Predicting Power Scalability in a Reconfigurable Platform

A thesis submitted in fulfillment
of the requirements of the
degree of Doctor of Philosophy

August 2007

Paul Beckett

B. Eng (Comm.), M. Eng.

School of Electrical & Computer Engineering

Science, Engineering and Technology Portfolio

RMIT University

Contents

List of Figures

List of Tables

Declaration

This is to certify that:

1. This dissertation comprises only my original work towards the PhD degree carried out since the official commencement date of the research program;
2. Due acknowledgement has been made in the text to all other material used;
3. No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other University or Institute of learning.
4. No specific editorial assistance, either paid or unpaid, has been received during the preparation of this manuscript.
5. Ethics procedures and guidelines have been followed.

Paul Beckett

30 August 2007

Copyright

1. The Author asserts copyright over the text of this dissertation. Copies by any process either in full, or of extracts may be made only in accordance with instructions given by the Author.

2. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted provided that copies are not made or distributed for profit or commercial advantage and that the author's copyright is shown on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission.

3. A non-exclusive license is hereby granted to RMIT University or its agents to:

    a. archive and to reproduce this thesis in digital form;

    b. Communicate it to the public by making it available online through the Australian Digital Thesis Program. The author warrants that the thesis does not infringe the intellectual property rights of any person, and indemnifies RMIT University against any loss or liability it may incur in respect of a breach of this warranty.

4. The ownership of and rights to any intellectual property that may be described in this dissertation is vested in the RMIT University, subject to any prior agreement to the contrary, and may not be made available for use by third parties without permission of the University, which will prescribe the terms and conditions of any such agreement. Further information on the conditions under which disclosures and exploitation may take place is available from the Head of School, Electrical and Computer Engineering, RMIT University.

Acknowledgements

"Sometimes a scream is better than a thesis."

Ralph Waldo Emerson (1803-1882)

This work has relied heavily on simulation facilities provided by the Network for Computational Nanotechnology (NCN) at http://nanohub.org. It came as something of a surprise to realize that I had made it into the top 50 users launching jobs on the Nanohub, but I'm sure I needed every one of those 2900 runs to make sense of it all.

I have also used some of the Berkeley Predictive Technology Model (BPTM) work, now supported as an online simulation site by the Nanoscale Integration and Modeling (NIMO) Group at the Arizona State University, http://www.eas.asu.edu/~ptm/.

All of the curve-fits in Chapters 3–5 were performed at http://zunzun.com, a brilliant online curve fitting and statistics site created and maintained by James R. Phillips.

The ITRS predictive tool MASTAR (version 4.1.0.5, 2005 as well as the earlier version 2.0.7, 2003) was written by the Advanced Devices Research Team at STMicroelectronics. The tool was downloaded from the ITRS site at http://www.itrs.net/models.html.

I would like to thank Christophe Lallement, Fabien Prégaldiny and the others in the Electronics Group at EPFL for providing two of their advanced, "work-in-progress" versions of the EKV model written in VHDL-AMS. The symmetric double-gate model was particularly important to the final stage of this work.

I would particularly like to thank Dr. Seth Copen Goldstein, Carnegie Mellon University, for providing the original spark for the power-area-performance model and for allowing me to work on developing it into a more complete theory.

Dr. Mark Lundstrom, Purdue University, took time out of his busy schedule for some useful early discussions.

Finally, thanks to my two supervisors, Dr. Andrew Jennings and Dr. Mike Austin for their support over the past six years.

And special and heartfelt thanks to my wife, Alexis, for ongoing support and encouragement—and for kicks at just the right moments.

Summary

This thesis focuses on the evolution of digital hardware systems. A reconfigurable platform is proposed and analysed based on thin-body, fully-depleted silicon-on-insulator Schottky-barrier transistors with metal gates and silicide source/drain (TBFDSBSOI). These offer the potential for simplified processing that will allow them to reach ultimate nanoscale gate dimensions.

Technology CAD was used to show that the threshold voltage in TBFDSBSOI devices will be controllable by gate potentials that scale down with the channel dimensions while remaining within appropriate gate reliability limits. SPICE simulations determined that the magnitude of the threshold shift predicted by TCAD software would be sufficient to control the logic configuration of a simple, regular array of these TBFDSBSOI transistors as well as to constrain its overall subthreshold power growth. Using these devices, a reconfigurable platform is proposed based on a regular 6-input, 6-output NOR LUT block in which the logic and configuration functions of the array are mapped onto separate gates of the double-gate device.

A new analytic model of the relationship between power (P), area (A) and performance (T) has been developed based on a simple VLSI complexity metric of the form AT^{σ} = constant. As σ defines the performance "return" gained as a result of an increase in area, it also represents a bound on the architectural options available in power-scalable digital systems. This analytic model was used to determine that simple computing functions mapped to the reconfigurable platform will exhibit continuous power-area-performance scaling behavior.

A number of simple arithmetic circuits were mapped to the array and their delay and subthreshold leakage analysed over a representative range of supply and threshold voltages, thus determining a worse-case range for the device/circuit-level parameters of the model. Finally, an architectural simulation was built in VHDL-AMS. The frequency scaling described by σ, combined with the device/circuit-level parameters predicts the overall power and performance scaling of parallel architectures mapped to the array.

Chapter 1. Introduction

"…transistor scaling is approaching its limit. When that limit is reached, things must change, but that does not mean that Moore's law has to end."

Mark Lundstrom in [1]

1.1 Overview

Of all human inventions, perhaps the most astonishing is the integrated circuit. The first transistor, announced to the world on June 30, 1948, was a lump of germanium crystal that took its inventors more than four years to perfect. The first commercially available planar integrated circuit (IC), shipped by Fairchild Semiconductor Corporation in March 1961, comprised one transistor, three resistors and a capacitor [2]. It was largely ignored. To really consolidate the success and scalability of the IC took the development and refinement of processes such as masked diffusion, lithography, planar technology, isolation, high-quality oxide and epitaxy [3] but since then it has been a story of smaller, faster, cheaper to a point where in 2005, world semiconductor capacity was estimated to be more than 1.5 million wafers per week [4]—well over seventy billion transistors per second—in a global market worth more than $1 trillion a year.

This extravagant abundance has driven the emergence of the modern VLSI microprocessor in which vast numbers of practically identical transistor switches are interconnected to form complex computational networks. For example, in 1999 constructing the Alpha 21264 processor took some 15 million transistors [5]. By 2001 this had grown to 130 million in the 4th-generation Alpha [6]. In 2003, Intel released the Itanium® II processor with 410 million transistors on a single 374mm^2 chip [7] and the 2005 Montecito® processor contained 1.7 billion-transistors in a multi-core architecture operating at 1.8GHz [8]. It has been predicted that by 2012 a CMOS (or more likely SiGe) chip may comprise some 10^{10} transistors operating at speeds in the order of 10–15GHz [9], although this now appears unlikely due to power density constraints.

Devices with gate lengths of less than 100nm were commercially shipped in the year 2000, signalling the end of the "Microelectronics Era" and the start of the age of "Nanoelectronics" [10]. As a result, the International Technology Roadmap for Semiconductors (the ITRS, which focuses mainly on CMOS) [11] is now predicting what appears to be the end of the development path for silicon by 2020, when effective gate lengths are likely to be less than 5nm. There is anecdotal evidence[1] to suggest that funding for silicon research is already diminishing as the hunt intensifies for the next technology that will take the integrated circuit beyond that point.

However, even to reach the end of the silicon roadmap the challenges will be formidable. Amongst a long list of technical difficulties, the ITRS identifies the following issues:

- the rapid growth in power consumption at each successive technology node;
- the need for new architectures to overcome bottlenecks at interconnects;
- escalating difficulties in both lithography and fabrication, leading to spiralling costs.
- the need for more complex structures such as SOI or dual-gate transistors to work around the limitations of short device channels;

The likely nexus between power consumption and architecture has been articulated by the 2003 ITRS as follows: *"Below 65 nm, MPU designs hit fundamental walls of performance, power consumption and heat dissipation.…Power consumption can be managed only by careful application of on-chip processing parallelism…the future goal of system-level design is to map a maximally parallel function to a maximally parallel implementation.…Methodologically, this defines a new design domain that emphasizes distributed implementation over centralized implementation;…Given such trends, standalone MPU design style will likely evolve into a sea-of-processing elements design style." [12].*

It is this link between power and parallelism, especially in the context of very fine-grained computing structures built using simplified manufacturing technologies, which has been the primary motivation for this thesis.

[1] Dr. Mark Lundstrom, Purdue University, personal communication, 2003

1.2 Motivation and Scope

This work is concerned with the evolution of digital hardware systems as devices scale towards the end of the CMOS roadmap. Although there is a truly vast literature related to the problems to be overcome in order to reach this point, a number of general observations are already possible and these have motivated this research:

- Although the continued scaling of conventional CMOS will eventually reach fundamental physical limits, forcing a move to alternative materials and structures, there is currently still scope in CMOS for improved performance at nanoscale dimensions.

- Power density, both static and dynamic, will become *the* critical issue as device numbers scale that, in itself, has the potential to prevent the deployment of architectures at nano-scale dimensions [13].

- Even taking into account the impact of low-κ interconnect dielectrics, transistor delay will continue to improve with scaling at a faster rate than wire delay. As a result, communications will increasingly replace processing performance as the limiting factor in computer architectures [14].

- The rapidly escalating costs of IC design, fabrication and test will increasingly favour simple, regular structures that support flexible hardware configuration and design reuse and that may be reprogrammed and/or reconfigured post-manufacture. This appears to be inevitable for two main reasons:

    1. *Foundry Overheads:* as technology moves past the 90nm node, the high costs of establishing and running an advanced foundry as well as increasing non-recurrent engineering costs (mainly driven by lithography) mean higher fixed overheads on each chip produced.

    2. *Device Reliability:* The manufacture of chips at nanoscale dimensions with 100% working transistors will be prohibitively expensive, if not impossible. Devices

and their interconnections will exhibit lower intrinsic reliability and increased variability. To achieve reasonable yields will require flexible architectures that can "configure around the defects" [15].

- A strong case is emerging for the integrated use of reconfiguration in future nanoscale systems (e.g. [15-18]). While fine-grained array-based reconfigurable systems, such as field-programmable gate arrays, already offer the ability to customize a device to a specific application, their limitations are well known: poor area-delay performance, high (relative) power consumption and large reconfiguration and routing overheads (often more than 10 times the area of the logic [19]) making them a poor match to dense, regular computational structures such as ALUs or memory.

Based on these observations, the research described in this dissertation addresses the following questions:

1. Does the escalating cost of design, fabrication and test in future nanoscale systems justify a re-evaluation of homogeneous reconfigurable meshes and can nanoscale electronic devices offer new opportunities for developing these into low-power, low-overhead reconfigurable systems?

2. Can a simple, homogeneous, mesh-connected array of reconfigurable components efficiently support the sort of complex heterogeneous processing organizations that characterize typical high-performance computer architectures?

3. Can the scalability of reconfigurable meshes be predicted from an architectural perspective?

4. Can reconfigurable structures of this type be made sufficiently scalable in terms of performance and power such that very high levels of integration (e.g. $>10^{11}$ devices) might be achievable?

1.3 Thesis Statement

As CMOS technology scales towards the end of the silicon roadmap, simple reconfigurable logic function arrays with predominately nearest neighbour connectivity will become feasible building blocks for scalable, low-power digital hardware.

1.4 Research Approach

The issues raised in the research questions above have been addressed at both a device and architectural level, reflecting the observation made by the 2005 ITRS that: "*[t]hese challenges demand …continued mergers between traditionally separated areas of [design technology]*" [11]. A hierarchical simulation approach has been used in this thesis work, with each simulation stage being used to validate the models for the next and, at the same time, being cross-checked against results presented in the literature, where available.

As it is difficult to anticipate the impact of future materials and device-level discoveries, current predictions for highly scaled silicon devices must be considered to be speculative. On the other hand, the 2005 ITRS places the end of the silicon roadmap at around 2019–2020 which at the time of writing is at most four scaling generations away. Although it remains to be seen which of the many competing approaches will be successfully integrated into commercial CMOS fabrication lines, it is likely that the most *plausible* technology drivers (i.e., those with the highest likelihood of contributing to so-called "end-of-roadmap" devices) have already been described in some form or other. It is therefore possible to make some realistic predictions about these ultimately scaled devices and the architectures that will be created from them.

To answer the questions outlined in Section 1.2 above, this research has proceeded in four stages, as outlined below. The approach and objectives are summarized in Figure 1.

1. ***Can simple CMOS logic arrays become feasible building blocks at nanoscale dimensions?***

   The first stage of this research involved setting up a demonstration reconfigurable platform based on a hypothetical thin-body fully-depleted silicon-on-insulator transistor with

metal-gate and silicide source/drain (TBFDSBSOI). These were chosen to represent a plausible end-of-roadmap device technology. At the time this work was undertaken, a small number of planar devices with silicide source/drain had already been reported in the literature, and a similarly small number of nanoscale double-gate silicon devices, but no examples had been published of double-gate silicided source/drain transistors. Thus, the objective here was to characterize the likely performance of this technology and to use it to develop a reconfigurable test platform. The Technology Computer Aided Design (TCAD) results, derived in this work from a commercial TCAD simulator [2], predict that the threshold voltage in TBFDSBSOI devices will be able to be controlled by gate potentials that scale down with the channel dimensions and that are within appropriate gate reliability limits.

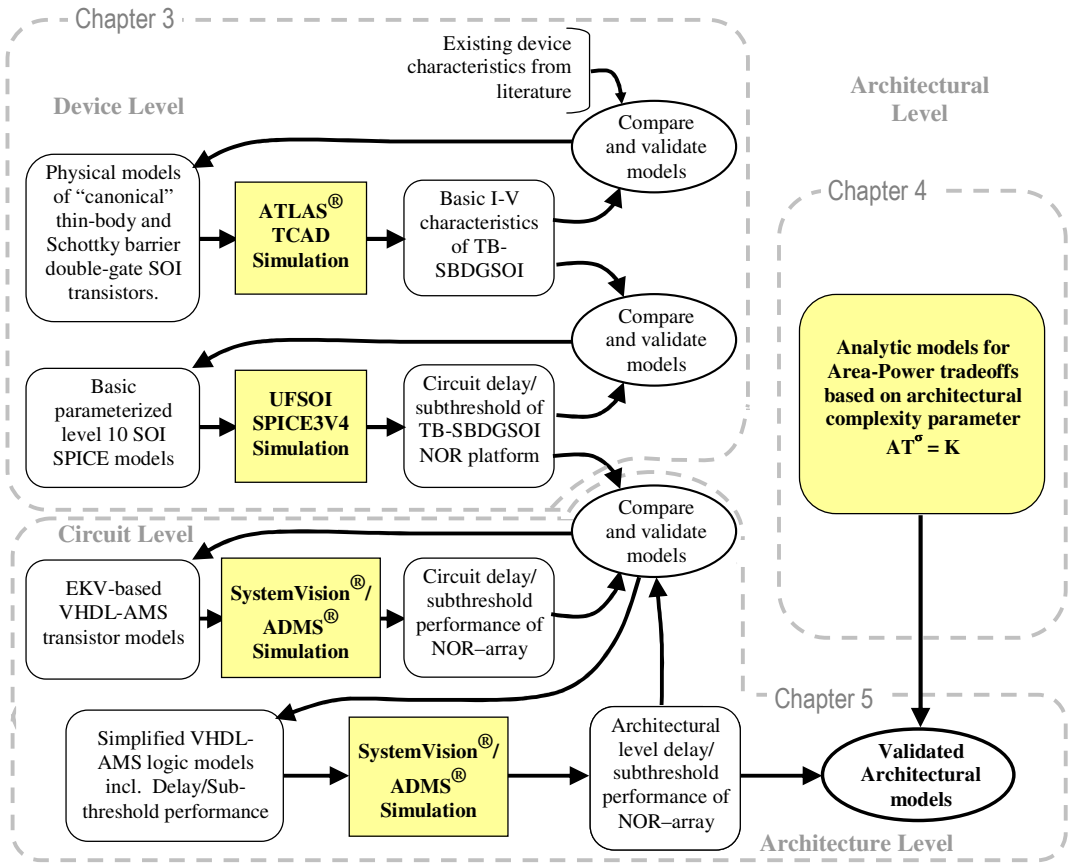


Figure 1. Research approach and objectives
encompassing Device, Circuit and Architectural-level simulation.

[2] Atlas/SPisces from Silvaco Inc., http://www.silvaco.com/products/device_simulation/atlas.html

2. ***Can heterogeneous processing organizations be set up using homogeneous meshes of reconfigurable components?***

   The approximate I-V characteristics derived from TCAD simulation were used to characterize SPICE models that, in turn, were used to show that the magnitude of the threshold shift will be sufficient to constrain the overall subthreshold power of arrays of these TBFDSBSOI transistors, as well as providing a mechanism to control the logic configuration. The result of this stage was the analysis of a highly regular 6-input, 6-output NOR LUT block in which the logic and configuration functions of the array are mapped onto separate gates of a double-gate device. In this way, the array can be configured using the threshold shifts seen at the logic gate resulting from bias changes on the configuration gate. An overall objective here was to determine how this simple array organization might support both combinational and sequential logic.

3. ***Is it possible to predict the scalability of these reconfigurable systems at an architectural level?***

   An analytic relationship between power (P), area (A) and performance (T) was developed based on a simple VLSI complexity metric of the form: $AT^{\sigma} = \text{constant}$. The complexity metric σ defines a bound on the architectural options available in power-scalable digital systems. The objective of this stage was to develop a set of metrics that could be used to evaluate the scaling performance of the reconfigurable array at an abstract level and to determine how the threshold/supply voltage relationship of future technology might impact on this behavior.

4. ***Can simple reconfigurable arrays scale with high performance and low power?***

   For this final stage, the model developed previously was used to determine under what circumstances the computing functions mapped to the reconfigurable platform would exhibit continuously scalable power-area-performance characteristics. This comprised two interrelated levels of simulation. Firstly, a device/circuit level simulation was created using simplified EKV transistor models [20] written in the VHDL-AMS mixed-signal lan-

guage (i.e., VHDL with Analog and Mixed-Signal extensions [21]). A representative arithmetic circuit was mapped to the array and its performance used to predict the technology-related parameters for the model designed in the previous stage. Finally, an architectural model was built, also in VHDL-AMS, and used to predict the power-area-performance characteristics of the reconfigurable fabric over the supply range expected for the remaining nodes of the CMOS roadmap.

1.5 Specific Outcomes and Contributions

The work reported in this dissertation has resulted in the following specific outcomes, many of which have been previously reported in the publications listed in Section 1.7:

- The demonstration, by TCAD simulation, that ultra-thin body, double-gated fully depleted SOI transistors will exhibit novel operating behavior that will, in turn, support simple reconfigurable computing meshes.
- The specification and analysis of a regular, mesh-connected array based on the TBFD-SBSOI devices, firstly by low-level TCAD and SPICE simulation and then via a register transfer level (RTL) simulation using behavioral models derived from the previous TCAD and SPICE work.
- A demonstration via high level simulation that this mesh-connected array is logically equivalent to more complex FPGA-like organizations and will support power-scalable reconfigurable systems.
- The development of a new analytic approach to power and energy vs. area based on a traditional architectural complexity metric of the form $AT^{\sigma} = K$. This defines the limits on the area-performance tradeoffs for architectures that will support massive area scaling.
- The verification by simulation that architectures mapped to the array may be described by the analytic relationship developed between area and power/energy, and that this will predict their ultimate scalability.

1.6 Dissertation Outline

This document is organized as follows:

- Chapter 2 first presents an overview of the general issues that are expected to impact on computer architecture as devices scale to the end of the silicon roadmap. This necessarily encompasses a fairly broad range of device/circuit/architecture considerations.
- Chapter 3 focuses on the ***device*** and ***circuit*** level. It describes and analyses the performance of a reconfigurable mesh based on thin-body, double gate silicide devices. This chapter includes TCAD results that characterize the basic TBDGSBSOI devices, as well as the SPICE simulations describing the circuit-level performance of combinational and sequential devices built using a simple 6-NOR building block.
- In Chapter 4, encompasses an ***architectural*** level analysis. A new theoretical framework is described that supports the evaluation of power-area-performance tradeoffs in future digital logic systems.
- Chapter 5 draws these device/circuit and architecture threads together by applying the analytic model of Chapter 4 to an evaluation of the scalability of the mesh-connected reconfigurable system proposed in Chapter 3.
- Finally, an overall summary, conclusions and outline for future work can be found in Chapter 6.

1.7 Publications

The following publications have resulted directly from the work described in this dissertation.

- P. Beckett and A. Jennings, "Towards Nanocomputer Architecture", presented at the Seventh Asia-Pacific Computer Systems Architecture Conference, ACSAC'2002, Melbourne, Australia, 2002.

- P. Beckett, "A Fine-Grained Reconfigurable Logic Array Based on Double Gate Transistors", presented at IEEE International Conference on Field-Programmable Technology, FPT2002, Hong Kong, 2002.

- P. Beckett, "A Polymorphic Hardware Platform", presented at the 10th Reconfigurable Architectures Workshop, RAW 2003, Nice, France, 2003.

- P. Beckett, "Exploiting Multiple Functionality for Nano-Scale Reconfigurable Systems", presented at the Great Lakes Symposium on VLSI, Washington, USA, 2003.

- P. Beckett, "Low-Power Circuits using Dynamic Threshold Devices", presented at the Great Lakes Symposium on VLSI, Chicago, Il, 2005.

- P. Beckett and S. C. Goldstein, "Why Area Might Reduce Power in Nanoscale CMOS", Presented at IEEE International Symposium on Circuits and Systems, ISCAS'05, Kobe, Japan, May 2005.

- P. Beckett, “A Nanowire Array for Reconfigurable Computing”, presented at TenCon 2005, Melbourne, 21- 24 November, 2005.

- P. Beckett, “A Low–Power Reconfigurable Logic Array Based on Double Gate Transistors”, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, accepted for publication February 2007.

Chapter 2. Scaling Issues for Future Computer Architecture

"It would appear that we have reached the limits of what it is possible to achieve with computer technology, although one should be careful with such statements, as they tend to sound pretty silly in 5 years."

Attributed to John von Neumann (1903-1957)
http://en.wikiquote.org/wiki/John_von_Neumann

This chapter examines a range of issues that will impact on computer architecture as it moves further into deep sub-micron and ultimately the nanoscale domain. There is a rich literature on the likely effect of technology scaling and the particular trends driving it. Broadly, the issues may be divided into fundamental, material/device and circuit/architecture considerations. The fact that many of these are interrelated and therefore cannot be considered in isolation is becoming a problem in itself as technology scales and many of the abstractions that have served the design community well in the past 20–30 years begin to break down. The following analysis is based very loosely on Meindl's hierarchy of limits [22], starting with *fundamental* (physical), *material* and *device* issues and concluding with some of the more abstract issues in nanoscale *circuits* and *architectures*.

2.1 Fundamental Limits to Device Scaling

The primary objective of any system for electronic information processing is the creation of controllable electron barriers [23]. The limits to maximum performance, density and minimum energy arise from the characteristics of these barriers and are ultimately constrained by the basic principles of thermodynamics, quantum mechanics and electromagnetics [22].

Thermodynamic Limits

Thermal noise will exist in any electronic system operating above a temperature of absolute zero. Thus, at a given temperature, T, an electron has a finite probability that it will be able to transition *over* a barrier with height E_b that is given by the classic models as:

$$\prod = \exp\left(-\frac{E_b}{k_B T}\right) \tag{2.1}$$

(k_B is the Boltzmann constant). In a binary system, it is reasonable to treat a probability of $\prod = 0.5$ as the point at which it becomes impossible to distinguish between one logic level and another (i.e. between the cases where the information-carrying electron is confined or not). This leads directly to the *Shannon–von Neumann–Landauer (SNL)* expression for the smallest energy required to process a bit at temperature T: $E_b = kT\ln(2) \approx 17\text{meV}$ at 300 K [24]. Energy transitions in CMOS right now are typically in the region of 10^7 times greater than this. Meindl and Davis [25] use this expression to compute an absolute minimum supply voltage (V_{DD}) for an ideal MOS device (i.e. one with a subthreshold slope of 60 mV/decade at 300 K) of $V_{DD}(\min) \approx 2(kT/q)\ln(2) \approx 36$ mV at 300 K. Similarly, it was determined in [26] that a $V_{DD}(\min)$ of approximately 83 mV at 300 K is necessary to maintain a logic gain $(A) > 4$ in a standard CMOS gate with a fanin of 3.

Quantum Mechanical Considerations

A second fundamental limit arises from quantum mechanics, or more particularly from the limitations imposed by quantum mechanical tunneling through the barrier. The (Heisenberg) uncertainty in momentum corresponding to a barrier of height E_b sets a lower bound on the barrier width (a_{min}) such that [24]:

$$a_{\min} \cong \frac{\hbar}{\sqrt{2m_e E_b}}, \tag{2.2}$$

(m_e is the effective electron mass and $\hbar$ the reduced Planck constant) which sets a minimum room temperature barrier width of:

$$a_{\min} \cong \frac{\hbar}{\sqrt{2m_e k_B T \ln 2}} \approx 1.45nm\,. \tag{2.3}$$

The smallest transit time for an electron through this barrier results in an absolute minimum delay time and is derived in [23] as:

$$\tau_p = \frac{\pi\hbar}{\Delta E} = \frac{\pi\hbar}{k_b T \ln 2} \approx 0.11 pS\,, \tag{2.4}$$

which is (coincidentally) almost the same as the ITRS prediction for gate delay at the 2018 (16nm) node. In contrast, using a method based on material electrostatics, Meindl has calculated a slightly higher unit transit time of 0.33pS for silicon and 0.25pS for GaAs [22].

Meindl also showed that the uncertainty in momentum leads to a bound on the average power (P) transferred during a switching transition (Δt), i.e., the switching transition of a single electron wave packet, such that:

$$P \geq \frac{h}{(\Delta t)^2}\,. \tag{2.5}$$

As the operation of *all* devices based on charge transport, including Field Effect Transistors plus all of the more esoteric technologies—Resonant Tunneling Devices, Single Electron Transistors, Quantum Cellular Arrays etc.—involves the charging and discharging of capacitances to change the height of the controlling barrier, the energy required to move the barrier is equivalent to the energy required to charge these control capacitances. This energy is eventually dissipated as heat. Using the smallest energy dissipation given by the SNL expression (~17meV), the maximum power density is derived in [23] as:

$$P = \frac{3 \text{ x } 10^{-21}\text{J}}{a_{\min}^2 \tau_p} \approx 1.2\text{x}10^6 \text{W/cm}^2\,. \tag{2.6}$$

Given that current cooling methods are limited to a few hundred watts/cm^2, this power density is clearly too large to be physically achievable. The inescapable conclusion is that all charge-based nanoscale electronic systems will be ultimately limited by power/energy density regardless of their implementation technology. It is worth noting that although (2.1) exhibits an exponential sensitivity to temperature, cryogenic operation will not change these energy constraints as it

simply exchanges chip dissipation for cooling energy. Refrigeration losses will always result in a substantial increase in the overall "wall-socket" power [23].

Electromagnetics

Electromagnetic considerations limit the propagation velocity (v) of a pulse to less than the speed of light in free space so that:

$$v = \frac{L}{\tau} \leq c_0 \tag{2.7}$$

where L is the line length and τ is the transit time across the line. While the free space limit is independent of materials or implementation structure, the propagation of an electromagnetic wave across an interconnect line will be ultimately constrained by the dielectric constant (κ) of the material surrounding the line. As a rule of thumb, propagation velocity is proportional to $3.3\sqrt{\kappa}$, which is approximately 6.5ps/mm for a SiO_2 with $\kappa \approx 3.9$.

2.2 Material Limits

Most of the major advances in semiconductor technology over the past 50 years have been achieved using the same basic metal oxide semiconductor [MOS] switching element and with a limited number of materials (primarily Si, SiO_2, Al, Cu, Si_3N_4, $TiSi_2$, TiN, and W) [27]. Manufacturing processes have obviously improved over that period, so that feature sizes have reduced by four orders of magnitude while wafer areas have grown by a factor of around 600. However, until the recent (2007) announcement of 45nm processes using a dual-metal gate with hafnium-based hi-κ dielectric [28], no fundamentally new materials or fabrication processes had been introduced that altered the basic transistor topology (Figure 2). To date, there are no obvious successors to silicon MOS technology that offer sufficient improvements to justify their costs. The run to the end of the roadmap will therefore comprise increasingly difficult incremental improvements to MOS, such as high and low-κ dielectrics, silicon-on-insulator and multiple-gate topologies (see Section 2.3, below).

The limits imposed by materials are determined by the properties of the particular materials but are essentially independent of the particular structural features and/or device dimensions [27, 29]. A key limitation to future scaling is the dielectric constant of the insulator materials used in the gate stack and as part of a multi-level interconnection network. The continued use of silicon also imposes limits on the basic switching energy, transit time and thermal conductance as well as on the fluctuations caused by dopant atoms.

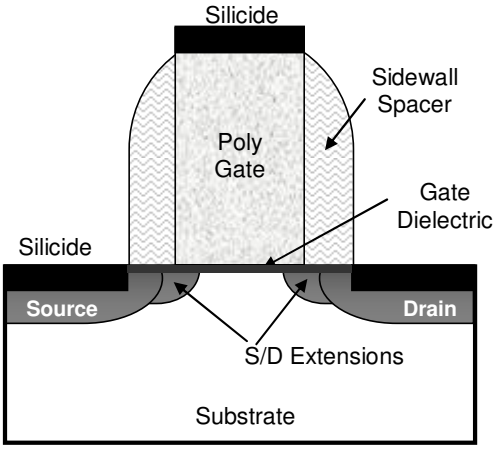


Figure 2. Simplified cross section of a modern MOS transistor.
(based on [27])

The interface between silicon and its native oxide, SiO_2, is atomically abrupt and is relatively easy to fabricate with small defect and charge trap densities. It will be difficult for any alternative material to match these almost ideal characteristics. To maintain good channel control, and restrict short-channel effects, the oxide thickness (T_{OX}) must scale with channel length. Various "rules-of-thumb" have been proposed, but it appears that roughly $T_{OX} < L_g/4$ will ensure adequate gate control. Gate oxide thicknesses will be ultimately constrained by quantum mechanical tunneling of carriers through the insulator. The direct tunneling probability (T) for a rectangular barrier has an exponential form similar to that of (2.1), i.e.:

$$T = e^{\left(-2\sqrt{\frac{2m^* qE_b}{\hbar^2}}T_{OX}\right)} \tag{2.8}$$

where m* is the electron effective mass, E_b the barrier height for the tunneling particle and $\hbar$ the reduced Planck's constant. Thus, gate current will increase exponentially with decreasing oxide

thickness, T_{OX}, resulting in excessive standby power at oxide dimensions of less than 1.0–1.5nm (about 5–7 atomic layers). Further, at these dimensions the wave functions of the gate and the silicon substrate begin to overlap, causing scattering and reduced mobility, significantly degrading the switching performance [30]. Insulators with higher dielectric constants will allow the same effective electric field at a thicker T_{OX}, thus reducing tunneling currents. Silicon nitride ($\kappa \approx 7.4$) is likely to be the first hi-κ dielectric to be widely adopted in mass-production, at the 65nm node [31, 32].

2.3 CMOS Device Scaling

CMOS has been the dominant technology in commercial VLSI systems for more than 25 years, during which time transistor gate lengths have shrunk from several microns to typical commercial dimensions of 180 to 130nm [33] and to 90–65nm in high performance systems [34]. Although the semiconductor industry ultimately expects to be able to scale CMOS gate lengths to a few nanometers [11], it is far from clear how this might be achieved. In the near term, it is almost certain to exploit what the ITRS calls “enhanced CMOS” i.e., the integration of new technologies into the standard CMOS fabrication process (Figure 3).

Some early predictions (e.g. [35]) suggested that gains in FET device performance might eventually stall as the minimum effective channel length approaches 30nm at a supply voltage of 1.0V and a gate oxide thickness of about 1.5nm. However, this shows no sign of occurring. Devices with physical gate lengths as small as 10nm have already been built on research lines (e.g. [36-39]) and by mid to late 2006 [40] manufacturers such as Intel, TSMC and Toshiba had successfully mass-produced transistors with sub-25nm gate lengths (i.e. at the 65nm node) and with core voltages of 1.0V. Intel’s mass-produced 45nm transistor mentioned above was developed some 2–3 years before the ITRS prediction for this technology, on bulk silicon rather than SOI. The 2005 ITRS now predicts that supply voltage scaling will tend to level out at about 0.5V for low-operating power (LOP) and 0.7V for high performance technology. This can be compared to the theoretical minimum of 2–3kT/q. It is suggested in [41] that when devices operate within their

ballistic region, the optimum value of V_{DD} *increases* slightly as the channel length decreases due to the effect of increasing static leakage power as gate length reduces, although the effect is small relative to the supply voltage.

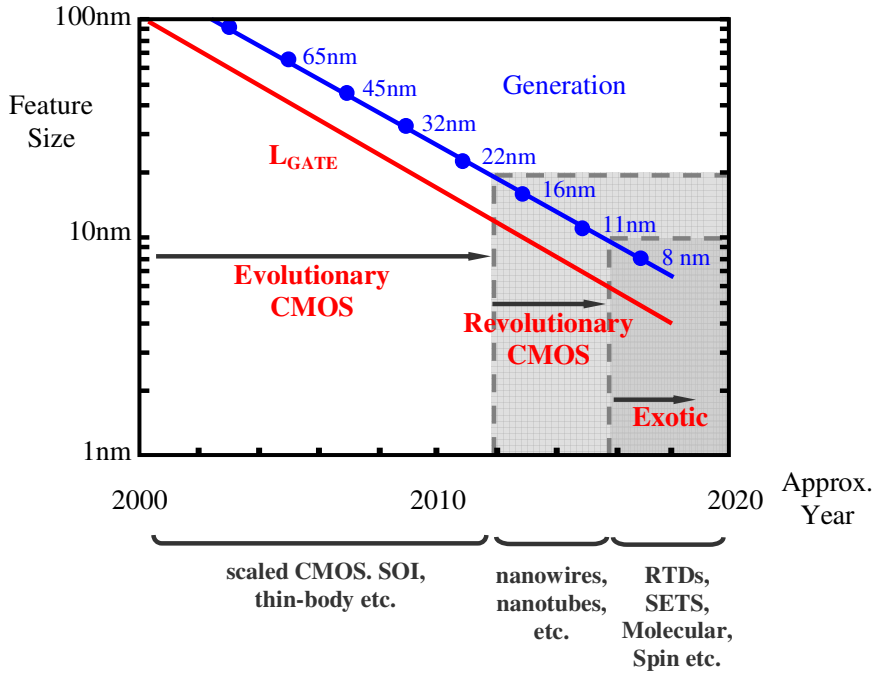


Figure 3. Predicted evolution of CMOS technology (adapted from [11] and [42]).

In order to constrain escalating power-densities and at the same time maintain adequate reliability margins, traditional CMOS scaling has relied on the simultaneous reduction of device dimensions, isolation, interconnect dimensions, and supply voltages [35]. Eventually, FET scaling will be limited by a combination of high fields in the gate oxide and channel plus short channel effects that reduce device thresholds and increased subthreshold leakage currents [43]. By 2020 the ITRS is predicting effective gate lengths of 7–12nm (Figure 3) with equivalent gate oxide thicknesses of 5–8Å. Beyond this point, any further performance growth will need to rely on either increased functional integration with an emphasis on circuit and architectural innovations, or on a move to a technology that is not based on the transfer of charge.

2.3.1 Silicon-on-Insulator

Silicon-on-Insulator (SOI) has been suggested as a solution for many of the problems with scaled CMOS. The ITRS predicts its commercial application as early as 2011. Theoretically, small SOI devices do not need channel doping and can therefore be scaled to dimensions below 10nm without running into problems of uncontrollable parameter variations due to the random distribution of dopant atoms [44]. However, difficulties in controlling device parasitics plus the need for tight dimensional control may sabotage potential performance gains [45]. Figure 4a shows a simple SOI device structure in which the thin film channel is totally isolated from the body by a thick oxide (the body oxide, or BOX).

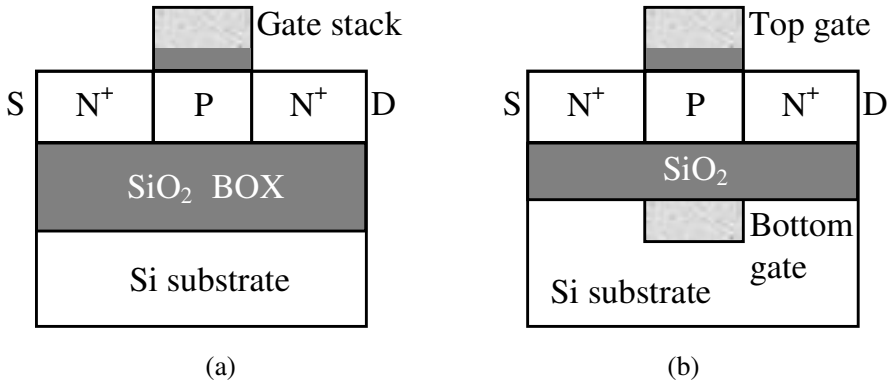


Figure 4. Conventional Silicon On Insulator (SOI) device topology
(a) Single-gate SOI and (b) Double-gate SOI.

The double-gate SOI transistor (Figure 4b) is inherently resistant to short-channel effects and can exhibit close to ideal subthreshold performance [46]. Ultra-thin body, fully-depleted, double-gate SOI has already been suggested as an effective low power technology [47] and double-gate devices exhibit additional functionality that make them well suited to reconfigurable architectures. In particular, they can theoretically be built on top of other structures in three-dimensional layouts and they do not require ancillary structures such as body contacts and well structures that enlarge traditional CMOS layouts. Finally, the second (back) gate offers a means of controlling the threshold of the logic device in a way that can used to configure the system. This threshold control mechanism forms the basis of operation of the reconfigurable mesh that will be described in Chapter 3.

2.3.2 Extreme Device Scaling—Schottky Barrier MOSFETs

The increased difficulty in maintaining low off currents (I_{OFF}) as channel lengths scale below 50nm has resulted in a revival of interest in Schottky barrier MOSFETs, first described almost 40 years ago [48], in which metal silicides (e.g. PtSi, ErSi etc.) replace the heavily doped silicon source and drain regions (Figure 5) [49-51]. Metal silicides form natural Schottky barriers to silicon substrates, acting to confine carriers and reducing or eliminating the need for impurities in the channel to prevent current flow in the "off" condition [52]. They exhibit several advantages when compared with conventional devices, including the elimination of punch-through and latch-up as well as offering a significantly simpler processing technology. As shown in Figure 5, they are also potentially more compact than conventional CMOS due to the elimination of the well(s), body contacts and isolation regions.

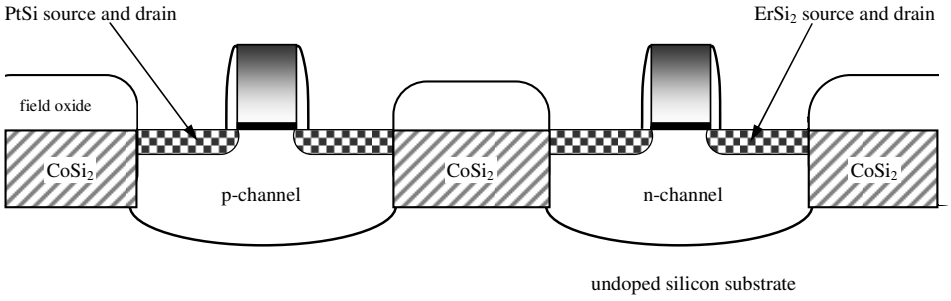


Figure 5. A Schottky barrier CMOS inverter
with buried epitaxial self-planarized $CoSi_2$ local interconnects (adapted from [49]).

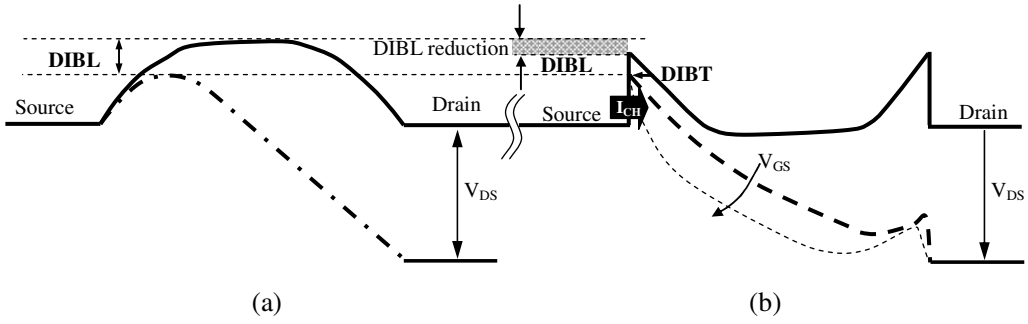


Figure 6. DIBL mechanisms in (a) Double-gate MOSFETs and (b) Schottky barrier FETs showing DIBL reduction in SBFETs (adapted from [53] and [54]).

In conventional short channel MOS devices, an increase in the drain voltage will cause a decrease in the built-in potential between the source and the channel resulting in increased subthreshold

current due to DIBL (Drain-Induced Barrier Lowering). In contrast, the subthreshold characteristics of a SB-MOSFET (including DIBL and subthreshold slope) are mainly determined by the barrier itself [53]. The primary switching mechanism (Figure 6) involves a reduction in the thickness of the tunneling barrier between the source and channel under the influence of the gate potential. It can also be seen that an increase in the drain voltage will cause both DIBL and Drain-Induced Barrier Thinning (DIBT) [54] effects to occur simultaneously. The DIBL effect increases the thermionic current over the barrier, whereas DIBT causes a decrease in threshold voltage due to the thinner tunneling barrier.

At the ultimate dimensions for this technology (i.e., gate lengths below 10nm [55]), the channels would effectively become undoped silicon wires with regular silicide patterns forming the source/drain regions. At this scale it is conceivable for them to approach densities of 10^8 gates/mm^2. However, this comes at a cost. The overall current drive of Schottky barrier devices can be significantly lower than MOS due to the very high resistance of their source/drain regions at low supply voltages [56] although it was shown in [57] that the barrier height (and therefore the junction resistance) can be substantially reduced by the inclusion of a thin insulating layer at the metal/semiconductor contact. Any loss in performance implied by an increased $\tau = CV/I$ would have to be made up either by reducing C (by using local interconnect, for example) or at other levels in the design hierarchy.

2.3.3 Device Variability

Uncontrolled variations in device performance are already critical to analog circuits and, as devices move into the nanoscale domain, they will become increasingly important to digital logic. Three main sources of variability are considered here: global manufacturing uncertainty, local random fluctuations and temperature. As scaling continues, new sources of variability that were negligible in previous generations will increase the local component of the total variance. In [58] it is suggested that the following effects will become dominant: over/under etching of small geometries, proximity effects, doping fluctuations along the channel and lateral diffusion of

dopants between adjacent high-energy implanted wells. To this list can be added atomic scale interface and line edge roughness and increased charge trapping [59, 60].

Traditionally, die-to-die (D2D) variation has been the primary concern although systematic within-die (WID) variation is likely to have a greater effect on future device scaling [61]. In addition, elevated operating temperatures and the presence of "hot-spots" will cause changes to mobility, threshold voltage and subthreshold slope across the surface of the chip. The cumulative effect of all of these variations will be to cause increasingly large uncertainty in key performance metrics including intrinsic delay, switching threshold, noise margins and dynamic and static power.

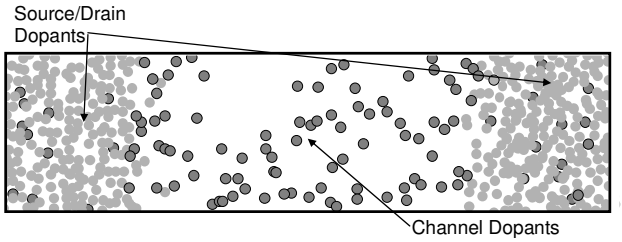


Figure 7. Random placement of impurities in device channel based on [66].

Doping Variability

Variations in the number and position of channel dopant atoms (Figure 7) can already cause significant intra-device variation in on-current, mobility and threshold voltage [62], and this effect will become more important in future devices. For example, at a doping level (N_a) of $2x10^{19}/cm^3$, a channel of L = W = 25nm with T_{SI} = 10nm may contain just over 100 dopant atoms. Standard models based on Poisson statistics (e.g., [63]) that assume that dopant induced variability is roughly proportional to $1/\sqrt{N_a}$ would predict around 10% mismatch between these devices. However, dopant induced fluctuations contribute only part of the overall variability (e.g., < 50–60% of ΔV_{TH} for the studies in [64] and [65]). Further, the simulations reported in [66] predict that V_{TH} will be almost insensitive to channel doping with concentrations below $2x10^{18}/cm^3$ and

below $10^{17}/cm^3$, the standard deviation of threshold variation (σ_{vth}) due to discrete dopant placement is predicted to fall to less than 7 mV.

In [67], it was shown that, by using a single-ion implantation technique to place individual dopant atoms in precisely controlled locations in a transistor channel, the standard deviation of the V_{TH} distribution could be reduced to a third of its value with random doping (approximately ±50 mV vs. ±150 mV). Also, the shift in V_{TH} achieved for a given implant density was about twice that of the random case (-0.4V vs. -0.2V) due to a lower uniform channel potential. Rather than offering a solution to the problem (even the authors admit that the technique is unlikely to be appropriate for mass production), this result serves to clearly demonstrate the importance of channel dopant distribution to the characteristics of semiconductor devices.

A rule of thumb is suggested in [68] that gives the standard deviation of the local threshold mismatch as $aT_{OX} / \sqrt{WL}$, where W, L = active area width, length in µm and oxide thickness, T_{OX} is in nanometers and *a* is an empirical constant. The reported range for *a* is 1–2mVµm for current technology but it may decrease below 1 mVµm in the future. Using ITRS bulk technology data as an illustration, a minimum size transistor in 65nm HP technology (L_{eff}= 21.6nm, T_{OX} = 1.2nm) might exhibit a $\pm 3\sigma_{vth}$ spread of as much as ±170 mV. For this reason a "constant fluctuation" scaling scheme has been proposed in [63], in which $N_a^{1/4} T_{OX}$ is reduced at each successive technology node by a factor proportional to the feature size.

Dimensional Variability

As just outlined, due to the increased difficulty of scaling planar bulk devices, scaled CMOS structures are tending to move towards ultra-thin body SOI with channel thicknesses in the range of nanometers, gate oxide dimensions equivalent to a few atomic layers and undoped or lightly doped channels. However, typical interface roughness values of the order of ±1 to 2 atomic layers (~±0.3-0.6nm) may result in significant variations in both body and oxide thickness between individual transistors. Further, line edge roughness (LER) will need to be reduced to well below 5nm to keep its effect below that of the remaining sources of uncertainty [69]. For example, the

quantum corrected simulations reported in [70] show that body thickness and length variations alone could result in one standard deviation of the threshold voltage variation (σ_{vth}) reaching ~25 mV for thin-body SOI devices with L_g = 5nm and T_{SI} = 2nm.

Figure 8 plots the sensitivity of the threshold voltage to both silicon thickness ($\Delta V_{TH}/\Delta T_{SI}$) and channel length ($\Delta V_{TH}/\Delta L_g$) for a small number of ultra-thin-body SOI devices. These are typical end-of-roadmap devices formed with intrinsic channels (e.g. ~10^{13}–10^{15} cm^{-3}), using the (metal) gate workfunction to set the threshold. The data for conventionally doped S/D devices are from [62] and [66]. The points labelled *silicide* represent devices with fully silicided source and drain regions of the sort described in [71] that have been simulated using classical drift-diffusion models. In this case, the source and drain are defined by the silicide/body boundary, which can be atomically sharp, resulting in well-defined channel lengths even under extreme scaling.

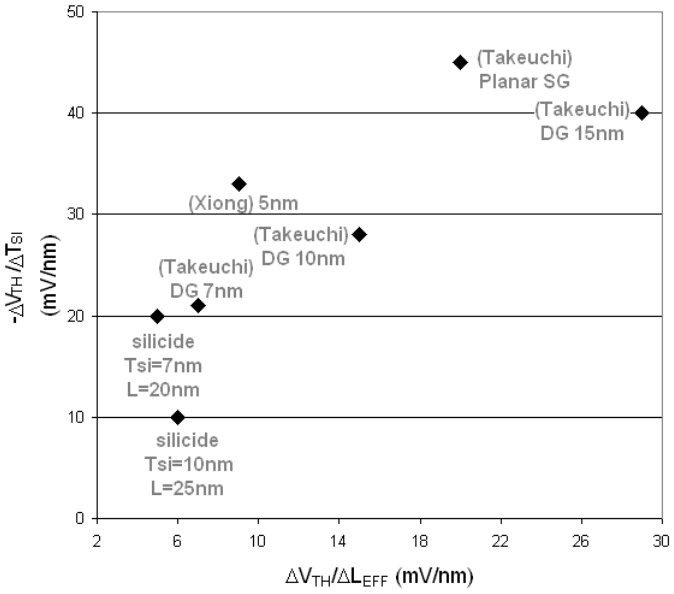


Figure 8. V_{TH} sensitivity for ultra-thin-body double gate devices (T_{SI} as shown). Data for doped S/D, undoped channel devices are from [62] (Takeuchi) and [66] (Xiong) are all with L_g = 25nm. Also shown are two 10nm Schottky barrier devices with completely silicided S/D simulated using drift-diffusion models in Atlas TCAD (silicide: L_g = 25nm and L_g = 20nm).

As gate lengths reduce it will be necessary to reduce the body thickness to a few nanometers (e.g. $T_{SI} \le 0.5L_g - 6T_{OX}$ [66]) in order to maintain good electrostatic control and to keep short-channel effects under control. It should be noted that these simulations do not account for the effects of

energy quantization in thin-body devices at high electric fields. The quantum effect on threshold voltage becomes non-negligible when the surface electric field of the inversion layer (E_S) is > 10^5V/cm and beyond $E_S = 10^6$V/cm, ΔV_{TH} may exceed 200 mV due to quantum effects [37]. Techniques such as retrograde [72] or super-halo doping profiles [73] have been proposed to minimize these effects in bulk devices.

Based on the $\Delta V_{TH}/\Delta L_g$ and $\Delta V_{TH}/\Delta T_{SI}$ figures for the 10nm silicide device and assuming that the ITRS targets for interface and line-edge roughness can be met, we may reasonably expect that a threshold variability figure in the range ±15–25% of V_{TH} could be achievable in the long term. This is consistent with the relative variations found in [66] and [74], and will be used in the analyses in the following chapters. For example, the 20nm symmetric FinFET simulations reported in [66] were much better than this, with a 30% variation in the relative gate oxide dimensions resulted in only minor changes to V_{TH} (~3 mV) and subthreshold slope (< 1 mV/dec.).

Temperature Induced Variability

The main temperature dependant device parameters of interest here are the threshold voltage, V_{TH} and subthreshold slope, S [75]. Mobility is also affected but with $V_{DD} < 1$V, the fall in V_{TH} given by $\Delta V_{TH} = V_{TH0} - \kappa \Delta T$ tends to dominate the effect of mobility degradation. The temperature coefficient, (κ, currently about 1–2 mV/K [76]) is a function of doping density and will therefore tend to diminish with reduced channel doping in future technology generations. This idea is reinforced by Figure 9, which is based on data from [77] that assumes a 5x increase in I_{OFF} at successive generations. The data in Figure 9 have been normalized to their room temperature values at 250nm and show the expected reduction in temperature sensitivity at future nodes. Calculating the equivalent shift in threshold voltage as $\Delta V_{TH} \approx n V_t \ln(I^2_{OFF} / I^1_{OFF})$, where the superscripts (1 and 2) represent successive technology nodes, and ignoring any subthreshold slope change, the equivalent ΔV_{TH} falls from around 1.5 mV/K at the 250nm node to less than 70 µV/K at the 45nm node, implying that temperature-induced variability will become a much smaller

component of overall variability as technology moves towards the end of the CMOS roadmap (although, clearly, the absolute value of I_{OFF} will remain an issue).

While the interaction between temperature and device performance is complex and technology-dependent, it may be collapsed to a simple empirical model relating maximum operating frequency (F_{MAX}) to temperature [78]:

$$\Delta F_{MAX} \propto \Delta T^{-a} \tag{2.9}$$

where T = temperature and 0.5< a <0.75 is an empirically derived exponent. Here, F_{MAX} takes its conventional definition of the maximum clock frequency that can be achieved without violating internal setup and hold times.

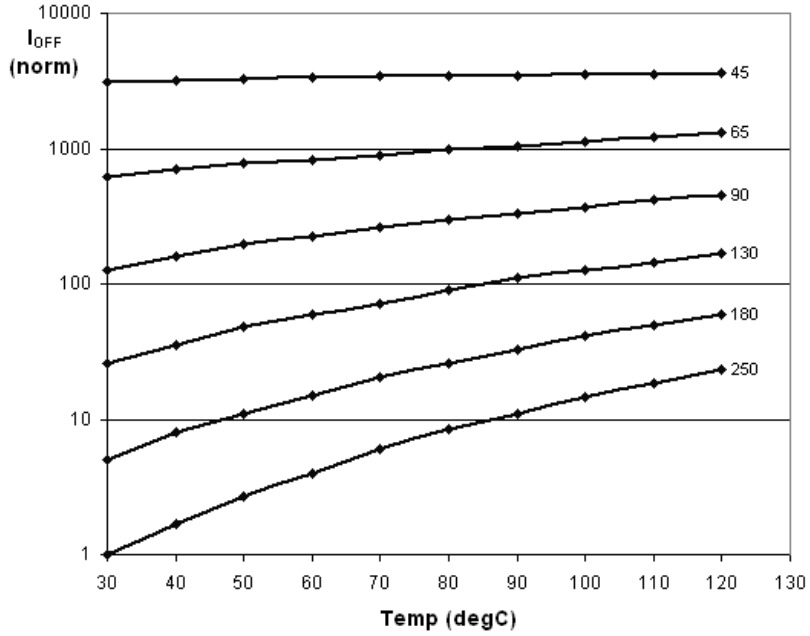


Figure 9. Leakage current temperature characteristics
from 250nm down to 45nm, I_{OFF} normalized to 30°C at 250nm. Data from [77].

This model predicts that at around room temperature, the operating frequency will fall only modestly with increasing temperature. For example, a rise from 300 K to 350 K will reduce the operating frequency by around 10% (given a = 0.63). Various studies (e.g., [79]) have found intra-die

temperature gradients of this order in high-performance processors. By contrast, subthreshold slope $S = (nkT/q)\ln(10)$ is directly proportional to temperature. Thus the same temperature gradient would cause an initial subthreshold slope of 65 mV/dec. to rise to ~75 mV/dec., implying an increase in subthreshold current of more than an order of magnitude.

2.4 Interconnect Scaling Limits

The propagation of a voltage down a generalized wire can be described by the Telegrapher's Equation:

$$\frac{\partial^2 V}{\partial x^2} = RC\frac{\partial V}{\partial t} + LC\frac{\partial^2 V}{\partial t^2} \tag{2.10}$$

where x is the distance along the wire, t is time, V is voltage, and R, L, and C are the distributed resistance, inductance and capacitance per unit length, respectively. At low frequencies, the first term on the RHS of (2.10) dominates and the wire behaves as a distributed RC network. As the frequency increases, the second term becomes significant, allowing for the propagation of an electromagnetic wave. As a result, the maximum phase velocity will be given by the speed of light in the dielectric that surrounds the interconnect line and this (as opposed to group velocity) will determine the signal propagation delay in high-speed transmission line systems.

Transmission Line Considerations

From conventional transmission line theory, the overall delay from the output of a driving gate to the input of its load will be minimal if the output resistance of the driver equals the characteristic impedance of the interconnect line, $Z_0 = \sqrt{L/C}$, and the total resistance of the line is small compared with Z_0. However, this is rarely the case in CMOS, and will become increasingly less so as device dimensions decrease. This can be seen in Table 1, which compares the effective output impedance of a unit square transistor using high-performance data from the ITRS [11] with $Z_0 = \sqrt{L/C}$ for a copper Metal 1 layer at the same technology node. In this case, the effec-

tive output resistance is taken to be $R_{ON}(\min) = KV_{DD} / I_D(sat)$, a first-order approximation to its lower bound.

In [81] the empirical constant (K) is given a range from 0.69–0.83 for 250nm down to 50nm. Assuming that it can be linearly extrapolated, K will be approximately 0.95 at the ITRS scaling limit. The characteristic impedance is estimated by applying representative data from [80] to the interconnect topology of Figure 10 and assuming that the dielectric constant will progressively reduce to about 1.7 by the end of the ITRS. It is evident that the effective output resistance is likely to remain remarkably constant (in the range of 10–20KΩ) across technologies and over the remaining years of the roadmap. Similarly, the characteristic impedance of the lowest interconnect layer will remain in the range 100–150Ω.

Table 1 Comparing Minimum Effective Output Resistance (R_{ON}) to Estimated Z_0 of M1 for some High Performance Technologies from the ITRS.

Year	HP Technology	L_{Drawn} (μm)	V_{DD}	I_D (sat) (μA/μm)	R_{ON} (min) Ω	Z_0 (M1) Ω	W/L
2004	Bulk	0.090	1.2	1024	1.3E+04	94	138
2007	Bulk	0.065	1.1	1197	1.4E+04	112	126
2010	SOI	0.045	1.0	1812	1.2E+04	120	102
2013	SOI	0.032	0.9	2212	1.3E+04	126	101
2016	DG	0.022	0.8	2763	1.3E+04	132	99
2019	DG	0.016	0.7	2677	1.7E+04	146	115

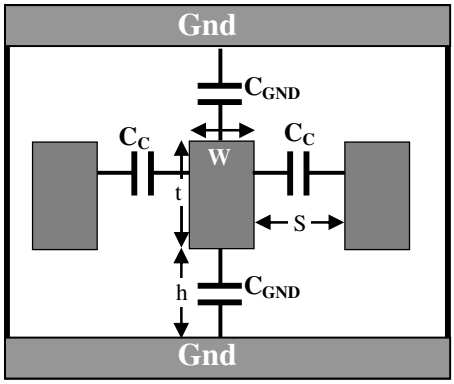


Figure 10. Interconnection capacitance model (from [80]).

The final column in Table 1 represents the width to length ratio that would be required to achieve $R_{ON} = Z_0$. This is consistently an order of magnitude greater than expected for a typical gate driving a local interconnect line. Although many researchers argue that a complete model of interconnect delay must include inductance, and by implication Z_0 (e.g. [82-86]), it can be seen from Table 1 that this will only be important to delay calculations where very low impedance drivers are used to drive long interconnection lines. In all other cases, the simple RC delay will be the dominant effect.

Crosstalk Issues

The mutual inductance and capacitance between lines will impose increasingly severe limits due to crosstalk. Its importance can be gauged using the approximate expression for normalized peak capacitive cross-talk noise (V_n/V_{DD}) developed in [87]:

$$\frac{V_n}{V_{DD}} \approx \left(\frac{\pi}{4}\right)\left(C_{GND} + C_M\right) \tag{2.11}$$

where C_M is the mutual capacitance between adjacent interconnects and C_{GND} is the capacitance between interconnect and the ground plane. In Figure 11, the curve for a conventional wire is derived using the simplified capacitance model of Figure 10. This is based on ITRS predictions at the (2019) 16nm node for minimum conductor width (W = 16nm), thickness (h = 32nm) with a best-case height above ground plane (t = 48nm) and plots $C_R = V_n / V_{DD}$ vs. the separation between adjacent conductors (S). Under these conditions, crosstalk would reach approximately 33% at a separation equal to the minimum contacted local metal pitch (MP = 3.17 x drawn feature size). On the other hand, constraining crosstalk to a more reasonable figure (e.g., around 25%) would require a separation of at least 4 x feature size (i.e. around 64nm). Thus, while the *minimum contacted local metal pitch* represents an absolute minimum dimension imposed by lithography constraints, the ultimate minimum separation may have to be larger (by ~20% in this example) to accommodate signal integrity constraints.

The second set of curves in Figure 11 show the results of applying (2.11) to the cross-bar structure described in [88], comprising ~30nm wires on a 60nm pitch (i.e., S = 30nm). The ground-plane separations were experimentally set to 50nm, 25nm and 10nm. As expected, the ground-plane insulator thickness is critical to achieving the objective of high density, low-noise interconnects. For example, to limit the crosstalk to less than 25% of V_{DD} these crossbar wires would have to be positioned less than 25nm (< S) above a ground plane. On the other hand, inductive crosstalk is unlikely to be an issue at any stage. Using the approximate crosstalk models in [89], the high impedances and short lengths of nanowire interconnections imply that signal edge rates would have to be in the order of 10^{14}V/sec before inductive crosstalk could become a serious problem.

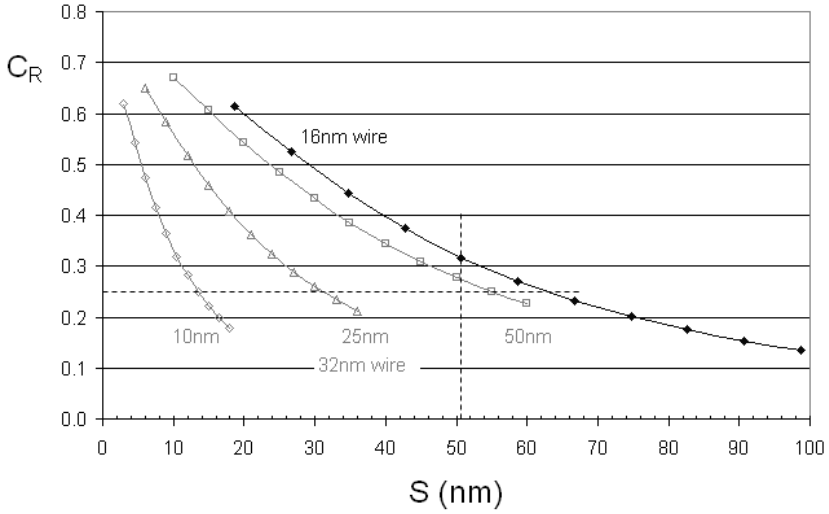


Figure 11. Crosstalk ratio ($C_R = V_n/V_{DD}$) vs. interconnect separation (S) for a 16 x 32nm wire and 32nm with Ground-plane height h =10, 25 and 50mn as shown. The vertical dotted line is at F = 3.17x16nm≈51nm = minimum contacted local metal pitch.

2.4.1 Interconnect Delay Scaling

The physical scaling of interconnections into the nanometer regime will face many technical challenges. Davis *et al*. [85] list a number of major problems such as resistivity degradation, material integration issues, high-aspect ratio via and wire coverage, planarity control, and reliability problems due to electrical, thermal, and mechanical stresses in a multilevel wire stack. At a basic level, the wiring delay problem is simple to articulate: as interconnection width and thickness decrease, resistance per unit length increases, while as interconnections become denser (and

oxide layers thinner), capacitance also tends to increase [90]. For example, if the RC delay of a 1mm metal line in 0.5µm technology is 15ps then at 100nm (in the same materials) the delay would be 340ps [91].

A common rule-of-thumb used in commercial designs is to insert a buffer wherever the intrinsic gate delay equals the RC time constant of its wire. This is the so-called *drive distance* at which the signal propagation speed is considered to be optimal [92]. Even assuming that wire technology stays the same, the combination of the thinner wires and thinner field oxides in advanced process technology will increase a wire's RC time constant per unit length, as identified in the previous section. This is equivalent to saying that to achieve a constant RC value the wire length must decrease as technology advances. At the same time, intrinsic gate delay is falling and thus drive distance will decrease at a rate set by the product of the two. As a result, the distance and number of gates that may be considered "local" i.e., that are directly reachable by a signal in a single gate delay, is rapidly shrinking, and so using this rule would require an increasing number of buffers just to implement the design. For example, under the assumptions of [92] at the 100nm generation, only 16 percent of the die was predicted to be reachable within a single clock cycle.

More detailed analyses of scaled wires in [93] and [94] have identified two distinct performance regions. For short connections (those that tend to dominate current ASIC wiring), the ratio of local interconnection delay to gate delay remains very close to unity so that interconnection delay closely tracks gate delay with scaling. On the other hand, global wiring tends to increase in length with increasing levels of integration, implying that the interconnection delay of these wires will increase relative to intrinsic gate delay. Sylvester and Keutzer [91] concluded that the scaling of global wires will be increasingly unsustainable below about 180nm due to the rising RC delays of scaled-dimension conductors. As interconnect delay appears to be tolerably small in blocks of 50 – 100,000 gates, they argue for hierarchical implementation methodologies based on macro-blocks of this size. Their results could equally be used to support a case for flat, locally connected organizations.

An estimate of the interconnect delay can be derived using the empirical formula developed by Fisher and Nesbitt [95]:

$$t_{int} = R_o C_T + 0.4((R_w C_w)^{1.6} + (tof)^{1.6})^{1/1.6} + 0.7 R_w C_g \quad (2.12)$$

where R_o is the output resistance of the device, C_T is the total line capacitance driven by the gate (= $C_o + foC_g + C_W$, fo = fanout, R_w and C_w relate to the wiring and C_g is the load gate capacitance). The time of flight (*tof*) could fall to ~4.9ps/mm with K = 2.2 and will therefore remain comparatively small for local connections. Thus, if *tof* is ignored, (2.12) simplifies to three terms:

$$t_{\text{int}} = R_o(C_o + foC_g + C_W) + 0.4 C_W R_W + 0.7 C_g R_W . \quad (2.13)$$

While the actual RC values clearly depend on the particular implementation technology, it is possible to derive some comparative insight using published as well as predicted data. Table 2 presents some examples with estimates of the output resistance and gate capacitance for the following three plausible ultimately scaled devices:

1. An end-of-roadmap double-gate silicon transistor [96] with metal gate, physical length = 5nm, $W/L_g = 3:1$;

2. The Metal-Insulator Tunneling Transistor described in [97, 98], based on a Ti/TiOx tunneling barrier and representing a simplified manufacturing technology;

3. A high performance Carbon Nanotube device with ~100nm gate length, 1.4nm tube diameter, Al_2O_3 gate dielectric ($\kappa = 9.4$) and thickness T_{OX} = 1.5nm [99], representing an advanced and (to date) completely speculative technology.

The last row of the table includes the RC figures derived from [80] for copper-based interconnects based on the dimensions described in Figure 12 using $\kappa = 2.2$. The general trend of the resulting interconnect delay is shown in Figure 13. This approximate analysis illustrates that, regardless of technology and even at line lengths below 1μm, interconnect delay has the potential to become significantly greater than the intrinsic gate delay.

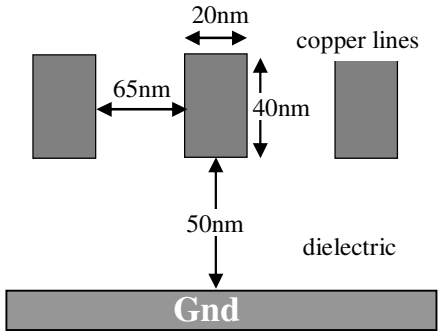


Figure 12. Example interconnect topology

Table 2 Estimated RC values of some potential implementation technologies

Device [reference]	R_o (on) (Ω)	C_g (aF)
DG silicon (V_{GS}=1.2V) MASTAR [96]	$1.2x10^4$	5.1
MITT (2nm SiO_2 gate insulator) [97]	$5x10^9$	0.0265
Carbon Nanotube (p-type, ∅=1.4nm) [100]	$1x10^5$	~0.2
Cu M1 local interconnect [80]	$2.75x10^4$ (Ω/mm)	80 (fF/mm)

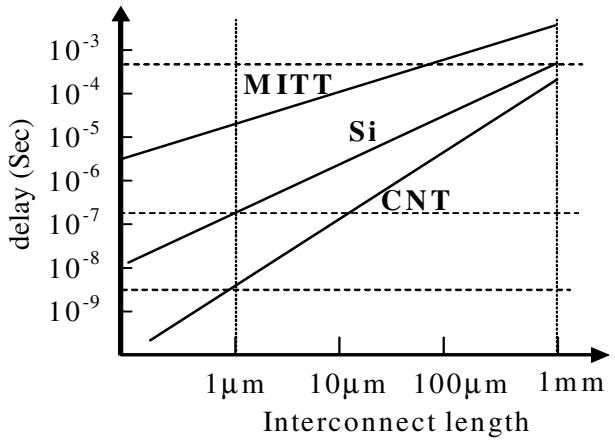


Figure 13. Estimated interconnect delay based on 10nm technology

At future (gigahertz) operating speeds, it is likely that both time-of-flight and signal attenuation will impact on the performance of long interconnect lines and will not be able to be ignored. As both of these depend on the dielectric constant of the propagation material, their solution will require significant changes to processing technology. For example, [101] describes a process that uses a gas-isolated, high-k gate dielectric, metal-gate, metal-substrate SOI scheme with thermally

conducting through-holes to reduce temperature variations and increase interconnection reliability. This complex, aggressive fabrication scheme contrasts markedly with the simplified self-assembly mechanisms proposed by advocates of chemically self-assembled techniques (e.g. [102]). A simpler solution would be to eliminate long interconnection lines from the architecture.

2.5 Performance Modeling in Advanced CMOS

The analyses undertaken later in this work depend on a number of simplified I-V models for advanced transistors. In particular it is assumed that the general form of the saturation drain current equation developed for DSM by Chen *et al.* [103] will remain valid to the end of the roadmap. This section looks briefly at this assumption and also on the general relationships between subthreshold slope, threshold voltage, saturation drive current and propagation delay.

2.5.1 Saturation Drain Current Models

The simplified saturation drive current equation developed for DSM in [103] is given by:

$$I_D(sat) \approx f(R_S) W L_g^{-0.5} T_{OX}^{-0.8} (V - V_{TH})^{\alpha} \tag{2.14}$$

where W is the gate width, L_g its physical gate length and T_{OX} the gate oxide thickness. The term $f(R_S)$ is related to the series resistance of the source/drain and will be small in the mesh configurations studied in this thesis. It will be ignored in all of the following analyses. In (2.14), V is the gate voltage V_G but is assumed to be very close to V_{DD} for CMOS i.e. V_G changes in lock-step with V_{DD}. In this form, Chen's simplified model is very similar to the (now classical) alpha-law model of Sakurai and Newton [104] where the exponent α describes the degree to which the transistors are velocity saturated [105]. Chen suggests a value of about 1.25–1.3 for current technology, but α has been progressively decreasing from around two in long-channel, micron-level devices (equivalent to the original Shockley MOS equation) and may approach unity in ballistic nanoscale technology [106].

To demonstrate this effect and at the same time validate this simplified model under future technology assumptions, a number of simulations were performed using the models built into the

ITRS predictive tool MASTAR (v4.1.0.5, 2005) [96] as well as using the 2D thin-body quantum device simulator nanoMOS3.0 available at http://nanohub.org [107]. Figure 14 shows the relationship between the gate overdrive (V_G-V_{TH}) and I_D(sat) for a bulk device with its supply set to 1.2V, 1.0V and 0.8V as well as a medium term DG-SOI device (~2011, V_{DD} = 1V) and finally an end-of-roadmap (2020) thin-body SOI transistor (V_{DD} = 0.7V). These three are MASTAR models. In all cases, the overdrive has been normalized to its maximum value so that the slopes are easier to compare. In the bulk case, the channel doping was adjusted to set a range of threshold voltages between 160 mV and 240 mV at the supply voltages shown. The SOI devices exhibit low channel doping so, instead, the gate workfunction was adjusted to set the same threshold range.

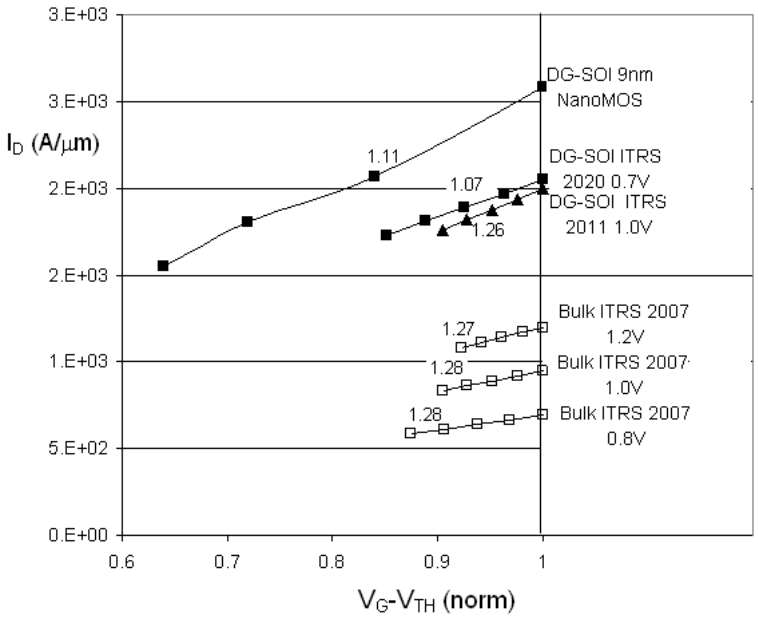


Figure 14. I_D(sat) vs. normalized gate overdrive (V_G-V_{TH}) for a number of advanced transistor models to 2020.

In all cases the slopes in Figure 14 closely follow $(V_G - V_{TH})^{\alpha}$ with α ranging from ~1.3 in bulk technology down to 1.07 in the advanced thin-body SOI device. A similar device simulated at V_{DD} = 0.5V using nanoMOS (with quantum ballistic models) exhibits $\alpha \approx 1.1$, implying that the MASTAR models slightly over-estimate ballistic performance towards the end of the roadmap where increased scattering will prevent truly ballistic performance [108]. In any case, these

experiments do indicate that it is safe to assume that the simplified form of $I_D(\text{sat}) \propto (V_{DD} - V_{TH})^{\alpha}$ will be valid to 2020, by which time α is likely to have fallen to around 1.05–1.1.

2.5.2 Subthreshold Current Models

Static leakage current is predominately a function of the threshold voltage and sub-threshold slope [109]. In bulk CMOS, there is a contribution from the reverse-biased diode junction at the drain/substrate junction, but it tends to be small and will disappear with a move to SIO technology to be replaced by the quantum mechanical tunneling of carriers through the source-to-drain barrier [110].

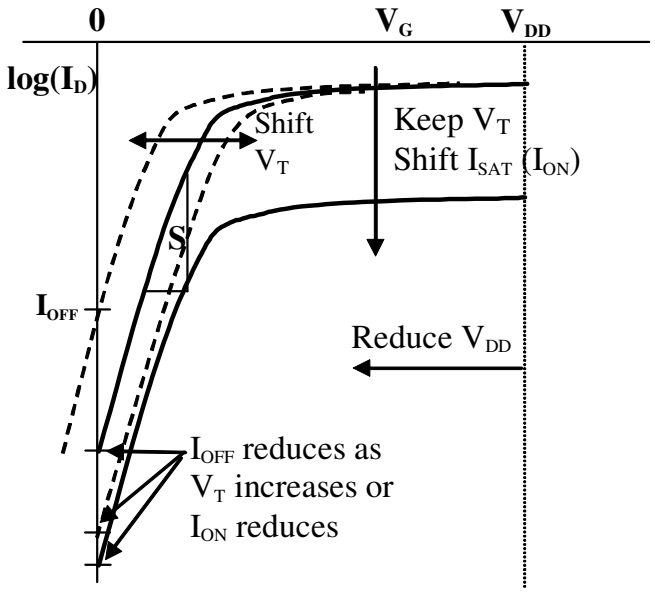


Figure 15. The general impact of shifts in transistor characteristics.

One of the major challenges of nanoscale CMOS will be the control of subthreshold slope, S. The fundamental limit for S in bulk CMOS is approximately 60mV/decade, which puts an ultimate limit on the ratio of I_{ON}/I_{OFF} at small values of supply voltage. Although it is theoretically possible for double gate devices to closely approach this figure, all such devices built to date have tended to exhibit slopes greater than 100 mv/decade. In ultra-thin-film double gate devices, the

subthreshold slope (when the back surface is in depletion) depends largely on the relative gate oxide thicknesses [111]:

$$S = \ln(10)\frac{k_B T}{q}\left(1 + \frac{C_{it_1}}{C_{OX_1}} + \frac{C_{OX_2} + C_{it_2}}{C_{SI} + C_{OX_2} + C_{it_2}}\frac{C_{SI}}{C_{OX_1}}\right)$$
$$\approx 2.3\frac{k_B T}{q}\left(1 + \frac{t_{OX_1}}{t_{OX_2}}\right) \quad (2.15)$$

where C_{it}, C_{OX} and C_{SI} are the capacitances of the interface traps, gate oxides and the silicon (channel) film respectively and the subscripts (1 and 2) indicate the front and back surfaces. Thus, if the thicknesses of the front and back gate oxides are approximately equal, (2.15) reduces to a constant of around 115 mV/decade, consistent with experimental values of S observed to date.

As illustrated in Figure 15, reducing I_{OFF} can be achieved by any combination of: *reducing* S (to its limit), *increasing* V_{TH} (within the constraints imposed by reducing V_{DD}) or *reducing* both I_{SAT} and I_{OFF} in the same ratio, effectively shifting the entire curve downwards. It is interesting to note here that technologies such as silicon nanowire (e.g. [112]) and Schottky barrier will exhibit just this sort of high impedance (low I_D(sat)) and as a result may be well suited to the trade off between operating frequency and leakage power.

2.6 High-Level Technology Drivers

In [113], Claasen defines *technology* as a composite function made up of a number of exponentially varying parameters that have improved over the past 25 years by factors roughly between 10^2 and 10^8 (Table 3), although as even Gordon Moore himself acknowledges, "*no exponential lasts forever*" [114]. However, as the performance of the "canonical" machine could be said to still be a factor of 10^{40} away from the theoretical limits of computing power [115], obviously there is scope for improvement. This section examines some of these higher-level design and technology issues to determine how they might impact future computer architectures.

Table 3 Approximate technology scaling with time (adapted from [113])

Parameter	Time Dependence	Approximate coefficient	Approx. 25 year scaling
Computing power (IPS)	$\exp(c_1 t)$	c_1 : 0.33/year	4×10^3
Solid-state memory	$\exp(m_1 t)$	m_1: 0.28/year	1×10^3
Computing Speed	$\exp(s_1 t)$	s_1: 0.75/year	1×10^8
Dynamic Power	$\exp(-p_1 t)$	p_1: 0.2/year	1×10^{-2}
Magnetic Storage Size	$\exp(m_2 t)$	m_2: 0.32/year	3×10^3
Storage Speed	$\exp(s_2 t)$	s_2: 0.75/year	1×10^8
Storage Price/Mb	$\exp(-p_2 t)$	p_2: 0.70/year	3×10^{-8}
Software Size	$\exp(c_2 t)$	c_2: 0.31/year	2×10^3

2.6.1 Reconfigurable Hardware

Reconfigurable organizations are important to hardware system designers because they offer a way of achieving performance and efficiency by matching algorithmic constructs with the appropriate architectures [116]. Reconfiguration provides an even greater marketing edge as the lifetime of digital consumer products becomes shorter. A typical product life-cycle is now characterized by a shorter start-up period, an earlier and higher manufacturing peak and a faster end-of-lifetime decline. Reconfigurable products are a good match to these characteristics as they can be developed quickly and may be updated and/or upgraded during their lifetime in the field.

Field Programmable Gate Arrays

The traditional approach to reconfiguration, in FPGAs for example, has been to build separate areas of programmable logic gates and interconnection blocks from transistors and to manage the two resources more-or-less separately during the configuration process. Therefore, much of the work on these platforms has been directed towards answering the questions: "how much of each and in what form?" (see, for example, [117-119]). This is largely because reconfiguration always imposes a cost – it increases the area and power while reducing the performance of a system compared to purpose-built solutions[3].

When Field Programmable Gate Arrays (FPGAs) were first introduced they were primarily considered to be just another form of (mask programmed) gate array, albeit without the large start-up

[3] In [120], Hauck proposes the "law" of FPGAs vs. ASICs: *"for any FPGA-based application, there is an ASIC implementation of the system that is AT LEAST as fast, dense, power-efficient, and cheap in (very) high volumes as the FPGA based solution"*.

costs and lead times. Since then FPGAs have moved beyond the simple implementation of digital ("glue") logic and into general purpose computation. Although offering flexibility and the ability to optimize an architecture for a particular application, programmable logic tends to be inefficient at implementing certain types of operations, such as loop and branch control [121].

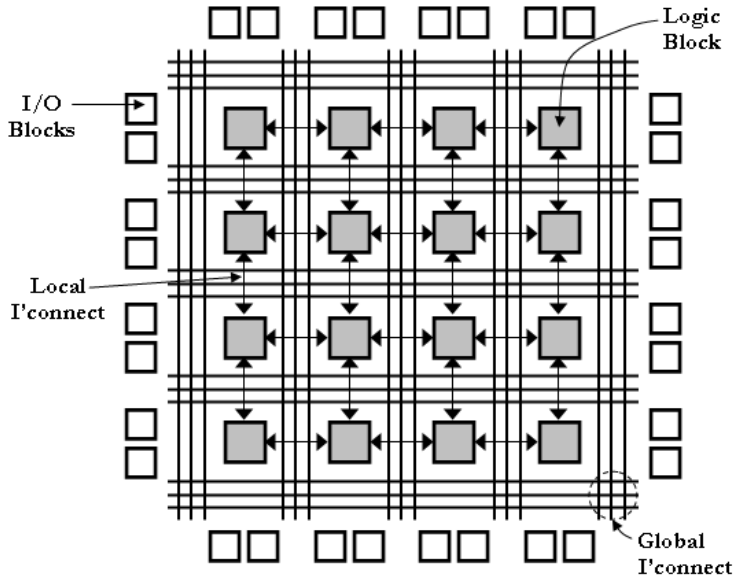


Figure 16. Basic FPGA architecture
(adapted from [124])

General-purpose FPGAs (e.g., Figure 16) are typically "fine grained" in that their logic blocks are mostly small, ranging from 2-input function generators to multiple 4-input Look-up Tables (LUTs). These organizations provide good support for irregular functions such as random logic but are widely considered be too generic in that they impose high routing and configuration overheads on regular elements such as adders, multipliers and shifters [122, 123]. This has been one reason for the move towards more "coarse-grained" (e.g., platform) structures, where the basic logic elements are multi-bit functional blocks (multipliers, ALUs etc.).

In many ways the "coarse-grained" versus "fine-grained" arguments for reconfigurable computing are reminiscent of the CISC vs. RISC debate [125]. This latter argument was largely about how a mapping from high-level language to machine code could be best achieved. Would it better to provide "solutions" i.e., complex features in the ISA that a compiler could use, or would a better

way be to provide "primitives" from which more complex instructions could be built? Many of the same arguments are now re-emerging, this time related to the hardware mapping process. Now the questions tend to be: what is the most appropriate grain size for general purpose systems [120]; and will high configuration and routing overheads [19] always favour coarse-grained architectures that provide operator-level configurable functional blocks and/or word-level data paths [122] over fine-grained organizations offering only logic primitives and interconnect from which these blocks can be built?

If the debate was to be based only on current FPGA organizations, then it might be said that the argument has already been fought and won: by coarse-grain style architectures [122]. A large number of platforms based on reconfigurable data path units of various granularities have been proposed along with a range of synthesis tools (e.g. [126], [127], [128], [129]) while, increasingly, commercial FPGA vendors are producing hybrid architectures incorporating both standard microprocessors and reconfigurable logic on the one chip. Examples include the Virtex-II Pro "platform FPGAs" from Xilinx® [130] and the "Excalibur" series from Altera®[131].

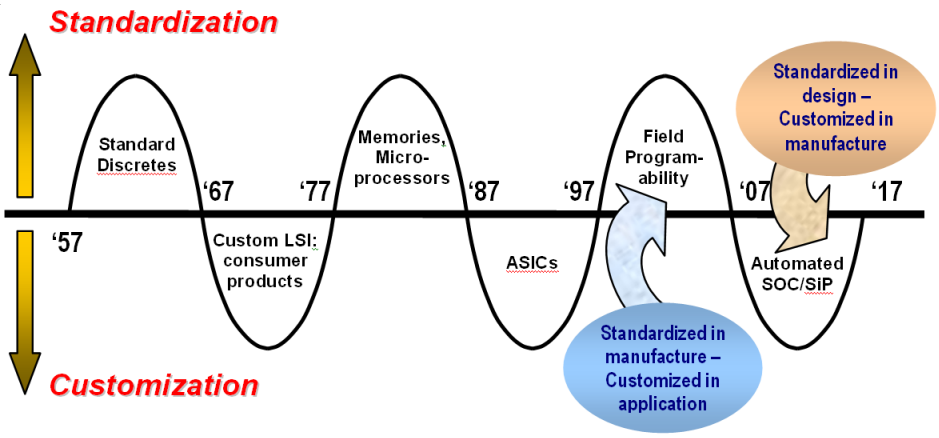


Figure 17. Cyclical semiconductor trends- Makimoto's Wave (adapted from [132]).

The so-called "Makimoto's wave" (Figure 17) [132] describes an empirical observation that the semiconductor business has tended to swing between standard and custom products in roughly 10-year cycles. Starting in 1957, the discrete transistor dominated semiconductor production. Then around 1967 production moved towards custom ICs for consumer products (televisions,

radios, clocks etc.) and another ten years later standard memories and microprocessors became dominant. Custom logic, Application Specific Integrated Circuits (ASICs), and systems-on-chip emerged in 1987 and by 1997 the FPGA had started to take market share away from ASICs and standard microprocessors.

According to this prediction the "next wave" (to commence as early as 2007) will see fine-grained FPGAs replaced by custom-programmable logic and configurable "platform" systems such as structured ASICs. This prediction appears to be based on a perception that platform systems will exhibit the same economies of scale that, for example, drove the original shift from custom LSI to standard microprocessors. Certainly, the same tension will exist between programmability (or *configurability* in this case) and architectural flexibility that allowed fine-grained FPGAs to capture at least part of the commodity microprocessor market.

For FPGAs using deep DSM technology, interconnect and wiring delays are already the dominant factor in the total delay figure [133], typically accounting for as much as 80% of the path delay [19]. As devices scale, the effect of distributed resistance and capacitance of both programmable interconnect switches and wiring will become worse. De Dinechin [119] has estimated that, if the general organization of FPGAs stays the same, their operating frequency will only increase $O(\lambda^{1/2})$ with reducing feature size (λ), leading to an widening gap between their performance and that of custom hardware. Indeed, ASIC designers face essentially the same problem and, as a result, future interconnect architectures are likely to include "fat" (i.e. un-scaled) global wires plus careful repeater insertion [91]. This observation has led some researchers to propose the idea of pipelining the interconnect as well as the logic [134], [135].

Studies of commercial FPGAs [117] have demonstrated that logic clusters are typically configured with more routing inputs than are strictly necessary and that these, in turn, have more configuration bits than necessary [136]. This contributes to the fact that 80–90% of the area of a typical FPGA is occupied by the interconnect switches and wires while most of the remaining area goes into configuration memory. The actual logic function occupies only a few percent of the area in a typical device [19]. A final observation about FPGA geometries is that their compo-

nents tend to be intrinsically large, with the area of a "typical" 4-input LUT (4-LUT) including its programmable interconnect and configuration memory estimated in [19] to be roughly $6x10^5\ \lambda^2$. The logic density of FPGAs is therefore likely to be limited even at very small feature sizes.

Structured ASICs

The emerging "Structured ASICs" [137] attempt to achieve faster and cheaper implementations using a predefined arrangement of late-stage mask-customizable logic and pre-diffused macros and IP blocks, albeit most often targeting a specific market segment (such as mobile and wireless applications). As such, they fit between standard cell ASICs with their long development times, high manufacturing costs but low per unit costs, and commodity FPGAs that have low design costs and short lead-times but also high per-unit costs and restricted design size, complexity, and performance.

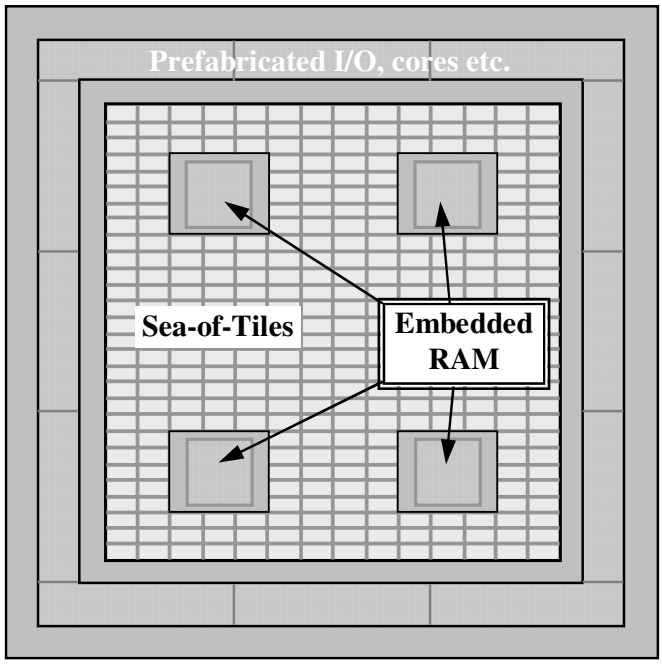


Figure 18. The Structured ASIC concept (from [137]).

Although a number of alternative architectures have been proposed [138], most tend to be built around an array of tiles or modules that, in a similar manner to the logic blocks in an FPGA, contain a small amount of generic logic often including one or more registers. Tiles are then

replicated across the surface of the chip (Figure 18). The tile logic may be augmented with a small amount of local RAM in some organizations.

A key difference from FPGAs is that Structured ASICs also typically contain additional prefabricated elements, such as configurable general-purpose I/O, microprocessor cores, gigabit transceivers and embedded (block) RAM. However, all of these elements are increasingly found in the newer so-called "Platform FPGA" systems [130, 139] so that, in reality, the Structured ASIC is simply one further member of a continuum of products that includes fabrics such as platform ASICs, array-based platforms, cell-based platforms, embedded FPGAs through to standard products embedded in FPGAs. All of these will offer features to support a particular market segment. As a result, future computer architectures may well be market application driven [140], with the characteristics of each market segment resulting in its own optimized micro-architecture. Finally, the resulting increase in design complexity is also likely to force a jump to the next level of design language abstraction, such as SystemC [141] or SystemVerilog [142].

Integrating Reconfigurable Hardware

A consensus appears to be emerging that most future systems will need to include some form of reconfigurable hardware [143]. Just as for previous decisions about FPGA granularity, there remain unresolved issues of built-in hardware functionality vs. mapping efficiency within a particular application domain. This is one issue that has motivated the research reported in this dissertation. If the area and performance overheads of reconfiguration can be reduced then fine-grained structures would offer a much more general solution and the "wave-after-next" could see a swing back to flexible, fine-grained computing platforms.

Reconfigurable hardware can be used in a number of ways: as a reconfigurable coprocessor unit that provides reconfigurable functional units either within a host processor or as an attached reconfigurable processor in a multiprocessor system; or as a loosely coupled external standalone processing unit [144]. One of the primary variations between these architectures is the degree of coupling (if any) with a host microprocessor. For example, the *OneChip* architecture [145] inte-

grates a Reconfigurable Functional Unit (RFU) into the pipeline of a superscalar RISC. The reconfigurable logic appears as a set of Programmable Function Units (PFU's) that operate in parallel with the standard processor. The Berkeley hybrid MIPS architecture, *Garp*, [146] includes a reconfigurable coprocessor that shares a single memory hierarchy with the standard processor, while the *Chimaera* system [147] integrates reconfigurable logic into the host processor itself with direct access to the host's register file.

2.6.2 Reliability and Defect Tolerance

The probability of failure for transistors in current CMOS manufacturing processes range from 10^{-9} to 10^{-7} [149] and it appears unlikely that currently available processes will support defect-free device structures at sub-100nm dimensions [150]. Thus any architecture built from large numbers of nanoscale components will necessarily contain a significant number of defects and an understanding of the role of these defects and how they affect yield will be important to future architectures. For example, it is shown in [149] that it is theoretically possible to produce working systems with defect rates as high as 10^{-5} to 10^{-4} if reconfiguration is used to bypass them. Existing static fault mapping techniques (such as are used in hard disk systems, for example) may represent a good starting point but ultimately, built-in self test (BIST) may be necessary to maintain system integrity in the presence of soft–errors and noise. There have been some initial studies into how to optimally configure BIST in an extremely large cellular arrays [151] but no general solutions have been developed as yet.

A closely related issue is the reliability of nanoelectronic technology over its operating lifetime. The reliability curve developed for ULSI logic in [148], based on the assumption that a single gate failure results in the failure of the entire system (Figure 19), indicates that at gate densities in the order of 10^7 almost half of systems can be expected to have failed within 10 years. Extrapolating these curves for transistor densities in the order of 10^9 (and assuming an average of five transistors per gate) implies that 90% of systems will have failed within about 1.3 years. To maintain the same reliability as a one million gate chip would require an error rate in the order of 10^{-16}/hour-gate, four orders of magnitude better than current technology. As a result, future

architectures will certainly need to be dynamically defect tolerant, with an ability to find defects as they develop and to reconfigure around them [152].

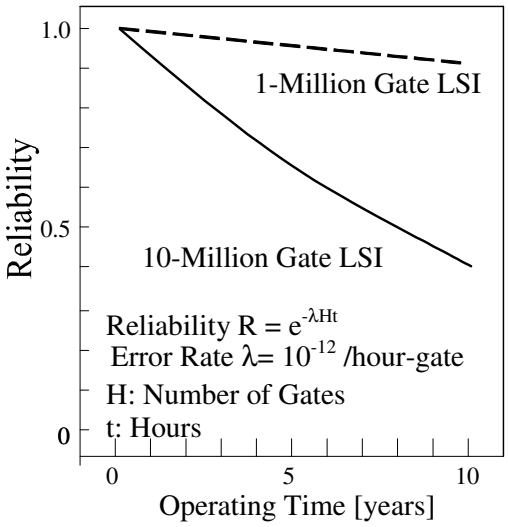


Figure 19. ULSI reliability curves
(adapted from [148])

Testing will also represent a major issue in nanoelectronic systems. Testing can already account for up to 70% of the total costs of production in 0.25µm CMOS ASICs [12], and this figure will become worse at higher densities. Techniques such as run-time self-test will therefore be increasingly important in the nanocomputer domain.

Defect Tolerant Architectures

As identified previously, defect tolerant architectures will be the only way to economically build computing systems with hundreds of billions of devices because any system using nanoscale components of any technology will contain significant numbers of defects. One example of an existing defect tolerant custom configurable system is the Teramac [153, 154]. The basic idea was to build a system out of cheap but imperfect components (FPGAs in this case), find the defects and configure the available good resources using software. The high routability of the Teramac is based on the availability of excessive interconnections due to its "fat-tree" routing configuration. However, it is possible that current methods for detecting defects such as those used in Teramac will not scale to devices with 10^{10} configuration bits [155]. Thus, novel parallel

defect mapping techniques will need to be developed, possibly built-in and coupled with self-configuration mechanisms of the type suggested by [156] or [157].

"Embryonics" [158] is a biologically inspired approach that aims to produce highly robust integrated circuits with self-repair and self-replication properties. In this case, the reconfiguration algorithm is performed on-chip in the form of an "artificial genome" containing the basic configuration of the cell. Its fault tolerance relies on fault detection and location via built-in self-test plus an ability to bypass faulty cells and to substitute spare cells in their place. However, the simplistic system employed (substituting entire columns of cells if just one cell is faulty) is not likely to scale successfully, due in part to the need to predetermine the number of standby logic cells that might be required in a typical implementation.

While the various demonstration systems have their limitations, they do illustrate that it is possible to build a computer system that contains defective components as long as there is sufficient communication bandwidth to support the discovery and use of working components plus the capacity to perform such a rearrangement of working components.

2.6.3 Issues in Design for Manufacture

It is fairly safe to predict that scaling in silicon will continue as long as the incremental performance gain per incremental cost incurred is greater than it would be for any alternative technology [159]. However, silicon manufacture will become more difficult with each successive reduction in feature size, and it is virtually impossible to forecast just when the “tipping point” will occur that makes any specific alternative more attractive.

Without a significant simplification in manufacturing processes [150], the rapid rise in costs caused by the use of finer design rules will become an increasing problem for semiconductor manufacturing plants. If current trends continue, the price of building an advanced fabrication facility will soon exceed $6 billion [160]. Similarly, non-recurring engineering (NRE) costs incurred in producing the first sample of a new chip have been increasing every year with mask costs alone representing in the order of $1.5 million of the NRE for a 90nm design [161]. The

overall result will be continuing economic pressure in the commodity device market which may ultimately tip the balance back towards fine-grained reconfigurable systems, continuing Makimoto's wave with a swing back towards standardized fine-grained array architectures.

Patterning Limits

A 90nm process requires over 35 masks and more than 800 individual processing steps [162]. These various steps already constrain most levels of the design process, and this will become much worse at smaller dimensions. For example, 90nm feature sizes are significantly smaller than the 193nm wavelength of conventional photolithography so various types of resolution enhancement techniques (RETs) must be used to ensure mask integrity. Techniques such as Off Axis Illumination (OAI), Optical and Process Correction (OPC) and Phase-shift Masks (PSM) [163] restrict the types of layouts that are suitable for manufacture [164] leading inexorably towards a preference for regular layouts. Layouts that might otherwise pass the design rules may lead to uncorrectable phase errors, or exhibit conflicts with adjacent areas. Also, techniques such as OAI are optimized for a single mask feature pitch so that pitches that are significantly different than this optimal value will see much less resolution enhancement.

Chemical-Mechanical Polishing (CMP) is another example where a process strongly interacts with layout [165]. CMP requires a uniform layout density in order to maintain a uniform thickness of the wafer [166]. Significant variations in layout will result in thickness variations across the wafer that will, in turn, influence intra-die variability in unpredictable ways. Dummy features might be added to each layer to help make irregular layouts more uniform for CMP [165] but this will become more difficult at smaller geometries.

Overall, the complexity of generating a reliable mask set that accurately describes the chip's drawn features will limit the ability to create arbitrary wiring patterns [167]. Some suggested solutions include restricted design rule sets with more awareness of OPC and PSM effects [164, 168] and process aware routing [167]. However, many of the patterning "work-arounds" devised to allow techniques such as OPC to work are rapidly becoming unsustainable and may cease to be

effective in mass production as early as the 65nm node. Ultimately, it will simply become easier and more cost-effective to restrict circuit designs to regular layouts only.

Complexity and Design Productivity

The 2005 ITRS identifies *complexity* in both silicon and systems and the resulting cost of design as a significant threat to the continuation of the roadmap. This is manifested as a *productivity gap* that threatens to prevent the exploitation of large transistor budgets, just at a time that they are becoming available [169]. Whereas the number of transistors per die has grown by some 58% per annum over the previous 20 years, designer productivity (CAD tools, etc) has only increased by about 21% p.a., leading to a productivity gap that is increasing by about 25% per year. Flynn and Hung [170] suggest that there is a tradeoff between design time and the system flexibility (i.e., the level of reuse or programmability) of the form:

Flexibility*Design Effort = constant.

By this they mean that the increased flexibility offered, for example, by reprogrammable or reconfigurable systems will reduce the design effort but at the cost of some other important optimizations, such as area–time–power tradeoffs, that might otherwise be important to a particular design.

Based primarily on an assumption that logic reuse can increase over time (from around 30% now to about 90% in 2020), the ITRS predicts that design effort for logic will remain approximately constant out to 2020. This implies that overall design productivity will need to increase some twenty-fold over the period 2005–2020 simply to maintain constant design effort. This will require combinations of approaches such as increases in design abstraction, levels of design and verification automation and reuse rates, along with reductions in reuse overheads.

2.7 Power–Area–Performance Scaling

It has been suggested [171] that the high-performance microprocessor has already passed through a number of identifiable scaling stages (Figure 20). Initially, Dennard's classical (1974) scaling theory meant that every feature size scaling of n resulted in $O(n^2)$ transistors that would perform O(n) times faster [172]. This was the N^3 era. Examples included the transition from the original

Intel 4004 through to the 386 architecture. In the N^2 era, the device speedup of n was maintained, but then only another n could be extracted from the n^2 transistors due to the effect of architectural features such as large on-chip caches for which miss rate might halve for each quadrupling of size. This era included the 486 through Pentium III/IV.

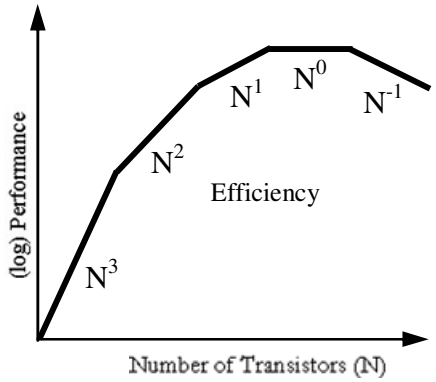


Figure 20. Jouppi's "Eras of Microprocessor Efficiency" (redrawn from [171])

In [171], Jouppi suggests that we may now be entering an N^1 era where improvements come entirely from the (linear) device frequency scaling and the n^2 improvement in transistor count might (at best) result in constant additional performance. For example, the additional complexity of very wide issue machines may be of little help to many applications. Taking this idea *ad extremum* (as in Figure 20) may result in an N^0 and even N^{-1} region, in which excessive increases in the size and complexity of various computing structures may result in a slowdown due to long global wires and increased memory access times.

The design of logic systems has traditionally been based on the assumption that overcoming the performance limitations imposed by fanout and interconnection parasitics was a straightforward matter of "tuning" the transistor width to length ratio to achieve the necessary delay. This simple assumption, built into techniques such as *Logical Effort* [173], is becoming less effective due to its effect on both dynamic and leakage power densities. In the Logical Effort model, the general form of the gate delay is related to the intrinsic delay (τ) as:

$$d = \tau\left(g\frac{W_{out}}{W_{in}} + p \right) \tag{2.16}$$

where g and p are constants that depend on the gate configuration and W_{in} , W_{out} are the input and load transistor widths. As both switching and subthreshold power are proportional to transistor width (i.e., $P_{static} \propto W_{in}$, $P_{dynamic} \propto W_{out}$), there is a clear delay/power tradeoff determined by the absolute and relative values of W_{out} and W_{in}.

The relationship between area, power and delay works at four primary levels:

- *device* – driven by technology choice and transistor sizing, as just mentioned;
- *circuit* – design style and layout;
- *micro-architecture* – encompassing implementation issues such as the Instruction–Set Architecture (ISA) as well as asynchronous vs. synchronous, or serial vs. parallel;
- *architectural* – including processor decisions: e.g., multi–core, super-scalar VLIW or strategies such as spatial computing.

Of these, the last two offer the greatest opportunity for power-performance tradeoffs, particularly at the instruction set and micro-architecture definition stages [174]. Even small design optimizations at this level can result in significant improvements to a processor's power-performance characteristics.

Using passive cooling and economical packaging, a maximum power density in the order of $100W/cm^2$ holds regardless of the actual density of the switching devices. Thus, as the device density increases, both the static power per device and the percentage of devices active at a given instant must fall. Further, a more realistic power target for portable devices might be more like $0.01W/cm^2$. As a result, interest must eventually switch to some other state variable that does not rely on the transfer of charge. There are no obvious contenders at present.

2.7.1 Low-Power Circuit Techniques

The two most important sources of power consumption in CMOS are dynamic switching power ($\propto aFCV^2$) and static (sub-threshold) leakage power ($\propto I_{OFF}V_{DD}$). In addition, there is a typically small contribution from short circuit current [175] plus a number of increasingly important tunneling effects through the gate oxide, for example, [176] and directly between the source and drain [177].

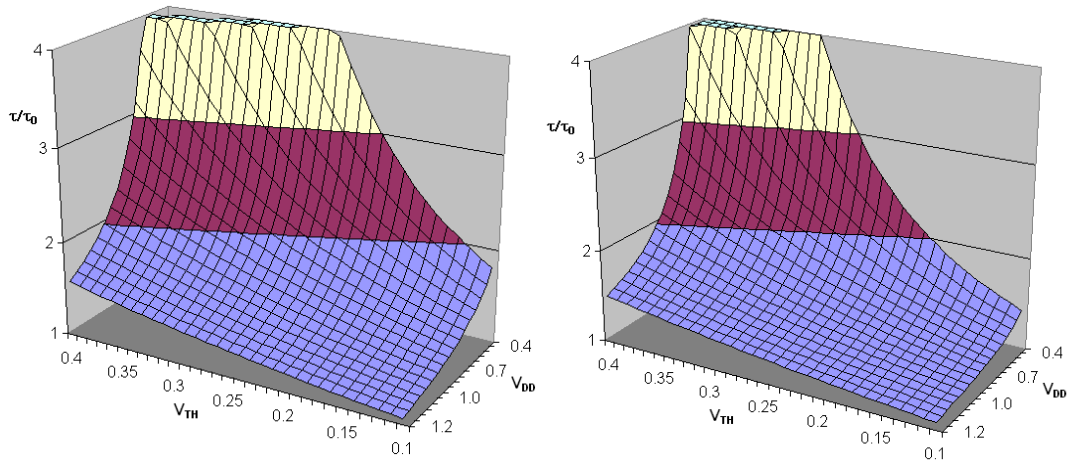


Figure 21. Delay scaling $\tau/\tau_0 \propto V_{DD}/(V_{DD}-V_{TH})^{\alpha}$ vs. V_{DD} and V_{TH}
(a) α=1.25; (b) α=1.05.

Given the square-law relationship between dynamic power and voltage in CMOS, and also that its performance is related to supply voltage and threshold (at constant load capacitance) by $F \propto (V_{DD}-V_{TH})^{\alpha}/V_{DD}$ [103], it will be decisions about supply (V_{DD}) and threshold voltage (V_{TH}) that will determine the static and dynamic power as well as its operating frequency. The general form of the performance-voltage tradeoff is illustrated in Figure 21, which plots the scaling of the delay term $\tau = \frac{1}{F} \propto V_{DD}/(V_{DD}-V_{TH})^{\alpha}$ over typical a range of V_{DD} and V_{TH} and with two values of α: 1.25 and 1.05. As supply voltage falls, thereby saving dynamic power, the impact of threshold voltage on delay becomes greater so that it will become increasingly difficult to find a *fixed* V_{TH} that optimizes both frequency and static power. In an extreme case, as V_{DD} is reduced to a very

low value (e.g. some small multiple of the thermal voltage, kT/q), only sub-threshold operation will be possible.

It can be seen from a comparison of the two curves in Figure 21, that the effect of reducing α is to reduce the dependence of delay on supply and threshold voltage, although across the expected range for α between the present and 2020 (approximately $1.25 \geq \alpha \geq 1.05$), its impact will be relatively small. For example, when $\alpha = 1.05$ the point at which delay is double the original value $(\tau / \tau_0 = 2)$ occurs at higher values of supply and/or threshold—$\Delta V_{DD} \approx 100$ mV and ΔV_{TH} up to 60 mV.

An obvious solution to these conflicting design requirements for subthreshold power and delay is to allow them to be optimized separately, thereby reducing the need to carefully manage the threshold voltage (and subthreshold slope) in order to achieve a particular power-delay design point. This is the basic idea behind the many examples of variable threshold devices that have been previously reported in both HEMT [178] and CMOS technology [179], as well as the circuit techniques that exploit variable threshold voltages (e.g. [180, 181]).

One of the simplest low-power techniques is to turn off parts of the circuit when they are not in use. For example, the MTCMOS technique (Figure 22a) uses a pair of high-V_{TH} sleep transistors to disable the power and ground lines of the low-V_{TH} (and therefore high-speed) operational gates. In this way, the overall subthreshold leakage of the circuit is governed by the sleep transistors. There are a number of potential problems with this approach, quite apart from the added manufacturing complexity imposed by the dual threshold voltages. Firstly, the internal nodes float during power-down, requiring special circuit design techniques for the state elements [184]. Secondly, the correct sizing of the sleep transistors can be difficult. They need to be large enough not to interfere with the performance of the logic circuits, but without adding significant area or incurring excessive energy losses when switching between modes. Finally, supply scaling is limited by the switching threshold of the high-V_{TH} (typically to no less than 0.6V). Super Cut-Off CMOS

(SCCMOS) [183] attempts to address this last issue by overdriving the gate of the (single) sleep transistor (Figure 22b).

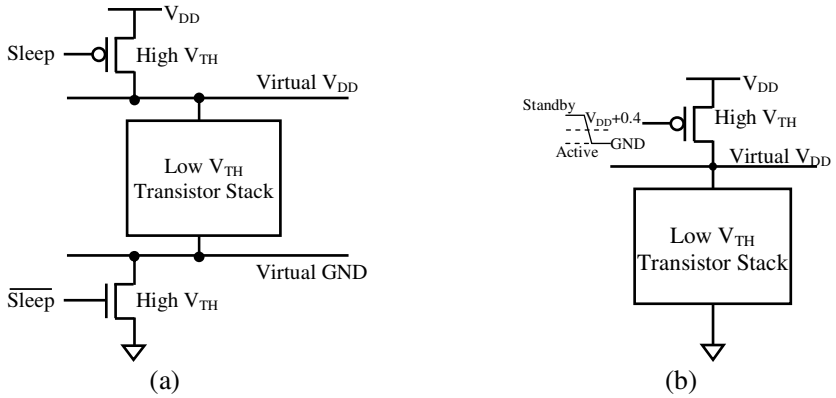


Figure 22. Alternative power-down circuits using high V_{TH} sleep-mode transistors (a) MTCMOS (from [182]); (b) SCCMOS (from [183]).

Various variable supply schemes have been proposed based on either permanently lowering the supply voltage in non critical parts of the circuit [185] or on dynamically setting a supply voltage (and clock frequency) to deliver a level of performance appropriate to the workload [186, 187]. In the former case, the objective is to balance the delays though the various critical paths to achieve the lowest power for a given performance. It therefore relies on careful circuit simulation using realistic logic vectors. Although the dynamic supply scheme is undoubtedly an effective low-power design technique, particularly for real-time embedded systems and multimedia workloads, it imposes the overhead of determining the workload behavior at each instance during actual operation [188].

An alternative to controlling the supply (using *fixed* threshold voltages) is to shift the threshold voltage between modes, either with or without supply variations. In Variable Threshold (VT) CMOS (Figure 23) [190], the threshold voltage is controlled by altering the well bias of the devices in a triple-well CMOS process (Figure 23). During the active state, a low threshold voltage is achieved by setting the well bias to V_{DD} + 0.5V for the p-well and -0.5V for the n-well. In standby mode, the source-body junction is strongly reverse biased to increase the threshold voltage and to reduce leakage current.

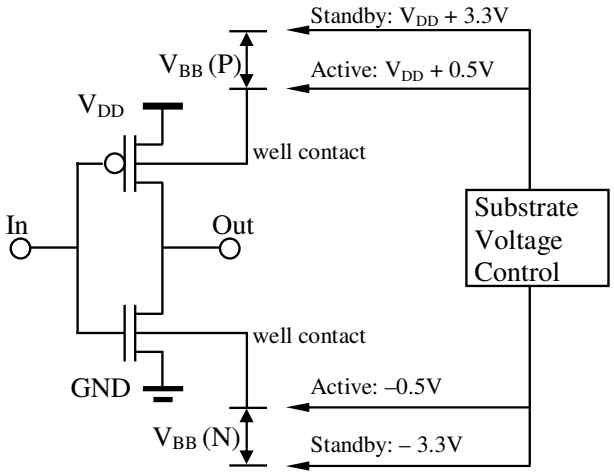


Figure 23. Variable threshold CMOS (VTCMOS) (adapted from [189]).

While the dual threshold characteristics of VTCMOS are effective at reducing leakage, there are still a number of problems with the technique. The threshold voltage $V_{TH} \propto \sqrt{V_{BS}}$ (V_{BS} = body-source voltage) and this body-effect appears to be reducing at successive generations. As a result increasing V_{BS} levels would have to be applied to change the threshold voltage, ultimately leading to excessive source-body and drain-body tunneling. In addition, the scheme relies on a complex triple-well process and the large capacitance of the wells prevents the threshold from being changed dynamically.

VTCMOS has also been investigated using fully depleted SOI devices [190]. In this case, the body effect $\Delta V_{TH} / \Delta V_{SUB}$ is replaced by a term proportional to the ratio of the gate-channel and the channel-substrate capacitances, so that:

$$\frac{\Delta V_{TH}}{\Delta V_{SUB}} = \frac{t_{OXF}}{t_{SI}/3 + t_{BOX}} \tag{2.17}$$

where t_{OXF}, t_{SI} and t_{BOX} represent the thickness of the front gate oxide, the channel and the body oxide, respectively [190]. It can be seen from (2.17) that in order to achieve a useful effect the dimensions of both the channel and the body oxide would have to be small compared to the front oxide. For example, the thin device described in [190] (t_{OXF} = 1.2nm, $t_{SI} = t_{BOX}$ = 10nm; $N_a \approx$

10^{15}cm^{-3}) exhibits $\frac{\Delta V_{TH}}{\Delta V_{BG}} \approx 0.07$ (i.e., ~70 mV/V) for -1.0V < V_{BG} < 0V (replacing V_{SUB} with the back-gate voltage, V_{BG}). The less aggressive SOIAS (Active Substrate) device in [180] (t_{OXF} = 9nm, t_{SI} = 40nm; t_{BOX} = 100nm; $N_a \approx 10^{17}$cm^{-3}) exhibits ~80mV/V over a similar range of V_{BG}. By contrast, in the SOI device in [181] ($t_{OXF} = t_{BOX}$ = 7nm, t_{SI} = 50nm; $N_a \approx 10^{15}$cm^{-3})), V_{TH} can be shifted by ~0.9V over -2.0V < V_{BG} < 0V and $\frac{\Delta V_{TH}}{\Delta V_{BG}} \approx 0.75$ (i.e., 750 mV/V) in the range -1.0V < V_{BG} < -0.6V, about 2.5 times the value of ~300 mV/V predicted by (2.17).

2.7.2 Adiabatic Systems

Adiabatic switching techniques are different to other low-power methodologies in that they attempt to recycle signal energies instead of allowing them to be dissipated as heat [191]. These "reversible" techniques attempt to reduce the dissipated energy by slowing the speed with which the charge is drained from the gate [192]. The objective is to reduce the switching energy of a gate from CV^2 to $2(RC/T_S)CV^2$ by making the switching period T_S large compared to its intrinsic time constant (RC). Under some circumstances, the approach can be used to reduce the power dissipation of digital systems.

However, as pointed out in [193], a basic assumption of reversible computing (as made in [194], for example) is that the system is completely isolated from the environment so that the energy of the system is totally conserved and the system can be described by reversible equations. No physical system can exist in complete isolation and adiabatic circuits make no attempt to do so. Thermal noise and errors due to thermal excitations are equivalent to information erasure and this, combined with the additional energy required by logic measurement and control, ensures that a "reversible" computation will always dissipate energy. As Frensley argues in [195], gain is a fundamental requirement of any computational system (not just electronic) and this implies the presence of an external energy source. As a result, systems exhibiting gain must be intrinsically dissipative.

Experimental results have demonstrated that the technique might be useful for high capacitance circuits where the losses do not account for a significant component of the total energy. For example, in [196] it was shown via simulation that a fully adiabatic approach might be useful in driving capacitances in excess of 0.1pF, but in those simulations this "low-power" technique was two orders of magnitude larger than conventional CMOS at typical local interconnect capacitance values. Regardless of the claims made by various authors (e.g. [192, 197-200]) it appears that the limitations of "reversible" or adiabatic systems, along with their additional complexity, will prevent them from making more than a theoretical contribution to future high-density charge-based systems.

2.7.3 Architectural Level Power/Energy Scaling Models

As mentioned in the previous section, it is decisions made at higher levels of design abstraction that will have the most impact on power-performance tradeoffs. For this reason, a number of architecture-level metrics have been developed that model the impact of architectural modifications on a range of general cost-functions: performance (IPC), power, clocking rate, etc.

Energy-efficient Scaling Models

In [201], Zyuban and Kogge introduce the idea of an *energy-efficient* circuit as one that delivers the highest performance amongst a group—an *energy-efficient family*—that dissipate the same power. Each circuit in the family is assumed to be optimal in terms of an energy-delay (ED) criterion given by ED^n, $n>0$ [4].

This is further developed in [202] to show that the ED^n metric characterizes any optimal tradeoff between energy and delay for an arbitrary computation. Selecting the exponent *n* represents a simple mechanism for optimizing in favour of either low power/energy or high performance. The metric ED^2 is considered by many to be ideal [203] as, under the assumption that $F \propto V$, it is largely independent of supply voltage. One example of its application is in the *PowerTimer* tool [204], which uses a parameterized set of energy functions in conjunction with a cycle-accurate

[4] Here, D is effectively equivalent to the generalized time parameter (T) used later in this thesis.

micro-architectural simulator to assess worse-case power swings resulting from typical workload sequences in terms of a $(CPI)^3P$ metric, which is equivalent to ED^2. However, with increasingly ballistic device operation, the $F \propto V$ assumption will become less valid and this will, in turn, tend to undermine the validity of the ED^2 metric. Units such as MIPS/watt are routinely used in low-power and embedded processor systems, while $(MIPS)^2$/watt and even $(MIPS)^3$/watt are used to measure systems for which performance is paramount [205].

Because it is often difficult to untangle the relative impacts of circuit and architectural-level tradeoffs, Zyuban has proposed two design metrics that include contributions from both [206, 207]. The first, Hardware Intensity (HI) measures the effect of design modifications at the architectural and micro-architectural levels on architectural performance, dynamic instruction count, average energy dissipated per executed instruction and the maximum clocking rate of the processor at a *fixed* supply voltage (V). It is intended to support the evaluation of power-performance tradeoffs at the architectural level before a design is fixed. HI is defined as a parameter (η) in a cost function of the form [206]:

$$F_C = \left(E / E_0\right)\left(D / D_0\right)^{\eta} \qquad 0 \le \eta \le +\infty \tag{2.18}$$

where D is the critical path delay, E is the average energy dissipated per cycle and D_0 and E_0 are the lower bounds on delay and energy that might be achieved for a fixed supply voltage. In effect, Hardware Intensity describes the relative increase in power and/or energy incurred when local logic restructuring and tuning is used to reduce the critical path delay at a fixed power supply voltage for an energy–efficient design [206]. It has the property that $\eta = -\left.\frac{D\partial E}{E\partial D}\right|_{\text{fixed V}}$.

The second metric, "Voltage Intensity", $\theta = -\left.\frac{D\partial E}{E\partial D}\right|_{\text{fixed circuit}}$ describes the energy–delay sensitivity of a fixed circuit to voltage scaling. Thus at each value of supply voltage there will be an optimal η and θ that can be determined from the ratio of the energy cost of a small perturbation around V

$(\partial E/\partial V)$ to the change in performance $(\partial D/\partial V)$. These cost functions would most easily be determined by circuit simulations around the intended operating conditions.

To apply these intensity metrics at an architectural level, [206] introduces the discrete architectural complexity (ξ) such that both the average power and the performance of a processor design can be expressed as functions of ξ, the supply voltage (V) and η. The impact of a proposed architectural feature may be evaluated by determining the values of V and η for which power remains constant. Under the assumption that for each architectural change the processor pipeline is re-tuned to give an optimal balance between hardware intensity and supply ($\eta = \theta$), the conditions under which a processor can increase performance while keeping power constant is given by:

$$\eta \frac{\Delta I}{I} - (\eta + 1)\frac{\Delta N}{N} > \sum_i \frac{\Delta E}{E} + \sum_i \eta_i w_i \frac{\Delta D}{D}. \tag{2.19}$$

Here, $(\Delta I / I)$ and $(\Delta N / N)$ are the relative changes in performance and dynamic instruction count resulting from an architectural (or micro-architectural) modification applied to pipeline stage *i*, evaluated for a fixed hardware intensity and power supply voltage. These estimates would typically be derived from an architectural-level or timing simulator. The term w_i is the energy weight applied to a particular pipeline stage and is typically available as part of the power budgeting early in the processor definition phase.

However, there are a few drawbacks in the formulation of both hardware and voltage intensity [208]. The derivation of the metrics assumes that the energy and delay of individual logic stages are independent of one another, which is only true in the case where transistor sizes are fixed (i.e., fixed input and load). Where this is not the case, and in particular where the adjustment of transistor sizing in one block impacts the load seen at an adjacent block, these analytic solutions tend to break down. Further, (2.19) is valid for small architectural changes that do not impact too greatly on the overall architectural complexity and where the resulting changes in energy ΔE and delay ΔD are similarly small.

Parallel Scaling Models

Since the early work by Chandrakasan *et al.* [209], it has been shown many times that parallel organizations may be used to exchange area for a reduction in supply voltage and therefore in power and/or energy. For example, in the simple duplicated/pipelined data path circuits studied in [209], power reductions of up to a factor of five were obtained compared to a single data path baseline. This general point can be illustrated with a simple example based on an arbitrary computational function of unit size, and assuming that this can be multiplied up as many times as necessary to achieve a particular performance point. Here the area is modeled as $A_{i+1} = A_i(N+O_a)$, where A_{i+1} and A_i are successive area terms and N is the number of replicated paths (e.g., 2). The area overhead, O_a, results directly from issues such as more and longer routing paths as well as additional hardware required to distribute the operands and to recombine the results back to a single data stream and is assumed to be a simple linear function of N. In [209], this overhead was significant (~70%) because of routing inefficiencies in the standard cell technique used to synthesize the data paths. This multiplicative series may be approximated by a simple power function of N, such that $A_N \approx N^{\alpha}$.

In a similar way, each area increase will result in an overall increase in the critical path delay due to the additional hardware and routing path lengths so that $T_{i+1} = T_i(1+O_T)$, where O_T is the time overhead at each stage, and thus $T_N \approx N^{\tau}$. However, at each stage the overall throughput will be $N^{\tau}/N = N^{\tau-1}$, which reduces the effective delay at the cost of a small increase in latency (Figure 24). This gain may be traded for a reduction in frequency and therefore dynamic power $P_D \propto FCV^2$. Figure 25 shows this area-performance tradeoff on a log scale for a fixed area overhead and for the three delay overhead figures of Table 4. It can be seen that these follow a simple power-law function of the form $T \propto A^{-1/\sigma}$, with the values of σ shown on each curve. As the overheads increase, so does the value of σ, implying that there is less performance "return" on each successive area scaling. This is the basis of the generalized Area-Power-Performance model developed in Chapter 4.

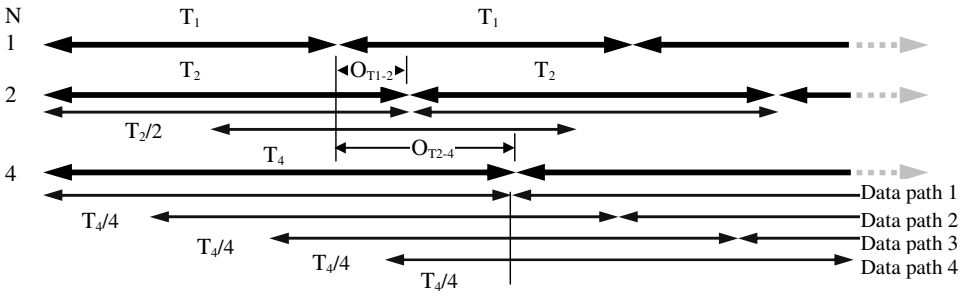


Figure 24. Overall performance speedup using parallel data paths.

Table 4 Example Area and Time Scaling vs. Delay Overhead.

N	$A=N^{1.07}$	$O_T=0.1$	$O_T=0.3$	$O_T=0.5$
1	1.0	1.0	1.0	1.0
2	2.1	1.10	1.30	1.50
4	4.4	1.21	1.69	2.25
8	9.3	1.33	2.20	3.38
16	19.4	1.46	2.86	5.06
32	40.8	1.61	3.71	7.59
64	85.8	1.77	4.83	11.39
128	180.1	1.95	6.28	17.09

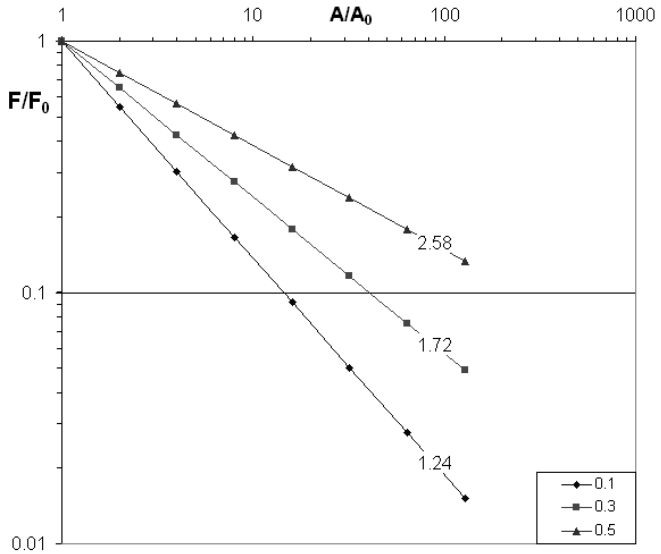


Figure 25. An area-frequency scaling example showing the area—performance tradeoff of the form $T \propto A^{-1/\sigma}$ for a fixed area overhead and for the three delay overhead assumptions of Table 4. The numbers on each curve give the corresponding value of σ.

Of course one major tradeoff here is dynamic vs. static power that both depend on the relative values of supply and threshold voltage. If V_{TH} is held constant, or reduces with V_{DD} to maintain performance (as predicted in the ITRS), then subthreshold power will increase to a point where it

becomes dominant (Figure 26). To maintain P_{SUB} with increasing device numbers, V_{TH} *must* increase, impacting on performance. For this reason, it is argued in [210] that this technique will become less useful in the future as decreasing V_{DD}/V_{TH} ratios will increase the performance penalty for a given reduction in supply voltage. That model suggests that the cross-over point, at which the use of parallelism could become ineffective, might occur as early as the 2012 ITRS node.

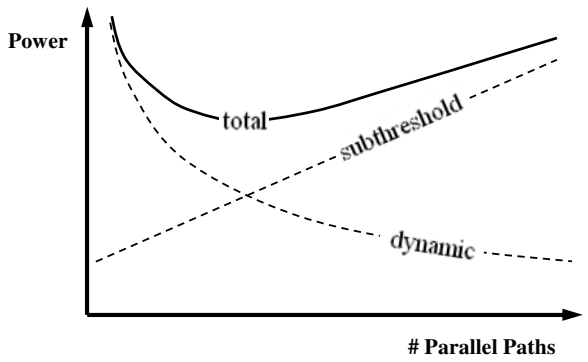


Figure 26. Generalized total power trajectory with parallel data paths assuming constant or reducing V_{TH}, causing increasing I_{OFF}.

On the other hand, a more recent analysis in [211] found that both parallelism and pipelining could still result in significant energy reductions while maintaining the same throughput. Because of the assumption that output load increases in proportion to parallelism, improvements in both energy and throughput tended to diminish at higher degrees of parallelism in that study. Nevertheless, they were still able to demonstrate energy reductions in excess of 50% over a wide range of organizations. Considering the *degree of parallelism* in [211] to be equivalent to area (A), those experiments (based on a 0.13µm, 1.2V technology) resulted in scaling functions relating area to energy (E), time (T) and power (P) of the form $ET^2 = PT^3 \propto A^{-1.58}$ for parallel and $\propto A^{-1.75}$ for pipelined organizations.

While all of these examples clearly demonstrate the area-power tradeoffs available by exploiting parallelism, none of them represents a formal framework which might guide how this could be achieved. As well as allowing circuits to be optimized such that they operate at the most efficient energy-delay point, the *Hardware* and *Voltage Intensity* methods of Zyuban and Strenski [207]

provide a mechanism for evaluating the impact of design modifications at the architectural and micro-architectural levels on architectural performance, dynamic instruction count, average energy dissipated per executed instruction and the maximum clocking rate of the processor at a fixed supply voltage [206, 207]. Although it can be argued that all such design modifications will have an area impact, the method does not explicitly include an area cost component.

2.8 Emerging Computer Architecture

Modern microprocessor architectures can be said to be *heterogeneous* at every level of their design hierarchy [212] in that they employ a wide variety of devices, circuits, and subsystems in their design. While this suits current fabrication techniques, it is likely to be beyond the capability of nano-fabrication. As a result, the development of nanocomputer architectures has tended to focus on simple homogenous hardware structures that avoid introducing additional hardware complexity and that can be configured post-manufacture. For example, [213] describes the greatest challenge in nanoelectronics as the development of logic designs and computer architectures necessary to link small, sensitive devices together to perform useful calculations efficiently. While an ultimate vision might be to construct a useful "Avogadro computer" [214] (i.e., one that efficiently exploits some 10^{23} switches) in more realistic terms the ITRS predicts that as early as 2012 even a standard CMOS chip may comprise in excess of 10^{10} transistors [9]. The primary question is still how to most efficiently exploit this number of switching devices.

2.8.1 Parallelism

To date, architecture research has responded to the opportunities and challenges offered by device scaling in two ways. The first approach simply increases existing machine resources – more or larger caches; more on-chip processors often including local DRAM [215], direct multi-threading support (i.e. exploiting parallelism between concurrently running tasks rather than within a single algorithm) and other similar techniques. While being effective for some applications, these can quickly run into some or all of the physical limitations outlined previously. In particular, the wire-length problems can result in unacceptable memory and I/O latency, although the 50 to

100,000-gate hierarchical design blocks suggested in [91] are certainly large enough to contain a small RISC processor or other quite sophisticated processing elements. Durbeck and Macias [214] put it this way: "... *there is no clear way for CPU/memory architectures to tap into the extremely high switch counts … available with atomic-scale manufacture, because there is no clear way to massively scale up the (CPU) architecture. … there is no such thing as "more" Pentium®. There is such a thing as more Pentiums®, however.*"

An alternative approach uses modular and hierarchical architectures to improve the performance of traditional single-thread architectures [216]. Table 5 (reproduced from [217]) compares the three main classes of parallel architectures in terms of characteristics applicable to the nanocomputer domain. They conclude that highly regular, locally connected, peripherally interfaced, data-parallel architectures offer a good match to the characteristics of nanoelectronic devices. However, it is worth noting that data-parallel architectures represent only a small portion of the interesting problems in computer architecture and are a poor match for most general purpose computing problems. In [218], Bilardi and Preparata come to a similar but stronger conclusion. Assuming an extreme view of the physical speed of light and "boundedness" limits, [218] contends that an "asymptotically" scalable parallel machine will be a nearest-neighbour mesh interconnection of small machines, each approximately the size of the synchronous region (i.e., within which the impact of interconnect delay is small; cf. *drive distance* [92]). In fact, as identified in Section 2.6, future computer architectures may yet turn out to be largely market-application driven, with the characteristics of each market segment resulting in its own preferred parallel micro-architecture.

Table 5 A Comparison of three parallel architecture classes (from [217])

Characteristic	Class		
	Data	**Function**	**Neural**
Degree of parallelism	High	Low	High
Processor Complexity	Low	High	Medium
Interconnect Density	Low	High	High
Amount of Interfacing	Low	High	Low
Extensibility	High	Low	Low

Locally Connected Machines

A common example of regular, locally connected, data-parallel architectures is the Single Instruction Multiple Data machine. SIMD machines exploit the inherent data parallelism in many algorithms - especially those targeting signal and image processing [219]. Fountain *et al.* [217] identify the characteristics that may make the SIMD topology suited to future computer architecture as:

- a regular and repetitive structure;
- *local* connections between all system elements;
- all external connections made at the array edge;
- the existence of feasible strategies for fault tolerance.

However, SIMD architecture still suffers from two major problems - global instruction issue as well as global control and clock signals. Global clocking is required by SIMD machines not only to drive each individual (synchronous) element but also to manage inter-element synchronization.

It is clear from the analysis of [217] that the interconnection costs of SIMD in the nano-domain are very high, with the majority of the die area in their experiments being taken up by control signal distribution. Numerous asynchronous design techniques (e.g. [220]) have been proposed to overcome the need for a global clock in SIMD machines. While it is still unclear whether, in practice, these asynchronous techniques actually offer improved performance, they are at least as good as the conventional synchronous approach and may offer the only means to overcome the constraints of global communication.

The same considerations appear to constrain other multi-processor architectures such as MIMD. In [221], a series of experiments were performed on various MIMD architectures and it was concluded that inter-processor communications will be limited by the availability of wider metal tracks on upper layers (the so-called "fat" wiring layers). The tradeoff here is between track resistance (and therefore delay) and interconnection density. It was also noted that complex

computational structures such as carry look-ahead begin to lose their advantages over simpler and smaller structures once wiring delays are factored in.

Propagated Instruction Processor

The Propagated Instruction Processor (PIP) was proposed by Fountain [222] as a way of avoiding the interconnection problem in SIMD arising from its global instruction flow. In the PIP architecture, instructions are pipelined in a horizontal direction such that the single-bit functional units can operate simultaneously on multiple algorithms. The technique shares many of the characteristics of SIMD, pipelined processors and systolic arrays. One of the primary advantages of the architecture is its completely local interconnection scheme that results in high performance on selected applications.

However, the architecture is still basically SIMD and thus will work best with algorithms from which significant data parallelism can be extracted such as Fountain's examples of point-wise 1-bit AND of two images, an 8-bit local median filter, 32-bit point-wise floating point division and an 8-bit global matrix multiplication [222]. In addition, the fault tolerance of the PIP may ultimately depend of an ability to bypass faulty processors without upsetting the timing relationship between propagating instructions, something that has not been reported to date.

Merged Processor/Memory Systems - IRAM and RAW

The structure and performance of memory chips are becoming a liability to computer architecture. There are two basic problems. The first is the divergence in the relative speed of processor and DRAM, the so-called "memory wall" outlined previously. Secondly, while DRAM size grows by an average of 60% per year, its fundamentally limited access bandwidth is becoming increasingly difficult to circumvent. This observation has led to the development of a number of merged memory/processor architectures. Two notable examples of this approach are the Intelligent RAM (IRAM) system [223], and the Reconfigurable Architecture Workstation (RAW) [224].

The IRAM system merges processing and memory onto a single chip. The objective is to lower memory latency, increase memory bandwidth, and at the same time improve energy efficiency.

The IRAM scheme revives the vector–style architecture originally found in supercomputers and implements it by merging at least 16MB of DRAM, a 64-bit two-way superscalar processor core with caches, variable width vector units, and a high-performance memory switch onto a single chip.

The RAW microprocessor chip comprises a set of replicated tiles, each tile containing a simple RISC like processor, a small amount of configurable logic, and a portion of memory for instructions and data. Each tile has an associated programmable switch which connects the tiles in a wide-channel point-to-point interconnect. The compiler statically schedules multiple streams of computations, with one program counter per tile. The interconnect provides register-to-register communication with very low latency and can also be statically scheduled. The compiler is thus able to schedule instruction-level parallelism across the tiles and exploit the large number of registers and memory ports.

Stream Processors

The recent emergence of Stream Processors [225] represents one way of accessing parallelism at the instruction-level (ILP), sub-word and data levels. Stream processors combine an interconnection bandwidth hierarchy made up of local register files, a global stream register file, and memory that keeps most data movements local with clusters comprising potentially thousands of functional units [226]. As a result, they are able to reveal and exploit instruction-level and sub-word parallelism within a cluster and data parallelism across clusters [227]. The application has to be organized into *streams* (a group of elements of the same type) and *kernels* (a sequence of operations on streams). Thus the technique is especially suited to applications such as media streaming where abundant parallelism is available along with minimal global communication and storage.

The studies in [226] indicate that stream processors may be scaled by increasing both the number of functional units per cluster (“intra-cluster” scaling) and the overall number of clusters (“inter-cluster” scaling). Under the particular modeling assumptions made in those studies, the impact of intra-cluster interconnect switching overheads placed the most area and energy efficient configu-

ration at about five functional units per cluster while the overall number of clusters could be scaled up to 128 with a small impact on area and energy. Ultimately, it is the availability of parallelism in the application that effectively limits any increase in the number of clusters [227]. For example, the wireless base-station code analysed in [227] exhibited data parallelism values between 32 and 256 with the majority of algorithms at 32. As a result, the observed performance increases almost linearly with increasing clusters to 32 then sub-linearly to 64 after which there was almost no additional performance improvement.

2.8.2 Spatial Architectures

Because they are used to working in a regime where devices are a scarce resource, computer architects have a strong tradition of trading off component area against performance [170]. However, as switching devices continue to shrink, micro-architecture is entering a resource-rich era, prompting an examination of alternative computing models. Spatial architecture [228-231], in which operations (and their operators) are connected in space rather than time, is one such model. Feldman and Shapiro [232] originally described an abstract Spatial Machine as a finite set of finite processors moving and exchanging messages in 3D unbounded empty space. Implementations of spatial organizations “unwind” a computation into hardware, exploiting the availability of resources to expose the full parallelism available in a task thus completing it in time proportional to its longest path rather than in a time proportional to the number of operations. The basic argument is twofold:

1. when die space is no longer at a premium, heavy multiplexing of the processor data paths is not necessary to keep the problem within the available silicon area [233] and:

2. although spatial designs may be much larger than the minimum sized temporal design, they automatically achieve high computational performance [229] without the need for the hardware to "best-guess" the temporal control flow of its software (e.g. with features such as pre-fetch, register renaming , branch prediction etc.).

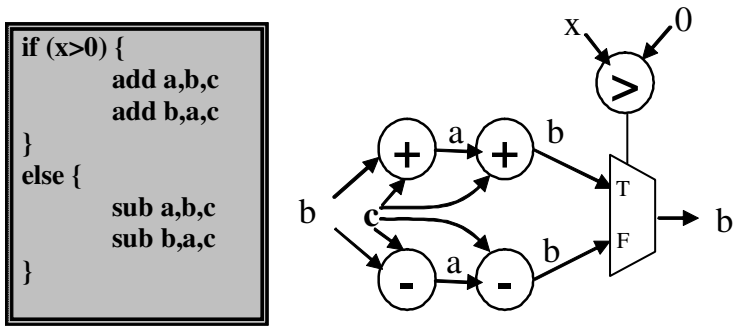


Figure 27. A processing graph fragment.

On the face of it, spatial computing approaches would seem to be a very wasteful, brute-force approach. By fully expanding operators (and their operations) into hardware, the overall size of a spatial computational fabric may be many times that of a conventional organization. One default starting point is to create all possible computation paths in the "program" and then simply select the single correct result at the end, as is illustrated by the graph fragment in Figure 27. Computational functions are created as and when required, along with any intermediate variables. The procedural flow of the software is effectively "wired-in" so that all static data transfer operations are embedded in the fabric of the computer and represent wires rather than logic gates. Of course, this will have important performance consequences as wires scale to nanometer dimensions.

Table 6 Dynamic Instruction frequency of MIPS-R3000 (based on [234])

Instruction Class	Dynamic Frequency
Load/store	36.94 %
Move register	20.27 %
All arithmetic/logic operations	13.03 %
Unconditional jump	10.36 %
NOP	7.26 %
Branch on status	6.69%
Shift operations	3.10 %
Jump relative	1.96 %
Other (e.g. mult, div)	0.21 %

As an illustration, Table 6 summarizes the average dynamic instruction frequencies for an R3000 processor, derived from its run-time performance on a number of integer benchmarks [234] and

sorted into approximate functional categories. It can be seen that almost 75% of instructions involve some form of control-flow, procedural or data-flow (i.e. register transfer) operations that are likely to end up wired into the processing fabric of a spatial architecture i.e., resulting in connections rather than gates. Indeed, as few as 15% of the instructions in these benchmarks involve functions that would appear as logic structures in a resultant spatial architecture. This observation clearly illustrates the way that spatial computing will change the balance between computation and interconnections in future architectures.

One difficulty here is that it is likely that computing tasks will continue to be described by software and nanocomputers will inherit a vast quantity of legacy software that cannot be ignored. Software is largely sequential in nature, dominated by control dependencies and tends to rely on dynamic data structures that do not map well to spatial architectures [235]. There is already a significant body of work relating to the transformation of high-level language to hardware [236], in terms of C [237] or C++ [238, 239]) or other variants of C [240-242] although the primary focus here tends to be on design productivity improvements rather than on optimization.

To date, the most complete series of analyses of a nanoscale spatial computing fabric is that by DeHon and his co-workers [243, 244]. Their demonstration platform comprises a large array of nanotube and/or nanowires crossbar structures that form the logical equivalent of the AND and OR planes within a conventional Programmable Logic Array (PLA). The architecture comprises four subsystems: (1) an array of crossed nanowire diodes used as a programmable OR-plane; (2) an inverting sub-array of crossed nanowire transistors; (3) a similar buffering sub-array, both of which are used to regenerate signals plus (4) an input/output decoder (Figure 28).

It needs to be remembered here that nanowire FETs will be constrained by the same I_{ON}, I_{OFF} and subthreshold slope considerations as conventional MOS devices. As a result, deHon's proposal employs dynamic logic as a way to maintain performance while minimizing overall system power. However, this introduces an additional multiphase clock distribution overhead, and is likely to have little impact on static power. Further, the simulation results of [212] show that the output logic states of the molecular diode-based circuits become indistinguishable with more than

about five diodes in the "on" state loading an individual restoring column. This imposes a hard limit on the complexity of the PLA circuits. By the final ITRS technology node (<22nm), all transistors will be formed from what are effectively nanotubes or nanowires in SOI structures and it will therefore be power density that sets the ultimate scaling limit rather than layout area. Thus it appears that mixing conventional and self-assembled manufacturing techniques will not support significant density improvements (as envisaged in [245], for example) and it is therefore unlikely that the technique would be worth the resulting increase in manufacturing complexity.

The clear message here is that moving to esoteric or non-CMOS technologies will not guarantee, *per se*, higher densities or improved performance. Many of the gains to be made towards the end of the roadmap will be achieved by exploiting innovative architectures to overcome the impact of poor device performance and reliability, increasing variability and, in particular, high power.

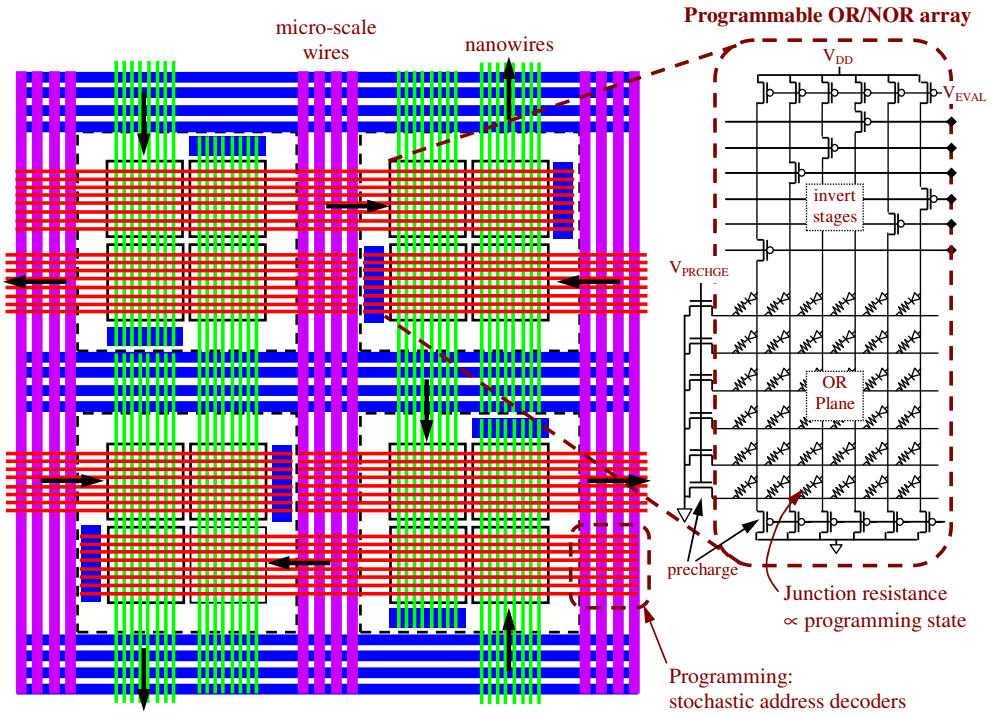


Figure 28. Nanoscale PLA architecture (adapted from [246], [244] and [212]).

2.8.3 Asynchronous Architectures

Asynchronous design has been an active area of research for many years but has failed to achieve widespread application. The elimination of the clock offers a number of immediate benefits [220]: no clock skew; no continuous power loss in the clock tree; average-case rather than worst-case performance; an easing of global timing issues including improved intra-die variability tolerance; automatic adaptation to parameter variations, such as temperature, supply voltage, stress and ageing; robust mutual exclusion and external input synchronization. On the other hand, the removal of the basic synchronous assumption makes the design process significantly harder. The detection and removal of logic hazards, for example, has proved to be very difficult to achieve in a standard CAD flow. Finally, regardless of their average-case performance, it is still not clear whether asynchronous designs offer improved performance in the general case due to interface signalling overheads. However, they have been found to be at least as good as an equivalent synchronous design, and the elimination of losses in the global clock tree, along with their inherently local interconnection topology, make them suitable candidates for future power-efficient computer organizations.

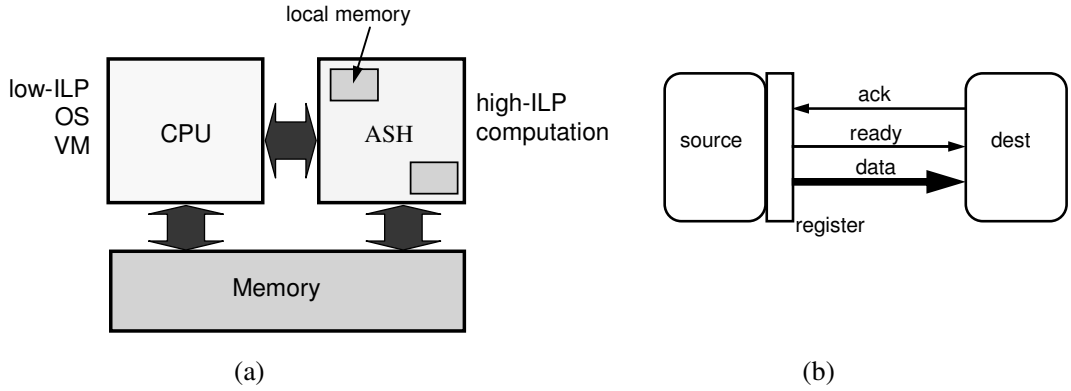


Figure 29. Application-specific hardware (ASH)
(a) The Basic Model (b) Abstract Signalling Protocol (adapted from [230]).

Application-Specific Hardware (ASH)

The Application-Specific Hardware (ASH) architecture [230] assumes that the delay, power and area cost of interconnect will increasingly outstrip that of gates. In the ASH model (Figure 29a), program operations are synthesized as different functional units that are not shared across opera-

tions. High-level language applications are directly compiled into hardware descriptions (without an intervening ISA), which are then synthesized directly into highly pipelined circuits that exhibit only localized, self-synchronizing communication (Figure 29b). The resulting architecture is completely distributed, featuring no global communication or broadcast, no global register files, no associative structures and uses resource arbitration only for accessing global memory [247]. As a result, the system tends to automatically identify and exploit instruction-level parallelism (ILP) where it exists in a particular program, although this model has been shown in [247] to be less efficient on control-intensive programs. In this latter case, the more traditional techniques used in superscalar processors such as branch prediction, control speculation and register renaming will result in superior performance. Ideally then, the code for a complete system would be partitioned between the conventional CPU and the reconfigurable fabric as shown in Figure 29a.

Wave pipelining

Wave pipelining, or maximum rate pipelining [248], is a technique which allows synchronous systems to be clocked at rates higher than can be achieved with conventional pipelining. By applying new data to the circuit faster than the propagation delay, the finite delay through the combinational logic effectively behaves as data storage. The most important design criterion is that the new data must be guaranteed not to interfere with the current data. Thus the clock speed is limited by the difference between the maximum and minimum path delays through the logic and achieving balanced paths (i.e., with equal path delays) will result in maximum performance [249]. The increasing difficulty in achieving and maintaining these balanced paths in the face of high intra-die variability is likely to make this technique less attractive for future architectures.

Asynchronous Wave-Pipelines [250] have been proposed as a method to overcome some of the variability and synchronization issues in conventional wave-pipelines. It recognizes that the technique is inherently asynchronous and thus synchronous boundary registers will impose unnecessary timing constraints on the system. The main difference here is that the synchronizing pipeline registers are controlled by a delayed request signal that propagates with the signal (Figure 30). The circuit will now automatically adjust for global changes in the logic propagation delay

(e.g., due to temperature). On the other hand, increasing physical variability (e.g., thickness variations, line-edge roughness etc.) will remain an issue with asynchronous wave-pipelining.

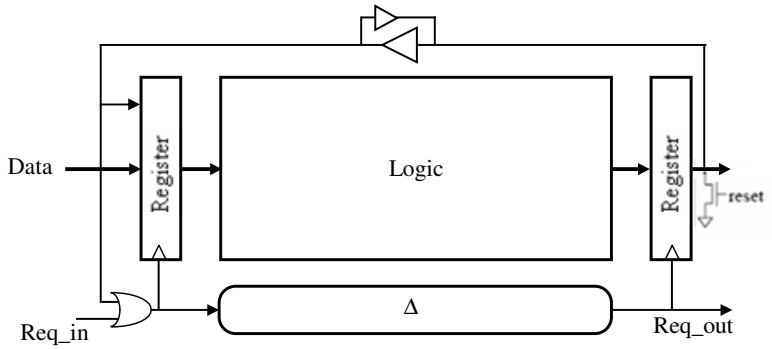


Figure 30. Generic asynchronous wave-pipeline as proposed in [250]

Dataflow Models

Just as the exponential rise in device number has sparked interest in Spatial Computing, it has also rekindled work on the Dataflow model that first emerged the 1970s (e.g. [251, 252]). Within the decentralized dataflow model, each instruction executes ("fires") as soon as all of its necessary inputs are available. As a result, dataflow can extract maximal parallelism where it is available. By contrast, the sequential program counter of the von Neumann processor tends to hide much of the available parallelism.

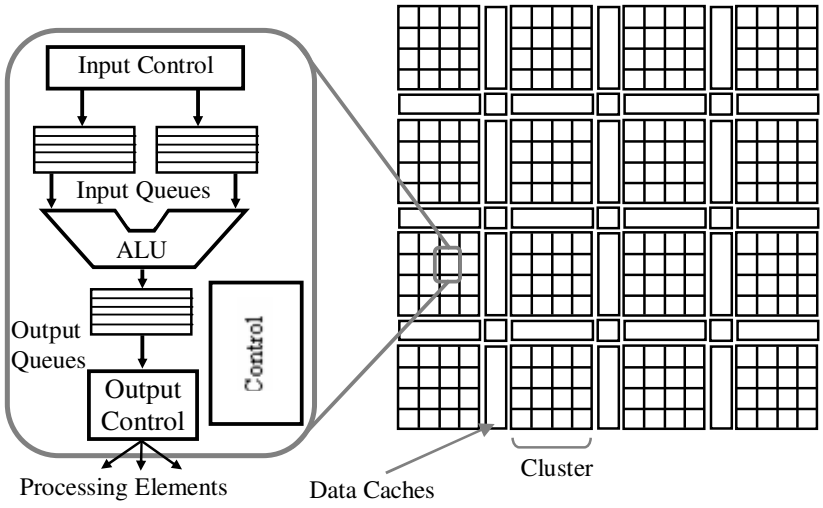


Figure 31. A WaveScalar processor implementation (from [253]).

WaveScalar [254] is a recent dataflow instruction set architecture and execution model that attempts to sidestep the need for instruction and data fetches that otherwise represent so-called "serialization" points in an architecture i.e., operations that create dependencies between instructions. The WaveScalar processor organization proposed in [253] and [254] (Figure 31) comprises a set of replicated tiles, each containing input and output instruction operand queues, communication logic, and an ALU along with dynamic configuration logic to control the placement of instructions. Processor elements within a cluster communicate via a set of shared buses while communication between clusters occurs over a dynamically routed on-chip network. A WaveScalar executable is essentially a fully encoded description of the program dataflow graph that is bound quasi-statically to a particular configuration of processing elements. Because of this, instruction placement impacts greatly on the communication latency and therefore the overall performance of the architecture.

2.9 Summary

This chapter has surveyed a number of the key challenges and opportunities as technology moves further into the nanoelectronic domain. A number of general trends have emerged from this analysis that make it possible to predict some likely characteristics of future architectures.

Integrated Reconfigurability

Reconfigurability offers two major advantages: the flexibility to optimally configure an operator for a particular operation and the ability to configure around physical defects, thereby potentially greatly improving fault tolerance. Examples where reconfigurable fabric can be more efficient than a general-purpose processor include those cases where bit widths are different from the processor's basic word size, where there is significant parallelism available such that multiple specialized function units operate in parallel, where basic operations exist that can be combined into a single specialized operation or where constant operands allow an operator to be greatly simplified. It is clear that the effectiveness of reconfigurable systems can depend greatly on the characteristics of the interface between the fabric and the remainder of the system.

Defect and/or Fault Tolerance

All future computer systems will contain faulty components. Defect/fault tolerance supporting the ability to detect and avoid defects at both the commissioning/ configuration stage and at run-time, is therefore of critical importance. As just mentioned, organizations such as reconfigurable systems can be set up to be intrinsically fault tolerant.

Simplified Manufacturing

The demands of finer design rules will become an increasing problem for semiconductor manufacturing from both a financial and technical perspective. Increasingly finer rules will demand uniform layout densities that will, in turn, restrict the ability to create arbitrary wiring patterns, thereby reinforcing a move towards more regular, mesh-based geometries.

Processor/Memory Convergence

Although the fabrication of RAM and digital logic are completely separated at present, and there is a vast and expensive infrastructure supporting both, the functions of logic and memory must eventually merge if the increasing gap in performance (the "memory gap") between the two is to be overcome. Low-overhead reconfigurability would support the creation of a merged memory/ processing structure in which the idea of "mass storage" is replaced by "mass reconfiguration" as program and data become indistinguishable from the processing mesh.

Coarse-Grain vs. Fine-Grain Architectures

It appears that an inevitable outcome of shrinking geometries, as devices evolve towards ultra thin-film silicon (and from there into molecular and quantum/single electron technologies), is that computer architectures will increasingly be formed from arrays of simple, repeated cells with highly localized interconnect. The current tendency towards heterogeneous, coarse-grained architectures (e.g. multiple CPU blocks, ALU arrays etc.) is being driven by high reconfiguration overheads in devices such as FPGAs. If this can be reduced then fine-grained structures may offer a much more general solution to the creation of flexible reconfigurable computing platforms. Using this model, an ultimate computing fabric might best comprise an homogenous, fine-

grained, non-volatile, fault tolerant, reconfigurable, processing array, exhibiting adjacent or nearest neighbour interconnect only and supporting heterogeneous structures that are derived by compiling a HLL program. Such a processing fabric would ideally be reconfigurable in a way that maximizes a system's ability to exploit parallelism, comprising as many individual processing elements as necessary, each configured in an optimal manner for the particular function.

To summarize, a "wish-list" of features for a future architectural platform might include the following:

- a simplified processing technology supporting a highly regular layout style;
- small logic and interconnect footprints, supporting high component densities;
- reconfigurability: with minimal reconfiguration area and/or performance overheads;
- support for flexible and efficient routing by allowing a continuous tradeoff between routing and logic.

There is a strong case emerging that the combined forces of design effort, manufacturability, reliability, variability and power are pushing future computer organization towards simple, reconfigurable, locally-connected hardware meshes that merge processing and memory. Power (or more strictly, energy-delay) will be the primary limiting factor and will need to be managed at all levels in the design hierarchy. If the overheads associated with reconfigurability can be reduced or hidden, architectures based on fine-grained meshes with rich, local interconnect offer a good match to the characteristics of nanoscale CMOS devices. The following chapter describes a reconfigurable nanocomputer platform which was developed to explore these issues.

Chapter 3. A Double-Gate Reconfigurable Platform

"You've done the simulations, but all it takes is one missing rectangle…"

Allan Strong, Sun Microsystems, quoted in [255]

This chapter proposes and analyses a simple reconfigurable CMOS platform and explores how heterogeneous functionality might then emerge from what is essentially an homogeneous mesh of simple switching devices. Remaining in the CMOS domain offers a number of advantages, including the availability of three terminal switching devices with intrinsic gain, a stable and well characterized manufacturing base plus compatibility with existing design tools. The disadvantage is that the design is constrained by lithographic patterning and alignment issues. While the 2005 ITRS forecasts that feature sizes for logic will approach 16–22nm by 2016 or 2018, it is not clear at the moment how this might be achieved. A premise of this research is that simplified, regular structures with a minimal number of interconnection layers will have a better chance of achieving sub-10nm feature sizes than the complex, heterogeneous layouts that characterize most existing micro-architectures.

The array could be said to be "polymorphic" in that its constituent blocks may be arbitrarily configured into state elements, logic, interconnect or combinations of all three. The system exploits the threshold shift seen at one gate of a double-gate transistor resulting from changes to the bias on the other gate. In this way, the overheads imposed by reconfigurability can be reduced to an extent where fine-grained organizations become viable for general computing. The proposed fabric is a highly regular, locally interconnected, homogeneous structure based on thin-body Schottky barrier technology and may therefore be considered to be representative of end-of-roadmap systems. Parts of this work have been published in [231], [256] and [257].

The approach taken here has involved a hierarchical device-circuit-architecture simulation where each level builds on the previous, thus ensuring the validity of the final architectural models. The overall objective has been to explore the operation and ultimate scalability of this nanoscale reconfigurable system beginning with physical TCAD modeling, via SPICE simulation and culminating in a high-level description implemented in VHDL-AMS, as follows:

1. *TBDGSBSOI TCAD Modeling*: at the time that this work was commenced (2002-3), there had been little published work from which performance predictions could be made for advanced double-gate SOI technology, particularly those with silicide source-drain regions that form the focus of the research here. This part of the work was intended to determine the likely range of threshold shifts achievable via back gate bias changes and to provide a means to approximately "calibrate" the higher level models that were subsequently developed.

2. *FDSOI SPICE Models*: the TCAD results were used to characterize in general terms some University of Florida, level 10 (UFSOI) SPICE models. Of course, the actual model details will depend on process characteristics that are unknown at this time. So the thrust here has been to look at the "macro" behavior, especially in the subthreshold region. These models supported the development of the reconfigurable array, along with some estimates of the power/ performance tradeoffs possible using double-gate technology.

3. *VHDL-AMS Models*: the final stage of this work was based on two successively more abstract models of the reconfigurable architecture created using the analog and mixed signal description language VHDL-AMS on the Mentor Graphics® ADMS platform. The key objective here was to characterize the power-area-performance tradeoffs for more complex circuits that could not be simulated in reasonable time using UFSOI SPICE. The VHDL-AMS model was used to directly support the analytic scaling model developed as part of this work and will be described in Chapter 5.

3.1 Thin-Body Double-Gate SOI

This section examines the basic operation of thin-body (TB), fully-depleted (FD) double-gate (DG) silicon-on-insulator (SOI) devices. These may ultimately become a preferred building block due to their potentially superior sub-threshold performance and better control over short-channel effects. It appears likely that most applications for double gate transistors will link the front and back gates to the same gate voltage as this leads to the best performance as a switching device [258]. However, if the two gates can be accessed independently, one can be used to set the operating point of the transistor thus affecting the behavior seen at the other gate. This can form the basis of novel circuit operation [259-261].

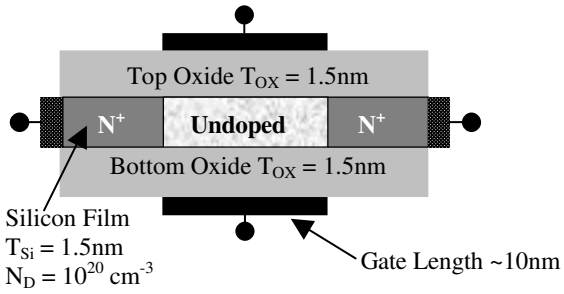
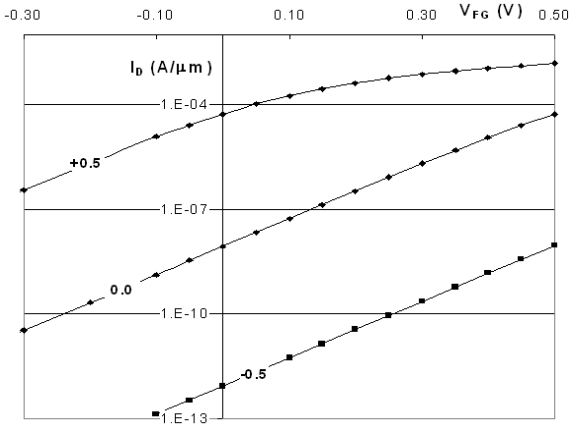


Figure 32. "Canonical" thin-body SOI double-gate NMOSFET (from [262])

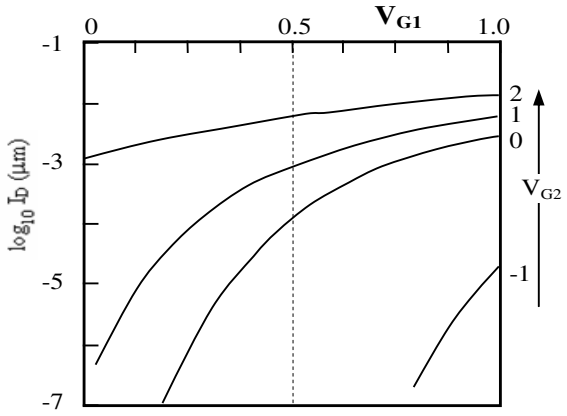
Figure 32 illustrates a generic planar style of a DG-SOI device based on [262], in which the channel is arranged as a horizontal layer of undoped silicon between two vertically opposed gates with conventionally doped source and drain regions. A number of alternative layout styles have been proposed, including vertical-channel [263] and finned (i.e. multi-gate) [263]. In a conventional double gate transistor, of the sort that have been used for many years in high frequency circuits such as RF mixers and oscillators, for example, the gates are typically separated horizontally and their fields act independently on the channel such that the two gates behave as if they are connected in series. In thick-body DG transistors, even of the general style shown in Figure 32, a localized inversion region exists under each gate and the transistor can be considered to comprise

two parallel channels [264]. On the other hand, in thin-body, fully-depleted devices, the gate fields interact in the channel and produce novel behavior not seen in conventional transistors.



Source/Drain doping conc.	(/cm^3)	1e20
Body doping conc.	(/cm^3)	1e13
Gate length (both)	(nm)	9.0
Channel thickness	(nm)	9.0
Gate oxide thickness (both)	(nm)	1.0
Oxide dielectric constant		3.9
Gate workfunction	(eV)	4.188
Longitudinal. rel. elect. mass ratio		0.98
Transversal. rel. elect. mass ratio		0.19
Conduction band offset (subs.-g)	(eV)	3.34

Figure 33. Simulated I_D-V_{FG} characteristics of an ultra-thin body FD-DGSOI transistor showing threshold voltage shift with back gate voltages of +0.5, 0 and -0.5.



Source/Drain doping conc.	(/cm^3)	N/A
Body doping conc.	(/cm^3)	5e14
Gate length	(μm)	30.0
Channel thickness	(nm)	1.0
Gate oxide thickness - top	(nm)	50
Gate oxide thickness - bottom	(nm)	62
Oxide dielectric constant		3.9

Figure 34. Measured I_D-V_{G1} characteristics of DGSOI transistor, T_{SI} = 1nm. Data from [111] for $0<V_{G1}<1$, V_D=50 mV, V_{G2} as shown. In this case, T_{OX} = 50nm.

Figure 33 illustrates this basic idea for the "canonical" device of Figure 32 with an intrinsic silicon channel ($N_a \approx 10^{15}$cm^{-3}) and conventionally doped source/drain regions. The plots were derived using NanoMOS 3.0 using a quantum ballistic transport model with the parameters shown in the adjacent table. NanoMOS is a 2-D simulator for thin body, fully depleted, double-gated n-MOSFETs available at *nanohub.org* [265]. The I_D–V_G curves are shown for back-gate voltages

(V_{BG}) of -0.5, 0 and +0.5, V_{DD} = 0.5V. It can be seen that the device exhibits a threshold shift at the front gate of approximately ±300 mV when the back-gate voltage is moved between ±0.5V.

This can be compared with Figure 34 which shows data from an experimental n-type SOI transistor with a 1nm thick intrinsic channel described in [111]. In this particular device, the body oxide could not be thinned below 62nm, so the top oxide thickness was grown to 50nm. The other parameters, where available, are listed in the accompanying table. This has resulted in a lower sensitivity to the back gate voltage, but a strong threshold effect is still clearly visible.

3.1.1 Thin-Body Silicide Source/Drain Devices

As previously described in Section 2.3.2, Schottky barrier MOSFETs employ metal silicides to form the source and drain regions so that electron confinement can be achieved without the need for doping in either the channel or the source/drain regions. Compared with conventional devices SB-SOI exhibits several advantages, including the elimination of punch-through and latch-up. It is therefore likely to support scaling to sub-10nm dimensions. At these dimensions, the channels would effectively become undoped silicon wires with regular silicide patterns forming the source/drain regions and it may be possible to approach densities of 10^8 gates/mm^2. For this reason ultra-thin body, Schottky-barrier SOI was considered to be an ideal technology on which to base an evaluation of homogeneous reconfigurable meshes at nanoscale dimensions. Although examples of planar Schottky devices have been reported in the literature, and there have been a number of conventionally doped double-gate silicon devices developed, no examples of double-gate Schottky technology have been fabricated to date. Thus the objective of this part of the research was to determine their likely characteristics, especially with respect to drive current (I_D), $\Delta V_{TH}/\Delta V_G$ and $\Delta S/\Delta V_G$ and whether these would support the proposed reconfigurable platform.

3.1.2 TCAD Modeling of TB-DGSOI

Thin-body double-gate p and n-type silicide S/D devices of the general form shown in Figure 35 were analysed with a commercial TCAD simulator [5] using classical drift-diffusion models. These

[5] Atlas/SPisces - Silvaco Inc.

models have been shown to be sufficiently accurate to around T_{SI} = 5nm [266], the limit of the work described in this dissertation. The devices were set up with uniform, lightly doped channels ($N_D = 10^{15}$ cm^{-3}) and with source/ drain regions formed from $ErSi_{1.7}$ (n-type: barrier height ϕ_{BN} = 0.28eV above Si) and PtSi. (p-type: ϕ_{BP} = 0.23eV). The gate work function (ψ) was adjusted to give the desired threshold voltage at $V_{BG} = 0$. The interface between the intrinsic silicon channel and the silicide source and drain regions was assumed to be atomically abrupt and the default gate length was fixed at $4t_{SI}$, (i.e. typically >20nm) to maintain good gate control and to reduce short channel effects that would unnecessarily complicate the analysis. The input decks for all of these experiments have the same form as that shown in Appendix A.

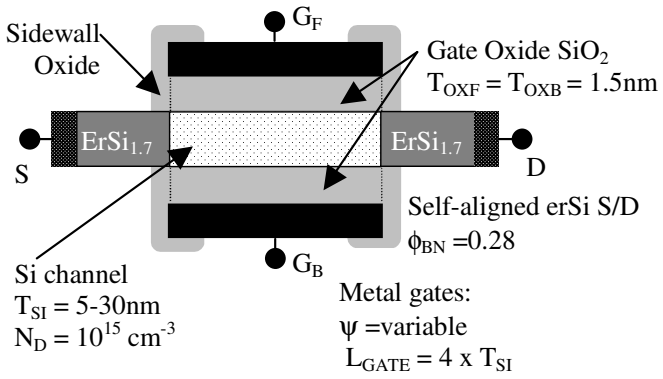


Figure 35. Simplified view of a double-gate n-channel TBFDSBSOI transistor. The general topology is the same as in Figure 32. The p-type uses PtSi source/drain.

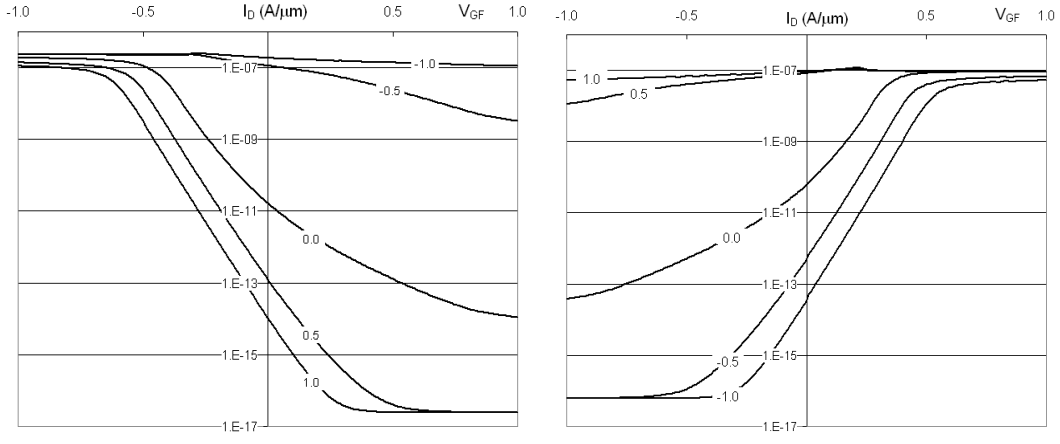


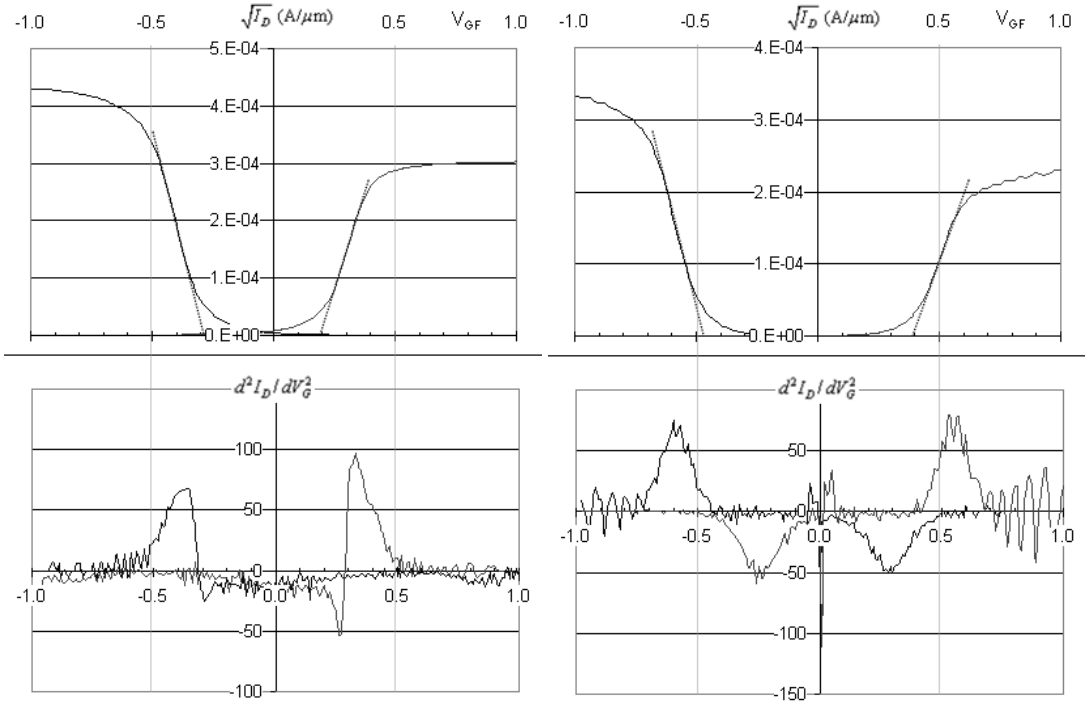
Figure 36. Simulated I_D/V_{FG} Characteristics with $-1.0 \leq |V_{BG}| \leq 1.0$ (a) P-Type; (b) N-Type. In both cases, $|V_{DS}|$=1.0V, T_{SI}=4nm, ψ_G=4.7.

The simulated results for both the P-type (PtSi S/D) and N-Type (ErSi S/D) (Figure 36) confirm that the behavior of this SB double-gate system is similar to its conventionally doped counterpart (cf. Figure 33). One significant difference is that the saturation current density of the Schottky barrier device is up to three orders of magnitude lower. It an also be seen that a ±1V shift in the back-gate bias is sufficient to change the subthreshold current by more than three orders of magnitude in both cases.

A Note on Threshold Voltage Extraction in TBSOI

The analysis in this work relies on the extraction of accurate threshold voltage figures from the TCAD simulations of the ultra-thin-body Schottky barrier devices over a wide range of bias conditions. A number of existing methods were tested against the simulation data, including the conventional constant-current and linear-extrapolation methods, the transconductance change or second-derivative method (i.e., $V_{TH} \approx \max(d^2 I_D / dV_G^2)$) [267] as well as some more complex methods (e.g. as proposed in [268], [269] and [270]). Of these, only the transconductance change method appeared to correctly handle the volume inversion behavior of double-gate MOSFETs [271], as well as being robust in the face of other effects such as interface states, mobility degradation and parasitic resistance [270].

This thesis work primarily takes a comparative approach, in that the threshold voltages are typically normalised to a reference value. Thus, the primary criterion was that the threshold extraction method gave consistent and comparable results across the expected range of bias voltages and drain currents, as opposed to requiring the correct absolute value. As the I-V curves of these nanoscale devices tend not to behave as "regularly" as conventional transistors, the constant-current and linear extrapolation methods did not give reliable results. On the other hand, while the second derivative method was relatively easy to implement and gave repeatable results, it has a tendency to amplify inaccuracies in the measurement of the subthreshold slope, especially numerical errors in the simulation.



(a) (b)

Figure 37. $\sqrt{I_D}$ and d^2I_D/dV_G^2 vs. V_{FG}
for (a) V_{BG}=0V and (b) V_{BG}(p)=+1V; V_{BG}(n)=-1V.

Figure 37 illustrates threshold extraction based on two methods: the extrapolation of the $\sqrt{I_D}$ curve to the V_G axis and the maximum value of the d^2I_D/dV_G^2 waveform. There are two major difficulties with applying the linear extrapolation method here. Firstly, the use of the $\sqrt{I_D}$ curve is based on the assumption that the quadratic (Shockley) model holds, which is increasingly inaccurate in nanoscale devices. More importantly, difficulties in consistently selecting an appropriate value of the maximum slope resulted in non-monotonic threshold values when the extraction was performed automatically (e.g., using the built-in measurement tool within the Atlas TCAD simulator). The second derivative curve show a strong peak at the nominal threshold voltage (where it exists in the measurement range) but also suffers from a severe amplification of noise on I_D (see, for example, $V_{FG} > 0.7$V for the n-type transistor in Figure 37).

In this application, the bias on the second gate can push the threshold well outside the range of the I_D curve, so it is possible for artefacts on the second derivative curve to be misinterpreted. In all

of the work that follows, a "hybrid" method was used. Linear extrapolation positioned the threshold within a range of "likely" values and the peak of d^2I_D/dV_G^2 then identified the actual threshold. Where simulation artefacts arose, such as the double peak in Figure 37b, an average value was used, so that the accuracy depends on the number of samples derived during simulation. It was found that a gate voltage resolution of 10 mV offered a reasonable compromise between accuracy and simulation time.

A Note on Quantum Correction

The finite ground-state electron energy (approximately proportional to $1/T_{OX}^2$) will result in a significant threshold voltage shift for thin-body devices [272]. Where the silicon film thicknesses is in the region of 5nm or less, V_{TH} will be higher than predicted by classical models. For example, the simulations of [273] showed a 20–25% reduction in I_D(sat) when quantum considerations were included, from which it was inferred that the threshold voltage increased with quantization. Similarly, the density-gradient quantum correction method used in [274] resulted in current drive predictions up to 60% less than classical models. On the other hand, these experiments as well as the preliminary NanoMOS simulations undertaken for this work (e.g., Figure 33) indicate that, while the absolute values of V_{TH} and I_D(sat) may change, the overall shape of the I-V characteristics remains largely the same. The drift-diffusion models built into the Atlas simulator were therefore considered to be sufficient for this work.

3.1.3 Threshold Behavior of Thin-Body Devices

Figure 38 shows the sensitivity of the threshold voltage (as seen at the front gate) to the back gate bias for various values of channel thickness (T_{SI}) between 5nm and 30nm for the "canonical" double-gate device of Figure 35. The threshold values have been normalized such that $\Delta V_{TH} = 0$ at $V_{BG} = 0$. These plots are for the n-type transistor, but those for the p-type have an identical form. It can be seen that as the channel thickness is reduced the threshold sensitivity increases to a point where at $T_{SI} = 5$nm, setting $V_{BG} = -1$V can produce ~0.45V shift in threshold voltage. A shift of similar magnitude is observed for the p-type device at $V_{BG} = V_{DD}+1$. Figure 39 plots the

same data vs. T_{SI} over the range $-1.0 \le V_{BG} \le 0.0$ so that the increased sensitivity (i.e., the slope of $\Delta V_{TH}/\Delta V_{BG}$) with reducing channel thickness can be more easily seen.

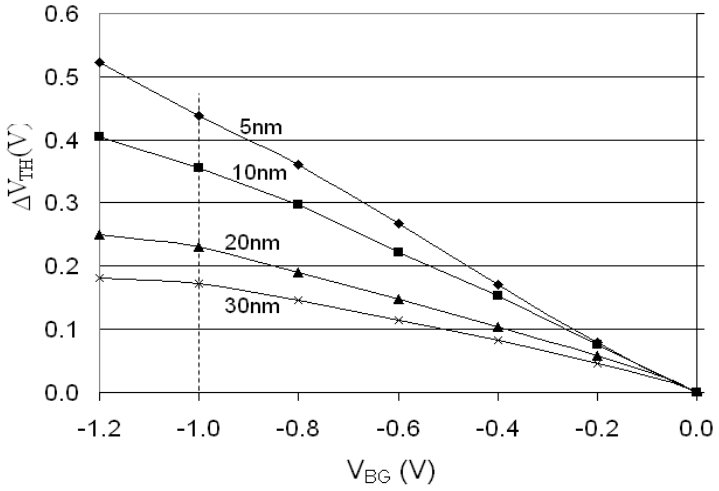


Figure 38. Threshold voltage change (ΔV_{TH}) vs. back gate voltage at various T_{SI} for the n-type device of Figure 35. The P-type device characteristics are similar.

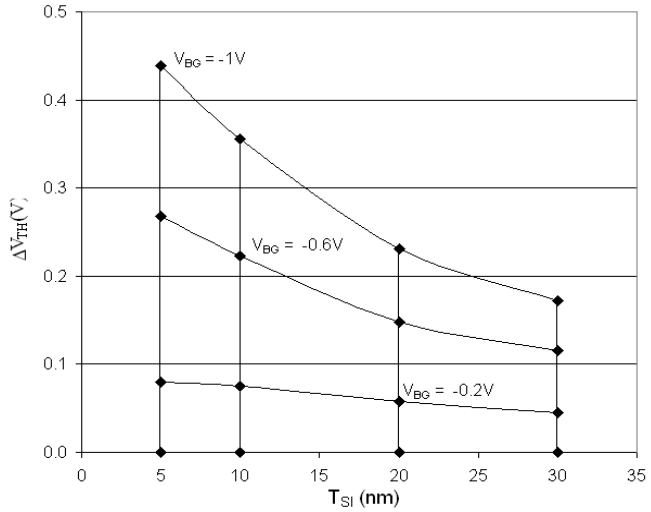


Figure 39. ΔV_{TH} vs. silicon film thickness, $5nm \le T_{SI} \le 30nm$.

In the ground-plane mode (i.e., with the back gate at a fixed potential), the behavior of the sub-threshold slope (S) is similar to that of planar devices and is given by [275]:

$$S = \frac{kT}{q}\ln(10)\left[1 + \frac{C_S}{C_{OXF}}\right] \tag{3.1}$$

where: C_{OXF} = the front gate oxide capacitance ≈ ε_{OX}/T_{OXF} and C_S is the effective body capacitance between the inversion layer and the back gate: $C_S \approx \varepsilon_{SI}/T_{SI}$ if the back surface is in accumulation, and $C_S = C_{SI}C_{OXB}/(C_{SI}+C_{OXB})$ in depletion. Substituting $\varepsilon_{SI} \approx k_r\varepsilon_{OX}$, ($k_r = \varepsilon_{SI}/\varepsilon_{OX} \approx 3$ for SiO_2), the slope becomes:

$$S \approx 60\left[1+\frac{k_r T_{OXF}}{k_r T_{OXB}+T_{SI}}\right] \quad mv/decade. \tag{3.2}$$

The term k_rT_{OXB} becomes zero if the back surface is in accumulation. In [111], it is shown that for ultra-thin body devices (e.g. T_{SI} < 3nm), it is more-or-less impossible to bias the back surface in accumulation as the required voltage (given approximately by C_{SI}/C_{OXB}) is very large and would cause the failure of the oxide.

While reducing the body thickness increases the threshold sensitivity ($\Delta V_{TH}/\Delta V_{BG}$), (3.2) implies that it also degrades the subthreshold slope. To maintain S < 100 mV/decade at T_{SI} = 5nm would require the effective oxide thickness to be less than 1nm, resulting in manufacturing difficulties, high gate tunneling currents and oxide reliability problems. While it is theoretically possible for DG-SOI transistors to approach the ideal subthreshold slope for MOS (~60 mV/decade) when used in double gate mode (both gates driven together), this is not the case for the ground-plane mode (see (2.15), page 37). No device fabricated to date has achieved this figure. The $YbSi_{2-x}$ S/D bulk device described in [276] achieved S = 75 mV/dec., while more typical values (e.g. [277], [278]) range between 100 and 150 mV/decade (i.e. C_S/C_{OX} between 0.6 and 1.5). It is suggested in [266] that this restricts the V_{TH} tuning range (e.g. to $|\Delta V_{BG}| < 0.25$ in that study). However, Figure 40 shows that useful subthreshold leakage reductions can still occur well outside this range for thin-body devices with T_{SI} in the range of 5–10nm.

Although the sensitivity of the subthreshold shift to back-gate bias ($\Delta V_{TH}/\Delta V_{BG}$) increases monotonically with reducing T_{SI}, the interaction between $\Delta S/\Delta V_{BG}$ and $\Delta V_{TH}/\Delta V_{BG}$ results in the minimum absolute subthreshold current being achieved at T_{SI} = 10nm. Equation (3.2) also implies that moving to hi-κ gate dielectrics (as well as serving to reduce gate leakage) can significantly

improve S. For example, with $HfSiO_4$ ($\varepsilon_r > 12$), k_r in (3.2) will become less than 1.0 and the worse-case slope will reduce to approximately 78 mV/decade, although this is likely to be at the expense of increased short channel effects [279].

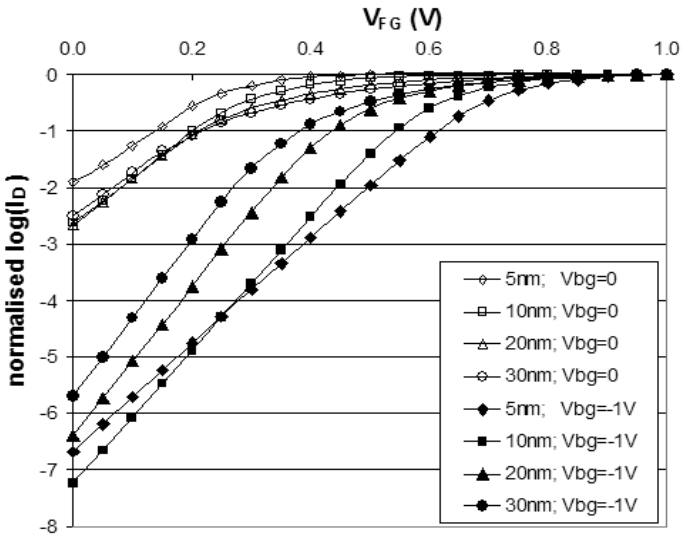
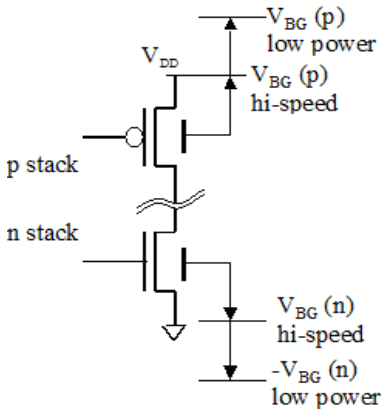


Figure 40. TCAD simulated log(I_D) vs. V_{FG} (n-type) for various body thickness values (T_{SI}). I_D has been normalized to its value at V_{FG}=1V, showing the relative effect on I_{OFF} achieved with a –1V shift in V_{BG}. V_{DD}=1V, and the initial threshold voltage for each curve (at V_{BG}=0) has been set to approximately 0.2V.

3.2 Physically Based SPICE Models for TB-SOI

In the second stage of the simulation, the TCAD results for devices with T_{SI} = 5nm were used to approximately calibrate full-depleted UFSOI SPICE models (University of Florida, level 10) [280]. All of the following simulations were run on *nanohub.org* [265]. The UFSOI models are charge-based with five terminals (two gates, source, drain and a reference bulk), and were used in their floating-body mode. The objective here has been to determine the fundamental power-performance tradeoffs in simple CMOS circuits. To achieve this, the SPICE models were applied to a number of static circuits with the general form shown in Figure 41. The simulated I_D vs. V_G for both p and n-type devices with T_{SI} = 5nm, T_{OX} = 1.5nm is shown in Figure 42 along with data from fabricated p-type silicide devices reported in [281] and n-type from [71] that have been included for comparison. Two examples of specific SPICE input decks can be found in Appendix B.



Mode	V_{GB}(p)	V_{GB}(n)
Low Threshold	1V	0V
High Threshold	+2V	-1V

Figure 41. The general form of the double-gate CMOS transistor stack.

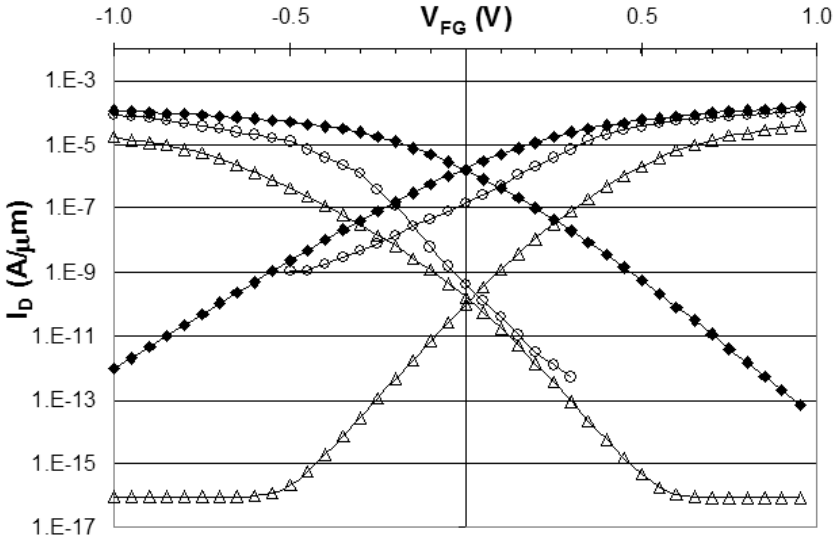
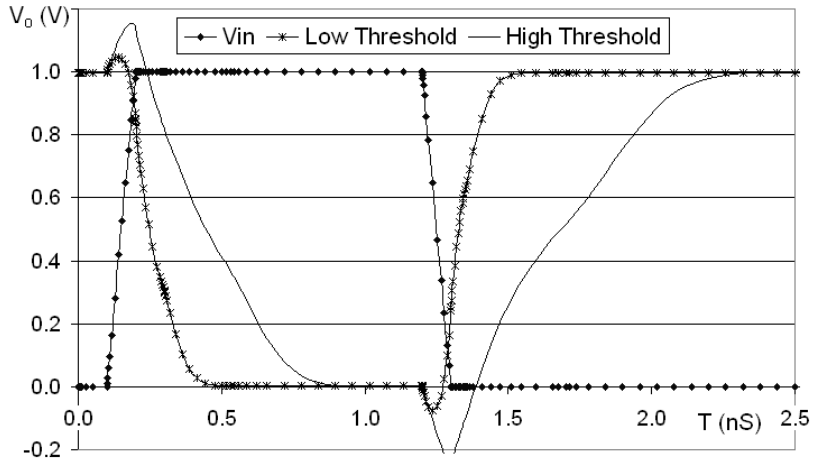


Figure 42. SPICE simulated I_D vs. V_{GS} for p and n-type double gate silicide S/D devices T_{SI}=5nm, T_{OX}=1.5nm; W/L=3. High Threshold Mode (open triangles) V_{DD}=1.0; V_{BGP}=2.0, V_{BGN}=-1.0; Low Threshold Mode (filled diamonds) V_{BGP}=1.0, V_{BGN}=0. Data from fabricated p-type silicide devices reported in [281] and n-type from [71] are included for comparison (open circles).

The I_D-V_{FG} characteristics derived using these SPICE models (Figure 42) exhibit the same general shape as the curves of Figure 36 and it can be seen that shifting the bias on the back gate (V_{BG}) by 1 volt above V_{DD} or below ground shifts the value of I_{OFF} (I_D at V_{GF} = 0) by approximately the same amount as determined previously in the TCAD simulations. For these particular devices, this is equivalent to increasing $|V_{TH}|$ by just over 0.3V. V_{BGP} = V_{DD} (p-type) and V_{BGN}=0 ground (n-type) sets the circuits into its low threshold/high performance (and high power) mode, while V_{BGP} = V_{DD}+1V and V_{BGN} = –1V sets the high threshold/low power mode (Figure 41).



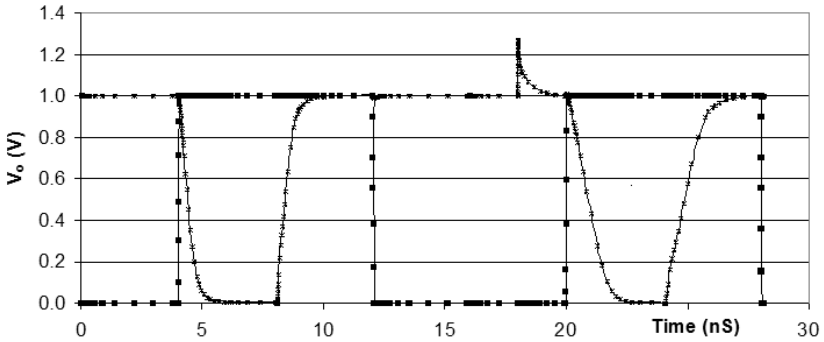
Mode	I_{OFF} (A/μm)	T_P (pS)
High speed – low threshold	5×10^{-7}	100
Low power – high threshold	1.1×10^{-11}	490
LSTP 2018 target	1×10^{-10}	10

Figure 43. Basic inverter characteristics (FO-4):
Curve 1 = high-speed mode; curve 2 = low-power mode. Also shown are the average I_{OFF} and average FO-4 propagation delays for both modes. The ITRS 2018 low standby power and delay targets for a single nMOS device are included for comparison.

The impact of this shift in threshold is shown in Figure 43 for a simple inverter circuit (long channel, L = 0.33μm, W_N = 1.0μm; W_P = 2.0μm; V_{DD} = 1.0, $V_{TH} \approx 0.3$; FO-4). In this example, the ratio of I_{OFF} between the two modes is more than 5×10^4. At the same time, the average FO-4 delay, $T_P = (T_{PLH}+T_{PHL})/2$, increases by a factor of almost five. Whereas the standby current in this case is an order of magnitude below the ITRS 2018 target for low standby power technology (at V_{TH} = 0.4V), the nominal saturation drive current of these silicide S/D devices is 11 times lower (~90μA/μm vs. 990μA/μm), implying that the FO-4 delay will be at least an order of magnitude greater than the ITRS target at the same fanout. To achieve low standby power at the same time as high performance using a fixed V_{TH} would require an almost ideal value of subthreshold slope. On the other hand, uncoupling these two objectives, so that power and subthreshold leakage may be optimised separately, will significantly relax this subthreshold slope target.

In a typical realisation of the dual-gate SOI transistor, the back gate presents a load to the bias circuit that is approximately the same as that of the front gate. Switching between modes can therefore occur at normal circuit rates with minimal disruption to the operation of the circuit. In

the 2-input NAND gate example shown in Figure 44, the back gate biases were switched from high to low power modes at 18nS with a rise time of 500pS. The table in Figure 44 shows the impact on the propagation delay and the subthreshold leakage. In the application envisaged here, the low-power delay times might be considered to be irrelevant as the gates will not be operated in this mode.



Mode	T_{PHL} (nS)	T_{PLH} (nS)	Subthreshold Leakage (nA)			
			00	01	10	11
Low threshold	0.32	0.37	536	1090	1090	3500
High Threshold	0.58	0.83	0.067	0.134	0.127	0.162

Figure 44. 2-NAND gate characteristics (all transistors: L=350nm, W=1.4µm). Mode switching from high speed to low power occurs at T=18nS. For clarity, only one input waveform is shown. The table shows the propagation delay times along with the subthreshold leakage for each input logic state.

Table 7 Subthreshold leakage power vs. supply voltage, 1-bit CMOS full-adder (Figure 45) under various threshold scaling regimes; T_{SI}=5nm; ΔV_{BG}=±1V.

Scaling	V_{DD}	V_{TH}	P_{OFF} (µW) (high-P)	P_{OFF} (nW) (Low-P)	Power Ratio
ITRS HP logic	1.0	0.15	28	5.5	$5.5x10^3$
	0.8	0.14	22	1.6	$1.3 x10^3$
	0.7	0.13	8.75	3.0	$2.8 x10^3$
ITRS LOP	0.7	0.22	6.4	0.5	$1.3x10^4$
	0.6	0.19	7.8	1.0	$7.8x10^3$
	0.5	0.17	7.0	1.0	$7.0x10^3$
Fixed V_{TH}	1.0	0.2	17	1.8	$9.4x10^3$
	0.8	0.2	9.5	1.1	$8.9 x10^3$
	0.7	0.2	8.5	0.9	$9.3 x10^3$

The results for a more complex circuit are shown in Table 7, in which a conventional 28 transistor static CMOS full-adder [282] (Figure 45) has been analysed over a range of supply and threshold

values. The threshold values shown here are for the nFET at 0V back-gate bias (the devices are symmetrical). The first two groups are based on the supply and threshold targets for the ITRS high-performance and low-operating power technologies from 2010 through 2018. Using a ±1V shift in back gate bias achieves an approximate shift in threshold of 0.45V (relative to the values listed in the table). As a result, I_{OFF} is typically reduced by a factor of more than 10^3 in these examples.

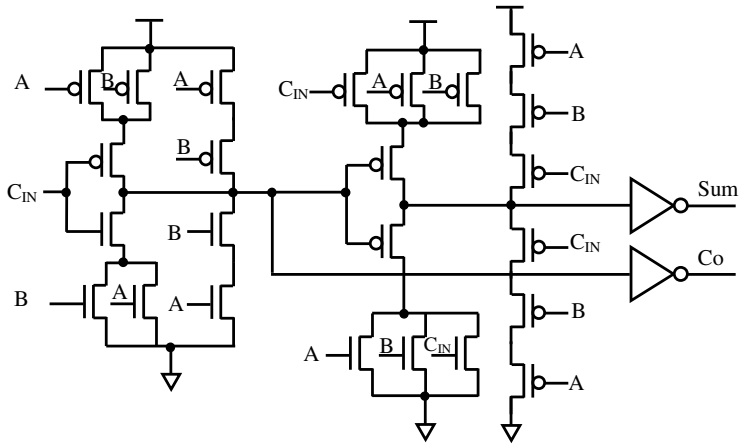


Figure 45. 28 transistor static CMOS Full Adder circuit

3.3 A Reconfigurable Array based on TBDGSOI Devices

The previous section illustrated the magnitude of the power reductions that can be expected from TBDGSOI devices by shifting the threshold voltage between active and standby modes. As the switching threshold of a CMOS logic gate is also a function of V_{TH}, low-power operation can be combined with an ability to functionalize a reconfigurable structure. This section describes the structure and operation of a proposed reconfigurable platform based on these Schottky barrier devices. The objective here has been to develop a *plausible* model of an end-of-roadmap system in order to evaluate the likely performance of architectures mapped to mesh-connected arrays of this type.

3.3.1 Reconfigurable Double-Gate Cell

Figure 46 shows the simulated DC transfer characteristics of an inverter circuit formed from these DG transistors in which the two back gate connections are tied together such that the threshold voltages of the p and n-type transistors shift in opposite directions for a given change in bias under five operating conditions (V_{G2} = 1.5, 0.5, 0, -0.5 and -1.5, V_{DD} = 1V, no load).

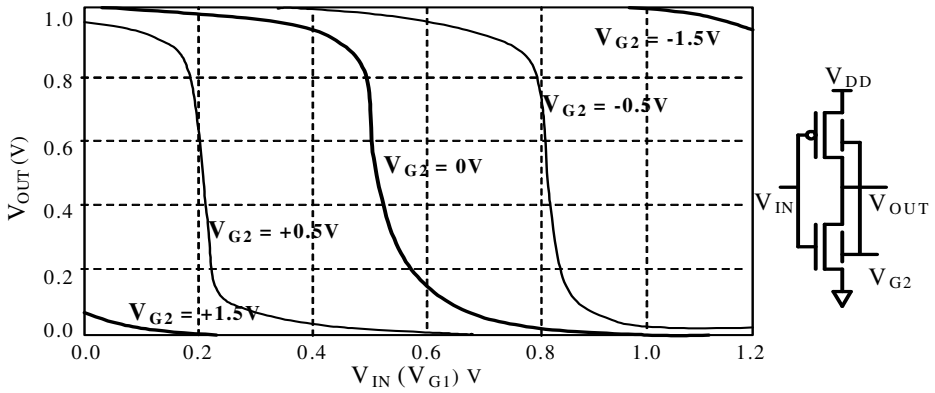


Figure 46. DC transfer characteristics of a variable switching threshold inverter with V_{G2} = 1.5, 0.5, 0, -0.5 and -1.5, V_{DD} = 1V, no load.

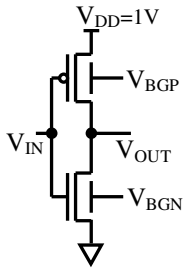
This simulation verifies that altering the back gate bias moves the switching threshold of the inverter such that, at the two extremes, the output stays high (V_o>0.8V) or low (V_o<0.1V) while for back-gate values around 0V, the output switches symmetrically. The DC transfer curves (Figure 46) are similar in general form to the characteristics of the planar "ground plane" (GP) CMOS inverter described in [179] and match fairly closely the relationship between the back gate (V_{G2}) bias and the shift in switching point observed in that study.

The static subthreshold current for the inverter of Figure 46 are listed in Table 8. In the first three rows the back gate voltages are coupled and set between -2 and +2, whereas the second three entries have separate back-gate voltage entries. It can be seen that back-gate voltages of +1 (p) and 0 (n) result in an active mode with more symmetrical subthreshold currents and also allow the gate to be placed in a standby mode (final row) in which the subthreshold currents are almost four orders of magnitude lower than in the active mode. These results indicate that for this 5nm device, setting the back-gate voltages to between ±1V (nMOS) and V_{DD}±1V (pMOS) will be suffi-

cient to both functionalize the device between active, standby and fixed output, and to constrain the static power levels of the gate.

Table 8 Subthreshold current vs. back-gate voltage for a simple inverter.

V_{IN} (V)	V_{OUT} (V)	V_{BGP} (V)	V_{BGN} (V)	Comments	I_{SUB} (A)
0	1	-2	-2	Back-gates tied V_{OUT} high	3.19E-16
1	1				3.60E-07
0	1	0	0	Back-gates tied Active mode	8.41E-07
1	0				2.42E-05
0	0	+2	+2	Back-gates tied V_{OUT} low	2.42E-05
1	0				1.14E-11
0	1	0	-1	Back-gates separated V_{OUT} high	1.13E-11
1	1				2.20E-05
0	1	1	0	Back-gates separated Active mode	8.41E-07
1	0				1.93E-07
0	0	+2	+1	Back-gates separated V_{OUT} low	2.40E-05
1	0				1.14E-11
0	1	+2	-1	Back-gates separated Standby mode	1.13E-11
1	0				1.14E-11

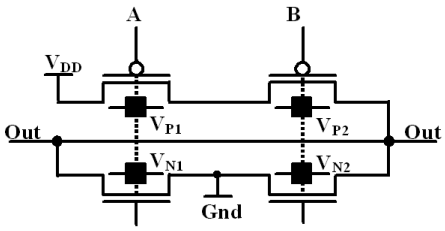


$W_p=2W_n=2\mu m$,

$L_{p,n}=330nm$

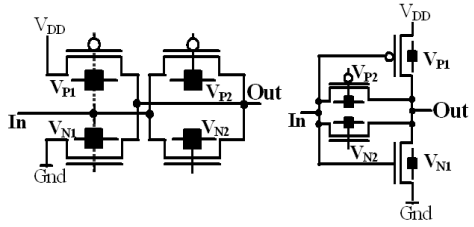
This basic mechanism can now be extended to form more complex logic circuits. The circuit of Figure 47a is essentially a 2-input NOR gate (2-NOR) in which each transistor is controlled by an individual back gate bias voltage (shown as black squares on the diagram). In normal operation, the switching threshold seen at each gate is set to about $V_{DD}/2$. Moving this switching point to $>V_{DD}$ (as in Table 8) effectively disables that input (i.e., the gate becomes insensitive to that line). In this NOR configuration, setting the switching thresholds of all inputs $>V_{DD}$ will cause the output to become unconditionally high while if any input is set < 0V the output will be unconditionally low. The table in Figure 47a describes the set of logic functions that can be developed using this technique. A pair of adjacent cells will support a generalized NOR-NOR LUT structure that can derive any sum-of-products logic function of an arbitrary set of its inputs.

The circuit of Figure 47b is a reconfigurable pass-gate/inverter block. In this case, the back-gate bias states may activate either the pass-gate or the inverter or neither. As such, the circuit can be used to transfer a signal (bidirectionally) or its complement (in one direction only) between adjacent cells or alternatively to provide isolation between the cells.



	Control gate voltages (V)			
f(A,B)	V_{P1}	V_{N1}	V_{P2}	V_{N2}
$\overline{A}$	1	0	0	-1
$\overline{B}$	0	-1	1	0
$\overline{A+B}$	1	0	1	0
1	0	-1	0	-1
0	2	1	2	1

(a)



	Control gate voltages (V)			
f(IN)	V_{P1}	V_{N1}	V_{P2}	V_{N2}
$\overline{IN}$	1	0	2	-1
IN	2	-1	0	1
O/C	2	-1	2	-1

(b)

Figure 47. TB-DGSOI transistor circuits
(a) Configurable 2-NOR Gate; (b) Configurable Inverting/Non-inverting Pass-gate

Using Chen's saturation drain current model described previously in Section 2.5.1, and equating the drain current in the usual way [283], an approximate formula for the switching threshold of each NOR gate or inverter is given as:

$$V_{SW}(NOR) = \frac{V_{THN} + n^{-1}K_r^{1/\alpha}(V_{DD} + V_{THP})}{1 + n^{-1}K_r^{1/\alpha}} \tag{3.3}$$

where $V_{THN(P)}$ is the threshold voltages for the N (P) devices, n is the number of transistors in the stack (e.g., n = 1 for an inverter, ≥2 for a NOR stack) and $K_r = K_P/K_N$. Thus, K_r will be a function of the transistor gain ratio, given by $(W/L)_P/(W/L)_N$ and the relative mobility values. In the case of the Schottky devices considered here, the effective mobility of the pMOS devices is greater than that of nMOS due to the lower barrier height of PtSi (0.23V vs. 0.28V for $ErSi_{1.7}$). It can be seen from (3.3) that for the symmetrical threshold case ($|V_{THN}| = |V_{THP}|$), the switching threshold (V_{SW}) becomes $V_{DD}/2$ when $n^{-1}K_r^{1/\alpha} = 1$ and thus $K_r = n^{\alpha}$. Further, $V_{SW} \geq V_{DD}$ when:

$$V_{THN} + n^{-1}K_r^{1/\alpha}V_{THP} \geq V_{DD} \tag{3.4}$$

and $V_{SW} \leq 0$ when:

$$V_{THP} + n^{-1}K_r^{1/\alpha}V_{THN} \leq -V_{DD} \tag{3.5}$$

Equations (3.4) and (3.5) illustrate two important points about this array. Firstly, as the supply reduces with scaling, the range of threshold shifts that will be required to configure the array will also scale down. Secondly, although the optimum value of K_r (i.e., such that $V_{SW} = V_{DD}/2$) is related to α, in common with all static gates we can adjust K_r over a wide range with a minimal effect on V_{SW} (but with an effect on performance [283]). It is also possible to achieve a shift in the switching threshold across the range V_{DD} to 0V with $\Delta V_{TH} \approx \pm 0.45V$. For the devices in Figure 47a and Figure 47b, this would mean modulating the front gate by approximately ±1.3V. The simulations described in Section 2.8 predict that a $\Delta V_{TH}/\Delta V_{FG}$ of ±0.45V can be expected with ultra-thin body (i.e., T_{body} = 5nm) DGSBSOI with T_{OX} = 1nm. Thus, gate biases in the range ±1V to ±2V, that are also compatible with oxide reliability [172], will be sufficient to configure the array. As a final note, an advantage of this organization is that complementary operation is maintained regardless of the logic configuration.

Threshold Variability Effects

As the functionality of the array depends on the relative threshold voltages of the p and n, it will also be sensitive to the various sources of threshold variability outlined in Section 2.3.3. The worse-case effect will occur in (3.3) when the two thresholds V_{THN} and V_{THP} are both offset in the same direction and to their maximum value. With $V_{THN} = -V_{THP}$, (3.3) becomes:

$$V_{SW} = \frac{(V_{TH} + \varepsilon V_{TH} + \Delta V_{FG}) + K_r'(V_{DD} + (-V_{TH} + \varepsilon V_{TH} + \Delta V_{FG})}{1 + K_r'} \tag{3.6}$$

where $K_r' = n^{-1}K_r^{1/\alpha}$, εV_{TH} is the offset from V_{TH} due to manufacturing spread and temperature and ΔV_{FG} is the effective (configuration) threshold shift due to the front-gate bias changes.

From (3.6), it can be seen that the overall impact of both variability and transistor gain ratio will be to increase the front-gate bias required to set a given switching threshold, V_{SW}. Figure 48 shows the switching threshold offset over K_r' at $\Delta V_{FG} = 0$. As K_r' decreases, a greater range of ΔV_{FG} is needed to overcome the increasing asymmetry in the switching point. For example, to achieve $V_{SW} = V_{DD}$ with εV_{TH} at +25% from the nominal threshold with K' = 1 requires a normalized bias voltage ($\Delta V_{FG}/\Delta V_{DD}$) of just under 0.64, whereas at K_r' = 0.5 this increases to around 0.7. Assuming that $\Delta V_{TH}/\Delta V_{FG} \approx \pm 0.45$V as previously, this raises the required configuration bias from around ±1.3V to approximately ±2.0V.

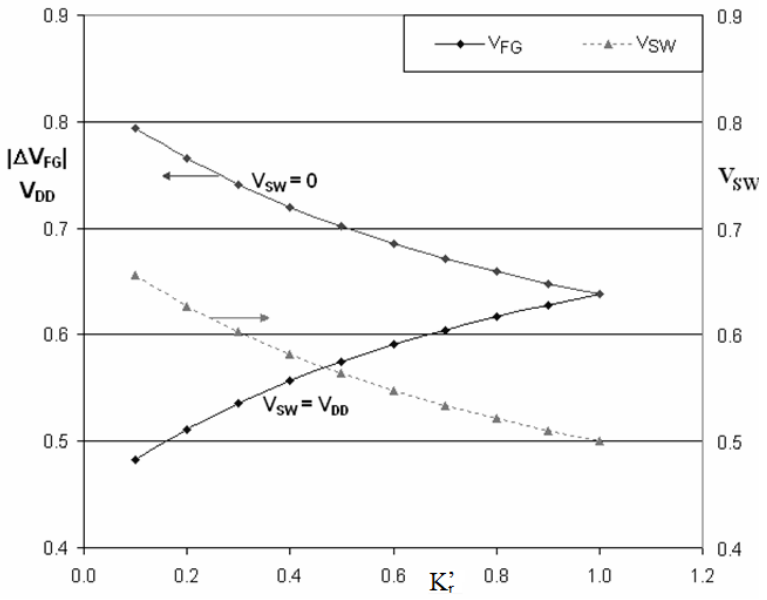


Figure 48. Normalized $|\Delta V_{FG}|$ vs. K_r' required to achieve $V_{SW}=V_{DD}$ or 0V at $\varepsilon V_{TH}=\pm 25\%$.

3.3.2 Reconfigurable Array Topology

Having created what is, in essence, an undifferentiated leaf-cell, the question remains as to the best way to deploy it. A detailed study of the various options is outside the scope of this work, which is aimed at a more abstract analysis of the performance of this homogeneous reconfigurable computing fabric. Instead, it is simply noted that various studies of reconfigurable systems such as FPGAs [284-286] and their interconnection structures [287-289] have indicated that a LUT size of between 4 and 7 inputs appears to offer an optimum area-delay tradeoff over a range of

applications. Similar studies on multiple output LUTs [290, 291] showed that a 4-input LUT gives the minimum area while in [292] He and Rose suggest that a mixture of different sized LUTs (for example 4-LUTs and 6-LUTs) may provide a better tradeoff between speed and density. More recently, an analysis that focused on yield and leakage [293] found that in FPGA devices with multiple supplies or power gating, a 4-LUT resulted in the highest leakage yield, whereas a size of seven gives the highest timing yield. That study recommended a LUT size of five as a good compromise between leakage and timing yield. In the following, a 6x6 logic block has been chosen as two adjacent blocks are then sufficient to implement either a small state machine structure such as a D-type flip-flop or transparent latch, or a conventional arithmetic circuit such as a single-bit full adder.

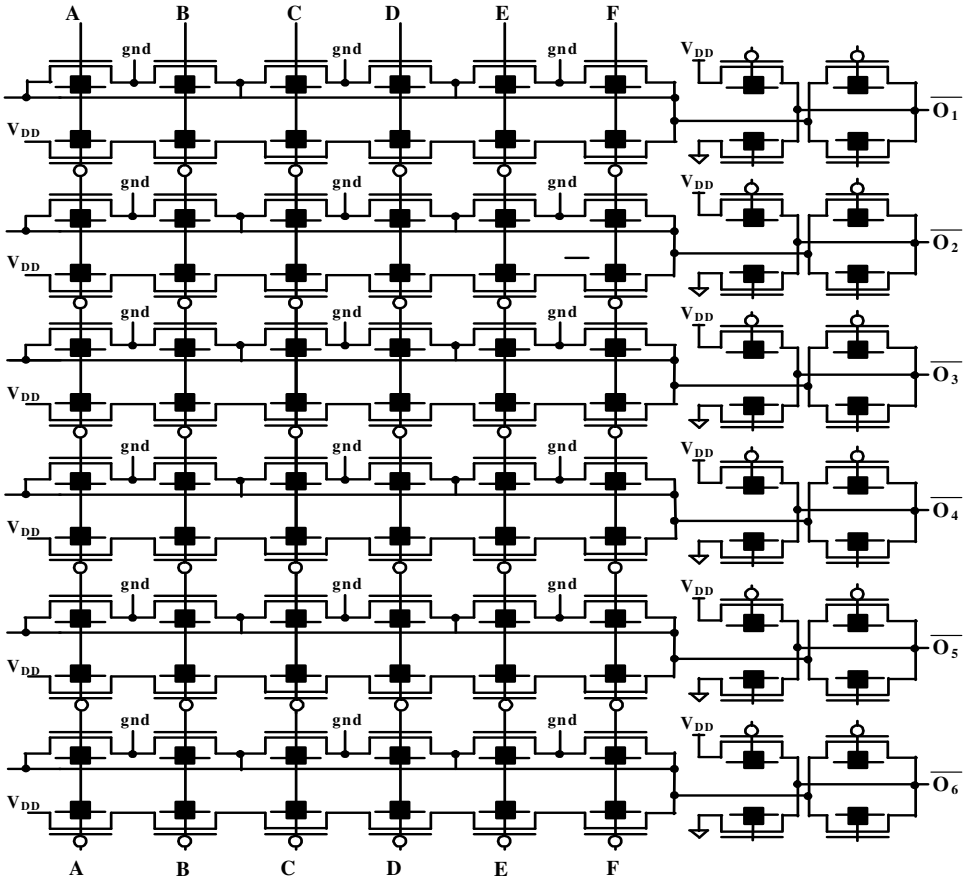


Figure 49. An example reconfigurable cell based on a 6x6 NOR organization.

A Basic Reconfigurable Block

The basic logic block (Figure 49) is arranged as a 6-input, 6-output NOR array with each (horizontal) output terminated in a configurable inverter/3-state interface described previously in Figure 47b. Other organizations (e.g. NAND) are possible with simple modifications to the internal connections of the array. In this case, a NOR topology was chosen as, unlike its planar counterpart, the effective mobility of the SB p-type transistor mobility is higher than that of the n-type, potentially leading to a more compact implementation. As outlined above, the interface circuit serves a number of purposes. In its off-state, it decouples adjacent cells and determines the direction of logic flow. Configured as an inverting driver, it supports the creation of more complex logic functions and, just as importantly, provides a buffer function that will allow any output line to be used as a data feed-through from an adjacent cell. Finally it can be configured as a (bi-directional) pass-transistor connection between neighboring cells.

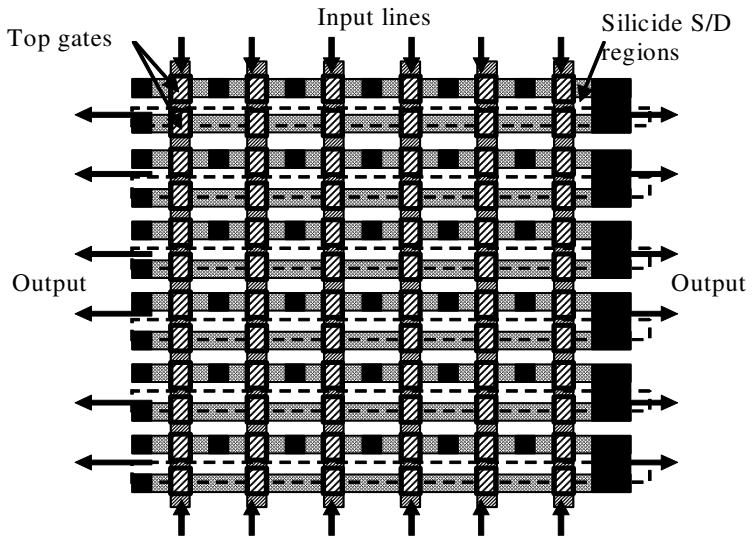


Figure 50. Simplified symbolic layout of the 6-input, 6-output array.

In the simplified example layout illustrated in Figure 50, pairs of p and n-type transistors have been formed on an array of undoped silicon nanowires by patterning erbium and platinum silicide in alternate rows. An alternative organization might be based on the Opposite Gate FLASH configuration illustrated in Figure 51 [294]. In this case, one gate (the bottom, floating gate in Figure 51) acts as the programming gate such that the change on the floating gate sets the thresh-

old seen at the top gate in the same way as a conventional FLASH. From an external viewpoint, this reconfiguration array appears as a simple 8x8 RAM block and would need to be controlled by set of (multi-valued) RAM drivers surrounding the array and forming a link to a reconfiguration bit stream. Each of the 6x6 NOR blocks would require 36x2x2 = 144 bits of reconfiguration data, in the same order (on a function-for-function basis) as the several hundred bits required by typical CLB structures and their associated interconnects in commercial FPGA devices [19].

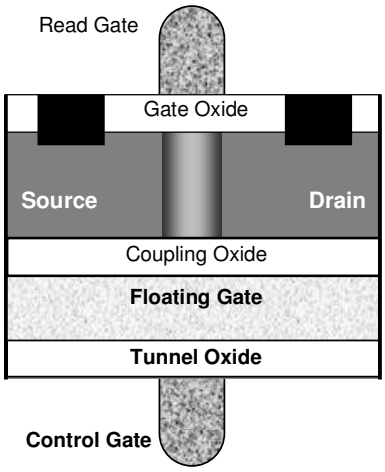


Figure 51. Layout cross-section of Opposite-Side Floating-Gate FLASH Memory (adapted from [294]).

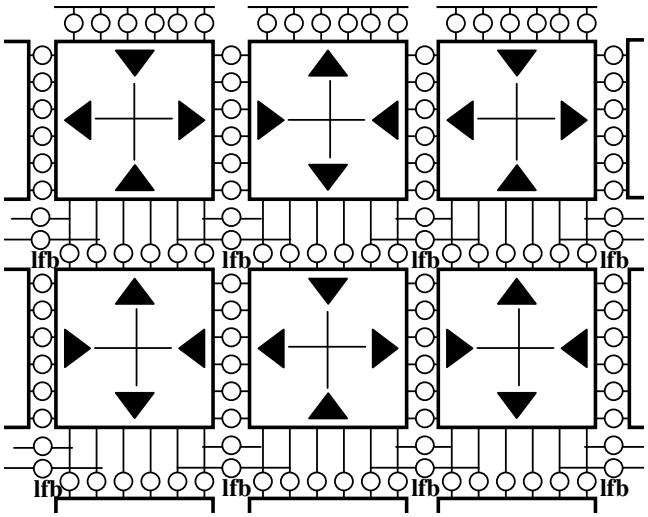


Figure 52. Simplified partial view of the array connectivity.

In Figure 52, the basic NOR cells are stacked into a regular square array somewhat reminiscent of an FPGA, in which each cell has the same functionality regardless of location but without any form of hierarchical routing. The open circles in Figure 52 represent the pass transistor switches connecting adjacent cells, while the black arrows indicate the potential flow directions of the logic or interconnect (i.e., with a 90° rotation at adjacent cells). Pairs of adjacent cells, configured together, represent the equivalent of a small LUT with 6 inputs, 6 outputs and 6 product-terms. The two local feedback lines (labelled *lfb*) can be used to form two transparent latches or a single edge triggered register. When configured in this way, a single cell (or a pair of adjacent cells) may form logic and/or interconnect, with or without latched/ registered outputs. Because of the regularity of the structure and the adjacent connectivity, the array has the potential to be very dense. For example, a 6-LUT cell-pair could occupy around $4–500\lambda^2$ (λ = feature size) compared with the $6x10^5\,\lambda^2$ for a typical 4-LUT estimated in [19].

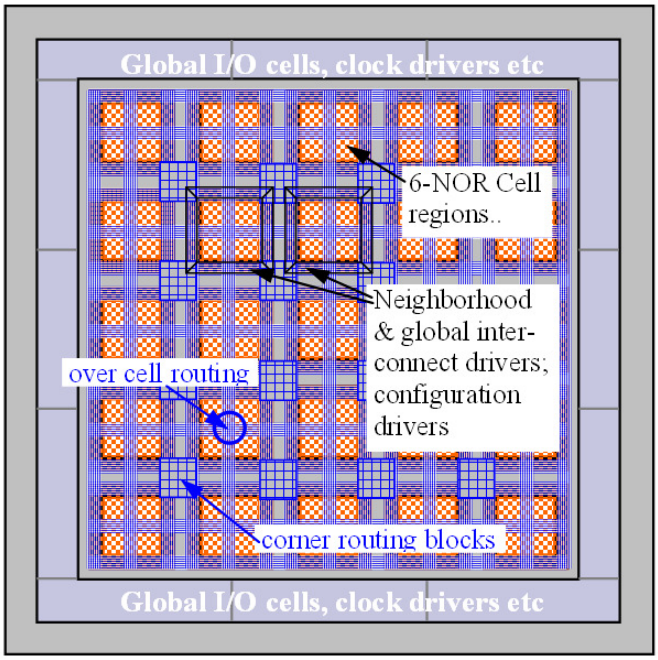


Figure 53. Generic floorplan of a reconfigurable fabric.

A Proposed Reconfigurable Fabric

The generic floorplan of Figure 53 illustrates one way that this array could be combined into an overall reconfigurable computing fabric. This proposal combines ideas from conventional FPGAs as well as structured ASIC systems (cf. Figure 18). The reconfigurable cells of Figure 49 are arranged in regions that are surrounded by an area containing interconnect drivers linking to both local and global lines as well as the configuration infrastructure (decoders, drivers, level shift etc.). The interconnect lines would route over the top of the cells and connect to global I/O cells on the periphery of the chip. Of course, there are many alternatives to this structure but a full exploration of these is outside the scope of this work. A major consideration here is that imposing any sort of hierarchical structure on the fabric moves it away from the main objective of this research, which has been to investigate the application of homogeneous reconfigurable arrays to computer architecture. Thus, the remainder of this thesis will focus just on the performance of the cells within a region.

3.3.3 Logic and Interconnect Mapping

A primary characteristic of this organization is that there is little intrinsic difference between logic and routing and each cell can be used for both simultaneously. A range of options available to merge logic fanin, fanout and routing are described in general terms below.

Logic Fanin and Fanout

The basic layout of each cell comprises six programmable 6-NOR gates. Input lines that are not part of a particular logic mapping on a 6-NOR are set to “don't-care” on that gate by shifting the effective switching threshold seen at that input to greater than V_{DD}. In the example of Figure 54, the nine inputs A…I are partitioned across cells 1, 3 and 5. Cell 1 creates the logic function G+H+I (normal threshold for these three inputs, high $|V_{TH}|$ for the remaining three). The term is then routed across cell 2 and 3 to merge with the terms A+B+C and D+E+F generated by cells 3 and 5 respectively. The resulting function appears on line ‘Y’ at the output of cell 4 and is then transferred to cell 9. In a similar manner, this term is distributed horizontally to cells 8 and 11

where it is combined with inputs routed through 2, 6, 7 and 11. The partial logic terms developed in cells 8 and 10 are then transferred via 12 and 14 to be recombined in cell 13 [6].

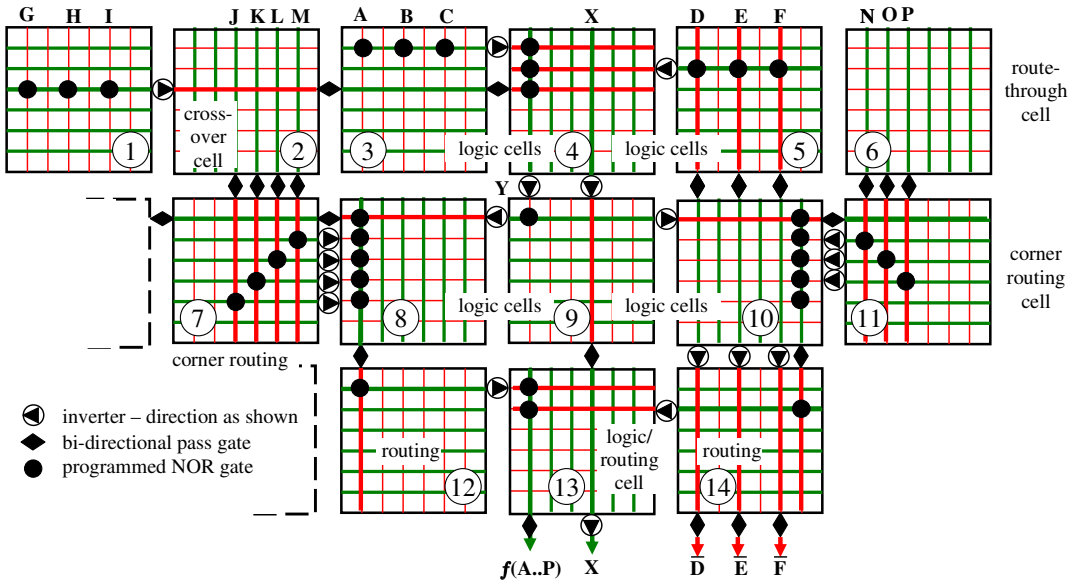


Figure 54. Example logic cell and interconnect topologies

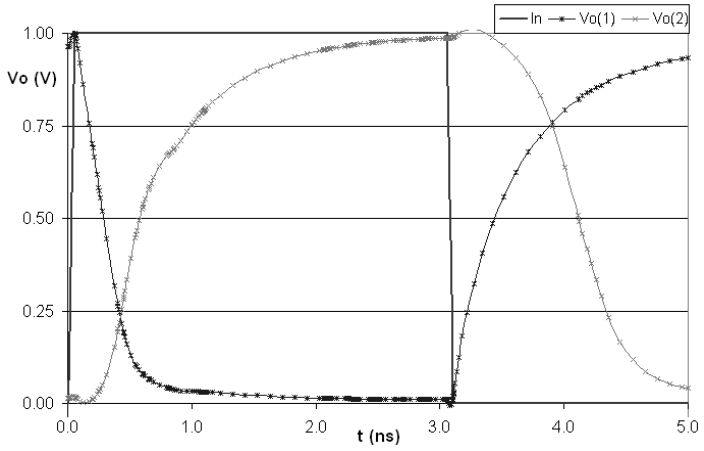


Figure 55. Simulated transient response of a single 6-NOR pair V_{DD}=1.0V; $(W/L)_P$:$(W/L)_N$ = 3:1.

Figure 55 shows the basic transient response of two consecutive 6-NOR cells for the proposed TB-SBSOI device (i.e., with the I-V characteristics given in Figure 42). Here, $V_o(1)$ and $V_o(2)$ are the outputs of the first and second cells respectively. Both cells are biased with a single gate

[6] Note that this example is illustrative only and is not intended to represent any particular logic function

on such that each cell forms a simple inverter function, representing the worst case falling edge delay configuration for this NOR organization.

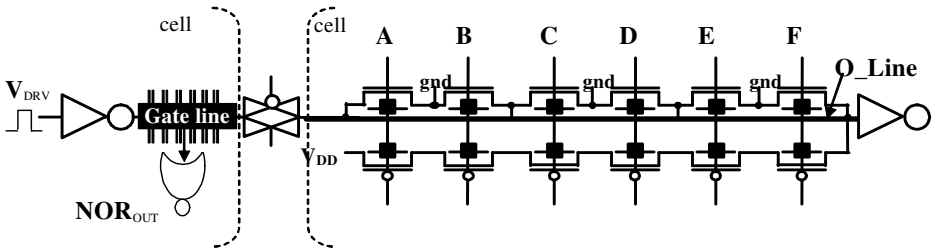
Routing

With some restrictions, signal routing can occur simultaneously in the two orthogonal directions within each cell. Firstly, as the gate lines traverse the entire cell, all input lines automatically carry their respective signal across the cell so that they become available to be connected to an adjacent cell via pass transistors. The partial directionality imposed by the inverting interface gates means that while the gate line drivers can be configured as either inverting or non-inverting, the output connections must be non-inverting (i.e., via the pass-gates). This is called *type 1* interconnect.

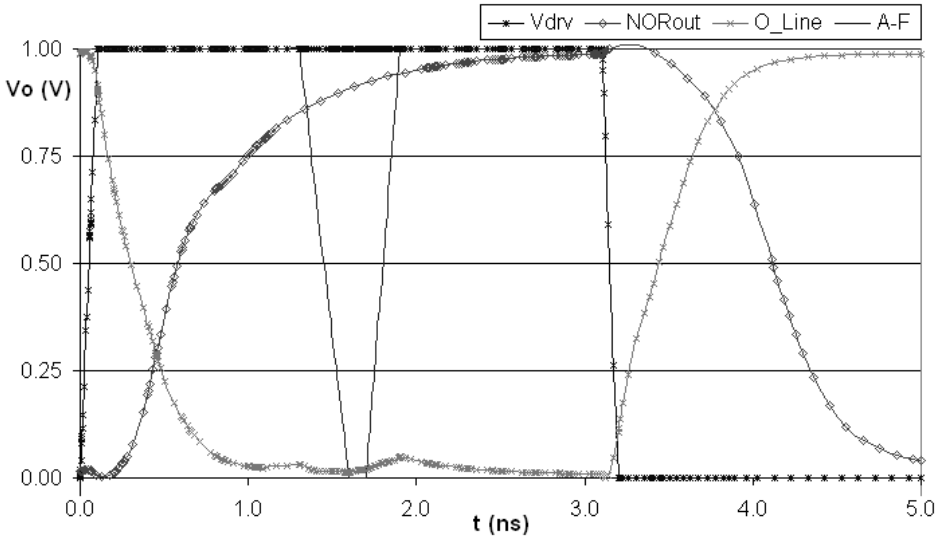
At the same time, any output line not involved in logic, for which all transistors are biased off (i.e. high $|V_{TH}|$) can also be used to route signals across the cell. This is called *type 2* interconnect. Again, the partial directionality of the interface gates requires the input to be non-inverting whereas the output connections can be either. Longer interconnect lines can be formed from alternating sequences of type 1 and 2 cells. These form part of the behavioral abstraction supporting the final architectural-level simulations presented in Chapter 5.

In Figure 56, a simulation of an interconnect line is shown using the thin-body (T_{SI} = 5nm) UFSOI transistor models (see Figure 42), compared with the performance of a single NOR cell driving another identical cell (from Figure 55). The line is assumed to be driven from an adjacent cell as illustrated in Figure 56a. In this example, the inputs A…F have been set to their high-threshold condition (V_{BGP} = +2,V_{BGN} = -1) and the output line (O_line, Figure 56a) is driven low between T ≈ 1 and 3nS (Figure 56b). The input lines A…F are driven low for approximately 250pS starting at 1.5nS. It can be seen that the output impedance of the driving inverter dominates and its output swings (low) to almost full rail, independent of the logic levels on the primary inputs and with negligible crosstalk from the input transitions. Cell 4 in Figure 54 represents an example of how this would be applied. In this particular case, the cell has already been partially

used to create a set of intermediate logic terms and is then used to transfer the signal 'X' vertically to cell 9, from where it continues through to cell 13.



(a)



(b)

	NOR Logic	Interconnect
T_{PHL}	980ps	310ps
T_{PLH}	524ps	300ps
T_P	752ps	305ps

(c)

Figure 56. Interconnect signals compared to basic NOR operation.
(a) Interconnect Circuit; (b) Interconnection waveforms. (c) Propagation delay.

The propagation delays for these interconnection and logic signals (Figure 56c) show that each segment of interconnect (O-Line) would add in the range of 13% to 18% to the basic logic delay. As the interconnection line capacitance is small in these locally connected cells, this is due simply to the ratio of the load capacitances: six gate loads in the case of a standard logic cell compared with a combination of a metal interconnection line and seven drain connections in the pass-

through connections. In Chapter 5 a worse-case lower bound of about 20% will be used for the relative delay between interconnect and logic.

Finally, the logic in each cell can be set up to be equivalent to a 6x6 universal routing block. These corner routing cells (e.g., cells 7 and 11 of Figure 54) may be inverting or non-inverting (or more correctly, twice inverting) and can be cascaded in two directions. These are termed *type 3* interconnect although they are essentially indistinguishable from logic cells. The overall result is a flexible organization that supports fine-grained tradeoffs between logic and routing. Unfortunately, this level of flexibility is also likely to make the task of automatic place and route extremely difficult and it is therefore unlikely that every individual cell would be able to be optimally configured at all times.

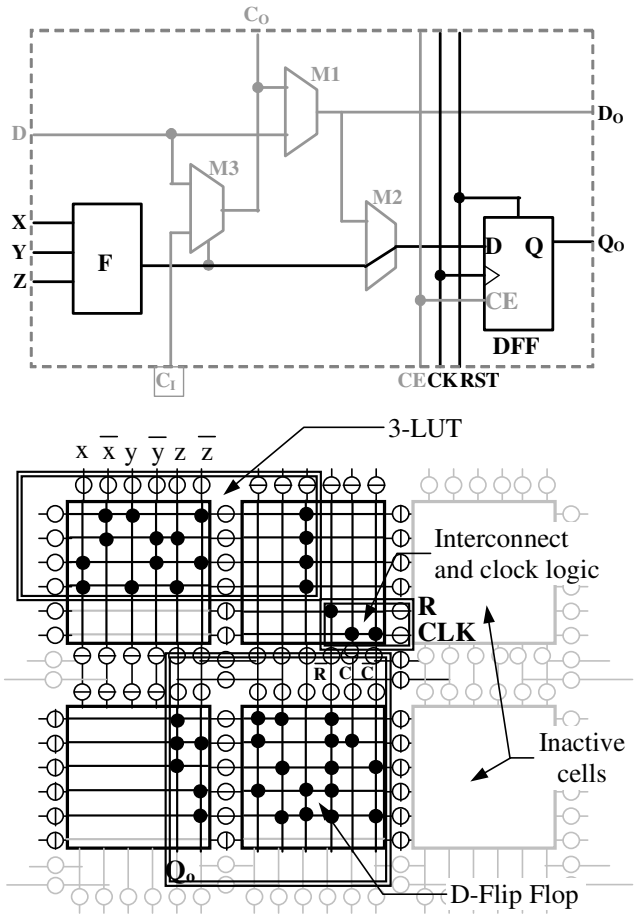


Figure 57. A configured logic cell forming a 3-LUT and Flip-Flop.

3.3.4 Combinational/Sequential Logic Mapping

In Figure 57, one functional pathway in a typical FPGA has been implemented to illustrate how the logic mapping in the proposed scheme compares to that of a conventional FPGA. The filled dots in this figure represent p-n transistor pairs that have been enabled. The remainder is configured off. Four of the NOR-cells form a 3-LUT (2 cells) plus an edge-triggered D-type flip-flop (2 cells). As the right-most LUT cell uses only four NOR-term lines, the remainder of that cell is used to bring in the reset line connection and to develop the complementary clock signals.

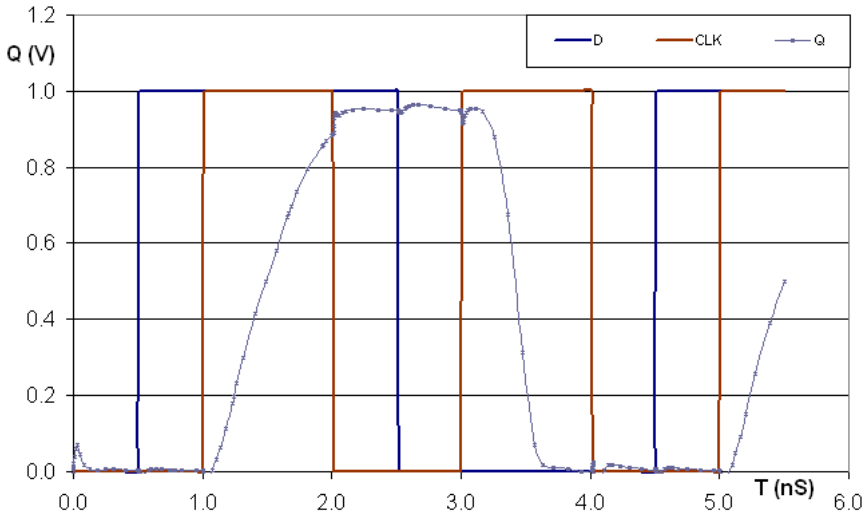


Figure 58. Simulated D-type FF operation.

The edge-triggered D-type equations were developed using conventional asynchronous circuit techniques and the simulation results for D-type FF (Figure 58) were derived using the UFSOI SPICE models. The two state variables, mapped onto the local feedback lines are given by:

$$\begin{aligned} \overline{q_1} &= \overline{\overline{(\overline{q_1}+clk)}+\overline{(\overline{q_0}+\overline{clk})}+\overline{(\overline{q_1}+\overline{q_0})}} \\ \overline{q_0} &= \overline{\overline{(\overline{D}+clk)}+\overline{(\overline{q_0}+\overline{clk})}+\overline{(\overline{D}+\overline{q_0})}} \end{aligned} \tag{3.7}$$

and $\overline{q_1}$ maps directly to the output (Q). Other state machines of equivalent complexity, for example level-triggered (transparent) latches, can be built using the same number of cells. The SPICE input decks for both Figure 56 and Figure 58 can be found in Appendix B.

Figure 59 extends this slightly to illustrate the formation of a cascaded data path. The sharing of terms between the sum and carry allows a full adder to be implemented in just five terms. Thus, if the two horizontal connections between adjacent cells are used to transfer the ripple carry between bits of the adder, each bit will fit within one 6-NOR cell pair. These layouts are reminiscent of the sort derived from an ASIC module generator [295] or a "sea-of-gates" gate-array implementation. Conversely, in a standard cell ASIC environment, it would make little sense to decompose to the level of NOR gates , as it would be likely to result in a very inefficient structure dominated by interconnect.

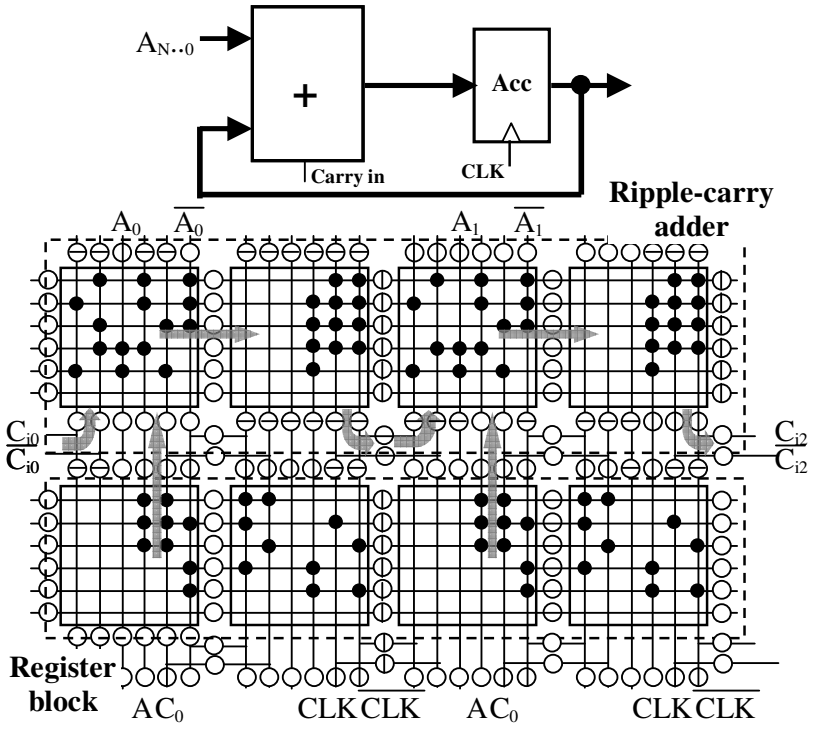


Figure 59. Simple data path example with cascaded cells (2 bits shown)

Of course, specialized support hardware such as fast carry chains will not be available in this organization. However, it should be noted that conventional high-level synthesis tools (VHDL, for example) take a generalized approach and tend not to use these specialized structures, except perhaps as part of a device-specific module generation process. Further, there is already evidence (e.g. [296]) that functionality of this sort will be less effective when interconnection delay domi-

nates and simpler techniques such as serial arithmetic [297] may exhibit similar levels of performance.

3.3.5 Registered or Non-Registered Logic?

Since research such as reported in [284] showed how the presence of a flip-flop in a logic block would reduce overall chip area, most fine-grained reconfigurable systems include at least one latch and/or flip-flop as part of their logic cell structure. A recent exception is the "ProASICPlus" Flash-based family from Actel [131], that combines a "sea-of-tiles" core architecture with a four-level interconnect hierarchy. Each tile in the core has four inputs and can be configured as either any three-input logic function (except, surprisingly, XOR) or a state element comprising a latch or D-type with set or clear [298].

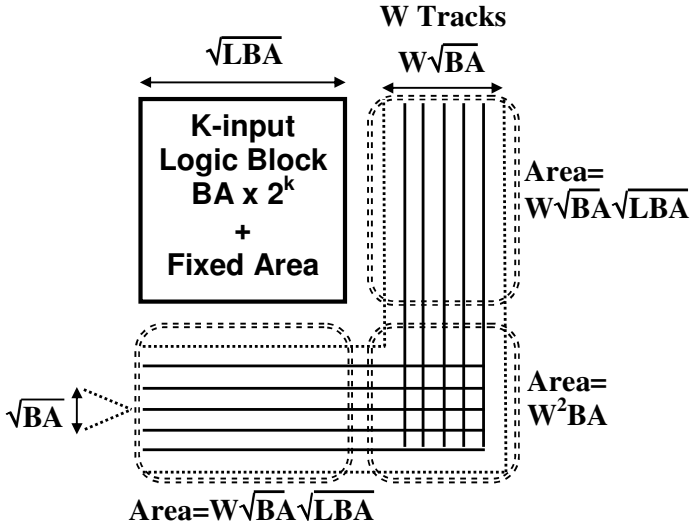


Figure 60. Interconnect area model of Rose *et al.* adapted from [284].

The original experiments in [284] were based on the area model shown in Figure 60 for which the routing area is given by $W^2BA + 2WS\sqrt{BA}$, where BA is the area required to store one bit in the given technology (bit-area), W is the width of the adjacent routing channel and $S = \sqrt{\text{Logic Block Area}}$. As the dominant term here is W^2BA, the routing area will be approximately proportional to W^2 (in units of bit-area, proportional to technology). Rose *et al.* found that, by removing the register and its associated multiplexers, the logic block size could be sub-

stantially reduced (by up to a factor of 2.5). However, this required an increase in the overall number of logic blocks in a typical design (between 1.4 and 2.3 times) that had the "flow-on" effect of increasing the required wire length and therefore the number of interconnect lines (W) per channel.

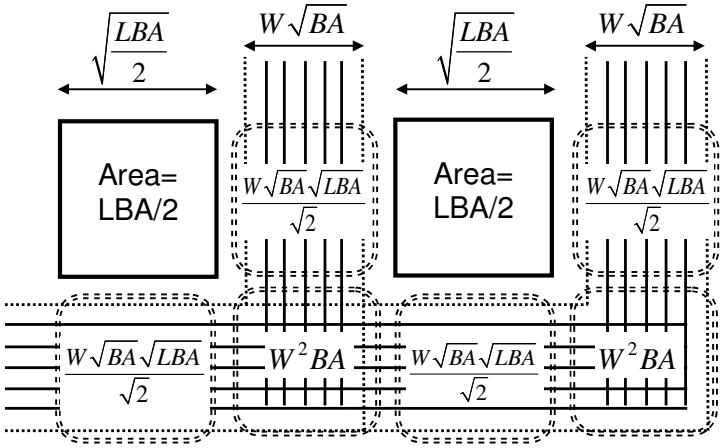


Figure 61. Modified interconnect area model based on the assumption of fixed routing width.

This is illustrated in Figure 61 where the logic block area has been reduced to half its original size (shown as LBA/2, where LBA is the original logic block area) then duplicated. The interconnect width becomes $W_{ND}\sqrt{BA}$, where W_{ND} (>W) is the width with no D-type. As a result, the routing area per logic block becomes $W_{ND}^2 BA + W_{ND} S\sqrt{BA}/\sqrt{2}$, consistent with [284] where it was found that that the overall chip area increased by a factor of up to almost three for some circuits.

It can be seen that this result relies implicitly on the assumption of an "island–style" organization in which each logic block is surrounded by interconnect so that the total number of blocks increases as the logic block size is reduced (e.g., by removing the flip-flop). In turn, the increased number of logic blocks results in corresponding increases in both channel width and the number of channels. Therefore, the overall routing area goes up. Removing this assumption significantly weakens the case for maintaining the flip-flop. For example, the ProASICPlus® FPGA mentioned above employs a partial island style layout in which clusters of tiles are surrounded by (global) long-line channels while local interconnect routes point-to-point over the top of adjacent cells.

None of the routing paths dominate and the device designers have opted not to include a state element but to allow it to be configured from a single logic tile.

Table 9 Area Comparison for LGSynth93 circuits

BLE mapping data derived from [299].

Circuit	Component			Area		Area PMA / Area BLE
	LUT Only	Latch Only	Both Used	BLE	PMA	
Alu4.net	1522	-	-	38050	24532	0.6
Apex4.net	1262	-	-	31550	19457	0.6
Ex5p.net	1064	-	-	26600	15758	0.6
Misex3.net	1397	-	-	34925	22063	0.6
Apex2.net	1878	-	-	46950	31833	0.7
Des.net	1591	-	-	39775	25917	0.7
Seq.net	1750	-	-	43750	29165	0.7
Pdc.net	4575	-	-	114375	96463	0.8
Spla.net	3690	-	-	92250	73757	0.8
Bigkey.net	1482	-	224	42650	31969	0.7
S298.net	1923	1	7	48275	33067	0.7
Diffeq.net	1120	3	374	37425	30024	0.8
tseng.net	662	1	384	26175	21109	0.8
Dma.net	8350	2	31	209575	206658	1.0
Elliptic.net	2482	2	1120	90100	93863	1.0
Frisc.net	2670	17	869	88900	87540	1.0
S38417.net	4943	310	1153	160150	171272	1.1

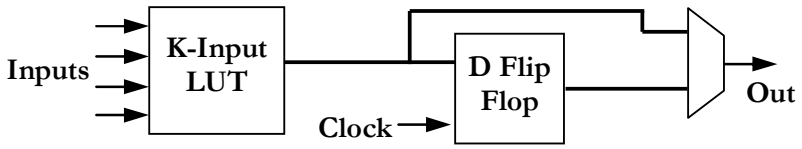


Figure 62. Basic Logic Element (BLE)

To more clearly compare the fixed-channel island-style organization with the proposed "poly-morphic" array, an analysis was performed on the netlists of a number of circuits drawn from the LGSynth93 archive used as benchmarks for the place-and-route challenge [299]. Each of these circuits had already been synthesized onto a standard basic logic element (BLE) comprising a 4-LUT and a flip flop (Figure 62). The component values in Table 9 represent the number of nets for that circuit that have been mapped a particular basic block configuration i.e., LUT only (by-passing the D flip-flop), latch only (bypassing the LUT) and both LUT and flip-flop simultane-

ously. The area results compare the basic logic element (BLE) of Figure 62 with the "polymorphic" array (PMA) cell (Figure 49). These area comparisons were based on a number of conservative assumptions:

1. in the "polymorphic" cell, a 4-LUT and D-type can each be configured using a pair of adjacent cells then interconnected by abutment, incurring no additional local routing overhead. As they are approximately the same complexity, one cell-pair and a BLE are assumed to occupy equal areas in the same technology.

2. the impact of the configuration mechanisms were assumed to be approximately the same in each case so that it could be ignored for this analysis. Conventional FPGA devices use sparse encodings, typically using at least a factor of 2–4 more configuration bits than necessary [19], so are likely to exhibit similar a number of configuration bits as the PMA for an equivalent function.

3. the width of the FPGA routing channels was estimated to be four times the original logic block width based on data from [284] with the LUT input size, K, in the range 3 to 6. The overall cost of using a BLE was therefore fixed at 25 units. This is an under-estimate. For large FPGA devices the cost will be significantly more than this.

4. the routing overhead for the "polymorphic" layout was estimated using the stochastic wire length model of [300], based on Rent's Rule [332], which models the interconnect density within a recursively partitioned netlist as a power law function of the form $R = kN^P$. In this case, the Rent exponent (p) was set to 0.8. This is an overestimate, more typical of random logic blocks. The true parameter will be smaller for all of these circuits. The wire length estimates (in units of cell pitches) were then scaled up by $\sqrt{2}$ to allow for the rectangular routing constraints then doubled again to account for placement inefficiencies and routing congestion. The overall cost of using a polymorphic cell-pair is therefore given by $(2\sqrt{2}\overline{R}+1)$ units, where $\overline{R}$ is the average in-

terconnect length. Routing is confined to the adjacent regions between merged cells in a similar way to a layout based around generated IP block within a standard ASIC.

Unlike a conventional FPGA, the array proposed here is not constrained by a fixed channel width set by the number of pins on a fixed logic block. The ability to configure routing only where necessary and to collapse low fanout cells into blocks connected by abutment, results in area ratios of one or less for all of the smaller benchmarks. The ratio of the areas reaches (and slightly exceeds) unity in the larger circuits with longer average wire lengths and in the case where a greater number of flip-flops increases the overall number of cells compared to the FPGA. Even though these results are only approximate, it can be seen that the proposed array is likely to be no worse that the fixed channel FPGA and will typically be better due to its greater flexibility in trading logic for interconnect area.

The general trend of these results is reinforced in Table 10 that shows a similar analysis of a number of arithmetic circuits written in VHDL, originally targeting a FPGA. In this case, wire lengths are estimated using a Rent exponent of 0.4 to reflect the regularity of these arithmetic circuits and their low fanout. These arithmetic circuits can therefore be readily grouped to form larger blocks, mainly by abutment and with a small ratio of internal routing. Table 10 shows that under the same assumptions as above this regularity and locality will support compact layouts, up to three times more compact than on conventional organizations with fixed routing.

Table 10 Area results for arithmetic circuit mappings

Circuit	LUTs	F/Fs	Area		Area PMA / Area BLE
			BLE	PMA	
64-bit ripple adder	165	0	4125	1186	0.3
32x32 array multiplier	2439	0	60975	22219	0.4
16x16 add/shift multiplier	85	51	3400	915	0.3
32x32 Booth multiplier	263	103	9150	2752	0.3

It should be noted that the objective of these experiments has not been to prove that the proposed array is more compact, *per se,* than existing organizations but to explore the question of whether complex heterogeneous circuits can be efficiently implemented using this flat, undifferentiated

device array. Because logic and interconnect are interchangeable, such that functional "clusters" may be formed as required, the array exhibits some of the characteristics of an ASIC formed from generated IP blocks. These results show that the array can be expected to support circuit mappings that are at least as good as current FPGA systems and can therefore form the basis of a useful reconfigurable platform.

3.4 Summary

In Chapter 2, it was determined that design effort, manufacturability, reliability, variability and power tend to be moving future architectures in the direction of simple, low-power, reconfigurable, locally-connected hardware meshes. In this chapter, a processing fabric based on thin-body double-gate Schottky barrier MOSFET devices has been proposed and analysed. These devices were chosen as representative of an end-of-roadmap technology with (potentially) greatly simplified manufacturing. Examples of single-gate Schottky barrier devices have already been built, as have a number of planar double-gate configurations. It is likely that these technologies could be combined to form a double-gate Schottky barrier topology before the predicted end of the CMOS roadmap.

A key problem with these devices is their very low I_D(sat), which severely limits their performance in anything but local interconnect organizations. This would seem, *prima face*, to rule them out as suitable candidates for future high-performance architectures. On the other hand, long channel SB devices have already been manufactured on experimental lines with I_{ON}/I_{OFF} ratios of 10^5 and, in general, the presence of the Schottky barrier results in much better short-channel behavior than is typically the case in conventional technology. As a result, these devices offer the promise of improved subthreshold performance and, in particular, low subthreshold leakage power. These characteristics will be critical for future "ubiquitous" low-power and hand-held systems.

The main objective of this chapter has been to analyze the primary characteristics of a proposed reconfigurable mesh. The operation of this fine-grained reconfigurable platform is based on a

novel characteristic of thin-body, double gate technology i.e., that the threshold voltage seen at one gate of the double gate transistor may be substantially shifted by altering the bias on its second gate. Varying the threshold voltage can greatly reduce the overall subthreshold power of the platform by uncoupling the conflicting requirements of high performance and low standby power. In addition, it allows the functionality of the reconfigurable cells to be established. The resulting organization might be described as "polymorphic" as each component cell is capable of being configured as logic, interconnect, or an arbitrary combination of both.

The envisaged application for the array encompasses the design space that FPGAs and, to a lesser extent, structured ASIC devices currently target. All of these organizations include other specialized functions such as memory and configurable I/O. There is certainly no current trend indicating that memories will disappear from VLSI circuits, although the issue of interconnect "locality" will make it increasingly costly to access memory in its present form. It is more likely that memory will continue to be organized into both distributed and block structures, just as it is in current FPGA systems and that distributed memory will become an increasingly important function in future architectures.

In the following chapter, an analytic model is developed that relates these issues of performance, area and power (primarily dynamic and subthreshold). This provides a framework within which the performance, area and power characteristics of the reconfigurable array can be analysed to determine its ultimate scalability.

Chapter 4. An Area–Power–Performance Model for CMOS

"Those who have knowledge, don't predict. Those who predict, don't have knowledge."

Lao Tzu, 6th Century BC Chinese Poet
http://www.met.rdg.ac.uk/cag/forecasting/quotes.html

The previous chapter proposed and analysed a number of device and circuit-level issues for a mesh-connected, fine-grained reconfigurable architecture based on a double-gate technology that offers the potential to be scalable to sub-10nm gate lengths. This chapter now examines the question of whether it is possible to predict at an abstract level the ultimate scalability of architectures mapped to this reconfigurable system. A premise here is that, as on-chip device density increases, there will be little choice but to manage power consumption by exploiting parallelism. Power (or energy)-delay tradeoffs will have to be biased in favor of power (energy) and the consequent loss of performance made up by architectural and circuit design level innovations. In this way, appropriate performance can be maintained while constraining power and managing heat dissipation. This trend is already emerging in the commercial microprocessor domain (e.g. [8, 301, 302]).

In this chapter, a model is proposed and analysed that relates area, performance and power in future CMOS devices. As both performance and power depend implicitly on the relationship between supply and threshold voltages, the model focuses on the optimization of these two variables. It is assumed here that both may be adjusted more-or-less at will in the double-gate reconfigurable circuits proposed previously in Chapter 3. Although the model was developed in order to estimate the scalability of that particular reconfigurable fabric, it is also likely to be more generally applicable to nanoscale digital logic circuits wherever the relationship between supply and threshold voltages can be independently controlled.

Some existing architectural metrics were outlined in Section 2.7.3 including, for example, hardware intensity (HI). Although these may be used to evaluate the impact of circuit changes on architecture, no current analytic framework exists that directly describes the tradeoffs between area and power in parallel organizations. The time-space-complexity observations of Flynn and Hung are useful here, but do not constitute a complete model. The model developed in this chapter builds on that work [169, 170, 303] as well as early work in VLSI complexity analysis (e.g., [304]) to predict the evolution of power/energy and area in digital logic systems both within a given technology node, and as technology scales to the end of the roadmap. Following this, in Chapter 5, it will be used to analyze the reconfigurable fabric described in Chapter 3. An early version of this chapter has been published in [305].

4.1 Architecture Level Area–Power–Delay Tradeoffs

As outlined previously in Section 2.7, the link between area and delay (and therefore power) operates at four primary levels: device, circuit, micro-architecture and architecture. Of these, it is the higher architectural levels that offer the greatest opportunities to manage power-performance tradeoffs. In general terms, the power-performance nexus is constrained by the cubic relationship between dynamic power and propagation delay, and the exponential nature of subthreshold current vs. threshold voltage, both intrinsic to MOS. As a result, it will become increasingly difficult to balance performance and static power in large systems.

The solution for any particular digital design can be considered to be situated somewhere in a space defined by the primary constraints of area, time (i.e., performance) and power. This solution space can be conceptualized as a theoretical 3D volume that relates the three. As an example, the space shown in Figure 63 has been adapted from [169] and shows the relationships between area and performance (time):

$$AT^{\sigma} = K_1 \tag{4.1}$$

and between power and performance:

$$PT^3 = K_2 \tag{4.2}$$

where both K_1 and K_2 are constants. Called the *par* curve in [303], (4.1) is a result derived from the analysis of the complexity of VLSI circuits [304]. This early work in VLSI theory (e.g., [306-309]) found that AT^σ typically represents a strong lower bound on the *best* circuits that may be constructed for a particular function, based on the flow of information through that circuit. An example is the form derived for binary multiplication in [307]: $\left(\frac{A}{A_0}\right)\left(\frac{T}{T_0}\right) \geq n^{(1+\sigma)}$, $0 \leq \sigma \leq 1$, where both A_0 and T_0 depend on technology but are independent of the operand size, n. This equation may also be interpreted as relating area scaling (A/A_0) to changes in performance (T/T_0) for an optimal design. An example of this interpretation is shown in Figure 64, based on five types of 32-bit adder written in VHDL and mapped to a Spartan 2S100 series FPGA. Area (in arbitrary CLB units in this case) is plotted against the worse-case delay reported by the synthesis tool for both the area-optimized and delay-optimized cases. The two trend lines represent the functions $T = 100A^{-0.56}$ (i.e. $\sigma \approx 1.82$) and $T = 90A^{-0.8}$ (i.e. $\sigma \approx 1.25$) that have been fitted to the area and delay optimized points respectively using a least-squares approximation are intended to illustrate the general trend of the five data points.

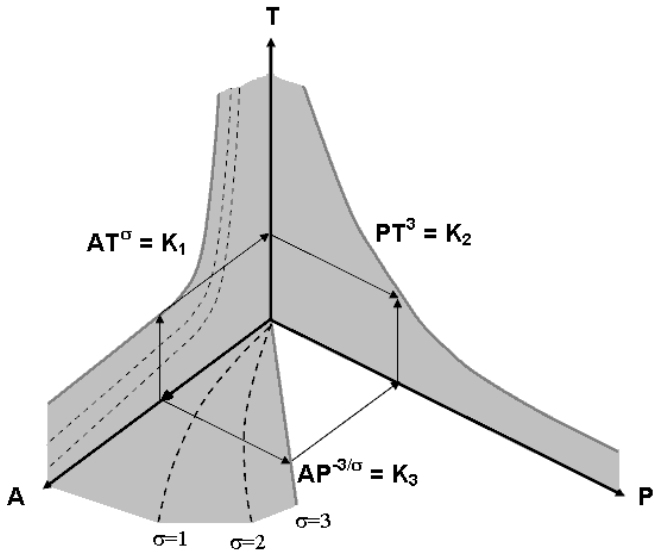


Figure 63. A hypothetical 3-dimensional Area-Time-Power space adapted from [169].

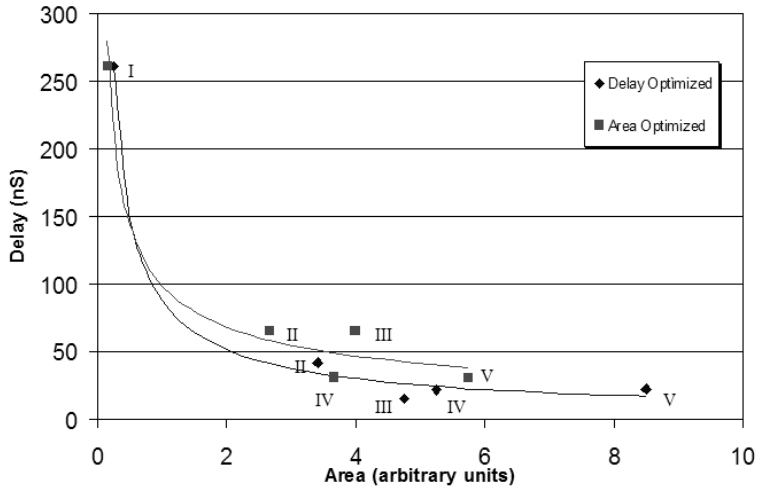


Figure 64. Area vs. Delay for five 32-bit adder styles
I – serial adder; II – ripple-carry; III - carry-save; IV – carry-select; V – carry lookahead

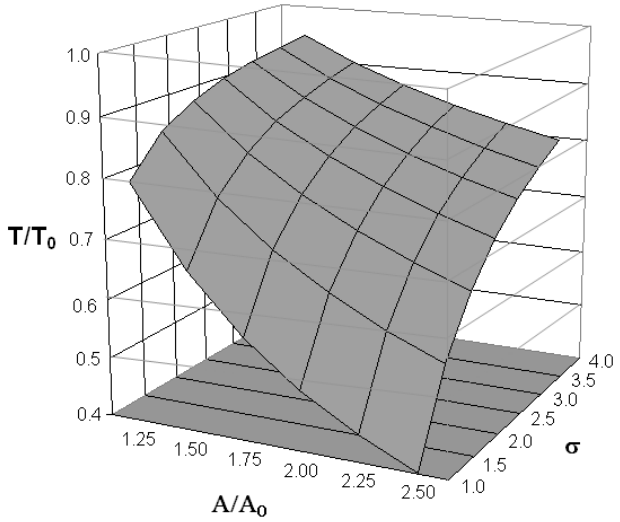


Figure 65. T/T_0 vs. $(A/A_0)^{-1/\sigma}$ for $1 \le \sigma \le 4$.

This observation describes the "classic" area-time tradeoff: it is usually possible to increase circuit area in order to reduce overall delay. It is this interpretation that guides the following analysis. From the area-performance relationship described by (4.1) over a range of σ (Figure 65), it can be seen that at successive (increasing) values of σ, the T vs. A curve is flatter, implying that there will be less architectural "return" (i.e. change in performance) resulting from a given change in area.

Strictly, (4.2) is an approximation and refers only to the dynamic power term ($P_{DYN} = aFC_L V_{DD}^2$) under a limited set of assumptions. If the threshold voltage, V_{TH}, is fixed at about 20–30% of supply, the operating frequency becomes an almost linear function of supply [103] over the range of interest (i.e., ~0.5V ≤ V_{DD} ≤ ~1V) and thus at the same feature size (C_L = constant), P_{DYN} becomes proportional to F^3 leading to the result in (4.2).

The curve on the AP axis can be derived by substituting (4.1) into (4.2) so that $T \propto P^{-1/3}$ and $T^{\sigma} \propto P^{-\sigma/3}$. Thus:

$$AP^{-3/\sigma} = K_3. \tag{4.3}$$

It can be seen from (4.1), as well as in Figure 63, that an increase in area can be traded off against a decrease in clock period $T \propto A^{-1/\sigma}$. As a result (from (4.1) and (4.2)), power will increase by $A^{3/\sigma}$. An example of this trajectory is shown by the arrows on Figure 63 and represents the so-called "power wall", outlined previously in Chapter 2, that threatens to stall the deployment of large numbers of devices on a single chip.

4.2 Scaling with Constant Performance

A similar analysis may be used where the objective is to achieve a constant overall performance as area increases. It is expected that by exploiting parallelism (thus increasing area) the required operating frequency might be reduced, in turn allowing a reduction in supply voltage and therefore in power consumption. This is already part of the design considerations for low-power microprocessors as well as real time processors and DSP [174], but is likely to become a global problem as designers increasingly hit the "power-wall". The particular case of constant total power at constant completion time with increasing area may be treated as a boundary of the desired solution space. Inside that boundary, power will be a decreasing function of area whereas outside this region, power increases with area.

As the completion time is assumed to be fixed and T ∝ 1/F, the frequency will be given by:

$$F \propto A^{-1/\sigma}. \tag{4.4}$$

Considering only dynamic power for the moment, and using the F ∝ V approximation, the power term $P_{DYN} = aFC_LV_{DD}^2$ reduces to $P_{DYN} \propto C_LF^3$. In this case, C_L is the *total* capacitance switched per cycle and is roughly proportional to area (or, more strictly, the number of transistors, N). Substituting $C_L \propto A \propto F^{-\sigma}$ results in:

$$PF^{\sigma-3} = K_2. \tag{4.5}$$

Finally, substituting (4.4) into (4.5) gives:

$$PA^{-(\sigma-3)/\sigma} = K_3. \tag{4.6}$$

It can be seen from (4.4) to (4.6) that under the two assumptions of fixed completion time and F ∝ V, frequency (and as a consequence dynamic power) may become reducing functions of area. The value of σ describes the effectiveness of the technique for a particular circuit and/or algorithm. In the following sections, the basic relationship given by (4.6) is first analysed in the context of the major sources of power consumption in CMOS. The F ∝ V assumption is then removed and a more general form is developed for the dynamic and subthreshold power components.

4.3 Modeling Power vs. Area in CMOS

Power consumption in CMOS arises primarily from four main sources (Figure 66):

1. Dynamic power (P_{DYN}), a function of capacitance (C), voltage (V), the activity factor (a), and switching frequency (F) such that $P = a\,FCV^2$.

2. Short circuit switching current ($P_{SS} = I_{SS}.V_{SW}$) with I_{SS} being a function of frequency and transistor size.

3. Subthreshold leakage: $P_{LEAKAGE} \propto I_{OFF} \cdot V_{DD}$

4. Gate current and GIDL, proportional to supply/logic level and transistor size.

Of these, the dynamic power terms (P_{DYN} and P_{SS}) are primarily a function of the switching frequency and capacitance (fanout and interconnect). One way to reduce dynamic power is to reduce the number of devices that switch per cycle, for example by using asynchronous circuits [310], [311], that eliminate the global clock (and its associated global wire), and are based on local communication and synchronization. An alternative approach is to allow clock frequency to decrease and recover the resulting loss of performance by exploiting parallelism.

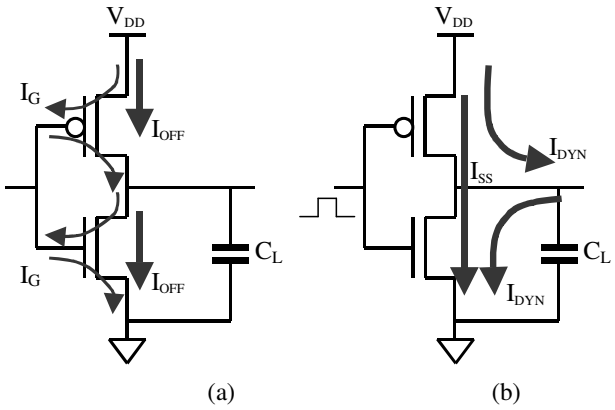


Figure 66. (a) Static and (b) Dynamic power loss mechanisms in CMOS

The remaining two terms (P_{SUB} and P_G) represent a static power loss that is largely unaffected by either of these techniques. Gate current and leakage current strongly interact [312] and since their total is a function not only of technology (e.g. oxide thickness, dielectric etc.) but the average gate voltage during operation, static leakage depends strongly on logic state. Static power is already recognized as a major constraint to future device scaling in CMOS [172]. The following sections explore these four different sources of power consumption and relate them to the area used to implement the circuit.

4.3.1 Subthreshold Leakage

Subthreshold leakage current arises mainly due to diffusion between the source and drain when the channel is in weak inversion. In bulk CMOS, there is a small contribution from tunneling through the reverse-biased diode junction at the drain/substrate junction, but it will be negligible

in future low-voltage SOI technology [172]. Direct source-drain tunneling will become important only at gate lengths of less than 10nm, projected to arise towards the end of the roadmap.

An analysis of the subthreshold leakage power can start with the BSIM3V3 transistor model for subthreshold drain current: [76]

$$I_{SUB} = I_{SO}\left(1 - e^{\frac{-V_{DS}}{V_t}}\right) e^{\frac{V_{GS} - V_{TH} - V_{OFF}}{nV_t}} \tag{4.7}$$

where I_{SO} is a function of the transistor geometry (W/L) plus a number of process parameters and V_{OFF} represents a small offset from V_{TH} to the subthreshold region. The parameter n (≈1 to 2) is related to technology and is adjusted to fit the slope of the curve such that $S = 2.3nV_t$ empirically describes $\Delta V_{GS}/\Delta I_{SUB}$ in mV/decade.

The worst-case current when the gate is off (I_{OFF}) occurs at the point at which the gate voltage is zero and the voltage drop from the drain to the source is highest, i.e., $V_{GS} = 0$ and $V_{DS} = V_{DD}$. Under these conditions the first exponential term becomes $e^{(-V_{DD}/V_t)}$. Since V_t (= kT/q) will always be small compared to V_{DD}, this term tends to zero. Assuming that V_{OFF} is small and setting V_t to its room temperature value of 0.025V, I_{OFF} is given by:

$$I_{OFF} \propto \left(\frac{W}{L_g}\right) e^{-\left(\frac{40}{n}\right)V_{TH}}. \tag{4.8}$$

Where the transistor length (L_g) is fixed to the minimum allowed by the technology, $|I_{OFF}|$ at a particular circuit node is determined primarily by the width (W). For a given *fixed* threshold, the exponential term is constant, independent of area. Thus, total subthreshold current will be a linear function of the number of devices (N) and therefore of increasing area, i.e., $P_{SUB} = NI_{OFF}V_{DD} \propto A$. Clearly, in order for I_{OFF} to reduce with N, V_{TH} must *increase* as supply reduces, notwithstanding its effect on performance.

If V_{TH} and V_{DD} are coupled via a function of the form:

$$V_{TH} = a - bV_{DD}, \text{ where } a \text{ and } b \text{ are constants,} \tag{4.9}$$

then, the behavior of the second exponential term changes and it becomes possible to trade reduced subthreshold currents for performance in a controlled manner. Given a pair of scaling factors (*a* and *b*), the exponential term e^{-V_{TH}/nV_t} is transformed to $e^{-(a-bV_{DD})/nV_t}$ and therefore to a product of two terms: e^{-a/nV_t} i.e., a constant at a fixed temperature, and $e^{bV_{DD}}$.

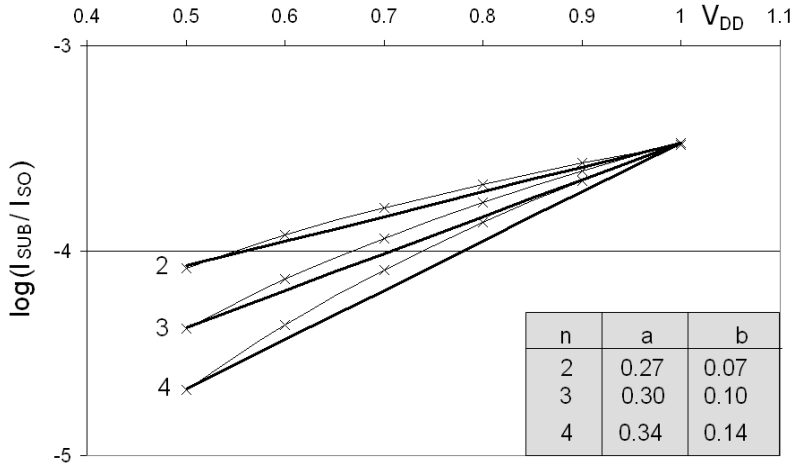


Figure 67. $I_{SUB}/I_{SO}=e^{-40a}e^{40bV_{DD}}$ (solid lines) and V^{n}_{DD} (dotted lines) for n=2–4. The inset shows values of a and b used to approximate V^{n}_{DD}.

As illustrated in Figure 67, values of a and b can always be chosen such that V_{DD}^{n} $(2 \le n \le 4)$ represents an upper bound on the subthreshold power down to $V_{DD} = 0.5$, the 2016-18 ITRS target for low-power SOC [11]. Thus, the subthreshold leakage current can be modeled as a simple power-law function of the form:

$$I_{OFF} \propto V_{DD}^{\eta} \tag{4.10}$$

For example, if *a* and *b* are set such that $I_{OFF} \propto V^{2}_{DD}$ then the overall subthreshold power becomes $\propto AV^{3}_{DD}$ and:

$$P_{SUB} \propto A^{(\sigma-3)/\sigma}, \tag{4.11}$$

which has the same form as (4.6). Thus, with careful management of the relationship between threshold and supply voltage subthreshold power can be made to be a reducing function of area, in this case for circuits that can achieve $\sigma \le 3$.

4.3.2 Saturation Drive Current

The analysis for saturation drive current is based on the general form of the equation developed for DSM [103], previously analysed in Section 2.5.1. The simplified saturation current relationship is given by:

$$I_D(sat) \approx WL_g^{-0.5}T_{OX}^{-0.8}(V - V_{TH})^{\alpha} \tag{4.12}$$

where W is the gate width, L_g its physical gate length, T_{OX} the gate oxide thickness and $V = V_G \approx V_{DD}$. As was identified in Section 2.5.1, it can be assumed that (4.12) will remain valid to the end of the roadmap, albeit with reducing values of α. Fixing α at 1.25, taking the average transistor width to be proportional to L_g and assuming that the average W/L ratios do not change through scaling, (4.12) becomes:

$$I_{D(sat)} \propto \sqrt{L_g}T_{OX}^{-0.8}(V - V_{TH})^{1.25}. \tag{4.13}$$

At a given technology node, and ignoring variability, the term $\sqrt{L_g}T_{OX}^{-0.8}$ will be a constant. Further, the ITRS predicts that the scaling of L_g and T_{OX} will continue to (non-monotonically) track one another such that $\sqrt{L_g}T_{OX}^{-0.8}$ may be considered to be approximately constant over the remaining nodes. This is illustrated in Figure 68, which plots $\sqrt{L_g}T_{OX}^{-0.8}$ vs. the physical gate length for both HP and LOP technology. The data source in this case is the ITRS 2006 update in which it is assumed that bulk CMOS will last until around 2010, and will be followed by two periods each where thin-body SOI and then DG-SOI will be the dominant technology, with the roadmap ending around 2020. The dotted lines in this figure show the mean values for both technologies and also that $\sqrt{L_g}T_{OX}^{-0.8}$ will be constrained to within about ±8% of the mean in both cases.

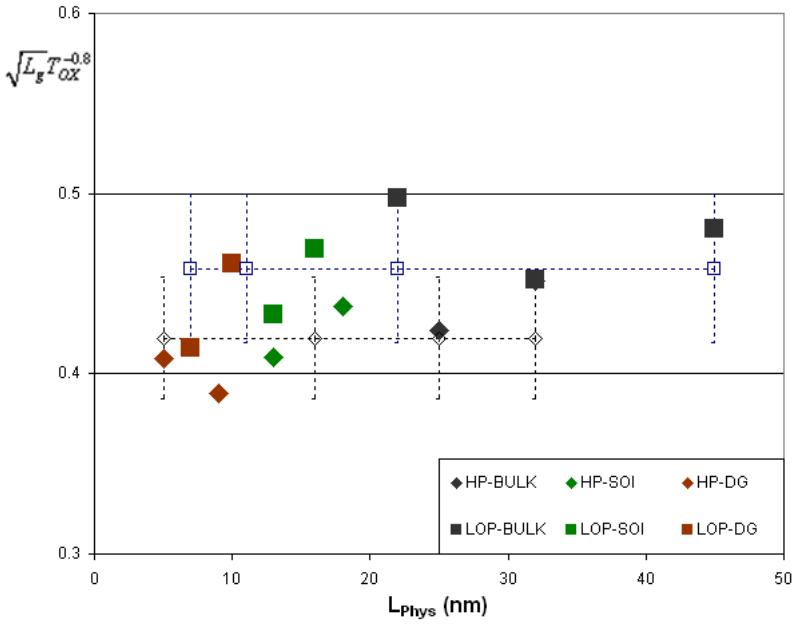


Figure 68. $\sqrt{L_g}T_{OX}^{-0.8}$ vs. L_{Phys} for some selected ITRS technologies.
The dotted lines show the mean and approximate spread (±8%) of the LOP and HP sets.

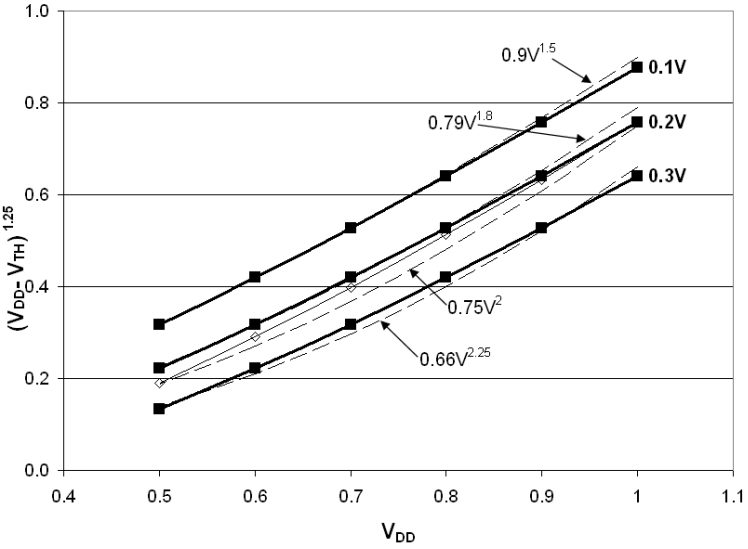


Figure 69. $(V_{DD} - V_{TH})^{1.25}$ vs. V_{DD} with V_{TH} = 0.1, 0.2 and 0.3V (filled squares).
Also shown (open diamonds) is the case where $V_{TH} = 0.27\text{-}0.07V_{DD}$, as in Figure 67.
The dashed lines represent a least-squares error fit to each curve.

Figure 69 plots $I_D(sat) \propto (V - V_{TH})^{1.25}$ vs. V_{DD} over the range $0.5 \le V_{DD} \le 1.0$ for various fixed V_{TH} as well as for the V_{TH} scaling case in the previous section that resulted in $I_{SUB}/I_{SO} \propto V_{DD}^2$. It can be seen that in each case a close approximation can be made of the form:

$$I_D(sat) \propto V^{\beta}. \tag{4.14}$$

This approximation, which can be compared to (4.10), is most valid where V_{TH} is small compared to the supply voltage or at larger V_{TH} where the relationship between the supply and threshold voltages is "tailored" to achieve it. It is obvious that the relationship between supply and V_{TH} affects both $I_D(sat)$ and I_{SUB}. For example, from Figure 69 we can see that, while setting a threshold scaling function of $V_{TH} \approx 0.27 - 0.07V_{DD}$ (that resulted above in $\eta = 2$) will have a large impact on subthreshold current, it increases the slope of the $I_D(sat)$ curve only slightly (from about $V^{1.8}$ to V^2) and results in both the static and dynamic currents becoming proportional to V_{DD}^2. In this case, assuming that $A \propto N$ (the number of switching devices) and $F \propto A^{-1/\sigma} \propto V$, the dynamic power term $NFCV^2$ becomes $P_D \propto AA^{-1/\sigma}A^{-2/\sigma}$ so that:

$$P_D \propto A^{(\sigma-3)/\sigma} \tag{4.15}$$

which is identical to (4.6). However, it will be seen below that although an achievable range appears to be $\sim 1.9 \le \beta \le \sim 3.5$, depending on a number of specific technology assumptions, the case where $\beta = \eta = 2$, and thus where the exponents in both (4.11) and (4.15) are $(\sigma\text{-}3)/\sigma$, is unlikely to be typical.

4.3.3 Modeling Variability

Some of the primary sources of variability have already been discussed in Section 2.3.3. In the model of [61], the density function of the critical path delay variation resulting from die-to-die (D2D) and within-die (WID) fluctuations is taken to be normally distributed. The WID fluctuations have been shown to directly affect the mean of the performance distribution, proportional to the nominal critical path delay ($T_{cp\ nom}$), while the D2D variability changes its variance. The worse-case performance for a particular chip is determined by the sum of the D2D and WID contributions [313] so that:

$$T_{cp\ \max} = T_{cp\ nom} + \Delta T_{D2D} + \Delta T_{WID} \tag{4.16}$$

As a result, the maximum operating frequency will be normally distributed about $1/T_{cp\ nom}$.. The standard deviation will be affected by both the average critical path logic depth and the total number of independent critical paths in the chip (N_{cp}). It was found in [61] that systematic WID fluctuations (and in particular its contribution to channel length variability) will have the most significant impact on future device scaling. These models predict a worse-case reduction in the maximum operating frequency of up to 40%, depending on the manner in which WID variation impacts on total channel length.

Combining this with the empirical temperature model of [78], $F_{MAX} \propto \Theta^{-a}$, where Θ = temperature and 0.5<a<0.75, the overall variability resulting from both process and temperature can be modeled as a normally distributed spread in the maximum operating frequency centred around a temperature-dependant mean. F_{MAX} is proportional to $\frac{I_D(sat)}{d_{cp}C_L V_{DD}}$, where C_L and V_{DD} are constant at a given node. It can also be assumed that d_{cp} will asymptote to an approximately constant value at high N [169]. Thus $I_D(sat) \propto V^{\beta}$ will exhibit the same distribution. Considering only the worse-case performance vs. area at each scaling point, the overall impact on the model is to add an uncertainty range to the $I_D(sat)$ curve. Variability may therefore be accommodated by introducing an additional term into (4.14) such that:

$$I_D(sat) \propto V^{\beta \pm \varepsilon}. \tag{4.17}$$

In a similar way, because the variation in both V_{TH} and S directly affects the worse-case slope of and intersection points of the curves in Figure 67, the subthreshold current can be modeled with variability as:

$$I_{OFF} \propto V^{\eta \pm \varepsilon'}. \tag{4.18}$$

The terms ε and ε' represent the bounds of the distribution around the mean of $I_D(sat)$ and I_{OFF} respectively due to all sources of parameter variability. In the case of a fully synchronous system, the most interesting point is the $+3\sigma_{vth}$ corner as this will set the lower limit on the operating

frequency and therefore on the architectural parameter σ. In contrast, asynchronous systems will tend to operate closer to the mean [313] with the impact of the $\pm 3\sigma_{vth}$ distribution tending to average out at the system level. Various studies identified previously in Section 2.3.3 have indicated that a variability figure of less than ±25% of V_{TH} will be achievable into the future. This is higher than the ITRS predictions (~12%) but is consistent with [61] as well as [66] and is the figure that will be used in Section 4.4.5 below.

4.3.4 Short Circuit Power

Short circuit power represents only a small percentage, typically 10–20%, of the overall dynamic power figure as long as the gate is loaded such that the input and output signals exhibit approximately equal rise and fall times. If this is not the case, for example with small fanout and local interconnect, then the short-circuit dissipation may exhibit the same order of magnitude as the switching power. The unloaded case therefore represents an upper bound on the short circuit power.

An analysis of the relationship between area and short-circuit power can start with the equation for average short circuit current (I_{AVE}) derived by Veendrick [175] for the unloaded case (i.e., $C_L = 0$) and with the widths of the P and N transistors adjusted to compensate for mobility differences:

$$I_{AVE} = \frac{1}{12}\frac{\beta}{V_{DD}}\left[V_{DD} - 2V_{TH}\right]^3 \frac{\tau}{T} \tag{4.19}$$

where β = device gain, τ = input rise/fall time and T is the clock period (1/F). A key assumption here is that the circuits of interest exhibit low Rent exponents, such that the average fanout and interconnect length asymptotes to a small fixed value as the size of the circuit increases. Further, given the assumption of a fixed technology, both β (∝ W/L) and the device capacitance can be considered to be constant. Thus, with (4.14), $\tau = CV_{DD}/I_D$ simplifies to $\tau \propto V_{DD}^{1-\beta} \propto V_{DD}^{-1}$.

Typical values of τ/T tend to be small (<0.1), which is why P_{SS} is ignored in most power analyses. Further, (4.19) holds only where V_{DD} is greater than the sum of the device thresholds. When V_{DD} falls below this value, I_{AVE} tends to zero, as it is not possible for both transistors to be on simultaneously over the full range of gate voltages. Substituting $\tau \propto 1/V_{DD}$ and eliminating both β and the constant (1/12), (4.19) simplifies to:

$$I_{AVE} \propto \left[V_{DD} - 2V_{TH}\right]^3 FV^{-2} \tag{4.20}$$

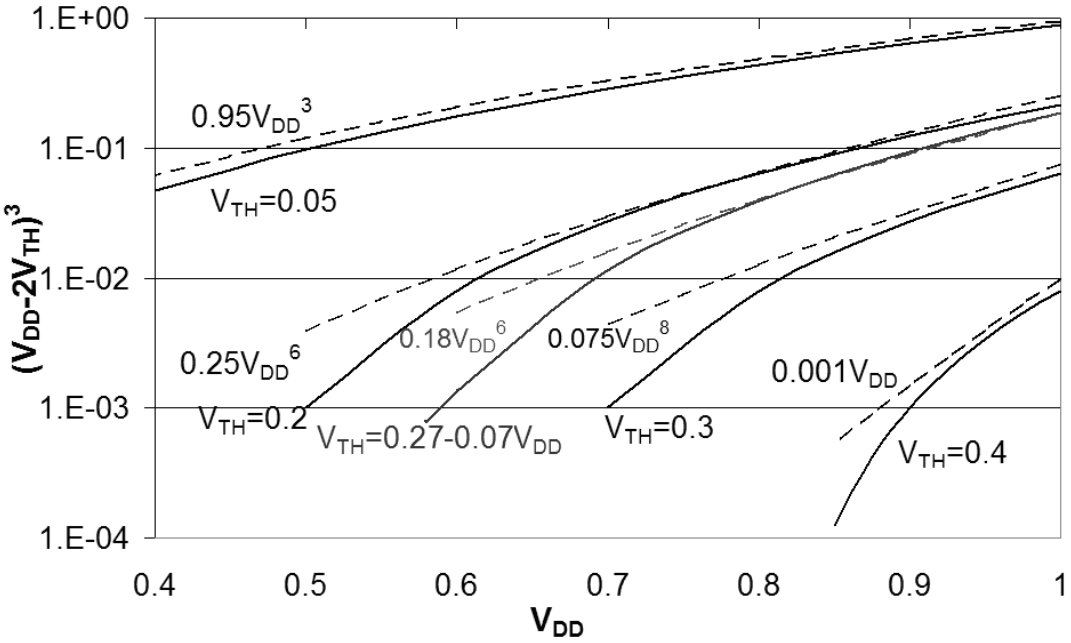


Figure 70. $(V_{DD}-2V_{TH})^3$ vs. V_{DD} for various V_{TH} functions
With either V_{TH} fixed (solid lines) or V_{TH}=0.29-0.075V_{DD} (red line at center), a value of n can be selected such that kV_{DD}^n (dotted lines, k=constant) represents a close upper bound on $(V_{DD}-2V_{TH})^3$.

Figure 70 plots $(V_{DD} - 2V_{TH})^3$ against V_{DD} over the range $0.4 \le V_{DD} \le 1.0$V, and for various fixed values of V_{TH} along with $V_{TH} = 0.27\text{-}0.07V_{DD}$. Also shown are plots of V_{DD}^n for various values of n. It can be seen that, just as for the subthreshold case, it is always possible to select a value of n (≥3) such that V_{DD}^n becomes an upper bound on $(V_{DD} - 2V_{TH})^3$. For example, with $V_{TH} = 0.27\text{-}0.07V_{DD}$, the term is bound by approximately $0.18V_{DD}^6$ and thus, $I_{AVE} \propto FV_{DD}^4$ ($V_{DD} > 0.6$), and $P_{SS} \propto V_{DD}^5$. Substituting $V \propto F \propto A^{-1/\sigma}$ and multiplying by A the average short circuit power becomes:

$$P_{SS} \propto A^{\sigma-5/\sigma} \tag{4.21}$$

The exponent $(\sigma\text{-}5)/\sigma$ implies that short circuit power will continue to contribute only a very small fraction of the overall dynamic term as area increases. Further, as V_{DD} approaches $2V_{TH}$, the short circuit current rapidly tends to zero. It is extremely sensitive to V_{DD} and, as for switching power, can be easily traded off against area.

4.3.5 Gate Leakage

Gate leakage has been predicted to exceed sub-threshold leakage at the 65nm technology node [12] although there is recent evidence that problems have already arisen at 90nm [314]. As it is due to direct tunneling through the gate oxide in the presence of high electric fields, it varies exponentially with oxide thickness and is extremely sensitive to gate voltage [315]. The current density J_{FN} at the transistor gate will have the general form of the Fowler-Nordheim tunneling equation:

$$J_{FN} = C_0 E^2 e^{-Y/E} \tag{4.22}$$

where $E \approx V_G / T_{OX}$ is the surface electric field; T_{OX} = oxide thickness while both $C_0 = \left(q^3 / 6\pi^2\hbar\right)\left(m_0 / m_{OX}\right)\Phi_b^{-1}$ and $Y = \left(4(2m_{OX})^{1/2}\right)\Phi^{3/2}$ are functions of the effective barrier height (Φ_b) between the oxide and the silicon surface as well as the effective electron mass (m_{OX}). Assuming that C_0 is fixed for a given technology, the gate current will have the form:

$$I_G \propto A\left(\frac{V_G}{T_{OX}}\right)^2 e^{-\frac{YT_{OX}}{V_G}} \tag{4.23}$$

The value of Y depends somewhat on the model of gate leakage and both C_0 and Y are dependent on temperature and the Schottky effect [316, 317]. In the current literature, Y varies from $1.9\text{x}10^8$V/cm [318] to a more recent value of $1.43\text{x}10^8$V/cm in [319]. It is possible to fit curves of the form aV_G^n to the $1/T_{OX}^2\, e^{-(YT_{OX}/V_G)}$ term in (4.23) for each particular value of T_{OX} and so $I_G \propto AV_G^2 V_G^n$. As T_{OX} decreases, the value of n tends to fall to about 5.

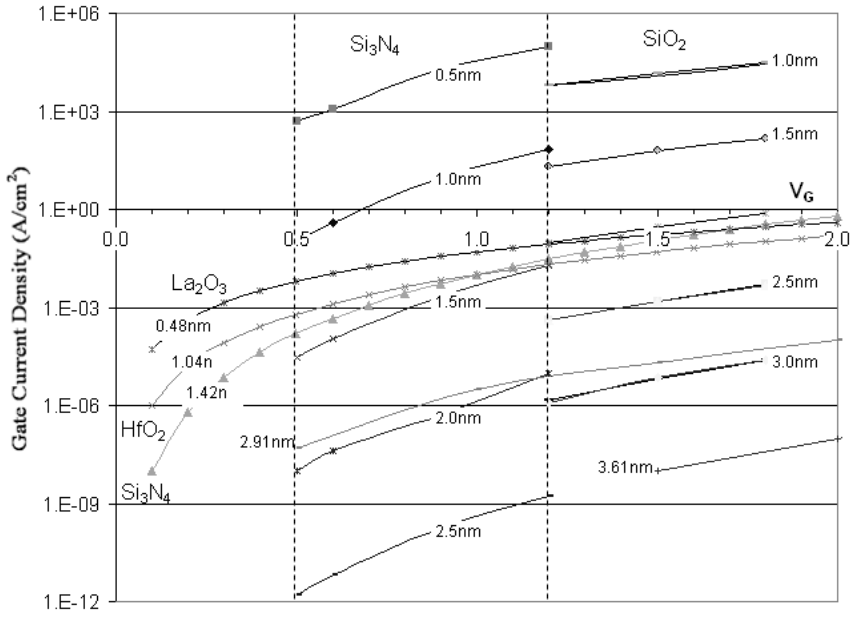


Figure 71. Gate current density (Amp/cm^2) vs. gate voltage for various materials and gate oxide thickness values as shown.

Figure 71 plots gate leakage data from a number of simulations and experiments reported in the literature [31, 176, 320]. Approximations of the same aV_G^n form can be fitted to all of these curves with an average fitting error typically less than ±15%. The worse-case exponent for SiO_2 at T_{OX} = 1.5nm is n ≈ 2.5 and, similarly, n ≈ 2.64 for the La_2O_3 dielectric described in [176] with an effective T_{OX} of 0.48nm. As will be seen in the next section, the ITRS is predicting that the path for gate dielectric will be a nitride such as Si_2N_3 ($\varepsilon_r \approx 7.5$) followed by a high-k material such as $ZrSiO_2$ ($\varepsilon \approx 15$) or more likely HfO_2 with a dielectric constant of about 25. The current density curve for the HfO_2 gate n-FET in [176] (effective T_{OX} = 1.4nm) can be approximated to $0.0186V^{6.3}$. Thus, it can be predicted that the gate current density of future oxides will tend to exhibit a range $\propto V_G^3$ to V_G^8.

As T_{OX} shrinks, gate leakage will become more of a constraint to lowering power by increasing area but will never become dominant. Substituting $F \propto V \propto A^{-1/\sigma}$, the gate leakage becomes $I_G \propto A^{(\sigma-5)/\sigma}$. The gate leakage component of power can be easily reduced with increasing area for all effective oxide thicknesses ≥ 0.2nm such that:

$$P_G \propto A^{(\sigma-\chi)/\sigma}, \quad \chi \geq 6 \tag{4.24}$$

Gate Leakage from the CMOS Roadmap

The following analysis uses the ITRS MASTAR tool (version 2.0.7) [321] to examine the likely development of gate power in HP logic. High Performance (HP) technology has been used here as it exhibits the worse-case gate leakage. The model for the growth in transistor numbers (N) is based on the ITRS assumptions for ASIC/MPU such that $N \approx 5.8x10^5 L^{-1.97}$ (L = physical gate length in micron). It is also assumed that the overall chip size stays roughly the same as the device size reduces and the number of transistors increases. The ITRS suggests a typical figure of ~280mm^2 for high volume microprocessor chips throughout the roadmap period.

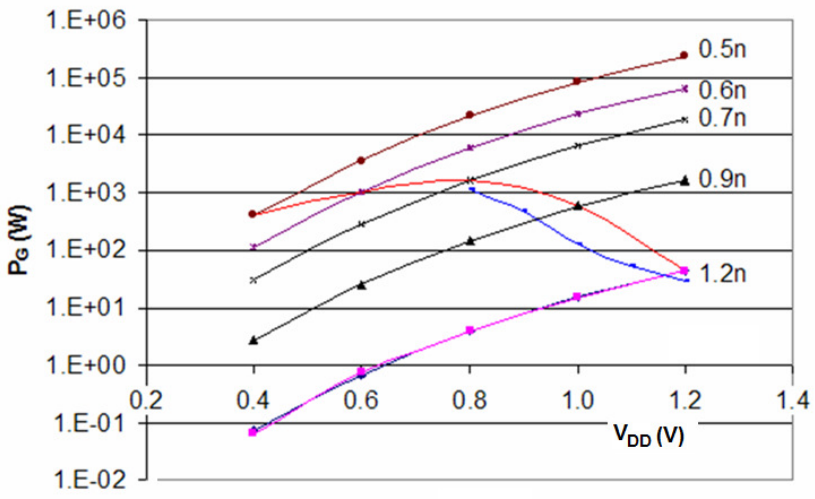


Figure 72. Total gate leakage power vs. supply (V_{DD}) at various T_{OX} as shown. The pink line is $14.6V_{DD}^{5.9}$. This matches the curve for L=37nm within 10%. The general shape of the curves is the same for all T_{OX} values. The red curve is the gate leakage power trajectory if both supply voltage and T_{OX} reduce at successive nodes.

In the previous section, it was assumed that gate leakage could be described using a F-N tunneling model such that $I_G \propto (V_G / T_{OX})^2 e^{-YT_{OX}/V_G}$ and a power-law function of the form aV_G^n fitted to the gate current such that n > 3 for a plausible range of gate materials and biases. Figure 72 shows the total gate power (= $V_{DD}I_G$) determined from MASTAR for the high performance technology nodes from 90nm down to 22nm over a range of supply voltages (V_{DD}) from 1.2V down to 0.4V. This is the direct tunneling component through the oxide when $V_G = V_{DD}$ and $V_{S/D} = 0$. Due to

the fixed tunneling parameters built into the MASTAR equations, the derived gate current figures are valid for SiO_2, but other dielectric materials have to be normalized to an equivalent oxide thickness (EOT) i.e., the thickness of SiO_2 that would result in an equivalent value of leakage current.

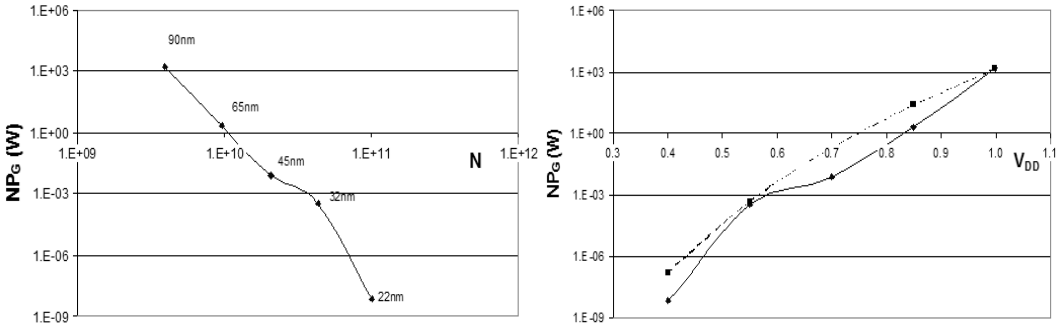
The blue curve overlaying Figure 72 indicates the gate power trajectory for the ITRS HP chart between 2004 and 2016 while the red curve represents the trajectory where supply voltage is aggressively scaled from 1.2V to 0.4V at successive technology nodes but SiO_2 is kept as the gate oxide. In either case, the gate current reaches more than 1KW and although the trend from the 45nm point is decreasing, its final (2016) value is still around 400W. Clearly, this is unsustainable and is therefore unlikely to be a trajectory in any practical technology.

The general shapes of the individual gate current curves in Figure 72 are a close match to $kV_{DD}^{5.9}$ (±10%). Only the multiplier *k* changes with technology node and thus the total gate power will be approximately $P_G \propto V_{DD}^7$. However, this only applies under the unrealistic assumption that the gate oxide thickness is held constant at successive technology nodes. The current roadmap for HP technology specifies that EOT will progressively thin to 0.5nm, although the likely scaling limit for SiO_2 is greater than 1nm. Silicon Oxynitride (SiON) material with multiple EOT between 1 and 2nm has already been introduced at the 65nm node and will probably continue to be used (in preference to Hafnium based oxides) at 45nm [322]. Further, gate tunneling in ultra-thin and double-gate devices is expected to be significantly better again than planar devices due to a combination of reduced vertical electric field and quantum confinement effects [323]. The simulations reported in [323] and [324] predict that undoped thin-body double-gate structures (e.g., HfO_2 insulator, T_{OX} = 1nm) could exhibit gate leakages an order of magnitude smaller than an equivalent bulk device. These considerations are not yet reflected in the ITRS estimates.

Hi-κ Dielectrics

To investigate the impact of dielectric constant (κ) on these models, successively higher dielectric values were substituted into the existing profiles on MASTAR. The general methodology was to

load the high performance profile and adjust the supply voltage to the desired value. This caused the threshold voltage (and therefore the off-current) to shift due to SCE and DIBL. The hi-κ dielectric was then switched in and the oxide thickness adjusted upwards until the original value of off current was achieved.



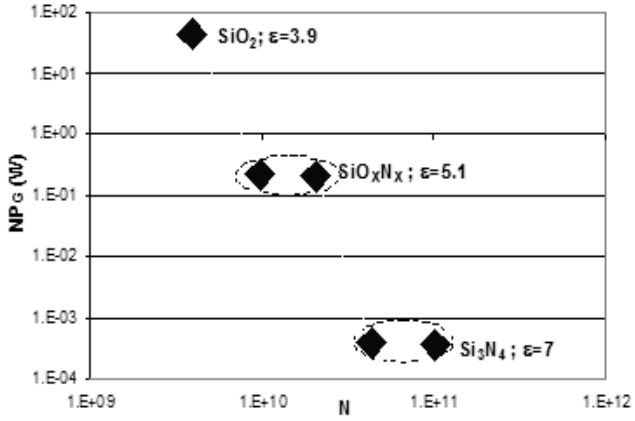
(a) (b)

Figure 73. Total gate leakage power (a) vs. N and (b) vs. V_{DD} assuming that gate dielectric constant increases at each node.

Values of dielectric constant for SiON ($\kappa = 5.08$) were used at the 65nm node, Si_3N_4 ($\kappa = 7.0$) at 45nm and 32nm, and $ZrSiO_2$ ($\kappa = 15$) at 22nm. The physical oxide thickness values were then used to extract comparative leakage figures from the literature, where available. Figure 73 illustrates the gate power under these scaling assumptions. The dotted curve of Figure 73b is $1.5x10^3 V_{DD}^{25}$, indicating that it will be possible to reduce the effect of gate power such that it makes an insignificant contribution to the overall power budget. It also clearly demonstrates the strong case for moving to Silicon Oxynitride from the 65nm technology node.

The total gate power (NP_G) vs. number of devices (N) (Figure 74) was derived using a combination of aggressive voltage and threshold scaling. Here, Silicon Oxynitride is assumed at 65nm and 45nm and Silicon Nitride for the 32nm and 22nm nodes. The values of leakage current relative to SiO_2 were taken from [325]. Clearly, the impact of gate power is greatly reduced. At the 65 and 45nm nodes, gate power will be less than 1% of the subthreshold current, while at the 32nm node it falls to five orders of magnitude less. It can be concluded that the gate power will always represent a small fraction of the overall static leakage power. Even in the HP roadmap, it is predicted to asymptote to about 30% of the subthreshold power, and it is reasonable to expect

that suitable mechanisms will become available to ensure that gate power makes only a small contribution to overall power [326], even if these still require significant engineering effort to integrate them into existing manufacturing lines.



L (μm)	V_{DD}	V_{TH}	N ($x10^9$)	τ (pS)
0.037	1.2	0.200	3.93	0.96
0.025	1.0	0.225	9.68	0.71
0.018	0.8	0.250	20.6	0.50
0.013	0.6	0.275	43.5	0.43
0.009	0.4	0.300	101	0.60

Figure 74. Total gate leakage power vs. N – gate materials as shown.
The table gives the corresponding values of supply, threshold and intrinsic delay

4.3.6 Gate Induced Drain Leakage (GIDL)

Gate Induced Drain Leakage (GIDL) originates from band-to-band tunneling involving carriers in the high field region of the drain that is overlapped by the gate and occurs when the gate is grounded and the drain is at V_{DD}. As it is a form of tunneling current, it can be modeled using a similar general form to the gate leakage. The main difference in this case is the linear electric field dependence:

$$I_{GIDL} = aE_S e^{(-b/E_S)} \tag{4.25}$$

where $E_S \approx \left(V_{GD} - V_{FB} + E_g/q\right)/3T_{OX} \approx (V_{DG} - 1.2)/3T_{OX}$ is the vertical surface electric field, both a $\left[= q^2 m_r^{1/2}/18\pi h^2 E_g^{3/2}\right]$ and b $\left[= \left(\pi m_r^{1/2} E_g^{3/2}\right)/2qh\right]$ are process dependent constants, m_r the effective electron mass, V_{GD} the gate-drain voltage and T_{OX} = gate oxide thickness and E_g/q the energy bandgap. Typical values of b range from 23 to 70 MV/cm [327, 328]. Strictly, the simplified electric field equation is not directly applicable to thin-body transistor structures where the electric field depends on the body thickness as well. In [281], GIDL was shown to be

significantly lower in thin-body transistors as compared with conventional bulk-Si MOSFETs due to a reduction in transverse electric field at the surface of the drain and an increase in the effective transverse electron mass with decreasing body thickness. Thus $E_S \approx (V_{DG} - 1.2)/3T_{OX}$ can be considered to be a worse-case field strength. From (4.25), it can be seen that GIDL is sensitive to the drain doping profile (which results in a non-uniform electric field), the transverse electrical field (dependent on gate-drain voltage and oxide thickness) as well as the effective mass of the tunneling electrons.

Although it is likely to remain a problem for memory devices [329], various analyses [281, 327, 330, 331] have shown that GIDL will be small (~10^{-12}A) for digital logic at $V_{DD} < 1.1$V as a band bending of at least 1.2eV (i.e. the approximate energy band gap of silicon) is necessary for band-to-band tunneling to occur [331]. Supply voltages (and therefore V_{DG}) have already fallen below a point where this can occur and for this reason GIDL will not be considered further in this analysis.

4.4 Dynamic and Subthreshold Power/Energy Scaling vs. Area

In this section the simple voltage-current-frequency relationships developed in the previous section are extended to derive power-area and energy-area functions for both the dynamic and static power/energy loss. It was determined in the previous sections that the dynamic and subthreshold currents will impose the most severe limits on power/energy scaling so these will be the main focus from here onwards. The threshold and slope variability will also be omitted for the moment and will be reintroduced later. The main objective here is to remove the assumption that $F \propto V$ and to identify the conditions under which it is possible to achieve constant or reducing power and/or energy with increasing area.

4.4.1 Capacitance Scaling

An implicit assumption of the analysis to this point has been that it relates to a single technology node with a fixed minimum gate size and where the overall area (chip size) increased with the number of devices (N). Thus the total switching capacitance (load + interconnect) can be ap-

proximated to a simple linear function of N. In this section, that constraint is removed and an additional parameter γ is introduced to account for changes to capacitance per unit area as a function of N.

Two main factors affect the general relationship between area and node capacitance:

1. the impact of changes in the interconnect loading due to increases in the number of devices and/or circuit complexity. Most wire length estimation techniques are based on variations of Rent's Rule [332] that models the interconnect density as a power law function of the number of devices (N) with the form $R = kN^P$;

2. changes in device and interconnect characteristics when migrating between fabrication technologies at successive nodes of the roadmap. The desire to continually improve performance is driving industry towards aggressive scaling of L_g as well towards a number of major material and process changes. These include the high-κ gate dielectrics discussed above, along with metal gate electrodes and low-κ interconnect materials. All of these design decisions will impact on the final node capacitance.

In the first case above, it can be expected that capacitance will follow a simple power-law function of N, due to the form of Rent's Rule. The second case is more complex, as it depends on a number of intangible design decisions at future technology nodes. However, the analysis in [93] of wires in scaled technologies has identified that for short connections (those that tend to dominate chip wiring), interconnection delay closely tracks gate delay with scaling. As the trend in gate capacitance is $C_g \propto L_g \propto N^\gamma$, overall scaling interconnect capacitance at each circuit node can be assumed to exhibit the same general form, i.e.:

$$C \propto N^\gamma \tag{4.26}$$

where γ incorporates contributions from the Rent exponent as well as reductions in minimum feature size. Although this is undoubtedly an oversimplification, it will be seen in Section 4.4.6, which looks at some examples using data drawn from the ITRS, that it sufficiently captures the

trends in circuit design, interconnect and fabrication technology for this work, and allows the estimation of realistic bounds on their likely impact at future technology nodes.

4.4.2 Dynamic and Subthreshold Scaling Models

Given that the architectural parameter σ that links performance (frequency) with area can also describe the tradeoff between power (and/or energy) and area, this can now be combined with the capacitance model just developed and two further device level parameters introduced that together with γ set an upper bound on the range of σ for which reducing power and/or energy can be achieved.

Using the voltage-current and capacitance approximations of (4.10) and (4.14), and ignoring variability for the moment, the frequency scaling factor, F, becomes $F \propto I/CV \propto V^{\beta}/N^{\gamma}V$ so that:

$$\begin{aligned} &F \propto N^{-\gamma}V^{(\beta-1)} \text{ or, equivalently,} \\ &V \propto (N^{\gamma}F)^{1/(\beta-1)}. \end{aligned} \tag{4.27}$$

This generalizes the F ∝V relationship and (4.27) can now be used to extend (4.4)–(4.6) to model the growth of power and energy with future device scaling.

Dynamic Switching Power

Starting with the dynamic power equation $P_D \propto FC_LV_{DD}^2$, multiplying by N and substituting (4.27) results in $P_D \propto NC_LF(C_LF)^{2/(\beta-1)}$, so that:

$$P_D \propto N(C_LF)^{\chi} \quad \text{where } \chi=\frac{\beta+1}{\beta-1}. \tag{4.28}$$

Here, it is also assumed that the activity ratio does not change with scaling (i.e., the same ratio of devices switching to total devices is maintained as area increases). This will be true of the parallel architectures analysed later in Chapter 5. Combining (4.28) with (4.4) and (4.26), then setting A ∝ N, the dynamic power becomes:

$$P_D \propto A^{\gamma\chi} A^{(\sigma-\chi)/\sigma}. \tag{4.29}$$

Further, by multiplying (4.29) and (4.4) the dynamic switching energy $E_D = P_D T = P_D A^{1/\sigma}$ is given by:

$$E_D \propto P_D A^{1/\sigma} \propto A^{\gamma\chi} A^{(\sigma-(\chi-1))/\sigma}. \tag{4.30}$$

Note that setting F ∝ V (i.e., $\beta = 2$) and $\gamma = 0$, $\chi = 3$ so that (4.29) reduces to (4.15). Under the same conditions, switching energy becomes $E_D \propto A^{(\sigma-2)/\sigma}$.

The necessary condition for constant or reducing dynamic power is that the sum of the exponents in (4.29) are ≤ 0, i.e.:

$$\sigma \le \frac{\chi}{\gamma\chi + 1} \tag{4.31}$$

Similarly, in the case of energy:

$$\sigma \le \frac{\chi - 1}{\gamma\chi + 1}. \tag{4.32}$$

Subthreshold Leakage Power

Under the assumption that supply and threshold voltages are related by $V_{TH} = a - bV_{DD}$, the subthreshold leakage current can be simply modeled as $I_{OFF} \propto NV_{DD}{}^{\eta}$ so that $P_{SUB} \propto NV_{DD}^{(\eta+1)}$. As for the dynamic power case, by substituting (4.27) subthreshold power becomes $P_{SUB} \propto N(CF)^{(\eta+1)/(\beta-1)}$ so that:

$$P_{SUB} \propto A^{\gamma\chi'} A^{(\sigma-\chi')/\sigma} \quad \text{where } \chi' = \frac{\eta+1}{\beta-1} \tag{4.33}$$

which can be compared to the form of the dynamic power case in (4.28). In the same manner as (4.30), subthreshold leakage energy becomes $E_{SUB} = P_{SUB}T$ so that:

$$E_{SUB} \propto A^{\gamma\chi'} A^{(\sigma-(\chi'-1))/\sigma}. \tag{4.34}$$

As a result, constant or reducing subthreshold leakage power can be achieved when:

$$\sigma \le \frac{\chi'}{\gamma\chi'+1} \tag{4.35}$$

and energy when:

$$\sigma \le \frac{\chi'+1}{\gamma\chi'+1}. \tag{4.36}$$

It can be seen that when $\beta = \eta = 2$ and $\gamma = 0$, the subthreshold power becomes proportional to NV_{DD}^3 and thus $P_{SUB} \propto A^{(\sigma-3)/\sigma}$, which is the result obtained previously in Section 4.3.1. Under the same conditions, the energy term becomes $E_{SUB} \propto A^{(\sigma-2)/\sigma}$, so that both energy and power exhibit identical forms to the dynamic case. However, it will be seen in Section 4.4.5 below that while it is always possible to find values of 'a' and 'b' such that $I_{SUB}/I_{SO} \propto V_{DD}^2$ (i.e., $\eta = 2$), these values will not automatically result in $F \propto V$ (i.e., $\beta = 2$) except under a very narrow set of assumptions regarding the scaling of V_{DD} and V_{TH}.

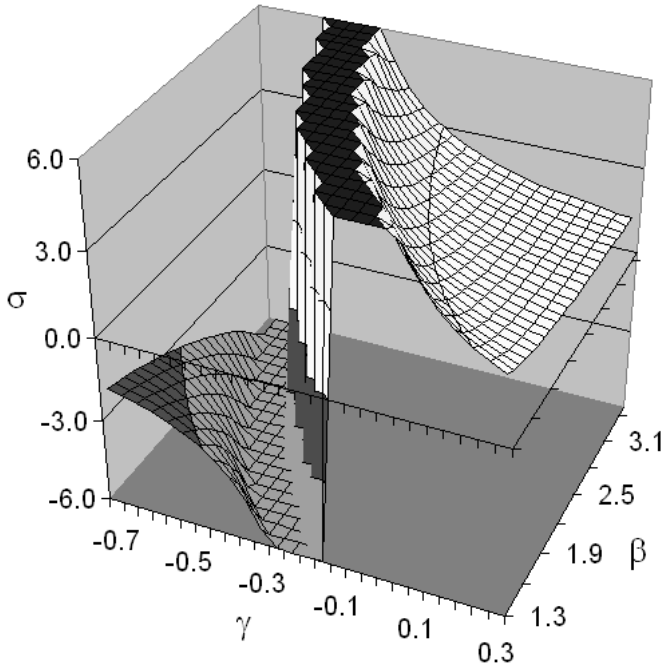


Figure 75. Surface defining $\sigma=\chi/(\chi\gamma+1)$ as a function of β and γ

In summary, the model is now based on a single architectural parameter (σ) and three device-level parameters (χ, χ' and γ) that interact to relate the power and energy of a circuit to its area. Figure 75 shows a surface defined by the equality in (4.31) i.e., $\sigma = \frac{\chi}{\gamma\chi + 1}$, $\chi = \frac{\beta + 1}{\beta - 1}$, over a range of β and γ. The singularity where $\gamma\chi + 1 \to 0$ represents the case where drive current, supply and capacitance track one another such that $F = \frac{I}{CV} \propto A^{-1/\sigma} \to 1$ and (4.29) reduces to $P_D \propto A^{\gamma\chi} \to 1$. Under these conditions, the maximum operating frequency and power scaling will be constant regardless of the architecture. This singularity divides the surface into two regions reflecting the general relationship between area and node capacitance outlined in Section 4.4.1 above. To the right of Figure 75, where $\gamma\chi \geq 1$, capacitance predominately increases due to higher interconnect loading resulting from the impact of Rent's Rule. This would be the case where the architecture is duplicated within a given technology, changing A (and possibly γ) but not χ or χ'. In the left region, $\sigma < 0$ due to the effect of reductions in device and interconnect capacitance at successive technology nodes i.e. $F = \frac{I}{CV} > 1$. As the form of (4.35) is the same as (4.31), an identical surface may be drawn for the subthreshold power case, describing σ as a function of χ' and γ.

Here, the parameter σ can be interpreted as a measure of how *serial* is a particular architecture. As σ increases, there is a smaller performance improvement for a given increase in area (see Figure 65). As both χ and χ' depend on the scaling behavior of V_{DD} and V_{TH} (defined by β and η), these may now be used to determine the overall relationship between σ, voltage and area.

4.4.3 Supply and Threshold Scaling vs. Area

The work in this chapter focuses specifically on the evolution of power and/or energy as device numbers (area) scale upward. The model assumes that as area increases, each component in the dynamic power equation will scale by some factor, thereby contributing to the overall power and

energy. With $A \propto N$ and using (4.4) and (4.26), $P_D = NFCV^2$ becomes $P_D = AA^{-1/\sigma}A^{\gamma}V^2$ and V represents a voltage scaling factor that can be related to the target power and area scaling as:

$$V = \sqrt{P_D A^{-\gamma}}\, A^{-\frac{(\sigma-1)}{2\sigma}} \tag{4.37}$$

where P_D is the target dynamic power *scaling* between successive nodes, ideally ≤ 1. By substituting (4.29), the voltage scaling becomes a function of both the architectural parameter σ and the circuit-level parameter χ:

$$V = A^{-\gamma X} A^{X/\sigma} = A^{(X-\gamma X\sigma)/\sigma} \quad , X=(1-\chi)/2. \tag{4.38}$$

In a similar way, the total subthreshold power scaling factor (P_S) for N devices will be linked to a change in threshold voltage (ΔV_{TH}) by:

$$P_S = NVe^{-\Delta V_{TH}/nV_t}. \tag{4.39}$$

Substituting the supply scaling from (4.37), the threshold voltage scaling function becomes:

$$\Delta V_{TH} = nV_t \ln\left(\frac{\sqrt{P_D A^{\gamma}}}{P_S} A^{\frac{\sigma+1}{2\sigma}}\right). \tag{4.40}$$

Equations (4.37) and (4.40) describe the area-voltage relationships that will allow constant or reducing power with increasing area, assuming constant overall completion time.

As an illustration, Figure 76 plots (4.37) vs. (4.4), with $\gamma = 0$ and $P_D = 1$, over a range of A and σ. At the point A = 2, $\sigma = 3$ (circled in Figure 76), $F = V = \sqrt[3]{2} \approx 0.79$, while from (4.40) (and with $P_S = 1$ and S = 100 mV/dec), $\Delta V_{TH} = nV_t \ln 2^{(2/3)} \approx +20$ mV. As a result, constant dynamic and subthreshold power may be achieved in the face of a successive doubling of device numbers (area) by adjusting the supply, frequency and threshold parameters as shown (i.e. at each doubling, F and V are scaled by 0.79 and $\Delta V_{TH} \approx +20$ mV). The resulting performance loss can be compensated at the architectural level as long as the chosen technique can achieve AT^3 = constant. However, it will be seen in Chapter 5 that the combined effects of lower gate overdrive and

device variability on typical low-power systems tends to push both χ and χ' towards values much smaller than 3, making it increasingly difficult to find architectures that are sufficiently parallel to satisfy the inequalities of (4.31) and (4.35).

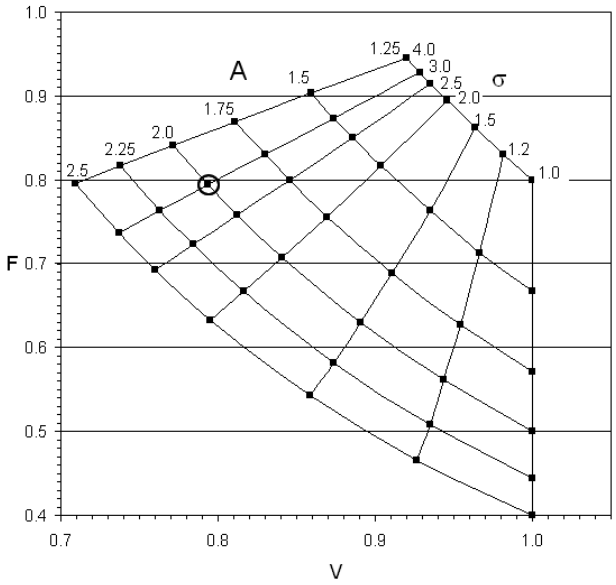


Figure 76. A constant dynamic power scaling surface defined by $F=A^{-1/\sigma}$ vs. $V=A^{-(\sigma-1)/2\sigma}$ ($\gamma = 0$, $P_D = 1$) across a range of area scaling factors (A) and σ.

4.4.4 Total Power vs. Area

As only the switching and subthreshold terms are being considered here, from (4.29) and (4.33), the total power at the next node *n+1* i.e., $(P_{total})_{n+1}$, will be related to the power at the present node *n* as:

$$\left(P_{total}\right)_{n+1} = \left(P_D\right)_n A^{\chi} A^{(\sigma-\chi)/\sigma} + \left(P_S\right)_n A^{\chi'} A^{(\sigma-\chi')/\sigma}. \tag{4.41}$$

If the target in this case is a constant scaling factor (P_T) for total power, $(P_{total})_{n+1} = P_T(P_{total})_n$, then (dropping the subscripts) (4.41) becomes $P_T\left[P_D + P_S\right] = P_D A^{\chi} A^{(\sigma-\chi)/\sigma} + P_S A^{\chi'} A^{(\sigma-\chi')/\sigma}$ and:

$$P_T = \frac{A^{\chi} A^{(\sigma-\chi)/\sigma} + P_R A^{\chi'} A^{(\sigma-\chi')/\sigma}}{P_R + 1} \tag{4.42}$$

where $P_R = P_S/P_D$ represents the ratio of the subthreshold and dynamic power contributions.

Traditionally, both the absolute value of subthreshold power (P_{SUB}) and the ratio P_R have been small so that P_{SUB} could be safely ignored. However, this is unlikely to continue into the future, particularly if the ITRS target of 14–17% performance improvement per node is used to constrain the relationship between the supply and threshold voltages.

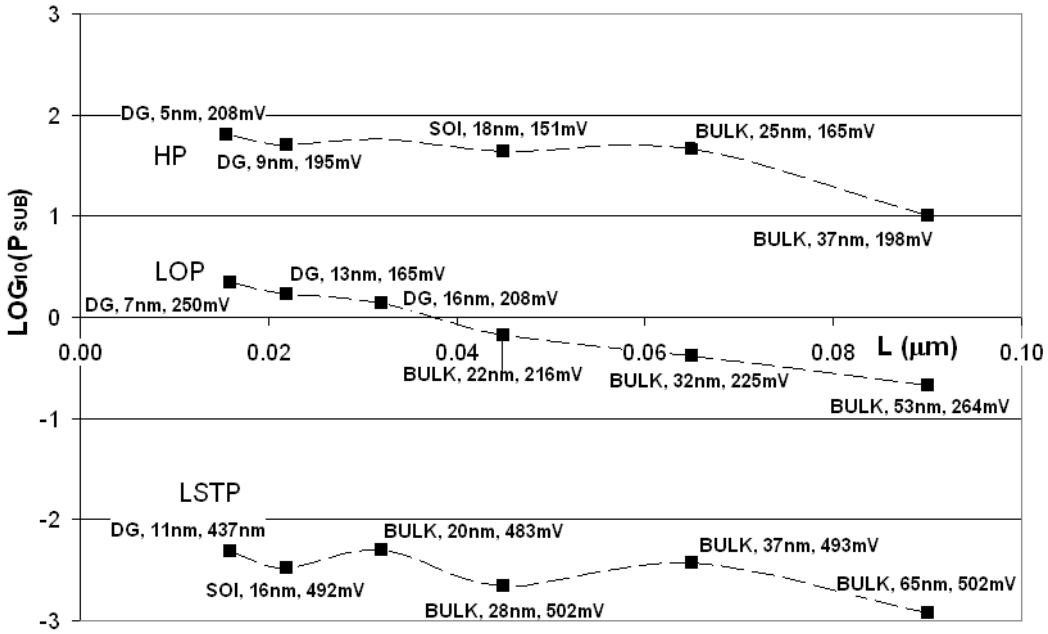


Figure 77. Some ITRS subthreshold current predictions vs. gate length

Figure 77 maps one possible trajectory of subthreshold power for the three ITRS classes of technology, High Performance (HP), Low Operating Power (LOP) and Low-Standby Power (LSTP) over the remaining scaling nodes. The supply and threshold voltages are drawn from ITRS data and an average transistor width of $3L_g$ is assumed. The growth in device numbers is modeled as $N \propto L_g^{-1.98}$ (L_g = drawn gate length) and the sequence of technologies—bulk followed by SOI and double gate—represents a realistic scaling scenario predicted by the roadmap. Under these assumptions, subthreshold power will grow by as much as an order of magnitude over the remaining nodes of the roadmap.

Table 11 shows the HP dynamic and subthreshold power estimates under the same assumptions and with a fixed activity factor of 0.1. In general, the ratio of subthreshold to dynamic power

ranges between 10% and 40%, although it obviously depends directly on specific technology parameters such as subthreshold slope and activity factor. On the other hand, for low-power technology its is reasonable to expect that P_S/P_D should remain below ~10%. The threshold voltage scaling proposed above is intended to limit the growth of both P_D and P_S, ensuring that P_S/P_D remains at or below its current (e.g. 90nm) value.

Table 11 Indicative dynamic and subthreshold power estimates for ITRS HP technology.

Year	HP Tech	L_{DRWN} (μm)	V_{DD} (V)	V_{TH} (mV)	N ($x10^9$)	I_D(sat)♣ (μA/μm)	I_{OFF} ♣ (nA/μm)	P_{DYN}= 0.1NVI_D (W)	P_{SUB} (W)	P_R
2004	Bulk	0.090	1.2	198	0.53	1024	58	84	10	0.12
2007	Bulk	0.065	1.1	165	1.10	1197	196	133	46	0.35
2010	SOI	0.045	1.0	180	2.21	1812	145	255	43	0.17
2013	SOI	0.032	0.9	188	4.40	2212	152	397	58	0.15
2016	DG	0.022	0.8	195	8.85	2763	108	610	50	0.08
2019	DG	0.016	0.7	208	17.5	2677	450	724	63	0.09

♣ Figures for I_D and I_{OFF} based on an average W/L = 3.

Although it is difficult to make generalized predictions about the absolute value of P_R at any particular node, it can be seen from (4.10) and (4.14) that it will change from node to node as $\left((P_S)_{n+1}/(P_S)_n\right)/\left((P_D)_{n+1}/(P_D)_n\right)$, so that $(V_{DD}I_S)_{n+1}(V_{DD}I_D)_n/(V_{DD}I_S)_n(V_{DD}I_D)_{n+1}$. As the current ratio $(I_D)_{n+1}/(I_D)_n$ is simply V^{β} and $(I_S)_{n+1}/(I_S)_n = V^{\eta}$, this becomes:

$$P_R = (P_R)_0 V^{\eta-\beta}. \tag{4.43}$$

Fixing an initial value for the power ratio $(P_R)_0$ also sets the power ratio at each scaling node depending on the successive values of β and η (determined, in turn, by the supply the threshold scaling, V and b), so that $P_R \propto V^{(\eta-\beta)}$. Substituting (4.43) into (4.42) results in:

$$P_T = \left(\frac{1}{1+(P_R)_0 V^{(\eta-\beta)}}\right)\left(A^{\chi}A^{(\sigma-\chi)/\sigma} + (P_R)_0 V^{(\eta-\beta)}A^{\chi'}A^{(\sigma-\chi')/\sigma}\right) \tag{4.44}$$

For a particular χ and χ', (4.44) implies that the scaling of dynamic power can be balanced against that for static power via a choice of σ and that its final value will depend on a combination

of the target power scaling (P_T), the subthreshold to dynamic power ratio P_R, plus the relative scaling of the supply and threshold voltages.

In summary, the conditions under which constant or reducing power may be achieved with increasing area are given by (4.29), (4.33) and (4.42). It can be seen that there are three basic contributions to all of these:

1. a parameter (σ) that links performance (frequency) to area, determined by the system architecture or micro-architecture;
2. two terms that describes the dynamic and subthreshold I-V scaling characteristics of a particular technology, i.e. $\chi = (\beta+1)/(\beta-1)$ for drive current and $\chi' = (\eta+1)/(\beta-1)$ for subthreshold current;
3. a parameter (γ) that models the impact of capacitance scaling with area.

Ideally, χ and χ' will be large as possible as it will then be easier to select an architecture that satisfies (4.31) and (4.35). From (4.42), a value of σ may exist that optimizes total power scaling. The following section will explore these relationships further using predictive data from various sources to determine some likely future values for all of these parameters. The objective is to forecast the constraints on power and energy vs. area for some realistic scaling scenarios.

4.4.5 Power and Energy vs. Area—Examples from the Roadmap

In this section, some estimates from the Predictive Technology Models (PTM) of [80] combined with ITRS figures are used to determine a realistic range for the three parameters β, η and γ, and thus their impact on power and energy in future architectures. The issue of variability is also reintroduced.

This analysis is based on the supply and threshold voltage characteristics of two of the ITRS technology classes: High Performance (HP) and Low Operating Power (LOP) and explores the effect of moving the behavior of the supply and threshold voltage scaling away from that predicted in the ITRS. The baseline used below is the (circa 2005) 90nm node where

V_{DD}(HP) = 1.2V, V_{TH}(HP) = 200 mV and V_{DD}(LOP) = 0.9V, V_{TH}(LOP) = 260 mV. All of the scaling parameters derived below are referenced to these points.

Low-Power Scaling and σ

Both χ and χ' depend directly on the scaling relationship between V_{DD} and V_{TH}, characterized in this model by β and η. The mechanism for determining β and η is illustrated in Table 12. The first column to the left of the table contains the supply voltages for the given scaling assumption (V = 0.85, in this example). Five successive supply entries are shown. In the adjacent rows are threshold voltages derived as $V_{TH} = a\text{-}bV_{DD}$ for the various values of a and b shown, and these result in the entries for $I_D(sat) = (V_{DD}\text{-}V_{TH})^{1.25}$ from (4.12) in the next five rows. The saturation current entries were approximated to kV^{β} using curve fitting software and the resulting approximations appear in the next five rows, followed by the relative error between each drive current and its fitted approximation. Finally, from (4.28), $\chi = \frac{\beta+1}{\beta-1}$. The maximum error tends to increase with more aggressive voltage and threshold scaling and reaches ±12% in this particular example. Around this error range (~±12–15%), the simple power-law approximations break down so that the actual values derived for β and η will be unreliable, although their general trend will still be indicative. The next set of entries in the table represent the subthreshold current, evaluated as $I_{SUB} = e^{-V_{TH}/nV_t}$ and its approximation $I_{SUB} \approx V^{\eta}$ fitted in the same way as before so that $\chi' = \frac{\eta+1}{\beta-1}$. Table 12 represents the case for one supply scaling value (V = 0.85) assuming no V_{TH} variability. Similar calculations were performed for each set of supply scaling values, firstly with no variability, then assuming a maximum threshold offset due to variability of +25%. Finally, the exponent α was varied in the range $1.05<\alpha<1.25$.

Table 12 Example supply-threshold voltage scaling, approximations with V_{DD}(90nm)=0.9V; V_{TH}(90nm)=0.26.

V_{DD}	0.215	0.26	0.282	0.305	0.327	0.350	0.372	0.395	0.417	0.440	a
0.85	**-0.05**	**0**	**0.025**	**0.05**	**0.075**	**0.1**	**0.125**	**0.15**	**0.175**	**0.2**	**b**
0.90	0.260	0.260	0.260	0.260	0.260	0.260	0.260	0.260	0.260	0.260	
0.77	0.254	0.260	0.263	0.267	0.269	0.273	0.276	0.280	0.282	0.286	
0.65	0.248	0.260	0.266	0.273	0.278	0.285	0.291	0.298	0.303	0.310	$V_{TH} = a - bV_{DD}$
0.55	0.243	0.260	0.268	0.278	0.286	0.295	0.303	0.313	0.321	0.330	
0.47	0.239	0.260	0.270	0.282	0.292	0.303	0.313	0.325	0.335	0.346	
	0.572	0.572	0.573	0.572	0.573	0.572	0.573	0.572	0.573	0.572	
	0.438	0.431	0.428	0.424	0.421	0.417	0.414	0.410	0.408	0.404	
	0.321	0.308	0.303	0.296	0.290	0.284	0.278	0.272	0.266	0.260	$I_D(sat) = (V_{DD}-V_{TH})^{1.25}$
	0.229	0.213	0.205	0.197	0.189	0.181	0.174	0.166	0.159	0.151	
	0.161	0.142	0.134	0.124	0.116	0.107	0.099	0.090	0.082	0.074	
K	0.723	0.746	0.759	0.772	0.788	0.804	0.823	0.843	0.866	0.892	
β	**1.96**	**2.14**	**2.24**	**2.35**	**2.46**	**2.58**	**2.71**	**2.85**	**2.99**	**3.16**	
	0.589	0.595	0.599	0.603	0.608	0.612	0.618	0.624	0.632	0.640	
	0.434	0.426	0.422	0.418	0.414	0.409	0.405	0.400	0.396	0.391	
	0.311	0.296	0.289	0.280	0.273	0.264	0.256	0.247	0.239	0.229	$I_D(sat) \approx kV^{\beta}$
	0.225	0.207	0.199	0.189	0.181	0.171	0.163	0.153	0.145	0.135	
	0.165	0.148	0.140	0.131	0.123	0.114	0.106	0.098	0.090	0.082	
	-2.84%	-3.92%	-4.54%	-5.27%	-6.04%	-6.93%	-7.92%	-9.06%	-10.31%	-11.78%	
	0.91%	1.21%	1.37%	1.56%	1.75%	1.98%	2.22%	2.50%	2.78%	3.15%	
ε	2.85%	3.92%	4.55%	5.27%	6.04%	6.94%	7.92%	9.07%	10.30%	11.79%	error (%) = $(V_{DD}-V_{TH})^{1.25} - V^{\beta}$
	1.90%	2.74%	3.28%	3.89%	4.57%	5.39%	6.34%	7.45%	8.73%	10.31%	
	-2.86%	-3.95%	-4.51%	-5.27%	-6.02%	-6.98%	-7.88%	-9.11%	-10.28%	-11.81%	
χ	**3.09**	**2.75**	**2.61**	**2.48**	**2.37**	**2.26**	**2.17**	**2.08**	**2.00**	**1.93**	$\frac{\beta+1}{\beta-1}$
	2.53E-03	2.53E-03	2.56E-03	2.53E-03	2.56E-03	2.53E-03	2.56E-03	2.53E-03	2.56E-03	2.53E-03	
	2.94E-03	2.53E-03	2.37E-03	2.18E-03	2.04E-03	1.88E-03	1.76E-03	1.61E-03	1.52E-03	1.39E-03	
	3.37E-03	2.53E-03	2.22E-03	1.90E-03	1.66E-03	1.42E-03	1.25E-03	1.07E-03	9.35E-04	8.01E-04	$I_{SUB} = e^{-V_{TH}/nV_t}$
	3.78E-03	2.53E-03	2.09E-03	1.69E-03	1.40E-03	1.13E-03	9.35E-04	7.56E-04	6.25E-04	5.05E-04	
	4.15E-03	2.53E-03	2.00E-03	1.54E-03	1.22E-03	9.41E-04	7.43E-04	5.74E-04	4.53E-04	3.50E-04	
	2.38E-03	2.53E-03	2.64E-03	2.69E-03	2.80E-03	2.85E-03	2.97E-03	3.02E-03	3.15E-03	3.21E-03	
η	**-0.761**	**0.000**	**0.380**	**0.762**	**1.142**	**1.522**	**1.903**	**2.284**	**2.664**	**3.045**	
	2.58E-03	2.53E-03	2.53E-03	2.48E-03	2.48E-03	2.43E-03	2.43E-03	2.38E-03	2.38E-03	2.33E-03	
	2.90E-03	2.53E-03	2.39E-03	2.20E-03	2.08E-03	1.91E-03	1.81E-03	1.67E-03	1.57E-03	1.45E-03	
	3.30E-03	2.53E-03	2.24E-03	1.93E-03	1.71E-03	1.48E-03	1.31E-03	1.13E-03	1.00E-03	8.64E-04	$I_{SUB} \approx kV^{\eta}$
	3.75E-03	2.53E-03	2.10E-03	1.70E-03	1.41E-03	1.15E-03	9.52E-04	7.72E-04	6.41E-04	5.20E-04	
	4.23E-03	2.53E-03	1.98E-03	1.51E-03	1.18E-03	9.03E-04	7.06E-04	5.39E-04	4.22E-04	3.22E-04	
	-2.00%	0.00%	0.99%	1.98%	2.98%	3.97%	4.96%	5.98%	6.96%	7.94%	
	1.10%	0.00%	-0.55%	-1.08%	-1.60%	-2.12%	-2.64%	-3.11%	-3.62%	-4.11%	
	1.99%	0.00%	-1.01%	-1.99%	-2.98%	-3.98%	-4.98%	-5.94%	-6.95%	-7.95%	error (%) = $e^{-V_{TH}/nV_t} - V^{\eta}$
	0.80%	0.00%	-0.40%	-0.75%	-1.11%	-1.48%	-1.84%	-2.14%	-2.50%	-2.83%	
	-1.98%	0.00%	0.97%	2.00%	3.00%	3.98%	4.96%	5.99%	6.95%	7.94%	
χ'	**0.25**	**0.87**	**1.11**	**1.30**	**1.47**	**1.59**	**1.70**	**1.78**	**1.84**	**1.87**	$\frac{\eta+1}{\beta-1}$

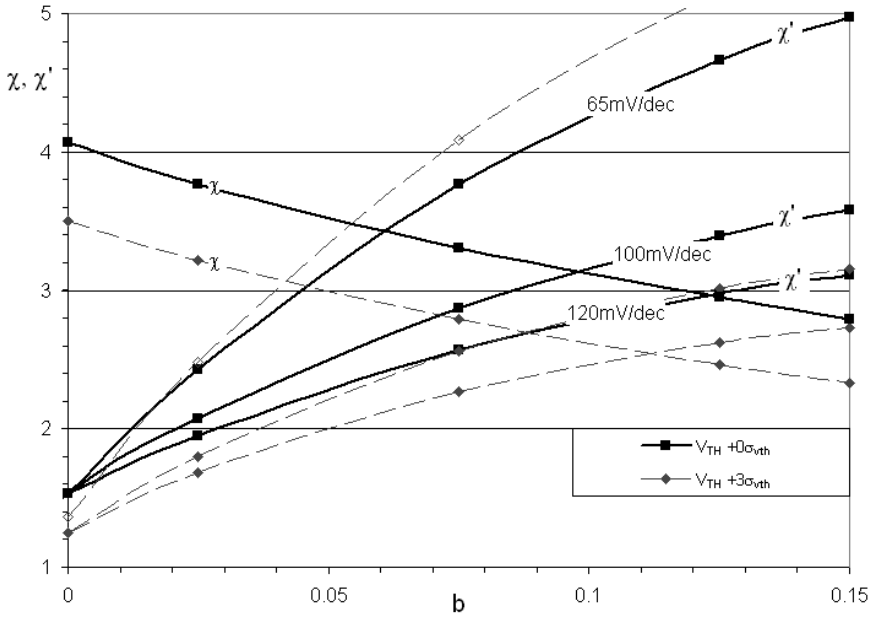
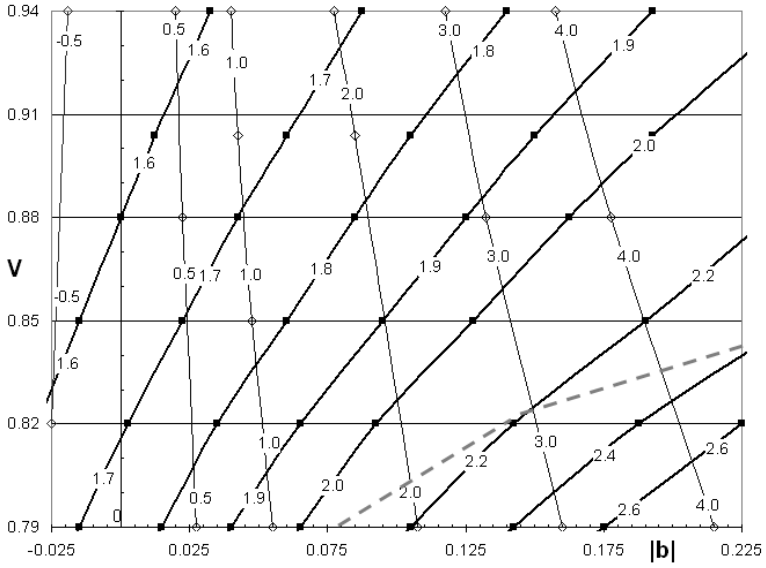


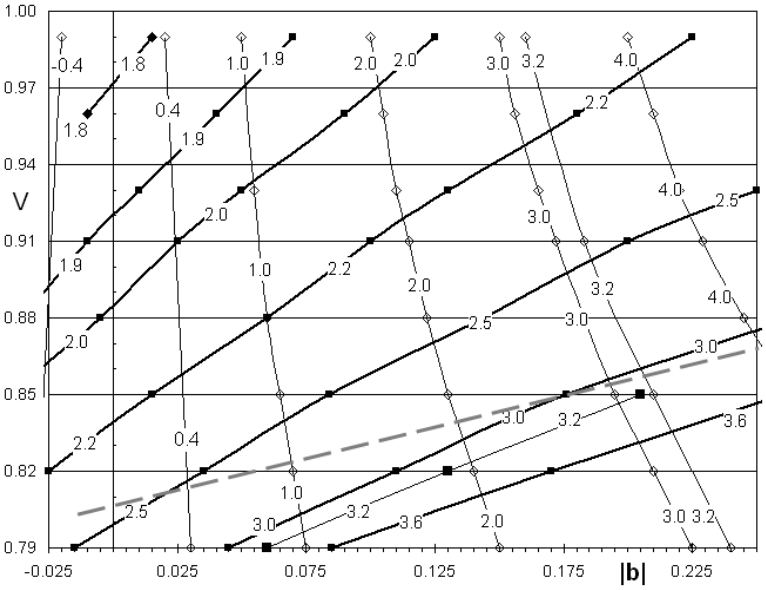
Figure 78. χ and χ' vs. *b* for supply scaling (V) = 0.84
HP technology (V_{DD0}=1.2V, V_{TH0}=0.2V) showing the effect of subthreshold slope and threshold variability.

The typical form of the scaling relationship is illustrated in Figure 78, which shows χ and χ' for a range of *b* (*>0*) at a single technology and supply scaling (in this example, V_{DD0} = 1.2, V = 0.84). In general terms, a more aggressive threshold scaling (higher b) results in higher values of β because of its impact on I_D(sat). Thus the value of χ falls with increasing b. From (4.31), the dynamic scaling term in (4.42) will be constant when $\sigma = \chi/(\gamma\chi + 1)$ (see Figure 75). An identical relationship exists between σ and χ' based on (4.35). Increasing *b* causes subthreshold current to fall at an increasing rate, so that χ' will also increase. The two curves intersect where $\chi = \chi$' or, equivalently, $\beta = \eta$. It can be seen that variations in subthreshold slope actually have a minor impact on σ as it affects only the χ' curve. In this example, moving the slope ±20 mV/decade centered around 100 mV/decade causes the intersection point to change less than ±6%. The impact of this slope variation will further reduce almost linearly with smaller P_R and thus will be ignored in the following analysis. On the other hand, threshold variability will have a larger impact as it reduces the values of both χ and χ' at a given value of *b*. It can be seen in Figure 78

that a +25% offset in V_{TH} will cause the intersection points to move down by about $\chi = 0.5$ (around 15%), independent of S.



(a)



(b)

Figure 79. Contour plots of β (filled squares) and η (open diamonds) vs. supply scaling factors (V) and V_{TH} scaling factor (b), $V_{TH} = a - bV_{DD}$, $-0.025 \le b \le 0.25$ (a) HP: $V_{DD}^{90nm} = 1.2V$, $V_{TH}^{90nm} = 0.2V$. (b) LOP: $V_{TH}^{90nm} = 0.26V$, $V_{DD}^{90nm} = 0.9V$.

The two contour plots in Figure 79 illustrate the likely range for β and η based on the initial (90nm) supply and threshold values for the ITRS LOP and HP technologies. These plots were

developed using the same method as in Figure 78, over various supply scaling factors (V) and threshold scaling factors (*b*) and with the initial values of V_{DD} and V_{TH} shown. The subthreshold slope was fixed at S = 100 mV/decade in all cases and variability ignored. As mentioned, the curve fitting errors increase with supply scaling and larger *b*, exceeding approximately ±12% below and to the right of the dashed line on each plot, representing the putative accuracy limit for this model. As expected, the primary difference between these plots is the range and slope of β and η with the reduced gate overdrive (V_{DD}-V_{TH}) of LOP technology increasing its sensitivity to movements in both supply and threshold.

The total power scaling at each (V, b) point in Figure 79 can be derived by substituting (4.27) and (4.43) into the equations used to derive (4.42) so that power scaling becomes:

$$P_T = \frac{A}{1+(P_R)_0 V^{\eta-\beta}}\left(V^{\beta+1}+(P_R)_0 V^{2\eta-\beta+1}\right) \tag{4.45}$$

where $(P_R)_0$ is the initial relative power (e.g. at 90nm) for the particular technology.

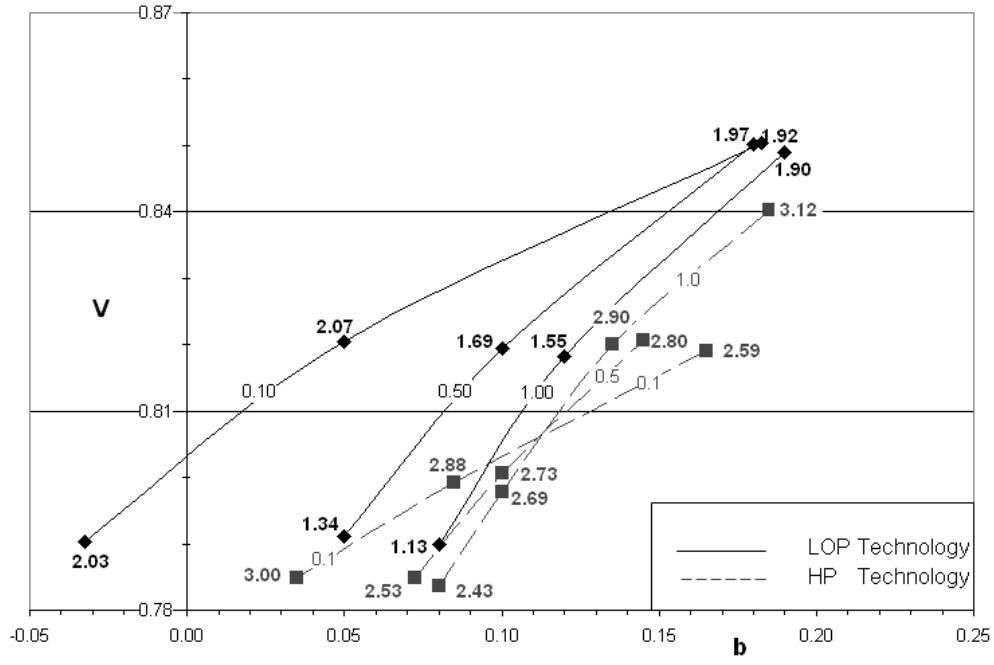


Figure 80. Approximate loci of P_T=1.0 in (4.45) for LOP and HP technologies, with initial relative power $(P_R)_0$=0.1, 0.5 and 1.0. A=1.92, no variability.

A set of (V, b) loci that result in $P_T = 1.0$ in (4.45) with A = 1.92 and $(P_R)_0$ values of 0.1, 0.5 and 1.0, are shown in Figure 80. The β and η data are drawn from Figure 79. Although the actual value of $(P_R)_0$ for a particular technology will depend on a number of indeterminate system-level parameters, it can be estimated from their initial V_{TH}/V_{DD} ratios that $(P_R)_0$ is likely to range around 0.5 for HP (see Table 11) and less than 0.1 for LOP technologies. From (4.45), it is expected that the curves for each P_R in Figure 80 will converge at the points where $\beta = \eta$ for a given technology. For HP this occurs at $\beta = \eta \approx 2.1$ at $V \approx 0.805$, whereas for the LOP case $\beta = \eta \approx 3.1$ with $V \approx 0.85$. At these points, the overall power will be independent of the relative power term.

The additional data labels on Figure 80 (in bold) are the values of σ that result in $P_T = 1$ in (4.42) at each point shown. In most cases, σ increases monotonically with increasing V and b. The exception is the LOP curve with $(P_R)_0 = 0.1$ where the peak occurs at $b \approx 0.05$, although the variation along this line is very small, less than 5% across the range of V and b. The impact of P_R on σ is most clearly seen in the HP curve. Where the relative contribution of subthreshold power is small, the best case (i.e., largest σ) tends to result primarily from scaling of supply voltage. This has been the case for traditional circuit design to date—adjusting supply with a fixed (or decreasing) threshold. One example of this is the point at (V, b) = (0.785, 0.035) for which $\sigma = 3$ (i.e., $\beta \approx 2$, so that $\chi \approx 3$). As the contribution from subthreshold power increases, the threshold voltage must increase to compensate, with an effect on performance so that the peak σ reduces and occurs at higher values of 'b'.

Figure 81 overlays the $P_T = 1$ solution curves from (4.45) onto the contour plots of χ and χ' at various area scaling factors, in this case including +25% variability (i.e., $V_{TH} = 1.25(a - bV_{DD})$. Each curve for $(P_R)_0 = 0.1$ and 0.5 therefore describes the scaling of supply and threshold that will result in constant power for the various area scaling factors. Comparing the curves for A = 1.9 between Figure 80 and Figure 81, it can be seen that the larger V_{TH} values in the latter case cause the intersection point to occur at a higher supply scaling value (0.87 vs. 0.85). The values of σ for

P_T = 1 are shown at a number of points on Figure 81 (enclosed by a rectangle) illustrating in general terms how σ reduces with increasing area scaling.

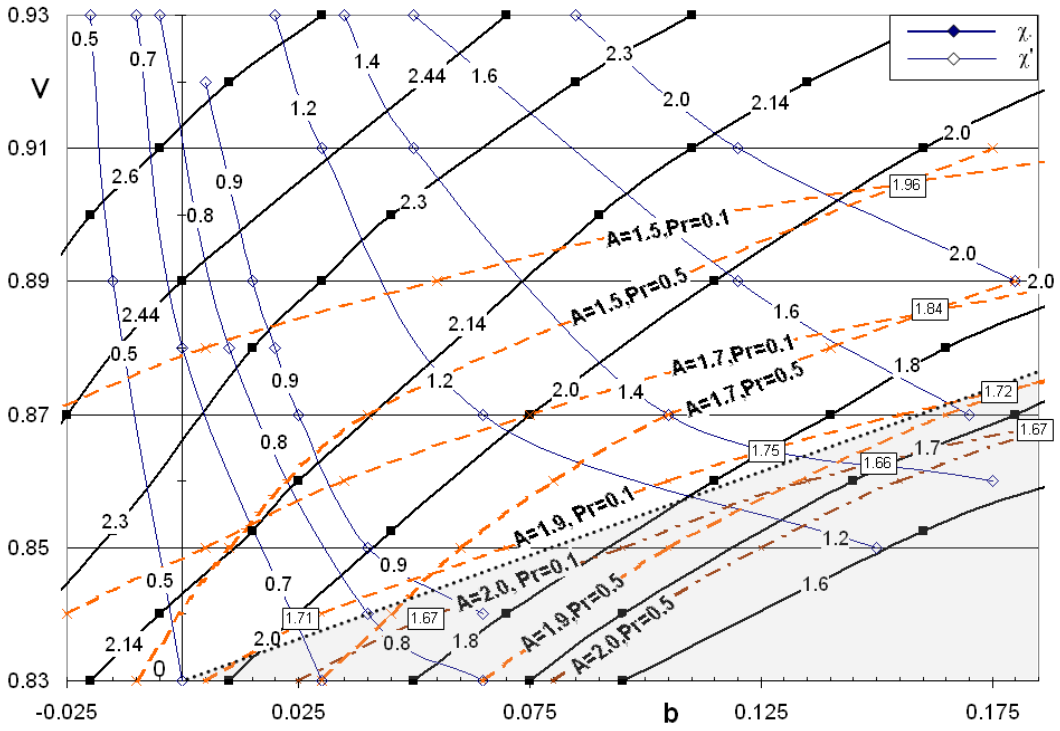


Figure 81. Contour plots of χ (filled squares) and χ' (open diamonds)
$V_{DD}^{90nm} = 0.9V$, $V_{TH}^{90nm} = 0.26V$, $V_{TH} = 1.25(a-bV_{DD})$, $-0.025 \le b \le 0.175$.
Enclosed numbers are σ_{MAX} resulting in P_T=1 at specific points
Curve fitting errors exceed ±12% inside the dotted line to the lower right.

Figure 82 shows χ and χ' vs. the threshold scaling factor 'b' for a single value of supply scaling (LOP technology, V = 0.85). The central pair of curves are with α fixed at 1.25 and the impact of variability is shown by the ±25% error bars around the V_{TH} curve (only a few are shown to avoid cluttering the diagram). As seen previously in Figure 78, the $+3\sigma_{vth}$ process corner reduces the worse-case χ and χ' values by as much as 25%, reflecting the fact that this corner constrains the operating frequency (for a synchronous system) and therefore its overall performance. Figure 82 also illustrates the general impact of reducing α, the velocity saturation exponent in (4.12). Compared to the original curve with α fixed at 1.25, the "reducing α" curves were derived by successively decreasing in α by 0.05 per node, from 1.25 to 1.05, simulating increasingly ballistic

transistor behavior. The result is a significant increase in both χ and χ' for a given value of *b*, something that is entirely expected given the improved I-V performance that will result from ballistic operation.

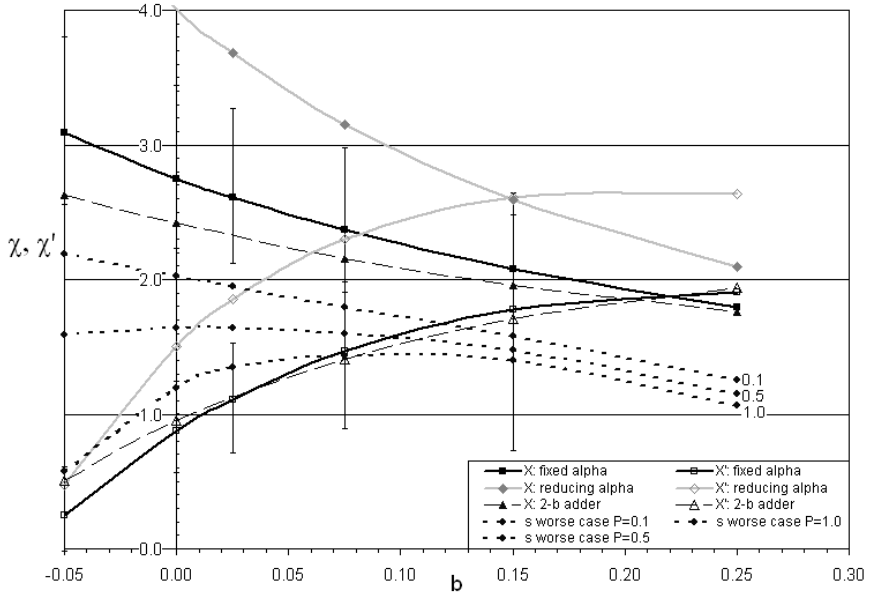


Figure 82. χ and χ' vs. *b* for supply scaling V = 0.85
(i) α fixed at 1.25; (ii) α decreasing by 0.05 at each successive node. The error bars show the shift caused by a ±25% variation in V_{TH} around its mean value. Also shown are χ and χ' derived from a simulated 2-bit adder circuit (dashed lines), along with the worse-case σ (α=1.25) that results in $P_S/P_D = 1$ with $P_T = 1$ in (4.42).

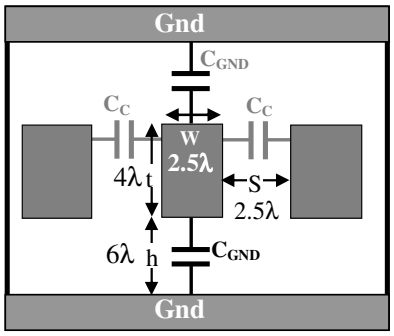
To verify the basic model (at fixed α), a simple 2-bit adder CMOS circuit was developed in VHDL-AMS using the EKV models of [333] and simulated with various supply voltages from 0.9V down to about 0.5V (V = 0.85) and with V_{TH} values determined by 'b'. Gate delay and static power were measured and least-squares error fits applied to V_{DD} vs. $I_D = V_{DD}/\tau$ (C_L = constant) and V_{DD} vs. $I_S = P_S/V_{DD}$ to determine β and η respectively. The results (dashed lines in Figure 82) show that both curves exhibit the same trend as the model. However, the range of α for these particular transistor models was determined to be 1.3 to 1.4, and it was also found to vary slightly with the value of gate overdrive (V_{DD}-V_{TH}). Thus while the curve for χ' is a close fit, χ is up to 15% lower than predicted by the model.

The final pair of (dotted) curves in Figure 82 represent the value of σ that satisfies (4.42) for constant total power scaling (i.e., $P_T = 1$) with $P_S/P_D = 0.1$, 0.5 and 1.0 as labelled, and assuming a worse-case $+3\sigma_{vth}$ deviation from the values of $\chi$and χ' predicted by the model. As expected, when the ratio of dynamic to static power is small (0.1), σ closely tracks χ. On the other hand, when these two components become equal, σ exhibits a broad maximum (approximately 1.4 in this example) across a wide range of 'b'. From Figure 76 it can be seen that $\sigma \approx 1.4$ means that each doubling in area (A = 2) allows the operating frequency to be reduced by around 0.6.

4.4.6 Node Capacitance Estimates

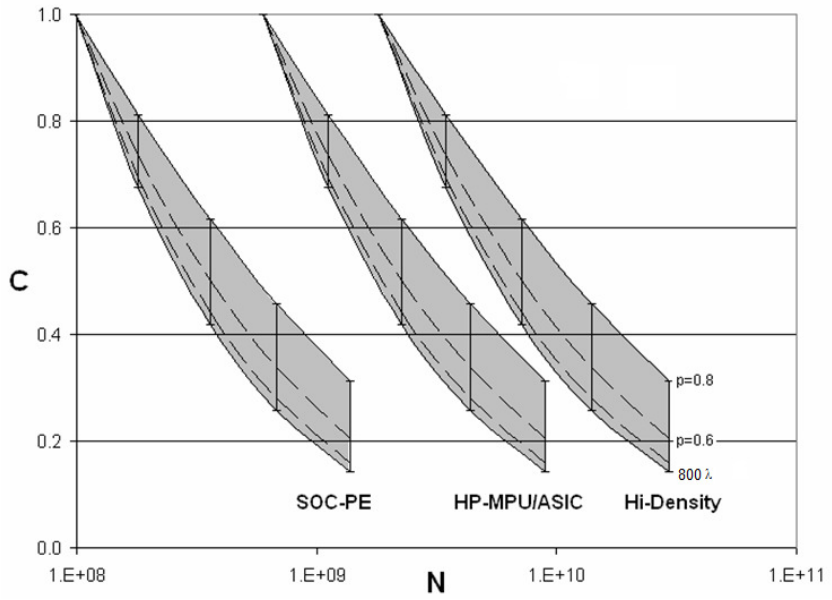
Under realistic assumptions, predictive technology models [80] forecast that the capacitance per unit area for an average interconnect will tend to reduce at successive technology nodes due to the combined effect of reductions in line dimensions, field oxide thickness and effective dielectric constant (κ_{EFF}). As it appears likely that the physical area and thickness terms will scale in approximately the same ratio, it will be reductions in effective dielectric that will have the primary impact on the interconnect capacitance.

For example, the total capacitance values derived in Figure 83a were generated using the models of [80] assuming that the line dimensions are proportional to feature size. Here, λ = 0.5*Feature Size and κ_{EFF} reduces at each node in line with ITRS predictions, so that the total interconnect capacitance reduces as: $C_{INT} \propto \lambda^{0.4}$. Figure 83b plots the capacitance of representative lengths of local interconnect against the predicted number of devices (N) at successive nodes for three ITRS system drivers: power efficient system on chip (SOC-PE), high performance microprocessor/ ASIC (HP-MPU) and high density (memory intensive) architectures (labelled 'hi-density'). In each case, the analysis uses the total capacitance per unit area in Figure 83a and starts with a fixed line 6λ wide by 800λ long which defines the lower boundary of each capacitance curve (labeled '800λ' in Figure 83b). The general trend here is $C_L \propto N^{-0.7}$ i.e., $\gamma = -0.7$.



Feature Size	90nm	65nm	45nm	32nm	22nm	16nm
W=2.5λ (nm)	110	76	54	38	27	20
S=2.5λ (nm)	110	76	54	38	27	20
T=4λ (nm)	180	130	97	72	54	40
H=6λ (nm)	270	195	135	96	66	48
$\mathbf{K_{EFF}}$	3.4	2.9	2.6	2.4	1.7	1.5
R (Ω/mm)	1111	2227	4200	8041	15090	27500
L (nH/mm)	1.9	1.9	2.0	2.1	2.1	2.2
$\mathbf{C_{GND}}$ (fF/mm)	24	24	21	20	14	11
$\mathbf{C_c}$ (fF/mm)	76	70	62	60	44	39
$\mathbf{C_{TOT}}$ (fF/mm)	200	187	167	160	117	100

(a) Interconnect model structure, dimensions and resultant RLC parameters.



(b) Normalized interconnect capacitance vs. N for various technologies.

Figure 83. Interconnect capacitance (C) at successive technology nodes vs. predicted device numbers (N) for some ITRS system drivers. The capacitance scaling region for each (shaded) is bounded by (i) a fixed 3F wide x 400F long interconnection line (F = minimum feature size) based on predictive technology models of [80] and (ii) load capacitance with L_{AVE} as predicted by [334] for Rent exponent p = 0.8. C is normalized to its value at the 90nm (2005) node.

The length of this line was then made a function of N using the stochastic wire-length model of [334]. This model predicts that the interconnect length (and therefore its capacitance) will grow slowly with the number of gates (N), especially for larger circuits. Although the full model is a complex function of N and the Rent exponent *P*, even for moderate values of N (e.g., N > ~10^3), it quickly asymptotes to a simpler power-law form $L_{AVE} = kN^{p'}$ ($p' << P$) where L_{AVE} is measured in

gate pitches, proportional to feature size (Figure 84). As interconnect capacitance will tend to dominate gate capacitance except for extremely localized connections, the general trend will be $C \propto N^{p'} \propto L_g L_{AVE} C_{INT}$ (L_g = gate length and C_{INT} = line capacitance per unit length). For example, the average interconnect length for a Rent exponent of p = 0.6 (typical for microprocessor and similar micro-architectures) is $L_{AVE} \propto N^{0.135}$ (gate pitches) implying that a doubling of N might result in an approximately 10% increase in interconnect capacitance. The upper bound on the shaded regions in Figure 83b represent the capacitance at P = 0.8, for which $C_L \propto N^{0.3}$ i.e., $\gamma = 0.3$.

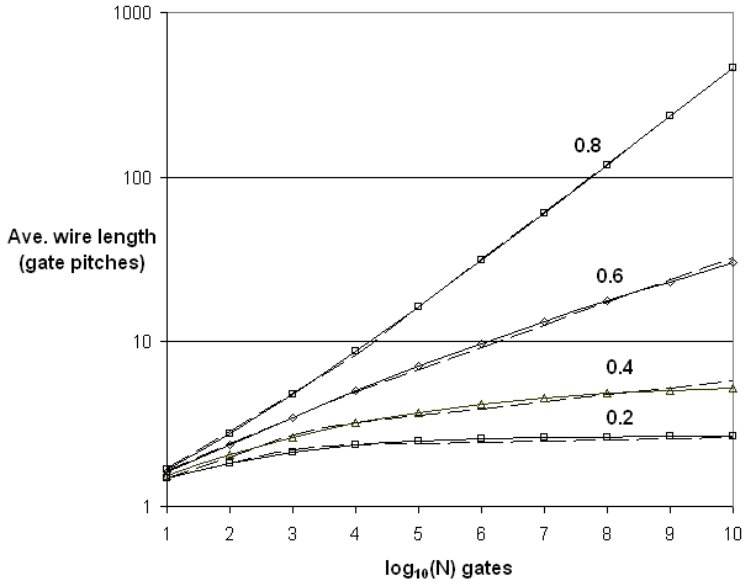


Figure 84. Average wire length as predicted by model of [334] (solid line) compared with simplified Rent model – $kN^{p'}$ (dashed line).

Although the evolution of γ will depend on a range of issues such as the integration of low-κ dielectrics, it appears that it will be relatively insensitive to the scaling of N. While the overall range shown here is about $-0.74 \leq \gamma \leq -0.42$ for the particular assumptions made regarding the interconnect growth (i.e., fixed length or fixed Rent exponent), γ remains remarkably constant, varying less than 5% across these three examples. For example, with p = 0.8, $-0.443 \leq \gamma \leq -0.428$ while at p = 0.4, $-0.694 \leq \gamma \leq -0.657$.

It can also be noted that, as expected, the curve for $p = 0.4$ is close to the fixed length curve, reflecting the low fanout and localized connectivity of architectures with small Rent exponents. As a result, it may be safely assumed that $-0.7 \leq \gamma \leq 0.3$ will encompass the overall range of capacitance scaling due to both technology and circuit considerations.

4.4.7 Applying the Model

The overall objective of this analysis is to determine the conditions under which parallelism may be exploited to reduce power. Thus it is necessary to ask whether it is actually worthwhile to invest additional transistors in producing a parallel version of the system. This will depend on both *architecture*: how much (performance) 'return' is received for an (area) investment; and *technology*: can β and η be set such that they satisfy the inequalities in (4.31) and (4.35) that will allow this extra performance to be exchanged for a reduction in power and/or energy?

Table 13 separates out the various technology impacts on the power calculations of (4.29) and (4.33). Here, $\sigma(\max)$ is the largest σ that results in $P_T = 1.0$, given the constraints on supply scaling, β and η shown in Figure 79. This was calculated by substituting a range of V ($0.79 \leq V \leq 0.91$), *b* ($-0.05 \leq b \leq 0.25$), γ and $(P_R)_0 \propto V^{\eta-\beta}$ into (4.42) and iteratively determining the largest σ for which $P_T = 1.0$ as the power balance was shifted between dynamic, at $(P_R)_0 = 0.1$ and sub-threshold, with $(P_R)_0 = 2$. At low $(P_R)_0$, power is mostly constrained by $NFCV^2$ and the most aggressive supply scaling available ($V = 0.79$) results in maximum σ. For $(P_R)_0 > 0.25$, the maxima of the σ curves for each V across the range of *b* are fairly constant, as was seen previously in Figure 82 for $V = 0.85$. This can be seen by comparing rows 4-7 in Table 13.

The HP technology case follows the same general form but at generally higher values of σ_{MAX}, as the larger initial gate overdrive (V_{DD}-V_{TH}) will support more aggressive supply and threshold scaling. Alternatively, it would be easier to exchange increased architectural performance (lower σ) for power in this technology. For example, with fixed α, $\gamma = 0$ and $P_R = 0.5$, the entries at row 25 imply that V and 'b' could be set such that an architecture with $\sigma \approx 2.8$ would support scaling at constant total power. With the supply and threshold settings of $V = 0.82$ and $b = 0.145$, the

scaling of dynamic power $P_{DYN} \propto A^{(2.8-2.66)/2.8} \approx 1.04$ is offset by that for the subthreshold case, $P_{SUB} \approx 0.93$, so that the total power remains at unity. On the other hand, an architecture capable of $\sigma = 2$ would allow frequency to be scaled by approximately 0.71 (from (4.4)) and supply voltage by 0.76 (from (4.37)), resulting in a total power scaling from (4.42) of $P_T \approx 0.77$. Under these conditions, total energy, given by the weighted sum of (4.30) and (4.34), would remain approximately constant.

With V_{TH} at its $+3\sigma_{vth}$ extreme, which models the impact of the maximum expected variability on the critical path, σ(max) reduces, but not by as much as might be expected. For example, comparing lines 25 and 31 for which $(P_R)_0 = 0.5$, σ(max) moves from ~2.8 to ~2.43, a shift of about 13% and still outside the range $1 < \sigma < 2$ typical of conventional architectures. Assuming increasingly ballistic device operation (variable α), increases this again to about ~2.6 (row 38), about 7% lower than the original value. It can be concluded that, although variability may have a direct and significant effect on circuit performance, it will have a smaller impact on power scaling. The combined effect of variability and α on σ(max) in Table 13 never exceeds ~15%, a figure that is likely to be recoverable at the architectural and/or micro-architectural level.

Table 13 also shows the impact of varying load capacitance, modeled here as γ. The two examples given in the table approximate the cases where load increases due to the effect of interconnection growth (γ = 0.25–0.3) and where it reduces at successive nodes due to smaller device sizes and advances in technology (γ = -0.65 to -0.7). In the case of LOP technology, the impact of $\gamma = 0.3$ (modeling increasing interconnect length) is fairly severe in that the worse-case σ(max) is reduced to less than 1.2 (see lines 11 and 12) which would severely limit the available architectural options. The equivalent effect in the HP case is smaller: σ(max) is reduced to ~1.6, in the middle of the range for typical circuits.

Table 13 Maximum σ Resulting in P_T=1 for various $(P_R)_0$, β, η and γ.

Tech/ Scaling	P_R	F	σ max	γ	β	V	η	*b*	ΔVth (mV)	χ	χ'	P_D	P_S	P_T (norm)	$NFCV^2$	$NVe^{-\Delta V_{TH}/nV_t}$	$NV^{(\beta+1)}$	$NV^{(\eta+1)}$	
LOP SOC-PE	V_{DD0}=0.9 A=1.92																		
	0.10	0.73	**2.07**	0.00	2.59	**0.82**	0.717	0.050	**6**	**2.28**	**1.05**	**0.95**	**1.39**	**1.00**	**0.95**	**1.32**	**0.95**	**1.39**	1
	0.10	0.60	**1.30**	0.30	2.56	**0.82**	0.645	0.050	**8**	**2.28**	**1.05**	**0.95**	**1.39**	**1.00**	**0.95**	**1.32**	**0.95**	**1.39**	2
	0.10	1.15	**-4.53**	-0.70	2.56	**0.82**	0.645	0.050	**8**	**2.28**	**1.05**	**0.95**	**1.39**	**1.00**	**0.95**	**1.32**	**0.95**	**1.39**	3
fixed α	0.25	0.72	**1.94**	0.00	3.09	**0.85**	2.89	0.180	**24**	**1.96**	**1.86**	**1.00**	**1.03**	**1.00**	**1.00**	**1.02**	**1.00**	**1.03**	4
V_{TH}(mean)	0.50	0.71	**1.92**	0.00	3.09	**0.85**	2.89	0.180	**24**	**1.96**	**1.86**	**0.99**	**1.02**	**1.00**	**0.99**	**0.95**	**0.99**	**1.02**	5
	1.00	0.71	**1.90**	0.00	3.09	**0.85**	2.89	0.190	**26**	**1.96**	**1.86**	**0.98**	**1.02**	**1.00**	**0.98**	**0.92**	**0.98**	**1.02**	6
	2.00	0.71	**1.90**	0.00	3.12	**0.87**	2.97	0.195	**26**	**1.94**	**1.87**	**0.99**	**1.01**	**1.00**	**0.99**	**0.92**	**0.99**	**1.01**	7
fixed α	0.10	0.70	**1.84**	0.00	3.14	**0.85**	1.427	0.025	**10**	**1.94**	**1.14**	**0.97**	**1.28**	**1.00**	**0.97**	**1.30**	**0.97**	**1.28**	8
V_{TH}+3σvth	0.50	0.68	**1.69**	0.00	2.94	**0.82**	1.427	0.100	**15**	**2.03**	**1.26**	**0.88**	**1.18**	**1.00**	**0.88**	**1.12**	**0.88**	**1.18**	9
γ = 0	1.00	0.71	**1.90**	0.00	3.09	**0.85**	2.89	0.190	**26**	**1.96**	**1.86**	**0.98**	**1.02**	**1.00**	**0.98**	**0.92**	**0.98**	**1.02**	10
	0.10	0.58	**1.18**	0.30	3.14	**0.85**	1.430	0.075	**10**	**1.94**	**1.14**	**0.96**	**1.28**	**1.00**	**0.96**	**1.29**	**0.96**	**1.28**	11
fixed α	0.50	0.56	**1.13**	0.30	2.94	**0.82**	1.434	0.10	**15**	**2.03**	**1.26**	**0.88**	**1.18**	**1.00**	**0.88**	**1.12**	**0.88**	**1.18**	12
V_{TH}+3σvth	0.50	1.13	**-5.44**	-0.70	3.09	**0.85**	2.89	0.180	**24**	**1.96**	**1.86**	**0.99**	**1.03**	**1.00**	**0.99**	**0.96**	**0.99**	**1.03**	13
γ ≠ 0	0.50	0.92	**7.82**	-0.70	3.59	**0.81**	0.877	0.05	**9**	**1.77**	**0.72**	**0.74**	**1.30**	**1.00**	**0.74**	**1.35**	**0.74**	**1.30**	14
	0.10	0.75	**2.27**	0.00	2.55	**0.83**	2.03	0.115	**18**	**2.29**	**1.95**	**0.99**	**1.09**	**1.00**	**0.99**	**1.08**	**0.99**	**1.09**	15
Red. α	0.25	0.74	**2.21**	0.00	2.67	**0.84**	2.80	0.195	**22**	**2.19**	**2.33**	**1.00**	**0.97**	**1.00**	**1.00**	**0.98**	**1.00**	**0.97**	16
V_{TH}(mean)	0.50	0.75	**2.24**	0.00	2.68	**0.84**	2.90	0.195	**22**	**2.19**	**2.33**	**1.01**	**0.97**	**1.00**	**1.01**	**0.99**	**1.01**	**0.97**	17
γ = 0	1.00	0.75	**2.27**	0.00	2.65	**0.84**	2.82	0.190	**21**	**2.21**	**2.32**	**1.00**	**0.99**	**1.00**	**1.00**	**1.02**	**1.00**	**0.99**	18
	0.1	0.59	**1.22**	0.30	2.70	**0.82**	0.356	0.035	**5**	**2.18**	**0.80**	**0.92**	**1.47**	**1.00**	**0.92**	**1.40**	**0.92**	**1.47**	19
red. α	0.50	0.56	**1.11**	0.30	3.37	**0.85**	2.28	0.140	**19**	**1.84**	**1.39**	**0.94**	**1.11**	**1.00**	**0.93**	**1.06**	**0.93**	**1.11**	20
V_{TH}+3σvth	0.50	0.68	**1.66**	0.00	3.37	**0.85**	2.28	0.140	**19**	**1.84**	**1.39**	**0.93**	**1.11**	**1.00**	**0.93**	**1.06**	**0.93**	**1.11**	21
γ ≠ 0	0.50	0.92	**7.82**	-0.70	3.39	**0.81**	0.877	0.050	**8**	**1.77**	**0.72**	**0.74**	**1.30**	**1.00**	**0.74**	**1.31**	**0.74**	**1.31**	22
HP MPU/ASIC	V_{DD0}=1.2 A=1.95																		
	0.10	0.80	**3.00**	0.00	1.93	**0.79**	0.71	0.050	**9**	**3.17**	**1.76**	**0.96**	**1.32**	**1.00**	**0.96**	**1.24**	**0.96**	**1.32**	23
fixed α	0.25	0.79	**2.79**	0.00	2.08	**0.80**	1.75	0.075	**18**	**2.85**	**2.55**	**0.98**	**1.06**	**1.00**	**0.98**	**1.05**	**0.98**	**1.06**	24
V_{TH}(mean)	0.50	0.79	**2.80**	0.00	2.24	**0.82**	2.76	0.145	**24**	**2.66**	**3.12**	**1.04**	**0.93**	**1.00**	**1.04**	**0.95**	**1.04**	**0.93**	25
γ = 0	1.00	0.81	**3.19**	0.00	2.28	**0.85**	4.48	0.190	**34**	**2.56**	**4.28**	**1.14**	**0.80**	**1.00**	**1.14**	**0.77**	**1.14**	**0.80**	26
fixed α	0.1	0.68	**1.71**	0.25	1.92	**0.785**	0.62	0.035	**7**	**3.17**	**1.76**	**0.96**	**1.32**	**1.00**	**0.96**	**1.32**	**0.96**	**1.32**	27
V_{TH}(mean)	0.50	0.68	**1.70**	0.25	2.28	**0.84**	4.48	0.220	**33**	**2.56**	**4.28**	**1.09**	**0.74**	**1.00**	**1.09**	**0.78**	**1.09**	**0.74**	28
γ ≠ 0	0.50	1.23	**-3.23**	-0.65	2.28	**0.84**	4.48	0.220	**34**	**2.56**	**4.28**	**1.09**	**0.74**	**1.00**	**1.09**	**0.78**	**1.09**	**0.74**	29
fixed α	0.10	0.76	**2.42**	0.00	2.39	**0.82**	2.37	0.100	**16**	**2.44**	**2.43**	**0.99**	**1.00**	**1.00**	**0.99**	**1.10**	**0.99**	**1.00**	30
V_{TH}+3σvth	0.50	0.76	**2.43**	0.00	2.57	**0.84**	3.36	0.165	**25**	**2.27**	**2.78**	**1.04**	**0.91**	**1.00**	**1.04**	**0.94**	**1.04**	**0.91**	31
γ = 0	1.00	0.77	**2.58**	0.00	2.60	**0.85**	3.86	0.190	**27**	**2.25**	**3.04**	**1.09**	**0.89**	**1.00**	**1.09**	**0.91**	**1.09**	**0.89**	32
fixed α	0.10	0.66	**1.58**	0.25	2.16	**0.8**	1.03	0.045	**8**	**2.73**	**1.76**	**0.97**	**1.24**	**1.00**	**0.97**	**1.31**	**0.97**	**1.24**	33
V_{TH}+3σvth	0.50	0.64	**1.50**	0.25	2.70	**0.85**	4.20	0.210	**30**	**2.18**	**3.06**	**1.06**	**0.83**	**1.00**	**1.06**	**0.85**	**1.06**	**0.83**	34
γ ≠ 0	0.50	1.20	**-3.74**	-0.65	2.47	**0.84**	3.61	0.145	**21**	**2.36**	**3.13**	**1.07**	**0.88**	**1.00**	**1.07**	**1.02**	**1.07**	**0.88**	35
	0.10	0.76	**2.47**	0.00	2.35	**0.82**	2.26	0.120	**20**	**2.48**	**2.41**	**1.00**	**1.02**	**1.00**	**1.00**	**1.02**	**1.00**	**1.02**	36
red. α	0.25	0.76	**2.43**	0.00	2.39	**0.82**	2.37	0.150	**24**	**2.44**	**2.42**	**1.00**	**1.00**	**1.00**	**1.00**	**0.92**	**1.00**	**1.00**	37
V_{TH}+3σvth	0.50	0.77	**2.59**	0.00	2.47	**0.84**	3.61	0.230	**33**	**2.36**	**3.14**	**1.06**	**0.87**	**1.00**	**1.06**	**0.76**	**1.06**	**0.87**	38
	1.00	0.79	**2.88**	0.00	2.42	**0.85**	3.81	0.150	**20**	**2.41**	**3.39**	**1.12**	**0.89**	**1.00**	**1.12**	**1.04**	**1.12**	**0.89**	39

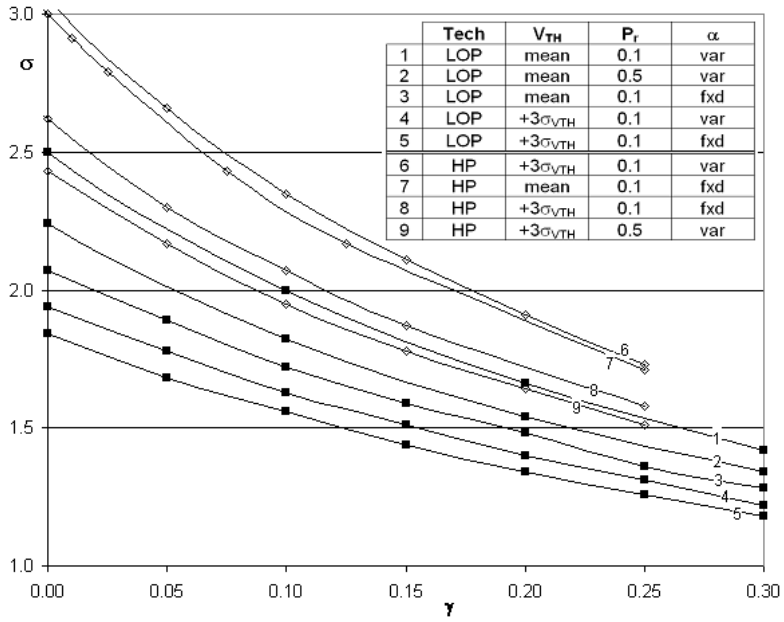


Figure 85. σ(max) vs. γ over a range of technology conditions.

Figure 85 expands on this issue and plots the maximum σ over a range of $\gamma \geq 0$. It can be seen that while the worse-case effect of γ for LOP technology under maximum V_{TH} variability reduces σ(max) to around 1.1, it is still above 1.4 for $\gamma < 0.15$. A figure $\gamma = 0.15$ implies that each doubling of device numbers will result in about a 10% increase in node capacitance. In Chapter 5, it is determined that the reconfigurable platform will exhibit a worst-case σ of approximately 1.4. Limiting the interconnection length such that $\gamma < 0.15$ will support σ values in this range even under maximum threshold variability. On the other hand, where load capacitance decreases at successive nodes, it is possible to exchange part of the expected performance increase for power. In the examples given in the table, the worse-case performance improvement may still be between 7% and 20%, comparable to the ITRS target of 14–17% per scaling node.

As outlined previously in Section 2.7.2, various researchers (e.g., [210]) have argued that the increased delay penalty resulting from lower V_{DD}/V_{TH} ratios will prevent the efficient exploitation of parallel organizations. However, the results of the model developed here indicate that these previous predictions may have been pessimistic. Figure 86 compares the trends of the operating

frequency derived from $F_{MAX} \propto I_D(sat)/CV \propto (V_{DD} - V_{TH})^{\alpha} / N^{\gamma} V_{DD}$, with $F \propto A^{-1/\sigma}$. The former term represents the capability of the technology to achieve a particular operating frequency (as a function of supply, threshold and capacitance) and assuming that the logic depth remains constant with scaling, while the latter is the target frequency scaling determined by power/energy considerations. It can be seen from Figure 86 that in all cases down to the final supply point at around 0.4V (Figure 86c and Figure 86d), $F_{MAX} \geq A^{-1/\sigma}$ indicating that these circuits will be capable of achieving that target frequency over the expected range of supply and threshold voltages.

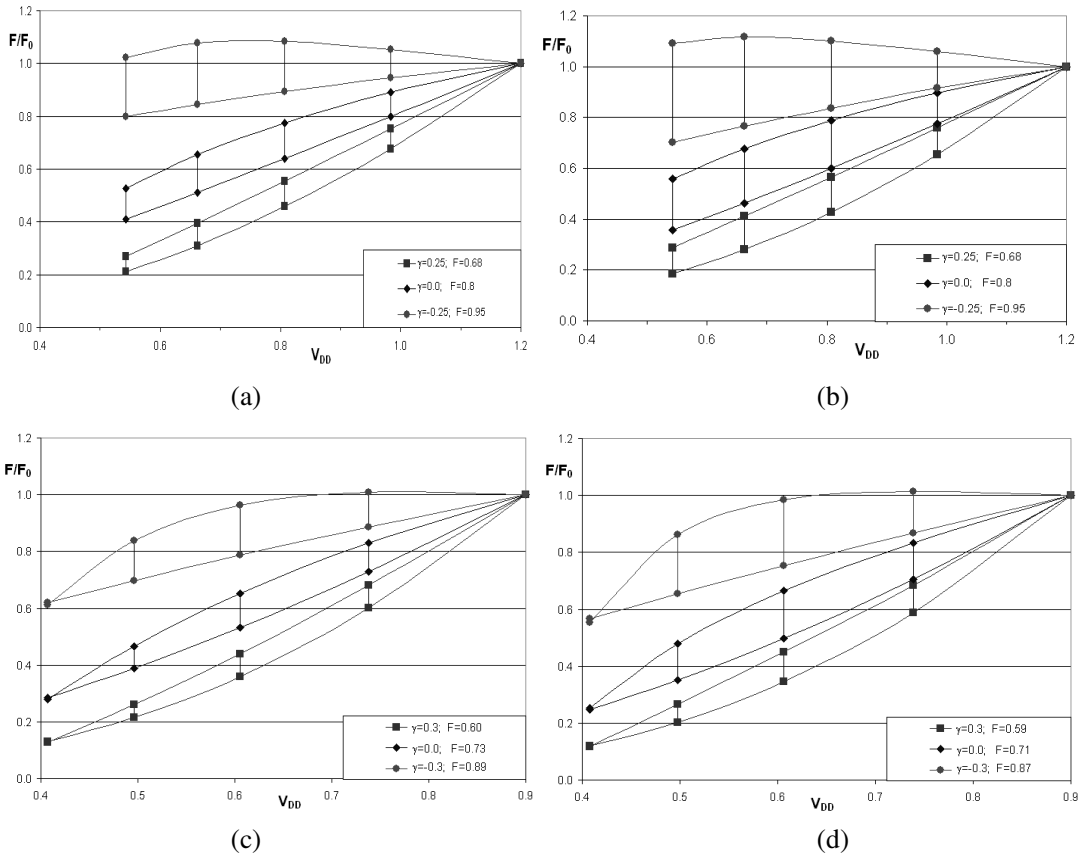


Figure 86. Frequency scaling vs. V_{DD} for P_T=1.0, P_R=0.1
(a) HP: V_{DD0}=1.2, V_{TH0}=0.20; mean (b) HP: V_{DD0}=1.2, V_{TH0}=0.20; V_{TH} =+25%
(c) LOP: V_{DD0}=0.9, V_{TH0}=0.26; mean (d) LOP: V_{DD0}=0.9, V_{TH0}=0.26; V_{TH} =+25%, reducing α.
In all cases, the upper curve of each pair is $F_{MAX} \propto I_D(sat)/CV_{DD}$.

4.5 Summary

It is becoming clear that as on-chip device numbers increase there will be little choice but to manage power consumption by exploiting parallelism. This chapter has developed a new model that describes how area can be traded off against all forms of power consumption in CMOS, especially the dynamic and subthreshold power terms. The objective here has been to determine a simple set of criteria under which this can occur.

It was found that for circuits and/or algorithms that can be characterized in terms of $AT^{\sigma} = K_1$, where $\sigma \leq 2$ is known to hold for many algorithms, dynamic and static power will scale with similar forms: $P_{DYN} \propto A^{\gamma}A^{(\sigma-\chi)/\sigma}$ and $P_{SUB} \propto A^{\gamma'}A^{(\sigma-\chi')/\sigma}$. Here, χ, χ' and γ describe the scaling of drive current, subthreshold current and load capacitance, respectively. The equations for P_{DYN} and P_{SUB} represent an *optimum* scaling case in which changes in supply and threshold voltages result in frequency reductions that are largely compensated by changes to the architecture (e.g., using replicated datapaths). It will be fairly straightforward to derive the values of χ and χ' from small simulations of representative devices and circuits for the given technology or sequence of technologies. Similarly, σ may be determined using a simple architectural–level simulator. Thus all of these parameters can be made available early in the design cycle, making the model a useful way of evaluating high-level architectural tradeoffs in the case where supply and threshold voltages can be adjusted at will.

It is important to note that the model says very little about the *actual* power consumption of a system, which will be a complex function of issues such as device technology, design and layout style, activity ratio as well as the specific values of supply and threshold voltage. The model parameters refer to various *ratios* normalized to a reference technology. The examples in this chapter used the initial (90nm) values for ITRS HP and LOP (i.e., $V_{DD} = 1.2V$, $V_{TH} = 0.2$ and $V_{DD} = 0.9$, $V_{TH} = 0.26$) to show how this choice impacts on the area-power tradeoffs. Similarly, the architectural parameter σ relates area to performance (as $F \propto A^{-1/\sigma}$) for a particular scalable organization, normalized to a baseline configuration.

The parameter σ is often taken to be a measure of the quality of a design and Flynn comments that *"Designs whose AT product is higher than the state of the art are inferior designs"* [170]. While this is undoubtedly true, the analysis in this chapter implies that not only should it be possible to trade area for power for traditional algorithms, but it should be possible to do so with some architectures that may have previously been discounted as sub-optimal (e.g. where $\sigma > 2$). The caveats appear to be the growth of interconnection capacitance and the impact of variability. However, this analysis has shown that one can be traded off against the other, especially with the reduced capacitance per unit area and increasingly ballistic operation likely with future technology.

The analysis also indicates that the overall impact of variability on the target operating frequency and voltage will not be as great as might be expected. For example, a worse-case +25% shift in V_{TH} (i.e., twice the ITRS prediction of ~12%) might move the target frequency scaling by less than 10–15%, a figure that might easily be reclaimed at the architectural level. Performance (operating frequency) will continue to improve simply from reductions in load capacitance at future technology nodes and it will be possible to exchange some or all of this improvement for constant power and/or energy.

In summary, the model describes limits on the ability to trade frequency for power (and/or energy), given a particular combination of technology and architecture and given the ability to choose a particular sequence of supply and threshold values. In the next chapter, the parameters χ, χ' and σ are derived for the reconfigurable fabric proposed in Chapter 3 and some predictions made regarding the ultimate scalability of this end-of-roadmap fabric.

Chapter 5. Power Scaling in the Reconfigurable Platform

"In Engineering, all other things being equal, simpler is always better, and sometimes much better."

Robert Colwell, in [335]

A key objective of this work has been to explore the scalability of locally connected computational structures at both device/circuit and architectural levels. In Chapter 3, a reconfigurable logic array based on thin-body Schottky-barrier transistors was proposed and analysed. While still challenging to produce, this does offer the promise of a simplified, regular and manufacturable fabric. The reconfigurable array is entirely locally connected, with the basic 6-NOR cell fulfilling the functions of logic, interconnect or arbitrary combinations of either. The analytic model developed in Chapter 4 provides a mechanism for predicting the evolution of power/energy in any digital design in terms of an architectural-level parameter, σ, and two circuit-level parameters, χ and χ'. This chapter brings these threads together by analyzing the characteristics of the reconfigurable platform in terms of that model. Given some assumptions relating to interconnect cost, it is determined here that appropriate values of these model parameters can be achieved such that architectures mapped to the platform may be continuously scalable in terms of power and performance.

This stage of the analysis has comprised two further levels of simulation, as follows:

1. ***Device/Circuit Level***: A single pair of 6-NOR cells was built using a modified version of a double-gate EKV transistor model sourced from the Laboratoire d'électronique at EPFL[7] [336]. The modifications made to this model, outlined below, are mainly related to the sensitivity of the threshold shift and the variation of the subthreshold slope, both functions of the control-gate voltage level on the double-gate device. These cells were

[7] Ecole Polytechnique Fédérale de Lausanne

configured into a simple representative circuit—in this case a simple 1-bit adder circuit—and its performance used to predict the range of the technology-related parameters χ, and χ'.

2. ***Architectural Level***: An architectural-level analysis was performed using a simple mixed digital/analog model. The propagation delay and subthreshold current of the NOR cells were modeled as simple functions of supply and threshold voltages, while a purely digital model described their logic behavior. This abstract architectural model was used to derive σ for the example circuits. The power-area-performance characteristics of the reconfigurable fabric could then be predicted over the supply range expected for the remaining nodes of the CMOS roadmap.

5.1 VHDL-AMS

Whereas the circuit-level analysis of the reconfigurable array described in Chapter 3 was based on University of Florida SOI (level 10) SPICE models, the analysis from here was transferred to a high-level design language (i.e., VHDL-AMS). This was necessary for two reasons. Firstly, the additional complexity of the circuits made the run times of SPICE untenable. Secondly, the functional analysis at this stage required more direct and independent control over specific aspects of the transistors than was available in the physically-based SOI SPICE models. VHDL-AMS offered a suitable solution to both issues.

Established as IEEE standard 1076.1 in 1999 [337], VHDL-AMS extends the VHDL language into the domain of analog and mixed-signal systems such that it supports the description of continuous-time behavior. VHDL-AMS adds the concept of a continuous *quantity* to the basic logic signal types in standard VHDL (e.g., *std_logic* in IEEE1164) and provides a generalized notation for describing Differential Algebraic Equations (DAEs). Multi-discipline networks can be described and simulated at two levels of abstraction: as either conservative-law networks (Kirchhoff's networks) or in terms of ideal signal-flow networks. In conservative-law networks, an *across* quantity represents an effort (e.g. a voltage in the case of electrical systems) while a

through quantity represents a flow (e.g. a current). This provides a straightforward mechanism to include analog effects such as the impact of drive current on performance (i.e., propagation delay) and logic state on average subthreshold current, within what otherwise would be a purely logic-level simulation.

5.2 Device/Circuit Level Modeling

As mentioned above, the objective here was to predict the expected range of the technology-related parameters χ, and χ'. Firstly, the modified EPFL-EKV model is described, followed by the results derived for a representative circuit mapped to the single pair of 6-NOR cells. These modifications were necessary because, although various groups (including the EPFL) are actively working on double-gate models, none had become available at the time this work was undertaken.

5.2.1 The EPFL Double-Gate Transistor Model

Just as for the UFSOI SPICE models used previously, the EPFL-EKV compact model [333] is based on physical properties of the MOS structure. It uses a charge sheet approach in which the drain current is derived as the sum of two components, diffusion and drift. The diffusion current is dominant in weak inversion, whereas the drain current in strong inversion is mainly due to drift current. These two components contribute continuously across the full device range, removing the need for transition corrections between the linear and saturation regions seen in piece-wise or "regional" models.

The complete model includes expressions for first-order derivatives such as transconductances and transcapacitances and is intended to be used for the design of low-voltage, low-current analog, and mixed analog-digital circuits using submicron CMOS technologies. However, this stage of the analysis focused only on the effect of peak normalized drive and subthreshold currents so these derivative terms were omitted and fixed capacitances used at each interface. The model equations assume that the silicon channel is undoped or lightly doped and that the mobility is constant along the channel. It also neglects both quantum and poly-depletion effects.

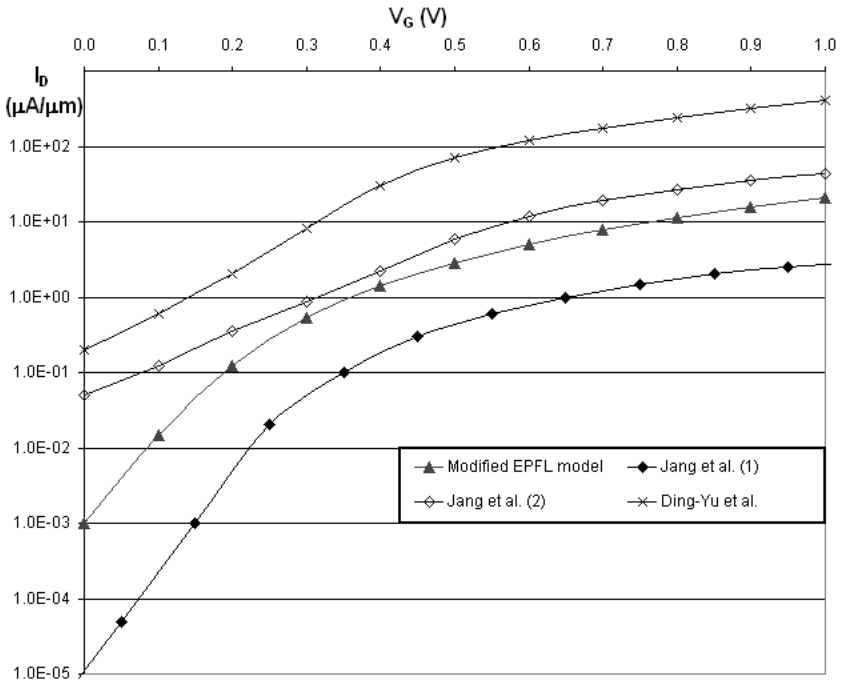


Figure 87. I_D vs. V_{BG} for the modified EPFL DGSOI model (V_{FG}=0). Also shown are data for ErSi-based devices derived from [338]: (Jang *et al.*(1)), [53]: (Jang *et al.*(2)) and [339]: (Ding-Yu *et al.*).

Figure 87 compares the I_D-V_{GS} curve derived using the modified EPFL model with some experimental *single-gate* devices reported in [53, 338] as well as the simulations of [339]. In [53], an annealing step in N_2 was used to greatly reduce the interface trap density allowing the subthreshold slope to approach its optimum 60 mV/decade value. On the other hand, the devices reported in [338] exhibit a saturation drive current more than an order of magnitude greater, but with a poorer off-current range. The device proposed in [339] suggests the use of both ErSi and $CoSi_2$ to set dual barrier heights at the source and drain such that the on-state and off-state currents can be optimised separately.

The curves for the devices shown in Figure 87 have been adjusted on the horizontal axis so that their nominal threshold voltages line up at $V_G \approx 0.2$V. Although done here just for ease of comparison, it is likely that this sort of threshold adjustment will be achievable in the future using metal gate technology with "tuneable" work functions [340]. The final curves for the system simulated below were set up with I_D(sat) towards the middle of the range of Figure 87, and with a

subthreshold slope of ~100 mV/decade. The absolute values of I_D(sat) and I_{OFF} here are some two orders of magnitude higher than suggested by the TCAD simulations of Chapter 3 (cf. Figure 36 with $V_{BG} = 0$). Moreover, the various devices drawn from the literature exhibit a range of I_D(sat) that varies by a factor of more than 10^2 and I_{OFF} as much as 10^4.

The model parameters presented in Chapter 4 are normalized such that it is the relative change in I_D(sat) and I_{SUB} with ΔV_{TH} and ΔV_{DD} with a given technology that is important. Thus, the key aspects of the device models here are those parts that impact on $\Delta V_{TH}/\Delta V_{BG}$ and $\Delta S/\Delta V_{BG}$. One difficulty with the original EPFL model is that it assumes symmetrical operation i.e., both gates driven together, and therefore does not model either of these characteristics. It also provides no explicit mechanism for adjusting S away from its initial value of ~60 mV/decade without the need to undertake a complicated physical calibration process for each new set point. Figure 88 is an annotated partial view of the EPFL model that has been modified to account for these effects. All of these alterations represent simple empirical corrections that are not intended to maintain the physical accuracy of the charge equations.

The EPFL model uses a simplified numerical solution to the relationship between charge densities and potentials given by [341]:

$$v_G^* - v_{CH} - v_{TH0} = 4q_G + \ln\left[q_G\left(1 + q_G\frac{C_{OX1}}{C_{SI}}\right)\right] \tag{5.1}$$

where v_G^* is the effective gate voltage (= v_G – $\Delta\psi_i$, with $\Delta\psi_I$ the gate-channel work function difference), v_{CH} is the electron quasi-Fermi potential, v_{TH0} the threshold voltage, q_G the charge density per unit surface on each gate and C_{OX1} and C_{SI} are the gate and silicon layer capacitance per unit area, respectively. The function *qi1n* in Figure 88 computes the normalized charge density given the potentials on each gate as well as on the source/drain. It is made up of terms for the solution of both components of (5.1) along with a transition potential that selects between them. Thus, in order to account for the threshold shift and changes to subthreshold slope, it is necessary to adjust these components along with the transition potential.

```
library IEEE;
use ieee.math_real.all;
library IEEE_PROPOSED;
use ieee_proposed.electrical_systems.all;
library MGC_AMS;
use MGC_AMS.conversion.all;
-- Double gate model based on Prégaldiny et al,
-- IJNM: Electronic Networks, Devices and Fields, vol. 19, pp. 239-256, 2006
-- this version does not contain transcapacitances
--
ENTITY dg_nmos IS
   generic (W :real:= 1.0e-6; -- Gate width [m]
                L :real:= 1.0e-6; -- Gate length [m]
                tox1 :real:= 2.0e-9; -- Top gate oxide thickness [m]
                tox2 :real:= 2.0e-9; -- Bottom gate oxide thickness [m]
                tsi :real:= 25.0e-9; -- Si film thickness [m]
                mu0 :real:= 0.01); -- Low-field mobility [m^2/Vs]
   port (quantity Vth0,Vdmax: in real;terminal d, g1, s, g2, b :electrical);
END ENTITY  dg_nmos;

ARCHITECTURE asymmetric OF  dg_nmos IS
--define constants : Physical;  Operating conditions; Derived parameters
--Define Quantities

--Definition of the qi1n function
pure function qi1n(vg,s,v :real) return real is
-- Local precomputed parameters

begin
   vt := s*(4.0*qt + bs);
   vp:= vg - vto;
   vtest := s*(vp - v);
   if (vtest>vt) then
   Calculate normalized charge when q>1
          return -q0*(1.0 + da*(1.0 + 0.13*da));
   else
   Calculate normalized charge when q<<1
   i.e., logarithmic term is dominant
          return -q0*(1.0 + da*(1.0 + 0.35*da));
   end if;
end qi1n;

BEGIN
   S_d==0.64/(vg2+1.23)**0.3;
   vgn == ((0.4-Vth0)+vg1+(Tox1/Tox2)*vg2)/UT;
   ids==ID0*(-4.0*(qi1n(vgn,S_d,vd/UT)**2-
                qi1n(vgn,S_d,vs/UT)**2)/(Vdmax**1.75) +
          4.0*(qi1n(vgn,S_d,vd/UT)-qi1n(vgn,S_d,vs/UT)) +
          2.0*(log(1.0-(alpha*qi1n(vgn,S_d,vd/UT)))-
                log(1.0-(alpha*qi1n(vgn,S_d,vs/UT))))/alpha);
END ARCHITECTURE asymmetric;
```

Entity Declaration

subthreshold slope term

back-gate threshold effect; set initial V_{TH} (V_{TH0}).

Unified Drain Current Equation

α correction term

Figure 88. Modified EPFL double-gate model.
The constant and quantity declarations have been omitted for clarity.

Firstly, the normalised gate voltage in Figure 88 was extended to include contributions from both gate voltages and a term (0.4-V_{TH0}) was included to offset the initial threshold voltage by 0.2V to align it more closely with the SPICE models used previously. The subthreshold slope correction equation $S_D = 0.64/(V_G + 1.23)^{0.3}$ was then derived empirically by measuring the change in sub-

threshold slope as the transition threshold was moved and curve-fitting the resulting data. It was designed to approximate the slope of $\Delta S/\Delta V_{BG}$ derived from the TCAD simulations i.e., $60 < S < 127$ mV/decade over a ±0.45V shift in back gate bias. Finally, experiments on the original model indicated that it follows the general trend of the alpha-law model $I_D \propto (V_D - V_{TH})^{\alpha}$ but with $\alpha \approx 1.4$. The quantity $V_{DMax}^{1.75}$ provides a correction to I_D(sat) to reduce α to around 1.25 as suggested by Chen (and verified in Section 2.5.1). V_{DMax} is the value of V_{DD} at the beginning of each simulation run, before it is scaled. The final results derived from the modified EPFL model are show in Figure 89. It can be seen that the general shape of the threshold shift and subthreshold slope mimics the shape of the previous TCAD simulation results (cf. Figure 36), albeit at higher absolute values of I_D(sat).

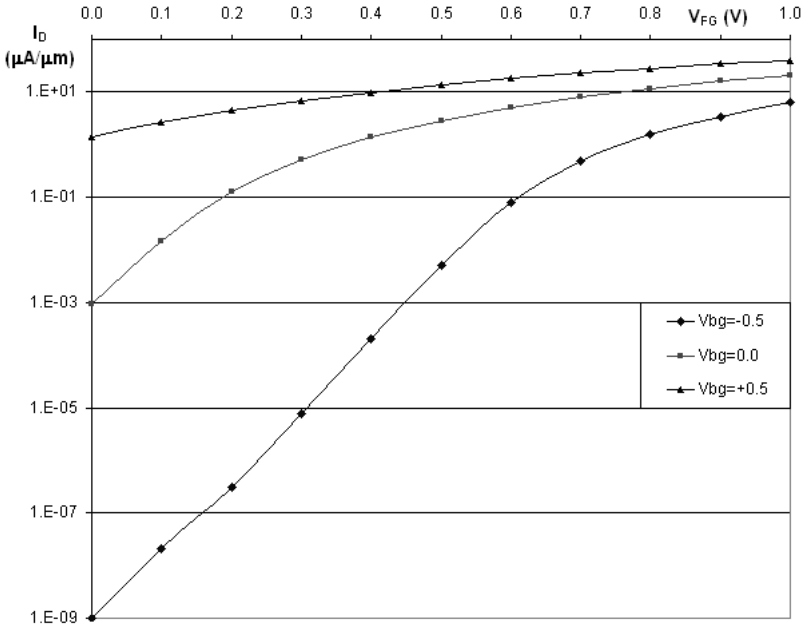


Figure 89. I_D(sat) vs. V_{GS} for the modified EPFL model.

The discussion to this point has focussed only on the nMOS device. In these VHDL-AMS simulations, the pMOS device is set up as a simple "mirror image" of the nMOS case. The only change is the polarity of the input signals and the signal reference terminal (Figure 90). It can be seen that in the case of the nMOS model, both the *across* and *through* quantities are defined

relative to a reference terminal *b*, which is usually tied to electrical_ref (ground). The corresponding quantities in the pMOS device are defined in the reverse sense with respect to the terminal *b*, which in most cases is set to V_{DD}. In all other respects (except for some of the user definable parameters such as the low-field mobility, μ_0), the models are identical.

```
ENTITY dg_pmos IS
   generic(W :real:= 1.0e-6; -- Gate width [m]
           L :real:= 1.0e-6; -- Gate length [m]
           tox1 :real:= 2.0e-9; -- Top gate oxide thickness [m]
           tox2 :real:= 2.0e-9; -- Bottom gate oxide thickness [m]
           tsi :real:= 25.0e-9; -- Si film thickness [m]
           mu0 :real:= 0.1); -- Low-field mobility [m^2/Vs]
   port (quantity Vth0, Vdmax: in real;
         terminal d, g1, s, g2, b :electrical);
--b is typically Vdd for pMOS, electrical_ref for nMOS
END ENTITY dg_pmos;

--Quantities definitions -
```

nMOS	pMOS
quantity vg1 across g1 to b;	quantity vg1 across b to g1;
quantity vg2 across g2 to b;	quantity vg2 across b to g2;
quantity vd across d to b;	quantity vd across b to d;
quantity vs across s to b;	quantity vs across b to s;
quantity ids through d to s;	quantity ids through s to d;

Figure 90. Interface quantities for nMOS and pMOS models.

5.2.2 Device/Circuit Level Parameter Extraction

The two model parameters χ and χ' can be readily derived from the mean critical path delay, (assuming average drive current $I_D(sat) \propto V/\tau$) and subthreshold current for a representative circuit mapped to the 6-NOR array. By measuring τ and I_{SUB} over a range of V_{DD} and V_{TH} values (with $V_{TH} = a\text{-}bV_{DD}$ as before), values for β and η were derived by fitting the I-V data using a least-squares error technique, from which $\chi = \frac{\beta+1}{\beta-1}$ and $\chi' = \frac{\eta+1}{\beta-1}$ (see Section 4.4.3). Figure 91 shows an example of the curves fitted to β and η for V_{DD} and V_{TH} scaling factors of V = 0.85 and $b = 0.05$, respectively. Here the vertical scale is current, normalized to its initial value at 1.2V. The maximum fitting errors are approximately ±5% and ±8% towards the centre of the β and η curves, respectively.

V_{DD}	V_{TH} (0.26-0.05V_{DD})	I_D (sat)	$0.74V_{DD}^{1.72}$	I_{OFF}	$0.87V_{DD}^{0.85}$
1.20	0.200	1.00	1.03	1.00	1.01
1.02	0.209	0.80	0.74	0.78	0.88
0.87	0.217	0.62	0.53	0.68	0.77
0.74	0.223	0.46	0.38	0.61	0.67
0.63	0.229	0.33	0.28	0.58	0.58

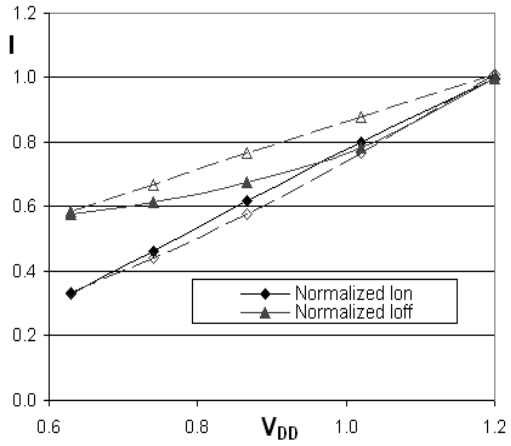


Figure 91. Normalized $I_D(sat) \propto kV_{DD}^{\beta}$ and $I_{OFF} \propto kV_{DD}^{\eta}$ with V = 0.85, b = 0.05.

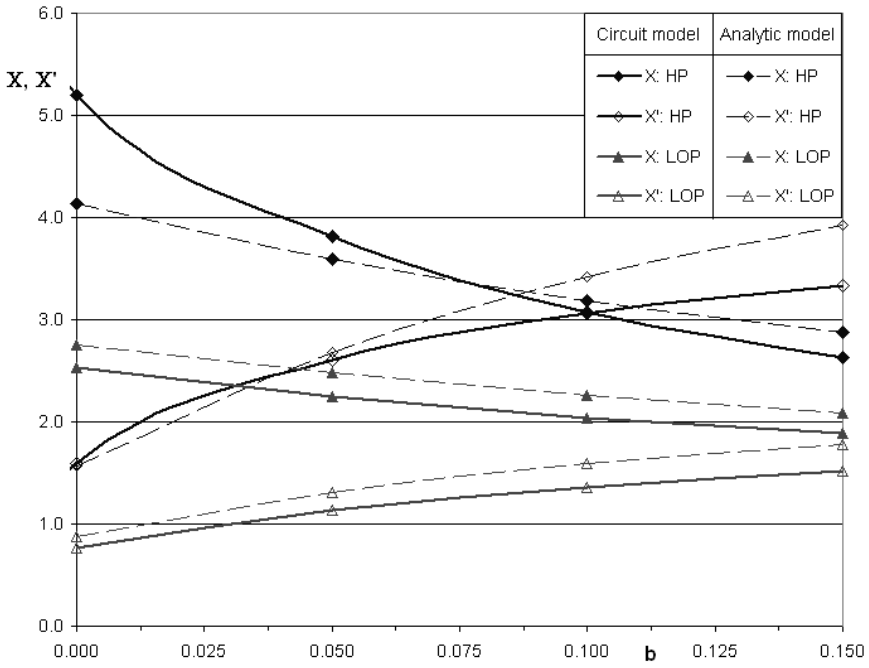
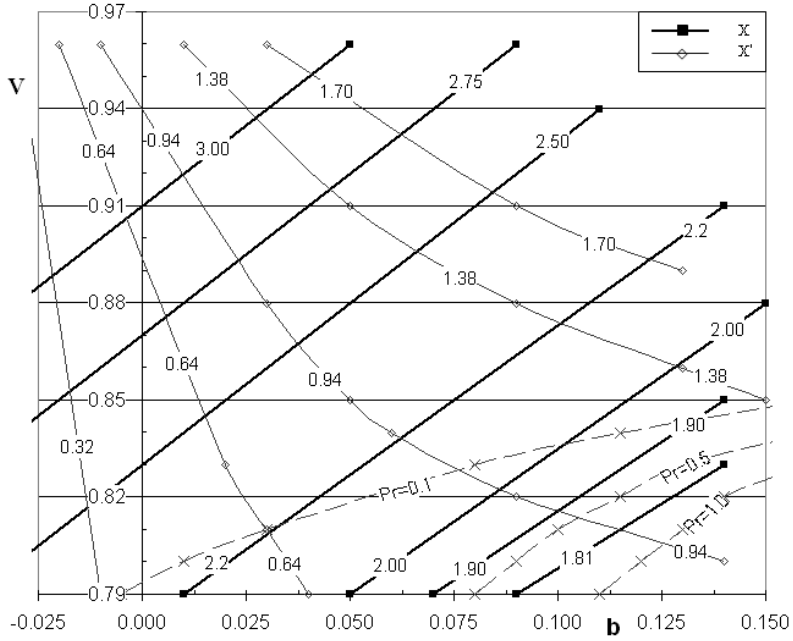


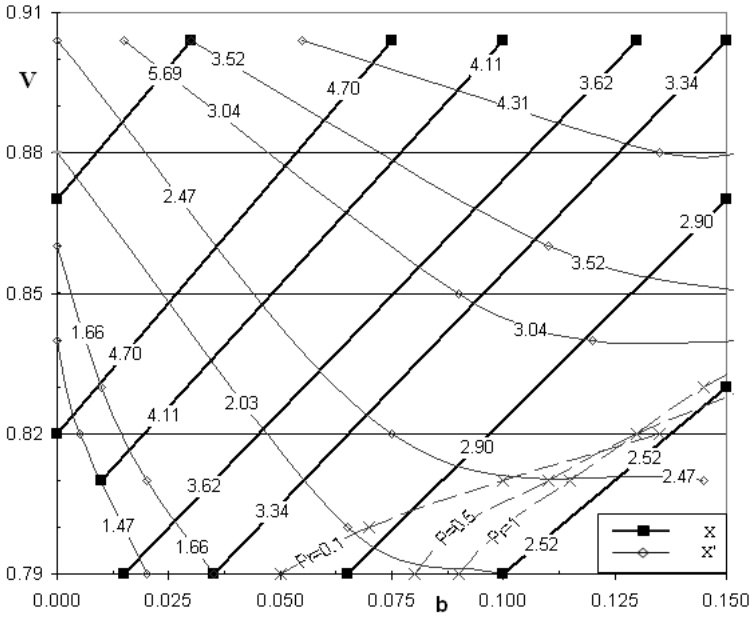
Figure 92. χ and χ' vs. *b* for V=0.85.
Solid lines: circuit model ; dashed lines: analytic model, technology assumptions as shown.

Figure 92 shows the χ and χ' curves at the same supply scaling (0.85) compared to those previously derived directly from the basic analytic models (cf. Figure 82). It can be seen that the general form is similar, although in the HP case, the slope of the curves implies a higher roll-off with threshold voltage than predicted for both $I_D(sat)$ and I_{OFF}. In the LOP case, the curves are simply offset by approximately 10%. These differences reflect the impact of the stacked transis-

tor topology of the NOR circuit compared to the original values that were determined from the characteristics of a single transistor.



(a) LOP Technology



(b) HP Technology

Figure 93. Contour plots of χ & χ' vs. supply (V) and threshold scaling (b), no variability (χ: filled squares; χ': open diamonds)

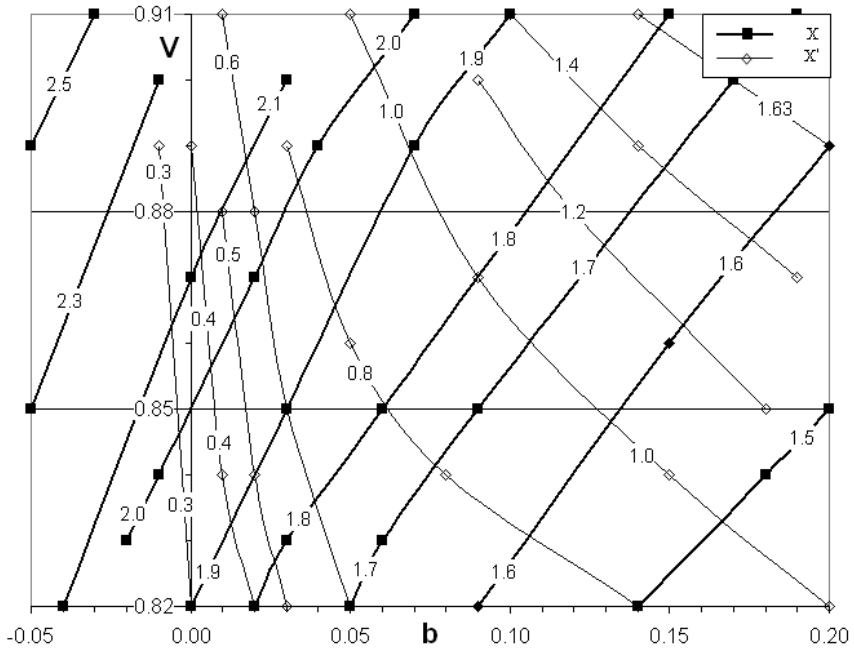


Figure 94. χ & χ' vs. supply and threshold scaling, variability = +25% σ_{VTH} (χ: filled squares; χ': open diamonds).

Just as with the analytic model in Chapter 4, it can be seen from Figure 93 and Figure 94 that the overall ranges available for both χ and χ' depend largely on the initial V_{DD} and V_{TH}. For example, using the initial values for ITRS LOP technology ($V_{DD0} = 0.9$, $V_{TH0} = 0.26$), Figure 93 exhibits $1.8 < \chi < 3$ and $0.65 < \chi' < 1.7$ over the experimental range of supply and threshold voltage scaling. Similarly, in HP technology ($V_{DD0} = 1.2$; $V_{TH0} = 0.20$) this expands to $2.5 < \chi < 5.6$ and $1.5 < \chi' < 4.3$. The dashed lines to the lower right of Figure 93a and Figure 93b represent the loci of $P_T = 1$ in (4.45) for $P_R = 0.1$, 0.5 and 1.0. In the same manner as Figure 80 (page 152), these converge to a single point where $\chi = \chi'$ (which implies $\beta = \eta$), where $P_T = 1$ for all P_R.

The impact of increasing variability is to reduce the range of χ and χ'. For the LOP case shown in Figure 94, the range reduces by 20–30% to around $1.5 < \chi < 2.1$ and $0.5 < \chi' < 1.6$, assuming that the worse-case V_{TH} variability is $+3\sigma_{VTH} = 25\%$ i.e., $V_{TH} = 1.25(a - bV_{DD})$. In general terms, it will then be more difficult to find an architectural solution that will allow a tradeoff between area and power. This would be entirely expected given the higher absolute threshold values under

the maximum variability assumptions. This tradeoff is examined in greater detail in the following section.

5.3 An Architectural Scaling Model

The objective of this final stage was to set up some representative computational structures in order to validate the predictions of the abstract architectural model developed in Chapter 4. A mixed signal approach has been adopted in which the basic 6-NOR cells of the reconfigurable platform were simulated at logic level, while the propagation delay and subthreshold current were modeled as continuous functions of V_{DD} and V_{TH}.

5.3.1 VHDL Behavioral Model

It was determined in Section 2.5.1, that the alpha-law model will be sufficiently predictive of saturation drive current with alpha in the range $1.05 < \alpha < 1.3$ so that the relative delay (τ_P) over the range of supply and threshold voltage can be modeled as:

$$\tau_p \propto \frac{CV}{I} = \frac{KCV}{(V - V_{TH})^{\alpha}}. \tag{5.2}$$

Here, device-level considerations such as transistor gain (i.e. $(W/L)_p$:$(W/L)_n$), mobility differences etc. are either accounted for in the parameter K, or are constant for a particular circuit configuration and therefore will be normalized out. Equation (5.2) is used to compute the propagation delay for both interconnect and logic. In Section 3.3.3, it was found that the delay of a type 2 interconnect is typically around 15–20% of the logic delay, but this ratio was made adjustable in order to measure the impact of varying interconnect costs.

The VHDL-AMS fragment in Figure 95 illustrates the basic mechanism. In line (1) the real quantity *delay* is evaluated as in (5.2). The ADMS® environment provides a set of explicit conversion utilities, including between the types *real* and *time*. In line 2, the function *real2time* is used to delay the assignment of the logic signal L(0) by the (continuous) value of *delay*. Each sum-of-products (SOP) function requires two adjacent cells to evaluate and the signal L repre-

sents an output from the first cell. An output signal from the second cell, Ao(1), is similarly delayed in line (3). It is assumed that both cells have the same delay function, and that signals on all inputs incur the same delay cost. Strictly, both of these assumptions are invalid: the first because of within-die variability and the second due to the impact of the circuit layout (e.g. stack effect, variations in switching threshold between inputs, variations in node capacitances, etc.). However, the model here is based on normalized aggregate characteristics so it is assumed that these effects will average out at the system level.

```
(1)     delay <= K1*Csqr*V/((V-Vth)**alpha);
(2)     L(0) <= not(inp(3) or inp(5) or inp(1))
                            after real2time(delay);
                            …
                            …
                            …
(3)     Ao(1) <= L(1) or L(2) or L(3) or L(4)
                            after real2time(delay);
```

Figure 95. Delay calculation and application in VHDL-AMS

In addition, while it is theoretically possible for the subthreshold slope, S, of double-gate devices to approach its optimum value of ~65 mV/decade, the proposed reconfigurable array uses the devices in their *ground-plane* mode for which S is expected to be significantly higher (> 100 mV/decade) if the gate oxide thicknesses (T_{OX1} and T_{OX2}) are approximately equal. To simplify the subthreshold analysis, S was fixed at 100 mV/decade in all of the following, consistent with the values observed over a range of experimental devices reported to date and a constant (room) temperature was assumed throughout. As a result, the subthreshold current was modeled as a simple function of V_{TH}:

$$I_{OFF} \propto K_{p,n} W_{p,n} e^{-V_{TH}/nV_t} \tag{5.3}$$

where $W_{p,n}$ is the effective width for the p or n block in a particular circuit topology (either NOT or NOR), and nV_t is set to achieve the fixed subthreshold slope of 100 mV/decade at 300 K. The parameter $K_{p,n}$ accounts for variations (in mobility etc.) between process gains of the p and n blocks. The magnitude of I_{OFF} is therefore data dependant. This could easily be extended to model $\Delta S/\Delta V_{TH}$ by making nV_t a function of a notional back gate bias, but this was not necessary

at this stage as it would simply add a constant "background" leakage term that is proportional to the number of inactive cells but some 10^6 times smaller than the derived I_{SUB} (i.e., in the range of 10^{-11}A). The model assumes that only the active (inverting) interfaces contribute to I_{SUB} and that, in SOI technology, the pass-gates have no leakage paths to ground. A VHDL-AMS description of a single-bit full adder used in this work can be found in Appendix C. All of the functions based on the 6-NOR array exhibit an identical form.

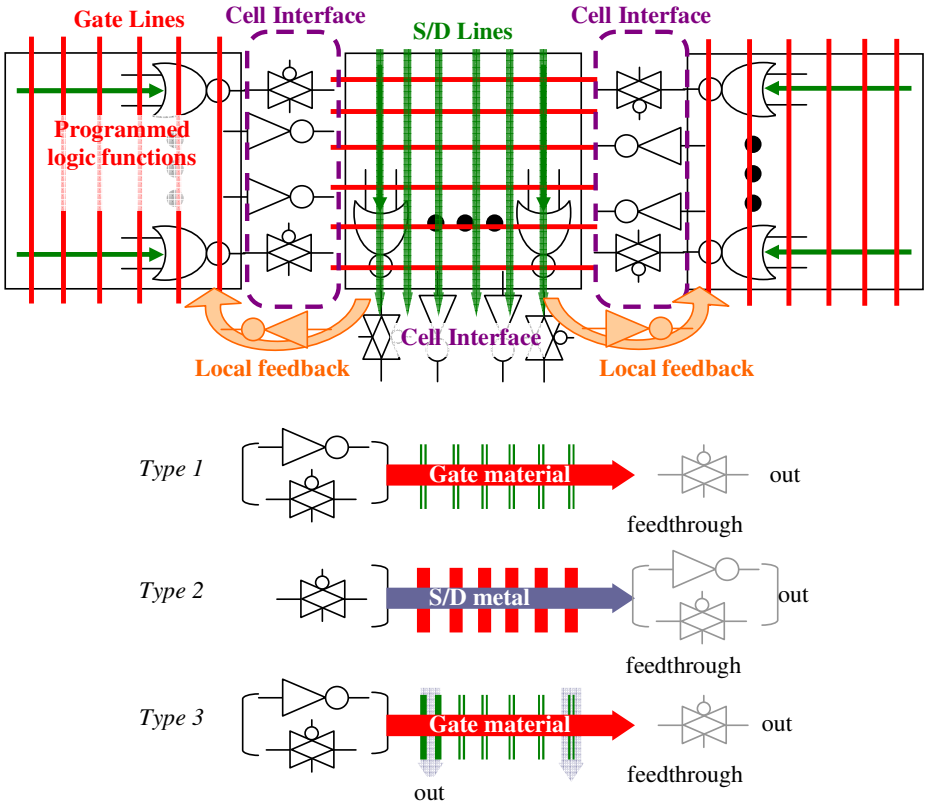


Figure 96. Abstract cell organization and interconnect types

Figure 96 shows the generic cell and interconnect organizations for the reconfigurable platform. In addition to the 6x6 programmable cell and interface, models for the three interconnect types were created. Thus, once a range of components (1-bit adders, D-type FF, multiplexers etc.) were defined, the final process of circuit generation became essentially a "floorplanning" exercise in which these components and their interconnections were instantiated on adjacent cells.

5.3.2 Parallel Architectures and σ

The question now is whether, assuming the various technology constraints on χ and χ', will parallel architectures mapped to the reconfigurable array result in a range of σ that will allow area to be traded for power. To explore this question, the simple data path of [209] (Figure 97) was used to analyze the behavior of the reconfigurable array. Although intrinsically simple, this data path was considered to be representative of a wide range of parallel architectural solutions, including those multiprocessor organizations outlined previously in Section 2.8, for which area is explicitly traded for performance.

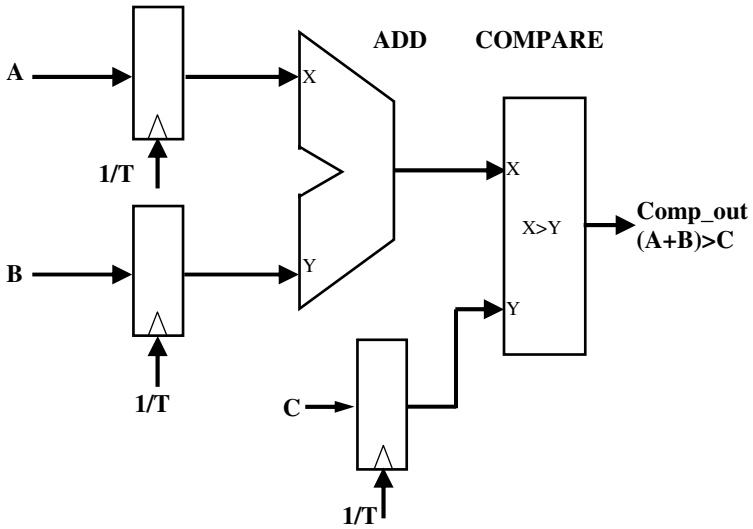


Figure 97. A simple data path (from [209]).

The basic architecture was replicated as many times as required and the operating performance derived for the same critical path in each configuration. The resulting area-time relationship gives σ for this parallel architecture. The expected power scaling is then given by (4.29) and (4.33). As the fabric is entirely mesh-connected, γ depends only on technology so that $\gamma = 0$ at a single technology node and $\gamma < 0$ where capacitance reduces at successive nodes due to changes in line dimensions and dielectric constant of the field oxide.

The original complexity measures on which the model is based were derived for various VLSI circuits under the assumption that propagation time across a wire could be made independent of

the wire length simply by adjusting the size of the drivers. The additional area incurred was assumed to be small compared to the wire area and by "fudging" [342] the feature size upwards, the area of the driver could be absorbed into the area of its wire. This is not the case for the proposed array, in which "units" of interconnect and logic are formed from the same cells and therefore may exhibit comparable area and delay dependencies. This is a more severe constraint than typical distributed computational platforms (e.g. FPGAs). The very high interconnect overheads exhibited here may be considered to be representative of ultimately scaled systems that are expected to exhibit similarly high interconnection costs.

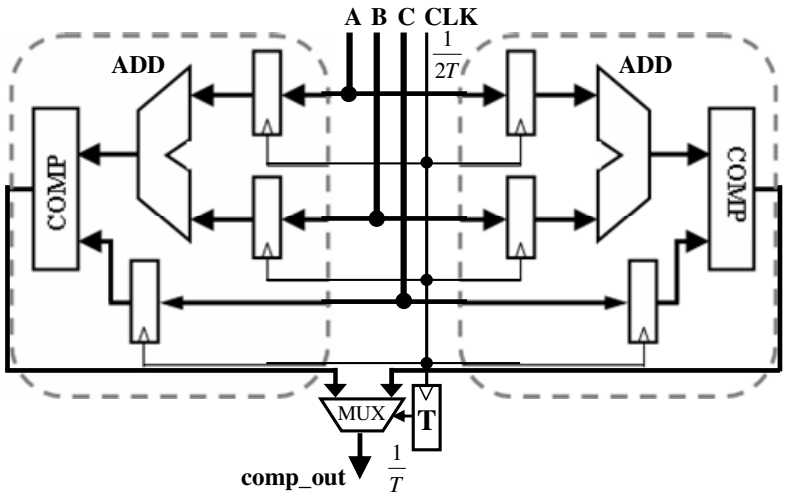


Figure 98. A duplicated version of the simple data path (modified from [209]).

It might reasonably be expected that the assumption of local connectivity in the array will impose severe delay overheads on an architecture (resulting in larger values of σ), thereby making the frequency--power tradeoff more difficult. However, it was found that this is not typically the case because the so-called "polymorphic" nature of the reconfigurable platform, in which logic and interconnect are (mostly) interchangeable can result in compact layouts which tends to offset the impact of the interconnect delay.

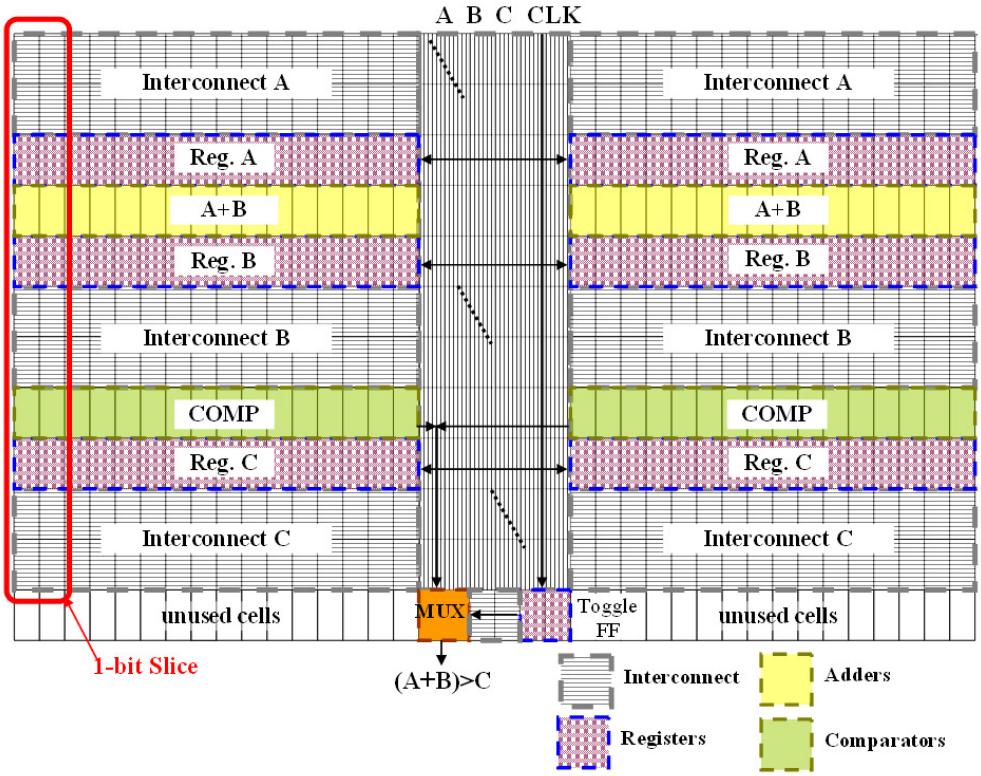


Figure 99. Simplified floorplan for parallel data path of Figure 98 (not to scale).

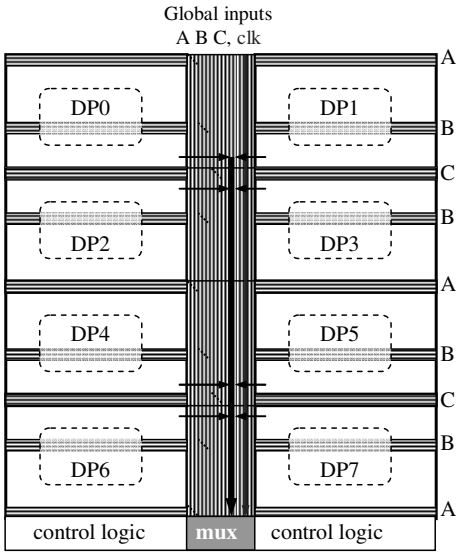


Figure 100. 8-way replicated data path layout.

For example, the duplicated 8-bit data path in Figure 98 can be synthesized with the floorplan shown in Figure 99. Here, the two data path blocks are formed from replicated 1-bit

add/compare/register slices mirrored across a common central interconnection bus thereby incurring little additional interconnection area or delay penalty apart from a negligible increase in the capacitive load on the "corner" routing cells in the central routing channel. This takes advantage of a characteristic of the particular organization chosen for the array i.e., that the dataflow direction can be reversed simply by shifting the slice mapping by one cell position in any direction. Continuing this idea, adjacent data path blocks may be connected by abutment, sharing interconnect channels where appropriate (Figure 100), thereby further limiting the growth of the interconnect.

Table 14 Normalized scaling characteristics of the simple parallel data path.

	# Parallel Data Paths (N)									
	2		4		8		16			
Architecture	**A**	**T**	**A**	**T**	**A**	**T**	**A**	**T**	***f*(A,F)**	**σ**
Chandrakasan ♣	2.06	0.58	4.17	0.33	8.67	0.19	17.66	0.11	$F \approx A^{-0.77} \pm 0.8\%$	1.3
FPGA baseline	2.11	0.50	4.23	0.28	8.45	0.14	16.93	0.07	$F \approx A^{-0.94} \pm 4\%$	1.06
PMA: $\Delta\tau_{INT}=0.2$	1.88	0.53	3.29	0.285	6.28	0.16	11.5	0.095	$F \approx A^{-1.0} \pm 2\%$	1.0
PMA: $\Delta\tau_{INT}=0.5$	1.88	0.53	3.29	0.33	6.28	0.21	11.5	0.125	$F \approx 0.95A^{-0.85} \pm 5\%$	1.18
PMA: $\Delta\tau_{INT}=1.0$	1.88	0.53	3.29	0.36	6.28	0.27	11.5	0.185	$F \approx 0.9A^{-0.7} \pm 10\%$	1.4

♣ circuit from [209]; estimates based on typical standard cell circuits.

Table 14 shows the results for various parallel implementations of this data path. In each case, both the area and critical path delay, measured between Register A clock rising edge ($\uparrow Reg_A$) and comp_out in Figure 97, have been normalized to their value with single path instantiated (N = 1). Normalization removes any technology impact here, so that the critical path depends only on the unit-delays over the various paths in the architecture in dimensionless units. The aggregate delay is given by $T = \frac{T_{cp}}{N} + T_{MUX}$, where T_{cp} is the critical path delay from $\uparrow Reg_A$ to the final multiplexer and T_{MUX} is the delay of the multiplexer where N > 1 (cf. Figure 24, page 60). The successive values of T also describe the frequency scaling that would result in constant overall performance, and thus σ for that circuit, as shown in the right-most two columns of the table.

The area and delay estimates for the add/compare circuit of [209] (labelled *Chandrakasan*) have been extrapolated well past those originally reported and are based on typical characteristics for static CMOS standard cell circuits. The area scaling result is quite different to that reported in [209] as it is based on the increase in switching device numbers and does not include the large area overhead incurred by the routing channels in a standard cell system. As a result, area (A) grows just slightly super-linearly, $O(N^{1.04})$. It was also assumed that the combination of additional routing and the output multiplexer adds $\log_2(N)$ unit gate delays to the critical path (N = number of parallel paths) so that the delay $T \propto N^{-0.8}$. A value of σ can be determined under these assumptions by equating the area and delay terms such that $N = A^{0.96} = T^{-1.25}$ and thus $F = A^{-0.77}$. The resulting value of $\sigma \approx 1.3$, places it towards the middle of the range for conventional architectures ($1 < \sigma < 2$).

To derive the results for the architecture labelled *FPGA baseline*, the replicated 8-bit add/compare circuit was implemented in VHDL using Xilinx® ISE 8.1i and mapped to a Spartan xc2s200e-6pq device. The area figures shown are based on the total equivalent gate count reported by the ISE as that figure is more directly comparable between these examples. The delay is based on the worse-case critical path from $\uparrow Reg_A$ to comp_out, as before, reported by the (unconstrained) post-place and route static timing report. For this FPGA implementation, area grows approximately $O(N)$ while the growth in delay is close to $O(N^{-0.92})$ giving $\sigma \approx 1.1$ over the range of parallel paths in this experiment. These two circuits—the original add/compare and the FPGA implementation—form a baseline against which the reconfigurable array will be compared in the following section.

5.3.3 Scalability Estimates for the Reconfigurable Platform

Having now simulated the reconfigurable array of Chapter 3 in terms of both its device/circuit behavior (Section 5.2.2) and its architectural characteristics (Section 5.3.2, above), it is now possible to derive the overall relationship between area and power/energy.

Architectural Parameter (σ)

As can be seen in Table 14, despite being constrained by its local interconnection topology, the proposed reconfigurable array exhibits $1 \leq \sigma \leq 1.4$ for the replicated architecture considered here. As the layout is based on a hierarchy of blocks clustered around common interconnection busses, total area growth is slightly sub-linear i.e., $A \approx N^{0.87}$. The three examples labelled PMA in Table 14 illustrate the impact of increasing interconnect cost, represented here by the ratio of the interconnect delay to the logic delay. It was determined above that the delay of the dominant interconnect (type 2, see Figure 96) is likely to be in the range 15–20% of the logic delay, due to the low RC time constant of its interconnect metal. This is the first PMA line in Table 14 ($\Delta\tau_{INT} = 0.2$), for which $T \approx N^{-0.88}$, resulting in $\sigma \approx 1.0$. At the other extreme, a maximum value of $\sigma = 1.4$ has been derived under the assumption that interconnect delay is equal to logic delay (i.e., $\Delta\tau_{INT} = 1.0$). These curves are illustrated in Figure 101, along with the baseline and can be compared to Figure 64 (page 119).

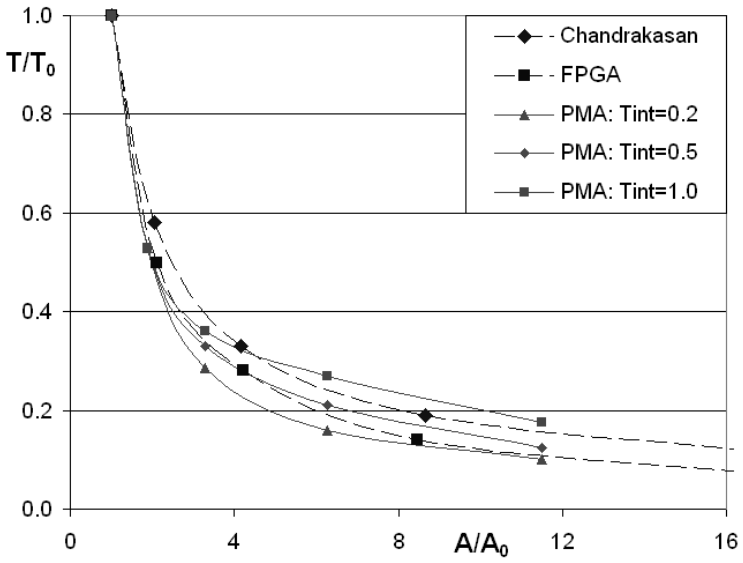


Figure 101. Area-Time relationship for the examples of Table 14.

Technology Parameters (χ and χ')

In Figure 102, the contour plots derived using the modified EPFL-EKV double-gate models assuming maximum V_{TH} variability (σ_{VTH} = +25%, cf. Figure 94) are overlayed with various solution loci from (4.45). Here, the dashed lines represent the values of σ that result in $P_T = 1.0$ for A = 1.8, 2.0 and 2.1, all at $P_R = 0.1$. These curves can be compared to those in Figure 80, which were derived using the device characteristics of Figure 79. In general terms, supply and threshold scaling values below these curves will result in reducing total power with area, although the individual dynamic or subthreshold terms may still be above 1.0. The (brown) dotted curves trace $P_T = 1.0$ for $P_R = 0.1$, 0.5 and 1.0 at A = 1.9, which is a slight overestimation of the area scaling observed for the three PMA circuits in Table 14.

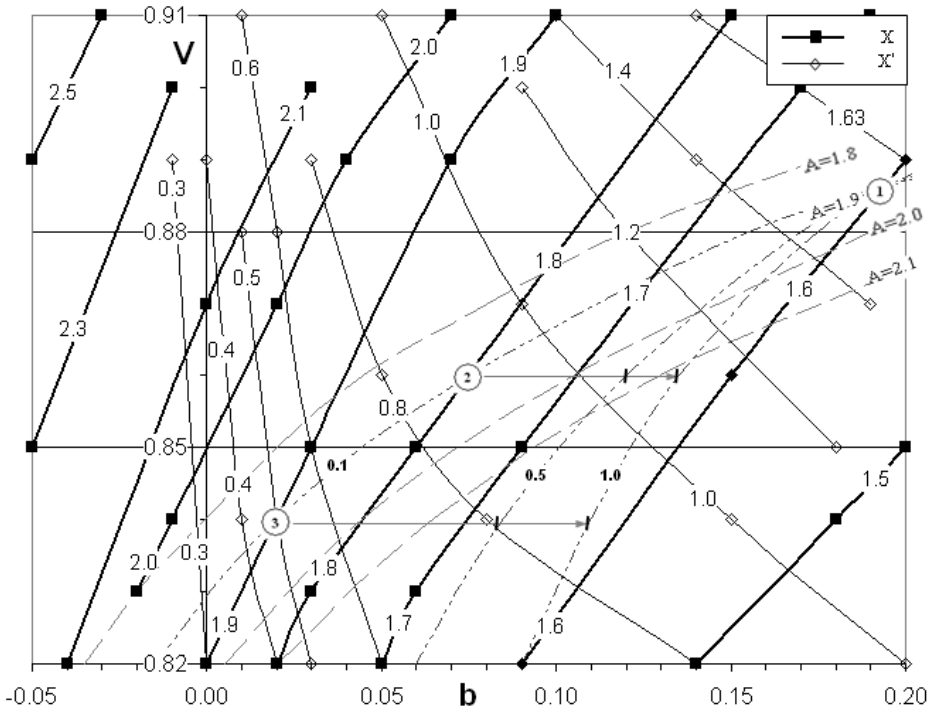


Figure 102. Contour plots for χ (filled squares) and χ' (open diamonds) as in Figure 94 (σ_{VTH}=+25%), overlayed with loci of P_T=1 at various A and P_R.

The normalized voltage and power predictions in Table 15 result from the application of the model at three points on the curve for A = 1.9, P_R = 0.1, as identified by the circled numbers 1–3 on Figure 102. In each case, χ and χ' are derived from the plot while the σ value is the maximum

value (σ_{MAX}) that results in $P_T \approx 1$ in (4.42). The table uses the frequency scaling given by (4.4) plus the supply and threshold voltage values from Figure 102 to predict the dynamic and subthreshold power scaling at that point. It can noted that in each row of the table, the target operating frequency is always less than is actually achievable by the given technology i.e., $A^{-1/\sigma} \le (V_{DD} - V_{TH})^{1.25}$, even where the worse-case V_{TH} becomes a significant fraction of the supply.

Table 15 Predicted voltage and power scaling at numbered points on Figure 102.
A = 1.9, P_R = 0.1, γ = 0.

1									
χ=1.60	σ=1.60	**Area**	V_{DD}	V_{TH}	P_{DYN}	P_{SUB}	P_{TOT}		
χ'=1.60	$F \propto \frac{(V_{DD}-V_{TH})^{1.25}}{V}$			**a=0.440**			P_R	P_R	P_R
$F \propto A^{-1/\sigma}$		**A=1.9**	**V=0.886**	**b=0.20**	$A^{(\sigma-\chi)/\sigma}$	$A^{(\sigma-\chi')/\sigma}$	**0.1**	**0.5**	**1.0**
1.00	1.0	1.00	0.90	0.325	1.00	1.00	1.00	1.00	1.00
0.67	0.8	1.90	0.80	0.350	1.00	1.00	1.00	1.00	1.00
0.45	0.7	3.6	0.71	0.373	1.00	1.00	1.00	1.00	1.00
0.30	0.5	6.9	0.63	0.393	1.00	1.00	1.00	1.00	1.00
0.20	0.3	13.0	0.55	0.411	1.00	1.00	1.00	1.00	1.00
2									
χ=1.79	σ=1.66			**A=0.328**			P_R	P_R	P_R
χ'=0.89			**V=0.86**	**b=0.075**	$A^{(\sigma-\chi)/\sigma}$	$A^{(\sigma-\chi')/\sigma}$	**0.1**	**0.5**	**1.0**
1.00	1.00	1.00	0.90	0.325	1.00	1.00	1.00	1.000	1.00
0.68	0.83	1.90	0.77	0.337	0.95	1.35	1.00	1.11	1.18
0.46	0.65	3.61	0.67	0.347	0.90	1.81	1.00	1.24	1.40
0.31	0.47	6.86	0.57	0.356	0.86	2.44	1.00	1.38	1.65
0.21	0.28	13.03	0.49	0.363	0.82	3.29	1.00	1.54	1.96
3									
χ=1.89	σ=1.63			**A=0.278**			P_R	P_R	P_R
χ'=0.5			**V=0.84**	**b=0.02**	$A^{(\sigma-\chi)/\sigma}$	$A^{(\sigma-\chi')/\sigma}$	**0.1**	**0.5**	**1.0**
1.00	1.00	1.00	0.90	0.325	1.000	1.00	1.00	1.00	1.00
0.67	0.81	1.90	0.76	0.327	0.903	1.56	1.00	1.21	1.32
0.43	0.61	3.61	0.64	0.328	0.815	2.44	1.00	1.46	1.74
0.28	0.41	6.86	0.53	0.329	0.736	3.80	1.00	1.76	2.29
0.19	0.21	13.03	0.45	0.330	0.664	5.93	1.00	2.12	3.03

Each of these three scaling assumptions represents a different tradeoff between dynamic and subthreshold power. At point 1, the terms χ and χ' are very close, so that the scaling of both the dynamic and subthreshold power will be similar. This is illustrated in the first group of figures in

Table 15, where both P_{DYN} and $P_{SUB} \approx 1$. This is also the point at which the total power P_T is largely independent of the relative power. Moving towards points 2 and 3 represents a trade-off between greater supply scaling (smaller V, leading to reduced dynamic power) and less aggressive threshold scaling. Between points 1 and 2, σ_{MAX} varies only a small amount, up to a maximum of 1.66.

In all cases, the trend for the absolute values of threshold voltage is quite different to that predicted by the ITRS (cf. Figure 77, page 145). The ultimate threshold voltage values illustrated in Table 15 (e.g., $V_{DD} = 0.54$ and $V_{TH} = 0.411$) can be contrasted with the ITRS predictions that assume that the target 14–17% performance increase at each generation mandates a constant or slightly reducing threshold voltage and that subthreshold power loss will be controlled by improving the subthreshold slope. This will be difficult to achieve with double-gate SOI technology, especially when the back and front gate oxide surfaces approach approximately equal thicknesses, in which case the sub-threshold slope is predicted to asymptote to a fixed final value in the range 100–115 mV/decade. The only remaining option to control subthreshold leakage will be to *increase* the threshold voltage as supply falls, with the resultant fall in performance being made up at other levels in the design hierarchy. This assumption has guided the model developed in this work. However, it makes little sense from a system point of view to simply exchange an increase in area for a constant level of power and performance. Ideally, it should be possible to choose an operating point that allows for both a reduction in power and an increase in performance. This issue is explored briefly in the following section.

5.3.4 Power–Performance Tradeoffs in Future Technology

Having now determined a range of σ achievable for the parallel data path implementation, this section compares how the reconfigurable platform might support scaling given some reasonable assumptions about the evolution of low-operating power (LOP) technology. An objective of the model is to allow quick evaluation of the tradeoffs between area, performance and power in the reconfigurable array. A number of examples of these tradeoffs are developed based around the technology parameters of Figure 102, which assumes a worse-case threshold variability of 25%.

Table 16 Baseline LOP scaling scenario.

A = 1.9, $\gamma = 0$ and -0.3, σ = 1.0 and 1.4.

γ=0													
N	**Area**	**V_{DD} (V)**	**V_{TH} (V)**	**C_{INT} $\gamma = 0$**	**Freq σ = 9.5**	**F_{MAX}**	**P_D=ACFV^2**	**P_{SUB}**	**Total Power**			**Perf.**	
									$(P_R)_0$=0.1	**$(P_R)_0$=0.5**	**$(P_R)_0$=1.0**	**σ=1.0**	**σ=1.4**
1	1.0	0.90	0.290	1.00	1.00	1.00	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.9	0.80	0.266	1.00	0.93	0.95	1.4	2.9	1.5	1.9	2.2	1.8	1.5
4	3.6	0.70	0.242	1.00	0.87	0.90	1.9	8.4	2.5	4.1	5.1	3.2	2.2
8	6.9	0.60	0.219	1.00	0.82	0.83	2.5	23.5	4.4	9.5	13.0	5.6	3.2
16	13.0	0.50	0.195	1.00	0.76	0.76	3.1	64.0	8.6	23.4	33.6	9.9	4.8
γ = -0.3													
N	**Area**	**V_{DD} (V)**	**V_{TH} (V)**	**C_{INT} γ = -0.3**	**Freq σ = -5.1**	**F_{MAX}**	**P_D=ACFV^2**	**P_{SUB}**	**Total Power**			**Perf.**	
									$(P_R)_0$=0.1	**$(P_R)_0$=0.5**	**$(P_R)_0$=1.0**	**σ=1.0**	**σ=1.4**
1	1.0	0.90	0.290	1.00	1.00	1.00	1.0	1.0	1.0	1.0	1.0	1.0	1.0
2	1.9	0.80	0.266	0.82	1.13	1.15	1.4	2.9	1.5	1.9	2.2	2.2	1.8
4	3.6	0.70	0.242	0.68	1.29	1.32	1.9	8.4	2.5	4.1	5.1	4.6	3.2
8	6.9	0.60	0.219	0.56	1.46	1.49	2.5	23.5	4.4	9.5	13.0	10.0	5.8
16	13.0	0.50	0.195	0.46	1.65	1.63	3.1	64.0	8.6	23.4	33.6	21.6	10.4

This analysis starts with a baseline scaling scenario (Table 16) derived from the supply and threshold voltage trends predicted for LOP technology in the 2005 ITRS [11] (cf. Figure 77). The magnitude of the threshold voltages have been increased by about 12% to account for the predicted level of threshold variability. Here it is assumed that the circuit will be operated at the maximum achievable frequency for the given technology assumptions so that σ is set such that $F_{MAX} \propto A^{-1/\sigma} = (V_{DD} - V_{TH})^{1.25} / C_L V_{DD})$. The growth in area, A = 1.9, is the same as in Table 15. The right-most *performance* column is $\frac{F_{MAX}}{A^{-1/\sigma}}$, the ratio of the normalized operating frequency ($\approx F_{MAX}$) to the frequency that results in constant performance in Table 14 with $\Delta\tau_{INT} = 0.2$ (i.e. $\sigma = 1.0$) and $\Delta\tau_{INT} = 1.0$ ($\sigma = 1.4$). It is evident that $\sigma = 1.0$ represents a lower bound for this platform, under the assumption that $\Delta\tau_{INT} \geq 0.2$. On the other hand, it is possible that an architecture that is grossly interconnect–limited may exhibit $\sigma > 1.4$, under worse-case assumptions for interconnect cost.

In mesh-connected organizations such as the PMA platform, the interconnection topology is fixed so that the scaling of load capacitance will depend only on technology. Thus, in the single technology case, $\gamma = 0$. This is the assumption for the first five entries in Table 16 and $\sigma_{MAX} \approx 9.5$

closely approximates the scaling of the maximum frequency, F_{MAX}, as constrained by V and b. Under these assumptions, performance will grow by a factor in the range of 1.5–1.8 for each increase in area, depending on the assumed interconnect cost ($\Delta\tau_{INT}$). At the same time, dynamic power grows by more than three times and subthreshold power by more than 60x over the supply range shown, consistent with the general trend illustrated in Figure 77. Setting $\gamma = -0.3$ (the remaining five lines in Table 16) results in a ~13% per node decrease in the intrinsic delay (approximately equal to the ITRS prediction for LOP technology) and $\sigma_{MAX} \approx -5.1$. As the increase in frequency is proportional to 1/C, the overall dynamic and subthreshold power will remain the same while performance can grow by as much as ~22x over the range of N.

Table 17 Some example power and performance predictions.

V	*b*	σ_{MAX}	χ	χ'	P_D	P_S	Total Power P_T			F_{MAX}	F	Perf.	Rel. Perf.
							$(P_R)_0$=0.1	$(P_R)_0$=0.5	$(P_R)_0$=1				
γ=0													
0.91	0.200	2.30	1.70	1.910	1.18	1.11	1.18	1.16	1.15	0.76	0.76	1.20	0.83
0.89	0.200	1.60	1.60	1.60	1.00	1.00	1.00	1.00	1.00	0.67	0.67	1.06	0.73
0.87	0.200	1.25	1.54	1.420	0.86	0.92	0.87	0.88	0.89	0.60	0.60	0.95	0.65
0.87	0.075	1.90	1.81	0.940	1.03	1.38	1.07	1.17	1.23	0.71	0.71	1.13	0.78
0.86	0.075	1.66	1.79	0.89	0.95	1.35	1.00	1.12	1.18	0.68	0.68	1.07	0.74
0.85	0.075	1.50	1.75	0.840	0.90	1.33	0.95	1.08	1.15	0.65	0.65	1.03	0.71
0.85	0.020	1.90	1.94	0.530	0.99	1.59	1.07	1.26	1.36	0.71	0.71	1.13	0.78
0.84	0.020	1.64	1.89	0.50	0.91	1.56	1.00	1.21	1.32	0.68	0.68	1.07	0.74
0.83	0.020	1.50	1.85	0.460	0.86	1.56	0.97	1.20	1.31	0.65	0.65	1.03	0.71
γ=-0.3													
0.91	0.200	9.00	1.70	1.910	1.21	1.15	1.21	1.19	1.18	0.93	0.93	1.47	0.84
0.89	0.200	3.00	1.60	1.595	1.00	0.99	1.00	0.99	0.99	0.81	0.81	1.28	0.73
0.87	0.200	1.95	1.54	1.420	0.85	0.91	0.86	0.87	0.88	0.72	0.72	1.14	0.65
0.87	0.075	4.20	1.81	0.940	1.02	1.37	1.06	1.16	1.22	0.86	0.86	1.36	0.77
0.86	0.075	3.50	1.79	0.885	0.97	1.36	1.02	1.13	1.20	0.83	0.83	1.32	0.75
0.85	0.075	2.70	1.75	0.840	0.89	1.32	0.95	1.08	1.15	0.79	0.79	1.25	0.71
0.85	0.020	4.10	1.94	0.530	0.97	1.58	1.05	1.24	1.35	0.86	0.86	1.35	0.77
0.84	0.020	3.20	1.89	0.500	0.90	1.56	1.00	1.21	1.32	0.82	0.82	1.29	0.74
0.83	0.020	2.60	1.85	0.460	0.84	1.55	0.95	1.18	1.30	0.78	0.78	1.24	0.70

Table 17 lists the total power scaling from (4.44), frequency from (4.4) and performance for some example (V, b) points on Figure 102, normalized here to the worse-case prediction in Table 14, $\sigma = 1.4$. The right-most column in this case (relative performance) is the normalized performance

(i.e., $\frac{A^{-1/\sigma}}{A^{-1/1.4}}$), divided by the relevant LOP baseline figure in Table 16. Relative performance therefore describes the impact on T of increasing V_{TH} while reducing V_{DD} in order to constrain subthreshold power. Each of the groups of three rows in Table 17 are centered on points 1–3 on Figure 102 (cf. Table 15) and illustrate the tradeoffs between supply scaling, power and performance. Where $\gamma = 0$, operating at any of the central (V, b) points (highlighted in grey on Table 17) will result in constant power scaling at $(P_R)_0 = 0.1$ while achieving a slight increase in performance, around 6–7% per node in Table 17, given by the difference between σ for constant power (e.g., ~1.6) and for constant performance (1.4 in this case). For example, in line 2 of Table 17 above, $\text{Perf.} = A^{-1/1.6} / A^{-1/1.4} \approx 1.9^{0.09} \approx 1.06$.

At the decreasing load capacitance implied by $\gamma = -0.3$, these parallel circuits can be still set up to exhibit constant power, but at a higher relative operating frequency resulting in about a 30% performance improvement per node. As expected, increasing V supports a higher intrinsic performance (larger σ_{MAX}) that can be traded for power in a controlled manner via more aggressive frequency scaling (i.e. by setting $\sigma \leq \sigma_{MAX}$). As σ_{MAX} is the maximum achievable value constrained by the scaling of supply and threshold, σ may be set to any value below this, resulting in a linear tradeoff ($\propto$ F) between performance and power.

5.4 Summary

As identified in Chapter 2, power/energy density will be a key issue affecting computer architecture as technology evolves towards the end of the roadmap. One way to address this issue is to trade off parallel operation (area) for power, for example by using replicated data paths. The effectiveness of this approach depends, in turn, on another important limitation in future architectures, that of interconnection cost. This chapter has analysed the power-area-performance scalability of a computational structure based on thin-body Schottky-barrier transistors that form a homogeneous, mesh-connected reconfigurable fabric using the analytic model developed in the previous chapter i.e., in terms of the circuit parameters χ, χ' and γ, and the architectural parameter

σ. In essence, χ describes the evolution of $I_D(sat)$ with supply and threshold voltages, and establishes bounds on the scaling of performance (frequency) and dynamic power. Similarly, χ' measures the impact of threshold voltage on subthreshold current for the given technology, thereby determining the scaling of subthreshold power. The exponent γ models the change in capacitance that will occur both at the device level, due to technology change, and at the circuit level due to the impact of the interconnection topology. For the mesh-connected topology considered here, load capacitance will depend only on technology, so $\gamma \leq 0$.

This chapter has explored the constraints on σ given the likely characteristics of the homogeneous reconfigurable platform with entirely local connectivity between adjacent cells. The most severe constraint in this case is the assumption that interconnect and logic are formed from the same reconfigurable cells and therefore exhibit similar area-performance tradeoffs. Local interconnect topology is also limited to two lines running in orthogonal directions, reminiscent of the early standard cell routing systems that had just a single metal layer plus polysilicon available for use as interconnect.

Despite worse-case assumptions with respect to variability and interconnect costs, the reconfigurable platform has been shown to support the synthesis of parallel architectures that will, in turn, allow reducing power with area. As an example, it was demonstrated that LOP devices at a single technology node will allow power to be exchanged completely for performance for any architecture capable of about $\sigma < 1.6$. Even under the most severe interconnection constraints, a replication of a simple data path on the reconfigurable array achieved $\sigma < 1.4$. It is highly likely that there are many other parallel organizations that can exhibit a similar or better range of performance. The situation improves further when capacitance reduces at successive nodes. In this case, it is possible to exchange a 2x area increase for a 30% increase on performance per scaling node, while maintaining constant power. Thus, since a characteristic of this platform is that the supply and threshold voltages may be set more-or-less as required, it can be seen that reconfigurable logic arrays of the type studied here will support a continuous tradeoff between power and area.

Chapter 6. Summary, Conclusions and Future Work

"Usefulness is a logarithmic function of technology."

Theo Claasen, in [113]

6.1 Summary

It is evident that a number of key challenges remain as technology moves further into the nanoelectronic domain. It is likely that the combined forces of design effort, manufacturability, reliability, variability and power will favour simple, reconfigurable, locally-connected hardware meshes that merge processing and memory. Power (or more strictly, energy-delay) will be a "first-class" constraint that will need to be managed at all levels in the design hierarchy.

This research has set out to examine whether simple, regular reconfigurable structures with a minimal number of interconnection layers (that therefore have a good chance of achieving sub-10nm feature sizes) will support scalable computer micro-architectures. To examine this issue, the work commenced with an analysis of the characteristics of a plausible (but still hypothetical) Schottky-barrier circuit technology. Thin-body or silicon nanowire devices of this type are beginning to emerge from research laboratories and, although they will still present a challenge to manufacture, at least offer the promise of simplified processing that will allow them to reach truly nanoscale gate dimensions.

A fine-grained, locally connected reconfigurable architecture has been proposed and analysed. The operation of the platform is based on a unique characteristic of thin-body double-gate devices i.e., that the threshold voltage seen at one gate of a double-gate transistor may be substantially altered using the bias on its second gate. This variable threshold concept can greatly reduce the overall subthreshold power of the platform by uncoupling the conflicting requirements of high performance and low standby power. In addition, it allows the functionality of the cells to be

configured. The resulting organization has been described as "polymorphic" as each component cell is capable of being configured as logic, interconnect, or a combination of both. The expected characteristics of this reconfigurable platform have been presented based on a thin-body double-gate silicided source/drain Schottky-barrier technology.

As identified in Chapter 2, power density will be a key issue affecting computer architecture as technology evolves towards the end of the roadmap. Due to the conflicting requirements of high performance and low subthreshold power, as device numbers increase it is likely that there will be little choice but to manage power consumption by exploiting parallelism. In Chapter 4, a new model has been developed that describes how area can be used to reduce power consumption in CMOS, in particular the dynamic and subthreshold power terms. This model identifies a simple set of criteria that describe scalable architectures. It was found that for circuits and/or algorithms that can be characterized in terms of $AT^{\sigma} = K_1$ where $\sigma \leq 2$ holds for many algorithms, dynamic power will scale as: $P_D \propto A^{\gamma\chi} A^{(\sigma-\chi)/\sigma}$ and subthreshold power as: $P_{SUB} \propto A^{\gamma\chi'} A^{(\sigma-\chi')/\sigma}$. The parameters χ, χ' and γ depend on physical issues such as technology and circuit design as well as the growth of interconnection capacitance and the impact of variability. The parameter σ can be interpreted as describing how *serial* an architecture is i.e., how much performance improvement can be traded for an increase in area. It must be remembered that the model says very little about the *actual* power consumption of a system, in that the technology parameters χ, χ' and γ refer to the ratio of saturation drive current, subthreshold current and capacitance to some notional reference. It characterizes the limits on the ability to trade frequency for power (and/or energy), given a particular combination of technology and architecture and given the ability to choose a particular sequence of supply and threshold values.

The model has been applied to the reconfigurable platform and its power-area-performance characteristics have been analysed in terms of the parameters σ, χ and χ'. While it might reasonably be expected that σ would be severely impacted by the assumption that the interconnect and logic for this platform are formed from the same reconfigurable cells and therefore exhibit similar area-performance tradeoffs, this was found not to be the case. Even under a worst-case assumption

where interconnect and logic delays are equivalent, it appears that $\sigma<1.4$ will be achievable for regular, parallel architectures of the type studied here.

6.2 Conclusions

The work in this thesis has been concerned with the evolution of digital hardware systems as devices scale towards the end of the CMOS roadmap. It has been motivated by two related questions: (1) can complex heterogeneous micro-architectures be based on regular homogeneous, reconfigurable circuit structures and (2) can these structures be made ultimately scalable to support the massive transistor counts expected at end-of-roadmap dimensions? These questions have been answered using a hierarchical simulation approach using a number of plausible assumptions about future device trends.

Using a combination of TCAD and SPICE circuit simulation, it has been shown that the characteristics of fully depleted dual-gate thin-body Schottky barrier silicon transistors will not only support low-power operation *per se*, but will also allow the development of a locally-connected reconfigurable computing mesh. Unlike a conventional FPGA, the fabric would be configured by changing the bias on the control gates of the dual-gate complementary devices such that the switching threshold seen at the other gate is moved, thereby altering the logic performed by the gate. Organizing these devices into simple arrays permits the development of complex *heterogeneous* computing functions from this homogeneous, mesh-connected organization.

The results derived from SPICE simulations of simple circuits have shown that dynamically shifting the threshold voltage of these TB-FDSBSOI circuits may reduce subthreshold power loss by in excess of 10^3 in some typical cases examined in this work. As the back gate of a double gate transistor presents approximately the same load as the front gate, such mode-switching can be achieved at normal circuit speeds. The TCAD simulations have also indicated that the magnitude of the threshold shift effect will scale with device dimensions and will remain compatible with device reliability constraints.

A simple array topology based on a 6-input, 6-output NOR block has been analysed and a number of combinational logic and asynchronous state machines have been demonstrated using the reconfigurable technique. Unlike conventional reconfigurable architectures (e.g., commercial FPGAs), this basic component is used interchangeably for both logic and interconnection. Even so it has proved fairly straightforward to derive circuits using conventional asynchronous design techniques that are more-or-less equivalent to current FPGA blocks comprising LUTs, flip-flops, multiplexers etc. This organization has been shown to offer substantial reductions in the overall implementation size of circuits exhibiting low Rent exponents due to its ability to form consolidated functional blocks as and where required.

It has been found in this work that for circuits and/or algorithms that can be characterized in terms of $AT^{\sigma} = K_1$ (where $\sigma \leq 2$ holds for many algorithms), dynamic and static power will scale as: $P_{DYN} \propto A^{\gamma\alpha}A^{(\sigma-\chi)/\sigma}$ and $P_{SUB} \propto A^{\gamma\alpha'}A^{(\sigma-\chi')/\sigma}$, respectively. Here, χ, χ' and γ describe the relationship between supply voltage and drive current, subthreshold current and load capacitance, respectively, while σ describes the relationship between performance (operating frequency) and the area of a given architecture or micro-architecture. These equations represent an *optimum* scaling case in which changes in supply and threshold voltages result in frequency reductions exactly track changes in performance. The values of χ and χ' may be obtained from small-scale SPICE simulations of representative devices, whereas σ may be derived using an abstract-level architectural simulator. Thus all of these parameters could be made available early in the design cycle, making them a useful way of gauging high-level architectural tradeoffs.

It must be reiterated that the model says very little about the *actual* performance and/or power consumption of a system, which depends on a range of issues such as technology, design and layout style, the activity ratio as well as the specific values of supply and threshold voltage. The parameters of this model refer to various *ratios* normalized a notional reference. The various examples in this work have shown how this choice impacts on the area-power tradeoffs. Similarly, the architectural parameter σ relates area to performance (as $F \propto A^{-1/\sigma}$) for a particular scalable organization, normalized to a baseline configuration. On the other hand, the analysis does

indicate that mesh-connected topologies of the type studied here will exhibit much the same area-performance tradeoffs as conventional organizations (i.e., the same range of σ).

The parameter σ is often taken to be a measure of the quality of a design. The higher the value of σ, the more *serial* a particular architecture is, and therefore the harder it will be to find a suitable tradeoff between area and power. The main impacts on σ in this work have been the growth of interconnection capacitance and the issue of threshold variability. The analysis shows that the overall impact of variability on the target operating frequency and voltage will not be as great as might be expected. For example, a worse-case +25% shift in V_{TH} will move the target frequency scaling by less than 10%. This sort of figure is well within the range that can be reclaimed at the architectural level.

Mesh-connected organizations automatically constrain the interconnect capacitance to a minimum value set primarily by the gate technology and the cell geometry (fanout), and therefore will at least partially compensate for the poor current drive behavior of end-of-roadmap technologies such as thin-body Schottky-barrier transistors. Performance (frequency) will continue to improve simply from reductions in gate capacitance and increasingly ballistic operation at future technology nodes and it will be possible to exchange some or all of this improvement for constant power or energy.

Ultimately, low-power/energy operation will require careful design at multiple levels of abstraction: for example, by using intrinsically low-leakage device technology, using localized asynchronous circuits to minimize the number of switching events per cycle while simultaneously exploiting area to implement highly parallel (i.e., low σ) versions of the particular algorithm. Taken together, these techniques will support reductions in operating voltage and frequency, allowing performance to be traded for power and/or energy in a controlled manner. Thus, given that these operating points may be set as required, locally-connected reconfigurable logic arrays of the type studied in this thesis represent scalable, low-power building blocks that will be well suited to future nanoscale computer architecture.

6.3 Summary of the Scalability Analysis Methodology

This section provides a short summary of the scalability analysis methodology developed in Chapter 4 and demonstrated in Chapter 5. The intention is to illustrate the technique of applying the model to the analysis of a real system. The overall power scaling function was shown in (4.44), which is reproduced below:

$$P_T = \left(\frac{1}{1 + (P_R)_0 V^{(\eta-\beta)}} \right) \left(A^{\chi} A^{(\sigma-\chi)/\sigma} + (P_R)_0 V^{(\eta-\beta)} A^{\chi'} A^{(\sigma-\chi')/\sigma} \right)$$

It can be seen that this power scaling model is based on three main parameters: an architectural parameter σ that links overall performance (operating frequency) to area, two terms that relate current to supply voltage (V) for a particular technology and design style, incorporating contributions from drive current: $\chi = \frac{\beta+1}{\beta-1}$ and subthreshold current: $\chi' = \frac{\eta+1}{\beta-1}$, and finally a parameter (γ) that models the effect of capacitance scaling with area. Where the analysis is undertaken at a fixed technology node, γ can be set to zero so that the terms A^{χ} and $A^{\chi'} \rightarrow 1$.

The values of χ and χ' would typically be obtained from small–scale SPICE simulations of representative devices. The initial (pre-analysis) design choices will encompass a choice of implementation technology, or desired sequence of technologies (e.g., drawn from the ITRS), a representative set of circuit designs and the desired range of supply voltage scaling values (potentially, also derived from the ITRS). The representative circuits chosen would need to be no more complex than a single–bit arithmetic function or flip-flop, which are amenable to rapid SPICE simulation. The requirement is to determine the average propagation delay (τ) and leakage current across the expected range of supply voltage (V) and threshold (V_{TH}). The parameter β can be determined from $I_D = \frac{CV_{DD}}{\tau} \propto V^{\beta}$ while η is obtained directly from the average value of $I_{SUB} \propto V^{\eta}$. In both cases, the exponent can be most easily determined by a curve fitting process relating V to I_D or I_{SUB} via simple power-law functions as shown previously (e.g., Figure 102). The ratio $\frac{I_{SUB}}{I_D}$ also

describes the power ratio P_R at a given V_{DD}. The load parameter γ will most easily be determined from the evolution of average gate and interconnect capacitance extracted directly from the characteristics of the technology and the expected scaling of device numbers (N, $\propto$ A). Where the technology does not change, $\gamma = 0$ whereas at successive technology nodes we would expect both N and the average load capacitance (C_L) to change together such that $C_L \propto A^{\gamma}$.

The architectural parameter, σ, would be derived most efficiently from a high–level architectural simulator. In this case, we are trying to reveal the relationship between performance and area in parallel architectures such as the examples in Chapter 5. Those examples were derived using a purpose-built simulation written in VHDL-AMS, as this allowed β and η to be extracted from the same simulations. However, these simple area–time relationships are actually the primary objective of virtually all architectural simulators and so these data would be readily derived using existing and standard EDA tools. Again, σ can be determined by curve fitting the area–time data to a simple power-law function.

In summary, it would be a fairly straightforward process for all of the necessary parameters to be made available early in the design cycle, so that the model as presented would support the efficient evaluation of tradeoffs between area, performance and power in a range of digital circuits. It should be noted, however, that because of some of the simplifying assumptions made, the model is most particularly applicable to circuits that exhibit the type of extreme regularity shown by the array developed in Chapter 3.

6.4 Summary of Contributions

The work reported in this dissertation has resulted in the following specific contributions:

- The demonstration, by TCAD simulation, that ultra-thin body, double-gated fully depleted Schottky barrier SOI transistors will support a low-overhead reconfigurable computing platform.

- The specification and analysis of a regular, adjacent-connected array based on the TBFD-SBSOI devices, firstly by low-level TCAD and SPICE simulation and then via a register transfer level (RTL) simulation using behavioral models derived from the previous TCAD and SPICE work.

- A demonstration via high level simulation that the resulting 6x6 LUT array is equivalent to more complex FPGA-like organizations and will support power-scalable reconfigurable systems.

- The development of a new analytic approach to power and energy vs. area based on a traditional architectural complexity metric of the form $AT^{\sigma} = K$. This defines the limits on the area-performance tradeoffs for architectures that will support massive area scaling.

- The verification by simulation that architectures mapped to the LUT array may be described by the analytic relationship developed between area and power/energy, and that this will predict their ultimate scalability.

6.5 Proposed Future Research

Achieving low-power operation using the double gate threshold technique proposed here would be especially applicable in cases where circuit activity can be easily detected, such as spatial computing architectures based on connected asynchronous operators that exhibit only localized communication. Previous work on the application of variable threshold techniques has tended to focus on the static assignment of the high and low threshold devices within the circuit. However, as activity in an asynchronous architecture is typically controlled by handshaking signals at the interface between operators, it would be straightforward to determine when a particular part of the system is active. Thus, it is an obvious strategy to use these signals to dynamically adjust the threshold according to the state of the circuits (active or inactive), especially given that double-gate circuits of the type analysed here can switch between high and low power modes at circuit speeds. It would be useful to examine both the dynamic and static behavior of memory circuits as well as more complex circuits compiled directly from high level language.

Although the move towards nanoscale dimensions has resulted in a renewed interest in spatial computing models, it is not yet clear whether these will be directly applicable to domains such as embedded computing, in which the key drivers are chip cost and power. It was observed in Chapter 2 that fewer than 15% of the instructions generated by a compiler for a standard ISA might be expected to appear as structures in a compiled spatial architecture, with the remainder representing communication channels between operators. Organizations such as Globally Asynchronous, Locally Synchronous (GALS) appear to offer a simple, locally connected framework that may support the efficient mapping of general-purpose code onto spatial architectures. As the reconfigurable platform developed in this work will directly support the alternative instantiation of small asynchronous state machines and/or interconnect, compiled code mapped to this platform using a GALS framework may turn out to be particularly area-efficient. Further research is needed to determine if this is, indeed, the case, and to evaluate the impact of the interconnection cost in systems such as this.

References

[1] M. Lundstrom, "Moore's Law Forever?," *Science*, vol. 299, pp. 210-211, 2003.

[2] M. Wright. (2006) Milestones That Mattered: The Planar IC—Revolution Underestimated. *EDN,* [Online]. Available: http://www.edn.com/index.asp?layout= article&articleid=CA6325586.

[3] B. T. Murphy, D. E. Haggan, and W. W. Troutman, "From Circuit Miniaturization to the Scalable IC," *Proceedings of the IEEE*, vol. 88, pp. 691-703, 2000.

[4] SIA. (2005) SICAS Capacity and Utilization Rates Q3 2005. [Online]. Available: http://www.sia-online.org/pre_stat.cfm?ID=278.

[5] R. E. Kessler, "The Alpha 21264 Microprocessor," *IEEE Micro*, vol. 19, pp. 24-36, 1999.

[6] A. Jain, W. Anderson, T. Benninghoff, D. Bertucci, B. M., J. Burnette, T. Chang, J. Eble, R. Faber, D. Gowda, J. Grodstein, G. Hess, J. Kowaleski, A. Kumar, M. B., R. Mueller, P. Paul, J. Pickholtz, S. Russell, M. Shen, T. Truex, A. Vardharajan, D. Xanthopoulos, and T. Zou, "A 1.2GHz Alpha Microprocessor with 44.8GB/s Chip Pin Bandwidth," presented at IEEE International Solid-State Circuits Conference, San Francisco, 2001.

[7] C. McNairy and D. Soltis, "Itanium 2 Processor Microarchitecture," *IEEE Micro*, vol. 23, pp. 44-55, 2003.

[8] C. McNairy and R. Bhatia, "Montecito: A Dual-Core, Dual-Thread Itanium Processor," *IEEE Micro*, vol. 25, pp. 10-20, 2005.

[9] R. Compano, *Technology Roadmap for Nanoelectronics*, 2nd ed: European Commission IST Programme - Future and Emerging Technologies, 2000.

[10] P. Gargini, "Silicon Nanoelectronics and Beyond," *Journal of Nanoparticle Research*, vol. 6, pp. 11-26, 2004.

[11] (2005) International Technology Roadmap for Semiconductors, 2005 Edition. [Online]. Available: http://www.itrs.net/Links/2005ITRS/Home2005.htm.

[12] (2003) International Technology Roadmap for Semiconductors, 2003 Edition. [Online]. Available: http://www.itrs.net/Links/2003ITRS/Home2003.htm.

[13] T. Mudge, "Power: A First-Class Architectural Design Constraint," *Computer*, vol. 34, pp. 52-58, 2001.

[14] J. D. Meindl, "Beyond Moore's Law: the Interconnect Era," *Computing in Science & Engineering*, vol. 5, pp. 20-24, 2003.

[15] P. Beckett and A. Jennings, "Towards Nanocomputer Architecture," presented at Seventh Asia-Pacific Computer Systems Architecture Conference, ACSAC'2002, Melbourne, Australia, 2002.

[16] J. Lach, W. H. Mangione-Smith, and M. Potkonjak, "Low Overhead Fault-Tolerant FPGA Systems," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 6, pp. 212 -221, 1998.

[17] S. Goldstein, M. Budiu, M. Mishra, and G. Venkataramani, "Reconfigurable Computing and Electronic Nanotechnology," presented at IEEE 14th International Conference on Application-specific Systems, Architectures and Processors, The Hague, Netherlands, 2003.

[18] J. Han and P. Jonker, "A Defect- and Fault-Tolerant Architecture for Nanocomputers," *Nanotechnology*, vol. 14, pp. 224–230, 2003.

[19] A. DeHon, "Reconfigurable Architectures for General-Purpose Computing," MIT, Massachusetts, A.I. Technical Report 1586, October 1996.

[20] C. Lallement, F. Pêcheux, and Y. Hervé, "VHDL-AMS Design of a MOST Model Including Deep Submicron and Thermal-Electronic Effects," presented at IEEE International Workshop on Behavioral Modeling and Simulation, BMAS 2001, FountainGrove Inn Santa Rosa, California, USA, 2001.

[21] Mentor Graphics. [Online]. Available: http://www.mentor.com.

[22] J. D. Meindl, "Low Power Microelectronics: Retrospect and Prospect," *Proceedings of the IEEE*, vol. 83, pp. 619-635, 1995.

[23] R. K. Cavin III and V. V. Zhirnov, "Future Devices for Information Processing," presented at 31st European Solid-State Circuits Conference, ESSCIRC 2005, pp. 7-12, 2005.

[24] V. V. Zhirnov, R. K. Cavin, III, J. A. Hutchby, and G. I. Bourianoff, "Limits to Binary Logic Switch Scaling - a Gedanken Model," *Proceedings of the IEEE*, vol. 91, pp. 1934-1939, 2003.

[25] J. D. Meindl and J. A. Davis, "*The* Fundamental Limit on Binary Switching Energy for Terascale Integration (TSI)," *IEEE Journal of Solid-State Circuits*, vol. 35, pp. 1515-1516, 2000.

[26] G. Schrom and S. Selberherr, "Ultra-Low-Power CMOS Technologies," presented at International Semiconductor Conference, pp. 237 -246 vol.1, 1996.

[27] J. D. Plummer and P. B. Griffin, "Material and Process Limits in Silicon VLSI Technology," *Proceedings of the IEEE*, vol. 89, pp. 240-258, 2001.

[28] M. Levenson. (2007) Intel Touts Working 45nm Chip with High-k, Metal Gates. *Solid State Technology,* [Online]. Available: http://sst.pennnet.com/Articles/ Article_Display.cfm?Section=ONART&PUBLICATION_ID=5&ARTICLE_ID=283213&C=TECHN.

[29] J. D. Meindl, Q. Chen, and J. A. Davis, "Limits on Silicon Nanoelectronics for Terascale Integration," *Science*, vol. 293, pp. 2044-2049, 2001.

[30] M. V. Fischetti, "Scaling MOSFETs to the Limit: A Physicists's Perspective," *Journal of Computational Electronics*, vol. 2, pp. 73-79, 2003.

[31] Y. C. Yeo, Q. Lu, W. C. Lee, T.-J. King, C. Hu, X. Wang, X. Guo, and T. P. Ma, "Direct Tunneling Gate Leakage Current in Transistors with Ultrathin Silicon Nitride Gate Dielectric," *IEEE Electron Device Letters*, vol. 21, pp. 540-542, 2000.

[32] B. Tavel, M. Bidaud, N. Emonet, D. Barge, N. Planes, H. Brut, D. Roy, J. C. Vildeuil, R. Difrenza, K. Rochereau, M. Denais, V. Huard, P. Llinares, S. Bruyere, C. Parthasarthy, N. Revil, R. Pantel, F. Guyader, L. Vishnubotla, K. Barla, F. Arnaud, P. Stolk, and M. Woo, "Thin Oxynitride Solution for Digital and Mixed-Signal 65nm CMOS Platform," presented at IEEE International Electron Devices Meeting, IEDM '03, pp. 27.6.1-27.6.4, 2003.

[33] P. P. Gelsinger, "Microprocessors for the New Millennium: Challenges, Opportunities, and New Frontiers," presented at International Solid-State Circuits Conference, ISSCC'01, San Francisco, USA, 2001.

[34] D. Boggs, A. Baktha, J. Hawkins, D. T. Marr, J. A. Miller, P. Roussel, R. Singhal, B. Toll, and K. S. Venkatraman, "The Microarchitecture of the Intel® Pentium® 4 Processor on 90nm Technology.," *Intel Technology Journal*, vol. 8, pp. 1-18, 2004.

[35] B. Davari, "CMOS Technology: Present and Future," presented at IEEE Symposium on VLSI Circuits, pp. 5-9, 1999.

[36] S. Asai and Y. Wada, "Technology Challenges for Integration Near and Below 0.1um," *Proceedings of the IEEE*, vol. 85, pp. 505-520, 1997.

[37] Y. Taur, D. A. Buchanan, W. Chen, D. J. Frank, K. E. Ismail, S.-H. Lo, G. A. Sai-Halasz, R. G. Viswanathan, H.-J. C. Wann, S. J. Wind, and H.-S. Wong, "CMOS Scaling into the Nanometer Regime," *Proceedings of the IEEE*, vol. 85, pp. 486 - 504, 1997.

[38] S. H. Tang, L. Chang, N. Lindert, Y.-K. Choi, W.-C. Lee, X. Huang, V. Subramanian, J. Bokor, T.-J. King, and C. Hu, "FinFET — A Quasi-Planar Double-Gate MOSFET," presented at IEEE International Solid State Circuits Conference, ISSCC 2001, San Francisco, USA, 2001.

[39] B. Doyle, R. Arghavani, D. Barlage, S. Datta, M. Doczy, J. Kavalieros, A. Murthy, and R. Chau. (2002) Transistor Elements for 30nm Physical Gate Lengths and Beyond. *Intel Technology Journal,* [Online]. Available: http://developer.intel.com/technology/itj/ archive/2002.htm.

[40] P. J. Silverman. (2002) The Intel Lithography Roadmap. *Intel Technology Journal,* [Online]. Available: http://developer.intel.com/technology/itj/archive/2002.htm.

[41] V. Sverdlov, Y. Naveh, and K. Likharev, "Nanoscale SOI Ballistic MOSFETs: an Impending Power Crisis," presented at IEEE International SOI Conference, pp. 151-152, 2001.

[42] K. Saraswat. (2006) EE311 Notes - Future Devices Part 1. [Online]. Available: www.stanford.edu/class/ee311/NOTES/Future Devices.pdf.

[43] G. W. McFarland, CMOS Technology Scaling and its Impact on Cache Delay, PhD Thesis, Electrical Engineering, Stanford University, Boston, 1997.

[44] A. R. Brown, J. R. Watling, and A. Asenov, "A 3-D Atomistic Study of Archetypal Double Gate MOSFET Structures," *Journal of Computational Electronics*, vol. 1, pp. 165-169, 2002.

[45] Z. Ren, R. Venugopal, S. Datta, M. Lundstrom, D. Jovanovic, and J. Fossum, "The Ballistic Nanotransistor: a Simulation Study," presented at International Electron Devices Meeting, IEDM 2000, San Francisco, CA, USA, pp. 715 - 718, 2000.

[46] J. P. Colinge, J. T. Park, and C. A. Colinge, "SOI Devices for Sub-0.1µm Gate Lengths," presented at 23rd International Conference on Microelectronics, MIEL 2002., Nis, Yugoslavia, pp. 109-113 vol.1, 2002.

[47] Y. Kado, "The Potential of Ultrathin-Film SOI Devices for Low-Power and High-Speed Applications," *IEICE Transactions on Electronics*, vol. E80-C, pp. 443-454, 1997.

[48] T. Lepselter and S. M. Sze, "SB-IGFET: An Insulated-Gate Field-Effect Transistor using Schottky Barrier Contacts for Source and Drain," *Proceedings of the IEEE*, vol. 56, pp. 1400–1401, 1968.

[49] J. R. Tucker and T. C. Shen, "New Approaches to Silicon Nanoelectronics," *Future Electron Devices Journal*, vol. 9, pp. 5-14, 1998.

[50] H.-C. Lin, M.-F. Wang, F.-J. Hou, J.-T. Liu, F.-H. Ko, H.-L. Chen, G.-W. Huang, T.-Y. Huang, and S. M. Sze, "Nano-Scale Implantless Schottky-Barrier SOI FinFETs with Excellent Ambipolar Performance," presented at 60th Device Research Conference, DRC2002, pp. 45-46, 2002.

[51] M. Nishisaka, S. Matsumoto, and T. Asano, "Schottky Source/Drain SOI MOSFET with Shallow Doped Extension," *Japanese Journal Applied Physics*, vol. 42, Part 1, pp. 2009-2013, 2003.

[52] J. R. Tucker, "Schottky Barrier MOSFETS for Silicon Nanoelectronics," presented at Advanced Workshop on Frontiers in Electronics, WOFE '97, pp. 97 - 100, 1997.

[53] M. Jang, Y. Kim, M. Jeon, C. Choi, I. Baek, S. Lee, and B. Park, "N_2-Annealing Effects on Characteristics of Schottky-Barrier MOSFETS," *IEEE Transactions on Electron Devices*, vol. 53, pp. 1821-1825, 2006.

[54] M. Jang, S. Lee, and K. Park, "Erbium Silicided n-Type Schottky Barrier Tunnel Transistors for Nanometer Regime Applications," *IEEE Transactions on Nanotechnology*, vol. 2, pp. 205-209, 2003.

[55] M. Bescond, J. L. Autran, D. Munteanu, N. Cavassilas, and M. Lannoo, "Atomic-scale Modeling of Source-to-Drain Tunneling in Ultimate Schottky Barrier Double-Gate MOSFET's," presented at 33rd Conference on European Solid-State Device Research, ESSDERC '03., pp. 395-398, 2003.

[56] M. Ieong, P. M. Solomon, S. E. Laux, H.-S. P. Wong, and D. Chidambarrao, "Comparison of Raised and Schottky Source/Drain MOSFETs Using a Novel Tunneling Contact Model," presented at International Electron Devices Meeting, IEDM '98, San Francisco, CA, USA, pp. 733 -736, 1998.

[57] D. Connelly, C. Faulkner, P. A. Clifton, and D. E. Grupp, "Fermi-Level Depinning for Low-Barrier Schottky Source/Drain Transistors," *Applied Physics Letters*, vol. 88, pp. 012105-3, 2006.

[58] A. J. Strojwas, M. Quarantelli, J. Borel, C. Guardiani, G. Nicollini, G. Crisenza, and B. Franzini, "Manufacturability of Low Power CMOS Technology Solutions," presented at International Symposium on Low Power Electronics and Design, pp. 225-232, 1996.

[59] A. Asenov, A. R. Brown, J. H. Davies, S. Kaya, and G. Slavcheva, "Simulation of Intrinsic Parameter Fluctuations in Decananometer and Nanometer-Scale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1837-1852, 2003.

[60] S. T. Ma, A. Keshavarzi, V. De, and J. R. Brews, "A Statistical Model for Extracting Geometric Sources of Transistor Performance Variation," *IEEE Transactions on Electron Devices*, vol. 51, pp. 36-41, 2004.

[61] K. A. Bowman, S. G. Duvall, and J. D. Meindl, "Impact of Die-to-Die and Within-Die Parameter Fluctuations on the Maximum Clock Frequency Distribution for Gigascale Integration," *IEEE Journal of Solid-State Circuits*, vol. 37, pp. 183-190, 2002.

[62] K. Takeuchi, R. Koh, and T. Mogami, "A Study of the Threshold Voltage Variation for Ultra-Small Bulk and SOI CMOS," *IEEE Transactions on Electron Devices*, vol. 48, pp. 1995-2001, 2001.

[63] T. Mizuno, J. Okumtura, and A. Toriumi, "Experimental Study of Threshold Voltage Fluctuation due to Statistical Variation of Channel Dopant Number in MOSFET's," *IEEE Transactions on Electron Devices*, vol. 41, pp. 2216-2221, 1994.

[64] H. Tuinhout, "Impact of Parametric Fluctuations on Performance and Yield of Deep-Submicron Technologies," presented at 32nd European Solid-State Device Research Conference, pp. 95-102, 2002.

[65] T. Mizuno, J. Okamura, and A. Toriumi, "Experimental Study Of Threshold Voltage Fluctuations Using An 8k MOSFET's Array," presented at Symposium on VLSI Technology, pp. 41-42, 1993.

[66] S. Xiong and J. Bokor, "Sensitivity of Double-Gate and FinFET Devices to Process Variations," *IEEE Transactions on Electron Devices*, vol. 50, pp. 2255-2261, 2003.

[67] T. Shinada, S. Okamoto, T. Kobayashi, and I. Ohdomari, "Enhancing Semiconductor Device Performance using Ordered Dopant Arrays," *Nature*, vol. 437, pp. 1128-1131, 2005.

[68] J. Pineda de Gyvez and H. P. Tuinhout, "Threshold Voltage Mismatch and Intra-die Leakage Current in Digital CMOS Circuits," *IEEE Journal of Solid-State Circuits*, vol. 39, pp. 157-168, 2004.

[69] S.-D. Kim, H. Wada, and J. C. S. Woo, "TCAD-Based Statistical Analysis and Modeling of Gate Line-Edge Roughness Effect on Nanoscale MOS Transistor Performance and Scaling," *IEEE Transactions on Semiconductor Manufacturing*, vol. 17, pp. 192-200, 2004.

[70] K. Samsudin, B. Cheng, A. R. Brown, S. Roy, and A. Asenov, "Integrating Intrinsic Parameter Fluctuation Description into BSIMSOI to Forecast Sub-15 nm UTB SOI Based 6T SRAM Operation," *Solid-State Electronics*, vol. 50, pp. 86-93, 2006.

[71] J. Kedzierski, P. Xuan, E. H. Anderson, J. Bokor, T.-J. King, and C. Hu, "Complementary Silicide Source/Drain Thin-Body MOSFETs for the 20 nm Gate Length Regime," presented at International Electron Devices Meeting, IEDM2000, San Francisco, CA , USA, pp. 57-60, 2000.

[72] B. Agrawal, V. K. De, and J. D. Meindl, "Device Parameter Optimization for Reduced Short Channel Effects in Retrograde Doping MOSFET's," *IEEE Transactions on Electron Devices*, vol. 43, pp. 365-368, 1996.

[73] Y. Taur, C. H. Wann, and D. J. Frank, "25 nm CMOS Design Considerations," presented at International Electron Devices Meeting, IEDM '98, San Francisco, CA , USA, pp. 789-792, 1998.

[74] K. Nagase, S. I. Ohkawa, M. Aoki, and H. Masuda, "Variation Status in 100nm CMOS Process and Below," presented at The International Conference on Microelectronic Test Structures, ICMTS '04, pp. 257-261, 2004.

[75] K. Nose and T. Sakurai, "Optimization of V_{DD} and V_{TH} for Low-Power and High-Speed Applications," presented at Asia and South Pacific Design Automation Conference, ASP-DAC 2000, Yokohama Japan, pp. 469-474, 2000.

[76] P. K. Ko, J. Huang, Z. Liu, and C. Hu, "BSIM3 for Analog and Digital Circuit Simulation," presented at IEEE Symposium on VLSI Technology CAD, pp. 400-429, 1993.

[77] S. Borkar. (2004) Exponential Challenges, Exponential Rewards - The Future of Moore's Law. [Online]. Available: http://www.nanohub.org/resources/?id=177.

[78] E. J. Nowak, "Maintaining the Benefits of CMOS Scaling when Scaling Bogs Down," *IBM Journal of Research & Development*, vol. 46, pp. 169-180, 2002.

[79] K. Skadron, M. R. Stan, W. Huang, S. Velusamy, K. Sankaranarayanan, and D. Tarjan, "Temperature-Aware Microarchitecture," presented at 30th Annual International Symposium on Computer Architecture, pp. 2-13, 2003.

[80] Predictive Technology Model (PTM). [Online]. Available: http://www.eas.asu.edu/~ptm/.

[81] P. D. Fisher and R. Nesbitt, "The Test of Time. Clock-Cycle Estimation and Test Challenges for Future Microprocessors," *IEEE Circuits and Devices Magazine*, vol. 14, pp. 37-44, 1998.

[82] A. B. Kahng and S. Muddu, "An Analytical Delay Model for RLC Interconnects," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 16, pp. 1507 - 1514, 1997.

[83] R. Sabelka, C. Harlander, and S. Selberherr, "The State of the Art in Interconnect Simulation," presented at International Conference on Simulation of Semiconductor Processes and Devices, SISPAD 2000, Seattle, WA, USA, pp. 6 -11, 2000.

[84] J. A. Davis and J. D. Meindl, "Compact Distributed RLC Interconnect Models-Part I: Single Line Transient, Time Delay, and Overshoot Expressions," *IEEE Transactions on Electron Devices*, vol. 47, pp. 2068-2077, 2000.

[85] J. A. Davis, R. Venkatesan, A. Kaloyeros, M. Beylansky, S. J. Souri, K. Banerjee, K. C. Saraswat, A. Rahman, R. Reif, and J. D. Meindl, "Interconnect Limits on Gigascale Integration (GSI) in the 21st Century," *Proceedings of the IEEE*, vol. 89, pp. 305 - 324, 2001.

[86] A. Naeemi, J. A. Davis, and J. D. Meindl, "Analysis and Optimization of Coplanar RLC Lines for GSI Global Interconnection," *IEEE Transactions on Electron Devices*, vol. 51, pp. 985-994, 2004.

[87] J. A. Davis and J. D. Meindl, "Compact Distributed RLC Interconnect Models-Part II: Coupled Line Transient Expressions and Peak Crosstalk in Multilevel Networks," *IEEE Transactions on Electron Devices*, vol. 47, pp. 2078-2087, 2000.

[88] W. Wu, G.-Y. Jung, D. L. Olynick, J. Straznicky, Z. Li, X. Li, D. A. A. Ohlberg, Y. Chen, S.-Y. Wang, J. A. Liddle, W. M. Tong, and R. S. Williams, "One-Kilobit Cross-Bar Molecular Memory Circuits at 30-nm Half-Pitch Fabricated by Nanoimprint Lithography," *Applied Physics A: Materials Science & Processing*, vol. 80, pp. 1133-1389, 2005.

[89] H. W. Johnson and M. Graham, *High-Speed Digital Design: A Handbook of Black Magic*: Prentice Hall PTR, 1993.

[90] S. Borkar, "Design Challenges of Technology Scaling," *IEEE Micro*, vol. 19, pp. 23-29, 1999.

[91] D. Sylvester and K. Keutzer, "Impact of Small Process Geometries on Microarchitectures in Systems on a Chip," *Proceedings of the IEEE*, vol. 89, pp. 467 - 489, 2001.

[92] D. Matzke, "Will Physical Scalability Sabotage Performance Gains?," *IEEE Computer*, vol. 30, pp. 37-39, 1997.

[93] R. Ho, K. W. Mai, and M. A. Horowitz, "The Future of Wires," *Proceedings of the IEEE*, vol. 89, pp. 490 -504, 2001.

[94] D. Sylvester and K. Keutzer, "Rethinking Deep-Submicron Circuit Design," *IEEE Computer*, vol. 32, pp. 25 - 33, 1999.

[95] P. D. Fisher and R. Nesbitt, "Clock-cycle Estimation and Test Challenges for Future Microprocessors," *IEEE Circuits and Devices*, pp. 37, 1998.

[96] T. Skotnicki, F. Boeuf, M. Müller, A. Pouydebasque, C. Fenouillet-Béranger, R. Cerutti, S. Harrison, S. Monfray, B. Dumont, and F. Payet. (2005) MASTAR 4.0 User's Guide. [Online]. Available: http://www.itrs.net/Links/2005ITRS/Linked%20Files/2005Files/PIDS/MASTARDownload.htm.

[97] K. Fujimaru and H. Matsumura, "Theoretical Consideration of a New Nanometer Transistor Using Metal/Insulator Tunnel-Junction," *Japanese Journal of Applied Physics*, vol. 35, Part 1, pp. 1781-1786, 1996.

[98] K. Fujimaru, R. Sasajima, and H. Matsumura, "Nanoscale Metal Transistor Control of Fowler–Nordheim Tunneling Currents Through 16 nm Insulating Channel," *Journal of Applied Physics*, vol. 85, pp. 6912-6916, 1999.

[99] A. Bachtold, P. Hadley, T. Nakanishi, and C. Dekker, "Logic Circuits with Carbon Nanotube Transistors," *Science*, vol. 294, pp. 1317-1320, 2001.

[100] S. J. Wind, J. Appenzeller, R. Martel, V. Derycke, and P. Avouris, "Vertical Scaling of Carbon Nanotube Field-Effect Transistors using Top Gate Electrodes," *Applied Physics Letters*, vol. 80, pp. 3817-3819, 2002.

[101] T. Ohmi, S. Sugawa, K. Kotani, M. Hirayama, and A. Morimoto, "New Paradigm of Silicon Technology," *Proceedings of the IEEE*, vol. 89, pp. 394 - 412, 2001.

[102] R. C. Merkle, "Design Considerations for an Assembler," *Nanotechnology*, vol. 7, pp. 210-215, 1996.

[103] K. Chen, C. Hu, P. Fang, M. R. Lin, and D. L. Wollesen, "Predicting CMOS Speed with Gate Oxide and Voltage Scaling and Interconnect Loading Effects," *IEEE Transactions on Electron Devices*, vol. 44, pp. 1951-1957, 1997.

[104] T. Sakurai and A. R. Newton, "Alpha-Power Law MOSFET Model and its Applications to CMOS Inverter Delay and other Formulas," *IEEE Journal of Solid-State Circuits*, vol. 25, pp. 584-594, 1990.

[105] K. A. Bowman, B. L. Austin, J. C. Eble, X. Tang, and J. D. Meindl, "A Physical Alpha-Power Law MOSFET Model," *IEEE Journal of Solid-State Circuits*, vol. 34, pp. 1410-1414, 1999.

[106] M. Lundstrom and Z. Ren, "Essential Physics of Carrier Transport in Nanoscale MOSFETs," *IEEE Transactions on Electron Devices*, vol. 49, pp. 133-141, 2002.

[107] Z. Ren, R. Venugopal, S. Goasguen, S. Datta, and M. S. Lundstrom, "nanoMOS 2.5: A Two-Dimensional Simulator for Quantum Transport in Double-Gate MOSFETs," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1914-1925, 2003.

[108] A. Svizhenko and M. P. Anantram, "The Effect of Scattering on Drive Current of Nanotransistors," presented at 60th Device Research Conference, DRC 2002, pp. 91-92, 2002.

[109] A. P. Chandrakasan and R. W. Brodersen, "Minimizing Power Consumption in Digital CMOS Circuits," *Proceedings of the IEEE*, vol. 83, pp. 498-523, 1995.

[110] D. Munteanu and J. L. Autran, "Two-Dimensional Modeling of Quantum Ballistic Transport in Ultimate Double-Gate SOI Devices," *Solid-State Electronics*, vol. 47, pp. 1219-1225, 2003.

[111] T. Ernst, S. Cristoloveanu, G. Ghibaudo, T. Ouisse, S. Horiguchi, Y. Ono, Y. Takahashi, and K. Murase, "Ultimately Thin Double-Gate SOI MOSFETs," *IEEE Transactions on Electron Devices*, vol. 50, pp. 830-838, 2003.

[112] S.-W. Chung, J.-Y. Yu, and J. R. Heath, "Silicon Nanowire Devices," *Applied Physics Letters*, vol. 76, pp. 2068-2070, 2000.

[113] T. Claasen. (1998) The Logarithmic Law of Usefulness. *Semiconductor International,* [Online]. Available: http://www.reed-electronics.com/semiconductor/article/CA164016?text=claasen.

[114] G. E. Moore, "No Exponential is Forever: but "Forever" can be Delayed!," presented at IEEE International Solid-State Circuits Conference, ISSCC 2003, pp. 20-23 vol.1, 2003.

[115] H. Bar-Lev, P. W. F. Rutten, A. M. I. Sonnino, and M. Tauman, "The Future of Semiconductor Technology," in *Kellogg on Technology & Innovation 2002*, M. Sawhney, R. Gulati, and A. Paoni, Eds.: Wiley, 2002.

[116] H. Zhang, V. Prabhu, V. George, M. Wan, M. Benes, A. Abnous, and J. M. Rabaey, "A 1 V Heterogeneous Reconfigurable Processor IC for Baseband Wireless Applications," presented at IEEE International Solid-State Circuits Conference, ISSCC 2000, pp. 68 -69, 448, 2000.

[117] V. Betz and J. Rose, "How Much Logic Should Go in an FPGA Logic Block?," *IEEE Design & Test of Computers*, vol. 15, pp. 10-15, 1998.

[118] A. Takahara, T. Miyazaki, T. Murooka, M. Katayama, K. Hayashi, A. Tsutsui, T. Ichimori, and K.-n. Fukami, "More Wires and Fewer LUTs: A Design Methodology for FPGAs," presented at ACM/SIGDA Sixth International Symposium on Field programmable Gate Arrays, pp. 12 - 19, 1998.

[119] F. de Dinechin, "The Price of Routing in FPGAs," *Journal of Universal Computer Science*, vol. 6, pp. 227-239, 2000.

[120] S. Hauck, "The Future of Reconfigurable Systems," presented at 5th Canadian Conference on Field Programmable Devices, Montreal, 1998.

[121] R. W. Hartenstein, "Coarse Grain Reconfigurable Architectures," presented at Proceedings of the ASP-DAC 2001 Design Automation Conference, pp. 564 -569, 2001.

[122] R. Hartenstein, "The Microprocessor is No Longer General Purpose: Why Future Reconfigurable Platforms Will Win," presented at Second Annual IEEE International Conference on Innovative Systems in Silicon, IEEE, pp. 2 -12, 1996.

[123] K. Eguro and S. Hauck, "Issues and Approaches to Coarse-Grain Reconfigurable Architecture Development," presented at 11th Annual IEEE Symposium on Field-Programmable Custom Computing Machines, FCCM 2003., pp. 111-120, 2003.

[124] S. Brown and J. Rose, "Architecture of FPGAs and CPLDs: A Tutorial," *IEEE Design and Test of Computers*, vol. 13, pp. 42-57, 1996.

[125] T. Jamil, "RISC versus CISC," *IEEE Potentials*, vol. 14, pp. 13-16, 1995.

[126] R. Hartenstein, M. Herz, T. Hoffmann, and U. Nageldinger, "Mapping Applications onto Reconfigurable KressArrays," presented at 9th International Workshop on Field Programmable Logic and Applications, FPL '99, Glasgow, UK, 1999.

[127] T. J. Callahan and J. Wawrzynek, "Adapting Software Pipelining for Reconfigurable Computing," presented at International Conference on Compilers, Architecture, and Synthesis for Embedded Systems, CASES, 2000.

[128] C. Ebeling, D. C. Cronquist, and P. Franklin, "RaPiD Reconfigurable Pipelined Datapath," presented at Field-Programmable Logic: Smart Applications, New Paradigms, and Compilers. 6th International Workshop on Field-Programmable Logic and Applications, 1996.

[129] H. Singh, G. Lu, E. Filho, R. Maestre, M.-H. Lee, F. Kurdahi, and N. Bagherzadeh, "MorphoSys: Case Study of a Reconfigurable Computing System Targeting Multimedia Applications," presented at Proceedings of the 37th Conference on Design Automation, Los Angeles, CA USA, pp. 573 - 578, 2000.

[130] Xilinx. [Online]. Available: http://www.xilinx.com.

[131] Actel. [Online]. Available: http://www.actel.com/.

[132] T. Makimoto, "(Keynote) Towards the Second Digital Wave. The Future of the Semiconductor Business as Predicted by "Makimoto's Wave"," presented at IEEE Field Programmable Technologies Conference, Hong Kong, 2002.

[133] O. Agrawal, H. Chang, B. Sharpe-Geisler, N. Schmitz, B. Nguyen, J. Wong, G. Tran, F. Fontana, and W. Harding, "An Innovative, Segmented High Performance FPGA Family with Variable-Grain-Architecture and Wide-Gating Functions," presented at ACM/SIGDA Seventh International Symposium on Field Programmable Gate Arrays, pp. 17 - 26, 1999.

[134] A. Singh, A. Mukherjee, and M. Marek-Sadowska, "Interconnect Pipelining in a Throughput-Intensive FPGA Architecture," presented at Ninth International Symposium on Field Programmable Gate Arrays, Monterey, CA, pp. 153 - 160, 2001.

[135] P. D. Singh and S. D. Brown, "The Case for Registered Routing Switches in Field Programmable Gate Arrays," presented at Ninth International Symposium on Field Programmable Gate Arrays, Monterey, CA, pp. 161 - 169, 2001.

[136] A. DeHon, "Entropy, Counting, and Programmable Interconnect," presented at ACM/SIGDA International Symposium on Field Programmable Gate Arrays (FPGA'96), pp. 73 - 79, 1996.

[137] B. Zahiri, "Structured ASICs: Opportunities and Challenges," presented at 21st International Conference on Computer Design, pp. 404-409, 2003.

[138] Structured ASIC Association. [Online]. Available: http://www.lightspeed.com/.

[139] Altera Inc. [Online]. Available: http://www.altera.com.

[140] R. Ronen, A. Mendelson, K. Lai, S.-L. Lu, F. Pollack, and J. P. Shen, "Coming Challenges in Microarchitecture and Architecture," *Proceedings of the IEEE*, vol. 98, pp. 325 - 340, 2001.

[141] http://www.systemc.org.

[142] http://www.systemverilog.org/.

[143] R. W. Hartenstein, "A Decade of Reconfigurable Computing: A Visionary Retrospective," presented at Design, Automation and Test in Europe Conference and Exhibition, pp. 642 - 649, 2001.

[144] C. Compton and S. Hauck, "An Introduction to Reconfigurable Computing," *IEEE Computer*, 2000.

[145] J. E. Carrillo and P. Chow, "The Effect of Reconfigurable Units in Superscalar Processors," presented at Ninth International Symposium on Field Programmable Gate Arrays, FPGA 2001, Monterey, CA, USA, pp. 141 - 150, 2001.

[146] J. R. Hauser and J. Wawrzynek, "Garp: A MIPS Processor with a Reconfigurable Coprocessor," presented at IEEE Symposium on FPGAs for Custom Computing Machines, FCCM'97, pp. 12-21, 1997.

[147] S. Hauck, T. W. Fry, M. M. Hosler, and J. P. Kao, "The Chimaera Reconfigurable Functional Unit," presented at 5th Annual IEEE Symposium on Field-Programmable Custom Computing Machines, FCCM'97, pp. 87 - 96, 1997.

[148] A. Shibayama, H. Igura, M. Mizuno, and M. Yamashina, "An Autonomous Reconfigurable Cell Array for Fault-Tolerant LSIs," presented at IEEE International Solid State Circuits Conference, ISSCC97, pp. 230 - 231, 462, 1997.

[149] M. R. B. Forshaw, K. Nikolic, and A. Sadek, "3rd Annual Report, Autonomous Nanoelectronic Systems With Extended Replication and Signalling, ANSWERS," University College London, Image Processing Group, London, U. K., Technical Report July 2000-July 2001 2001.

[150] V. Parihar, R. Singh, and K. F. Poole, "Silicon Nanoelectronics: 100nm Barriers and Potential Solutions," presented at IEEE/SEMI Advanced Semiconductor Manufacturing Conference, pp. 427-433, 1998.

[151] S. C. Goldstein, H. Schmit, M. Budiu, S. Cadambi, M. Moe, and R. R. Taylor, "PipeRench: A Reconfigurable Architecture and Compiler," *IEEE Computer*, vol. 33, pp. 70-77, 2000.

[152] I. Koren and Z. Koren, "Defect Tolerance in VLSI Circuits: Techniques and Yield Analysis," *Proceedings of the IEEE*, vol. 86, pp. 1819 - 1836, 1998.

[153] D. Clark, "Teramac: Pointing the Way to Real-World Nanotechnology," *IEEE Computational Science and Engineering*, vol. 5, pp. 88 -90, 1998.

[154] J. R. Heath, P. J. Kuekes, G. S. Snider, and S. Williams, "A Defect-Tolerant Computer Architecture: Opportunities for Nanotechnology," *Science*, vol. 280, pp. 1716-21, 1998.

[155] S. C. Goldstein, "Electronic Nanotechnology and Reconfigurable Computing," presented at IEEE Computer Society Workshop on VLSI, Orlando, Florida, pp. 10-15, 2001.

[156] N. J. Macias, "The PIG Paradigm: The Design and Use of a Massively Parallel Fine Grained Self-Reconfigurable Infinitely Scalable Architecture," presented at First NASA/DoD Workshop on Evolvable Hardware, 1999.

[157] M. G. Gericota, G. R. Alves, M. L. Silva, and J. M. Ferreira, "DRAFT: An On-line Fault Detection Method for Dynamic & Partially Reconfigurable FPGAs," presented at Seventh International On-Line Testing Workshop, pp. 34 -36, 2001.

[158] D. Mange, M. Sipper, A. Stauffer, and G. Tempesti, "Toward Robust Integrated Circuits: The Embryonics Approach," *Proceedings of the IEEE*, vol. 88, pp. 516-543, 2000.

[159] G. I. Bourianoff, "The Future of Nanocomputing," *IEEE Computer*, vol. 36, pp. 44-53, 2003.

[160] M. Kanellos. (2003) Soaring Costs of Chipmaking Recast Industry. [Online]. Available: http://news.com.com/Semi+survival/2009-1001_3-981418.html.

[161] P. Mead. (2004) A Low-Cost, Low-Risk Alternative To ASIC And ASSPs For Communication Systems. [Online]. Available: http://www.epn-online.com/page/12761/a-low-cost--low-risk-alternative-to-asic-and-assps-for-communication-systems-the-real-cost-of-asics.html.

[162] H. Stork. (2005) Economies of CMOS Scaling. [Online]. Available: http://www.eeel.nist.gov/812/conference/2005_Talks/Stork.pdf.

[163] F. M. Schellenberg and L. Capodieci, "Impact of RET on Physical Layouts," presented at International Symposium on Physical Design, Sonoma, California, United States, pp. 52 - 55, 2001.

[164] L. W. Liebmann, "Layout Impact of Resolution Enhancement Techniques: Impediment or Opportunity?," presented at International Symposium on Physical Design, Monterey, CA, USA, pp. 110 - 117, 2003.

[165] V. Singh, "The Importance of Layout Density Control in Semiconductor Manufacturing," presented at Electronic Design Processes, Monterey Beach Hotel, Monterey, CA, 2003.

[166] P. Gupta and A. B. Kahng, "Manufacturing-Aware Physical Design," presented at International Conference on Computer Aided Design, ICCAD-2003, pp. 681-687, 2003.

[167] H. K.-S. Leung, "Advanced Routing in Changing Technology Landscape," presented at International Symposium on Physical Design, Monterey, CA, USA, pp. 118 - 121, 2003.

[168] L. Capodieci, P. Gupta, A. B. Kahng, D. Sylvester, and J. Yang, "Toward a Methodology for Manufacturability-Driven Design Rule Exploration," presented at 41st Annual Conference on Design Automation, San Diego, CA, USA, pp. 311 - 316, 2004.

[169] M. J. Flynn and P. Hung, "Microprocessor Design Issues: Thoughts on the Road Ahead," *IEEE Micro*, vol. 25, pp. 16-31, 2005.

[170] M. J. Flynn, "Area – Time – Power and Design Effort: the Basic Tradeoffs in Application Specific Systems," presented at 16th International Conference on Application-Specific Systems, Architecture and Processors (ASAP'05), pp. 3-5, 2005.

[171] N. P. Jouppi, "(Keynote) The Future Evolution of High-Performance Microprocessors," presented at 38th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO'05), pp. 155, 2005.

[172] D. J. Frank, R. H. Dennard, E. Nowak, P. M. Solomon, Y. Taur, and H.-S. P. Wong, "Device Scaling Limits of Si MOSFETs and their Application Dependencies," *Proceedings of the IEEE*, vol. 89, pp. 259-288, 2001.

[173] D. Harris, R. F. Sproull, and I. E. Sutherland, *Logical Effort: Designing Fast CMOS Circuits*: Morgan Kaufmann Publishers, 1999.

[174] V. Zyuban, "Unified Architecture Level Energy-Efficiency Metric," presented at 12th ACM Great Lakes Symposium on VLSI, New York, USA, pp. 24 - 29, 2002.

[175] H. J. M. Veendrick, "Short-Circuit Dissipation of Static CMOS Circuitry and its Impact on the Design of Buffer Circuits," *IEEE Journal of Solid-State Circuits*, vol. 19, pp. 468-473, 1984.

[176] Y.-C. Yeo, T.-J. King, and C. Hu, "Direct Tunneling Leakage Current and Scalability of Alternative Gate Dielectrics," *Applied Physics Letters*, vol. 81, pp. 2091 - 2093, 2002.

[177] H. Kawaura and T. Baba, "Direct Tunneling from Source to Drain in Nanometer-Scale Silicon Transistors," *Japanese Journal of Applied Physics*, vol. 42, Part 1, pp. 351-357, 2003.

[178] N. J. Collier and J. R. A. Cleaver, "Novel Dual-Gate HEMT Utilising Multiple Split Gates," *Microelectronic Engineering*, vol. 41-42, pp. 457-460, 1998.

[179] M. Ieong, E. C. Jones, T. Kanarsky, Z. Ren, O. Dokumaci, R. A. Roy, L. Shi, T. Furukawa, Y. Taur, R. J. Miller, and H.-S. P. Wong, "Experimental Evaluation of Carrier Transport and Device Design for Planar Symmetric/Asymmetric Double-Gate/Ground-Plane CMOSFETs," presented at International Electron Devices Meeting, IEDM 2001, Washington, DC, pp. 6.1 - 6.4, 2001.

[180] I. Y. Yang, C. Vieri, A. Chandrakasan, and D. A. Antoniadis, "Back-Gated CMOS on SOIAS for Dynamic Threshold Voltage Control," *IEEE Transactions on Electron Devices*, vol. 44, pp. 822-831, 1997.

[181] U. Avci and S. Tiwari, "Back-Gated MOSFETs with Controlled Silicon Thickness for Adaptive Threshold-Voltage Control," *Electronics Letters*, vol. 40, pp. 74-75, 2004.

[182] J. Kao, S. Narendra, and A. Chandrakasan, "Subthreshold Leakage Modeling and Reduction Techniques," presented at 2002 IEEE/ACM International Conference on Computer-aided Design, San Jose, California, pp. 141 - 148, 2002.

[183] H. Kawaguchi, K. Nose, and T. Sakurai, "A Super Cut-Off CMOS (SCCMOS) Scheme for 0.5-V Supply Voltage with Picoampere Stand-By Current," *IEEE Journal of Solid-State Circuits*, vol. 35, pp. 1498-1501, 2000.

[184] B. H. Calhoun, F. A. Honore, and A. Chandrakasan, "Design Methodology for Fine-Grained Leakage Control in MTCMOS," presented at International Symposium on Low Power Electronics and Design, Seoul, Korea, pp. 104--109, 2003.

[185] T. Kuroda and M. Hamada, "Low-power CMOS Digital Design with Dual Embedded Adaptive Power Supplies," *IEEE Journal of Solid-State Circuits*, vol. 35, pp. 652-655, 2000.

[186] T. D. Burd, T. A. Pering, A. J. Stratakos, and R. W. Brodersen, "A Dynamic Voltage Scaled Microprocessor System," *IEEE Journal of Solid-State Circuits*, vol. 35, pp. 1571-1580, 2000.

[187] L. Yuan and G. Qu, "Analysis of Energy Reduction on Dynamic Voltage Scaling-Enabled Systems," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 24, pp. 1827-1837, 2005.

[188] T. Sherwood, E. Perelman, G. Hamerly, and B. Calder, "Automatically Characterizing Large Scale Program Behavior," presented at 10th International Conference on Architec-

tural Support for Programming Languages and Operating Systems, ASPLOS'02, San Jose, California, pp. 45 - 57, 2002.

[189] T. Kuroda, T. Fujita, S. Mita, T. Nagamatu, S. Yoshioka, F. Sano, M. Norishima, M. Murota, M. Kako, M. Kinugawa, M. Kakumu, and T. Sakurai, "A 0.9 V 150 MHz 10 mW 4 mm^2 2-D Discrete Cosine Transform Core Processor with Variable-Threshold-Voltage Scheme," presented at 43rd IEEE International Solid-State Circuits Conference, ISSCC 1996, pp. 166-167, 437, 1996.

[190] T. Ohtou, T. Nagumo, and T. Hiramoto, "Short-Channel Characteristics of Variable-Body-Factor Fully-Depleted Silicon-On-Insulator Metal-Oxide-Semiconductor-Field-Effect-Transistors," *Japanese Journal of Applied Physics*, vol. 44, pp. 3885-3888, 2005.

[191] W. C. Athas, L. J. Svensson, J. G. Koller, N. Tzartzanis, and E. Ying-Chin Chou, "Low-power Digital Systems based on Adiabatic-Switching Principles," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 2, pp. 398-407, 1994.

[192] R. C. Merkle, "Reversible Electronic Logic Using Switches," *Nanotechnology*, vol. 4, pp. 21-40, 1993.

[193] R. K. Cavin III, V. V. Zhirnov, J. A. Hutchby, and G. I. Bourianoff, "Energy Barriers, Demons, and Minimum Energy Operation of Electronic Devices (Plenary Paper)," presented at Noise in Devices and Circuits III, Austin, TX, USA, pp. 1-9, 2005.

[194] R. Landauer, "Uncertainty Principle and Minimal Energy Dissipation in the Computer," *International Journal of Theoretical Physics*, vol. 21, pp. 283-297, 1982.

[195] W. R. Frensley, "Gain In Nanoelectronic Devices," presented at Workshop on Physics and Computation, PhysComp '92, pp. 258 - 261, 1992.

[196] B. Voss and M. Glesner, "Adiabatic Charging of Long Interconnects," presented at 7th IEEE International Conference on Electronics, Circuits and Systems, ICECS2000, pp. 835-838 vol.2, 2000.

[197] R. C. Merkle, "Two Types of Mechanical Reversible Logic," *Nanotechnology*, vol. 4, pp. 114-131, 1993.

[198] J. S. Hall, "Nanocomputers and Reversible Logic," *Nanotechnology*, vol. 5, pp. 157-167, 1994.

[199] W. C. Athas and L. J. Svensson, "Reversible Logic Issues in Adiabatic CMOS," presented at Workshop on Physics and Computation, PhysComp '94, Dallas, TX, pp. 111-118, 1994.

[200] A. De Vos, "Reversible computing," *Progress in Quantum Electronics*, vol. 23, pp. 1-49, 1999.

[201] V. Zyuban and P. Kogge, "Optimization of High-Performance Superscalar Architectures for Energy Efficiency," presented at International Symposium on Low Power Electronics and Design, ISLPED '00, pp. 84-89, 2000.

[202] P. Penzes, "Energy-Delay Complexity of Asynchronous Circuits," California Institute of Technology, Technical Report CaltechCSTR:2002.010, 26 September 2002, http://resolver.caltech.edu/CaltechCSTR:2002.010.

[203] A. J. Martin, "Towards an Energy Complexity of Computation," *Information Processing Letters*, vol. 77, pp. 181-187, 2001.

[204] D. Brooks, P. Bose, V. Srinivasan, M. K. Gschwind, P. G. Emma, and M. G. Rosenfield, "New Methodology for Early-Stage, Microarchitecture-Level Power–Performance Analysis of Microprocessors," *IBM Journal of Research and Development*, vol. 47, pp. 653-670, 2003.

[205] D. M. Brooks, P. Bose, S. E. Schuster, H. Jacobson, P. N. Kudva, A. Buyuktosunoglu, J. Wellman, V. Zyuban, M. Gupta, and P. W. Cook, "Power-Aware Microarchitecture: Design and Modeling Challenges for Next-Generation Microprocessors," *IEEE Micro*, vol. 20, pp. 26-44, 2000.

[206] V. Zyuban and P. Strenski, "Unified Methodology for Resolving Power-Performance Tradeoffs at the Microarchitectural and Circuit Levels," presented at International Symposium on Low Power Electronics and Design, ISLPED '02, pp. 166-171, 2002.

[207] V. Zyuban and P. N. Strenski, "Balancing Hardware Intensity in Microprocessor Pipelines," *IBM Journal of Research & Development*, vol. 47, pp. 585-598, 2003.

[208] H. Q. Dao, B. R. Zeydel, and V. G. Oklobdzija, "Energy Minimization Method for Optimal Energy-Delay Extraction," presented at European Solid-State Circuits Conference, Estoril, Portugal, pp. 177-180, 2003.

[209] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, "Low-Power CMOS Digital Design," *IEEE Journal of Solid-State Circuits*, vol. 27, pp. 473-484, 1992.

[210] A. J. Bhavnagarwala, B. L. Austin, K. A. Bowman, and J. D. Meindl, "A Minimum Total Power Methodology for Projecting Limits on CMOS GSI," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, pp. 235-251, 2000.

[211] H. Q. Dao, B. R. Zeydel, and V. O. Oklobdzija, "Architectural Considerations for Energy Efficiency," presented at International Conference on Computer Design, pp. 13-16, 2005.

[212] S. Das, G. Rose, M. M. Ziegler, C. A. Picconatto, and J. C. Ellenbogen, "Architectures and Simulations for Nanoprocessor Systems Integrated on the Molecular Scale," in *Introducing Molecular Electronics*: Springer LNiP, 2004.

[213] M. Montemerlo, C. Love, G. Opiteck, D. Goldhaber-Gordon, and J. Ellenbogen. (1996) Technologies and Designs for Electronic Nanocomputers. [Online]. Available: http://www.mitre.org/tech/nanotech/index.html.

[214] L. J. K. Durbeck and N. J. Macias. (2000) The Cell Matrix: An Architecture for Nanocomputing. [Online]. Available: http://www.cellmatrix.com/entryway/products/pub/#papers.

[215] C. E. Kozyrakis and D. A. Patterson, "A New Direction for Computer Architecture Research," *IEEE Computer*, vol. 31, pp. 24 -32, 1998.

[216] S. Vajapeyam and M. Valero, "Early 21st Century Processors," *IEEE Computer*, vol. 34, pp. 47-50, 2001.

[217] T. J. Fountain, M. J. B. D. Duff, D. G. Crawley, C. Tomlinson, and C. Moffat, "The Use of Nanoelectronic Devices in Highly-Parallel Computing Systems," *IEEE Transactions on VLSI Systems*, vol. 6, pp. 31-38, 1998.

[218] G. Bilardi and F. P. Preparata, "Horizons of Parallel Computation," *Journal of Parallel and Distributed Computing*, vol. 27, pp. 172-182, 1995.

[219] E. S. Gayles, T. P. Kelliher, R. M. Owens, and M. J. Irwin, "The Design of the MGAP-2: A Micro-Grained Massively Parallel Array," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 8, pp. 709 - 716, 2000.

[220] S. Hauck, "Asynchronous Design Methodologies: An Overview," *Proceedings of the IEEE*, vol. 83, pp. 69 - 93, 1995.

[221] D. Crawley, "An Analysis of MIMD Processor Node Designs for Nanoelectronic Systems," Image Processing Group, Department of Physics & Astronomy, University College, London, Internal Report 97/3, 1997.

[222] T. J. Fountain, "The Propagated Instruction Processor," presented at Workshop on Innovative Circuits and Systems for Nanoelectronics, Delft, pp. 69-74, 1997.

[223] D. Patterson, T. Anderson, N. Cardwell, R. Fromm, K. Keeton, C. Kozyrakis, R. Thomas, and K. Yelick, "Intelligent RAM (IRAM): Chips that Remember and Compute," presented at 43rd IEEE International Solid-State Circuits Conference, ISSCC 1997, pp. 224 - 225, 1997.

[224] E. Waingold, M. Tayor, D. Srikrishna, V. Sarkar, W. Lee, V. Lee, J. Kim, M. Frank, P. Finch, R. Barua, J. Babb, S. Amarasinghe, and A. Agarwal, "Baring It All to Software: Raw Machines," *IEEE Computer*, vol. 30, pp. 83 - 96, 1997.

[225] U. J. Kapasi, S. Rixner, W. J. Dally, B. Khailany, J. H. Ahn, P. Mattson, and J. D. Owens, "Programmable Stream Processors," *Computer*, vol. 36, pp. 54-62, 2003.

[226] B. Khailany, W. J. Dally, S. Rixner, U. J. Kapasi, J. D. Owens, and B. Towles, "Exploring the VLSI Scalability of Stream Processors," presented at The Ninth International Symposium on High-Performance Computer Architecture, HPCA-9, pp. 153-164, 2003.

[227] S. Rajagopal, J. R. Cavallaro, and S. Rixner, "Design Space Exploration for Real-Time Embedded Stream Processors," *IEEE Micro*, vol. 24, pp. 54-66, 2004.

[228] S. C. Goldstein and M. Budiu, "NanoFabrics: Spatial Computing Using Molecular Electronics," presented at 28th International Symposium on Computer Architecture, Goteborg, Sweden, pp. 178 - 189, 2001.

[229] A. DeHon, "Very Large Scale Spatial Computing," presented at Third International Conference on Unconventional Models of Computation, UMC'02, 2002.

[230] M. Budiu, G. Venkataramani, T. Chelcea, and S. C. Goldstein, "Spatial Computation," presented at International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS'04, Boston, MA, 2004.

[231] P. Beckett, "Low-Power Spatial Computing using Dynamic Threshold Devices," presented at International Symposium on Circuits and Systems, ISCAS'05, Kobe, Japan, pp. 2345-2348, 2005.

[232] Y. Feldman and E. Shapiro, "Spatial Machines: a More Realistic Approach to Parallel Computation," *Communications of the ACM*, vol. 35, pp. 60 - 73, 1992.

[233] A. DeHon, "Trends Toward Spatial Computing Architectures," *IEEE International Solid-State Circuits Conference, ISSCC'99*, pp. 362 - 363, 1999.

[234] H. Lee, P. Beckett, and B. Appelbe, "High-Performance Extendable Instruction Set Computing," presented at 6th Australasian Computer Systems Architecture Conference, ACSAC 2001, Gold Coast, Qld., Australia, pp. 89 -94, 2001.

[235] W. Mangione-Smith, "Configurable Computing: Concepts and Issues," presented at Thirtieth Hawaii International Conference on System Sciences, Wailea, HI, USA, pp. 710 -712 vol.1, 1997.

[236] N. Wirth, "Hardware Compilation: Translating Programs into Circuits," *Computer*, vol. 31, pp. 25-31, 1998.

[237] W. R. Davis, "Getting High-Performance Silicon from System-Level Design," presented at IEEE Computer Society Annual Symposium on VLSI, ISVLSI'03, pp. 238 -243, 2003.

[238] T. Isshiki and W. Wei-Ming Dai, "Bit-Serial Pipeline Synthesis for Multi-FPGA Systems with C++ Design Capture," presented at IEEE Symposium on FPGAs for Custom Computing Machines, pp. 38 - 47, 1996.

[239] A. Ghosh, J. Kunkel, and S. Liao, "Hardware synthesis from C/C++," presented at Design, Automation and Test in Europe, DATE'99, Munich, Germany, pp. 82-84, 1999.

[240] X. Zhu and B. Lin, "Hardware Compilation for FPGA-Based Configurable Computing Machines," presented at 36th ACM/IEEE conference on Design automation conference, New Orleans, Louisiana, United States, pp. 697 - 702, 1999.

[241] D. Sulik, M. Vasilko, D. Durackova, and P. Fuchs, "Design of a RISC Microcontroller Core in 48 Hours," presented at Embedded Systems Show 2000, London Olympia, 2000.

[242] K. Okada, A. Yamada, and T. Kambe, "Hardware Algorithm Optimization Using Bach C," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. E85-A, pp. 835-841, 2002.

[243] A. DeHon and M. J. Wilson, "Nanowire-Based Sublithographic Programmable Logic Arrays," presented at ACM/SIGDA 12th International Symposium on Field-Programmable Gate Arrays, FPGA2004, pp. 123 - 132, 2004.

[244] A. DeHon, "Nanowire-Based Programmable Architectures," *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, vol. 1, pp. 109 - 162, 2005.

[245] A. DeHon, "Array-Based Architecture for FET-Based, Nanoscale Electronics," *IEEE Transactions on Nanotechnology*, vol. 2, pp. 23-32, 2003.

[246] A. DeHon, C. M. Lieber, P. Lincoln, and J. E. Savage, "Sub-lithographic Semiconductor Computing Systems," presented at HotChips 15, Stanford University, 2003.

[247] M. Budiu, Spatial Computation, PhD Thesis, School of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 2003.

[248] K. J. Nowka and M. J. Flynn, "System Design using Wave-Pipelining: a CMOS VLSI Vector Unit," presented at IEEE International Symposium on Circuits and Systems, ISCAS '95, pp. 2301-2304 vol.3, 1995.

[249] F. Klass, M. J. Flynn, and A. J. van de Goor, "A 16×16-bit Static CMOS Wave-Pipelined Multiplier," presented at IEEE International Symposium on Circuits and Systems, ISCAS '94, London, U. K., pp. 143 - 146 vol.4, 1994.

[250] O. Hauck, A. Katoch, and S. A. Huss, "VLSI System Design Using Asynchronous Wave Pipelines: A 0.35 mu CMOS 1.5GHz Elliptic Curve Public Key Cryptosystem Chip," presented at Sixth International Symposium on Advanced Research in Asynchronous Circuits and Systems, (ASYNC 2000), pp. 188 -197, 2000.

[251] J. B. Dennis, "A Preliminary Architecture for a Basic Dataflow Processor," *Proceedings of the 2nd Annual Symposium on Computer Architecture*, pp. 126-132, 1975.

[252] J. B. Dennis, "Data Flow Supercomputers," *Computer*, vol. 13, pp. 48-56, 1980.

[253] S. Swanson, K. Michelson, A. Schwerin, and M. Oskin, "Dataflow: The Road Less Complex," presented at Workshop on Complexity-Effective Design, San Diego, California, 2003.

[254] S. Swanson, K. Michelson, A. Schwerin, and M. Oskin, "WaveScalar," presented at 36th International Symposium on Microarchitecture, MICRO-36, 2003.

[255] L. Geppert, "Sun's Big Splash [Niagara microprocessor chip]," *Spectrum, IEEE*, vol. 42, pp. 56-60, 2005.

[256] P. Beckett, "Low-Power Circuits using Dynamic Threshold Devices," presented at Great Lakes Symposium on VLSI, Chicago, Il, pp. 213-216, 2005.

[257] P. Beckett, "A Nanowire Array for Reconfigurable Computing," presented at 2005 IEEE Region 10 Conference, TENCON 2005, Melbourne, Australia, 2005.

[258] K. W. Guarini, P. M. Solomon, Y. Zhang, K. K. Chan, E. C. Jones, G. M. Cohen, A. Krasnoperova, M. Ronay, O. Dokumaci, J. J. Bucchignano, C. Cabral Jr., C. Lavoie, V. Ku, D. C. Boyd, K. S. Petrarca, I. V. Babich, J. Treichler, and P. M. Kozlowski, "Triple-Self-Aligned, Planar Double-Gate MOSFETs: Devices and Circuits," presented at International New Electron Devices Meeting, Washington, DC, pp. 19.2.1 -19.2.4, 2001.

[259] L. S. Y. Wong and G. A. Rigby, "A 1V CMOS Digital Circuits with Double-Gate-Driven MOSFET," presented at IEEE International Solid-State Circuits Conference, pp. 292 - 293, 1997.

[260] K. Roy, H. Mahmoodi, S. Mukhopadhyay, H. Ananthan, A. Bansal, and T. Cakici, "Double-Gate SOI Devices for Low-Power and High-Performance Applications," presented at 19th International Conference on VLSI Design, pp. 8 pp., 2006.

[261] S. Mukhopadhyay, H. Mahmoodi, and K. Roy, "A Novel High-Performance and Robust Sense Amplifier using Independent Gate Control in Sub-50-nm Double-Gate MOSFET," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 14, pp. 183-192, 2006.

[262] S. Hasan, J. Wang, and M. Lundstrom. (2000) A Well-Tempered 10nm scale Double-Gate N-MOSFET, Rev. 10-28-02. [Online]. Available: falcon.ecn.purdue.edu:8080/mosfet/10nmstructure.pdf.

[263] T. Schulz, W. Rösner, E. Landgraf, L. Risch, and U. Langmann, "Planar and Vertical Double Gate Concepts," *Solid-State Electronics*, vol. 46, pp. 985-989, 2002.

[264] L. Chang, "Scaling Limits and Design Considerations for Double-Gate MOSFETs," Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, Research Project 5 October 2001, http://hkn.eecs.berkeley.edu/~leland/ research.html.

[265] The Nanotechnology Simulation Hub. [Online]. Available: http://www.nanohub.org/.

[266] H.-S. P. Wong, D. J. Frank, and P. M. Solomon, "Device Design Considerations for Double-Gate, Ground-Plane, and Single-Gated Ultra-Thin SOI MOSFET's at the 25 nm Channel Length Generation," presented at International Electron Devices Meeting, IEDM '98, San Francisco, CA, USA, pp. 407 -410, 1998.

[267] H.-S. Wong, M. H. White, T. J. Krutsick, and R. V. Booth, "Modeling of Transconductance Degradation and Extraction of Threshold Voltage in Thin Oxide MOSFET's," *Solid-State Electronics*, vol. 30, pp. 953-968, 1987.

[268] A. Ortiz-Conde, E. D. G. Fernandes, J. J. Liou, M. R. Hassan, F. J. Garcýa-Sanchez, G. D. Mercato, and W. Wong, "A New Approach to Extract the Threshold Voltage of MOSFET's," *IEEE Transactions on Electron Devices*, vol. 44, pp. 1523-1528, 1997.

[269] X. Zhou, K. Y. Lim, and W. Qian, "Threshold Voltage Definition and Extraction for Deep-Submicron MOSFETs," *Solid-State Electronics*, vol. 45, pp. 507-510, 2001.

[270] W. Y. Choi, B. L. Hwi Kim, J. D. Lee, and B.-G. Park, "Stable Threshold Voltage Extraction Using Tikhonov's Regularization Theory," *IEEE Transactions on Electron Devices*, vol. 51, pp. 1833-1839, 2004.

[271] P. Francis, A. Terao, D. Flandre, and F. Van de Wiele, "Modeling of Ultrathin Double-Gate nMOS/SOI Transistors," *IEEE Transactions on Electron Devices*, vol. 41, pp. 715 - 720, 1994.

[272] Y. Taur, "An Analytical Solution to a Double-gate MOSFET with Undoped Body," *IEEE Electron Device Letters*, vol. 21, pp. 245 - 247, 2000.

[273] S. Ahmed, C. Ringhofer, and D. Vasileska, "An Effective Potential Approach to Modeling 25 nm MOSFET Devices," *Journal of Computational Electronics*, vol. 2, pp. 113-117, 2003.

[274] R.-C. Chen and J.-L. Liu, "A Quantum Corrected Energy-Transport Model for Nanoscale Semiconductor Devices," *Journal of Computational Physics*, vol. 204, pp. 131 - 156, 2005.

[275] D. J. Wouters, J.-P. Colinge, and H. E. Maes, "Subthreshold Slope in Thin-Film SOI MOSFETs," *IEEE Transactions on Electron Devices*, vol. 37, pp. 2022-2033, 1990.

[276] S. Zhu, J. Chen, M.-F. Li, S. J. Lee, J. Singh, C. X. Zhu, A. Du, C. H. Tung, A. Chin, and D. L. Kwong, "N-type Schottky Barrier Source/Drain MOSFET using Ytterbium Silicide," *IEEE Electron Device Letters*, vol. 25, pp. 565-567, 2004.

[277] J. Kedzierski, P. Xuan, V. Subramanian, J. Bokor, T.-J. King, C. Hu, and E. Anderson, "A 20 nm Gate-Length Ultra-Thin Body p-MOSFET with Silicide Source/Drain," *Superlattices and Microstructures*, vol. 28, pp. 445-452, 2000.

[278] J. Kedzierski, E. Nowak, T. Kanarsky, Y. Zhang, D. Boyd, R. Carruthers, C. Cabral, R. Amos, C. Lavoie, R. Roy, J. Newbury, E. Sullivan, J. Benedict, P. Saunders, K. Wong, D. Canaperi, M. Krishnan, K.-L. Lee, B. A. Rainey, D. Fried, P. Cottrell, H.-S. P. Wong, M. Ieong, and W. Haensch, "Metal-Gate FinFET and Fully-Depleted SOI Devices using Total Gate Silicidation," presented at International Electron Devices Meeting, IEDM '02, pp. 247-250, 2002.

[279] Q. Chen, L. Wang, and J. D. Meindl, "Impact of High-K Dielectrics on Undoped Double-Gate MOSFET Scaling," presented at IEEE International SOI Conference, pp. 115-116, 2002.

[280] J. G. Fossum, "A Unified Process-Based Compact Model for Scaled PD/SOI and Bulk-Si MOSFETs," presented at International Conference on Modeling and Simulation of Microsystems, NanoTech 2002 - MSM 2002, San Juan, Puerto Rico, 2002.

[281] Y.-K. Choi, D. Ha, T.-J. King, and J. Bokor, "Investigation of Gate-Induced Drain Leakage (GIDL) Current in Thin Body Devices: Single-Gate Ultra-Thin Body, Symmetrical Double-Gate, and Asymmetrical Double-Gate MOSFETs," *Japanese Journal Applied Physics*, vol. 42, Part 1, pp. 2073-2076, 2003.

[282] N. H. E. Weste and K. Eshraghian, *Principles of CMOS VLSI Design: A System Perspective*, 2nd ed: Reading MA: Addison-Wesley, 1993.

[283] S.-M. Kang and Y. Leblebici, *CMOS Digital Integrated Circuits: Analysis and Design*: McGraw-Hill, 1996.

[284] J. Rose, R. J. Francis, D. Lewis, and P. Chow, "Architectures of Field-Programmable Gate Arrays: The Effect of Logic Functionality on Area Efficiency," *IEEE Journal of Solid-State Circuits*, vol. 25, pp. 1217 - 25, 1990.

[285] E. Ahmed and J. Rose, "The Effect of LUT and Cluster Size on Deep-Submicron FPGA Performance and Density," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 12, pp. 288-298, 2004.

[286] Y. Lin, F. Li, and L. He, "Circuits and Architectures for Field Programmable Gate Array with Configurable Supply Voltage," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, vol. 13, pp. 1035-1047, 2005.

[287] J. Rose and S. Brown, "Flexibility of Interconnection Structures for Field-Programmable Gate Arrays," *IEEE Journal of Solid-State Circuits*, vol. 26, pp. 277 - 282, 1991.

[288] J. Rose and D. Hill, "Architectural and Physical Design Challenges for One-Million Gate FPGAs and Beyond," presented at ACM Symposuim on FPGAs, FPGA `97, pp. 129-132, 1997.

[289] N. Vassiliadis, S. Nikolaidis, S. Siskos, and D. J. Soudris, "The Effect of the Interconnection Architecture on the FPGA Performance and Energy Consumption," presented at 12th IEEE Mediterranean Electrotechnical Conference, MELECON 2004, pp. 213-216 Vol.1, 2004.

[290] J. L. Kouloheris and A. El Gamal, "FPGA Performance versus Cell Granularity," presented at IEEE Custom Integrated Circuits Conference, pp. 6.2/1 -6.2/4, 1991.

[291] J. L. Kouloheris and A. El Gamal, "PLA-based FPGA Area Versus Cell Granularity," presented at IEEE Custom Integrated Circuits Conference, pp. 4.3.1 -4.3.4, 1992.

[292] J. He and J. Rose, "Advantages of Heterogeneous Logic Block Architecture for FPGAs," presented at Custom Integrated Circuits Conference, pp. 7.4.1-7.4.5, 1993.

[293] H.-Y. Wong, L. Cheng, Y. Lin, and L. He, "FPGA Device and Architecture Evaluation Considering Process Variations," presented at IEEE/ACM International Conference on Computer-Aided Design, ICCAD-2005, pp. 19-24, 2005.

[294] X. Lin and M. Chan, "An Opposite Side Floating Gate FLASH Memory Scalable to 20 nm Length," presented at IEEE International SOI Conference, pp. 71-72, 2002.

[295] E. Goetting, D. Schultz, D. Parlour, S. Frake, R. Carpenter, C. Abellera, B. Leone, D. Marquez, M. Palczewski, E. Wolsheimer, M. Hart, K. Look, M. Voogel, G. West, V. Tong, A. Chang, D. Chung, W. Hsieh, L. Farrell, and W. Carter, "A Sea-of-Gates FPGA," *IEEE International Solid-State Circuits Conference*, vol. XXXVIII, pp. 110 - 111, 1995.

[296] V. Agarwal, S. W. Keckler, and D. Burger, "The Effect of Technology Scaling on Microarchitectural Structures," University of Texas at Austin, Austin, Technical Report TR2000-02, 2000, http://citeseer.ist.psu.edu/agarwal00effect.html

[297] A. Djupdal, S. Aunet, and V. Beiu, "Ultra Low Power Neural Inspired Addition: When Serial Might Outperform Parallel Architectures," presented at International Workshop on Artificial Neural Networks, IWANN'05, Barcelona, Spain, 2005.

[298] (2005) General Description: ProASICPLUS® Flash Family FPGAs. [Online]. Available: http://www.actel.com/documents/ProASICPLUSGenDes.pdf.

[299] The FPGA Place-and-Route Challenge. [Online]. Available: http://www.eecg.toronto.edu/~vaughn/challenge/challenge.html.

[300] J. A. Davis, V. K. De, and J. D. Meindl, "A Stochastic Wire-Length Distribution for Gigascale Integration (GSI). I. Derivation and Validation," *IEEE Transactions on Electron Devices*, vol. 45, pp. 580 - 589, 1998.

[301] P. Kongetira. (2004) A 32-way Multithreaded SPARC® Processor. [Online]. Available: http://www.hotchips.org/archives/hc16/3_Tue/11_HC16_Sess9_Pres2_bw.pdf.

[302] D. C. Pham, T. Aipperspach, D. Boerstler, M. Bolliger, R. Chaudhry, D. Cox, P. Harvey, P. M. Harvey, H. P. Hofstee, C. Johns, J. Kahle, A. Kameyama, J. Keaty, Y. Masubuchi, M. Pham, J. Pille, S. Posluszny, M. Riley, D. L. Stasiak, M. Suzuoki, O. Takahashi, J. Warnock, S. Weitzel, D. Wendel, and K. Yazawa, "Overview of the Architecture,Circuit Design, and Physical Implementation of a First-Generation Cell Processor," *IEEE Journal of Solid-State Circuits* vol. 41, pp. 179-196, 2006.

[303] S. Fu, Cost Performance Optimization of Microprocessors, PhD Thesis, Computer Systems Laboratory, Department of Electrical Engineering, Stanford University, 2001.

[304] J. D. Ullman, *Computational Aspects of VLSI*: Computer Science Press, Rockville, Md, 1984.

[305] P. Beckett and S. C. Goldstein, "Why Area Might Reduce Power in Nanoscale CMOS," presented at International Symposium on Circuits and Systems, ISCAS'05, Kobe, Japan, pp. 2329-2332, 2005.

[306] R. P. Brent and H. T. Kung, "The Chip Complexity of Binary Arithmetic," presented at Twelfth Annual ACM Symposium on Theory of Computing, Los Angeles, California, United States, pp. 190 - 200, 1980.

[307] R. P. Brent and H. T. Kung, "The Area-Time Complexity of Binary Multiplication," *Journal of the ACM*, vol. 28, pp. 521 - 534, 1981.

[308] H. Abelson and P. Andreae, "Information Transfer and Area-Time Tradeoffs for VLSI Multiplication," *Communications of the ACM*, vol. 23, pp. 20 - 23, 1980.

[309] R. J. Lipton and R. Sedgewick, "Lower Bounds for VLSI," presented at Thirteenth Annual ACM Symposium on Theory of Computing, Milwaukee, Wisconsin, United States, pp. 300 - 307, 1981.

[310] I. E. Sutherland, "Micropipelines," *Communications of the ACM*, vol. 32, pp. 720 - 738, 1989.

[311] A. Iyer and D. Marculescu, "Power and Performance Evaluation of Globally Asynchronous Locally Synchronous Processors," presented at 29th Annual International Symposium on Computer Architecture, ISCA'02, Anchorage, AK, USA, 2002.

[312] D. Lee, W. Kwong, D. Blaauw, and D. Sylvester, "Analysis and Minimization Techniques for Total Leakage Considering Gate Oxide Leakage," presented at 40th Conference on Design Automation, Anaheim, CA, USA, pp. 175 - 180, 2003.

[313] D. Marculescu and E. Talpes, "Energy Awareness and Uncertainty in Microarchitecture-Level Design," *IEEE Micro*, vol. 25, pp. 64-76, 2005.

[314] R. Ball. (2003) Smaller Processes Lead to FPGA Leakage Crisis. *Electronics Weekly,* [Online]. Available: http://www.edn.com/article/CA332188.html.

[315] C.-K. Huang and N. Goldsman, "Modeling the Limits of Gate Oxide Scaling with a Schrodinger-Based Method of Direct Tunneling Gate Currents of Nanoscale MOSFETs," presented at 1st IEEE Conference on Nanotechnology, IEEE-NANO 2001, pp. 335 -339, 2001.

[316] N. M. Ravindra and J. Zhao, "Fowler-Nordheim Tunneling in Thin SiO_2 Films," *Smart Materials and Structures*, vol. 1, pp. 197-201, 1992.

[317] Y. Khlifi, K. Kassmi, L. Roubi, and R. Maimouni, "Modeling of Fowler-Nordheim Current of Metal/Ultra-Thin Oxide/Semiconductor Structures," *The Moroccan Journal of Condensed Matter*, vol. 3, pp. 53-57, 2000.

[318] S. Keeney, R. Bez, D. Cantarelli, F. Piccinini, A. Mathewson, L. Ravazzi, and C. Lombardi, "Complete Transient Simulation of Flash EEPROM Devices," *IEEE Transactions on Electron Devices*, vol. 39, pp. 2750-2757, 1992.

[319] K. Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, "Leakage Current Mechanisms and Leakage Reduction Techniques in Deep-Submicrometer CMOS Circuits," *Proceedings of the IEEE*, vol. 91, pp. 305-327, 2003.

[320] H.-S. P. Wong, D. J. Frank, P. M. Solomon, C. H. J. Wann, and J. J. Welser, "Nanoscale CMOS," *Proceedings of the IEEE*, vol. 87, pp. 537 - 570, 1999.

[321] T. Skotnicki, F. Boeuf, and M. Müller, "MASTAR 2.0 User's Guide," International Technology Roadmap for Semiconductors 2003.

[322] S. Tyagi, C. Auth, P. Bai, G. Curello, H. Deshpande, S. Gannavaram, O. Golonzka, R. Heussner, R. James, C. Kenyon, S.-H. Lee, N. Lindert, M. Liu, R. Nagisetty, S. Natarajan, C. Parker, J. Sebastian, B. Sell, S. Sivakumar, A. S. Amour, and K. Tone, "An Advanced Low Power, High Performance,Strained Channel 65nm Technology," presented at IEEE International Electron Devices Meeting, IEDM2005, pp. 245-247, 2005.

[323] L. Chang, K. J. Yang, Y.-C. Yeo, Y.-K. Choi, T.-J. King, and C. Hu, "Reduction of Direct-Tunneling Gate Leakage Current in Double-Gate and Ultra-Thin Body MOSFETs," presented at International Electron Devices Meeting, IEDM 2001, pp. 5.2.1-5.2.4, 2001.

[324] L. Chang, K. J. Yang, Y.-C. Yeo, I. Polishchuk, T.-J. King, and C. Hu, "Direct-Tunneling Gate Leakage Current in Double-Gate and Ultrathin Body MOSFETs," *IEEE Transactions on Electron Devices*, vol. 49, pp. 2288-2295, 2002.

[325] X. Guo and T. P. Ma, "Tunneling Leakage Current in Oxynitride: Dependence on Oxygen/Nitrogen Content," *IEEE Electron Device Letters*, vol. 19, pp. 207-209, 1998.

[326] Y.-C. Yeo, T.-J. King, and C. Hu, "MOSFET Gate Leakage Modeling and Selection Guide for Alternative Gate Dielectrics Based on Leakage Considerations," *IEEE Transactions on Electron Devices*, vol. 50, pp. 1027-1035, 2003.

[327] S. A. Parke, J. E. Moon, H. C. Wann, P. K. Ko, and C. Hu, "Design for Suppression of Gate-Induced Drain Leakage in LDD MOSFETs using a Quasi-Two-Dimensional Analytical Model," *IEEE Transactions on Electron Devices*, vol. 39, pp. 1694-1703, 1992.

[328] J. G. Fossum, K. Kim, and Y. Chong, "Extremely Scaled Double-Gate CMOS Performance Projections, Including GIDL-Controlled Off-State Current," *IEEE Transactions on Electron Devices*, vol. 46, pp. 2195-2200, 1999.

[329] K. Saino, S. Horiba, S. Uchiyama, Y. Takaishi, M. Takenaka, T. Uchida, Y. Takada, K. Koyama, H. Miyake, and C. Hu, "Impact of Gate-Induced Drain Leakage Current on the Tail Distribution of DRAM Data Retention Time," presented at IEDM Technical Digest. International Electron Devices Meeting, San Francisco, CA, pp. 837-840, 2000.

[330] J. Chen, F. Assaderaghi, P.-K. Ko, and C. Hu, "The Enhancement of Gate-Induced-Drain-Leakage (GIDL) Current in Short-Channel SOI MOSFET and its Application in Measuring Lateral Bipolar Current Gain β," *IEEE Electron Device Letters*, vol. 13, pp. 572-574, 1992.

[331] O. Semenov, A. Pradzynski, and M. Sachdev, "Impact of Gate Induced Drain Leakage on Overall Leakage of Submicrometer CMOS VLSI Circuits," *IEEE Transactions on Semiconductor Manufacturing*, vol. 15, pp. 9-18, 2002.

[332] B. S. Landman and R. L. Russo, "On a Pin versus Block Relationship for Partitions of Logic Graphs," *IEEE Transactions on Computers*, vol. C-20, pp. 1469–1479, 1971.

[333] M. Bucher, C. Lallement, C. Enz, F. Théodoloz, and F. Krummenacher, "The EPFL-EKV MOSFET Model Equations for Simulation, Version 2.6, Revision II," Electronics Laboratories, Swiss Federal Institute of Technology (EPFL), Lausanne, Switzerland, Technical Report July 1998.

[334] J. A. Davis, V. K. De, and J. D. Meindl, "A Stochastic Wire-Length Distribution for Gigascale Integration (GSI). II. Applications to Clock Frequency, Power Dissipation, and Chip Size Estimation," *IEEE Transactions on Electron Devices*, vol. 45, pp. 590 -597, 1998.

[335] R. P. Colwell, *The Pentium Chronicles: The People, Passion, and Politics Behind Intel's Landmark Chips*: Wiley-IEEE Computer Society Press, 2005.

[336] F. Prégaldiny and C. Lallement, "Fourth Generation MOSFET Model and its VHDL-AMS Implementation," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 18, pp. 39-48, 2005.

[337] "IEEE Standard VHDL Analog and Mixed-Signal Extensions," in *IEEE Std 1076.1-1999*: IEEE Press, 1999.

[338] M. Jang, Y. Kim, M. Jun, C. Choi, T. Kim, B. Park, and S. Lee, "Schottky Barrier MOSFETs with High Current Drivability for Nano-regime Applications," *Journal Of Semiconductor Technology and Science*, vol. 6, pp. 10-15, 2006.

[339] L. Ding-Yu, S. Lei, Z. Sheng-Dong, W. Yi, L. Xiao-Yan, and H. Ru-Qi, "Schottky Barrier MOSFET Structure with Silicide Source/Drain on Buried Metal," *Chinese Physics*, vol. 16, pp. 240, 2007.

[340] K. Sano, M. Hino, N. Ooishi, and K. Shibahara, "Workfunction Tuning Using Various Impurities for Fully Silicided NiSi Gate," *Japanese Journal of Applied Physics*, vol. 44, pp. 3774-3777, 2005.

[341] F. Prégaldiny, F. Krummenacher, B. Diagne, F. Pêcheux, J.-M. Sallese, and C. Lallement, "Explicit Modelling of the Double-Gate MOSFET with VHDL-AMS," *International Journal of Numerical Modelling: Electronic Networks, Devices and Fields*, vol. 19, pp. 239-256, 2006.

[342] C. D. Thompson, "Area-Time Complexity for VLSI," presented at Eleventh Annual ACM Symposium on Theory of Computing, Atlanta, Georgia, United States, pp. 81 - 88, 1979.

Appendix A: TCAD Input Decks

Example input deck for TCAD Simulation of thin-body double-gate Schottky nMOS device.

```
go atlas
TITLE SOI device simulation
# This simulates a L=5nm thin-body, double-gate MOSFET device
# with Schottky barrier source and drain.  Its loosely based on a
# combination of Guo & Lundstrom's ultimate 10nm device and
# Saitoh metal gate device.
#
# The device comprises a 5nm thick x 20nm long channel of i-Si on 40nm BOX
# over a Si substrate.  Two gates are set into SiO2 (which forms the
# gate insulator as well.  Tox (gate) = 1.5nm.
#
mesh space.mult=1.0
#
x.mesh loc=0.00     spac=0.005
x.mesh loc=0.014   spac=0.002
x.mesh loc=0.020   spac=0.0001
x.mesh loc=0.030   spac=0.005
x.mesh loc=0.040   spac=0.0001
x.mesh loc=0.046   spac=0.002
x.mesh loc=0.060   spac=0.005
#
y.mesh loc=-0.0015   spac=0.00025
y.mesh loc=0.0065   spac=0.00025
y.mesh loc=0.010    spac=0.005
y.mesh loc=0.04      spac=0.015
#
eliminate columns y.min=0.02
# first place the oxide and channel regions
region      num=1 y.max=0.04  oxide
region      num=2 y.min=0  y.max=0.005 silicon
#
# the electrodes
# #1=front (control) gate; #2=back gate;#3=source;#4=drain #5=substrate
electrode     name=cgate   num=1 x.min=0.020 x.max=0.040 y.min=-0.0015 y.max=-
0.0015
electrode     name=pgate   num=2 x.min=0.020 x.max=0.040 y.min=0.0065
y.max=0.0065
electrode     name=source  num=3 x.min=0.0  x.max=0.020  y.min=0.0 y.max=0.005
electrode     name=drain    num=4 x.min=0.040 x.max=0.060 y.min=0.0 y.max=0.005
electrode     substrate
#
# now specify the physical attributes of these electrode - in particular the
metallic workfunctions
# we are simulating a silicide schottky device;
# source and drain will be erbium silicide erSi2 (workfun=0.28V above n-Si)
# workfun=Aff+dWf = 4.17+0.28 = 4.45 for n-type
# gate materials: Au/Cr - wf=4.7; p.poly - wf=5.06
#I've fiddled with these WF values to get Vth=0.25V (approx)
contact number=1 workfun=4.45
contact number=2 workfun=4.45
contact number=3 workfun=4.25
contact number=4 workfun=4.25
#
# region 2 is the intrinsic silicon channel
# Background doping set to 10^15cm^3 for convenience
doping       uniform conc=1e15 p.type  region=2
#
save outf=DG_5nm-1.5nmSOI_erSi_SD.str
tonyplot DG_5nm-1.5nmSOI_erSi_SD.str
# set interface charge separately on front and back oxide interfaces
```

```
#
Interf       qf=3e10 y.max=0
Interf       qf=1e11 y.min=0.005
#
# select models
models MOS
#
#
method gummel newton   gum.init=5 trap

solve init
# now look at threshold shift Vd=1V;
# #1=cgate; #2=pgate; #3=source;#4=drain
# set drain to 1V and back gate to 0
# compute the threshiold voltage - tie cgate to drain for this part i.e. Vgd=0
#solve    v2=0.0 electr=2
#log       outf=DG5nm-1.5nm_erSiSD_Vbg_0.0.log master
#solve     v1=0.05 vstep=0.05  vfinal=1.0 electr=1
#log   off
#extract init infile="DG5nm-1.5nm_erSiSD_Vbg_0.0.log"
#extract name="subvt" \
#        1.0/slope(maxslope(curve(v."cgate",log10(abs(i."drain")))))
#extract name="vt" (xintercept(maxslope(curve(v."cgate",abs(i."drain")))))
#
solve     v4=0.1 vstep=0.2 vfinal=1 elect=4
#now ramp back gate to -1.2
solve     v2=0.0 vstep=-0.1 vfinal=-1.2 electr=2
#
#start logging the Id/Vg result for various back gate values
log       outf=DG5nm-1.5nm_erSiSD_Vbg-1.2.log master
solve     v1=-0.1 vstep=0.05  vfinal=1.0 electr=1
log    off
solve     v2=-1.0 electr=2
log       outf=DG5nm-1.5nm_erSiSD_Vbg-1.0.log master
solve     v1=1.0 vstep=-0.05  vfinal=-0.1 electr=1
log    off
solve     v2=-0.8 electr=2
log       outf=DG5nm-1.5nm_erSiSD_Vbg-0.8.log master
solve     v1=-0.1 vstep=0.05  vfinal=1.0 electr=1
log    off
solve     v2=-0.6 electr=2
log       outf=DG5nm-1.5nm_erSiSD_Vbg-0.6.log master
solve     v1=1.0 vstep=-0.05  vfinal=-0.1 electr=1
log    off
solve     v2=-0.4 electr=2
log       outf=DG5nm-1.5nm_erSiSD_Vbg-0.4.log master
solve     v1=-0.1 vstep=0.05  vfinal=1.0 electr=1
log    off
solve     v2=-0.2 electr=2
log       outf=DG5nm-1.5nm_erSiSD_Vbg-0.2.log master
solve     v1=1.0 vstep=-0.05  vfinal=-0.1 electr=1
log    off
solve     v2=0.0 electr=2
log       outf=DG5nm-1.5nm_erSiSD_Vbg_0.0.log master
solve     v1=-0.1 vstep=0.05  vfinal=1.0 electr=1
log    off
solve     v2=0.2 electr=2
log       outf=DG5nm-1.5nm_erSiSD_Vbg_0.2.log master
solve     v1=1.0 vstep=-0.05  vfinal=-0.1 electr=1
log    off
solve     v2=0.4 electr=2
log       outf=DG5nm-1.5nm_erSiSD_Vbg_0.4.log master
solve     v1=-0.1 vstep=0.05  vfinal=1.0 electr=1
log    off
quit
```

Appendix B: SPICE Input Decks

Simulation Deck for Figure 56

```
* dc vin-vo characteristics of a nand gate formed from fd/soi n/pmos devices
* modified 18-09-02 9:00pm
* transistor size n=120nm p=100nm x 1um
* floating body
* we are using the characteristics of thin-body dual-gate FET devices
* to simulate the interaction of the back and front gates.
* In this case, the configuration is a 6-input NOR gate which we are
* exploring to see if it can form the basis of a reconfigurable
* cell for something like an FPGA.
* The idea is that the back-gate biases establish the threshold voltages
* for the front-gate, thereby allowing the logic mesh to be configured to
* perform a particular logic function.

.width in=120 out=120

vgf1    1  0   dc 0.0 pulse (0 1.0    10p 50p 50p 3n 6n)
vgf2    2  0   dc 0.0 pulse (0 1.0   200p 50p 50p 3n 6n)
vgf3    3  0   dc 0.0 pulse (0 1.0   400p 50p 50p 3n 6n)
vgf4    4  0   dc 0.0 pulse (0 1.0   600p 50p 50p 3n 6n)
vgf5    5  0   dc 0.0 pulse (0 1.0   800p 50p 50p 3n 6n)
vgf6    6  0   dc 0.0 pulse (0 1.0  1000p 50p 50p 3n 6n)
vgf8    8  0   dc 0.0 pulse (1.0 0     1n 50p 50p 2n 6n)
vgf9    9  0   dc 0.0 pulse (1.0 0   1.5n 50p 50p 0.5n 6n)
* these are the back-gate biases that set up the logic conditions
* 0 is normal logic bias, -2/+2 sets nMOS off, pMOS off, +1/0 =vice versa
vgb1  91  0   dc 2
vgb2  92  0   dc 1
vgb3  93  0   dc 0
vgb4  94  0   dc -1
vgd5  95  0   dc -2

vds   7  0   dc 1
vdum1 7 17   dc 0

*six input NOR gate structure
.SUBCKT 6NOR A B C D E F Z VDD
+ BN1 BN2 BN3 BN4 BN5 BN6
+ BP1 BP2 BP3 BP4 BP5 BP6
mp1  1 A VDD BP1 sem05p w=600n l=200n ad=0.4p as=0.4p
mp2  2 B   1 BP2 sem05p w=600n l=200n ad=0.4p as=0.4p
mp3  3 C   2 BP3 sem05p w=600n l=200n ad=0.4p as=0.4p
mp4  4 D   3 BP4 sem05p w=600n l=200n ad=0.4p as=0.4p
mp5  5 E   4 BP5 sem05p w=600n l=200n ad=0.4p as=0.4p
mp6  Z F   5 BP6 sem05p w=600n l=200n ad=0.4p as=0.4p
*
mn1  Z A  0 BN1 sem05n w=200n l=200n ad=0.4p as=0.4p
mn2  Z B  0 BN2 sem05n w=200n l=200n ad=0.4p as=0.4p
mn3  Z C  0 BN3 sem05n w=200n l=200n ad=0.4p as=0.4p
mn4  Z D  0 BN4 sem05n w=200n l=200n ad=0.4p as=0.4p
mn5  Z E  0 BN5 sem05n w=200n l=200n ad=0.4p as=0.4p
mn6  Z F  0 BN6 sem05n w=200n l=200n ad=0.4p as=0.4p
*
.ENDS

.SUBCKT NOT A Z VDD BN1 BP1
mp1  Z A VDD BP1 sem05p w=200n l=200n ad=0.4p as=0.4p
mn1  Z A 0   BN1 sem05n w=200n l=200n ad=0.4p as=0.4p
.ENDS

X1 1 2 3 4 5 6 25 17
+93 93 93 93 93 93
+92 92 92 92 92 92 6NOR
* this cell is used as a load for the first
X2 25 25 25 25 25 25 26 17
+93 93 93 93 93 93
+92 92 92 92 92 92 6NOR
* and this one is a model of an interconnect
*line over an adjacent cell - the load is the
```

```
*sources and drain capacitances see by the output of the driving cell
* back gates off, input held high (worse case?)
X3 8 29 17 93 92 NOT
X4 9 9 9 9 9 9 28 17
+95 95 95 95 95 95
+91 91 91 91 91 91 6NOR
R1 29 28 5K
C1 28 0 0.1e-16 ic=1.0

* this cell is used as a load for the interconnect line
X5 28 30 17 93 92 NOT

.option acct list node nopage reltol=3e-4 numdgt=4 itl1=1e3
+  gmin=1e-20 abstol=1e-15 chgtol=1e-15 vntol=1e-6 pivtol=1e-24
+  method=gear maxord=2

.op
.tran 50p 5n
.print tran v(1) v(6) v(8) v(25) v(28) v(29)
*.print tran i(vdum1)

* note: nfdmod=0 is not recognised
```

```
.model sem05n nmos level=10 selft=0
+  toxf=1.5n
+  toxb=2n
+  nsub=1e15
+  ngate=1e20
+  nds=1e20
+  tb=5n
+  nbody=1e16
+  lldd=0 nldd=1e20
+  dl=0.0
+  dw=0.0
+  nqff=2e10
+  nqfb=1e11
+  nqfsw=1e11
+  qm=0
+  uo=100.0
+  theta=1.0e-6
+  vsat=2.0e7
+  vo=0.0
+  alpha=0
+  beta=0
+  bgidl=4.0e9
+  gamma=1.0 kappa=1.0
+  jro=9.0e-11
+  m=1.5
+  ldiff=1e-7
+  seff=2e4
+  cgfdo=0.27e-9
+  cgfso=0.27e-9
+  cgfbo=0.0
+  rd=120e-6
+  rs=120e-6
+  fnk=3.0e-27
+  fna=1.0
+  rhob=30k
+  wkf=-0.9 wkb=-0.9
+  tpg=1.0
+  tps=-1.0
```

```
.model sem05p pmos level=10 selft=0
+  toxf=1.5n
+  toxb=3n
+  nsub=1e13
+  ngate=1e19
+  nds=1e19
+  tb=7.5n
+  nbody=1e16
+  lldd=0 nldd=1e20
+  dl=0
+  dw=0
+  nqff=-1e10
+  nqfb=-1e11
+  nqfsw=0
+  qm=0
+  uo=175
+  theta=1.0e-6
+  vsat=6.0e6
+  vo=0.0
+  alpha=2.45e6
+  beta=1.92e6
+  bgidl=4.0e9
+  gamma=1.0 kappa=1.0
+  jro=9.0e-11
+  m=1.5
+  ldiff=1e-7
+  seff=2e4
+  cgfdo=0.27e-9
+  cgfso=0.27e-9
+  cgfbo=0.0
+  rd=80e-6
+  rs=80e-6
+  fnk=3.0e-27
+  fna=1.0
+  rhob=30k
+  wkf=0.9 wkb=0.9
+  rhob=30e3
+  tpg=1.0
+  tps=-1.0
```

```
.end
```

Simulation Deck for Figure 58

```
* dc vin-vo characteristics of a nand gate formed from fd/soi n/pmos devices
* modified 03-01-03
* floating body
* we are using the characteristics of thin-body dual-gate FET devices
* to simulate the interaction of the back and front gates.
* In this case, the configuration is a 6-input NAND gate which we are
* exploring to see if it can form the basis of a reconfigurable
* cell for something like an FPGA.
* The idea is that the back-gate biases establish the threshold voltages
* for the front-gate, thereby allowing the logic mesh to be configured to
* perform a particular logic function.
* The DG transistors have been arranged into a 6x6 array that will be used to
* experiment with various DFF circuits.
*
.width in=120 out=120

v-din   1  0   dc 0 pulse (0 1   10p 20p 20p 1.5n 3n)
v-clk   2  0   dc 0 pulse (0 1   1.2n 20p 20p 600p 1.5n)
v-nclk  3  0   dc 1 pulse (1 0   1.2n 20p 20p 600p 1.5n)
v-nrst  4  0   dc 1 pulse (0 1   600p 20p 20p   4n 4n)
v5      5  0   dc 1
v6      6  0   dc 1

* these are the back-gate biases that set up the logic conditions
* 0 is normal logic bias, -2 sets nMOS off/pMOS on, +2 vice versa
vgb1  91  0   dc -2
vgb2  92  0   dc 0
vgb3  93  0   dc 2
*
vds   7  0   dc 1
vdum1 7 17   dc 0

* body contact not specified => floating
* six input NAND gate structure
*
.SUBCKT 6NAND A B C D E F Z VDD B1 B2 B3 B4 B5 B6
mp1  Z A VDD B1 sem12p l =0.2u w =0.32u
mp2  Z B VDD B2 sem12p l =0.2u w =0.32u
mp3  Z C VDD B3 sem12p l =0.2u w =0.32u
mp4  Z D VDD B4 sem12p l =0.2u w =0.32u
mp5  Z E VDD B5 sem12p l =0.2u w =0.32u
mp6  Z F VDD B6 sem12p l =0.2u w =0.32u
*
mn1   Z A  1 B1 sem12n l =0.12u w =0.2u
mn2   1 B  2 B2 sem12n l =0.12u w =0.2u
mn3   2 C  3 B3 sem12n l =0.12u w =0.2u
mn4   3 D  4 B4 sem12n l =0.12u w =0.2u
mn5   4 E  5 B5 sem12n l =0.12u w =0.2u
mn6   5 F  0 B6 sem12n l =0.12u w =0.2u
.ENDS
*
.SUBCKT 6X6NAND A B C D E F VDD O1 O2 O3 O4 O5 O6
+ B11 B12 B13 B14 B15 B16
+ B21 B22 B23 B24 B25 B26
+ B31 B32 B33 B34 B35 B36
+ B41 B42 B43 B44 B45 B46
+ B51 B52 B53 B54 B55 B56
+ B61 B62 B63 B64 B65 B66
*
X1 A B C D E F O1 VDD B11 B12 B13 B14 B15 B16 6NAND
X2 A B C D E F O2 VDD B21 B22 B23 B24 B25 B26 6NAND
X3 A B C D E F O3 VDD B31 B32 B33 B34 B35 B36 6NAND
X4 A B C D E F O4 VDD B41 B42 B43 B44 B45 B46 6NAND
X5 A B C D E F O5 VDD B51 B52 B53 B54 B55 B56 6NAND
X6 A B C D E F O6 VDD B61 B62 B63 B64 B65 B66 6NAND
.ENDS
*
* instantiate two adjacent 6x6 blocks and connect them up to form a +ve edge dff
* Q1 = /(/(Q1./clk) . /(Q0.clk) . /(Q1.Q0)
* Q0 = /(/(Din./clk) . /(Q0.clk) + /(Q0.Din)
* 1=Din, 2=clk, 3=nclk (i.e. not clock) 4 not reset, 5=Q1, 6=Q0
* 25=nclk.nr.Q1, 26=clk.nr.Q0, 27=nr.Q1.Q0,
* 28=Din.nclk.nr, 29=Din.nr.Q0
```

```
* 35=Q0; 36=Q1=Q(out);
* remember that a connection to node 93 turns a line off, to 92=logic active
*
X11 1 2 3 4 36 35 17 25 26 27 28 29 30
+ 93 93 92 92 92 93
+ 93 92 93 92 93 92
+ 93 93 93 92 92 92
+ 92 93 92 92 93 93
+ 92 93 93 92 93 92
+ 93 93 93 93 93 93 6X6NAND
X12 25 26 27 28 29 30 17 35 36 37 38 39 40
+ 93 92 93 92 92 93
+ 92 92 92 93 93 93
+ 93 93 93 93 93 93
+ 93 93 93 93 93 93
+ 93 93 93 93 93 93
+ 93 93 93 93 93 93 6X6NAND
*
*
.IC v(25)=0 v(26)=0 v(27)=0 v(28)=0 v(29)=0 v(30)=0
+   v(35)=0 v(36)=0 v(37)=0 v(38)=0 v(39)=0 v(40)=0
.op
*
.tran 100p 5n
.print tran v(1) v(2) v(3) v(4) v(35) v(36)
*.print tran i(vdum1)

.option acct list node nopage numdgt=4 itl1=5e2 itl2=50 itl4=20
+ gmin=1e-12 abstol=1e-8 vntol=1e-6 pivtol=1e-14
+ method=gear maxord=2 reltol=5e-4 stepgmin
* note: nfdmod=0 is not recognised
```

```
.model sem12n nmos level=10 selft=0 body=0

+  nqfsw=1e11
+  qm=0.4
+  ngate=1e20
+  bgidl=4.0e9
+  fna=1.0
+  fnk=3.0e-27
+  lldd=0.0 nldd=1e20
+  gamma=1 kappa=1
+  rhob=30k
+  toxf=0.002u
+  toxb=0.003u
+  nqff=2e10
+  nqfb=1e11
+  nsub=1e16
+  tb=5n
+  nbody=1e10
+  uo=600.0
+  theta=1.0e-6
+  vsat=6.0e6
+  vo=0.2
+  tps=1.0
+  tpg=1.0
+  nds=5e19
+  alpha=2.45e6
+  beta=1.92e6
+  cgfdo=0.27e-9
+  cgfso=0.27e-9
+  rd=120e-6
+  rs=120e-6
+  dl=0.04u
+  jro=9.0e-11
+  m=1.5
+  seff=2e4
+  ldiff=1e-7
```

```
.model sem12p pmos level=10 selft=0 body=0
+  toxf=2n
+  toxb=3n
+  nsub=1e15
+  ngate=1e20
+  nds=1e20
+  tb=5n
+  nbody=1e10
+  lldd=0 nldd=1e20
+  dl=0.004u
+  dw=0.0
+  nqff=-2e10
+  nqfb=-1e11
+  nqfsw=-1e11
+  qm=0.4
+  uo=200.0
+  theta=1.0e-6
+  vsat=6.0e6
+  vo=0.0
+  alpha=2.45e6
+  beta=1.92e6
+  bgidl=4.0e9
+  gamma=1.0 kappa=1.0
+  jro=9.0e-11
+  m=1.5
+  ldiff=1e-7
+  seff=2e4
+  cgfdo=0.27e-9
+  cgfso=0.27e-9
+  cgfbo=0.0
+  rd=120e-6
+  rs=120e-6
+  fnk=3.0e-27
+  fna=1.0
+  rhob=30k
+  tps=-1.0
+  tpg=1.0
```

```
.end
```

Appendix C: VHDL-AMS 6-NOR Adder Description

VHDL-AMS description of an adder formed from 2x 6NOR reconfigurable blocks.

```
library IEEE;
use ieee.std_logic_1164.all;
use ieee.math_real.all;
library IEEE_PROPOSED;
use ieee_proposed.electrical_systems.all;
library MGC_AMS;
use MGC_AMS.conversion.all;
library arithmetic;
use arithmetic.std_logic_arith.all;

entity fn_6NOR_adder is
      generic(Wp: real := 3.0;
              Wn: real := 1.0);--these are W/L
            port(terminal Vd: electrical;
                  signal V, alpha, Vth, CSqr: in real;
                  signal inp: in std_logic_vector(5 downto 0);
                   signal outp: out std_logic_vector(5 downto 0));
end entity fn_6NOR_adder;

architecture behavior of fn_6NOR_adder is

      constant PHYS_K: real := 1.380_6503e-23;
      constant PHYS_Q:real := 1.602_176_462e-19;
      constant TempC: real := 27.0;-- Ambient Temperature [Degrees]
      constant TempK: real:= 273.0 + TempC;--Temperature [Kelvin]
      constant vt: real:= PHYS_K*TempK/PHYS_Q;  -- Thermal Voltage
      constant oneon_nvt: real := 1.0/(1.74*Vt);--S=100mV/dec
      constant K1: real := 1.0e-10; --delay scaling factor
      constant K2: real := 1.0e-6; -- isub scaling factor
      --these are the output lines from block 0
      signal L: std_logic_vector(4 downto 0) := (others=>'0');
      signal Ao: std_logic_vector(5 downto 0) := (others=>'0');

      quantity Vdd across Idd through Vd to electrical_ref;
      signal delay: real := 0.0;
      signal iddL: real_vector(4 downto 0) := (others=>0.0);
      signal iddO: real_vector(3 downto 0) := (others=>0.0);
      signal iddt: real := 0.0;

begin
--Functional simulation of a 1-bit full adder
-- formed from two 6NOR blocks
--the delay is a function of supply, threshold and alpha
--ain       0     1     2     3     4     5
--          Cin  /Cin   A     /A    B     /B
--aout      0     1     2     3     4     5
--          Co  /Co     X     X     S     /S
-------------------------------------------------------------------
      delay <= K1*Csqr*V/((V-Vth)**alpha);
      --/(/A+/B+/C)
      L(0) <= not(inp(3) or inp(5) or inp(1)) after real2time(delay);
```

```
      --/(A+B+/C)
      L(1) <= not(inp(2) or inp(4) or inp(1)) after real2time(delay);
      --/(A+/B+C)
      L(2) <= not(inp(2) or inp(5) or inp(0)) after real2time(delay);
      --/(/A+B+C)
      L(3) <= not(inp(3) or inp(4) or inp(0)) after real2time(delay);
      --/(A+B+C)
      L(4) <= not(inp(2) or inp(5) or inp(0)) after real2time(delay);
      --Co
      Ao(0) <= not(L(1) or L(2) or L(3) or L(4)) after real2time(delay);
      --/Co
      Ao(1) <= L(1) or L(2) or L(3) or L(4) after real2time(delay);
      Ao(2) <= '1';--unused   Ao(3) <= '1';--unused
      --S
      Ao(4) <= L(0) or L(1) or L(2) or L(3) after real2time(delay);
      --/S
      Ao(5) <= not(L(0) or L(1) or L(2) or L(3)) after real2time(delay);
-----------------------------------------------------------------
-- and, finally, the subthreshold current
-- this simple model assumes that each intermediate
-- line (L) and output line contributes to Idd weighted by the number of
-- cells back-gated on.  The off cells are assumed to make insignificant
-- Idd contribution.  There are 3 cells gated on for each L and 4
-- on each output line.

      with L(0) select
            iddL(0) <= 3.0*Wn*K2*exp(-Vth*oneon_nvt) when '1',
            K2*Wp*exp(-Vth*oneon_nvt) when others;
      with L(1) select
            iddL(1) <= 3.0*Wn*K2*exp(-Vth*oneon_nvt) when '1',
            K2*Wp*exp(-Vth*oneon_nvt) when others;
      with L(2) select
            iddL(2) <= 3.0*Wn*K2*exp(-Vth*oneon_nvt) when '1',
            K2*Wp*exp(-Vth*oneon_nvt) when others;
      with L(3) select
            iddL(3) <= 3.0*Wn*K2*exp(-Vth*oneon_nvt) when '1',
            K2*Wp*exp(-Vth*oneon_nvt) when others;
      with L(4) select
            iddL(4) <= 3.0*Wn*K2*exp(-Vth*oneon_nvt) when '1',
            K2*Wp*exp(-Vth*oneon_nvt) when others;
      with Ao(0) select
            iddO(0) <= 3.0*Wn*K2*exp(-Vth*oneon_nvt) when '1',
            K2*Wp*exp(-Vth*oneon_nvt) when others;
      with Ao(1) select
            iddO(1) <= 3.0*Wn*K2*exp(-Vth*oneon_nvt) when '1',
            K2*Wp*exp(-Vth*oneon_nvt) when others;
      with Ao(4) select
        iddO(2) <= 3.0*Wn*K2*exp(-Vth*oneon_nvt) when '1',
        K2*Wp*exp(-Vth*oneon_nvt) when others;
      with Ao(5) select
            iddO(3) <= 3.0*Wn*K2*exp(-Vth*oneon_nvt) when '1',
            K2*Wp*exp(-Vth*oneon_nvt) when others;

      iddt <= iddL(0)+iddL(1)+iddL(2)+iddL(3)+iddL(4)+iddO(0)+
                  iddO(1)+iddO(2)+iddO(3);
      outp<=Ao;
      idd == iddt'ramp;
end behavior;
```