-the-hood and other hostile environments

-

References

This relatively short reference list closes *Part I*. Each detailed paper in *Part II* has its own reference list.

- [Anderson, 2001] 'Physical Design of a Fourth-Generation Power GHz Microprocessor', *ISSCC Digest of Technical Papers*, February 2001, p. 232–233
- [Bakoglu, 1990] 'Circuits, Interconnections, and Packaging for VLSI', *Addison-Wesley*, 1990
- [Banerjee, 1999] 'On Thermal Effects in Deep-Submicron VLSI Interconnects', *Digest DAC conference*, p. 885-891,1999
- [Banerjee, 2001] 'Analysis and Optimization of Thermal Issues in High-Performance VLSI', *Digest ISPD Symposium*, p. 230-237, 2001
- [Berkhoff, 1983] 'Applications of Picture Memories in Television Receivers', *IEEE Trans. on Consumer Electronics,* August 1983, p. 251–255
- [Cataldo, 2001] 'SRAM soft errors cause hard network problems', *EETIMES.com: http://www.eetimes.com/story/OEG20010817S0073*
- [Chaney, 1973] 'Anomalous behaviour of synchronizer and arbiter circuits', *IEEE Trans. Comput.*, April 1973, p. 421–422
- [Chen, 1992] 'A reconfigurable Multiprocessor IC for Rapid Prototyping of Real-Time Data Paths', *ISSCC Digest of Technical Papers*, February 1992, p. 74–75
- [Choudhury, 1995] 'A 300MHz CMOS microprocessor with Multi-Media', *ISSCC Digest of Technical Papers*, 1995, p. 170–171
- [Dennard, 1997]'Future CMOS scaling: approaching the limits?', *Future Fab International, 1997*
- [Duchene, 1989] 'A highly flexible sea-of-gates structure for digital and analog applications', *IEEE J. Solid-State Circuits*, vol. 24, June 1989, p. 576–584
- [Edelstein, 1997] 'Full Copper Wiring in a sub-0.25μm CMOS ULSI Technology', IEDM 1997, *Digest of Technical Papers*, p. 773
- [Fisher, 1982] 'What is the Impact of Digital TV?', *IEEE Trans. on Consumer Electronics*, August 1982, p. 423–429
- [Gray, 1996] 'Analysis and Design of Analog Integrated Circuits', *Third Packaging Edition, John Wiley & Sons,*1996
- [Hunter, 1995], *IEDM Digest*, p. 483-486, 1995
- [Izumikawa, 1997] 'A 0.25μm 0.9V 100MHz DSP core', *IEEE-JSSC*, January 1997, p. 52–61
- [Jain, 2001] 'A 1.2GHz Alpha Microprocessor with 44.8GB/s Chip Pin Bandwidth', *ISSCC Digest of Technical Papers,* February 2001, p. 240–241
- [Kuroda, 1996] 'A 0.9V, 150MHz, 10mW, 4mm^2, 2-D DCT Core processor with variable Threshold Voltage Scheme', *IEEE-JSSC*, November 1996, p. 1770–1779
- [Lo, 1997] 'Quantum-Mechanical Modeling of Electron Tunneling Current in MOSFETs', *IEEE Electron Device Letters*, Vol. 18, No 5, 1997, pp 209-211
- [Mavor, 1983] 'Introduction to MOS LSI DESIGN', *Addison-Wesley,* 1983

- Mead, et al. ‘Introduction to VLSI systems’, *Addison-Wesley*, 1980
- [Nvidea] ‘RIVA 128ZX Processor’, *www.nvidea.com/products/frames_riva 128zx.html*
- [Ohkura, 1982] ‘Gate-isolation – A novel basic cell configuration for CMOS gate arrays’, *proceedings CICC*, 1982, p. 307–310
- [Pechoucek, 1976] ‘Anomalous response times of input synchronizers’, *IEEE Trans. Comput.*, February 1976, p. 133–139
- [Pelgrom, 1989] ‘Matching properties of MOS transistors’, *IEEE Journal of Solid-State Circuits,* vol.24, October 1989, p. 1433-1440
- [Roizen, 1986] ‘Dubrovnik impasse puts high-definition TV on hold’, *IEEE Spectrum,* September 1986, p. 32–37
- [Schaller, 1997] ‘MOORE’s LAW: past, present and future’, *IEEE Spectrum*, June 1997, p. 53–59
- [SIA, 1999] ‘The International Technology Roadmap for Semiconductors’, 1999-edition
- [Slotboom, 1981] ‘Leakage Current in High-Density CCD Memory Structures’, *IEDM Digest of Technical Papers*, 1981
- [R.Streiter, 1998] ‘Application of Combined Thermal and Electrical Simulation for Optimization of Deep Submicron Interconnect Systems‘ *www.infotech.tu-chemnitz.de/~zfm/pdf/semi01_pr.pdf*
- [Takahashi, 1985] ‘A 240k transistor CMOS array with flexible allocation of memories and channels’, *ISSCC Digest of Technical Papers*, February 1985, p. 124–125
- [Veendrick, 2000] ‘Deep-Submicron CMOS ICs, from Basics to ASICs’, *Kluwer Academic Publishers, Dordrecht (NL) London Boston*, 2000, chapter4
- [Veendrick, 2002] ‘Noise in Digital Circuits @ Low Voltage’, Invited paper in special low-voltage session. *International Solid-State Circuits Conference, San Francisco*
- [Vertregt, 1999]‘Embedded Analog Technology’, *IEDM short-course on ‘System-on-a-Chip Technology’*, December 5, 1999
- [Wong, 1986] ‘A high-performance 129k gate CMOS array’, *Proc. CICC*, 1986, p. 568–571

Part II

Detailed scientific work

Chapter 1.1A

An nMOS Dual-Mode Digital Low-Pass Filter for Colour TV

Harry J.M. Veendrick

IEEE Journal of Solid-State Circuits

June 1981, Volume SC-16, Number 3

Abstract

Implementation of digital colour television, to perform the total video processing digitally, requires several building blocks. This chapter deals with one of them: a low-pass filter for separation of the luminance and chrominance spectra. After a short introduction, showing the place and the function of the filter, and a few words on the implementation of the filter on a chip, the circuit design is discussed in some detail. The chip, which runs at frequencies up to 40MHz, has been designed with two-phase race-compensated MOS logic in a 4μm E/D nMOS technology with implanted under-crossings.

1 Introduction

To implement colour television in a digital way, several building blocks are needed to perform the total video processing digitally.

Because of its linear phase capability, a non-recursive digital low-pass filter is particularly suited for separation of the luminance and chrominance spectra.

In this introduction, the place and the function of this low-pass filter will be discussed.

The spectrum of a normal black-and-white video signal ranges from 0 to 5MHz (Fig. 1). Transmission of the colour information in the PAL system takes place as a quadrature modulation of the carrier frequency (4.43MHz) with two band-limited (≈ 1MHz) signals. The spectrum of the colour information lies between approximately 3 and 5MHz and is added to the black-and-white signal. This addition causes interference between the two spectra. The cross-talk into the black-and-white signal is called cross-luminance, the reverse one cross-colour. Sophisticated manners of separating these two spectra can greatly reduce this mutual cross-talk.

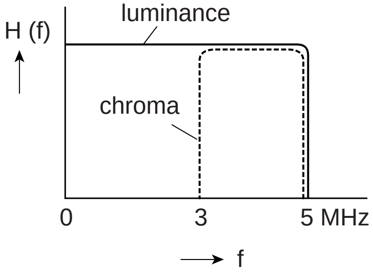


Fig. 1 Video band

The exact choice of the sub-carrier frequency is such that its phase shifts 90° from line to line (Fig. 2). By combining picture elements with a mutual delay of two television lines (2 T_1) or a frame delay (312 T_1) – that is to say the ones which have the colour information in the opposite phase –, it is possible to separate the two spectra simply by adding and subtracting these mutually delayed samples.

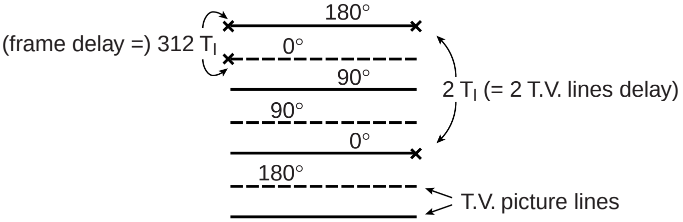


Fig. 2 Phase of the carrier in a television picture

The maximum vertical spatial frequency is obtained when the black-and-white information changes every line, the first line for instance being black, the second white, the third black, etc. It is clear now that if we combine picture elements [1]–[4] with a mutual delay of two television lines, the black-and-white changes in between are not detected. Combining picture elements with a mutual delay thus means filtering in the vertical direction. This results in a loss in vertical resolution. To prevent this loss in vertical resolution for the complete luminance band, comb filtering – which is in fact adding and subtracting mutually delayed elements – has only to be applied in the chrominance band (Fig. 3).

This requires a band-splitting function. In Fig. 3 the low-pass filter performs this in combination with a subtractor.

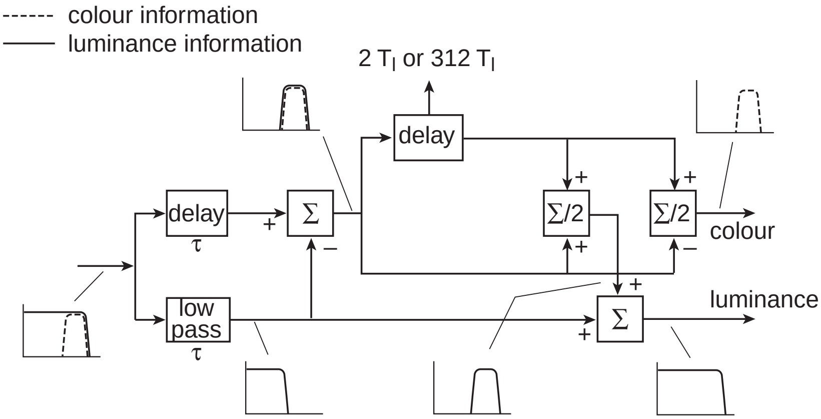


Fig. 3 Block diagram of digital separation of the luminance and chrominance band

Combining points with a mutual delay of two television lines results in a greater loss in vertical resolution than in the case of a frame delay (312 TV lines). Both situations, therefore, require different filters. If a delay of two television lines is used, a low-pass filter with a narrower transition band is required than in the case of a frame delay.

Fig. 4 shows the frequency response of both filters. Curve *a* refers to the two television lines delay case, curve *b* to the frame delay case. The corresponding impulse responses are given in Fig. 5.

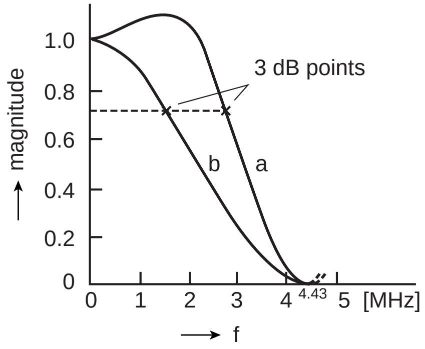


Fig. 4 Frequency response of both filters

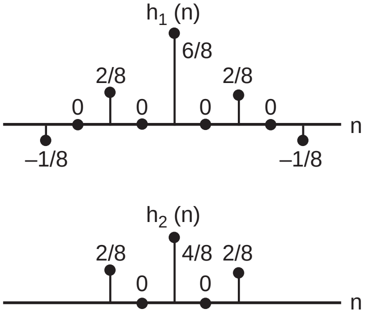


Fig. 5 Impulse response of both filters

The coefficients, which are a compromise between hardware and transition band, are determined by means of a computer program in multiples of powers of two to make the implementation easier [5]. Thus, a nine-coefficient version exists with a narrow transition band and a five-coefficient version with a somewhat wider transition band. Both filter versions are implemented in a dual mode chip and can be selected via an extra pin.

2 Implementation

The structure of the nine-coefficient linear transversal filter as shown in Fig. 6.

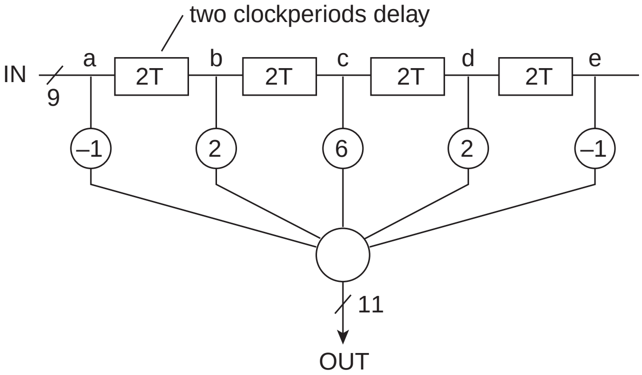


Fig. 6 Structure of the nine-coefficient linear transversal filter

Because every second coefficient is zero, a two-period delay is used between the nonzero coefficients. As can be seen, the coefficients have been multiplied by eight, which will be corrected at the output by shifting the binary point three-bit positions to the left. For input and output a nine- and eleven-bit representation is chosen, respectively, to meet accuracy requirements. As the input may also have negative values, the two’s complement code is used. Fig. 7a shows the bit representations that have to be added. For proper two’s complement addition we have to expand each bit pattern with the most significant bit to thirteen bits.

Multiplying by minus one (–*a* and –*e*) in the two’s complement code means inverting the corresponding bit pattern and adding a ‘1’ at the least significant bit position. Because this happens twice, we simply add a ‘1’ at the second bit position (Fig. 7a: ‘extra bit’). Multiplying by two (2*b*, 2*c*, and 2*d*) is done by shifting the corresponding bit pattern one bit to the left. Multiplying by four means a two-bit shift, etc. Thus, implementing these filter coefficients is just a matter of interconnecting.

Changing the filter function into the five-coefficient version (Fig. 5b) is easily done by forcing all bit positions, that corresponds to the values –*a*, 2*c*, and –*e* (Fig. 7a) to zero. For this filter the resulting bit patterns that have to be added are shown in Fig. 7b.

$$\begin{array}{r l l l}
 & \bar{a}_8\ \bar{a}_8\ \bar{a}_8\ \bar{a}_8\ \bar{a}_8\ \bar{a}_7\ \bar{a}_6\ \bar{a}_5\ \bar{a}_4\ \bar{a}_3\ \bar{a}_2\ \bar{a}_1\ \bar{a}_0 & \longleftarrow & -a \\
 & b_8\ b_8\ b_8\ b_8\ b_8\ b_7\ b_6\ b_5\ b_4\ b_3\ b_2\ b_1\ 0 & \longleftarrow & 2b \\
 & c_8\ c_8\ c_8\ c_8\ c_8\ c_7\ c_6\ c_5\ c_4\ c_3\ c_2\ c_1\ 0 & \longleftarrow & \left.\begin{array}{l} 2c \\ 4c \end{array}\right] 6c \\
 & c_8\ c_8\ c_8\ c_8\ c_8\ c_7\ c_6\ c_5\ c_4\ c_3\ c_2\ 1\ 0 & \longleftarrow & \\
 & d_8\ d_8\ d_8\ d_8\ d_8\ d_7\ d_6\ d_5\ d_4\ d_3\ d_2\ d_1\ 0 & \longleftarrow & 2d \\
 + & \bar{e}_8\ \bar{e}_8\ \bar{e}_8\ \bar{e}_8\ \bar{e}_8\ \bar{e}_7\ \bar{e}_6\ \bar{e}_5\ \bar{e}_4\ \bar{e}_3\ \bar{e}_2\ \bar{e}_1\ \bar{e}_0 & \longleftarrow & -e \\
\hline
 & .\ \ .\ \ .\ \ .\ \ .\ \ .\ \ .\ \ .\ \ .\ \ .\ \ ,.\ \ .\ \ .\ \ . & \longleftarrow & \text{out}
\end{array}$$

extra bit

a

$$\begin{array}{r l l l}
 & b_8\ b_8\ b_8\ b_8\ b_8\ b_7\ b_6\ b_5\ b_4\ b_3\ b_2\ b_1\ 0 & \longleftarrow & 2b \\
 & c_8\ c_8\ c_8\ c_8\ c_8\ c_7\ c_6\ c_5\ c_4\ c_3\ c_2\ 0\ 0 & \longleftarrow & 4c \\
 + & d_8\ d_8\ d_8\ d_8\ d_8\ d_7\ d_6\ d_5\ d_4\ d_3\ d_2\ d_1\ 0 & \longleftarrow & 2d \\
\hline
 & .\ \ .\ \ .\ \ .\ \ .\ \ .\ \ .\ \ .\ \ .\ \ .\ \ ,.\ \ .\ \ .\ \ . & \longleftarrow & \text{out}
\end{array}$$

b

Fig. 7 (a) Bit representation of the signal values that have to be added to obtain the filter of Fig. 5a
(b) Bit representation of the signal values that have to be added to obtain the five-coefficient filter (Fig. 5b)

Realisation of the filters now consists of a shift register per bit position, representing the delays, and a pipeline array of full adders and latches to perform the addition. Fig. 8 shows a microphotograph of the chip, which has been designed with two-phase 'race-compensated' MOS logic in 4μm E/D technology with implanted under-crossings.

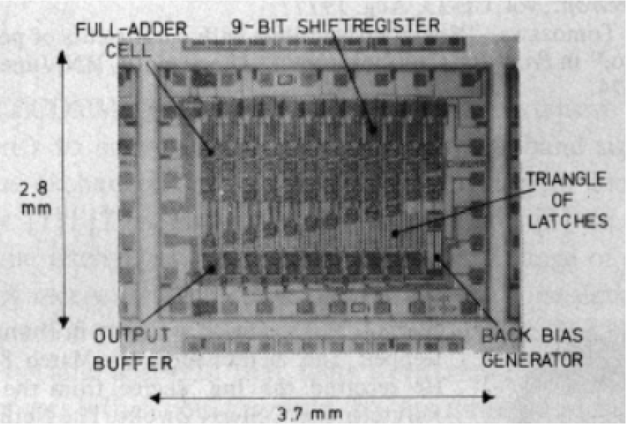


Fig. 8 Photograph of the chip

3 Circuit Design – 'Race-Compensated' MOS Logic

As mentioned, the addition is done with a pipeline array of full adders and latches. For this purpose a new full adder cell has been designed which can perform the full adder function (sum as well as carry) in half a clock period. This full adder cell is designed with two-phase 'race-compensated' MOS logic, which will be discussed based on the electrical circuit of the sum function of the full adder (Fig. 9).

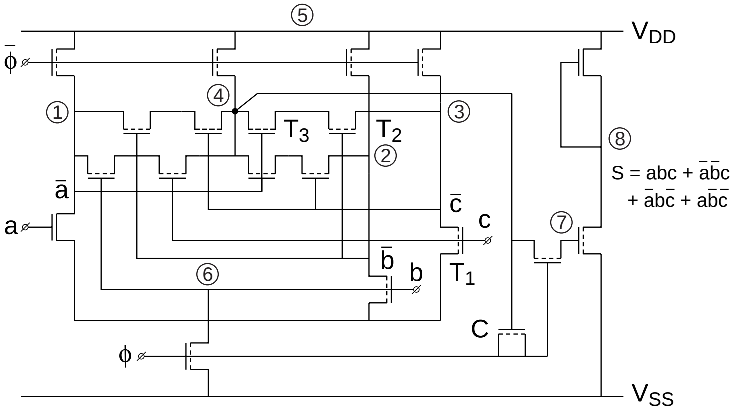


Fig. 9 Electrical circuit for the sum function of the full adder

During the clock $\overline{\Phi}$, nodes ①, ②, ③, and ④ are pre-charged. During Φ, the data at the input nodes *a*, *b*, and *c* are sampled and at the same moment the inverse data $\bar{a}, \bar{b}$ and $\bar{c}$ are generated and used to control other gates. Under certain input conditions this introduces a race by causing a voltage drop at node ④ – just after clock Φ is switched on – at a moment that it is not allowed. For instance, if the inputs *a* and *c* are at high level and input *b* is at low level, then transistors T_1 and T_2 must conduct while T_3 should not. However, if the sample clock Φ is switched on, it will take some time before the gate of transistor T_3 is at low level. This means that transistor T_3 is conducting during a very short time, which causes a conducting path of short duration from node ④ through transistors T_3, T_2, and T_1 to ground. This results in a voltage drop at node ④. For this reason the bootstrap capacitor *C* is used to compensate any voltage drop at node ④ caused by charge sharing or by a race.

During the pre-charge moment, node ④ is at high level and clock Φ at low level. Thus, capacitance *C* is charged. A possible race occurs right after clock Φ is switched on, which would cause a voltage drop at node ④, if it were not that at the same time this voltage drop is compensated via the capacitor *C*, the value of which depends on the magnitude of the voltage drop.

Normally, when a logic function uses the direct inputs and their inverse ones, it takes two clock periods to perform this function. Using race-compensated MOS logic means that the inverse data is made and used to control other gates at the same clock period. This results in a hardware reduction and in a decrease in time delay.

At the same time during the sample clock Φ (Fig. 9), the information at node ④ is stored on the gate of the output latch. The output of the full adder is used as an input to another full adder, which has Φ as pre-charge and $\overline{\Phi}$ as sample clock.

Thus, during one clock period the data ripple through two full adder layers. The least significant bit will be ready first. Clocked carry propagation occurs in diagonally arranged full adders. To arrange that for each bit position the output data will be available at the same time, a triangle of latches is needed.

A summary of filter characteristics is given in Fig. 10. The chip runs at frequencies up to 40MHz, but will be used in a system with a system clock of 17.7MHz, which is equal to four times the sub-carrier frequency. The internal delay from input to output equals seven clock periods, but in video processing this does not cause difficulties, as the total video signal is only shifted over 0.4μs in time.

– 3 dB points:	filter 1 (nine coef's)	: 2.8 MHz *
	filter 2 (five coef's)	:1.6 MHz *
– zero point (both filters)		: 4.43 MHz *
– clock frequency (max)		: 40 MHz
– chip size		: 10 mm^2
– dissipation (5V power supply)		: 300 mW *
– internal delay		: 0.4 msec *
– two external clocks	(20 MHz)	: 5 V
	(40 MHz)	: 7 V

* at 177 MHz clocks

Fig. 10 Summary of filter characteristics

4 Technology

The chip has been designed in 4μm E/D nMOS technology with implanted undercrossings.

A cross-section of a transistor and an under-crossing is given in Fig. 11. An under-crossing is marked by an extra mask and consists of an arsenic implant on top of which LOCOS (*LOC*al *O*xidation of *S*ilicon) is grown.

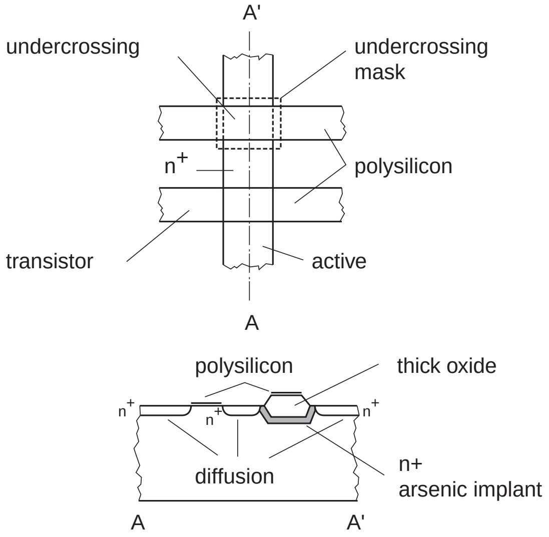


Fig. 11 Cross-sectional view of a transistor and an under-crossing

Use of under-crossings means the availability of an extra interconnection layer. Particularly for circuits with many crossings and interconnections (PLAs, full adder arrays, etc.), a designer can use under-crossings to great advantage.

5 Conclusion

A dual-mode digital low-pass filter for use in digital colour television equipment has been implemented in a chip, which runs at clock frequencies up to 40MHz.
A new full adder cell has been designed in two-phase race-compensated MOS logic to perform the full adder function with less hardware in only half a clock period.

6 Acknowledgement

The author wishes to thank N.F. Benschop and J.G. Raven for many stimulating discussions and P. Welles for determination of the filter coefficients.

7 References

[1] J.P. Rossi, ‘Digital T.V. comb filter with adaptive features’, in *Proc. IERE Conf. Video and Data Recording*, Birmingham, England, July 20-22, 1976.

[2] A.G. Deczky, ‘Color demodulation of an NTSC television signal using digital filtering techniques’, in *Proc. Int. Conf. Commun.*, San Francisco, CA, June 16-18, 1975.

[3] J.P. Rossi, ‘Digital television image enhancement’, *J. Soc. Motion Picture and Television Engineers*, vol. 84, pp. 545-551, July 1975.

[4] R. Turner, ‘Some thoughts on using comb filters in the broadcast television transmitter and at the receiver’, *IEEE Trans. Consumer Electron*., vol. CE-23, Aug. 1977.

[5] A. Tomozawa, ‘Non-recursive filters with coefficients of powers of two’, in *Proc. Int. Conf. Commun.*, Minneapolis, MN, June 17-19, 1974.

Chapter 1.1B

A 40MHz Multi-applicable Digital Signal Processing Chip

Harry J.M. Veendrick and Leo C. Pfennings

IEEE Journal of Solid-State Circuits

February 1982, Volume SC-17, Number 1

Abstract

A multi-applicable building block capable of performing several typical signal-processing (sub) operations such as addition, multiplication, multiplexing, etc. is discussed. The chip, made in 4μm E/D nMOS technology with implanted under-crossings, runs at clock frequencies up to 40MHz, which is particularly suited for digital video signal processing.

1 Introduction

Every digital signal processor uses (sub) operations such as addition, multiplication, multiplexing, etc., or a combination of a few of them [1].
A multi-applicable building block able to perform several of these operations (see the 'Applications' section) might be of general interest.
A digital processor with the following function can carry out the operations:

$$Z = K \cdot A + L \cdot B$$

where $K = (K'/n)$. K' is an integer ($0 \le K' \le n$), $L = 1 - K$ and n represents the resolution of K.
The output (Z) is an adjustable (K) weighted sum of the two inputs (A and B).
Although this building block is more generally applicable, the 40MHz data rates (inputs A and B as well as the parameter K) make it particularly suited for digital video signal processing.

2 Implementation

The inputs are represented with ten bits in two's complement, allowing negative inputs.
For rounding-off reasons, the output is represented with 11 bits.
A resolution of $\frac{1}{8}$ is chosen for K so, for instance,

$$Z = \frac{1}{8} \cdot A + \frac{7}{8} \cdot B \quad \text{or} \quad Z = \frac{2}{8} \cdot A + \frac{6}{8} \cdot B, \text{etc.}$$

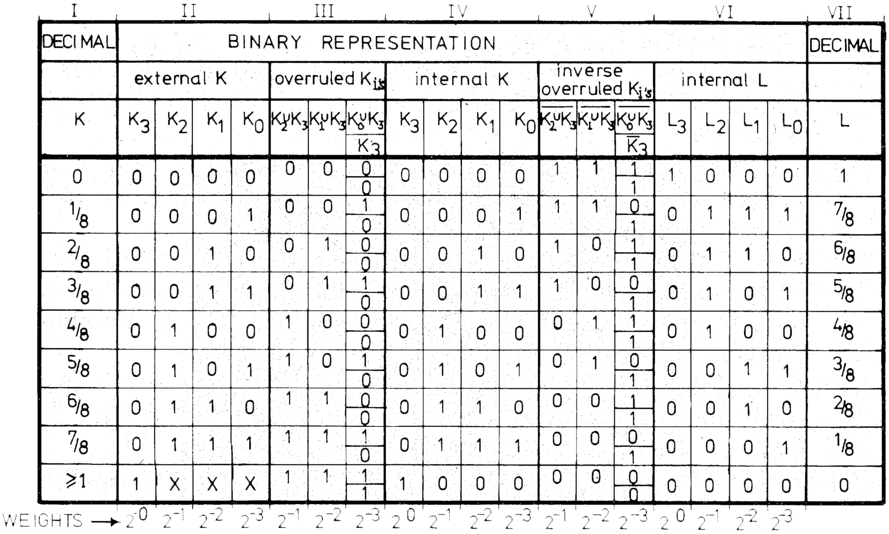
Some applications, however, also require inclusion of both extremes $K = 0$ ($Z = B$) and $K = 1$ ($Z = A$). Consequently, one is forced to a rather inefficient use of four bits to represent the nine possible K values.
Generally, the implementation of a function such as

$$Z = K \cdot A + L \cdot B \tag{1}$$

takes two multipliers, an adder, and a small piece of ROM to store the L weights.

However, the weights K and L (= $1 - K$) are each other's complement, which allows us to reduce the amount of hardware. The procedure of how to generate the internal L (= $1 - K$ [internal]) from the external K can best be explained on the basis of Table 1. With the terms ‘internal’ and ‘external’, we mean internally generated on the chip and externally offered to the chip, respectively.

Table 1



I	II				III			IV				V			VI				VII
DECIMAL	BINARY REPRESENTATION																		DECIMAL
	external K				overruled K_i's			internal K				inverse overruled K_i's			internal L				
K	K_3	K_2	K_1	K_0	$K_2 \cup K_3$	$K_1 \cup K_3$	$K_0 \cup K_3$ K_3	K_3	K_2	K_1	K_0	$\overline{K_2 \cup K_3}$	$\overline{K_1 \cup K_3}$	$\overline{K_0 \cup K_3}$ $\overline{K_3}$	L_3	L_2	L_1	L_0	L
0	0	0	0	0	0	0	0 0	0	0	0	0	1	1	1 1	1	0	0	0	1
1/8	0	0	0	1	0	0	1 0	0	0	0	1	1	1	0 1	0	1	1	1	7/8
2/8	0	0	1	0	0	1	0 0	0	0	1	0	1	0	1 1	0	1	1	0	6/8
3/8	0	0	1	1	0	1	1 0	0	0	1	1	1	0	0 1	0	1	0	1	5/8
4/8	0	1	0	0	1	0	0 0	0	1	0	0	0	1	1 1	0	1	0	0	4/8
5/8	0	1	0	1	1	0	1 0	0	1	0	1	0	1	0 1	0	0	1	1	3/8
6/8	0	1	1	0	1	1	0 0	0	1	1	0	0	0	1 1	0	0	1	0	2/8
7/8	0	1	1	1	1	1	1 0	0	1	1	1	0	0	0 1	0	0	0	1	1/8
⩾1	1	X	X	X	1	1	1 1	1	0	0	0	0	0	0 0	0	0	0	0	0
WEIGHTS →	2^0	2^{-1}	2^{-2}	2^{-3}	2^{-1}	2^{-2}	2^{-3}	2^0	2^{-1}	2^{-2}	2^{-3}	2^{-1}	2^{-2}	2^{-3}	2^0	2^{-1}	2^{-2}	2^{-3}	

The first step is to overrule the three least significant bits of the external K (K_2, K_1, and K_0) by its most significant bit (K_3), the result of which can be represented by $K_2 \cup K_3$, $K_1 \cup K_3$, $K_0 \cup K_3$, respectively (column III).

Next, K_3, is given the weight of the least significant bit position and will be added later on at this position. The result of this addition will be an internal K (column IV), which is equal to the external K (except for the case $K \geq 1$). Upon inversion of the overruled K bits (column III), column V appears. The internal L (column VI) is now generated from column V in the same way as the internal K (column IV) was generated from column III, i.e., by adding $\overline{K_3}$ at the least significant bit position ($\overline{K_0 \cup K_3}$; column V).

The above-discussed procedure is best illustrated by an example. Suppose the external K equal $\frac{3}{8}$ (decimal representation), which means that $K_3 = 0$, $K_2 = 0$, $K_1 = 1$, and $K_0 = 1$ (binary representation). As, in this case, $K_3 = 0$, the overruled three least significant bits ($K_2 \cup K_3$, $K_1 \cup K_3$, and $K_0 \cup K_3$) keep the same value as K_2, K_1, and K_0, respectively.

Next, the addition of K_3 (= 0) at the least significant bit position ($K_0 \cup K_3$; column III) does not affect the values of the three significant bits. The result of this addition is the internal K (column IV).

Now, the inverse of column III is generated (column V): $\overline{K_2 \cup K_3} = 1$, $\overline{K_1 \cup K_3} = 0$, $\overline{K_0 \cup K_3} = 0$, and $\overline{K_3} = 1$. The addition of $\overline{K_3}$ at the least significant bit position of column V completes this procedure, resulting in an L (L_3 = 0, L_2 = 1, L_1 = 0, and L_0 = 1) equal to $\frac{5}{8}$.

The overrule mechanism ($K_2 \cup K_3$, $K_1 \cup K_3$, $K_0 \cup K_3$) only plays a role in case $K_{\text{ext}} \geq 1$ (column I), which means that $K_{3_{\text{ext}}} = 1$. In all cases, the external K_3 equals 1; the three least significant bits are DON'T CARE'S and will be overruled by a 1. So in case the external $K \geq 1$, the internal K equals 1 and the internal L equals 0.

The above-discussed procedure can also be expressed in mathematical formulas:

$$K = K_3 \cdot 2^0 + K_2 \cdot 2^{-1} + K_1 \cdot 2^{-2} + K_0 \cdot 2^{-3}$$

$$K = (K_2 \cup K_3) \cdot 2^{-1} + (K_1 \cup K_3) \cdot 2^{-2} + (K_0 \cup K_3) \cdot 2^{-3} + K_3 \cdot 2^{-3} \qquad (2)$$

and

$$\begin{aligned} L &= 1 - K = L_3 \cdot 2^0 + L_2 \cdot 2^{-1} + L_1 \cdot 2^{-2} + L_0 \cdot 2^{-3} \\ &= \left(\overline{K_2 \cup K_3}\right) \cdot 2^{-1} + \left(\overline{K_1 \cup K_3}\right) \cdot 2^{-2} \left(\overline{K_0 \cup K_3}\right) \cdot 2^{-3} + \overline{K_3} \cdot 2^{-3}. \end{aligned} \qquad (3)$$

It can be seen that the bits of K that have to be added and the bits of L are each other's complement.

This means that implementation of the function $Z = K \cdot A + L \cdot B$ results in a multiplier array whose product bits have the following function:

$$P_{n,m} = (K_n \cup K_3) \cdot a_m + \left(\overline{K_n \cup K_3}\right) \cdot b_m \qquad (4)$$

Consequently, only one multiplier is required because $P_{n,m}$ is, in fact, a switch, since either of the two terms always equals 0. Depending on the input ($K_n \cup K_3$), the output $P_{n,m}$ equals either a_m or b_m.

A block diagram of the processor is given in Fig. 1. The input processing is shown in some detail in Fig. 2. The outputs of the $P_{n,m}$ circuits are connected to a pipeline array of full adders and latches to perform the addition.

As an example, the output (S) of the first full adder at the least significant bit position is given by

$$\begin{aligned} S &= K_3 \cdot a_0 + \overline{K_3} \cdot b_0 + \left(K_0 \cup K_3\right) \cdot a_0 + \left(\overline{K_0 \cup K_3}\right) \cdot b_0 \\ &= \left\{K_3 + \left(K_0 \cup K_3\right)\right\} \cdot a_0 + \left\{\overline{K_3} + \left(\overline{K_0 \cup K_3}\right)\right\} \cdot b_0 \end{aligned} \tag{5}$$

which is, of course, represented by a sum and a carry. This shows how the addition of K_3 at the least significant bit position of the overruled K_i's (Table 1; column III) and the addition of $\overline{K_3}$ at the least significant bit position of the inverse overruled K_i's (column V) are implemented.

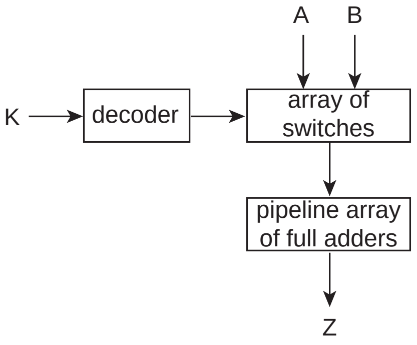


Fig. 1 Block diagram of the processor

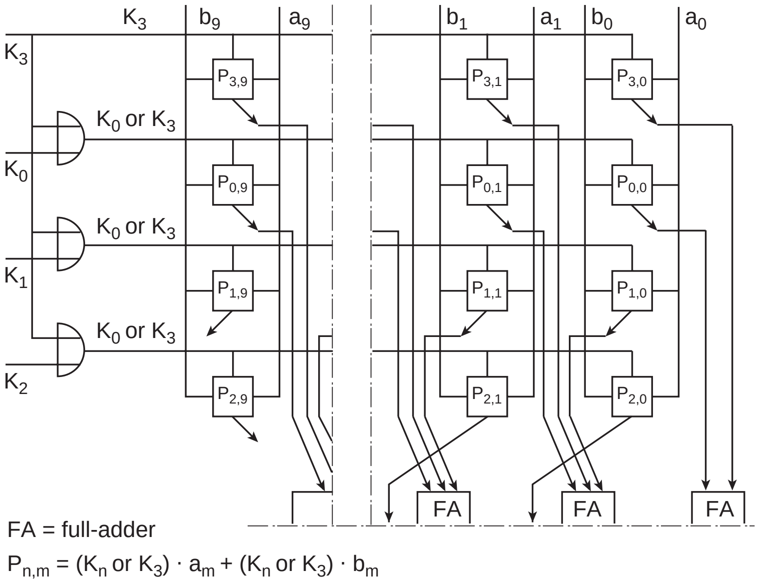


Fig. 2 The input processing, shown in some detail

3 Circuit Design

The electrical circuits (products bits ($P_{n,m}$), full adders ($F{\cdot}A$); Fig. 2) are designed with two-phase race-compensated MOS logic [2].
Fig. 3 shows the electrical circuit of the product bit $P_{n,m}$ designed with this type of logic, which will be discussed now some detail.

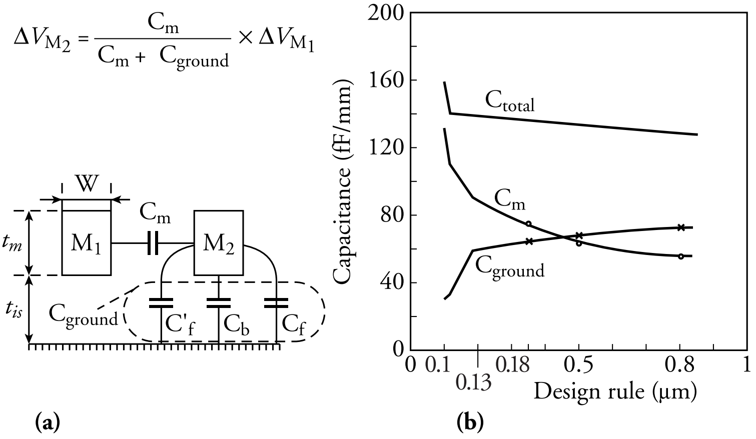


Fig. 3 Electrical circuit of a product bit $P_{n,m}$ designed with two-phase race-compensated MOS logic

During the high level of clock Φ, nodes ① and ② are pre-charged via transistors T_6 and T_7, respectively. During the high level of $\overline{\Phi}$, the data on the input nodes a, b and ($K_n \cup K_3$) are sampled and, within a small time delay (depending on the dimensions of transistors T_1 and T_2), the inverse of ($K_n \cup K_3$) is generated and used at the same time to control another gate. In case the input ($K_n \cup K_3$) is at high level, the gate (node ①) of transistor T_4 should be at low level during $\overline{\Phi}$.

However, after the clock $\overline{\Phi}$ switches on, it will take some time to discharge node ① via transistors T_1 and T_2. In case b_m is at high level, this may introduce a race by creating a conducting path of short duration from node ② via transistor T_1, T_4, and T_5 to ground, resulting in a voltage drop at node ②. Therefore, the bootstrap capacitor C is used to compensate any voltage drop at node ② caused either by charge sharing or by a race.

At the same time during the high level of clock $\overline{\Phi}$, the information at node ② is stored via a transfer gate (T_8) on the gate of the driver (T_9) of the output latch. The output of such a product bit is connected to an input of a full adder (Fig. 2) which has $\overline{\Phi}$ as pre-charge and Φ as a sample clock [2].

By using this type of logic (two-phase race-compensated MOS logic), the data flow is twice that of a straightforward design, since it is no longer necessary in synchronously clocked systems to generate and use the inverse data at different clock periods. It is obvious that this also leads to a hardware reduction.

A microphotograph of the chip is shown in Fig. 4. It is designed in a 4μm E/D technology with implanted under-crossings [2]. In Table 2 the performance of this processor is given. The chip needs two external clocks because of speed and system requirements. A high level of 7V for these clocks allows the chip to run at clock frequencies up to 40MHz.

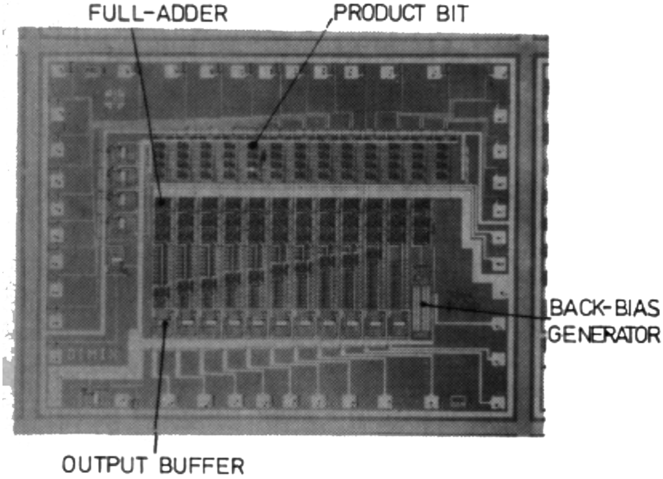


Fig. 4 Microphotograph of the chip

Table 2

Performance		
– clock frequency (max)		: 40 MHz
– chip size		: 15 mm^2
– dissipation		: 300 mW
– power supply		: 5 V
– two external clocks	(20 MHz)	: 5 V
	(40 MHz)	: 7 V
– back-bias generator on chip		
– internal delay		: 8 clock periods

4 Applications

Although the implemented function $Z = K \cdot A + (1 - K) \cdot B$ is a very simple one, the chip has a variety of applications, mainly in the field of digital signal processing. A brief discussion of the most important applications will now be given to show their diversity.

4.1 A Digital Mixer: $K = 0, \frac{1}{8}, ..., 1$

As an example, let K be equal to $\frac{1}{8}$. In this case, Z equals $\frac{1}{8}A + \frac{7}{8}B$. The resolution of $\frac{1}{8}$ can simply be improved by using more chips.

Fig. 5 shows a configuration, which generates the function

$$Z = \frac{n}{16} \cdot A + \frac{(16 - n)}{16} \cdot B.$$

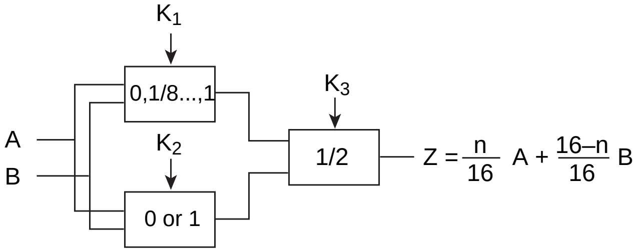


Fig. 5 Configuration of three chips to improve the resolution to $\frac{1}{16}$

However, in this case, the internal delay will be increased by a factor of two. Particularly, when in this application one input (A or B) equals zero, the result is a digital attenuator (potentiometer), or also called a digital mixer (DIMIX).

4.2 A Two Words of Ten Bits Adder: $K = \frac{1}{2}$

Here the output $Z = \frac{1}{2} \cdot (A + B)$ must be shifted over one bit position to get the real sum $Z = A + B$.

4.3 A Two Words of Ten Bits Multiplexer: *K* = 0, *K* = 1, *K* = 0, etc.

In this case, the output Z changes from $Z = B$ to $Z = A$ and vice versa every other clock. In digital video processing, one often wants to switch between two signals, so this processor can serve very well to perform this function. Using two chips, one can even build a de-multiplexer.

4.4 A Six by Ten Bit Multiplier

Fig. 6 shows the configuration using three chips. The input ($I_0 \cdot\cdot\cdot I_9$) is offered simultaneously to two chips where the input patterns are mutually shifted over three bit positions. Thus, it now appears that the three least significant bits of K_1 have an eight times (which equals three bit positions) higher weight than the three least significant bits of K_2. So the overall K seems to consist of six bits. This is expandable by using more chips. However, the output cannot represent more than 11 bits.

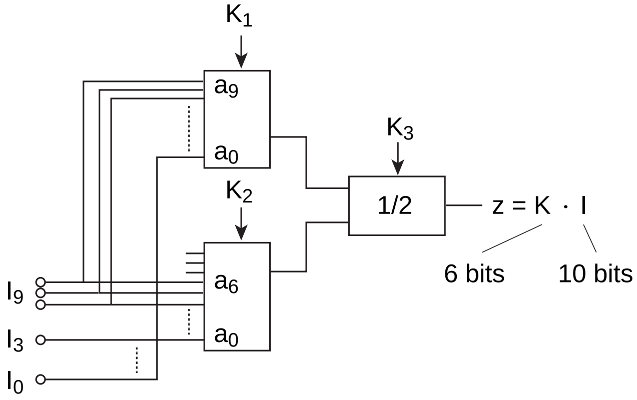


Fig. 6 Application as a six by ten bit multiplier

4.5 A Recursive Filter with Variable Bandwidth

As shown in Fig. 7a, the chip can be used as a recursive filter with variable bandwidth by using a delay in the feedback loop.
Its transfer function can be written as

$$H(z) = \frac{K \cdot z^{-n}}{1 + (K - 1) \cdot z^{-n}}. \tag{6}$$

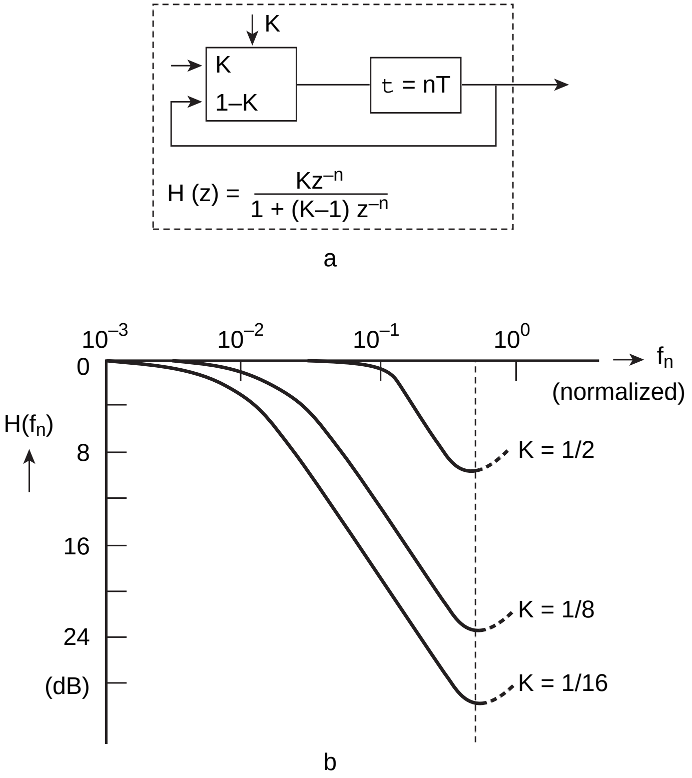


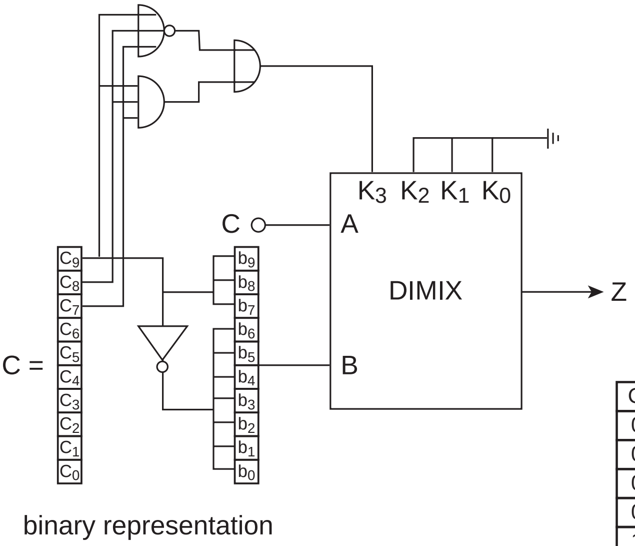
Fig. 7 Applications as a recursive digital filter with variable bandwidth

For different values of *K*, the filter characteristic is shown in Fig. 7b. It can be seen that the 3dB point scales down and the attenuation increases with decreasing *K*.
Particularly when the *K*-factor is controlled by a combination of movement and noise detection and a television frame memory is used in the feedback loop, this configuration can serve very well as a noise reducer in digital video signal processing [3].

4.6 A Digital Signal Limiter

It often happens after some digital signal processing that digital signals might get negative values and therefore lose their physical meaning.
In cases like this, a signal limiter is used, preventing those signals from being negative.

As an example, a ten-bit input in two’s complement which must be (two sides) limited to eight bits in two’s complement is shown in Fig. 8.



C_9	C_8	C_7	K_3	B	Z
0	0	0	1	x	C
0	0	1	0	B_p	B_p
0	1	0	0	B_p	B_p
0	1	1	0	B_p	B_p
1	0	0	0	B_n	B_n
1	0	1	0	B_n	B_n
1	1	0	0	B_n	B_n
1	1	1	1	x	C

$B_p \equiv b_0....b_6$ = "1" and $b_7....b_9$ = "0"

$B_n \equiv b_0....b_6$ = "0" and $b_7....b_9$ = "1"

Fig. 8 Application as a digital signal limiter

If the input (*C*) is more positive than 0001111111, the output (*Z*) must be equal to 01111111; on the other hand, if the input is more negative than 1110000000, then the output must be equal to 10000000.
The logic diagram of this configuration is also shown in Fig. 8.
The only additional hardware for this application consists of three logic gates and an inverter.

The above-mentioned applications, which can all be performed at data rates (*A* and *B* as well as *K*) up to 40MHz (Fig. 9), are not discussed in detail as they are only meant to show their diversity.

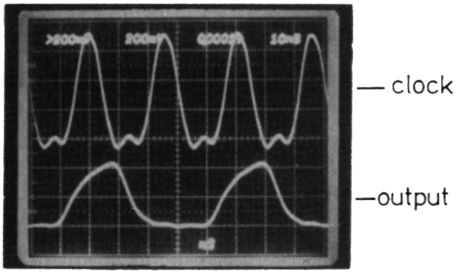


Fig. 9 If all input bits a_i, b_i, and K_i are ‘0’ except for K_0 and a_9, which are ‘1’ every other clock, then the output must be 11110000000 every other clock. This picture shows the response of one of the four most significant output bits at a clock frequency of 40MHz, measured with a ten times attenuating probe

5 Conclusions

A multi-applicable digital signal processor, designed with two-phase race-compensated MOS logic in a 4μm E/D nMOS technology, has been implemented in a chip which runs at clock frequencies up to 40MHz. A special property of the implemented function allowed a great reduction in the amount of hardware to realise this function on a chip. Although this processor is more generally applicable, its 40 MHz data rate makes it particularly suited for digital video signal processing.

6 Acknowledgement

The authors would like to thank J.G. Raven and A.H.H.J. Nillesen for their important contributions to this subject.

7 References

[1] L.R. Rabiner and B. Gold, *Theory and Application of Digital Signal Processing*, Englewood Cliffs, NJ: Prentice-Hall, 1975, ch. 8.

[2] H.J.M. Veendrick, ‘An nMOS dual-mode digital low-pass filter for color TV’, *IEEE J. Solid-State Circuits*, vol. SC-16, pp. 179-182, June 1981.

[3] R.H. McNann *et al.*, ‘Digital noise reducer for encoded NTSC signals’, *SMPTE J.*, vol. 87, Mar. 1978.

Chapter 1.1C

A 1.5 GIPS video signal processor (VSP)

Harry J.M. Veendrick, O. Popp, G. Postuma and M. Lecoutere

IEEE 1994 Custom Integrated Circuits Conference

Abstract

A general purpose, programmable, digital video signal processor has been designed for efficient processing of real-time video signals. The chip is fabricated in a 0.8μm CMOS technology on a die of 156mm^2. The parallel architecture of 28 Processing Elements realises a throughput of 1.5GIPS at a clock frequency of 54MHz. To achieve such a throughput, many of the blocks are custom designed. Special attention has been given to clock routing. Dedicated measures have been taken to reduce power and ground bounce. One of them is the inclusion of a low-voltage I/O mode, which also limits the power consumption. A set of programming tools supports the development of applications.

1 Introduction

One of the most important characteristics of real-time digital video signal processing is its very high sample rate. This may range from 3MHz for a single chrominance signal in a conventional TV receiver, up to 108MHz, or even higher. for advanced high-definition TV signals.
Accordingly, the required processing power and communication bandwidth for picture processing can be very high [1,2]. For this application a programmable general-purpose 1.5GIPS VSP chip has been developed, which runs at clock frequencies up to 54MHz. Higher throughputs can be obtained by the use of (de-) multi-plexers.

2 Architecture

Several architectures have been proposed in the literature [3,4,5].
The architecture of the chip is compatible with the first generation [6] and is characterised by its modular structure (Fig. 1).

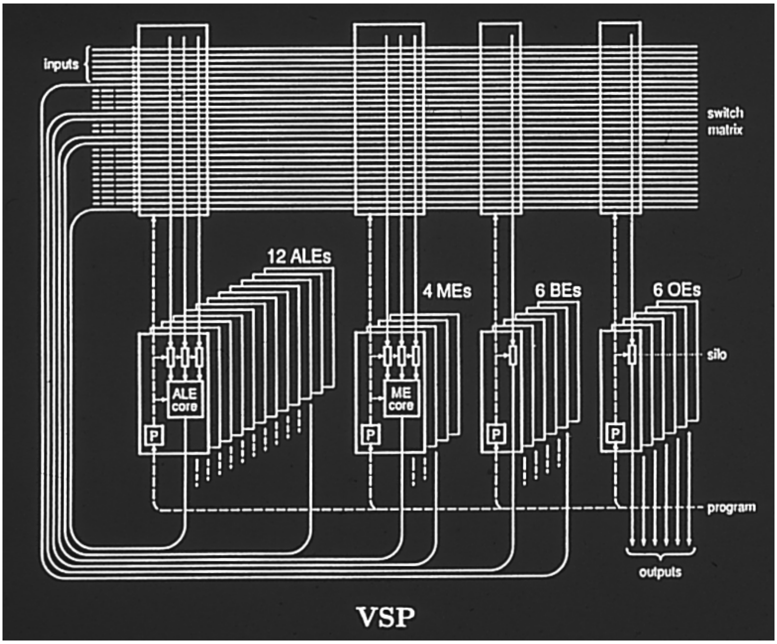


Fig. 1 Architecture of the VSP chip

The data-path is twelve-bit wide. The chip contains a number of pipelined processing elements (PE) operating in parallel and a switch-matrix, which provides programmable communication paths. Every clock cycle, each PE executes a new instruction. The instructions are stored in local program memories (P) and are cyclically executed. The length of each program is user defined, with a maximum of 32 instructions.

The instruction words are 10 to 60bit wide, depending on the type of PE. Each program memory has its own address generation, so that each PE runs an individual program. Conceptually, all programs running in parallel would compare with a VLIW architecture of 1200bit.

Each PE contains one or more programmable-delay registers (silos). These 32 × 12 bit dual-port memories act as buffer memories. They can realise sample (or pixel) delays, equalise pipeline delays in data-paths or ease the scheduling task.

In total there are 28 processing elements on the chip: twelve Arithmetic Logic Elements (ALE), four Memory Elements (ME), six buffer elements (BE) and six output elements (OE).

Fig. 2 shows the configuration of the ALE.

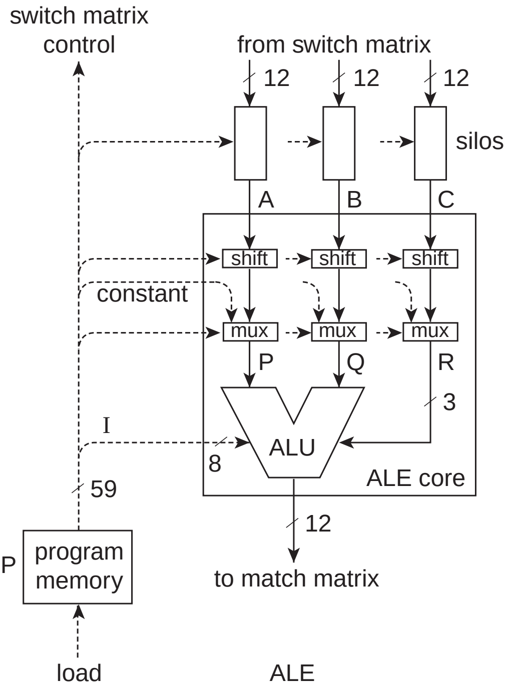
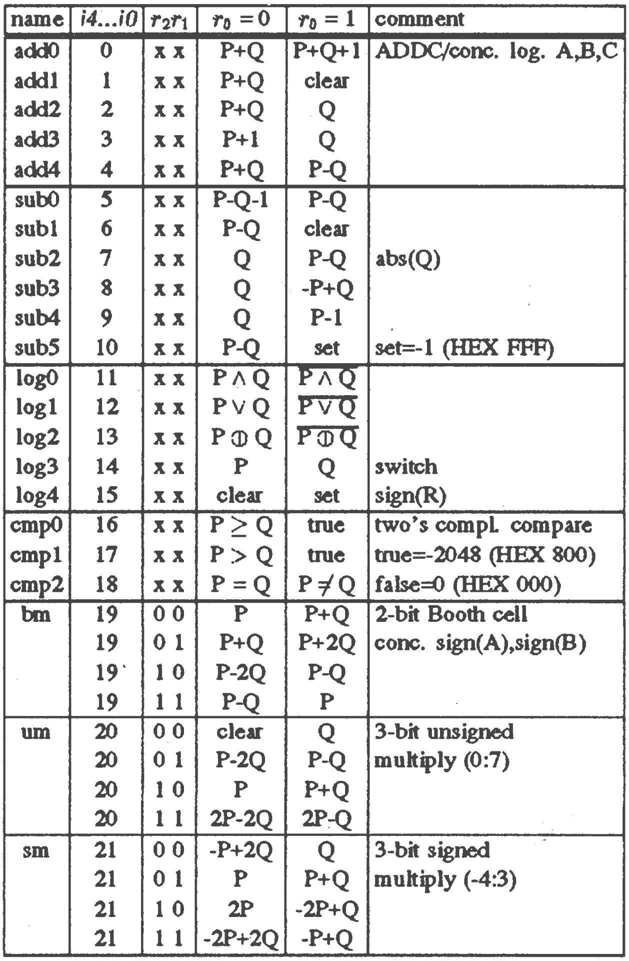


Fig. 2 Configuration of the Arithmetic Logic Element

An ALE contains a program memory (P), three silos and an ALE core. The ALE core consists of an Arithmetic Logic Unit (ALU) and three barrel shifters and can perform several arithmetic, logic, compare and shift operations. The instruction set of the ALU is shown in Fig. 3.



name	$i4...i0$	r_2r_1	$r_0 = 0$	$r_0 = 1$	comment
add0	0	x x	P+Q	P+Q+1	ADDC/conc. log. A,B,C
add1	1	x x	P+Q	clear	
add2	2	x x	P+Q	Q	
add3	3	x x	P+1	Q	
add4	4	x x	P+Q	P-Q	
sub0	5	x x	P-Q-1	P-Q	
sub1	6	x x	P-Q	clear	
sub2	7	x x	Q	P-Q	abs(Q)
sub3	8	x x	Q	-P+Q	
sub4	9	x x	Q	P-1	
sub5	10	x x	P-Q	set	set=-1 (HEX FFF)
log0	11	x x	$P \wedge Q$	$\overline{P \wedge Q}$	
log1	12	x x	$P \vee Q$	$\overline{P \vee Q}$	
log2	13	x x	$P \oplus Q$	$\overline{P \oplus Q}$	
log3	14	x x	P	Q	switch
log4	15	x x	clear	set	sign(R)
cmp0	16	x x	$P \geq Q$	true	two's compl. compare
cmp1	17	x x	$P > Q$	true	true=-2048 (HEX 800)
cmp2	18	x x	$P = Q$	$P \neq Q$	false=0 (HEX 000)
bm	19	0 0	P	P+Q	2-bit Booth cell
	19	0 1	P+Q	P+2Q	conc. sign(A),sign(B)
	19	1 0	P-2Q	P-Q	
	19	1 1	P-Q	P	
um	20	0 0	clear	Q	3-bit unsigned
	20	0 1	P-2Q	P-Q	multiply (0:7)
	20	1 0	P	P+Q	
	20	1 1	2P-2Q	2P-Q	
sm	21	0 0	-P+2Q	Q	3-bit signed
	21	0 1	P	P+Q	multiply (-4:3)
	21	1 0	2P	-2P+Q	
	21	1 1	-2P+2Q	-P+Q	

Fig. 3 Instruction set of the ALU

The instruction word for the ALU is obtained by combining the instruction I from the program memory with the three-bits operand R. This allows data-dependent operations like partial multiply operations (Booth multiplication). The three ALU operands P, Q and R can either be fetched from the switch-matrix, via silos and barrel-shifters or from the program memory (constants). The τ_i bits in the instruction table represent the three bits of the third operand. The ME (Fig. 4) consists of a 2k × 12bit two-port SRAM, three silos, a program memory and some logic for address calculation. It can perform both a read and a write operation in one clock period. The read and write addresses are formed by adding the sum of an address fetched from the switch-matrix via a silo, to an address stored in the program memory. This allows paging or fine-addressing facilities. During the program download a look-up table (LUT) can be stored in the memory. To prevent this data from being overwritten during normal operation, an address-range checker is included in the write-address path. It disables a write operation whenever the write address is within the address range of the fixed data (LUT).

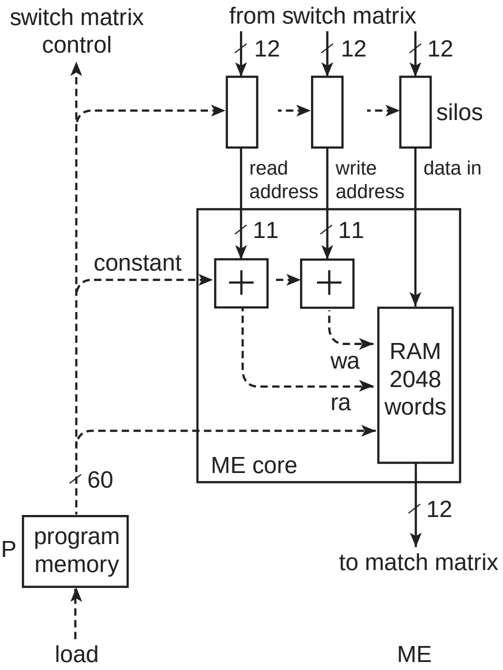


Fig. 4 Configuration of the Memory Element

Another processing element is the BE (Fig. 5a). Its function can be illustrated in the following way. The high processing power of this chip allows the execution of complex algorithms.

Therefore the scheduling task can also be very complex. To achieve an efficient schedule for the execution of such algorithms, the silos, which are used in every

PE, should have a certain storage capacity. The choice for a storage capacity of 32 words is a trade-off between chip area and scheduling efficiency. A BE, therefore, can be regarded as a ‘floating silo’ which can be used in series with any other PE to increase its silo-delay. Six BEs are included in the architecture. Along with a silo, a BE also contains a barrel shifter to perform shift operations thereby freeing an ALE. Like the ALE and the ME, the BE has its input and output also connected to the switch-matrix.

The external communication is realised by the six output elements (OE) allowing, for example, parallel processing of two streams of RGB (or YUV) signals. The chip thus requires 72 input and 72 output pads, resulting in an I/O-bandwidth of 7.7Gbit/s.

Reduction of the dissipation of an OE is achieved by two modes of operation: a normal (TTL-voltage) mode and a low-voltage (1V) mode, in which one VSP can communicate with another VSP. The mode selection for each individual OE is set during initialisation (program loading) of the chip. The configuration of the Output Element is shown in Fig. 5b.

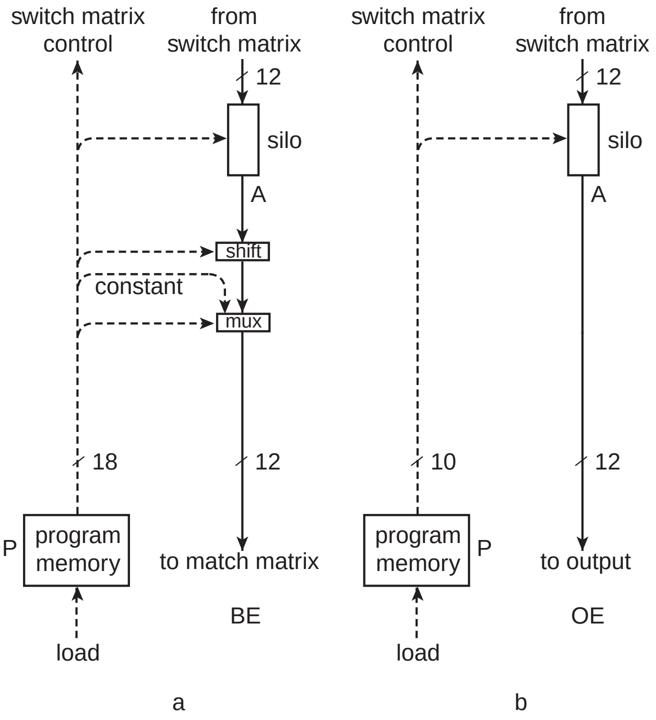


Fig. 5 (a) Buffer Element
(b) Output Element

As discussed before, communication between the 28 PEs is controlled by means of a switch-matrix. The total number of PE inputs equals 60. This results in a switch-matrix of 28 input busses (of 12 bits each) and 60 output busses. By means of extensive area optimisation, this custom designed switch-matrix is very compact and offers a flexible and programmable communication structure between the PEs. Each PE input can select either an output of any PE or any of the six chip inputs.

3 Design Methodology

In a 0.8μm double-metal CMOS process the design of this highly parallel architecture, where all PEs operate at the same speed, the 54MHz clock frequency is a real challenge. Therefore, many of the blocks are custom designs. The ALU, silo, different instances of the program memory, switch-matrix and barrel-shifters are all created from procedural layout descriptions, called generators. These custom generators are written in the L language of the GDT IC Design system. The dual-port data RAM in the ME is a full custom design, optimised for this chip in this technology. The address calculation logic in the ME is realised with standard cells, as is the initialisation and test-control logic. The special high-speed, low-voltage swing option of the I/O circuits also required a dedicated design. Block and chip assembly was done by channel routing. Over the cell routing was used to save chip area.

A set of tools for mapping and scheduling of the algorithms onto the hardware support application of this chip. For verification, these tools have been used to map several algorithms onto the design. These algorithms have been chosen such that the response of every Processing Element can be simulated and verified in GDT. An in-house tool was used to compare the final net-list, extracted from the layout with the GDT net-list.

4 Design implementation

The chip contains about 1.15 million transistors and has an area of 156mm^2. Its maximum dissipation (TTL-swing outputs at 54MHz) is less than 5W. Fig. 6 shows a chip micrograph.

Because of the combination of high complexity, high power dissipation and high performance, special clock and power routing strategies have been adopted during the design. The synchronous two-phase clock system is driven by six parallel clock buffers, which are located symmetrically on the chip. This, combined with very symmetrical and structured clock routing, resulted in an overall clock skew of less than 0.3ns. This is sufficiently less than the non-overlapping time of 0.7ns of the two-phase clock signals to prevent timing problems.

The supply and ground routing is very structured, with multiple supply and ground pads, which are located close to the centre of the chip edges. This reduces bond and lead-frame wire lengths. On-chip capacitances (10nF) are put close to the most bounce-sensitive circuits: the MEs (RAMs). Moreover, the inclusion of a low-voltage (1V) I/O mode also benefits the requirements on ground bounce.

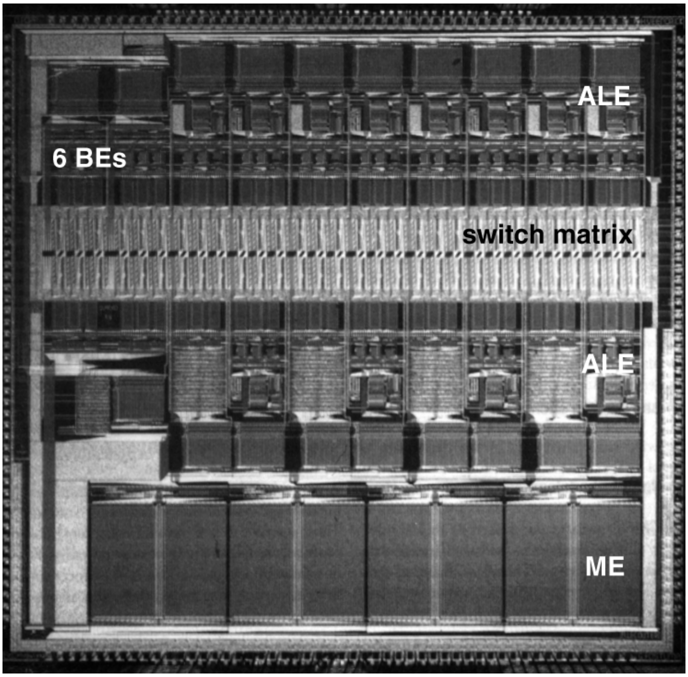


Fig. 6 VSP chip micrograph

Testability has been ensured by the inclusion of eight scan-chains. Observability has been increased with the implementation of Multiple Input Signature Registers (MISRs), which allow compression of data over a number of clock cycles. The final signature is then scanned out.

Table 1 shows the VSP characteristics.

Table 1 Characteristics of the VSP chip

Technology	0.8 μm double metal CMOS
Number of devices	1,115,000 transistors
Chip size	156 mm^2
Clock frequency	54 MHz
Processing power	1.5 GIPS (28 PEs at 54MHz)
Active power	< 5 Watt at 54MHz
I/O modes	low voltage swing (1V) TTL voltage swing
I/O bandwidth	7.7 Gbit/s
Package	208 pins QFP 208 pins PGA

5 Applications

A real-time video-processing algorithm can be regarded as a combination of several operations (add, subtract, multiply, store, etc.). These operations can have different repetition frequencies depending on the sampling frequencies of the individual signals. Each operation must be assigned to one of the processing elements in the proper time slot, such that the processing elements are optimally used. In order to support the use and programming of this chip, a set of tools has been developed to allow efficient mapping and scheduling of the algorithms onto the hardware (one or more chips). The designer interacts with the tools in a graphical way.

The programmable architecture, combined with the available high processing power, allows this chip to be used in a broad spectrum of real-time video applications, e.g. data compression and expansion, motion estimation and compensation, scan-rate conversion, noise reduction, multi picture-in-picture, sample-rate conversions, multidimensional filtering, etc.

6 Conclusions

A powerful chip, capable of processing several different video signals in parallel, has been developed for real-time high-speed video signal processing applications. Its modular parallel architecture, combined with its software programmability, guarantees a wide application area. A complete set of tools has been developed to program this VSP in a user-friendly way and to support its applications.

7 Acknowledgement

The authors gratefully acknowledge G. Essink, P. van Gerwen, C. Hemeryck, K. Huizer, A. van Roermund, P. Snijder and K. Vissers for co-defining the architecture of this chip. We also thank P. Janssen for the development of de test software.

8 References

[1] Roizen, ‘Dubrovnik impasse puts high-definition TV on hold’, *IEEE Spectrum,* September 1986, pp. 32-37.

[2] Chen and Rabaey, ‘A Reconfigurable Multiprocessor IC for Rapid Prototyping of Real-Time Data Paths’, *ISSCC Digest of Technical Papers*, Feb. 1992, pp. 74-75.

[3] Chatterjee, ‘The polycyclic processor’, *Proc. of the IEEE Int. Conf. on Computer Design: VLSI in computers,* Port Chester, New York, 31 Oct.-Nov. 1983, pp. 84-87.

[4] Yamashina *et al.*, ‘A Microprogrammable Real Time Video Signal Processor (VSP) LSI’, *IEEE Journal of Solid-State Circuits,* vol. SC-22, no. 6, December 1987, pp. 1117-1122.

[5] Maruyama *et al.*, ‘A 200 MIPS Image Signal Multiprocessor on a single chip’, *ISSCC Digest of Technical Papers,* Feb. 1990, pp. 122-123.

[6] Van Roermund *et al.*, ‘A general-purpose programmable Video Signal Processor, *IEEE Trans. on Consumer Electronics,* Aug. 1989, pp. 249-257.

Chapter 1.2

Short-Circuit Dissipation of Static CMOS Circuitry and Its Impact on the Design of Buffer Circuits

Harry J.M. Veendrick

IEEE Journal of Solid-State Circuits

August 1984, Volume SC-19, Number 4

Abstract

This paper gives a detailed discussion of the short-circuit component in the total power dissipation in CMOS circuits, on the basis of an elementary CMOS inverter. Design considerations are given for CMOS buffer circuits, based upon the results of the dissipation discussion, to increase circuit performance.

List of Parameters Used

A	process-determined constant defined in (11)
a	ratio between L_p en L_n [(A9)]
b	ratio between parasitic nodal capacitance and load capacitance
β	gain factor (μA/V^2) of an MOS transistor
β_n	β of nMOS transistor
β_p	β of pMOS transistor
β_N	β of nMOST and pMOST of the Nth (symmetrical) inverter of a string
$\beta_{\square}$	β of a transistor with equal channel length and channel width
C_{g_N}	input gate capacitance of the Nth inverter of a string
C_L	load capacitance
C_N	total capacitance on node N
C_o	input capacitance of the first inverter of a string
C_{ox}	gate oxide capacitance
$C_{\mathrm{par}\,N}$	parasitic capacitance on node N
f	frequency (= 1/T)
I	short-circuit current
I_{mean}	mean value of the short-circuit current
I_{max}	maximum value of the short-circuit current
L_n	gate length of the nMOST
L_p	gate length of the pMOST
ΔL_n	gate length minus effective channel length of the nMOST
ΔL_p	gate length minus effective channel length of the pMOST
N	number of inverters
P	total power dissipation
P_1	dynamic power dissipation
P_2	short-circuit power dissipation
T	period-time of a signal (= 1/f)
t	time
τ	rise or fall time of a signal[1]
τ_f	fall time of a signal[1]
τ_i	rise or fall time of an input signal[1]
τ_o	rise or fall time of an output signal[1]
τ_r	rise time of a signal[1]
V_{dd}	supply voltage
V_{in}	input voltage
V_{out}	output voltage
V_T	threshold voltage
V_{T_n}	threshold voltage of nMOST
V_{T_p}	threshold voltage of pMOST
W_{nN}	channel width of the nMOST of the Nth inverter
W_{pN}	channel width of the pMOST of the Nth inverter

[1] Although the rise and fall times are commonly defined to be the time between the 10 and 90 percent level of the signal extremes, in this chapter these parameters are defined as the total duration of a linearised edge.

1 Introduction

During the last five years CMOS technology has become one of the most dominant technologies for VLSI circuits.

The most important reason for this is its low static power dissipation, due to the absence of dc-currents during periods when no signal transients occur. However, during an edge of an input signal there will always be a short-circuit current flowing from supply to ground in static CMOS circuits. So far only limited analyses and discussions have appeared in the literature on this power component of static CMOS circuits [1].

In integrated circuits it is always necessary to drive large capacitances (bus lines, ‘off-chip’ circuitry, etc.), often at high clock frequencies. Such driving circuits (buffers) will take a relatively large part of the total power consumption of the chip. It is clear that optimisation of such circuits requires a different approach as compared to optimisation of CMOS logic [2]. These buffer circuits need extra attention to obtain minimum power dissipation. Therefore, a detailed discussion on power dissipation of a basic CMOS inverter will be given first.

2 Dissipation of a Basic CMOS Inverter

A static CMOS inverter does not dissipate power during the absence of transients on the inputs: when the input (Fig. 1) is at high level (V_{dd}), only the nMOS transistor conducts, and when the input is at low level, only the pMOST will conduct. However, during a transient on the input, there will be a time period in which both the nMOST and pMOST will conduct, causing a short-circuit (I) to flow from supply to ground, as shown in Fig. 2 for an inverter without load. This current flows as long as the input voltage (V_{in}) is higher than a threshold voltage $\left(V_{T_n}\right)$ above V_{dd} and lower than a threshold $\left(\left|V_{T_p}\right|\right)$ below V_{dd}.

If we load the inverter output of Fig. 1 with a capacitance C_L, then the dissipation of the circuit consists of two components:

dynamic dissipation: $P_1 = C_L \cdot V^2 \cdot f$ and (1)

short-circuit dissipation: $P_2 = I_{\text{mean}} \cdot V.$ (2)

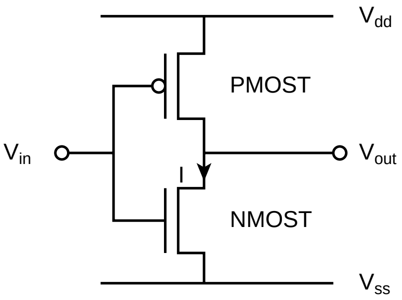


Fig. 1 Basic CMOS inverter

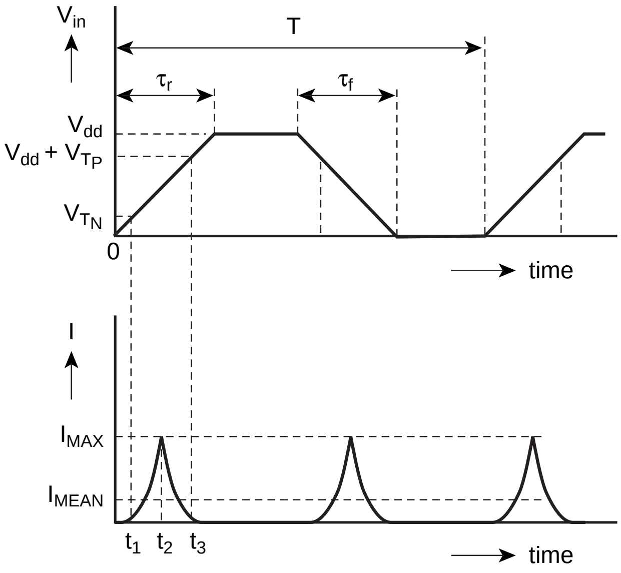


Fig. 2 Current behaviour of an inverter without load

Clearly, the dynamic component P_1 does not depend on the inverter design (apart from contributions due to parasitic output capacitances, such as junction capacitances). The second component P_2, however, strongly depends on the inverter design.

Since there is a difference in the short-circuit dissipation of an inverter without load and that of an inverter with load, we start our discussions on the basis of an inverter with zero load capacitance. For simplicity we assume that the inverter is symmetrical (an asymmetrical inverter is not fundamentally different), which means that

$$\beta_n = \beta_p = \beta \text{ and } V_{T_n} = -V_{T_p} = V_T . \tag{3}$$

During the period ($t_1 - t_2$; Fig. 2) in which the short-circuit current I increases from 0 to $I_{\max}$, the output voltage (V_{out}) will be larger than the input voltage (V_{in}) minus the threshold voltage (V_T) of the nMOST. As a consequence, the nMOS transistor will be in saturation during this period of time.
Using the simple MOS formula, this leads to

$$I = \frac{\beta}{2}(V_{\text{in}} - V_T)^2 \text{ for } 0 \le I \le I_{\max} . \tag{4}$$

This current will reach its maximum value when V_{in} equals half the supply voltage ($V_{\text{in}} = V_{dd}/2$), due to the assumption that the inverter was symmetrical. Another result of this assumption is that the current behaviour during the time period $t_1 - t_3$ will be symmetrical with respect to the time t_2.
The mean current during a time T (equal to one period of the input signal) can thus be written as

$$I_{\text{mean}} = 2 \cdot \frac{2}{T}\int_{t_1}^{t_2} I(t)\,\mathrm{d}t = \frac{4}{T}\int_{t_1}^{t_2} \frac{\beta}{2}(V_{\text{in}}(t) - V_T)^2\,\mathrm{d}t. \tag{5}$$

Assuming equal rise and fall times[2] ($\tau_r = \tau_f = \tau$) of the input signal (symmetrical) and a linear relation between the input voltage (V_{in}) and time (t) during its transients

$$V_{\text{in}}(t) = \frac{V_{dd}}{\tau} \cdot t, \tag{6}$$

it can be derived from Fig. 2 that

$$t_1 = \frac{V_T}{V_{dd}} \cdot \tau \text{ and } t_2 = \frac{\tau}{2}. \tag{7}$$

Equations (5), (6) and (7) lead to

$$I_{\text{mean}} = \frac{2\beta}{T}\int_{\tau/2}^{V_T \cdot \tau / V_{dd}} \left(\frac{V_{dd}}{\tau} \cdot t - V_T\right) \mathrm{d}\left(\frac{V_{dd}}{\tau} \cdot t - V_T\right), \tag{8}$$

[2] The definitions of rise and fall times used here are different from those in common use (see the list of parameters used).

which has the solution

$$I_{\text{mean}} = \frac{1}{12} \cdot \frac{\beta}{V_{dd}} \cdot (V_{dd} - 2V_T)^3 \cdot \frac{\tau}{T}. \tag{9}$$

From (2) and (9) the following expression can be derived for the short-circuit dissipation of a CMOS inverter without load:

$$P_2 = \frac{\beta}{12} \cdot (V_{dd} - 2V_T)^3 \cdot \frac{\tau}{T}. \tag{10}$$

As $1/T = f$, (10) shows that this dissipation component is also proportional to the frequency of switching. Because V_{dd} and V_T are process-determined, the only design parameters that affect P_2 are β and the input rise and fall times (τ) of the inverter.

For an inverter with capacitive load, the β's of the transistors are determined by requirements on output rise and fall times. In this case the short-circuit dissipation depends only on the duration of the input signal edges. As will be shown further on, these edges should not be too long, especially in the case of driver circuits that have a large β-value. In the derivation of (10), we started with an inverter without load. The following example examines what happens when we load the inverter with different capacitances.

Example

The discussions that follow are based upon the inverter whose parameter values are shown in Fig. 3. Its operation was simulated with a circuit analysis program. Some results are presented in Fig. 4. The figure shows the short-circuit current behaviour, during a time interval $t_1 - t_3$ (see Fig. 2), as a function of the load capacitance C_L, for input rise and fall times of 5ns. Curve ① shows the behaviour of the inverter without load. At any time this current is the maximum short-circuit current that can occur. This means that all other current characteristics for different load capacitances must be within this curve. Curve ④ shows the short-circuit current behaviour of the inverter when it is loaded with a characteristic capacitance C_L of 500fF. In this case the rise and fall times on the output node are equal to the rise and fall times on the input.

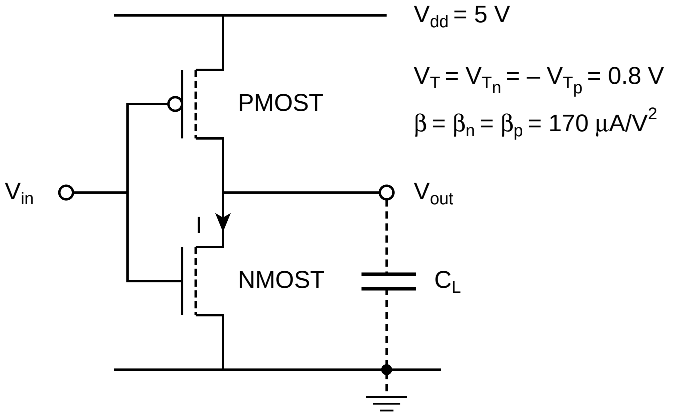

$V_{dd} = 5$ V
$V_T = V_{T_n} = -V_{T_p} = 0.8$ V
$\beta = \beta_n = \beta_p = 170$ µA/V^2
PMOST
NMOST
V_{in}
V_{out}
I
C_L

Fig. 3 The inverter used in the example

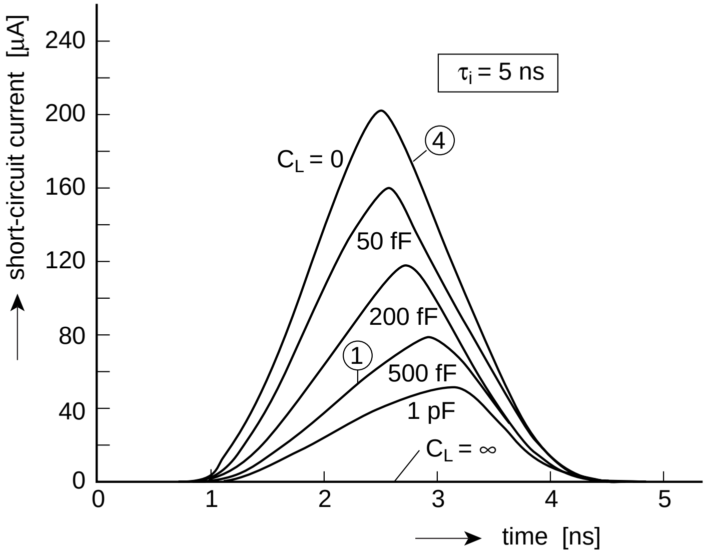


Fig. 4 Short-circuit current as a function of different inverter load capacitances

Fig. 5 shows the output transient curves for different values of the load capacitance when the input is switched from high level to ground with a fall time of 5ns. As expressed in (10) for a CMOS inverter with zero load capacitance, it is obvious that the dissipation versus load capacitance characteristic will depend on the rise and fall times of the input signal. Fig. 6 shows this characteristic for different values of the input rise and fall times (τ_i). The dashed line shows the dynamic dissipation (f = 10MHz), while the solid lines show the actual inverter dissipation (dynamic plus short-circuit dissipation). The points where the load capacitance corresponds to equal input and output rise and fall times for the different characteristics are indicated on the figure.

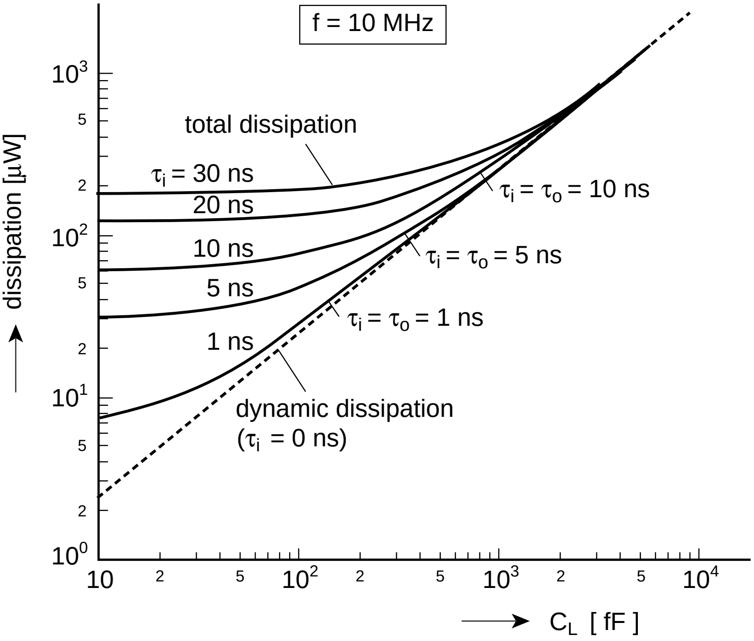


Fig. 5 Inverter output voltage behaviour for different inverter load capacitances

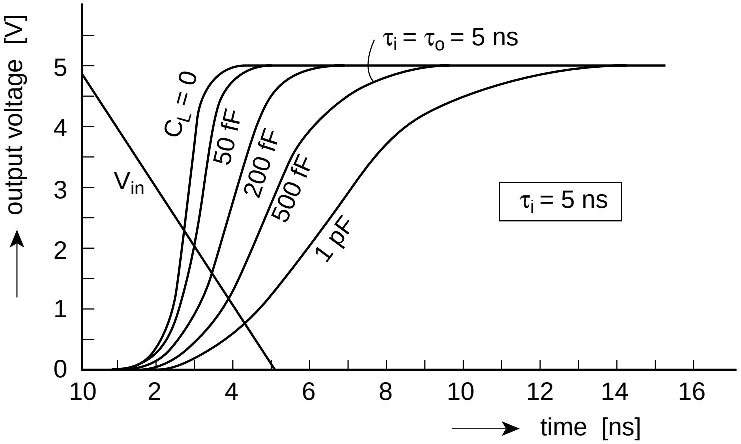


Fig. 6 Inverter dissipation as a function of the inverter load capacitances

From these characteristics we can conclude that if the operation of the inverter is such that the output signal and input signal have equal rise and fall times, the short-circuit dissipation will be only a fraction (< 20 percent) of the total dissipation. However, if the inverter is more lightly loaded, causing output rise and fall times that are relatively short as compared to the input rise and fall times, then the short-circuit dissipation will increase to the same order of magnitude as the dynamic dissipation. Therefore, to minimise dissipation, an inverter used as part of a buffer should be designed in such a way that the input rise and fall times are less than or about equal to the output rise and fall times in order to guarantee a relatively small short-circuit dissipation.

Fig. 7 shows the linear relationship [according to (10)] between the short-circuit dissipation and the input rise and fall times (τ_i) derived by means of circuit simulations. In this case the inverter of Fig. 3 was loaded with a capacitance of 500fF. From Fig. 5 it was known that a load capacitance of 500fF caused 5ns rise and fall times of the output signal, when the inverter input rise and fall times are equal to 5ns. This point, $\tau_i = \tau_0 = 5$ns, is indicated on Fig. 7 and the corresponding short-circuit dissipation is only about 10 percent of the dynamic dissipation.
This result of designing a string of inverters in such a way that the input and output rise and fall times of each inverter are equal to obtain minimum dissipation can very well be applied to the design of static CMOS buffers.

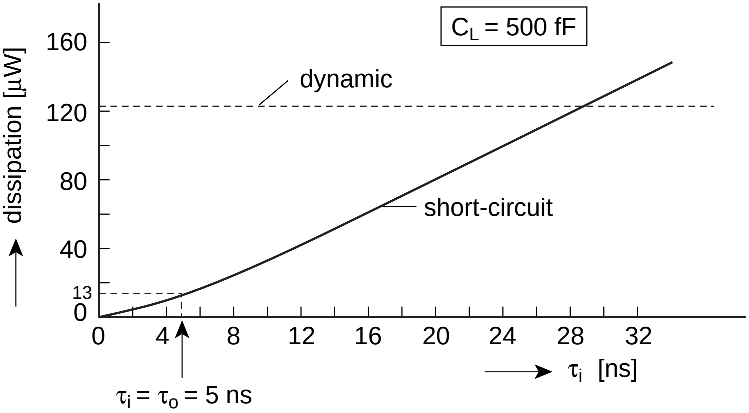


Fig. 7 Inverter dissipation as a function of the input rise and fall times

Although (10) was derived for an inverter with zero load capacitance and therefore is for the maximum short-circuit dissipation, it can also be applied to inverters designed with $\tau_i = \tau_0$. It has empirically been found that for such designs the short-circuit dissipation is half of the maximum, calculated with (10).

3 Design Considerations

In integrated circuits it is always necessary to drive large capacitances, likes bus lines or ‘off-chip’ circuitry. Moreover, this must often occur at high speed, which will take a relatively large part of the total power dissipation of the chip. Particularly in the case of bus lines, which control a large number of inputs of the different sub-circuits on a chip, it is necessary to have short signal rise and fall times [(10)] to minimise dissipation.
Suppose we want to drive such a bus interconnection line or ‘off-chip’ circuitry with a signal coming from an internal node *A*. Let us assume that the logic gate, having node *A* as output, is capable of charging a capacitance C_0 in a time *t* to 95 percent of the supply voltage. From the previous results we know that if we use an inverter string as a buffer circuit between node *A* and the bus line, the rise and fall times on each node of the string should be equal to the required rise and fall times on the bus line (or bonding pad) to be driven.
The problem now is how to design an inverter string (Fig. 8) loaded with a capacitance C_L, with τ ns rise and fall times on each node, driven from an internal logic gate capable of charging and discharging the input capacitance C_0 in the same time. In [3] it was derived that a factor of *e* between the β’s of the successive inverters (tapering factor) was needed to guarantee a minimum propagation delay time for such an inverter string. However, it is well understood that in terms of dissipation and silicon area this will not lead to an optimum design. Design optimisation for minimum dissipation and silicon area requires a different approach, as will be shown in the following.

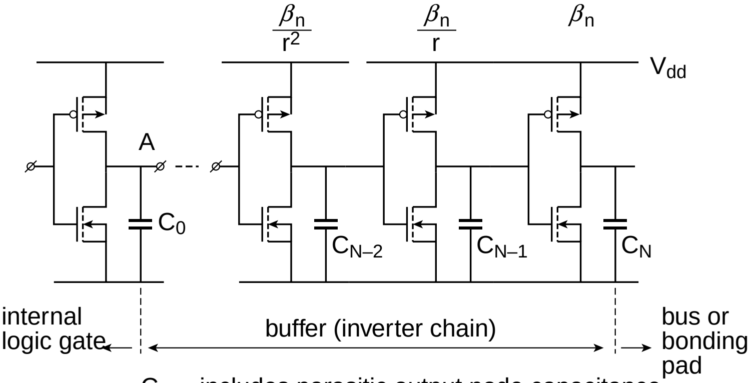
It has been derived in the Appendix included in this paper that a ‘minimum dissipation design’ of the inverter string is completely determined by the following three equations:

$$\beta_N = \frac{C_N}{\tau_N} \cdot A,$$

where

$$A = \frac{1}{V_{dd} - V_{T_n}} \cdot \left\{ \frac{2V_{T_n}}{V_{dd} - V_{T_n}} + \ln\left[\frac{2\left(V_{dd} - V_{T_n}\right) - V}{V} \right] \right\} \tag{11}$$

is constant for a given technology.



C_N = includes parasitic output node capacitance
r = tapering factor

*) the nodal capacitances include the parasitic nodal capacitances (C_{par_i})

$C_N = C_L + C_{par_N}$
$C_{N-1} = C_{g_N} + C_{par_{N-1}}$

Fig. 8 Inverter string acting as a buffer circuit

$$\frac{\beta_{N-1}}{\beta_N} = \frac{A}{\tau\cdot\beta_{n_\square}}\cdot(1+b)\cdot C_{ox}\cdot L_{n_N} \cdot\left\{L_{n_N}\left(1+3a^2\right) - \Delta L_{n_N} - 3a\cdot\Delta L_{p_N}\right\} \tag{12}$$

and

$$\left(\frac{\beta_N}{\beta_{N-1}}\right)^N = \frac{C_N}{C_0} \tag{13}$$

where β_N represents the β of the last inverter stage, β_{N-1}/β_N is the tapering factor at equal input and output rise and fall times and N is the number of inverters of the string. For a practical application the following assumptions are made for the parameters of (11), (12), and (13):

$$\left.\begin{aligned} V_{T_n} &= -V_{T_p} = 1\,\text{V} \\ V &= 0.05\cdot V_{dd}\ \text{[in (11)]} \\ V_{dd} &= 5\,\text{V} \end{aligned}\right\} A \approx 1$$

$$\beta_{n_\square} = 42\ \mu\text{A/V}^2 \qquad \beta_{p_\square} = 14\ \mu\text{A/V}^2$$
$$C_{\text{ox}} = 700\,\mu\text{F/m}^2$$
$$L_n = 2.5\,\mu\text{m} \qquad \Delta L_n = 0.5\,\mu\text{m} \qquad \Delta L_p = 1\,\mu\text{m}$$

Required rise and fall times: τ = 5ns. Practical values for the constants *a* and *b* are *a* = 3/2.5 and *b* = 0.1. With (12) this leads to

$$\frac{\beta_N}{\beta_{N-1}} = 11.5.$$

This tapering factor, as it is often called, is strongly process dependent; nearly all parameters in (12) are determined by the process. Different CMOS processes may therefore lead to different tapering factors.
If, in a practical situation, C_o = 100fF and C_L (Fig. 8) equals 10pF and we want 5ns rise and fall times (τ) on the output of the driver (inverter string), then the design procedure is as follows (with the above assumptions):

according to Fig. 8:

$$C_N = C_L + C_{\text{par}_N} = (1+b)\cdot C_L = 11\,\text{pF}$$

according to (11):

$$\beta_N = \frac{C_N}{\tau_N}\cdot A = \frac{C_N}{\tau}\cdot A = 4.4\cdot 10^{-3}\ \text{A/V}^2.$$

Thus, the last inverter stage (*N*) is determined by

$$\beta_{n_N} = \beta_{p_N} = \beta_N = 4.4\cdot 10^{-3}\ \text{A/V}^2,$$

or (see Appendix 1)

$$\left(\frac{W_n}{L_n - \Delta L_n}\right)_N = \frac{\beta_N}{\beta_{p_\square}} = 105, \quad \text{so} \left(\frac{W_n}{L_n}\right)_N = \frac{210\,\mu\text{m}}{2.5\,\mu\text{m}}$$

and

$$\left(\frac{W_p}{L_p - \Delta L_p}\right)_N = \frac{\beta_N}{\beta_{n_\square}} = 315, \quad \text{so} \left(\frac{W_p}{L_p}\right)_N = \frac{630\mu\text{m}}{3\mu\text{m}}.$$

The $(N - i)$ inverters are now determined by the tapering factor

$$\frac{\beta_N}{\beta_{N-1}} = 11.5.$$

With the given C_0 = 100fF and C_N = 11pF we find from (13) that the number of inverters needed for this example will be equal to N = 1.83. (This, of course, has to be rounded off to N = 2.)

For this example, therefore, the inverter string should be designed as shown in Fig. 9 to guarantee a very small short-circuit dissipation and a minimum area consumption. The parasitic nodal capacitances are also depicted in Fig. 9.

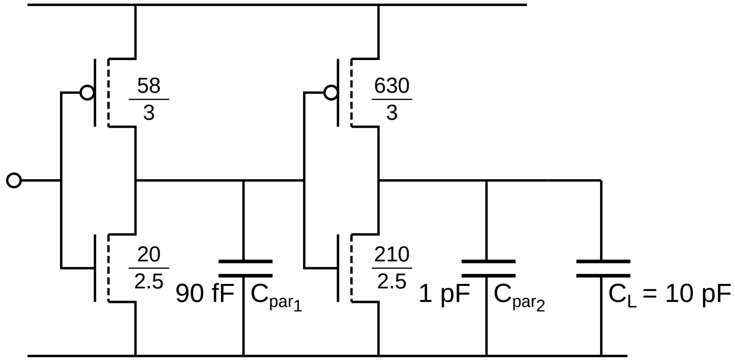


Fig. 9 The designed inverter string for a practical example

By means of circuit simulations, the mean power dissipation has been calculated at a clock frequency of $f = 1/T$ = 10MHz. Table 1 shows the results of a comparison of two tapering factors: a factor of 11.5, which is derived in this chapter (process-determined) from optimisation of power dissipation and area, and a factor e, which is derived [3] from optimisation of the propagation delay.

In this example the most important improvement due to choosing a tapering factor equal to 11.5 instead of a factor e is a much smaller area (< 1/4) and a reduced parasitic power consumption (≈ 1/4). This parasitic power consumption is the total power consumption minus the power dissipation, which was actually meant $(C_L V^2 f)$.

Table 1 Comparison of the performance of two inverter strings with different tapering factors

	Factor 11.5	Factor e
Number of inverters	2	5
Size of de pMOST in the last inverter	630/3	2660/3
Dynamic power dissipation Parasitic power dissipation	2.5mW 1.4mW	2.5mW 5.4mW
Propagation delay	6.5nsec	5.5nsec

In our case, the propagation delay has only been increased by 1ns as compared to a tapering factor equal to *e*.

In summarising, we can state that optimisation of the power dissipation of CMOS driving circuits, like buffers, will lead to a better overall circuit performance (power, delay, and area) then can be achieved by optimisation of the propagation delay.

4 Summary and Conclusions

In this chapter a simple formula is derived for a quick calculation of the maximum short-circuit dissipation of static CMOS circuits. A detailed discussion of this short-circuit dissipation is given based upon the behaviour of the inverter when loaded with different capacitances. It was found that if each inverter of a string is designed in such a way that the input and output rise and fall times are equal, the short-circuit dissipation will be much less than the dynamic dissipation (< 20 percent). This result has been applied to a practical design of a CMOS driving circuit (buffer), which is commonly built up of a string of inverters. An expression has also been derived for a tapering factor between two successive inverters of such a string to minimise parasitic power dissipation. Finally, it is concluded that optimisation in terms of power dissipation leads to a better overall performance (in terms of speed, power, and area) than is possible by minimisation of the propagation delay.

Appendix

It can easily be derived [1] what transistor is needed to discharge a capacitive load C_N from the supply voltage V_{dd} to a voltage V in the time τ_N:

$$\beta_N = \frac{1}{\tau_N} \cdot \frac{C_N}{V_{dd} - V_{T_n}} \cdot \left\{ \frac{2V_{T_n}}{V_{dd} - V_{T_n}} + \ln\left[\frac{2\left(V_{dd} - V_{T_n}\right) - V}{V} \right] \right\} \tag{A1}$$

where

$$\frac{1}{V_{dd} - V_{T_n}} \cdot \left\{ \frac{2V_{T_n}}{V_{dd} - V_{T_n}} + \ln\left[\frac{2\left(V_{dd} - V_{T_n}\right) - V}{V} \right] \right\} = A \tag{A2}$$

is a constant for given technology. Thus,

$$\beta_N = \frac{C_N}{\tau_N} \cdot A \tag{A3}$$

and for the (N – 1)-th inverter:

$$\beta_{N-1} = \frac{C_{N-1}}{\tau_{N-1}} \cdot A. \tag{A4}$$

Again, assuming the inverter to be symmetrical:

$$\beta_n = \beta_p = \beta; \quad V_{T_n} = -V_{T_p} \quad \text{and}\, \tau_r = \tau_f = \tau \tag{A5}$$

and, because of the difference in mobility of holes and electrons:

$$\frac{W_p - \Delta W_p}{L_p - \Delta L_p} = 3 \cdot \frac{W_n - \Delta W_n}{L_n - \Delta L_n}. \tag{A6}$$

As $W_n >> \Delta W_n$ and $W_p >> \Delta W_p$, (A6) reduces to

$$\frac{W_p}{L_p - \Delta L_p} = 3 \cdot \frac{W_n}{L_n - \Delta L_n}. \tag{A7}$$

With

$$W_n = \left(L_n - \Delta L_n\right)\frac{\beta_N}{\beta_{n_\square}} \tag{A8}$$

and given a linear relation between L_n and L_p

$$L_p = a \cdot L_n \tag{A9}$$

we find

$$W_p \cdot L_p = W_n \cdot L_n \cdot 3a \cdot \frac{a \cdot L_n - \Delta L_p}{L_n - \Delta L_n}. \tag{A10}$$

From Fig. 8 it is known that

$$C_{N-1} = C_{g_N} + C_{\mathrm{par}_{N-1}}. \tag{A11}$$

In a practical design the parasitic capacitance $C_{\mathrm{par}_{N-1}}$ of node N – 1 will be proportional to its load capacitance C_{g_N} , so that

$$C_{\mathrm{par}_{N-1}} = b \cdot C_{g_N} \text{ and } C_{\mathrm{par}_N} = b \cdot C_L. \tag{A12}$$

From (A11) and (A12) we derive

$$C_{N-1} = (1+b) \cdot \left(W_{p_N} \cdot L_{p_N} + W_{n_N} \cdot L_{n_N}\right) \cdot C_{\mathrm{ox}}. \tag{A13}$$

This, combined with (A10) yields

$$C_{N-1} = \left(W_{n_N} \cdot L_{n_N}\right) \cdot \left\{1 + a \cdot \frac{\left(3a \cdot L_{n_N} - 3\Delta L_{p_N}\right)}{L_{n_N} - \Delta L_{n_N}}\right\} \cdot (1+b) \cdot C_{\mathrm{ox}}. \tag{A14}$$

Equations (A4), (A5) and (A14) result in

$$\begin{aligned} \beta_{N-1} = {} & \frac{A}{\tau} \cdot (1+b) \cdot C_{\mathrm{ox}} \\ & \cdot \frac{W_{n_N} \cdot L_{n_N}}{L_{n_N} - \Delta L_{n_N}} \cdot \left(L_{n_N} - \Delta L_{n_N} + 3a^2 \cdot L_{n_N} - 3a \cdot \Delta L_{p_N}\right) \end{aligned} \tag{A15}$$

Finally, from (A8) and (A15) we derive

$$\frac{\beta_{N-1}}{\beta_N} = \frac{A}{\tau \cdot \beta_{n_\square}} \cdot (1+b) \cdot C_{ox} \cdot L_{n_N} \cdot \left\{ L_{n_N} \left(1+3a^2\right) - \Delta L_{n_N} - 3a \cdot \Delta L_{p_N} \right\} \tag{A16}$$

With equations (A1) and (A16) the inverter string is completely determined.
The number of inverters (*N*), which depends on the ratio in (A16), is now determined by

$$\left(\frac{\beta_N}{\beta_{N-1}} \right)^N = \frac{C_N}{C_0} \tag{A17}$$

where C_0 and C_N represent the input capacitance and the output load capacitance of the inverter string, respectively (Fig. 8).

Acknowledgement

The author wishes to acknowledge the helpful suggestions and contributions made by A.T. van Zanten and C. Hartgring.

References

[1] M.I. Elmasry, ‘Digital MOS integrated circuits: A tutorial’, *Digital MOS Integrated Circuits*. New York: IEEE Press, pp. 4-27.

[2] A. Kanuma, ‘CMOS circuit optimization’, *Solid-State Electron.*, vol. 26, no. 1, pp. 47-58, 1983.

[3] C. Mead and L. Conway, *Introduction to VLSI Systems*. New York: pp. 12-15.

Chapter 2.1A

A 40MHz 308Kb Charge Coupled Device (CCD) Video Memory

Harry J.M. Veendrick, Leo C. Pfennings, Marcel J.J.C. Annegarn,
Hendrick A. Harwig, Marcel J.M. Pelgrom, Henk Jan F. Peuscher,
Jan G. Raven, Arie Slob, Jan W. Slotboom
Philips Research Laboratories

ISSCC 84 / February 23, 1984

1984 IEEE International Solid-State Circuits Conference

A 40MHz 308Kb CCD memory (34.8mm^2), fabricated in a modified 2μm nMOS process (only one extra mask) will be discussed. Dedicated I/O control and a 20ms refresh time (at 90°C) assures flexible use in digital video signal processing.

Advances in VLSI technology have opened the way for video signal processing to go digital. In the near future this will be lead to reduced costs and an improved reliability. For picture quality improvement and additional features, the use of field memories becomes a necessity[3,4]. The serial character of video data implies a preferred use of line oriented serial memories (CCDs) over RAMs. Fig. 1 shows the basic I/O and control signals needed to operate a video CCD memory, which is synchronised to the line frequency.

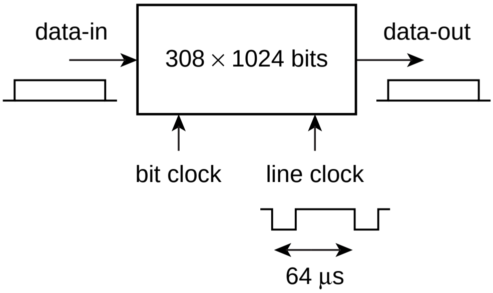


Fig. 1 Basic I/O and control signals for operation of the video CCD memory

As compared to buried channel CCDs, surface CCDs have the advantage of a relatively simple technology and a higher charge capability; moreover, they are less affected by narrow channel effects. This allows implementation in a standard E/D nMOS process (2μm) with only slight modifications (one extra mask to define the CCD region).

The elementary CCD cell (only 7 × 7μm^2), shown in Fig. 2, consists of a polysilicon storage gate (on 500Å oxide) and an aluminium transfer gate (on 1200Å oxide), both controlled by the same clock. The storage capability of this cell is 15fC.

[3] Berkhoff, E.J., Kraus, U.E., Raven, J.G., 'Applications of Picture Memories in Television Receivers', *IEEE Trans. Consumer Electronics*, Vol. CE-29, p. 251-255; Aug., 1983.

[4] Fisher, Th., 'What is the Impact of Digital TV?', *IEEE Trans. Consumer Electronics*, Vol. CE-28, p. 423-429; Aug., 1982.

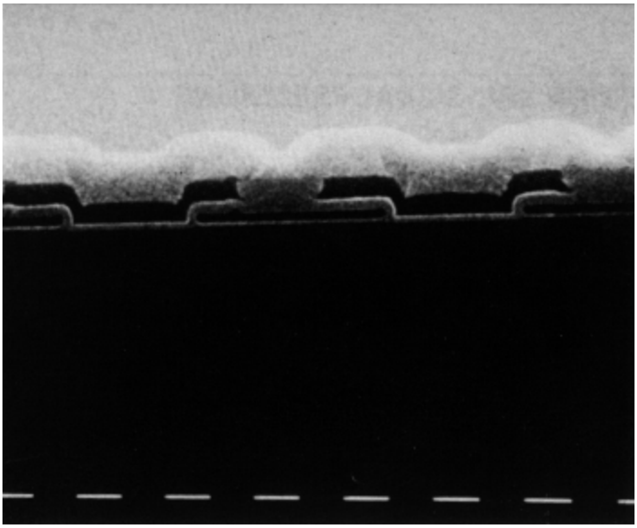


Fig. 2 CCD memory cell consisting of a polysilicon storage gate and an aluminium transfer gate

Storage of one bit of a digitised TV field requires a memory capacity of 308 lines of 1024 b[5] with a storage time of 10 or 20ms, depending on the application (40MHz of 20MHz). To overcome problems with speed, power and a limited number of transfers, which are related to a single serial-parallel-serial (SPS) structure solution, data (40MHz) is de-multiplexed over 8 SPS memory blocks (5MHz) of 39Kb each; Fig. 3. Both the input and output serial CCD registers are implemented as a two-phase 128 b register, requiring 256 storage gates. The parallel registers contain 170 storage gates each. By means of data interlacing and using a ten-phase ripple clock in the parallel register, a nearly one-electrode-per-bit storage density is achieved. As a consequence, each data bit is stored up to 20ms without refresh. The clocks for the serial registers are deduced from a Gray-code counter via a decoder circuit. The ripple clocks are generated via this decoder and a divider circuit. Each SPS block has its own sense amplifier and reference circuit (Fig. 4).

[5] In the PAL/SECAM system: the number of TV lines can be virtually reduced[5] for applications in the NTSO system.

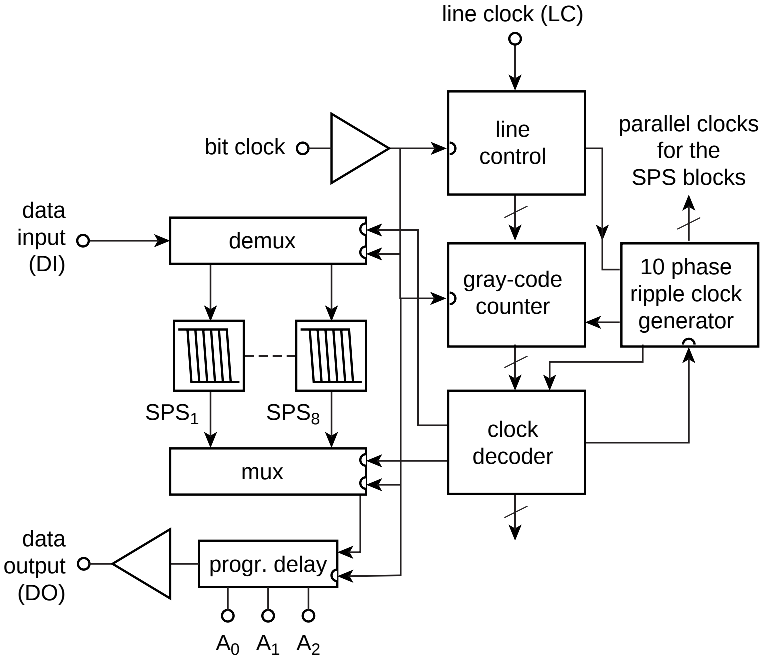


Fig. 3 Block diagram of the memory design

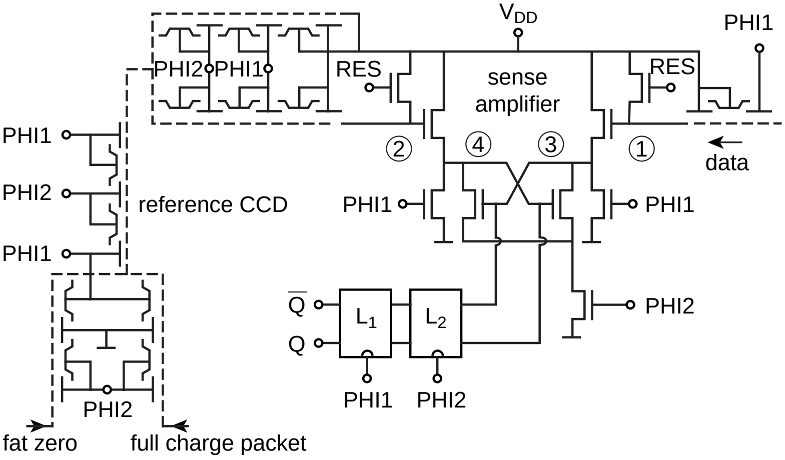


Fig. 4a Sense amplifier and reference circuit

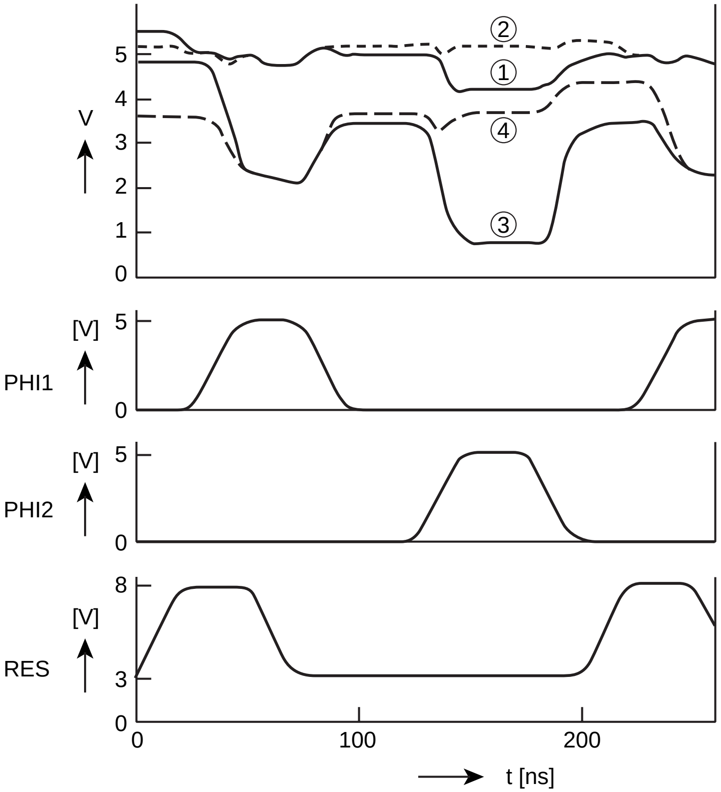


Fig. 4b Clock and output waveforms during sense amplifier operation

A high transfer efficiency ($\varepsilon \approx 5.10^{-4}$) combined with a very low leakage current[6] (0.2μA/cm^2 at 90°C), realised in our SPS structures, allow for only a short reference CCD generates instead of a complete SPS block. This reference CCD generates charge packets, which contain the mean of a fat zero and a full charge packet.
The sensitivity of the amplifier is about 100mV (1.5fC). Clock and output waveforms are shown in Fig. 4(b). Low level (3V) and high level (8V) of the reset (RES) signal respectively overcome problems with cross-talk and threshold losses on the sense nodes ① and ②. The sense amplifier outputs are latched and then fed to the output multiplexer; Fig. 3.
To be able to synchronise the input and output (for re-circulate applications), a programmable 7b-delay has been included in the design.

[6] Slotboom, J.W., Harwig, H.A., Pelgrom, M.J.M., 'Leakage Current in High Density CCD Memory Structures', *IEDM Digest of Technical Papers*; 1981.

The line clock feature ad flexibility to the use of this serial memory. First, the field memory can easily be locked to the TV picture and secondly, independent control of both input and output allows compensation of processing time losses[7].
Fig. 5 shows a chip photograph; its performance is summarised in Table 1.

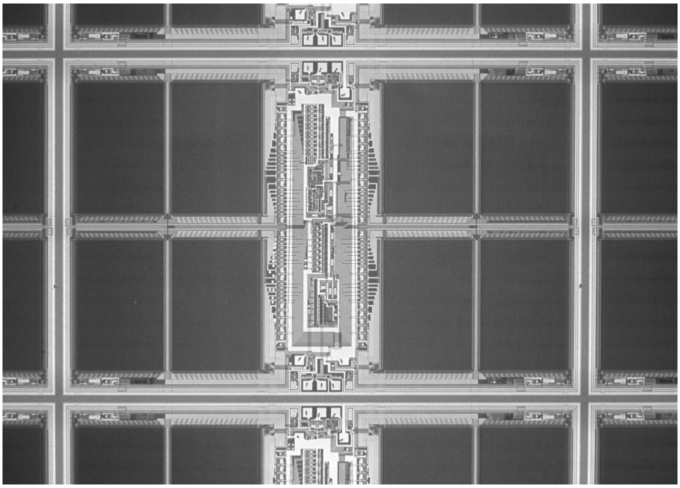


Fig. 5 Photomicrograph of the 308Kb CCD video memory

Table 1 Performance parameters

308Kbit CCD memory	Performance parameters
Memory capacity	308×1024bits
Supply voltages	+5, –2.5V
Power dissipation	350mW
Max. shift frequency	40MHz
Refresh	20ms
I/O and clock levels	TTL-compatible
Ambient temperature	0–70°C
Technology	CCD-nMOS (2μm)
Chip size	34.8mm^2 (53.100 mil^2)
Pinning	Clocks 2 Voltage supply 3 I/O 2 Prog. delay 3

[7] Pelgrom, J.M., Annegarn, M.J., Harwig, H.A., Peuscher, H.J.F., Pfennings, L.C.M.G., Raven, J.G., Slob, A., Slotboom, J.W., Veendrick, H.J.M., ‘A Digital Field Memory for Television Receivers’, *IEEE Trans. Consumer Electronics*, Vol. CE-29, p. 242-250, Aug. 1983.

Chapter 2.1B

An 835Kb Video Serial Memory

Harry J.M. Veendrick et al.

ISSCC 88 / February 17, 1988

1988 IEEE International Solid-State Circuits Conference

The introduction of digital memories in TV and VCR equipment has made it possible to enhance TV pictures with additional features[8]. The reduction of line and field flicker (PAL: 100Hz TV), and noise and cross-colour effects require the storage of part of the TV picture. On the other hand, features like still pictures, fast page access for teletext, multiple picture-in-picture, can also be implemented.

Recent studies of these video concepts revealed an increased demand for high-density low-cost digital video memories.

The CCD concept is known to offer two to three times higher bit density compared to DRAM implementation in the same technology. Moreover, an application-specific design combined with a high bit-rate of the VSM, results in a reduction of system overhead at the printed-circuit board level.

Therefore, an 835Kb surface channel CCD memory has been designed and fabricated in a 1.2μm double-polysilicon/single metal n-well CMOS process. Storage of these 835Kb chips can perform that function.

The elementary CCD cell (16μm^2) consists of a first polysilicon storage gate (on 250Å oxide) and a second polysilicon transfer gate on 400Å (Fig. 1). The 4μm pitch of the double-poly CCD memory cell in both X and Y directions has been realised by anisotropic plasma etching of both polysilicon layers, in combination with an interpoly isolation using CVD oxide spacers[9]. This isolation technique allows the growth of a 400Å thin second-gate oxide, after the formation of 0.2μm thick CVD spacers on the sides of a stack of 0.4μm poly with 0.25μm oxide on top (Fig. 2). The use of CVD oxide spacers is compatible with the formation of LDD N-channel transistors.

Basic CCD operation is performed by four-phase pull-push clocking (Fig. 3).

The memory was organised as 4-bit wide and operates at up to 30MHz. According to the CCIR standard, in the PAL system each field contains 288.5 lines of 720 active samples. Thus, each bit-plane is implemented as a memory block of 290 lines of 720 bits (208.800 bits). To avoid problems with speed, dissipation and transfer efficiency, the data-flow within a bit-plane is de-multiplexed over 6 SPS memory arrays of 26Kb. Such an SPS register is built up out of a serial input register, a parallel transfer register and a serial output register. Operation is comparable with that of large de-multiplexer/multiplexer system. For redundancy purposes, each SPS array was equipped with additional column shift register. If one column fails, the corresponding data bit is repeated at the SPS input and placed in the next column. At the output, data from the faulty column is skipped. This redundancy

[8] Berkhoff, E.J., Kraus, U.E., Raven, J.G., ‘Applications of Picture Memories in Television Receivers’, *IEEE Trans. Consumer Electronics*, Vol. CE-29, p. 251-255; Aug., 1983.

[9] Levin, R.M., Sheng, T.T., ‘Oxide Isolation for Double-Polysilicon VLSI Devices’, *J. Electrochem. Soc.*, Vol. 130, No. 9, 1984.

operation is controlled by a counter whose state is compared with the contents of a fuse ROM in which the number of the faulty columns is stored during testing.

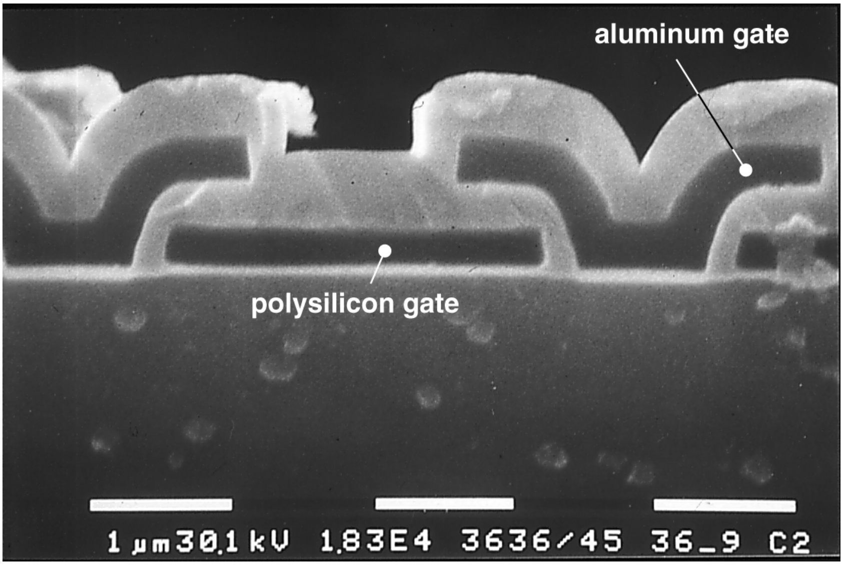


Fig. 1 SEM photograph of double-polysilicon CCD structure

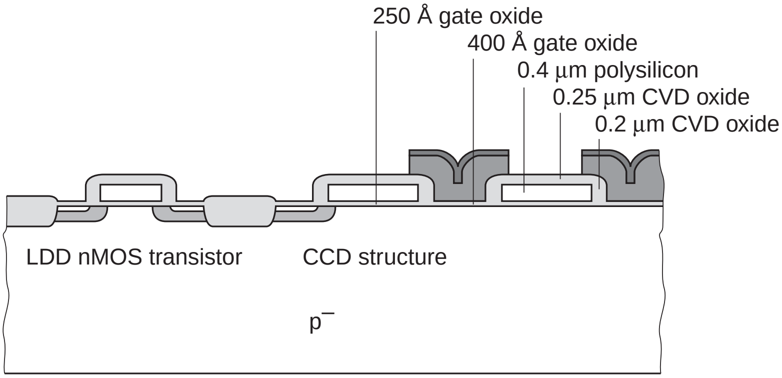


Fig. 2 Process cross-section

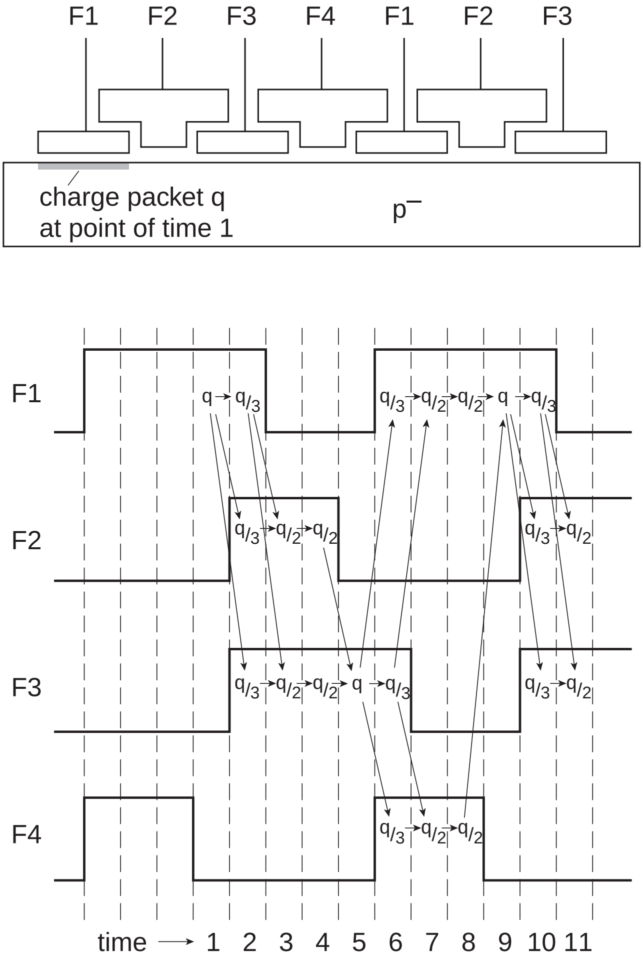


Fig. 3 Four phase pull-push clocking

A block diagram of the chip is shown in Fig. 4. It has several operation modes to facilitate its use in digital video systems:

1- In the 258-lines mode the chip can be used in NTSC systems where the number of TV lines is 525. One field contains 240 active lines. In the chip the number of lines can virtually be reduced to 258 per field. The remaining 18 lines can be shifted out by means of a fast *feed-through* operation at increased (parallel) clock frequencies.

2- The normal configuration for field storage is a parallel mode (switches S in position O and R in position 1). In this mode the memory is used as a 208Kb by a 4 shift register with four inputs and four outputs running at a clock frequency of 15MHz.

3- As pinning becomes a severe problem, the four inputs and outputs can be time multiplexed over two pins each, at a clock frequency of 30MHz saving four I/O pins per memory.

4- The serial more (S = R = 1) allows an 835Kb by 1 operation of the chip. This mode is used in teletext storage applications.

5- In the scan mode it is possible now to select one out of four 208Kb array outputs. This selection is controlled by a *running-one* in an on-chip 4-bit shift register.

6- The re-circulate mode (R = O) forces the data to be re-circulated internally. This allows for still picture applications.

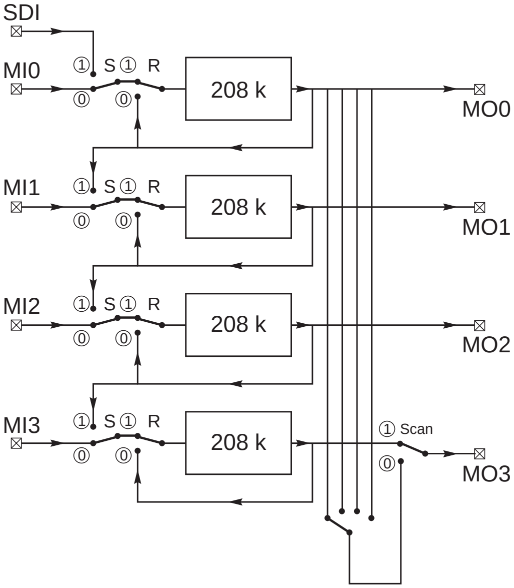
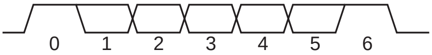


Fig. 4 Block diagram

The memory requires no addressing and is controlled with two clocks: memory clock and memory gating. The modes of operation are controlled via 5 message bits in the gating (Fig. 5).

Fig. 6 shows a microphotograph of the chip. The memory consists of 32 arrays, each containing 26Kb. The control logic is situated at the top. The total number of bonding pads is 18. At a maximum clock frequency of 30MHz the chip typically consumes 250mW. Size is 7.8mm by 3.7mm, which results in a die area of 29mm^2.



NUMBER	MESSAGE
0	start
1	258/ 290 mode
2	tristate output control
3	serial / parallel mode
4	normal / time multiplex data
5	data input control
6	stop

Fig. 5 Gating signal containing a coded message

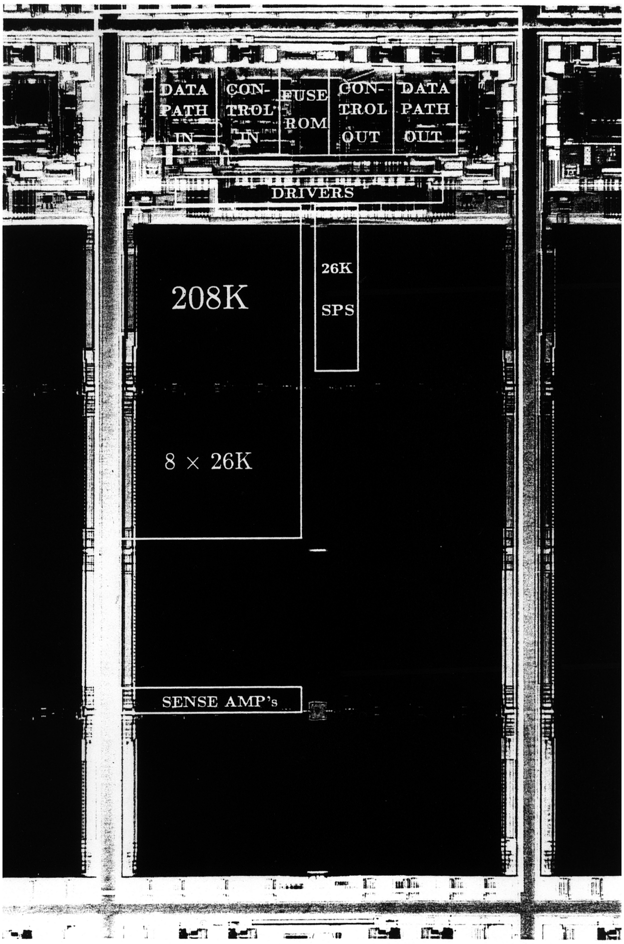


Fig. 6 Microphotograph of chip

Chapter 2.2

An Efficient and Flexible Architecture for High-Density Gate Arrays

Harry J.M. Veendrick, Member, IEEE, Dré A.J.M. van den Elshout,
Dick W. Harberts, and Teus Brand

IEEE Journal of Solid-State Circuits

October 1990, Volume 25, Number 5

Abstract

During the past decade several high-density gate array (HDGA) architectures have been proposed to accommodate a dense integration of logic circuits. This chapter describes an efficient and flexible HDGA architecture (or sea of transistors) with cells containing three common-gate wide and small transistors on which both logic and memory functions can be relatively densely mapped. The use of titanium-silicide straps for local interconnect, as an alternative to the third metal layer, is evaluated through different designs. Finally, the design and performance of an experimental chip in 0.8μm CMOS technology is discussed.

1 Introduction

In many high-density gate array (HDGA) architectures, a basic cell or gate usually consists of four to eight transistors [1,2]. A typical example of such an architecture is shown in Fig. 1.

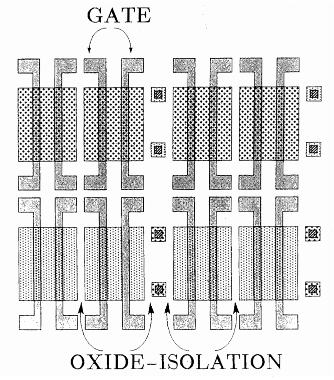


Fig. 1 Typical example of a sea-of-gates architecture

These gates are isolated by means of oxide isolation. This type of HDGA is often referred to as ‘sea of gates’. Another type of isolation, which allows for a more flexible placement of the logic cells, is the gate-isolation technique [3]. Most HDGAs, which use this technique, have an architecture similar to that shown in Fig. 2a. Such a ‘sea-of-transistors’ array, however, offers nMOS and pMOS transistors of only one size each. There are circuits like transmission-gate flipflops, ROMs, RAMs, and PLAs, which especially require transistors of different sizes. Dynamic CMOS circuits also benefit from this flexibility. Fig. 2b shows a sea-of-transistors architecture, as proposed in [4], in which the nMOS transistor is split up into two smaller ones in parallel. Each transistor gate has its own connections. However, in memory arrays many transistor gates usually share a common word line and thus do not require individual connections. This basic idea is applied in the development of a new HDGA architecture.

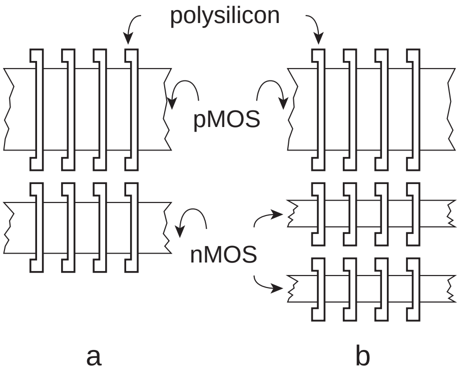


Fig. 2 (a) Typical example of a sea-of-transistors architecture
(b) Sea-of-transistors architecture with multiple nMOS transistors

2 The New Architecture and its Usage for Memory Structures

Fig. 3 shows the new architecture in which each basic cell provides three nMOS and three pMOS transistors. Every wide nMOS transistor lies in between two smaller ones. Similarly, every wide pMOS transistor lies in between two smaller ones. The transistor sizes in the figure are related to a 0.8μm CMOS technology. Both the nMOS and pMOS transistors each share a common gate. The contact positions in both horizontal and vertical directions are on the same grid, defined by a metal pitch (including contact holes) of 3.6μm. This is so designed as to reduce the dependency on specific software. The advantages of such an architecture can be fully exploited in memory and logic array structures like ROM, RAM, and PLAs. A ROM array, as shown in Fig. 4, uses all available transistors of both n- and p-type. The bit lines run horizontally in first metal, while the word lines run vertically in second metal. The common-gate architecture can lead to a three times denser ROM array than in former architectures. Performance figures will be presented in combination with the experimental chip design. The implementation of static RAM cells is shown in Fig. 5. These cells only consist of small transistors. To guard against parasitic writing during a READ operation, transistor N_2 must be at least twice the width of transistor N_1 [5]. In conventional HDGAs the required ratio between transistors N_2 and N_1 can be achieved by implementing N_1 as a series connection of two wide transistors. This, however, requires a large cell area. In the presented common-gate architecture, transistor N_2 can be implemented by a parallel connection of two small nMOS transistors, reducing the RAM cell area by half.

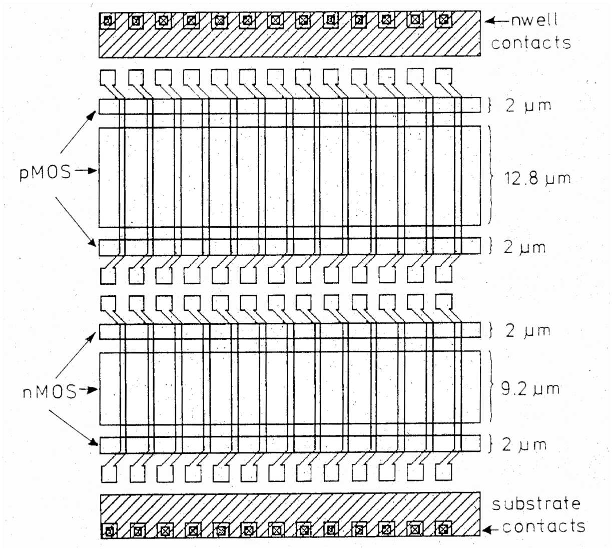


Fig. 3 The new common-gate HDGA architecture

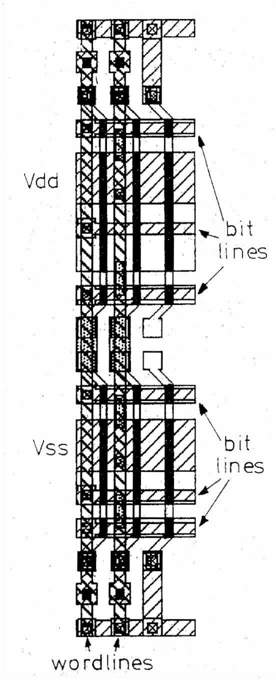


Fig. 4 Implementation of ROM cells on the common-gate architecture

In Fig. 5, word lines run horizontally in first metal, while shared bit lines run vertically in second metal. Performance figures will also be presented further on in this chapter (see Table 1).

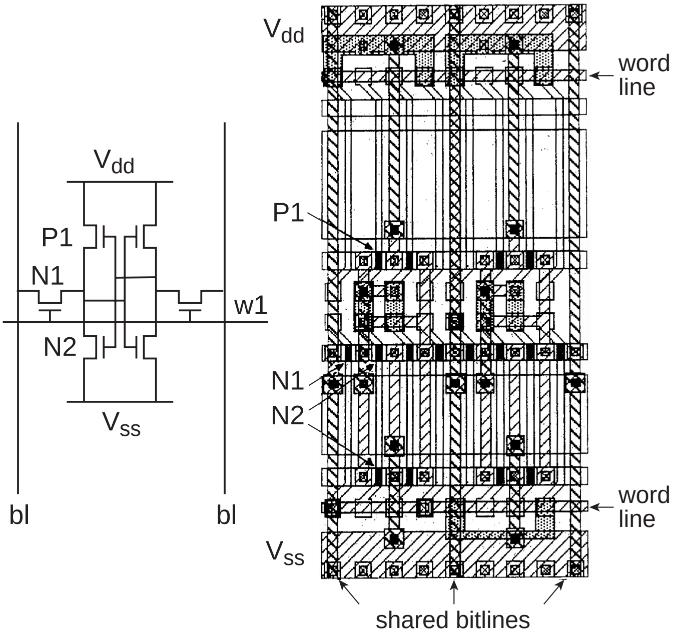


Fig. 5 Implementation of RAM cells on the common-gate architecture

3 Routing Aspects in Relation with Technology Options

An HDGA not only benefits from dense memory structures, but also from its ability to generate an efficient implementation of logic circuits. The transistor utilisation depends on the flexibility of the architecture for routing. Therefore different technology options have been used to investigate routing aspects. One of the more complex cells in a library is the *D*-type flipflop, which consists only of logic gates and two driver circuits (Fig. 6) and numbers 30 transistors.

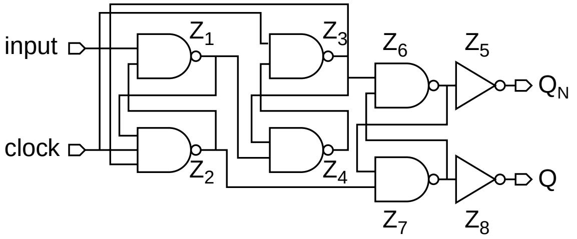


Fig. 6 Logic diagram of a *D*-type flipflop

This *D*-type flipflop is mapped onto the common-gate architecture and routed with two metal layers (Fig. 7). For such normal logic functions only the wide nMOS and pMOS transistors are used. The small sized transistor regions are then used for wiring and/or allocation of second to first metal vias. All potential contact locations, both horizontally and vertically, are on the same grid of 3.6μm. First metal runs horizontally, while second metal is used for vertical wiring.

The figure shows that when we route the library cells with two metal layers, only four spare horizontal tracks in first metal remain for inter-cell wiring. Thus, the transparency of the cell is about 25% of the horizontal pitches and about 50% of the vertical pitches. A two-metal-layer technology is thus not suited for an efficient implementation of logic circuits. Its transistor utilisation will always be less than 50%.

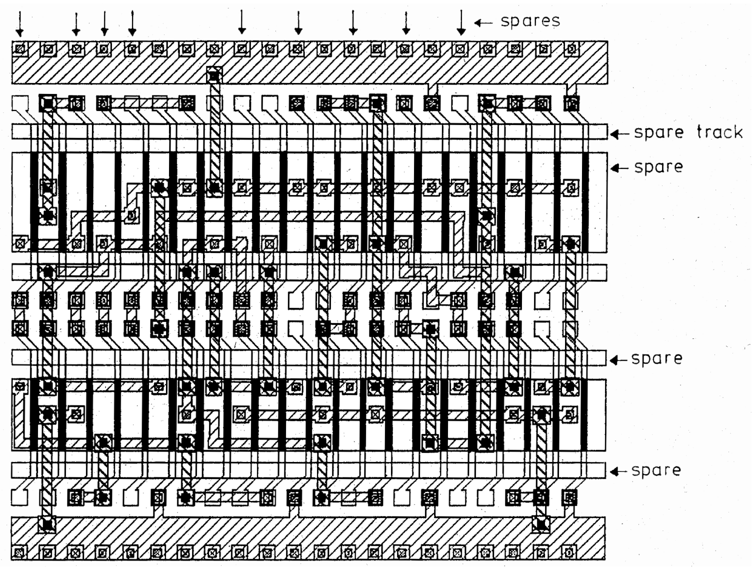


Fig. 7 *D*-type flipflop realisation in two-metal-layer technology

Triple-metal (Bi)CMOS processes are at the present used to implement HDGAs. However, the additional third metal layer for customisation can increase the silicon cost by up to 25% and processing turnaround time by up to 35%. Replacing this expensive third metal layer with a titanium-silicide ($TiSi_2$) layer increases the silicon cost and processing time by no more than 5%. Connections in this layer, called straps, are used to bridge only short distances, such as those within library cells. In the example of the *D*-type flipflop, several interconnections in metal 1 (Fig. 7) are replaced by straps.

The result is shown in Fig. 8 in which the straps are used for polysilicon-to-polysilicon interconnections. The straps can also be used for connecting transistors in parallel for increased driving capability.

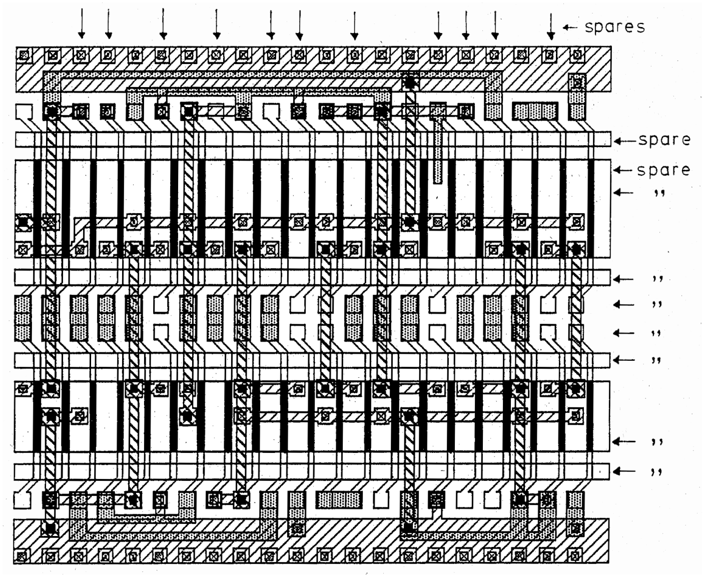


Fig. 8 *D*-type flipflop realisation in a two-metal-layer and straps technology

As a result of combining two metal layers with strap routing, the transparency of the cell for global interconnect has increased to 50% of both horizontal and vertical pitches. As will be shown later, this number is high enough to fulfil the global routing needs of a large part of today’s HDGAs.

To complete the routing comparison the *D*-type flipflop was also routed using a three-metal-layer process (Fig. 9). Now, the cell is completely routed in first metal only and it thus remains 100% transparent in both directions for second- and third-metal inter-cell wiring. This, of course, allows for more complex wired circuits in the future.

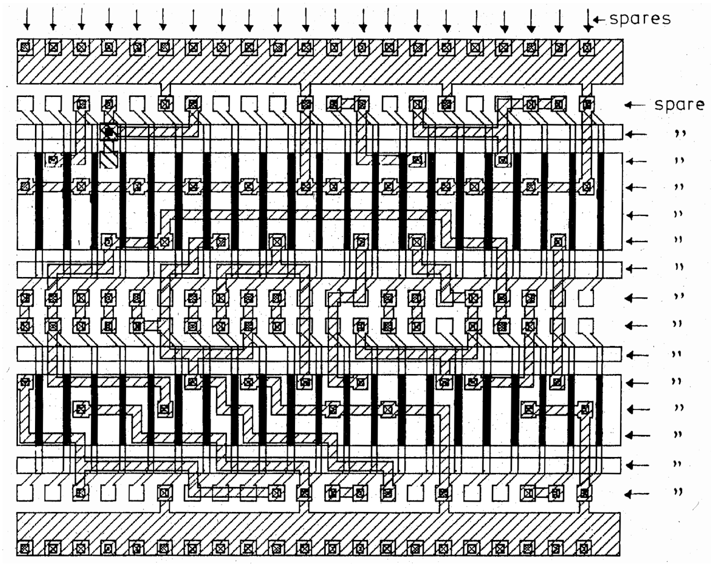


Fig. 9 *D*-type flipflop realisation in a three-metal-layer technology

4 Experimental Chip Design and Results

To prove the flexibility of the common-gate architecture, an experimental chip was developed. It included several implementations of a 10 × 10bit fully pipelined multiplier, a 24kb ROM, and some performance and technology evaluation modules.

First, a custom standard-cell version of the 10×10bit multiplier was designed with commercially available place and route software. This design consisted of only five different logic cells, which were mapped onto the common-gate architecture in a 'two metal layers with straps technology'. Fig. 10 shows the five different cells, which include the previously discussed *D*-type flipflop, an EXCLUSIVE OR, a NAND, an AND, and an inverter. In each cell the same nine horizontal tracks are transparent for inter-cell routing. The complexity and relative placement of these cells is kept exactly the same as in the standard-cell version. Then this multiplier was routed with internally available routing software [6].

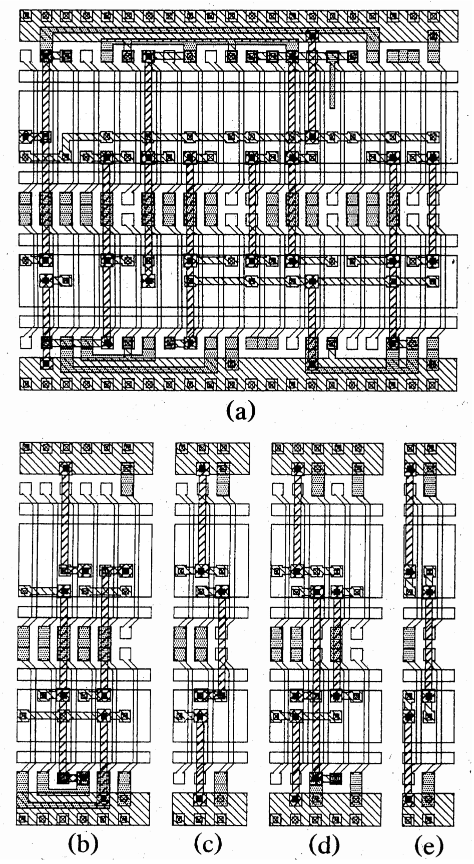


Fig. 10 The five different cells of the experimental multiplier: (a) *D*-type flipflop, (b) EXCLUSIVE OR, (c) NAND, (d) AND, and (e) inverter

The result is shown in the photograph of the experimental chip in Fig. 11. A three metal-layer version of the multiplier is also mapped onto the same common-gate architecture and shown in the photograph. It shows that the chip areas of both HDGA versions are equal to that of the standard-cell version. Because their net lengths and transistor sizes are also about equal, no major performance difference can be concluded.

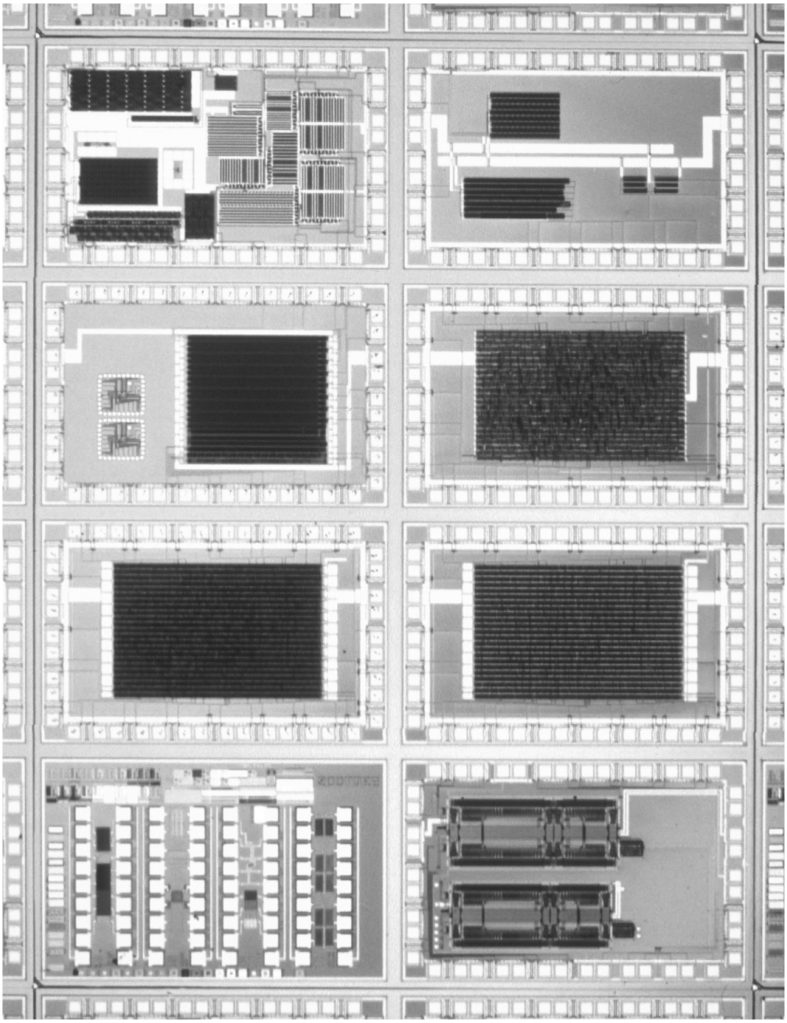


Fig. 11 Experimental chip, containing 24kb ROM, three versions of a 10×10bit multiplier, and performance and technology test modules

Fig. 12a and 12b shows the wiring complexity of the 10×10bit multiplier routed with three metal layers and two metal layers with straps, respectively. For each technology the middle and bottom figures represent the inter-cell routing layers. The difference in wiring densities of these layers shows that the three-metal-layer technology still allows integration of much more complex circuits. Another chip module (Fig. 11) shows a 24bit ROM, in which each bit line contains 128 ROM cell connections. The measured access time, worst case via the small pMOS transistors, was 30ns Its worst-case cycle time is 50ns. The other chip modules contain technology and performance modules such as ring oscillators and full-adder delay chains to determine figures on power dissipation and delay. The average gate delay (of a two-input NAND with fan-out of 2) is 430ps.

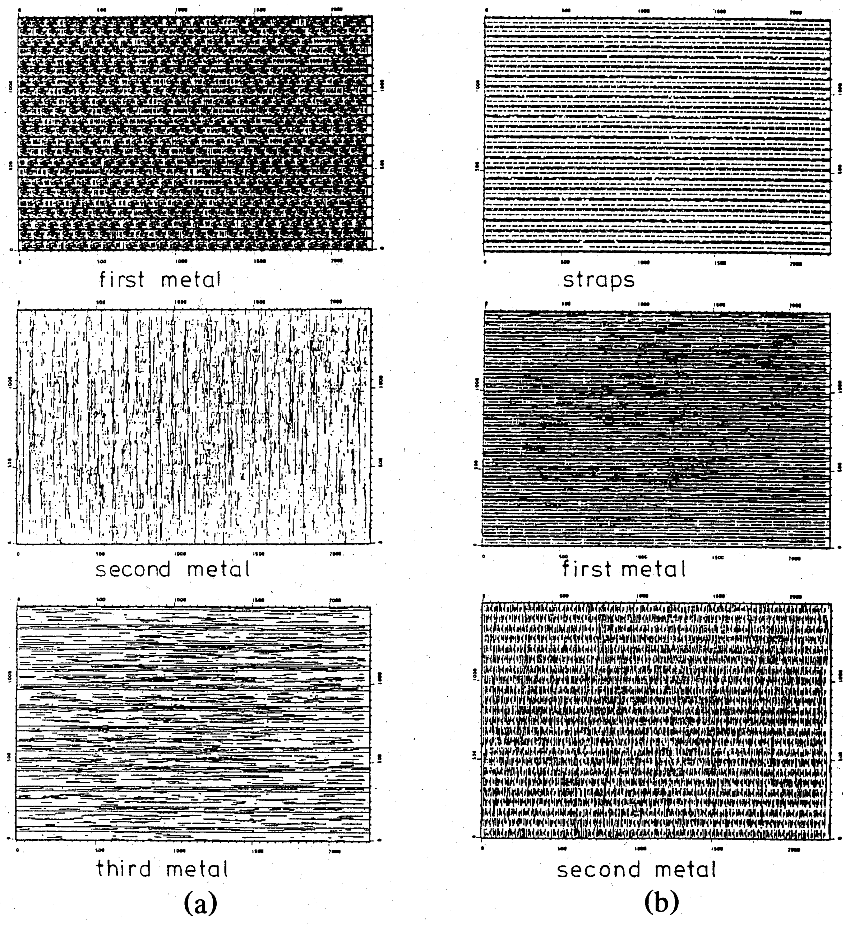


Fig. 12 Wiring levels of different multiplier realisations in: (a) three metal-layer technology, and (b) two metal-layer and $TiSi_2$-straps technology

The chip is fabricated in a 0.8μm CMOS technology with a titanium-silicide layer and triple metal layers with a pitch of 3.6μm, when including contact holes. For the different designs several technology options were used. The ROM was designed with two metal layers and straps and fills an area of 2.2mm^2 including the control logic. Table 1 summarises the chip data and performance. Note that the density of ROM and RAM cells is relatively high compared to former sea-of-gates architectures.

Table 1 Chip data and performance figures

Technology	• **0.8μm CMOS** – **option: 2 metal layers and TiSi2 straps** – **option: 3 metal layers** • **metal pitch: 3.6μm for all layers (including 2 contacts)**	
24kbit ROM	area: 2.2mm^2 (control logic included)	
10 by 10 MPY	area	wiring
Full custom HDGA 1 HDGA 2	3.4mm^2 3.5mm^2 3.5mm^2	2 metal layers 2 metal layers + straps 3 metal layers
Logic density ROM density sRAM density	1610gates/mm^2 (2-input NAND equivalents) 17k cells/mm^2 1.1k cells/mm^2	
Performance	delay/gate: 430psec (2-input NAND: fanout 2) power/gate/MHz: 7.5μW	

5 Evaluation of the architecture for complex wired circuits

The previously discussed multiplier was taken as an example because of its ease of design: it contained only five different library cells. However, the complexity and regularity of the multiplier is probably not representative for random logic circuits. So, a number of different designs were evaluated. These results are listed in Table 2.

Table 2 Comparison of various designs in standard cells and high-density gate array

		Standard cell			**Common-gate HDGA**			
design	#gates (2-nand)	aspect ratio	tracks/ row	# rows	aspect ratio	# rows	area ratio HDGA/SC	transistor utilisation
MPY (20×20)	20800	1.50	9	48	1.50	48	0.97	96%
MPY (10×10)	5000	1.51	8	23	1.51	23	1.00	95%
BLOCK1	1350	2.06	13	12	2.34	12	1.00	94%
FFT A	2637	1.02	13	20	1.07	20	0.68	87%
FFT B (1)	1980	1.06	14	19	1.03	20	0.77	86%
FFT B (4)	1980	3.92	15	9	3.92	11	0.82	74%
FFT B (1/4)	1980	0.33	9	38	0.27	40	0.82	81%

Table 2 shows a comparison of different logic designs implemented in standard cells and on an HDGA architecture in the ‘two-metal-layer and straps common-gate CMOS technology’. The discussed 10×10b fully pipelined multiplier and a more complex 20×20b multiplier occupy about the same area as their standard-cell counterparts. A complex logic block of a compact disc servo control chip, indicated as ‘block 1’ here, occupies the same area as the standard-cell version. FFT *A* represents a fast Fourier transform design of which the HDGA implementation is only 68% of the standard-cell version. Another fast Fourier design (FFT *B*) containing 1980 gates has been implemented with different aspect ratios. This is done to investigate the aspect ratio influence on the area. The table shows here also that the HDGA implementations occupy less area than their standard-cell counterparts for aspect ratios of one, four, and a quarter. In these design examples, which were placed and routed with commercially available software, the transistor utilisation varies from 74% for an elongated shape of a fast Fourier transform design to 96% for a complex multiplier design.

6 Conclusions

A common-gate high-density gate array architecture has been presented that allows a relatively dense integration of both logic and memory functions. The use of titanium-silicide straps for local interconnect is a good and cheap alternative to an additional third metal layer for a wide range of today’s circuits. This architecture has been evaluated through the development of an experimental chip containing memory and logic modules, and through a comparison of many different standard-cell and common-gate HDGA designs. The HDGA implementations showed equal performance at comparable or even smaller chip areas.

7 References

[1] T. Wong *et al.*, ‘A high performance 129k gate CMOS array’, in *Proc. CICC*, 1986, pp. 568-571.

[2] H. Takahashi *et al.*, ‘A 240k transistor CMOS array with flexible allocation of memory and channels’, in *ISSCC Dig. Tech. Papers*, Feb. 1985, pp. 124-125.

[3] I. Okhura *et al.*, ‘Gate isolation –A novel basic cell configuration for CMOS gate arrays’, in *Proc. CICC*, 1982, pp. 307-310.

[4] P. Duchene *et al.*, ‘A highly flexible sea-of-gates structure for digital and analog applications’, *IEEE J. Solid-State Circuits*, vol. 24, pp. 576-584, June 1989.

[5] E. Seevinck *et al.*, ‘Static-noise margin analysis of MOS SRAM cells’, *IEEE J. Solid-State Circuits*, vol. SC-22, pp. 748-754, Oct. 1987.

[6] J. Jess *et al.*, ‘A gate array design system adapted to many technologies’, in *Proc. ICCAD*, 1984, pp. 338-343.

Chapter 3.1

The Behaviour of Flipflops Used as Synchronisers and Prediction of Their Failure Rate

Harry J.M. Veendrick

IEEE Journal of Solid-State Circuits

April 1980, Volume 15, Number 2

Abstract

This chapter deals with the behaviour of flipflops, used as input synchronisers, in particular when they operate in the meta-stable region.

It is shown first, theoretically as well as experimentally, that the average rate of system failures, due to the occurrence of meta-stable states (MSSs), is independent of circuit noise.

A formula, which describes the probability of occurrence of a meta-stable state, has been derived. To verify the theory, measurements have been made on a flip-flop made in n-channel MOS technology. Although the theory is more generally applicable, in this chapter it will mainly be applied to the design of a synchroniser made in MOS technology.

A method is also given for predicting the average number of system failures, for a given flipflop, occurring over a year. This method is applied to predict this average failure for the designed synchroniser.

1 Introduction

In the communication between digital subsystems that do not share a common time reference, signals may occur which are not logically defined [1]. The interactions between these systems are asynchronous.

An example is a multi-microprocessor system in which the microprocessors do not share a common clock. This means that signals are generated randomly in time. It is particularly possible, that an input signal will change during a sample clock edge. This kind of situation can cause a system failure. The usual treatment of this problem is to design a synchroniser (most commonly a flipflop) which has to take care of reliable communication between asynchronous subsystems.

Synchronisers have been the subject of some papers [1]–[6]. This chapter draws attention to the aspect, that circuit noise does not affect the average number of meta-stable states (MSSs in Fig. 1a), which occur during a certain period of time. Couranz and Wann [2] have already mentioned this phenomenon, but only as a special case. In addition to this, design considerations are given for synchronisers made in MOS technology. According to these considerations a synchroniser has been designed for four-phase MOS logic. Moreover, a method for predicting the average number of system failures occurring in the course of a year is given. This method has been applied for the designed synchroniser.

2 Theory

2.1 Noise Independence

In digital systems which communicate asynchronously, situations may occur in which the input signal is changing (e.g., from a logical 0 to a logical 1) at the falling edge of the sample clock Φ, depicted in Fig. 1b. For better understanding, slow rising and falling edges of the input signal are drawn in this figure. The flipflop connected to the sample transistor (drawn as an MOS transfer gate in the figure) will reach a MSS (in Fig. 1a) when the falling edge of the sample clock appears within a small time interval $\delta(t)$ during a rising or falling edge of the input signal. Because input signal edges appear randomly in time, the sample values during these edges are uniformly distributed.

Circuit noise, which has a normal distribution with a zero mean, is superimposed on all these uniformly distributed sample values of the input signal. This superposition again has a nearly uniform distribution. Fig. 2 gives an illustration of this superposition and it shows that the distribution remains uniform over a wide range of sample values.

Let V_m be the sample voltage at which both flipflop nodal voltages are of the same height ($V_1 = V_2 = V_m$ = meta-stable voltage). The interesting region $\delta(v)$ around $V = V_m$ is the meta-stable region in which the superposition has a uniform distribution function (Fig. 2). Noise does not therefore affect the distribution of sample values at the sample moment.

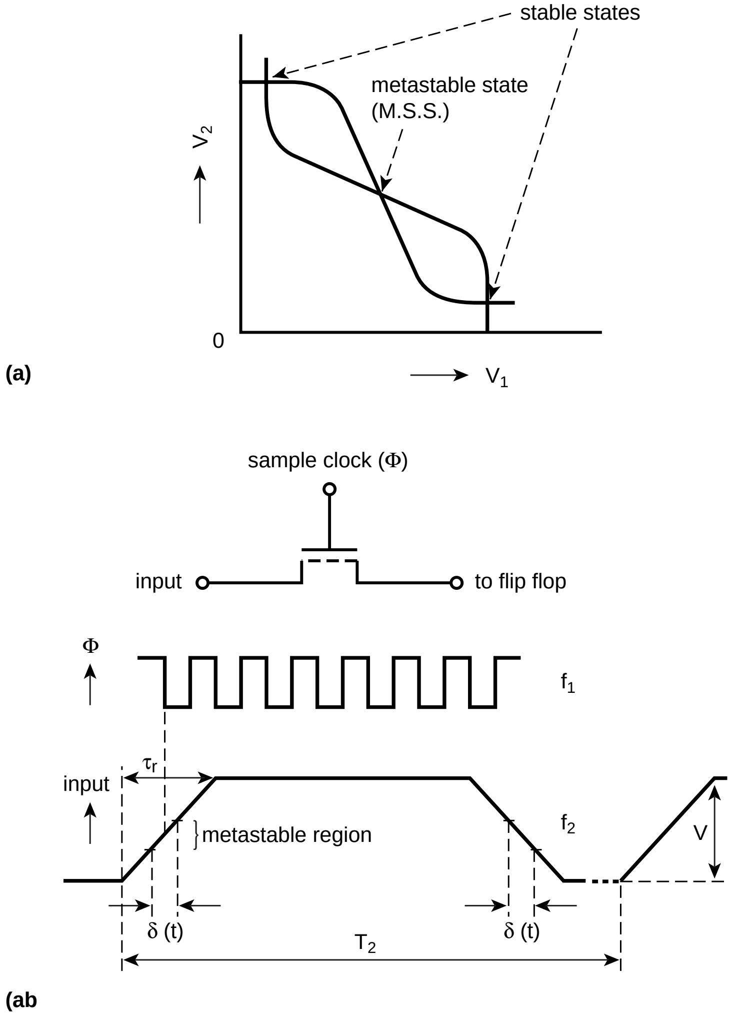


Fig. 1 (a) Possible states of a flipflop, consisting of two inverters of which the transfer characteristics are given
(b) Possible signal configuration at asynchronous communication

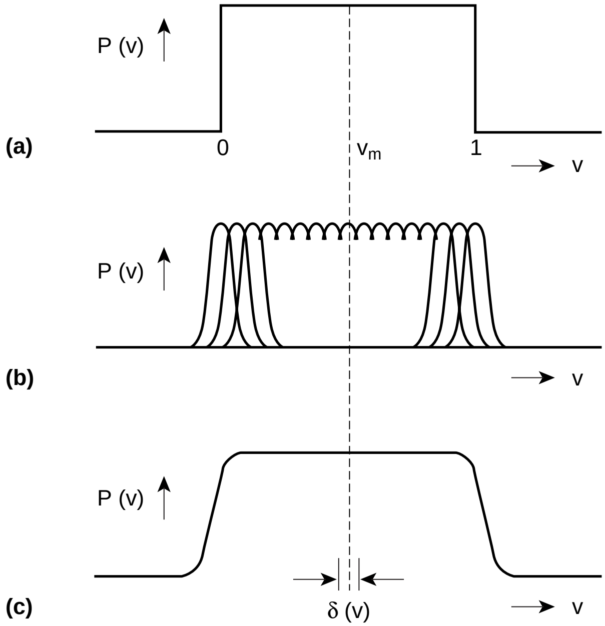


Fig. 2 (a) The (uniform) distribution function of the sample values
(b) Circuit noise effect on the uniform distribution function
(c) Superposition of circuit noise on a uniform distribution function

To show that the distribution of voltage values after the sample moment remains uniform, we must first consider the solutions of the differential equations, which describe the flipflops small signal behaviour (around the MSS).

As a first approximation, a flipflop can be modelled by two stages, each containing an inverter with amplification –*A* followed by an *RC*-filter with $\tau = RC$ (Fig. 3). The solutions of the differential equations are [7] (all voltages are with respect to the voltage V_m)

$$v_1 = \lambda_1 \cdot \exp\left(\frac{A-1}{\tau} \cdot t\right) + \lambda_2 \cdot \exp\left(\frac{-A-1}{\tau} \cdot t\right) \tag{1}$$

and

$$v_2 = -\lambda_1 \cdot \exp\left(\frac{A-1}{\tau} \cdot t\right) + \lambda_2 \cdot \exp\left(\frac{-A-1}{\tau} \cdot t\right), \tag{2}$$

where λ_1 and λ_2 are integration constants.

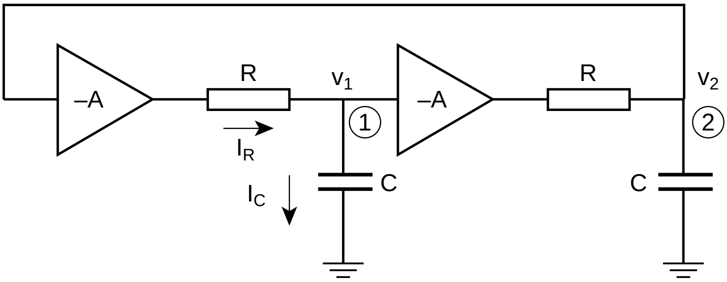


Fig. 3 First-order small-signal model of a flipflop

The initial conditions (at t = 0) can be expressed as

$$v_1 = v_{01} = \lambda_1 + \lambda_2 \qquad (3)$$

and

$$v_2 = v_{02} = -\lambda_1 + \lambda_2; \qquad (4)$$

hence

$$v_1 = \frac{v_{01} - v_{02}}{2} \cdot \exp\left(\frac{A-1}{\tau} \cdot t\right) + \frac{v_{01} + v_{02}}{2} \cdot \exp\left(\frac{-A-1}{\tau} \cdot t\right) \qquad (5)$$

and

$$v_2 = \frac{-v_{01} + v_{02}}{2} \cdot \exp\left(\frac{A-1}{\tau} \cdot t\right) + \frac{v_{01} + v_{02}}{2} \cdot \exp\left(\frac{-A-1}{\tau} \cdot t\right). \qquad (6)$$

If we only consider node ① and, without loss of generality, we assume

$$v_1 = v_{01} \cdot \exp\left(\frac{A-1}{\tau} \cdot t\right). \qquad (7)$$

Thus, at a time $t = t_c$ after the sample moment (t = 0) each sample value $v = v_{01}$ is multiplied by a factor $\exp((A\text{-}1)\cdot t_c/\tau)$.

As the values of v_{01} are uniformly distributed, the values of $v_1 = v_{01}\cdot\exp((A\text{-}1)\cdot t_c/\tau)$ are uniformly distributed too, at any time. This behaviour of v_1 is illustrated in Fig. 4a. In equally sized regions $\delta(v)$ there are an equal number of states N, owing to the uniform distribution (see Fig. 4b). At a time $t = t_c$ we superimpose a disturbing voltage with amplitude v_A at node ① of the flipflop. The number of states N_1 within region R_1 of size $\delta(v)$ that is forced out of the meta-stable region is replaced by the number of states N_2 within region R_2 of the same size that is forced back into the meta-stable region. As $N_1 = N_2$, it follows that such a disturbing voltage does not affect the average number of MSSs, over a large number of events, that lasts longer than a certain time. The same considerations apply to circuit noise, which can be considered as a sequence of disturbing voltages with a zero mean. Summarising, the following statement can be made. At the sample moment, as well as at any later time, circuit noise does not affect the average state distribution, and consequently does not affect the average duration of MSSs that are sampled with the same v_0.

In section 3.2, measurements are described which agree with the above analyses.

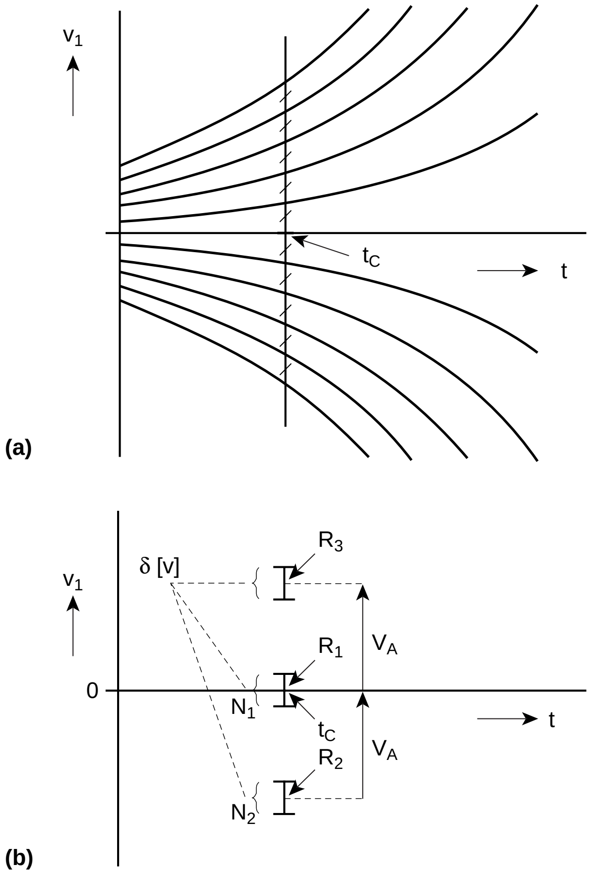


Fig. 4 (a) Behaviour of the nodal voltages of the flipflop at uniformly distributed sample values (at t = 0)
(b) The number of states N_1 that is forced out of the meta-stable region by a disturbing voltage V_A at a time $t = t_c$, is replaced by the number of states N_2 that is forced back into the meta-stable region

2.2 The Average Number of MSSs Versus Time

As the sample values during the edges (at t = 0) are uniformly distributed, the probability of sampling a voltage $v \leq v_n$ (Fig. 5) is

$$P\left(|v| \leq v_n\right) = \frac{v_n}{v_d} \cdot P\left(|v| \leq v_d\right), \tag{8}$$

where v_n and v_d are arbitrary voltages (with $v_n < v_d$). In Appendix 2 the following relationship between v_n and v_d is derived:

$$v_n = v_d \cdot \exp\left(-\frac{A-1}{\tau} \cdot \left(t_n - t_d\right)\right), \tag{9}$$

where t_n and t_d are the corresponding durations of meta-stable states which were sampled with voltages v_n and v_d, respectively.

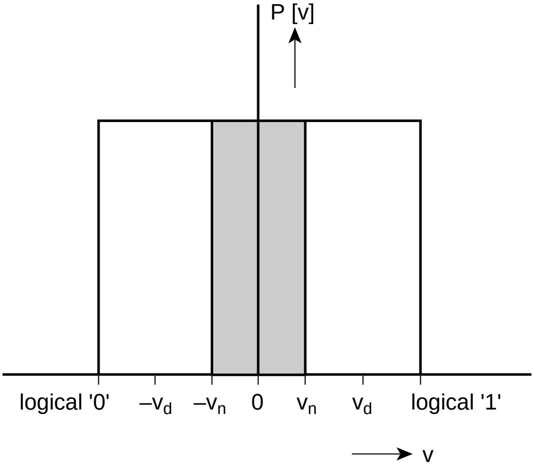


Fig. 5 Uniform distribution function of the sample values

Formulas (8) and (9) lead to

$$P\left(|v| \leq v_n\right) = \exp\left(-\frac{A-1}{\tau} \cdot \left(t_n - t_d\right)\right) \cdot P\left(|v| \leq v_d\right). \tag{10}$$

If v_d is equal to a logical one, and $-v_d$ is equal to a logical zero, then: $P\left(|v| \leq v_d\right) = 1$ and $t_d = 0$. Thus, (10) reduces to

$$P(|v| \le v_n) = \exp\left(-\frac{A-1}{\tau} \cdot t_n\right). \quad (11)$$

Now, the probability of sampling a value $|v| \le v_n$ is equivalent to the probability of occurrence of an MSS, whose duration t is longer than t_n; hence

$$P(t > t_n) = \exp\left(-\frac{A-1}{\tau} \cdot t_n\right). \quad (12)$$

So, if we take N_0 samples, uniformly distributed between logical '0' and logical '1', the average number M_n of MSSs lasting longer than a time t_n is

$$M_n = N_0 \cdot \exp\left(-\frac{A-1}{\tau} \cdot t_n\right). \quad (13)$$

This formula is verified in the measurements described in Section 3.1.

3 Measurements

Measurements have been made on an existing flipflop made in n-channel MOS technology (Fig. 6).

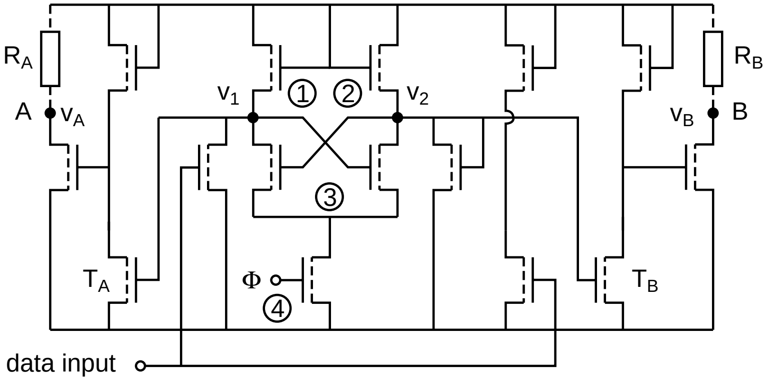


Fig. 6 Experimental MOS device

It should be noted that this flipflop was not intended to be used as a synchroniser; it merely served as a good experimental device. The design of the flipflop is such,

that the threshold voltage of both transistors T_A and T_B is much less than the metastable voltage $V_1 = V_2 = V_m$; this excludes the situation $V_A = 0$ and $V_B = 0$.
The resistors R_A and R_B are connected externally. The output nodes A and B are connected to an AND gate (see Fig. 7a). The output of the AND gate, together with a strobe pulse, is applied to a NAND gate whose output is connected to a counter. The strobe signal is applied a time Δt after the rising edge of the clock Φ (see Fig. 7b), in order to count the number of MSSs whose duration is longer than the time Δt.

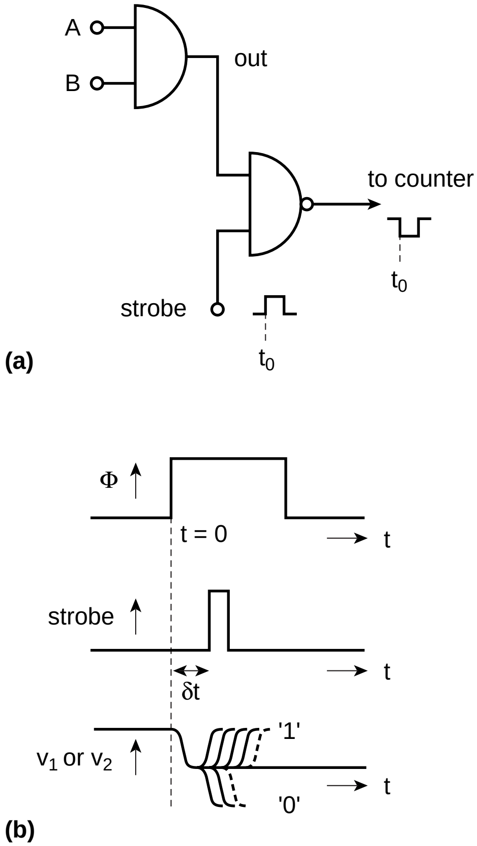


Fig. 7 (a) Counting circuit
(b) Signal configuration for counting

3.1 The Number of MSSs Versus Time

To stimulate a uniform distribution of sample values, we adjust the flipflop in the meta-stable region by means of a dc-voltage on the data input (Fig. 6), at which a saw-tooth voltage is superimposed. The frequency of this saw-tooth voltage is not correlated with the sample frequency, so a uniform distribution of sample values is obtained. During the measurements, the clock frequency at node ④ was equal to 2.5MHz. The measurements were made for 1 min at several values of Δ*t*, which means for each Δ*t* there are 1.5×10^8 states generated. Only those MSSs lasting longer than the time Δ*t* are counted. The results are given in Fig. 8.

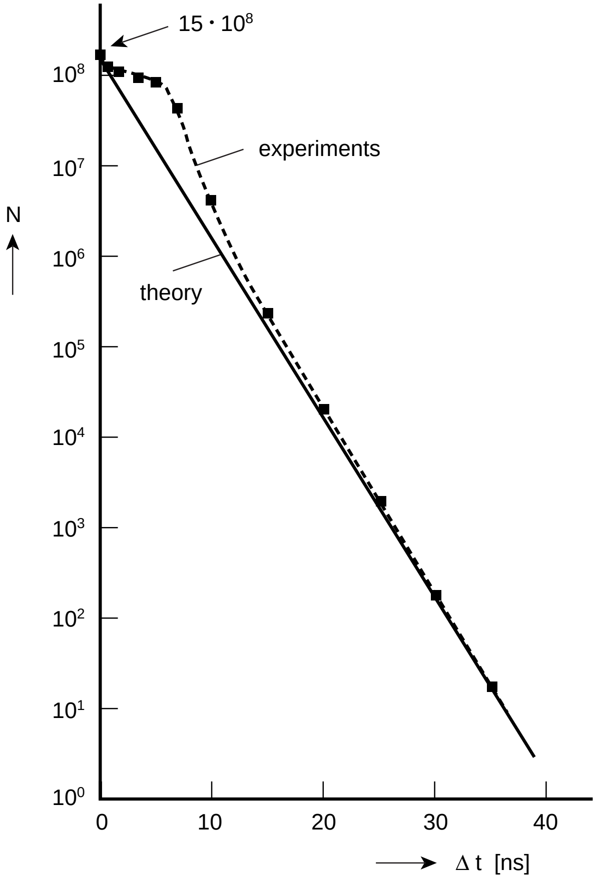


Fig. 8 The average number of MSSs which lasts longer than a time Δ*t* versus Δ*t*, over a large number of events, theoretical and experimental results

The theoretical results for this flipflop (with τ = 12ns and A = 3.4), in accordance with (13), are also given in this figure.
The difference between theory and measurement for small values of Δt is caused by the fact that the second term in (5) may not be neglected – which was done in the derivation of (13) (Appendix 2) – for these short MSS durations. For larger values of Δt theory and experimentation are in very close agreement.

3.2 Verification of Circuit Noise Independence

The adjustment of the flipflop around its meta-stable region has already been described in Section 3.1. Because of the fact that the saw-tooth generator is affected with noise – which can be seen as circuit noise on the flipflop nodes – the experiments described in Section 3.1 were done in a noisy environment. This noise was filtered away by means of a small *RC*-filter. Then the measurements described in Section 3.1 were made again. The resulting diagram was exactly the same as the one obtained in a noisy environment. The measurements show the statement made at the end of Section 2.1 to be true.

4 Design Aspects of a Synchroniser Made in MOS Technology

The probability of occurrence of an MSS whose duration t is longer than a time t_n is given by (12)

$$P(t > t_n) = \exp\left(-\frac{A-1}{\tau}\cdot t_n\right) \tag{12}$$

To decrease this probability the factor $(A - 1)/\tau$ has to be maximised. For flipflops made in MOS technology it is very practical not to use the small signal model of Fig. 3, but the one of Fig. 9a instead. So for MOS flipflops, the factor $(A - 1)/\tau$ can be written as

$$\frac{A-1}{\tau} = \frac{SR-1}{RC} = \frac{S}{C} - \frac{1}{RC} \tag{14}$$

where S represents the transconductance of the driver transistors, R the differential resistance of the loads, and C the nodal capacitance.

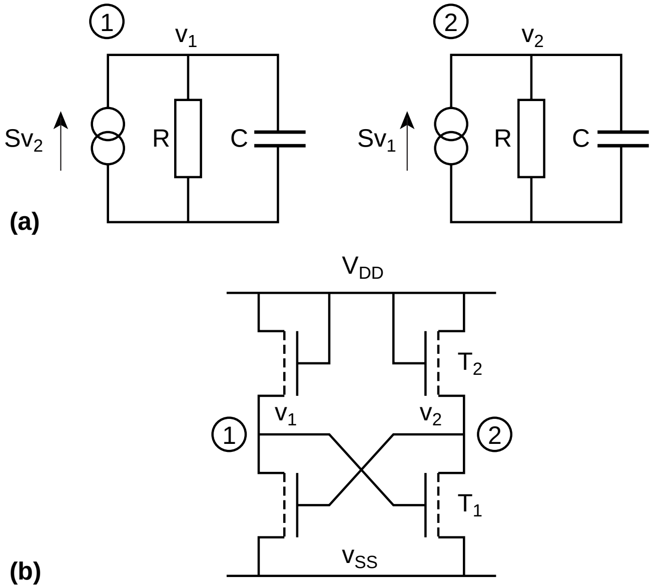


Fig. 9 (a) Small-signal model for flipflops made in MOS technology
(b) A flipflop made in MOS technology

The following considerations apply to a flipflop with enhancement transistors only. Transistor T_2 in Fig. 9b is in saturation, so (from simple MOS formula)

$$R = \frac{\partial V_{gs_2}}{\partial I_{ds_2}} = \frac{L_2}{\mu \cdot W_2 \cdot C_{\mathrm{ox}} \cdot \left(V_{gs_2} - V_{T_2}\right)} \tag{15}$$

where W_2 and L_2 represent the channel width and channel length, respectively, μ the mobility of the carriers in the channel, C_{ox} the oxide capacitance per unit area, V_{gs_2} the gate-source voltage, and V_{T_2} the threshold voltage of transistor T_2. In the meta-stable region ($v_1 \approx v_2$) transistor T_1 is also in saturation

$$S = \frac{W_1}{L_1} \cdot \mu \cdot C_{\mathrm{ox}} \cdot \left(V_{gs_1} - V_{T_1}\right). \tag{16}$$

The nodal capacitance C is equal to the sum of the gate capacitance C_g of the driver transistor T_1 and the junction capacitance to the substrate which is assumed to be equal to $\alpha \cdot C_g$, so

$$C = (1 + \alpha) \cdot C_{\mathrm{ox}} \cdot W_1 \cdot L_1. \tag{17}$$

In the meta-stable region we obtain

$$V_{gs_1} + V_{gs_2} = V_{DD} \tag{18}$$

and

$$\frac{V_{gs_2} - V_{T_2}}{V_{gs_1} - V_{T_1}} = \sqrt{\frac{W_1 / L_1}{W_2 / L_2}} = \sqrt{B}. \tag{19}$$

Assume, as a first approximation, that $V_{T_1} = V_{T_2}$, then (14)–(19) lead to

$$\frac{A-1}{\tau} = \frac{\mu \cdot (V_{DD} - 2V_T)}{(1+\alpha) \cdot L_1^2} \cdot f(B), \tag{20}$$

where

$$f(B) = \frac{1}{\sqrt{B}} \cdot \frac{\sqrt{B}-1}{\sqrt{B}+1}.$$

The factor $(A – 1)/\tau$ has a maximum, for $B \approx 5.8$, equal to

$$\left.\frac{A-1}{\tau}\right|_{\max} = 0.17 \cdot \frac{\mu}{(1+\alpha) \cdot L_1^2} \cdot (V_{DD} - 2V_T). \tag{21}$$

From a circuit simulation the duration of an MSS versus *B* is obtained for two flipflops made in different MOS processes. The MSS duration is defined as the time needed for the nodal small-signal voltage of the flipflop to increase from 1 μV to the arbitrary value of 4 percent of the supply voltage.

The results are given in Fig. 10. The MSS duration is minimal when *f*(*B*) is maximal. In Fig. 10, however, these extremes do not coincide. This is a consequence of the fact that in the derivation of (20) the simple MOS formula was used and the body effect neglected, or because of other model inaccuracies. We can conclude from Fig. 10 that designing flipflops with values of *B* greater than about 8 does not result in a better performance.

With an optimal design (which means $B \approx 8$) and an optimal layout (α is as small as possible), the maximum of the factor $(A – 1)/\tau$ is process-determined (see Fig. 10 and (21)).

Comparison of curves 1 and 2 in Fig. 10 shows an improvement in the factor $(A – 1)/\tau$ of about a factor of 2 in favour of the 2μm/5V process. It should be noted, that comparison of the circuit design in two processes, is only intended to show the process dependency. In a real design in a 2μm–5V nMOS process, however, it is obvious to use depletion loads!

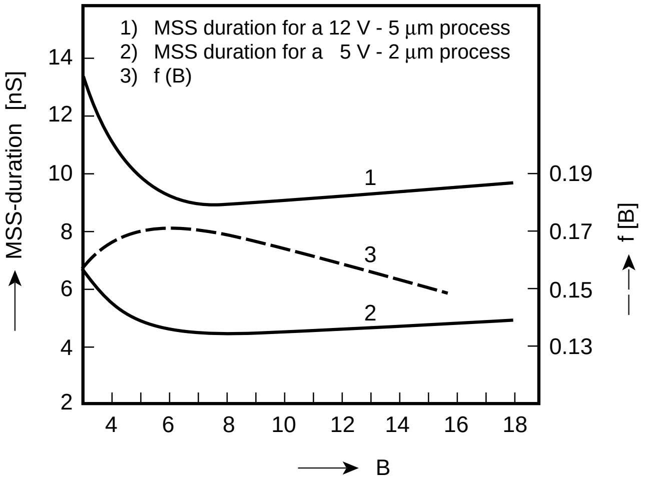


Fig. 10 MSS duration of two flipflops made in different MOS processes versus the ratio (*B*) between the aspect ratio (*W*/*L*) of the driver and the load transistor

5 Application of the Theory and Reliability of the Synchroniser

Proceeding from the above considerations, a synchroniser has been designed for four-phase MOS logic (Fig. 11). A bootstrapping technique has been applied in the flipflop to speed up the output push-pull stages.

We can now estimate the number of system failures due to the occurrence of a MSS lasting longer than a certain allowed time. The following typical values are assumed (see also Fig. 12):

data input rise/fall time	t_r = 10ns;
meta-stable region	Δv (to be determined later);
clock frequency	f_1 = 4MHz;
average data-in frequency	f_2 = 400kHz;
data-in amplitude	V = 10V;
allowed MSS duration	40ns (depends on f_1).

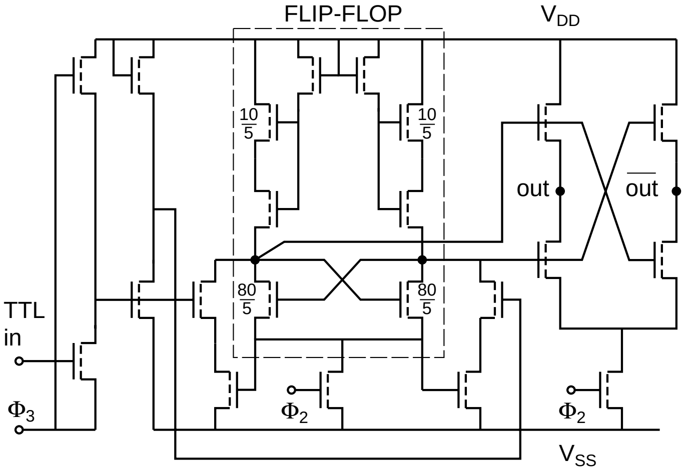


Fig. 11 The designed synchroniser for four-phase logic

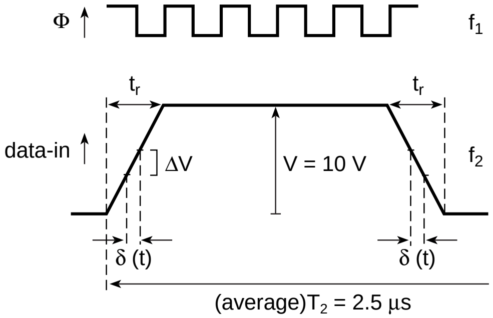


Fig. 12 Signal configuration for a typical application

As the sample values during the edges of data-in are uniformly distributed, the probability P that the falling edge of the sample clock will appear within a small-time interval $\delta(t)$ over one (average) period T_2 of data-in is

$$P = \frac{2 \cdot \delta(t)}{T_2} \cdot \frac{f_1}{f_2} = 2 \cdot \frac{\Delta v}{V} \cdot \frac{t_r}{T_2} \cdot \frac{f_1}{f_2} = 2 \cdot \frac{\Delta v}{V} \cdot t_r \cdot f_1. \qquad (22)$$

With n seconds a year (n = 31.5×10^6), the average number N of occurrences of such MSSs during one year is given by

$$N = P \cdot f_2 \cdot n = 2 \cdot \frac{\Delta v}{V} \cdot t_r \cdot f_1 \cdot f_2 \cdot n. \qquad (23)$$

The determination of the meta-stable region Δv is as follows: suppose the allowed duration of an MSS is 40ns, after which the information on the flipflop nodes should be valid. From a circuit simulation of the flipflop, the duration t_c of an MSS as a function of the sampled voltage v_0 (at $t = 0$) (see Fig. 13b) is obtained. Duration t_c is defined here to be the time at which the tangent to the voltage curve, in its bending point, crosses the time axis (see Fig. 13a). The probability of occurrence of an MSS, whose duration is longer than 40ns is, according to an extrapolation in Fig. 13b, equal to the probability of sampling a voltage $v_0 < 6 \times 10^{-15}$V, so $\Delta v = 6 \times 10^{-15}$V. Sample voltages within this region Δv cause MSSs that last longer than the allowed duration of 40ns. So, with the above assumptions, MSSs of a duration longer than 40ns occur at a rate of 6×10^{-4} times/year.

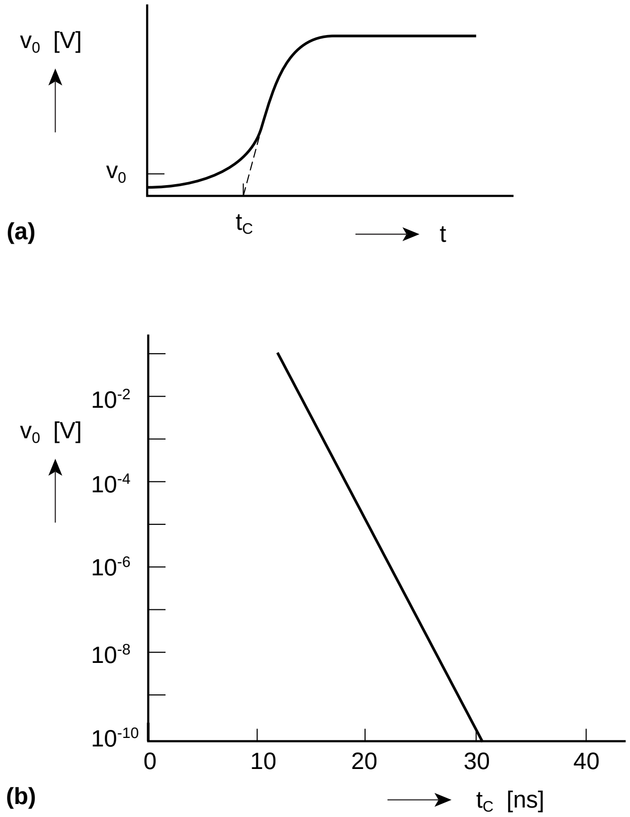


Fig. 13 (a) Definition of the MSS duration t_c
(b) Duration t_c of a MSS versus the sample voltage v_0

For applications in this clock-frequency range (several megahertz) this failure rate is quite acceptable. Formula (23) shows that the failure rate is directly proportional to the clock frequency and the frequency of the asynchronous inputs. However, if the clock frequency f_1 increases, the allowed duration of an MSS decreases proportional to f_1^{-1}. This leads to an exponential increase of the failure rate. So, if the clock frequency and/or the number of asynchronous inputs increase, an optimal synchroniser design and layout will no longer yield a reasonable failure rate. Formula (21) shows that the limits of improvement of the failure rate, by optimisation of the design and layout of the synchroniser, are process-determined.

6 Conclusions

Although each individual MSS is affected by circuit noise, the average number of MSSs during a constant time period is independent of this noise.

A flipflop can be modelled by two stages, each consisting of an inverter with amplification $-A$ followed by a first order RC-filter with $\tau = RC$. This model gives a good description of flipflop behaviour when it is operating in the meta-stable region. For the uniform distribution of sample values during the input signal edges, the probability of occurrence of an MSS whose duration t is longer than a time t_n is given by

$$P(t > t_n) = \exp\left(-\frac{A-1}{\tau} \cdot t_n\right)$$

To decrease this probability, the factor $(A - 1)/\tau$ has to be maximised. For synchronisers made in MOS technology this maximum is obtained by optimisation of design and layout. If enhancement loads are used, this optimum is obtained when the ratio between the W/L of the driver transistors and the W/L of the load transistors of the flipflop is equal to about 8, where W and L represent the channel width and channel length, respectively. At the same time, in the layout, the flipflop nodes should be as small as possible. Based on these considerations, a synchroniser has been designed and for a typical application its expected failure rate is 6.10^{-4} times/year. The limits of improvement of the failure rate, by optimisation of the design and layout, are process-determined.

With the above considerations, the failure rate can be reduced, but not to 0.

To solve the problems associated with asynchronous communication fundamentally, a circuit is needed that transduces a continuous signal into a discontinuous one, hence, this transducer must have an infinite factor $(A - 1)/\tau$. In practice, this means that such transducers cannot be made.

7 References

[1] T.J. Chaney and C.E. Molnar, ‘Anomalous behavior of synchronizer and arbiter circuits’, *IEEE Trans. Comput.*, vol. C-22, pp. 421-422, Apr. 1973.

[2] G.R. Couranz and D.F. Wann, ‘Theoretical and experimental behavior of synchronizers operating in the meta-stable region’, *IEEE Trans. Comput.*, vol. C-24, pp. 604-616, June 1975.

[3] S.H. Unger, ‘Asynchronous sequential switching circuits with unrestricted input changes’, *IEEE Trans. Comput.*, vol. C-20, pp. 1437-1444, Dec. 1971.

[4] M. Pechoucek, Anomalous response times of input synchronizers, *IEEE Trans. Comput.*, vol. C-25, pp. 133-139, Feb. 1976.

[5] L.R. Marino, ‘The effect of asynchronous inputs on sequential network reliability, *IEEE Trans. Comput.*, vol. C-26, pp. 1082-1090, Nov. 1977.

[6] T.J. Chaney, S.M. Ornstein, and W.M. Littlefield, ‘Beware the synchronizer’, in *COMPCON-72 IEEE Comput. Soc. Conf.*, San Francisco, CA, Sept. 12-14, 1972, pp. 317-319.

[7] J. Millmann and H. Taub, *Pulse and Digital Circuits*. New York: McGraw-Hill, 1965.

Appendix 1

Essentially, the nodal sample voltages v_{01} and v_{02} are derived from the same input data: if node ① is sampled with the input signal, then node ② must be sampled with the inverted input signal.
Assuming that the required inverter operates in the linear region with an amplification equal to –*D*, the following small signal relation is obtained:

$$v_{02} = -D \cdot v_{01}. \tag{A1}$$

Formula (5) then becomes

$$v_1 = \frac{v_{01}}{2}\left\{(1+D)\cdot\exp\left(\frac{A-1}{\tau}\cdot t\right) + (1-D)\cdot\exp\left(\frac{-A-1}{\tau}\cdot t\right)\right\}. \tag{A2}$$

At a time $t = t_c$ after the sample moment ($t = 0$) each sample value $v = v_{01}$ is multiplied by the same factor

$$\left\{(1+D)\cdot\exp\left(\frac{A-1}{\tau}\cdot t_c\right) + (1-D)\cdot\exp\left(\frac{-A-1}{\tau}\cdot t_c\right)\right\}.$$

The values of v_{01} are uniformly distributed, so the values

$$v_1 = \frac{v_{01}}{2}\left\{(1+D)\cdot\exp\left(\frac{A-1}{\tau}\cdot t_c\right) + (1-D)\cdot\exp\left(\frac{-A-1}{\tau}\cdot t_c\right)\right\}$$

remain uniformly distributed too, at any given time. For simplicity we may assume $D = 1$. Formula (A2) thus becomes

$$v_1 = v_{01}\cdot\exp\left(\frac{A-1}{\tau}\cdot t\right).$$

Appendix 2

In Appendix 1 it is derived that

$$v_1 = \frac{v_{01}}{2}\left\{(1+D)\cdot\exp\left(\frac{A-1}{\tau}\cdot t\right)+(1-D)\cdot\exp\left(\frac{-A-1}{\tau}\cdot t\right)\right\}. \tag{A2}$$

After a short time $t > (A + 1)/\tau$ the second term in (A2) can be neglected; thus (A2) then reduces to

$$v_1 = \frac{v_{01}}{2}\cdot(1+D)\cdot\exp\left(\frac{A-1}{\tau}\cdot t\right). \tag{A3}$$

If t_n is the time needed for the nodal small-signal voltage of the flipflop to increase from v_{01} to v_n and t_d is the time needed for the nodal small-signal voltage of the flipflop to increase from v_{01} to v_d, we get

$$v_n = \frac{v_{01}}{2}\cdot(1+D)\cdot\exp\left(\frac{A-1}{\tau}\cdot t_n\right), \tag{A4}$$

respectively,

$$v_d = \frac{v_{01}}{2}\cdot(1+D)\cdot\exp\left(\frac{A-1}{\tau}\cdot t_d\right). \tag{A5}$$

From (A4) and (A5) we obtain

$$v_n = v_d \cdot\exp\left(-\frac{A-1}{\tau}\cdot(t_n - t_d)\right).$$

Conclusions

There are two important factors that drive the process of continuous scaling of the minimum feature sizes on a chip: reduction of the costs per chip and improvement of its performance. These two factors have also been the drive behind the research *to get the most out of the MOST*.

Cost reduction is probably the most important drive for scaling. As a result of scaling, more dies (ICs) fit on the same wafer. Moreover, since the yield (which is the percentage of dies on a wafer that are correctly operating) is exponentially proportional to the die area, area reduction allows a major decline in the costs of silicon. Another way to improve the yield and thus to reduce the costs is to improve the density of a design. In this respect, CCD memories, which were developed for video applications, showed improvements in density, compared to existing video DRAMs, close to a factor of two. About the same density improvement was achieved in gate arrays, where a new basic cell concept enabled a flexible and dense realisation of both logic and memory functions.

The second important drive for scaling is the improvement in performance (speed and power) that results from it. Particularly in full-enhancement nMOS logic, the threshold voltage loss at the logic gates' outputs prevented the use of the full performance potentials of these technologies. A bootstrap mechanism, which boosts these output nodes to even above the supply voltage, compensated potential races in the logic trees of a logic gate and more than doubled its speed.

Although the speed of the individual transistors and logic functions increase every technology, it does not lead to an overall speed improvement, if the speed of communication does not keep pace with it. A full-custom designed switch-matrix that was used in a complex video signal processor, served as a flexible and programmable high-bandwidth communication bus between a large number of different video processing elements. Although most of the design efforts were on achieving the highest speed, also the density of the switch-matrix was an important target, since it had much influence on the overall processor floorplan and chip size.

Performance is not only about speed; it is also about power consumption and, even more important, it is about power efficiency. One of the contributions to the total power consumption of a chip is the short-circuit power. During an input transient, both the pMOS and nMOS logic gate parts temporarily conduct simultaneously, thereby causing a short-circuit current to flow directly from supply to ground. This current does not contribute to any functionality or performance increase and is a pure waste of energy. The largest short-circuit power consumption will therefore occur in circuits that consist of large transistors, such as on-chip drivers and output buffers. An expression to estimate this power has been developed and a design method is presented to limit it. It turns out, that with a good design approach, this short-circuit component can be reduced to only 5% to 10% of the total power consumption.

Getting the most out of the MOST, generally may result in a reduction of both the operating margins and reliability tolerances of an IC. For this reason, robust design, with a focus on reliability and signal integrity, is of major importance. One example, in this respect, discusses reliable communication between asynchronous systems. Due to this asynchronous communication, sampling flipflops may reach a meta-stable state, when the input signal is changing at the moment that it is being clocked. Such a meta-stable state may lead to logic failures. It has been proven that random noise has no influence on the average number of meta-stable states that sampling flipflops may reach during sampling and a certain period of time after that. A synchroniser has been developed to minimise this number. With a failure rate of only $6{\cdot}10^{-4}$ times/year, the developed synchroniser guarantees complete reliable operation.
Over the last decade, IC performance continuously increased, while the supply voltage reduced. This leads to much higher current densities in the supply lines on a chip and may result in a potential reliability problem in the form of electromigration. It is shown, that the contribution of wire self-heating to the actual wire temperature, due to the power dissipated in these wires, is negligible in today’s deep-submicron designs. As a result, and in contrast with existing publications, the wire self-heating hardly contributes to an increased manifestation of electromigration.

Scaling has important consequences for the operation of integrated circuits. Since the second half of the 90-ties the supply voltage is scaled proportionally to transistor channel length. This so-called constant-field scaling, particularly today, has major negative effects on the performance, density and robustness of deep-submicron ICs. An understanding of these effects is essential for the efficient exploitation of the full potential of modern deep-submicron IC manufacture processes. It is shown that a lot of additional measures in the design are needed to maintain its robustness at a sufficiently high level. These measures require additional chip area and hamper an efficient exploitation of future process generations.
Before the year 2010, we will face the fact that a move to the next process generation will no longer be commercially attractive for a lot of products. For cheap high-volume consumer products, however, this point of time will already be reached within only a few process generations.

Summary

Over the years CMOS designs required different design approaches and design priorities, depending on the application area. Particularly the drive for high performance and high density requires optimisation of the elements from which a VLSI chip is built: the basic electrical circuits and their interconnections. Although the selected subjects of research are very diverse, they share the same motivation: *getting the most out of the MOS transistor (MOST)*. This thesis describes the various subjects, which are categorised into four main topics: performance improvement, density improvement, robustness improvement and, finally, the effects of scaling.

1 Performance improvement

Much of the work has been done to support the integration of high-speed video signal processing. In this respect, one topic describes a method to improve the speed of the logic circuits, while another one presents the implementation of a very flexible high-bandwidth switch matrix, serving as a communication interface between different processing elements on a video signal processor.

Performance improvement not only means speed improvement, but it may also mean power reduction. This will probably be the most important topic of this decade. One of the components contributing to the total power consumption in CMOS ICs is the short-circuit power consumption, which occurs during input transients on logic gates during which the pMOS and nMOS transistors are temporarily conducting simultaneously. An expression is derived for this short-circuit power component, together with a method how to limit it.

2 Density improvement

Next to the fact that video signal processing requires high-speed circuits and high-bandwidth communication, it also requires the storage of huge amounts of data. Because of their serial character, charge-coupled devices (CCDs) are extremely suited to implement video memories. With only minor additions to a baseline CMOS process (only two more masks), high-density data storage (CCD) could be combined with video signal processing (CMOS) on one single chip, resulting in a 40% reduction of the memory size compared to a DRAM implementation.

In the past, density did not get much attention in the area of fast-prototyping ASICs, such as gate arrays. Research in this respect has led to the development of a flexible architecture for high-density gate arrays. This architecture

includes narrow and wide transistors in the basic cell, thereby supporting an efficient implementation of both memory and logic circuits. The realised logic densities are comparable with those of standard cell designs, meaning an improvement of almost a factor of two compared to existing gate-arrays.

3 Robustness improvement

In electronic circuits, robustness may deal on the one hand with the reliability of circuit operation, while on the other hand it also includes the integrity of the signals that propagate through them. In the design of a circuit that needs to synchronise asynchronous incoming signals, reliable operation and a low number of failures determine the design requirements of a synchroniser. Next to the fact that the presented synchroniser shows extremely reliable operation, it also has been proven, in contrast with previous publications, that noise can be neglected in the determination of the average number of these failures.

In today’s integrated circuits the continuous increase in current density really has become one of the most important design challenges of the 21-st century. Large current densities through the interconnect lead to large voltage drops in the supply lines and to an increased probability of occurrence of electromigration, due to the expected self-heating of these supply lines. It is proven that in the design of CMOS ICs, this wire self-heating may be neglected when using common design practices.

Finally, the continuous reduction of the minimum feature sizes had and still has a huge impact on the operation of the basic transistors, as well as on the circuits that are built from them. Particularly the smaller sizes, combined with an ever increased speed, has led to a dramatic increase of the noise, while at the same time the noise margins reduced due to the scaling of the applied voltages. A discussion is therefore presented, which is particularly focussed on noise and signal integrity. Also alternatives are included to keep signal integrity at a sufficiently high level.

4 The effects of scaling and future design challenges

During the four decades of its existence, the semiconductor industry has managed to reduce the transistor feature sizes by a factor of 0.7 about every 18 to 24 months. This resulted, everytime, in a doubling of the number of components per unit area. This scaling has major effects on the performance and robustness of integrated circuits (ICs). Particularly today, when we also scale the voltage at the same pace with the feature sizes, the robustness of operation is dramatically hampered. This section, therefore, discusses the effects and challenges of scaling with respect to the three previous subjects: performance, density and robustness of future ICs.

Samenvatting

Het ontwerpen van CMOS-schakelingen vereist een verschillende aanpak en verschillende prioriteitsstellingen, afhankelijk van hun applicatiegebied. In het bijzonder vereist het streven naar hogere prestaties en grotere dichtheden een optimalisatie van de elementen waaruit een chip is opgebouwd: de basisschakelingen en hun onderlinge verbindingen. Hoewel de gekozen onderwerpen nogal uiteenlopend zijn, delen ze toch een gezamenlijke motivatie: hoe kan ik zoveel mogelijk uit een MOS transistor 'persen'. De verschillende onderwerpen in dit proefschrift zijn ingedeeld in vier hoofdcategorieën: prestatieverbetering, dichtheidvergroting, verhoging van de robuustheid en, ten slotte, de effecten van schaling.

1 Prestatieverbetering

Een groot deel van het werk is besteed aan integratie ten behoeve van snelle video signaalbewerking. Terwijl één onderwerp een methode ter verhoging van de schakelsnelheid van logische poorten beschrijft, behandelt een volgend onderwerp de realisatie van een flexibele schakelmatrix. Deze dient als communicatie-interface tussen verschillende processingelementen op een video signaalprocessor en zorgt daarbij voor een zeer hoge bandbreedte.

Het verbeteren van de prestaties van een chip betekent niet alleen het proberen te vergroten van de snelheid, maar het betekent ook het verminderen van het vermogensverbruik. Waarschijnlijk zal dit laatste een van de belangrijkste eisen zijn van de komende 10 jaar. Een van de bijdragen aan het vermogens verbruik van een CMOS-chip is de zogenaamde kortsluitdissipatie. Deze treedt op wanneer, gedurende de flanken van de ingangssignalen, een of meerdere pMOST-en en een of meerdere nMOST-en tijdelijk tegelijkertijd geleiden. Er is een uitdrukking voor deze kortsluitdissipatie afgeleid, samen met een methode om deze dissipatie ook te beperken.

2 Dichtheidsverbetering

Naast het feit dat video signaalbewerking snelle schakelingen en een grote communicatiebandbreedte vereist, zijn er altijd ook grote hoeveelheden data bij betrokken. Charge Coupled Devices (CCDs) zijn, vanwege hun seriële karakter, bijzonder geschikt om videodata in op te slaan. Met slechts een paar toevoegingen aan een standaard CMOS-technologie, kon dataopslag met een hoge dichtheid gecombineerd worden met video signaalbewerking op één chip. Dit resulteerde in een 40% kleiner geheugen in vergelijking met een DRAM-implementatie.

Voor prototype-ASICs, zoals gate arrays, was een grote dichtheid van de transistoren gedurende lange tijd geen belangrijke eis. Echter, door de toenemende aantallen per product, komt hier ook verandering in. Onderzoek aan dit type ASICs (sea-of-gates) heeft geleid tot de ontwikkeling van een flexibele architectuur voor gate arrays met hoge dichtheden. Deze architectuur bevat zowel smalle als brede nMOS- en pMOS-transistoren in de basiscel. Hierdoor zijn ze bijzonder geschikt voor het efficiënt realiseren van zowel logica als geheugen. De behaalde dichtheden in de logica zijn vergelijkbaar met die van standaardcelontwerpen. Dit betekende een verbetering van een factor twee ten opzichte van bestaande gate arrays.

3 Robuustheidverbetering

De robuustheid van elektronische schakelingen heeft te maken met zowel de betrouwbare werking van deze schakelingen, als ook met de integriteit van de signalen die er door heen propageren.

Een schakeling die binnenkomende asynchrone signalen moet synchroniseren met een kloksignaal, vereist bijvoorbeeld een betrouwbare werking en een lage kans op foutieve beslissingen. De hier gepresenteerde synchroniser vertoont een zeer betrouwbare werking. Daarnaast blijkt, in tegenstelling tot andere publicaties, dat storing (noise) geen invloed heeft op het gemiddelde aantal foutieve beslissingen van synchronisers in het algemeen.

Door de steeds maar groter wordende stroomdichtheid in geïntegreerde schakelingen, behoort het tot de grootste uitdagingen van de huidig IC-ontwerp. Een grote stroomdichtheid door metaallijnen kan niet alleen leiden tot een grote spanningsval in de toevoerdraden op een chip, maar ook tot een verhoogde kans op electromigratie ten gevolge van de verwachte zelfverwarming (wire self-heating) van deze toevoerdraden. Er is afgeleid, dat, wanneer normale ontwerpmethoden worden gebruikt, deze wire self-heating kan worden verwaarloosd.

Ten slotte, het steeds kleiner worden van de minimale afmetingen op een chip heeft enorme gevolgen voor zowel werking van de basistransistoren als ook voor de gehele systemen die er op een chip mee worden gebouwd. In het bijzonder leiden het toegenomen aantal componenten en hun toegenomen schakelsnelheid tot een enorme toename van de storing, terwijl tegelijkertijd de spanningen en dus de storingsmarges afnemen. Daarnaast neemt de overspraak van signalen over lange draden toe. Naast een beschrijving van oorzaken van de toegenomen storingen in een digitale schakeling wordt ook ingegaan op ontwerpmethoden om de signaalintegriteit op een voldoende niveau te houden.

4 De invloed van schaling en toekomstige uitdagingen

Gedurende de vier decennia van zijn bestaan, is de halfgeleiderindustrie er in geslaagd om de minimale afbeeldingen op een IC elke 18 tot 24 maanden met een factor 0.7 te schalen. Dit betekende telkens een verdubbeling van het aantal transistoren per oppervlakte-eenheid. Daarnaast heeft deze schaling een enorme invloed op de prestaties en robuustheid van IC’s. Wanneer we nu ook nog de spanning even snel schalen als de minimale afbeelding, wat reeds gedurende een viertal CMOS-technologiegeneraties gebeurd is, vormt dat een extra bedreiging voor de robuustheid. Dit hoofdstuk beschrijft daarom de effecten en uitdagingen van schaling met betrekking tot de drie voornoemde onderwerpen: de prestaties, de dichtheid en de robuustheid van toekomstige IC’s.

Publications of the author

Conference contributions and other publications from 2000 onwards

Since much of the published work is included in this thesis, only the publications since 2000 are listed here.

12-06-2000 *100nm CMOS Technology; a Design Perspective,* short course at the VLSI Technology Symposium, Honolulu

01-12-2000 ‘*Deep-Submicron CMOS ICs; from Basics to ASICs*’, ISBN 9044001116, Kluwer Academic Publishers, Dordrecht, London, Boston

06-02-2001 ‘*100 Cube; Science or Fiction?*’, panel presentation at the International Solid-State Circuits Conference (ISSCC), San Francisco

25-04-2001 ‘*Future System-on-Chip Performance: Killed by the Backend?*’, Invited Paper SEMICON conference, Munich

06-05-2001 ‘*The Future of Semiconductors, Moore or Less?*’, International Symposium on Circuits and Systems (ISCAS), Sydney

03-02-2002 ‘*Noise in Digital Circuits @ Low Voltage*’, Invited paper in special low-voltage session at the International Solid-State Circuits Conference, San Francisco

Sept. 2002 ‘*Wire Self-Heating in supply lines on bulk-CMOS ICs*’, submitted to the European Solid-State Circuit conference, Florence (I)

Courses and published books

In parallel to the research work, many (C)MOS courses have been given at most of the Philips Research and Semiconductor sites, today accounting for a total of about 3000 participants. The continuous development and adaptation of the course material to a state-of-the-art level in all disciplines of CMOS manufacturing and design has resulted in the development of many course notes and finally in the publication of three books:

1 *Geïntegreerde MOS schakelingen*, Delta Press, Amerongen, Netherlands, 1990

2 *MOS ICs, from Basics to ASICs*, VCH, Weinheim, Germany, 1992

3 *Deep-Submicron CMOS ICs, from Basics to ASICs*, Kluwer, Deventer, 1998 with a second edition in 2000.

Granted patents and patent applications

Many of the ideas that form the basics behind the design solutions or implementations discussed in this thesis, have resulted in fourteen granted US patents with another six pending.

US06081149	2000	‘Electronic circuit with a clock switch’ H.J.M.Veendrick
US05264738	1993	‘Flipflop circuit having transfer-gate delay’ H.J.M.Veendrick, A.A.J.M. van den Elshout, C.M.Huizer
US05250823	1993	‘Integrated CMOS gate-array circuit’ H.J.M.Veendrick, A.A.J.M. van den Elshout, D.W. Harberts
US05053648	1991	‘Master slice CMOS array having complementary columns’ A.A.J.M. van den Elshout, H.J.M.Veendrick, D.W.Harberts
US04947380	1990	‘Multi-mode memory device’ A.T.van Zanten, H.J.M.Veendrick, F.A.Steenhof, P.H.Frencken, A.H.H.J.Nillesen, C.G.L.M.van der Sanden
US04918331	1990	‘Logic circuits with data resynchronization’ A.T.van Zanten, H.J.M.Veendrick, L.C.M.G.Pfennings
US04820936	1989	‘Integrated CMOS circuit comprising a substrate bias voltage generator’ H.J.M.Veendrick, C.G.L.M.van der Sanden, A.Slob
US04817090	1989	‘Integrated electronic multiplex circuit’ L.C.M.G.Pfennings, H.J.M.Veendrick, A.T.van Zanten
US04775806	1988	‘Integrated circuit having capacitive process-scatter compensation’ L.C.M.G.Pfennings, H.J.M.Veendrick, A.T.van Zanten
US04730266	1988	‘Logic full adder circuit’ J.L.van Meerbergen, H.J.M.Veendrick, F.P.J.M.Welten, F.van Wijk
US04727560	1988	‘Charge-coupled device with reduced signal distortion’ A.T.van Zanten, H.J.M.Veendrick, L.C.M.G.Pfennings
US04707844	1987	‘Integrated circuit having reduced cross-talk’ H.J.M.Veendrick, A.T.van Zanten, L.C.M.G.Pfennings
US04697111	1987	‘Logic bootstrapping circuit having a feedforward kicker circuit’ A.T.van Zanten, H.J.M.Veendrick, L.C.M.Pfennings, W.C.H.Gubbels
US04513388	1985	‘Electronic device for the execution of a mathematical operation on sets of three digital variables’ H.J.M.Veendrick, L.C.M.G.Pfennings, J.G.Raven, A.H.H.J.Nillesen

About the author

Harry Veendrick graduated from the Technical University Eindhoven, the Netherlands, in 1977. In the same year he joined Philips Research Laboratories, also in Eindhoven, where he has been involved in the design of memories, gate arrays and complex video signal processors. He holds 14 patents in the area of CMOS circuit design with another six patents pending and is the (co-)author of several publications on robust, high-performance, high-density and low-power CMOS IC design. In this respect, he has contributed to various conferences and conference workshops, as speaker, invited speaker, panellist, organiser and Program Committee member. In addition, he is the author of a book on ‘Deep-Submicron CMOS ICs’ and has been actively involved in the training of about 3000 design, test and product engineers.

His principal research interests include the design of low-power and high-speed complex digital CMOS circuits, with an emphasis on deep-submicron physical effects and scaling aspects. Currently he leads the Deep-Submicron Electrical Design Cluster within the research group Digital Design and Test at Philips Research Labs. Complementary to his experience in IC design is his interest in IC technology, which allows him to act as an interface between digital IC design and IC process technology.

He is a Research Fellow at Philips Research Laboratories and a Visiting Professor to the Department of Electronic and Electrical Engineering of the University of Strathclyde, Glasgow, Scotland, UK.

Stellingen

behorende bij het proefschrift

Semiconductor-Technology Exploration
getting the most out of the MOST

van

Harry Veendrick

Eindhoven, Juni 2002

0

Het gering aantal sociaal-maatschappelijke stellingen in recente proefschriften van bètawetenschappers duidt op een te hoge werkdruk voor deze promovendi.
[*S.W.S. Gussekloo*]

I

Het opvoeden van kinderen lijkt op het programmeren van computers. Bij hetzelfde programma zou je dezelfde respons verwachten, behalve wanneer ze, meestal in de puberteit, met een virus besmet raken.

II

Sommige mensen vinden het belangrijker om te tonen wat ze hebben, dan wie ze zijn.

III

Het leven is als een film. Een reünie is het proberen deze terug te spoelen naar een fragment.

IV

Er is bijna geen beroep te vinden, waarin zoveel fouten gemaakt en getolereerd worden als die van weerman in Nederland.

V

Het verkeerd aansluiten van slechts één transistor in een digitale schakeling leidt meestal tot foutief gedrag. In een analoge schakeling kan hetzelfde leiden tot een patent.

VI

Elk IC-ontwerp is correct totdat het tegendeel bewezen is.

VII

Niets is zo voorspelbaar als de progressie in de halfgeleidertechnologie.
[*ITRS-roadmap; dit proefschrift*]

VIII

Omdat in elke nieuwe halfgeleidertechnologie-generatie de circuitstoring toeneemt, terwijl de marges juist afnemen, zal er alleen al om deze reden een eind komen aan de wet van Moore.
[*Dit proefschrift*]

IX

Een correct ontwerp van een synchronisatieschakeling maakt zijn foutkans bijzonder klein, doch nooit gelijk aan nul.
[*Dit proefschrift*]

X

Om de toekomstige complexiteit van het IC-ontwerp en productieproces beheersbaar te maken, moet op alle ontwerpniveaus de regelmaat vergroot worden.
[*Dit proefschrift*]

XI

Om dezelfde reden dat er een *Known Good Die*-parameter behoort bij elke chip in een MCM, moet er een *Known Good Core*-parameter bestaan bij elk re-usable IP-blok in een SoC.
[*Veendrick, 2000*]