

Chapter 2

Design Challenges in Subthreshold Interconnect Circuits

Keywords CMOS buffer · Interconnect parasitics · Power dissipation · Weak inversion · Ultra-low-power

As technology advances to giga-scale integration level, global interconnect resource becomes increasingly valuable in a VLSI chip. This is due to the exponential growth of the total number of interconnects/wires as the feature size of MOS transistors decreases in scaled deep submicron CMOS technologies. Interconnect length, however, has not scaled down with feature size and remains long relative to other on-chip geometries. Interconnects are metal or polysilicon wires which connect billions of active devices to carry signals within a VLSI chip. There are a number of such wires in the whole chip. Of these, the length of long interconnects in large chips is of the order of 10 mm.

Interconnect and device performance in VLSI circuits depends on materials, geometry, and technology. With the dimensional scaling of technology, technological device and interconnect challenges have been closely examined by different researchers [28–33]. The delay in VLSI chips is due to active devices and interconnects. To avoid prohibitively larger delays, designers scale down global interconnect dimensions more slowly than the transistor dimensions [34]. Rather, reverse scaling is preferred for the global interconnects.

Interconnects also cause excessive power to be dissipated. In recent years, there has been a compelling demand for ultra-low-power devices to ensure longer battery lifetimes. Subthreshold circuits are ideally suited for applications where minimizing energy per operation is of prime importance [35–36]. Subsequently, the benefits from ultra-low energy operation have carved out a significant niche for subthreshold circuits. Furthermore, subthreshold circuits show exponential susceptibility to the process and temperature variations. Therefore, subthreshold operating region has made the design of energy-constrained robust ultra-low-power systems a very challenging design task. The present chapter reviews in detail the various aspects of buffer-driven long interconnect under subthreshold for ultra-low-power logic and the other associated problems. These are presented in the subsequent sections.

2.1 Interconnects for VLSI Applications—A Review

The long interconnects connect a larger number of active devices on a chip. These long interconnects distribute clock and signals and provide power and ground to the various circuits on a chip. Moreover, the associated parasitic impedance parameters increase as interconnect length increases. An overview of interconnect parasitics has been given in the following section.

2.1.1 Parasitic Impedance Parameters

Interconnect parasitics namely resistance, inductance, and capacitance lead to various undesirable effects in VLSI circuit design. These result in signal delay, power dissipation, distortion, and crosstalk. These problems are due to fundamental, material, device, circuit, and system physical limitations and need to be addressed while designing VLSI chips [37–39]. Interconnect has been represented by parasitic equivalent electrical components viz. resistance, inductance, and capacitance [40]. Such a lumped representation of the interconnect model is appropriate for medium and long interconnects at low-frequency applications. The parasitic impedance parameters are frequency dependent and responsible for decreased circuit efficiency and performance [41].

Eo and Eisenstadt [42] have developed models for high-speed and high-density VLSI circuit and found that interconnect circuit parameters vary with frequency. The model considers the silicon substrate properties, pad parasitics, fringing effects, and frequency variant properties of the circuit parameters. The model parameters are compared to scattering parameter measurements as well as PISCES-II simulations, and a good agreement is obtained with s -parameter measurements. Qian et al. [43] have developed an analytical expression for the effective load capacitance of resistive–capacitive (RC) interconnects. It is proved that, when there is a significant shielding, the response waveforms at the gate output may have a large exponential tail. This in turn can strongly influence the delay of RC interconnects. The concept of effective capacitance is extended to develop an equation on the basis of a two-piece gate-output approximation. The equation is solved accurately to obtain response waveform.

Delmore et al. [44], Moll et al. [45] and Wong et al. [46] have derived set of formulas to model capacitance and inductance in sub-half-micrometer VLSI circuits. Quasi three-dimensional (3D) modeling has been used for extracting the interconnect capacitance [47–48]. In the capacitance model, concept of effective width for a 3D wire has been used. This is derived from the combination of an analytical two-dimensional (2D) and ‘wall-to-wall’ model. The effective width provides a physics-based approach to decompose any 3D structure into a series of 2D segments, resulting in efficient and accurate capacitance extraction. Three-dimensional capacitance model for full-chip simulation has also been proposed in [49]. Huang et al. [50]

have carried out interconnect modeling for multi-gigahertz clock. However, accurate estimation of these interconnect parasitics requires details of the interconnect geometry, layout, technology, the current distributions and switching activities of the wires, which are difficult to predict and require more research. Rosa [51] has given the formula for self and mutual inductance using Biot–Savart law for linear conductors. Banerjee and Mehrotra [52] have introduced an accurate analysis of on-chip inductance effects for distributed interconnects that takes into account the effect of series resistance and output parasitic capacitance of the driver. The expressions for the transfer function of distributed interconnect lines, their time-domain responses, and computationally efficient performance optimization techniques have been presented. Closed-form approximation of frequency-dependent mutual impedance per unit length of lossy silicon substrate coplanar-strip IC interconnects has been developed in [53]. The derivation is based on a quasi-stationary full-wave analysis and Fourier integral transformation.

Sylvester and Hu [34] have considered the characterization of interconnect with particular attention to ultra-small capacitance measurement and in-situ noise evaluation techniques. An approach called the charge-based capacitance measurement technique, to measure Femto-Farad level wiring capacitances, has the advantages of being compact, having high-resolution and being very simple. Cong and Pan [54] have presented a set of interconnect performance estimation models for design planning with consideration of various effective interconnect layout optimization techniques. These models can be used efficiently during high-level design space exploration, interconnect-driven design planning, and synthesis- and timing-driven placement to ensure design convergence for deep sub-micrometer designs. A systematic method for deriving the characteristic model of interconnects from time-domain vector fitting has been investigated in [55]. The method is based on the iteration and convolution of time series by recursion. The approach extracts model parameters from terminal voltage waveforms directly by time-domain vector fitting so that the transformation of frequency loading can be simulated efficiently in SPICE-compatible simulator. The contributive interconnect parasitic impedance parameters contribute significantly to delay in VLSI chips. Estimation of propagation delay through interconnect has been of great concern for VLSI designers. Therefore, consideration of interconnect delay has been developed next.

2.1.2 Interconnect Delay

Interconnect delay modeling has been a subject of research since 1970s. Estimation of propagation delay through interconnect requires accurate models for the propagation path. Over the years, several models for interconnect delays have been proposed and tested. Resistive interconnect optimization under Elmore delay model [56] is carried out by Sapatnekar [57]. Gupta et al. [58] have proved that the Elmore delay measure is an absolute upper bound on the actual 50 % delay of RC tree response. Moreover, this bound holds for input signals other than steps. The actual

delay asymptotically approaches the Elmore delay as the input signal rise time increases. A lower bound on delay is also developed using the Elmore delay and the second moment of the impulse response. Brocco et al. [59] have investigated macro-modeling and RC tree approaches giving a unified timing simulation method. The simulation method is faster than SPICE by two orders, for 2 μm CMOS technology. O'Brien and Savarino [60] have modeled the driving point characteristics of resistive interconnects for delay estimation. Compact expressions for worst-case time delay and crosstalk of coupled RC lines are proposed by Sakurai [61]. Kahng and Muddu [62] have developed an analytical delay model based on first and second moments to incorporate inductance effects in the delay estimation with step input. Delays estimated are within 15 % of SPICE-computed delay across a wide range of interconnect parameter values. A stochastic wiring distribution based upon Rent's Rule has been derived by Davis and Meindl [63–64]. The distribution determines wire-length frequency and enables a priori estimation of the local, semi-global, and global wiring requirements for future GSI systems. Brachtendorf and Laur [65] have provided analytical models by discretization of the telegrapher's equations, for the transient simulation of lossy interconnects. Chiprout [66] has presented guidelines for modeling on-chip interconnects for accurate simulation of high-performance ultra-large-scale integration designs. Pamunuwa and Tenhunen [67] have discussed the delay model for coupled interconnects. Analytical expressions for delay, buffer size, and number that are suitable in a priori timing analyses and signal integrity estimations have been developed.

Davis and Meindl [68, 69] have extended Sakurai's work [61] by including self and mutual inductance. The compact analytical expressions derived give an explanation for the transient response of high-speed distributed resistive–inductive–capacitive interconnect. Simplified expressions enable physical understanding and accurate estimation of transient response, propagation delay and crosstalk for global interconnects. Venkatesan et al. [70, 71] have significantly extended the work reported in Refs. [68, 69]. They have developed a new physical model for the transient response of distributed interconnects with a capacitive load. The solutions are verified by HSPICE simulations. These solutions are used to derive novel expressions for the propagation delay, optimum number, and size of buffers for buffer inserted distributed lines. The analysis defines a design space that reveals the trade-off between the number of buffers and wire cross section for specified delay and crosstalk constraints.

Xu and Mazumder [72] have introduced the passive discrete modeling technique using the numerical approximation method. This is called the differential quadrature method for estimating signal propagation delays through on-chip long interconnects. This delay modeling generates equivalent circuit interconnect models consisting of current and voltage sources, which can be directly incorporated into circuit simulators such as SPICE. Current sensing, model-reduction-based algorithms, etc., are some other delay analysis methods which have been proposed in [73]. Worst-case delay has been estimated by Chen et al. [74]. Singhal et al. [75] have presented a twofold approach for evaluating the signal and data carrying capacity of on-chip interconnects. In the first approach, the wire is modeled as a

linear time invariant system and frequency response is studied and higher transmission rate is achieved using ideal signal shape. The second approach addresses delay and reliability in interconnects. Lehtonen et al. [76] have presented a self-contained adaptive system for detecting and bypassing permanent errors in on-chip interconnects. The proposed system reroutes data on erroneous links to a set of spare wires without interrupting the data flow. An improved syndrome storing-based detection method is presented and compared to the in-line test method. In the presence of permanent errors, the probability of correct transmission in the proposed systems is improved by up to 140 % over the standalone Hamming code. These methods achieve up to 38 % area, 64 % energy, and 61 % latency improvements at comparable error performance. Morgenshtein et al. [77] have presented a unified logical effort delay model for paths composed of CMOS logic gates and resistive wires. The method provides conditions for timing optimization while overcoming the limitations of standard logical effort in the presence of interconnects. The condition of optimal gate sizing in a logic path with long wires is also given and the condition is achieved when the delay component due to the gate input capacitance is equal to the delay component due to the effective output resistance of the gate.

2.1.3 CMOS Buffer

It is an important technique in VLSI to drive interconnects by buffers. Buffers have been realized using CMOS inverters [41]. Researchers have modeled buffers differently and much work has been reported in literature about CMOS buffers. Shockley [78] and Shichman and Hodges [79] have developed square law models for MOSFETs in which drain current varies as a square of the effective gate voltage. These models have been extensively used in computer-aided analyses of CMOS switching circuits. However, the proposed models do not give accurate results as channel length is reduced. Sakurai and Newton [80] developed alpha-power model which defines current–voltage characteristics for short-channel transistors. In this model, the input waveform slope effects and parasitic drain/source resistance effects are included. It has been observed that neglecting p -channel transistor (PMOS) is not valid when the input ramp is very slow compared to the output waveform. However, the approximation is valid if the input slope exceeds one-third of the output slope, which is usually true in VLSI. Various approaches for taking into consideration the non-ideal effects in short-channel MOSFETs have been considered [80].

Deng and Shiau [81] have used the linear RC delay method to empirically calculate the delay in digital CMOS circuits. This generic linear RC model has the advantage of being simple and reliable. The empirical model, which is a high-dimensional function of various circuit and device parameters, is simplified to a 2D model that estimates the delay of CMOS circuits. SPICE simulation is used to verify the analytical results. Chung et al. [82] have carried out a comprehensive study of the performance and reliability design issues for deep submicrometer

MOSFET. The performance criteria viz. current-driving capability, ring-oscillator switching speed, and small-signal voltage gain are studied. In this context, the allowable choice of MOSFET channel length, oxide thickness, and power supply voltage is examined. Dutta et al. [83] have developed an analytical and comprehensive scheme to evaluate the delay and the output transition time of buffer for any input ramp and different fan-outs. Turn points for the infinitely fast and infinitely slow input rise times have been identified. A smooth curve fitting is used to predict the delay and the transition time over a large range of input signal slopes and output loading. The accuracy of the analysis is within 3 % of SPICE results. Bisdounis et al. [84] have suggested analytical transient and propagation delay models for short-channel CMOS inverter with fast and slow ramp inputs. They have used alpha-power law MOSFET model and taken gate-to-drain coupling capacitance into account. The analytical results show an error less than 3 %. The reduction of transistor-level models of CMOS logic gates to equivalent inverters, for the purpose of computing the supply current, power and delay in digital circuits has been carried out by Nabavi-Lishi and Rumin [85]. Hirata et al. [86] have derived propagation delay for static CMOS gates considering short-circuit current and the currents through capacitive load and gate capacitance. They demonstrated that the influence of short-circuit power on delay becomes large with slow input transition and small output load capacitance. The accuracy of this analytical method is better than that reported in [80], especially when the velocity saturation is large. The error of the analysis is within 8 % of SPICE results. Pattanaik et al. [87] have used geometric programming for the optimization of delay and power of nanoscale CMOS inverter.

Daga and Auvergne [88] have demonstrated a design-oriented comprehensive analytical model for CMOS inverter delay considering input slope, input-to-output capacitive coupling, short-circuit current, and short-channel effects. Gate input dependency and the input-slope-induced nonlinearity are considered. The overall calculated results are within 10 % of SPICE simulation results. Raja et al. [89] have given a new CMOS gate design that has different delays along various inputs to output paths. The delays are accomplished by inserting selectively sized permanently on series transistors at the inputs of a logic gate. The use of variable input delay CMOS gates for total glitch-free minimum dynamic power implementations of digital circuits has been demonstrated. Using c7552 benchmark circuit and described gates, power saving of 58 % is obtained. CMOS gate sizing, taking the dependence of fan-out, spurious capacitances and the slope of the input waveforms to optimize delay has been presented [90]. The alpha-power law model equations have been used.

2.2 Coupling Capacitance Noise

Interconnect coupling noise or crosstalk refers to the voltage induced on the victim node due to capacitive coupling from a switching aggressor node. The coupling capacitance causes disastrous effect on the functionality and reliability of digital integrated circuits. It induces a voltage glitch in one or more adjacent quiet

interconnects and may become a cause for circuit failure [91]. It also leads to false propagation delay times and increased power dissipation. In modern interconnect design, interconnects in adjacent metal layers are kept orthogonal to each other. This is done to reduce crosstalk as far as possible. But with growing interconnect density and reduced chip size, even the non-adjacent interconnects exhibit coupling.

The extent of coupling is dependent upon the nature of the signal transitions [92–94]. If both interconnects switch in the same direction, the coupling capacitance (C_c) is approximately zero and the total capacitance of each interconnect is approximated by the line-to-ground capacitance. If one interconnect is switching and the other is quiet, the total capacitance of each interconnect is determined by the capacitance ($C + C_c$). On the other hand, if the signals on each interconnect switch out of phase, the effective coupling capacitance approximately doubles to $2 \times C_c$. Thus, the coupling capacitance changes the effective load capacitance, depending upon the signal switching activity. Buffer insertion is a technique commonly used for the reduction of crosstalk. However, the buffer insertion technique in sub-threshold is not a feasible technique contrary to the super-threshold region. Another useful approach of reducing crosstalk is to use shielding wires, which also increases the capacitive load and therefore delay. A more suitable approach is to increase the spacing between the wires.

Extensive research has been carried out regarding crosstalk and delay estimation of CMOS gate-driven coupled interconnects. Xie and Nakhla [95] have proposed a method for crosstalk and delay estimation in high-speed VLSI interconnects with nonlinear components. The solution of the mixed frequency and time-domain problem by replacing the linear subnetworks, with a set of ordinary differential equations using the asymptotic waveform evaluation technique, has been obtained. Poltz [96] has given electromagnetic modeling of VLSI interconnects and the Helmholtz equation is used to build models which include eddy current loss and dielectric loss. Equivalent circuits with high cutoff frequencies and the smallest possible number of components are assembled. The performance of a VLSI interconnect at different clock rates is analyzed. Kuhlmann et al. [97] have proposed a time-efficient method for the precise estimation of crosstalk noise. A metric to compute coupling noise according to the sink capacitances and conductances of the aggressor and victim nets has been reported. The noise waveform is computed using a closed form leading to short computation time. The problem of crosstalk computation and reduction using circuit and layout techniques has been addressed in [98–99]. Expressions have been provided for noise amplitude and pulse width in capacitively coupled resistive lines. The estimation is based upon the RC transmission line model. A three-line structure of coupled RC interconnects using the transmission line model is presented in [100]. However, MOS transistor has been approximated by a linear resistor. Ling et al. [101] have developed a method to estimate the coupling noise in the presence of multiple aggressor nets. Authors have reported a novel technique for modeling quiet aggressor nets based on the concept of coupling point admittance and a reduction method to replace tree branches with effective capacitors. The proposed method has been tested for noise-prone interconnects from an industrial high-performance processor in 0.15 μm technology.

The worst-case error of 7.8 % and an average error of 2.7 % are observed. Devgan [102] has presented a metric for estimation of coupled noise in on-chip interconnects. This noise estimation metric is an upper bound known as the Devgan metric for RC circuits, being similar in spirit to Elmore delay in timing analysis. An enhancement to the Devgan metric has been proposed in [103] to improve the accuracy for fast input signals. The coupling noise voltage on a quiet interconnect line has also been analyzed by Shoji using a simple linear RC circuit [104]. Hashimoto et al. [105] have proposed a method to capture crosstalk-induced noisy waveform for crosstalk aware static timing analysis. The static timing analysis is performed with the consideration of dynamic delay variation due to crosstalk noise. Eo et al. [106] have proposed a simple closed-form crosstalk model and experimentally verified the model with 0.35 μm CMOS process-based interconnect test structures having two, three and five coupled lines with different switching scenarios. Becer et al. [107] have presented a complete crosstalk noise model which incorporates all victim and aggressor driver/interconnect physical parameters including coupling locations on both victim and aggressor nets. The validity of given model against SPICE has been demonstrated and has a good trade-off between accuracy and completeness, having an average error of 10 % with respect to SPICE for 130 nm technology. Hasan et al. [108] have derived and analyzed the crosstalk noise effect on a single victim line. An accurate and flexible decoupled transient model for victim wire is introduced. The model can be used to compute the maximum delay and glitch effect due to crosstalk under different slew rates. Tuuna et al. [109] have given an analytical model for the current drawn by on-chip bus. The model is combined with an on-chip power supply grid model in order to analyze noise caused by switching buses in a power supply grid. The buses are modeled as distributed lines that are capacitively and inductively coupled to each other. Different switching patterns and driver skewing times are also included in the model.

Bazargan-Sabet and Renault [110] have presented closed-form formulas to estimate capacitive coupling-induced crosstalk noise for distributed RC coupling trees. The formulas are simple enough to be used in the inner loops of performance optimization algorithms or as cost functions to guide routers. Kaushik et al. [111] have considered the effect of crosstalk-induced overshoot and undershoot generated at noise-site. The false switching occurs when the magnitude of overshoot or undershoot is beyond the threshold voltage of the gate. The peak overshoot and undershoot generated at noise-site can wear out the thin gate oxide layer resulting in permanent failure of the VLSI chip. Agarwal et al. [112] have analyzed a simple crosstalk noise model for coupled on-chip interconnects. The model is based on coupled-transmission-line theory and is applicable to asymmetric driver and line configurations. The noise waveform shape is captured well and yields an average error of 6.5 % for noise peak over a wide range of test cases. Chen and Sadowska [113] have proposed closed-form formula to estimate capacitive coupling-induced crosstalk noise for distributed RC coupling trees. The efficiency of the approach stems from the fact that only the five basic operations are used in the expressions viz. addition, subtraction, multiplication, division, and square root. Lee et al. [114]

have given crosstalk estimation method using coupled inductive tree models in high-speed VLSI interconnect. The recursive formulas for moment computation of coupled inductive interconnect trees with self and mutual inductances have been generalized. Nieuwoudt et al. [115] have given a comprehensive investigation of crosstalk-induced delay, noise, and capacitance for 65 nm process technology. Naeemi et al. [116] have described an analytical model that describes distributed inductive interconnects with ideal and non-ideal return path to optimize crosstalk and time delay of high-speed global interconnect structures such that the crosstalk and delay reduce by 38 and 12 %, respectively.

Vittal et al. [117] have addressed the problem of crosstalk computation and reduction using circuit and layout techniques. The expressions for crosstalk amplitude and pulse width in capacitively coupled resistive lines have been provided. The expressions hold good for nets with arbitrary number of pins and of arbitrary topology under any specified input excitation. The experimental results show that the average error is about 10 % and the maximum error is less than 20 %. Avinash et al. [118] have proposed a spatiotemporal bus encoding scheme to minimize crosstalk in interconnects. The scheme eliminates crosstalk in the interconnect wires, thereby reducing delay and energy consumption. The technique is evaluated by focusing on L1 cache address/data bus of a microprocessor using SPEC2000 CINT benchmark and suites for 90 and 65 nm technologies.

Nuroska et al. [119] have given a technique that reduces crosstalk noise on buses based on profiling the switching behavior. Based on this profiling information, an architecture configuration obtained using a genetic algorithm is applied that encodes pairs of bus wires, permutes the wires, and assigns an inversion level to each wire in order to optimize for noise and power. Hanchate and Ranganathan [120] have proposed a methodology for wire sizing with simultaneous optimization of interconnect crosstalk noise and delay in deep submicron VLSI circuits. The wire sizing is modeled as an optimization problem, formulated as a normal form game, and solved using the Nash equilibrium. Game theory allows the optimization of multiple metrics with conflicting objectives. Lienig [121] presented a novel approach to solve the VLSI channel and routing problems. The approach is based on a parallel genetic algorithm which runs on a distributed network of workstations. The algorithm optimizes physical constraints such as the length of nets, number of vias and is able to significantly reduce the occurrence of crosstalk.

Rao et al. [122] have proposed a bus encoding algorithm and circuit scheme for on-chip buses that eliminates capacitive crosstalk while simultaneously reducing total power. The encoding scheme significantly reduces total power by 26 % and runtime leakage power by 42 % while eliminating capacitive crosstalk. Zhang and Sapatnekar [123] have presented a method for incorporating crosstalk reduction criteria into the global routing under a broad power supply network paradigm. The method utilizes power/ground wires as shields between the signal wires to reduce capacitive coupling, while considering the constraints imposed by limited routing and buffering resources. An iterative procedure is employed to route signal wires, assign supply shields, and insert buffers. Wu et al. [124] have proposed a probabilistic model-based approach for crosstalk mitigation at the layer assignment. The

approach aims to discover and reduce crosstalk at the pre-detailed-routing level. Ho et al. [125] have given a novel framework for fast multilevel routing considering crosstalk and performance optimization. An intermediate stage of layer/track assignment has been incorporated into the multilevel routing framework. Compared with the state-of-the-art multilevel routing, the experimental results show that their approach achieved a $6.7\times$ runtime speedup, reduced respective maximum and average crosstalk by about 30 and 24 %, and reduced respective maximum and average delay by about 15 and 5 %.

Yoshikawa and Terai [126] have examined crosstalk-driven placement procedure based on genetic algorithm. For selection control, objective functions are introduced for improving crosstalk noise, reducing power consumption, improving interconnection delay, and dispersing wire congestion. Authors in [127] have proposed a coupling-driven data encoding scheme for low-power data transmission in deep submicron buses. The encoding scheme reduces the coupling transitions by 23 % for a deep submicron bus compared to the non-coded data transmission. It has been found that 75 % of the power consumption is due to coupling capacitance, whereas 25 % is due to self capacitance.

2.3 Power Dissipation

With the emergence of portable computing and communication equipments, low-power design has become a principal theme of the VLSI industry. The need for portability has caused a major paradigm shift in which power dissipation is as important as speed and area. The most demanding applications of low-power microelectronics have been battery-operated wrist watches, hearing aids, implantable cardiac pacemakers (a few μW power consumption), pocket calculators, pagers, cellular telephones (a few mW), and prospectively the hand-held multimedia terminals (10–20 W). The power dissipation in VLSI circuits is reviewed in this section. Various methodologies for reduction of power dissipation in VLSI circuits are also examined.

Powers [128] discussed the existing and emerging battery systems in terms of energy content, shelf and cycle life besides other characteristics. Progress in battery technology is still far behind than that in the field of electronics. Packaging has resulted in significant changes in the older systems such as C–Zn, alkaline, Zn–Air, NiCd, and lead acid which continue to get better. Chandrakasan et al. [129] have presented an analysis of low-power CMOS digital design, giving the techniques for low-power operation that use the lowest possible supply voltage coupled with architectural, logic style, circuit, and technology optimizations. The optimum voltage for 2 μm technology is 1.5 V and for 0.8 μm technology is 1 V, with power dissipation reduction by a factor of 10. The architectural-based scaling strategy indicates that the optimum voltage is much lower than that determined by other scaling considerations. Davari et al. [130] have given guidelines of CMOS scaling for low-power design. Comparisons are given for CMOS technologies ranging from

0.25 μm at 2.5 V to sub-0.1 μm at 1 V. It is shown that over two orders of magnitude improvement in power-delay-product are expected by such scaling compared to 0.6 μm devices at 5 V supply. Meindl [131] meticulously discussed the pros and cons of future opportunities for low-power GSI which are governed by a hierarchy of (i) theoretical and practical, (ii) material, (iii) device, (iv) circuit, and (v) system limits.

Low power is an essential requirement of biomedical electronic devices. Bhattacharyya et al. [132] have developed low-power hearing aid circuit based on 1 V supply voltage and adaptive biasing. Corbishley et al. [133] have proposed an ultra-low-power analog system to provide adaptive directionality in digital hearing aids. Power reduction is obtained by designing all the circuit blocks, viz. filters, multipliers, and dividers, in CMOS technology using transistors in weak inversion region. The total power consumption of the complete system is 5 μW at a scaled supply voltage of 0.9 V in 0.35 μm technology. Various power estimation techniques have been surveyed by Najm [134]. Rajput and Jamuar [135, 136] have reported low-voltage analog VLSI circuit design techniques and their applications. Power dissipation analysis of DSM CMOS circuits is carried out by Gu and Elmasry [137]. Borah et al. [138] and Heulser and Fichtner [139] considered transistor sizing for minimizing power consumption of CMOS circuit under delay constraint.

Authors in [140–143] have described several methodologies for low-power VLSI design. To contain the adverse effects of power dissipation, low-voltage operation of circuits, along with variable threshold and multiple threshold CMOS techniques, is often resorted to. System level architectural measures such as pipelining approach and parallel processing or hardware replication technique are used in the trade-off areas for low-power dissipation. Reduction of switching activity by algorithmic, architectural and circuit level optimization by proper choice of logic topology reduces power dissipation. Delay balancing, glitch reduction, and use of conditional or gated clock signals are some of the useful architectural measures to reduce switching activity. Switched capacitances play a significant role in switching power dissipation. Reduction of switched capacitances is a major step for low-power design of digital ICs. This can be accomplished (i) at system level by limiting the use of shared resources, e.g., by partitioning the global bus into smaller dedicated local buses to handle data transmission between nearby modules, (ii) by using proper logic style e.g., pass transistor logic reduces load capacitance, and (iii) by reducing parasitic capacitance at physical design level by keeping transistors at minimum dimensions whenever possible.

Kang [144] has reported an accurate method for simulating the power dissipation in an IC by the use of a dependent current source and a parallel RC circuit. The steady-state voltage across the capacitor reads the average power drawn from the supply voltage source. Simulation results are shown for CMOS circuits. This subcircuit can be inserted into any VLSI circuit model without causing interference while the circuit is simulated with a simulator such as SPICE. Yacoub and Ku [145] envisioned a circuit simulation technique which permits the measurement of short-circuit power dissipation component in ICs using SPICE. This technique is most

appropriate for low-power CMOS circuit design that does not permit current flow, other than leakage current, during steady-state operation.

Constandinou et al. [146] implemented an ultra-low-power consuming, simple, and robust circuit for edge-detection in integrated vision systems in 0.18 μm CMOS technology. Kim et al. [147] presented a low-power smallest area, delay-locked loop-based clock generator. Fabricated in a 0.35 μm CMOS process, clock generator occupies 0.07 mm^2 area and consumes 42.9 mW power and operates in the frequency range of 120 MHz–1.1 GHz. Bhaumik et al. [148] implemented a divided word line scheme to bring down power dissipation in 256 kB static random access memory design. Mitra and Chandorkar [149] designed a low-voltage CMOS amplifier with rail-to-rail input common mode range. Alternative methods were applied for obtaining high common mode range, good common mode rejection ratio, and output swing at such low supply voltage. Hwang et al. [150] reported a self regulating CMOS voltage-controlled oscillator with low supply voltage sensitivity. Lidow et al. [151] examined future trends in Internet appliances, portable electronic appliances, and silicon-based power transistors and diodes. It is discussed how the changing requirements of end users are driving state-of-the-art devices, new analog ICs as well as different power management architectures. Methodologies and projections related to power dissipation in CMOS circuits have been specified by Bhavnagarwala et al. [152].

Mutoh et al. [153] have proposed circuit by inserting high-threshold devices in series into low-threshold circuitry. A sleep control scheme is introduced for efficient power management. Kawaguchi et al. [154] have suggested super cutoff CMOS circuit that uses low-threshold voltage transistor with an inserted gate bias generator. In the standby mode, the voltages are applied to transistors to fully cut off the leakage current. Wei et al. [155] have implemented the dual-threshold technique to reduce leakage power by assigning a high-threshold voltage to some transistors in non-critical paths and using low-threshold transistors in the critical path. An algorithm for selecting and assigning an optimal high-threshold voltage is also given. The reduction in leakage power is more than 80 % and total active power saving is around 50 and 20 %, respectively, at low- and high-switching activities for ISCAS benchmark circuits. In [156], the authors have presented architectures for low power and optimum speed for image segmentation using Sobel operators.

Pant et al. [157] have presented algorithms that can be used to design ultra-low-power CMOS logic circuits by joint optimization of supply voltage, threshold voltage, and device widths. Various components of power dissipation are considered and an efficient heuristic is developed that delivers over an order of magnitude savings in power over conventional optimization methods. The authors have also proposed a heuristic technique for minimizing the total power consumption under a given delay constraint. The approach simultaneously determines transistor power supply, threshold voltage, and device width by two distinct phases. The proposed approaches trade off energy and delay invariably by tuning variables (supply voltage, threshold voltage, transistor size, etc.). Chi et al. [158] have proposed a multiple supply voltage-scaling algorithm for low-power design. The algorithm combines a greedy approach and an iterative improvement optimization approach.

Deodhar and Davis [159] have suggested voltage-scaling and repeater insertion for throughput-centric low-power global interconnects. It is assessed that repeater insertion improves throughput. Using 180 nm technology, it is illustrated that 1 V supply voltage can reduce power dissipation up to 25 % of that with 2.5 V supply, for 2 Gbps throughput. The results are compared with SPICE simulations and show a good agreement. The possibility of applying the buffer insertion technique to reduce power dissipation and delay in interconnects in voltage-scaled environment has been carried out in [160, 161]. Analytical approaches for optimum design and optimum number of buffers in low-power environment have been developed. Buffer sizing for minimum power and delay in voltage-scaled environment has also been carried out. The analytical results are within 10 % of the SPICE simulation results. Banerjee and Mehrotra [162] have addressed the problem of power dissipation during the buffer insertion phase of interconnect design optimization. Since all global interconnects are not on the critical path, a small delay penalty can be tolerated on these non-critical interconnects. A delay penalty of 5 % for lesser power dissipation at different MOS technologies has been included. It is proved that there exists a potential for large power saving by using smaller buffers and larger inter-buffer interconnect lengths. Wang et al. [163] have represented signals by localized wave packets that propagate along the interconnect lines at the speed of light to trigger the receivers. Energy consumption is reduced through charging up only part of the interconnect lines. Voltage doubling property of the receiver gate capacitances is used. Zhong and Jha [164] demonstrated the importance of optimizing on-chip interconnects for power reduction. It is concluded that significant spurious switching activity occurs in interconnects.

Tajalli and Leblebici [165] experimentally and analytically showed that scaling supply voltage in deep subthreshold region increases energy consumption and also investigated that optimum supply voltage for minimum energy consumption lies in moderate subthreshold region. Moreover, digital circuits operated in deep subthreshold region will have significant delay and noise margin penalties along with robustness issues that cannot be ignored for portable devices with real-time applications [166]. Hence, the designing of robust and moderate performance subthreshold field programmable gate arrays, real-time portable devices, buses, and clock signal is uncertain at such low bias [167].

Bol et al. [168] investigated the interests and limitations of technology scaling for subthreshold logic from 0.25 μm to 32 nm nodes. It is shown that scaling from 90 to 65 nm nodes is highly desirable for medium-throughput applications (1–10 MHz) due to great dynamic energy reduction. Upsizing of the channel length as a circuit level technique has been proposed to efficiently mitigate short-channel effects.

Thus, from the literature, it is clear that reducing power dissipation has been a crucial parameter for low-power VLSI designs. Also, energy-constrained VLSI applications have emerged for which the energy consumption is the key metric and speed of operation less relevant. The power consumption of these systems should decrease to the extent so as to extend the battery life and theoretically have unlimited lifetimes [169]. To cope with such ultra-low-power applications, design

of digital subthreshold logic was investigated. In the next section, fundamental aspects of subthreshold design for ultra-low-power circuits have been provided.

2.4 Weak Inversion for Ultra-Low-Power Logic

When gate-to-source voltage is less than or equal to the transistor threshold voltage, transistor is said to be biased in subthreshold. The transistor conducts current through an inverted channel between the source and drain caused by a voltage applied to the gate. The majority carriers in the substrate are repelled from the surface directly below the gate. A depletion charge of immobile atoms forms a depletion layer beneath the gate. The minority carriers in the depletion layer are made to move by diffusion and induce a drain current when a voltage that is less than the threshold voltage is applied between the drain and source terminals in the MOS device. This current is referred to as weak inversion current or subthreshold current. Due to small drive current, the subthreshold logic only fits in designs, where the performance is secondary and not the main concern. Since the leakage current is orders of magnitude lower than the drain strong inversion current and the power supply is reduced, subthreshold logic dissipates ultra-low-power [12, 13, 170]. The subthreshold circuit designs therefore offer significant savings in energy because reduction in power consumption outweighs the increase in delay by an order of magnitude [171]. These also provide near ideal voltage characteristic curve, a requirement for digital circuits [172]. Furthermore, in the subthreshold region, the transistor input capacitance is lesser than its strong inversion counterpart [36]. The low-operating frequency, low supply voltage, and smaller input gate capacitance combine together to reduce both dynamic and leakage power. A number of other advantages in subthreshold operation include improved gain, better noise margin, and tolerant to higher stack of series transistors.

Subthreshold digital operation was first examined theoretically in 1970s in the context of studying the limits of voltage scaling [173]. Subthreshold design was explored for low-power analog applications such as amplitude detector, quartz ring oscillator, band pass amplifier, and transconductance amplifier [174–176]. Digital subthreshold circuits were slower to catch on.

In the past years, a growing number of successful implementations of digital subthreshold systems viz. biomedical devices, fast fourier transform (FFT) processors, sensors, and static random access memory (SRAM) [18, 177–181] have occurred. An ultra-low-power delayed least mean square adaptive filter for hearing aids that uses parallelism is reported in [14]. The adaptive filter achieves 91 % improvement in power compared with a non-parallel CMOS implementation. The filter gives the desired performance of 22 kHz and operates at 400 mV. In 2001, Paul et al. designed an 8×8 array multiplier in 0.35 μm technology to operate in subthreshold operation [182]. The power-delay-product of this multiplier is around 25 times lower than its strong inversion operation. Body biasing is used to reduce the multiplier delay occurring due to temperature changes. A 2.60 pJ/instruction

subthreshold sensor processor in 0.13 μm technology has been fabricated in [183]. The minimum energy consumption is improved 10 times that of the previous sensor processor. A 180 mV subthreshold processor using FFT in 0.18 μm technology has been fabricated by Wang and Chandrakasan [16, 184]. The FFT processor dissipates 155 nJ for 16 bits and 1,024 point FFT at the optimum supply voltage. Besides the subthreshold static logic, other logic families such as subthreshold pseudo-NMOS, variable threshold voltage subthreshold CMOS, subthreshold dynamic threshold voltage MOS, and subthreshold dynamic logic have also been proposed [185, 186].

Thus, ultra-low-power applications have established a significant niche for subthreshold circuits [11]. In future CMOS technologies, domination of subthreshold logic over super-threshold logic for ultra-low-power moderate throughput applications is expected. However, process and temperature variations have become one of the most challenging obstacles in subthreshold circuits in recent deep sub-micron technologies. The process variations are dramatically accentuated in subthreshold designs. This topic is addressed in the following section.

2.5 Variability in Subthreshold Design

The variation occurring in the various design parameters of transistor viz. threshold voltage, oxide thickness, channel length, and mobility during the IC fabrication is termed as process variation. It may also be defined as fluctuations around the desired value of design parameters introduced during chip device fabrication [187]. The process variation issue is important in present day IC design [4]. This section briefly introduces this topic and thereafter, a survey of the literature that addresses variability in subthreshold circuits is conducted.

The impact of process variations on power and timing has become significant especially beyond 90 nm since the fabrication process tolerances have not scaled proportionally with miniaturization of the device dimensions [21]. As CMOS devices are further scaled in the nanometer regime, variations in the number and placement of dopant atoms in the channel region, called random dopant fluctuation (RDF), cause random variations in the threshold voltage. RDF makes it increasingly difficult to achieve threshold voltage accuracy. This further exaggerates the variability problem by producing variations in the subthreshold swing, drain current, and subthreshold leakage current [188]. Shockley [189] during his research on random fluctuations in junction breakdown first discovered random variation phenomenon in semiconductor devices. He explained that variations in the threshold voltage are randomly distributed according to Poisson distribution. Keyes further extended Shockley's work by studying the effect of variability on the electrical characteristics of a MOSFET [190]. These variations cause different relative strengths of the constituent transistors, thereby causing functional failure of logic gates [191]. Consequently, the output voltage rise and fall times differ, thus

impacting the switching frequency or power consumption. Body bias compensation circuits have been used to mitigate mismatch [192].

Authors in [193] have reduced the sensitivity to RDF through circuit sizing and the choice of circuit logic depth. The statistical models for circuit delay, power, and energy efficiency have also been derived. Kim et al. [194] have reduced the impact of RDF by device sizing optimization process which uses the reverse short-channel effect present in standard CMOS non-uniform halo doping profile devices. A transistor-level yield optimization technique to suppress process-induced variability has been proposed in [195].

Besides process variations, temperature variation has also a significant impact on the performance of subthreshold systems. The sources of temperature variations in VLSI circuits include ambient temperature and self heating. An increase in temperature increases the subthreshold leakage current and leakage power. This leakage power can be several orders of magnitude at higher temperatures.

2.5.1 Process Variations

The subthreshold current is exponentially dependent on the transistor threshold voltage. The threshold voltage is strongly related to the various device parameters such as effective gate length, oxide thickness, and doping concentration. These device parameters vary considerably in the DSM regime [196]. For example, a 10 % variation in the transistor effective length can lead to as much as a threefold difference in the amount of subthreshold leakage current [197]. Thus, subthreshold circuit designs are prone to process variations since this current drives the circuits.

The process variations have been classified into random and systematic variations [198]. Random variations can cause a device mismatch of identical and adjacent devices. The mismatch in the threshold voltage caused due to random variations decreases with decrease in doping and gate oxide thickness and increases when effective gate length and width decrease. This has been shown by Stolk et al. [199]. Systematic variations have been classified into across-field and layout-dependent variations. A cross-field variations cause identical devices in different parts of the chips to behave differently. Layout-dependent variations cause different layouts of the same device to have different characteristics. Even for the chips that meet the required operating frequency, a large portion dissipates very large amount of leakage power. This makes ICs unsuitable for commercial use [200]. These errors are caused due to photolithographic and etching sources, lens aberrations, mask errors, and variations in etch loading [201, 202]. Authors in [203] have carried out Monte Carlo analysis for a small MOSFET. They showed that controlling the process variation parameters to ± 10 % yields a threshold voltage variation of ± 15 %. They also showed that 95 % of the variance was around ± 100 mV about the mean threshold voltage with normal distribution. Bauer et al. [204] have shown that threshold voltage depends upon the depth of penetration of ions during ion implantation. Schemmert and Zimmer [205] introduced a procedure for

minimizing this threshold voltage sensitivity of ion-implanted MOSFETs. Their results showed a maximum deviation of 10 %. Kuhn et al. [206] have demonstrated that process variations also affect high-dielectric metal gates resulting in parametric variations in drive current, gate tunneling current, and threshold voltage. Recent work on process variation in subthreshold has focused upon the design of ultra-low-voltage SRAM with techniques reported for improving robustness [17, 207]. However, it fails to operate in the presence of process variability. Zhai et al. have highlighted three design challenges to ultra-low-voltage subthreshold SRAM [178]. These design challenges are (i) increased sensitivity to process variations, (ii) reduced on-current to off-current ratio which leads to difficulty in distinguishing between the read current of an accessed cell and the leakage current in the unaccessed cell, and (iii) the change in gate sizing requirements. The read and write stabilities of SRAM are heavily dependent upon the pull-up, pull-down, and pass transistors. The authors have presented a six-transistor SRAM design in subthreshold capable of overcoming aforementioned design challenges. The proposed design provides 36 % improvement in energy over other proposed SRAM designs with less area overhead.

Thus, process variation plays a key role in deciding robustness and energy efficiency of subthreshold designs. CMOS literature has always shown process variation as a critical element in semiconductor fabrication. The next section conveys some considerations regarding the effect of temperature in subthreshold VLSI circuits.

2.5.2 Temperature Variations

The temperature effects on subthreshold circuit operation have been investigated by Datta and Burleson [208]. It has been found that current exhibits positive temperature coefficient and increases exponentially with temperature while the on-to-off current ratio degrades by 0.52 %/°C due to the relative increase in leakage currents. Effect of temperature on subthreshold interconnect performance has also been carried out. It is found that optimal energy-delay-product can be achieved in the high temperature range of 75–90 °C. Authors in [209] have proposed an architecture and circuit for temperature sensors in 0.8 μm CMOS technology for ultra-low-power applications. The sensor draws 40 nA current from 1.6 to 3 V supply at room temperature. The circuit is suitable for temperature sensing in the 290–350 K temperature range. A temperature sensor suitable for passive wireless applications has been fabricated in 0.18 μm CMOS technology [210]. The temperature sensor consumes only 220 nW at 1 V at room temperature. It exhibits temperature inaccuracy of $-1.6\text{ }^{\circ}\text{C}/+3\text{ }^{\circ}\text{C}$ from 0 to 100 °C. Hanson et al. [211] have designed a processor for sensor applications. The processor is capable of working at 350 mV operating voltage in the subthreshold region while consuming only 3.5 pJ of energy per cycle.

The effect of temperature on the various leakage current components has also been explained by the various researchers. The thermal dependence of punch-through current has been explained in [212]. It is found that punch-through current reduces at low temperatures. The temperature dependence of drain-induced barrier lowering (DIBL) current in deep submicron MOSFETs has been extensively investigated in [213]. It is found that the DIBL coefficient is nearly insensitive to temperature reduction in the temperature interval from 300 to 50 K. Authors in [214] show that DIBL coefficient increases nearly 2.5 times under temperature reduction from 150 to 25 °C. The dependence of impact ionization current component has been studied in [215–217]. The leakage current due to impact ionization is temperature independent in the temperature interval from 300 to 77 K and significantly increases under technology scaling.

2.6 Concluding Remarks

From the literature survey, it is observed that parasitics are associated with VLSI global interconnects, which hamper the performance of integrated circuits. Significant research has been carried out on optimal interconnect design in super-threshold region. However, very limited literature deals with interconnect design challenges under subthreshold conditions. Further study of interconnect design, for ultra-low-power environment, is needed. Increased delay and crosstalk have become challenging design issues particularly for subthreshold interconnects. The driver delay rather than the interconnect delay dominates under subthreshold conditions. Subthreshold attracted attention in the digital domain in the late 1990s. Since then, several subthreshold systems have been implemented with standard deep submicron technologies. Subthreshold circuits have been best suited to meet the growing demand for battery-operated portable ultra-low-power VLSI applications. Process, voltage, and temperature variations on the subthreshold circuit behavior have been analyzed. It is observed that variability is accentuated in subthreshold designs and is one of the most challenging obstacles in deep submicron technologies. From the literature review, it is thus concluded that buffer-driven interconnects under subthreshold need investigation as these are useful for energy-constrained ultra-low-power applications. Alternatively, there are more precise existing circuit MOS models such as EKV, BSIM, and high-level empirical models implemented in HSPICE for the evaluation of CMOS circuit performance. These models do not provide a closed-form expression for the characteristics of the MOSFET. However, the relationship between the geometric structure and the electrical behavior can be elucidated appropriately only by compact analytical techniques.

Compact Models and Performance Investigations for
Subthreshold Interconnects

Dhiman, R.; Chandel, R.

2015, XIII, 113 p. 45 illus., Hardcover

ISBN: 978-81-322-2131-9