Chapter 3
Design Technologies for Nanoelectronic

**Haykel Ben Jamaa, Bahman Kheradmand Boroujeni,
Giovanni De Micheli, Yusuf Leblebici, Christian Piguet,
Alexandre Schmid, and Milos Stanisavljevic**

Introduction

As already explained in the introduction to Chap. , the development of economically feasible nanoelectronic systems requires a tight interplay between materials and fabrication technologies on the one hand and design technologies on the other. In particular, it is quite essential to explore *circuit-level measures* to mitigate the limitations of process variations (PVs), leakage, and reduced device reliability and, finally, to explore *system-level design* approaches that are better adapted to the constraints imposed by the materials, technology, and device physics. This chapter largely deals with some of these key questions that relate to design technologies for nanoelectronic systems.

Fault-tolerant design approaches for regular arrays based on silicon nanowires are discussed in detail in the section entitled "Fault-tolerant Design Approaches for Regular Arrays Based on Silicon Nanowires." Turning the focus more toward conventional technologies, the next section explores a novel technique for minimizing local delay variations for nanoscale CMOS technologies. Finally, the last section presents the adaptive V_{gs} technique for controlling the power and delay of nanometer-scale logic gates operating in a sub-VT regime.

Fault-Tolerant Design Approaches for Regular Arrays Based on Silicon Nanowires

With the progress of manufacturing technologies, many one- and zero-dimensional electronic devices have been designed and their operation demonstrated. These devices, which include nanowires (NWs) [1] and molecular switches [2], promise an ultrahigh integration density and push the fabricated circuits a few steps closer to

---

Yusuf Leblebici (✉)
EPFL – Swiss Federal Institute of Technology, Lausanne, Switzerland

the natural limits imposed by the physics of electron-based systems. In their mature stage, they did not reach such a level that their placement could be controlled in an accurate and cost-effective way. One promising design paradigm is based on a regular arrangement of these devices into arrays [3], which offers an easier fabrication approach and a higher fault tolerance and design flexibility thanks to the inherent redundancy level.

In this section we investigate different design approaches for regular arrays based on silicon NWs. We first present a CMOS-compatible fabrication technique for regular silicon NW arrays. Then we present some design challenges for regular NW circuits based on the example of crossbar memories. The challenges covered here include decoder design and reliable circuit testing.

Fabrication of Nanowire Arrays

A crossbar circuit is formed by two perpendicular layers of parallel NWs with molecular switches at their cross-points (Fig. 3.1). These molecular switches can store information (for memories) or perform computation (for logic).

It is of considerable interest to build arrays of dense parallel NWs while meeting the following requirements. On the one hand, it is desirable that the fabrication process be compatible with standard CMOS processes, not only for cost reasons, but also in order to integrate crossbar circuits onto CMOS chips. On the other hand, the NWs, which have a sublithographic resolution, need to be contacted by the lithographically defined outer circuit.

We propose the use of the spacer patterning technique (SPT) in order to address these issues. This technique was successfully used to build FinFET with sublithographic dimensions [4] by transforming a thin vertical dimension into a narrow horizontal dimension. An iteration of this technique [5] allows the fabrication of dense arrays of a few nanometer-wide parallel wires (Fig. 3.2).

The addressing of the NWs can be performed by a lithographically defined single-gate electrode laid out over the whole NW set. The current flow in the NWs is field-effect-controlled. Assuming that the NW FETs have different threshold voltages, one single-gate electrode is able to modulate the current flow through the
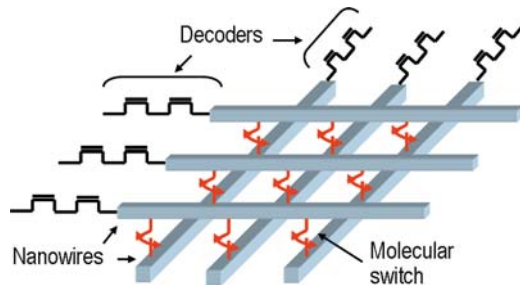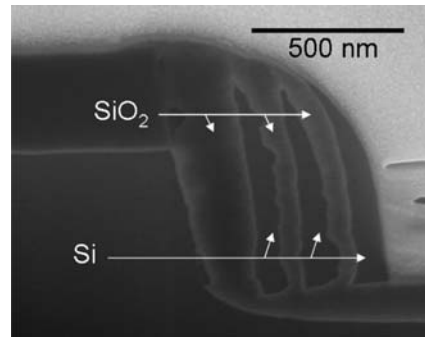

Fig. 3.1 Overall organization of a crossbar circuit
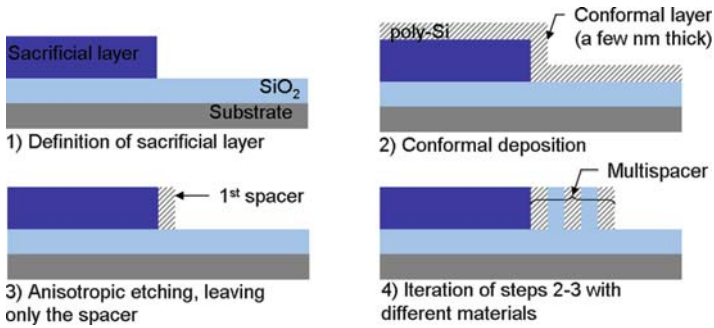
Fig. 3.2 Multispacer patterning technique

**Fig. 3.3** SEM cross-section of multispacer $(3 \times poly - Si + 2 \times SiO_2)$

whole set of NWs. The threshold voltage modulation can be achieved by different means. For instance, the variation of the NW thickness, or the doping level in the NWs, may induce the expected effect.

We applied this approach to fabricate arrays of long NWs and to address them. In Fig. 3.3 we show some scanning electron microscopy (SEM) pictures of the array cross-section with three to five parallel spacers.

Addressing Nanowire Arrays

The pitch of NW arrays is not dependent on the (photo-) lithography limit. For instance, the fabricated array shown above has a pitch of 150 down to 25 nm, and it is below the lithography pitch, as drawn in the layout. It is therefore necessary to address the issue of contacting and addressing every NW independently of the others in the same array. In what follows, we investigate the design aspects of the decoder that guarantees a unique addressing of NWs under high variability conditions. In order to investigate the design space with a concrete circuit, we assumed that the NW array operated as a crossbar memory.
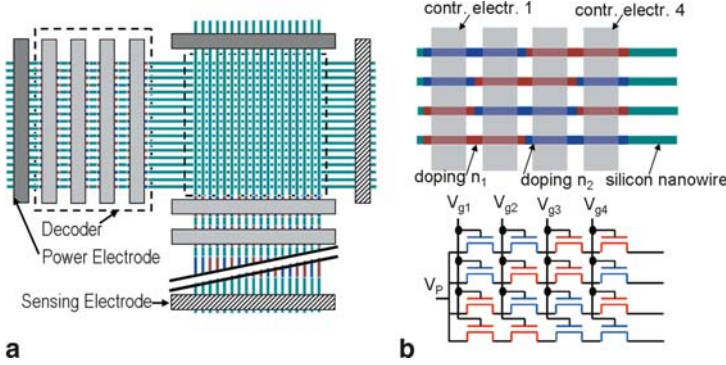
Fig. 3.4 **a** Crossbar memory with decoder. **b** Decoder layout and circuit

Nanowire Array Model and Technology

The crossbar memory architecture is depicted in Fig. 3.4a. The studied architecture has two parts, organized in an identical way and laid out perpendicularly to each other. Each part is a plane of N parallel NWs. M mesowires are used to address the NWs within each group. Then the NW decoder (Fig. 3.4b) has the size $N \times M$. The area sandwiched between the NW arrays is the actual memory, in which the information is stored in bistable switches grafted at the cross-points.

Each NW has to be addressed. This operation is performed by the NW decoder, which is formed by a set of parallel mesowires (or control wires) crossing the NW plane. The part of NWs under the decoder is coated by a dielectric, thus allowing a field-effect control of the NWs by the mesowires of the decoder. The NW technology assumed for this decoder exploits the bulk silicon fabrication platform reported in [6]. The proposed access devices in the decoder are gate-all-around (GAA) FETs whose threshold voltage depends on the doping level of the channel. We considered the use of a multiple-threshold voltage process that enables the fabrication of a multivalued logic decoder.

Multivalued Logic Codes for Nanowire Arrays

Each NW has a series of differently doped regions, defining a multivalued logic pattern. Each applied address is a series of voltages defining a multivalued logic code that covers a pattern, switches on all the transistors in the NW, and lets the current flow through it. A NW is said to be uniquely addressed by a code if this code covers only this NW pattern.

We can generalize the notion of binary code to multivalued logic by defining two types of codes: the n-ary hot code and the n-ary reflexive code. Both of them are addressable, i.e., a NW array coded with either an n-ary hot or reflexive code has

every NW addressed uniquely for any applied address. However, when defects affect the array, the addressing becomes more complex, as explained in what follows.

The control of silicon NWs is based on the modulation of the threshold voltage of the controlling transistors. The encoding schemes impose a distribution of the applied control voltages between the successive threshold voltages (V_T). The main issue with V_T is its variability and process-dependency. It is generally assumed that it follows a normal distribution.

We define a single-digit error as follows. If V_T exceeds a certain value, then the corresponding digit with value i will be detected as $(i + 1)$, and this is called a flip-up defect. The complementary case is when the threshold voltage drops beyond a fixed value, and then digit i is detected as $(i - 1)$, and this is called a flip-down defect.

If V_T varies within a small range close to its mean value, then the pattern does not change since the NW still conducts under the same conditions. Then, a one-to-one1 mapping between the code and the pattern space holds. By contrast, if the V_T variation is large, then some digits may be shifted up or down, as explained above. When a pattern has a sequence of errors, either it can be covered by one or more codes or it can be covered by no code. When we consider the codes, some of them cover one or more patterns and others cover no pattern under the error assumptions.

We defined the algorithms that estimate for every code type and length the part of NW patterns that become nonaddressable under given variability conditions [8]. The results could be confirmed by Monte Carlo simulations, which enable a more accurate assessment of the code behavior and a better exploration of the decoder design space.

Design of Nanowire Array Decoders

We considered a crossbar memory based on a double layer of NW arrays including the decoders, and we estimated the effective memory capacity in terms of defect-free and addressable bits per unit area. The effective memory capacity for different codes and technologies is depicted in Fig. 3.9a and shows an area saving up to ∼20% depending on the decoder design choices. Figure 3.5b shows the impact of the variation in the applied voltages at the decoder from its nominal value (υ) for a high variability level; the figure demonstrates that an optimized choice of the applied voltages improves the effective memory capacity.

Testing Nanowire Memories

Even though there are no complete memory systems based on the crossbar architecture yet, we believe that such systems will have the same architecture as CMOS memories [7, 9] (Fig. 3.6a). Unlike conventional RAM, crossbar memories have
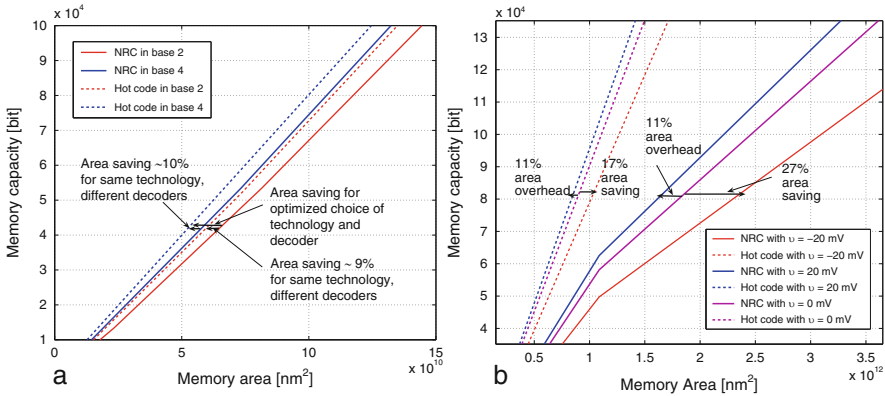
Fig. 3.5 a Area-capacity tradeoff for different codes. **b** Area-capacity tradeoff for different applied voltages

Fig. 3.6 Crossbar system
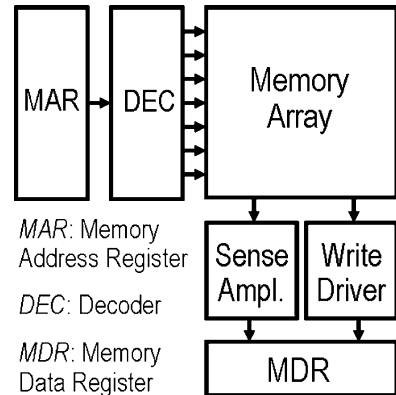
two parts: a sublithographic part formed by the decoder and the memory array and fabricated using one of the emerging technologies described in the section titled "Fabrication of Nanowire Arrays" and a lithographic part formed by the rest of the circuit and fabricated using CMOS technology.

The information is assumed to be stored in molecular switches grafted to every pair of crossing NWs. In the on-state, the molecule is conducting (logic 1), and in the off-state, it is highly resistive (logic 0). The writing operation is performed by first selecting the bit to be written, and then by applying a large positive or negative voltage at the pair of NWs connected by the molecular switch in order to set the molecular state, i.e., the bit value. On the other hand, the reading operation is current-based. In fact, if the molecule is in the off-state, then the NW in the lower level is almost floating and no correct voltage level can be sensed. Consequently, the reading operation is performed by selecting the bit to be read, then by measuring
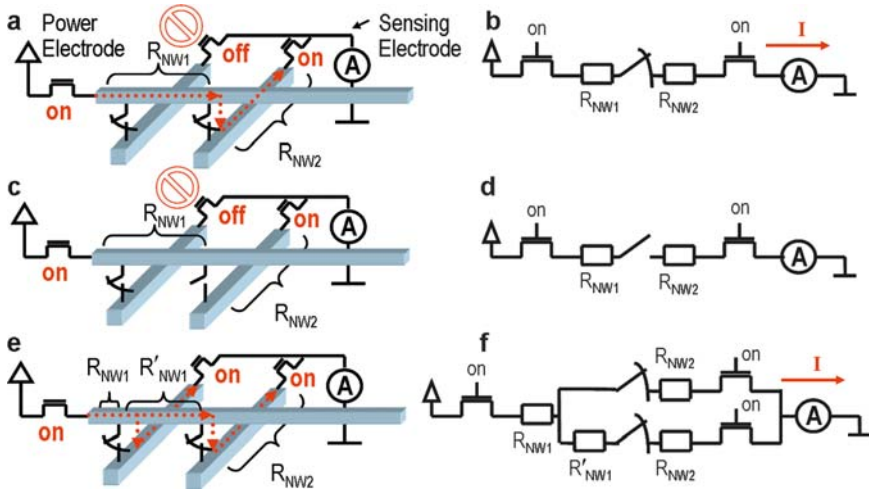
Fig. 3.7 Read operation in a 2-bit memory

the current through the sensing electrode (Fig. 3.7a–d). Thus, the current-based read operation in crossbar memories necessitates a thresholder as a part of the sense amplifier in order to set the limit between the logic values 0 and 1 and to translate them into logic levels that can be stored in the memory data register.

Variability-Induced Errors in Crossbar Memories

The threshold voltage variation was shown to cause defects in the decoder in such a way that by applying an address, any number of NWs could be activated instead of one single NW. Figure 3.7e, f shows an example of defective addressing in the second NW layer. Thus, the sense amplifier reads the superposition of the information stored in two bits. The thresholder cannot properly distinguish between the sensed signals resulting from the following cases: (a) one bit with a value of 1 and (b) the superposition of two bits whereby at least one of them has a value of 1. In such a situation, the read operation of the first bit yields a result depending on the state of the second bit, which causes coupling faults (CF) in the memory [7]. Considering the fact that decoder defects typically make two, three, or more NWs in each array active with the same address [8], the number of interdependant bits can be as large as 4 to 9 or even more, without necessarily having neighboring locations. This leads to the more critical pattern sensitivity faults [7].

In order to avoid complex and exhaustive PSF test procedures on the whole memory [9], one may try to resolve the PSF caused by the decoder defects before performing the conventional memory test. The thresholder can carry out this operation by checking the addresses of all NWs in every layer (after separating them) and

keeping only the addresses that activate one single NW. This procedure has a linear complexity with N, the number of NWs in a layer (where N^2 is the number of bits in the memory). While it represents an additional testing step, this testing procedure, which we call a *nanowire test*, obviates the necessity of an exhaustive PSF testing of the whole memory whose complexity is exponential with N^2. However, we expect that the molecular switches will also induce PSFs that we do not consider in this paper. Since only neighboring molecules are likely to interact with each other, one can assume neighborhood patterns for the PSFs caused by the molecules. Therefore, simplified PSF procedures having a linear complexity with N^2 can be applied [9].

Thresholder Optimization

For every address applied at the decoder, a validation signal is given by the thresholder indicating whether (a) a single NW is addressed or (b) no NW or more than one NW is addressed. The thresholder senses I_s, after a possible amplification, then it compares it to two reference values (I_0 and I_1 with $I_0 < I_1$). If the sensed current is smaller than I_0, then no NW is addressed. If the sensed current is larger than I_1, then at least two NWs are activated with the same address. If the sensed current is between the reference current levels, then only one NW is activated and the address is considered to be valid.

In our defect model we assume two sources of variability of the sensed current: (a) variation of the threshold voltages of the transistors in the decoder: we assume, as for conventional MOS devices [4], that V_T follows a normal distribution with known mean value V_T and standard deviation σ_T; (b) variation of the NW resistance in the memory array: in the mathematical model, we assume that this resistance is fixed. Then, we investigate the impact of its variation on the results.

By assuming that the V_Ts are stochastic variables and the NW resistance is fixed, we model I_s as a stochastic variable whose distribution parameters depend on the NW resistance. Thus, the calculation of the thresholder parameters I_0 and I_1 results from a stochastic optimization. Their optimal values are obtained by maximizing the probability that a correct address is detected (P_1: the conditional probability that I_s is between I_0 and I_1 given that only one NW is activated) and the probabilities that a defective address is identified as such (P_0 and P_2: the conditional probability that I_s is below I_0 or beyond I_1 given that no NW or more than one single NW is activated, respectively). Then, the probability that all three events happen simultaneously is given by $P_0 \times P_1 \times P_2$ (assuming that the considered events are independent). Consequently, we can define the error probability of the thresholder as $\varepsilon = 1 - P_0 \times P_1 \times P_2$. In order to optimize the design of the thresholder with the smallest error, we developed a model that calculates the optimal values of I_0 and I_1.

Fig. 3.8 Circuit model for a
NW under test
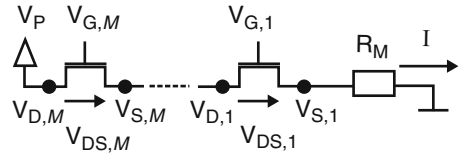
Stochastic and Perturbative Model

It is possible from a mathematical point of view to derive the exact stochastic distribution of the sensed current as a function of the stochastic distributions of the threshold voltages. However, the nonlinear effects, especially those affecting the short channel transistors, make an analytical solution impossible. Nevertheless, the model can be kept simple by linearizing the decoder circuit around the operating point. Each addressed NW was modeled as shown in Fig. 3.8. From this circuit model, the linearized relations between current and voltages can be derived.

On the other hand, the sensed current (either before or after linearization) can be modeled with different components. The first component is the useful signal, which is the current flowing through a correctly (i.e., intentionally) addressed NW. The second component is the defect-induced signal, which is represented by the current flowing through a badly (i.e., unintentionally) addressed NW. The third component is the intrinsic noise, which is the (subthreshold) current flowing through nonaddressed NWs.

We quantified these three stochastical components of the current and linearized them around the operating point. The obtained linearized stochastical current model was used in order to optimize the thresholder parameters I_0 and I_1 [10].

Thresholder Design

The design of the thresholder means in this context the optimization of its parameters I_0 and I_1 and the minimization of the error ε. The optimal value of I_1 is shown for different design and technology parameters in Fig. 3.9a. It is fairly robust and constant for a low memory resistive load. The value $I_0 = 0.8 \times I_1$ was found to be a good compromise between the cancellation of intrinsic noise and a signal that is useful for detection. The error of the decoder was assessed in Fig. 3.9b for different design and technology parameters. Adding redundancy to the decoder by increasing the code length is shown to be a very efficient way to minimize thresholder error. For instance, using 18 instead of 12 access transistors reduces the error by a factor of 60.
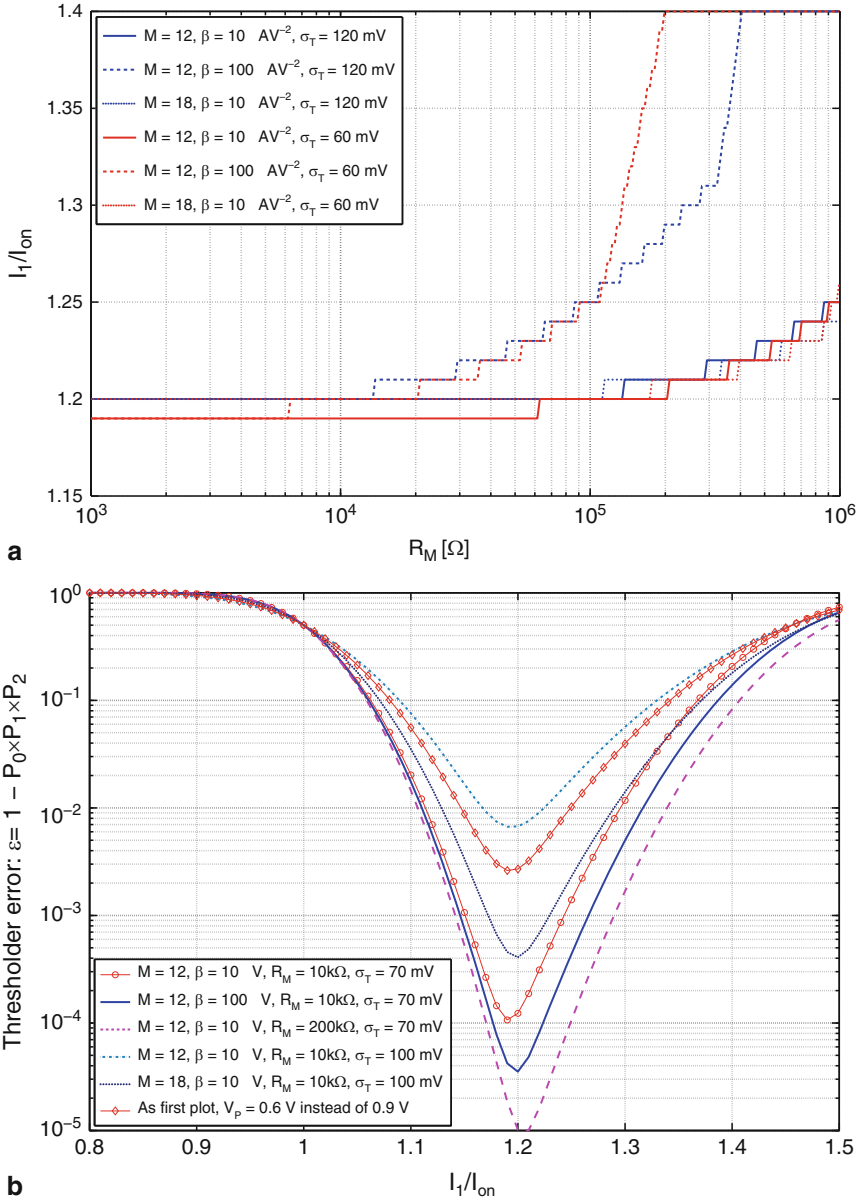
Fig. 3.9 **a** Optimal value of I_1. **b** Testing error

Summary

This section has shown the promises of crossbars when used as regular macros for computation and storage. The compact nature of crossbars leads to an efficient use of silicon areas. Nevertheless, the connection of these nanoelements to mesowires

requires specific addressing and decoding techniques. A complete design style has been described, including a procedure for testing.

Minimizing Local Delay Variations for Nanoscale CMOS Technologies

The technology scaling that has been the trend for decades is expected to continue at the same speed or possibly at a slightly slower pace for at least the next 10 years. The nano age has already begun (where typical feature dimensions are considered to be less than 100 nm). According to the ITRS roadmap, the operation frequency is expected to increase up to 12 GHz and a single chip could contain over 12 billion transistors in 2020 [11]. Future very-deep submicron and nanoelectronic fabrication technologies are expected to suffer from the dramatic dimensional scaling, which will strongly impact parameter variations.

The yield of low-voltage digital circuits is found to be sensitive to die-to-die (D2D) (interdie, global), and within-die (WID) (intradie, local) parameter variations in the manufacturing process. D2D variations act globally on the entire chip or on functional blocks, so that each device on one chip or in one block shows the same deviation. Interchip or interblock variations can be caused by systematic effects like process gradients over the wafer [12] with typical distances in the range of functional block sizes or above. Variations of the gate oxide thickness can be regarded as global variations. Sets of worst-case and best-case parameters are used during design verifications to mitigate the impact of global variations. The effects of WID variations are becoming more and more prominent with scaling and they have a direct influence on local gate delay variations. Numerous random factors such as statistical deviations of the doping concentration and imprecision of lithography lead to more pronounced delay variations for minimum transistor sizes [13, 14]. These factors are intrinsic since they cannot be eliminated by external control of conventional manufacturing. The increase of the path delay variations for smaller device dimensions and reduced supply voltages, as well as the dependence of path delay variations on the path length, are becoming more prominent with scaling. Circuits with a large number of critical paths and with a low logic depth are most sensitive to uncorrelated gate delay variations [15, 16].

As a first approximation, the gate delay of an inverter can be described by

$$T_{\text{gate}} \propto \frac{C_{\text{load}} V_{\text{DD}}}{\mu C_{\text{OX}} (W/L)(V_{\text{DD}} - V_{\text{th}})^{\alpha}}, \tag{3.1}$$

where μ is an effective mobility, V_{th} is the threshold voltage, C_{ox} is the gate capacitance per unit area, W and L are the transistor dimensions, and α is a power approximation coefficient that has a value between 2 and 1, with respect to the short channel effect [17]. A reduced supply voltage (V_{DD}) leads to an increased sensitivity $S_{T_{\text{gate}}}^{V_{\text{th}}}$ of gate delays to parameter variations [13]:

$$S_{T_{\text{gate}}}^{V_{\text{th}}} = \frac{V_{\text{th}}}{T_{\text{gate}}} \cdot \frac{\partial T_{\text{gate}}}{\partial V_{\text{th}}} = \frac{\alpha V_{\text{th}}}{V_{\text{DD}} - V_{\text{th}}}. \tag{3.2}$$

Furthermore, small transistor dimensions increase the effect of geometry-dependent parameter variations (variations in effective channel lengths). The impact of local variations may become significant since small-dimension devices operating at low supply voltages show an increased sensitivity to parameter variations. Therefore, for the design of low-voltage digital circuits the effect of intradie local parameter variations has to be minimized.

In this section, we present a novel technique that minimizes the impact of WID variations using redundancy applied only on critical parts of the circuit. First, the impact of WID parameter variations on gate and path delays is discussed with respect to critical path length. Moreover, a maximal critical path delay distribution has been derived. In the next section local delay variation minimization techniques, including majority and averaging gate, are presented. Last, a global analysis on the chip level is given followed by some conclusions.

Within-Die Parameter Variations and Their Effect on Gate and Path Delays

The two main parameters influencing gate delays are V_{th} and variations in effective channel length [18]. The impact of scaling on supply voltage and V_{th} mismatch variations is shown in Fig. 3.10. The data are obtained from [19]. With recent technology
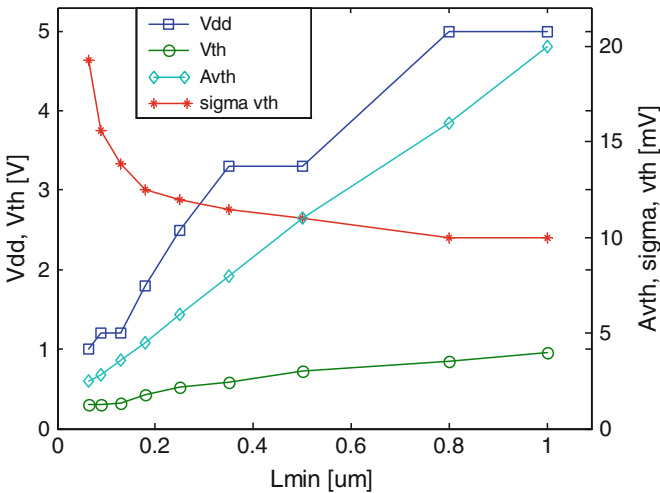
Fig. 3.10 Supply voltage (V_{DD}), threshold voltage (V_{th}), V_{th} mismatch coefficient $(A_{V_{\text{th}}})$, and standard deviation of V_{th} mismatch $(\sigma_{V_{\text{th}}})$ in different technology nodes

nodes (90, 65, and 45 nm) V_{th} mismatch ($\sigma_{V_{th}}$) value is significantly increased. The V_{th} impact of WID variations becomes very prominent with further reduction in the supply voltage.

Impact of the Critical Path Length and Gate Correlation on Delay Distribution

The critical path delay distributions resulting from D2D and WID parameter variations are calculated from D2D and WID statistical models using a SPICE-equivalent circuit simulator [20] with a 65-nm process file and a netlist containing modeled critical paths from a microprocessor.

To model a critical logic path in a design, a simple static inverter chain is used, containing a number n_{cp} of identical inverters, where each inverter drives four copies of itself yielding an FO4 load. FO4 is a common metric for the first-order analysis and evaluation of digital circuit performance in given process technologies [21]. The critical path delay $\left(T_{cp}\right)$ is calculated as

$$T_{cp} = n_{cp} T_{inv}, \tag{3.3}$$

where T_{inv} is an average propagation delay through an FO4 inverter.

The WID variations model represents systematic WID parameter variations by expressing the device-to-device correlation as a function of the distance between the devices. This correlation function, however, is significantly influenced by specific manufacturing capabilities. For future technology nodes relatively smaller gate-to-gate correlation factors are expected, considering that wire connections do not scale with the same factor as transistor sizes, which makes gate-to-gate transistor distances relatively larger with respect to previous technologies.

To simplify the analysis, two separate WID variations cases can be identified: (1) completely dependent gates (gate delay correlation equal to 1 and (2) completely independent gates (gate delay correlation equal to 0), which may be viewed as extreme conditions of systematic and random variations, respectively.

In the completely systematic case, the variations have the same impact on every element in a critical path:

$$\frac{\sigma_{T_{cp}}}{T_{cp}} = \frac{n_{cp}\sigma_{T_{inv}}}{n_{cp}T_{inv}} = \frac{\sigma_{T_{inv}}}{T_{inv}}, \tag{3.4}$$

where $\sigma_{T_{cp}}$ and $\sigma_{T_{inv}}$ are the standard deviations of the critical path delay distribution and the inverter gate delay distribution, respectively. This case, however, is not realistic in state-of-the-art technology nodes, where the distance between transistor gates relative to transistor sizes increases with respect to scaling.

In the case of completely random variations, however, the variations in the critical path delay are expected to have an averaging effect over the gates in the path [15]:

$$\frac{\sigma_{T_{cp}}}{T_{cp}} = \frac{\sigma_{T_{inv}}}{\sqrt{n_{cp}} T_{inv}}. \tag{3.5}$$

For completely random WID variations, the ratio of standard deviation to mean for the critical path delay distribution is inversely proportional to the square root of n_{cp} [13, 15]. This is a more realistic approximation of the circuit delay variations in state-of-the-art technology nodes than (3.4). To demonstrate this, the general case is compared to the simulated data. In the general case, the variance of the sum of n_{cp} identical random variables ($X_1, \ldots, X_n$) is given by [22]

$$\text{Var}\left(\sum_{i=1}^{n_{cp}} X_i\right) = \sum_{i=1}^{n_{cp}} \text{Var}(X_i) + 2 \sum_{i=1}^{n_{cp}-1} \sum_{j=i+1}^{n_{cp}} \sqrt{\text{Var}(X_i)\text{Var}(X_j)} \rho(X_i, X_j), \tag{3.6}$$

where $\rho(X_i, X_j)$ is a correlation factor between random variables X_i and X_j. If (3.6) is applied to gates in the critical path, $\text{Var}(X_i)$ becomes $\sigma_{T_{inv}}^2$ for every gate, $\text{Var}(\sum_{i=1}^{n_{cp}} X_i)$ becomes $\sigma_{T_{cp}}^2$, and $\rho(X_i, X_j)$ becomes $\rho_{i,j}$ – the correlation factor between the ith and jth gates in the critical path. With these substitutions (3.6) becomes

$$\sigma_{T_{cp}}^2 = n_{cp}\sigma_{T_{inv}}^2 + \sigma_{T_{inv}}^2 \times \sum_{i=1}^{n_{cp}-1} \sum_{j=i+1}^{n_{cp}} \rho_{i,j}. \tag{3.7}$$

After dividing (3.7) by T_{cp}, with respect to (3.3),

$$\frac{\sigma_{T_{cp}}}{T_{cp}} = \frac{\sigma_{T_{inv}}}{\sqrt{n_{cp}} T_{inv}} \sqrt{1 + \frac{2}{n_{cp}} \sum_{i=1}^{n_{cp}-1} \sum_{j=i+1}^{n_{cp}} \rho_{i,j}}. \tag{3.8}$$

If $\rho_{i,j} = 1$ for any pair of gates, then (3.7) becomes (3.4), otherwise for $\rho_{i,j} = 0$ for any pair of gates, (3.7) becomes (3.5). Due to the difference in their physical origins, variations of L and V_{th} exhibit different characteristics of correlation among transistors: lithography-induced variation of L is spatially correlated [23], while the V_{th} variation is mostly random due to dopant fluctuations [24]. For a small-size gate, a strong correlation is usually assumed to reduce the complexity of analysis. However, for a realistic circuit path that spans across larger distances, knowing the spatial correlation among gates is important for accurate statistical timing analysis [25]. A spatial correlation can be modeled as a linear function of distance [23]. Here we are using the following simplified model:

$$\rho_{i,j} = \begin{cases} \rho_0 \left(1 - \frac{j-i-1}{D}\right) & \text{for} \quad j - i \leq D \\ 0 & \text{for} \quad j - i > D \end{cases}, \tag{3.9}$$

Fig. 3.11 Spatial correlation modeled as a linear function of distance
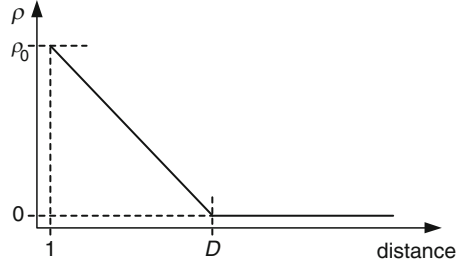
Table 3.1 Ratio of standard deviation to mean delay for WID and D2D variations and different critical path lengths (n_{cp})

$\sigma_{T_{cp}}/\mu_{T_{cp}}$ (%)	$n_{cp} = 1$	$n_{cp} = 6$	$n_{cp} = 8$	$n_{cp} = 10$
WID	11.48	5.44	4.78	4.44
D2D	7.96	7.94	8.12	8.22
	$n_{cp} = 12$	$n_{cp} = 16$	$n_{cp} = 20$	$n_{cp} = 24$
WID	4.08	3.62	3.32	2.98
D2D	8.29	8.38	8.43	8.46

where ρ_0 is the correlation factor between neighboring gates ($j = i + 1$) and D is the maximum distance between gates where correlation effects are still present. A spatial correlation model from (3.9) is illustrated in Fig. 3.11.

WID and D2D variations are acquired from Monte Carlo SPICE simulations using mismatch and parameter models for commercial 65-nm technology. Table 3.1 summarizes the statistical simulations for different critical path lengths providing the ratio of the standard deviation to the mean delay $\left(\sigma_{T_{cp}}/\mu_{T_{cp}}\right)$ corresponding to WID and D2D variations.

An almost constant value for the standard deviation of D2D variations for different critical path lengths confirms that D2D variations can be assumed as purely systematic.

Substituting (3.9) into (3.8) with assumption that $n_{cp} \leq D + 2$ yields

$$\rho_0 = \frac{\left(\frac{\sigma_{T_{cp}} T_{inv}}{T_{cp}\sigma_{T_{inv}}}\right)^2 n_{cp} - 1}{(n_{cp} - 1)\left(1 - \frac{n_{cp}-2}{3D}\right)}. \tag{3.10}$$

Eqation (3.10) has two parameters, ρ_0 and D, that need to be estimated for the given technology. Linear fitting is applied to (3.10) using the values for WID variations from Table 3.1. The following values are acquired (95% confidence parameter interval in brackets): $\rho_0 = 0.08$ (0.078–0.082) and $D = 8$ (7.3–8.3). Such a low correlation factor even for neighboring gates justifies the assumption that WID variations are mostly random.

The mean delay has been taken as a nominal critical path delay $T_{cp,nom}$. The WID and D2D nominal critical path standard deviations are $\sigma_{T_{cp},WID}$ and $\sigma_{T_{cp},D2D}$, respectively. The critical path delay probability density functions (PDFs) resulting from WID and D2D parameter variations are modeled as normal distributions [15, 18, 26] in (3.11) and (3.12), respectively:

$$f_{T_{cp,nom},WID} = N(T_{cp,nom}, \sigma_{T_{cp},WID}^2), \tag{3.11}$$

$$f_{T_{cp,nom},D2D} = N(T_{cp,nom}, \sigma_{T_{cp},D2D}^2). \tag{3.12}$$

Impact of Within-Die Variations on the Maximum Critical Path Delay Distribution

Following (3.11) and the procedure from [15] the probability of one critical path satisfying a specified maximum delay T_{max} is given with the cumulative distribution function (CDF) in (3.13):

$$F_{T_{cp,nom},WID}(T_{max}) = P_{T_{cp,nom},WID}(t < T_{max}) = \int_0^{T_{max}} f_{T_{cp,nom},WID} dt. \tag{3.13}$$

A large chip, however, contains *many* critical paths, all of which must satisfy the worst-case delay constraint [13–16, 27]. For completely dependent paths (path delay correlation equal to 1), the PDF given in (3.11) is valid for all the paths to model the worst-case delay. However, thanks to a very low gate correlation factor, as shown in the previous section, all the paths can be assumed to be independent (path delay correlation equal to 0). Assuming a number N_{cp} of independent critical paths for the entire chip [13], the probability that the whole chip satisfies the worst-case delay is given with a CDF for the whole chip $\left(F_{chip,WID}\right)$ in (3.14):

$$F_{chip,WID}(T_{max}) = P_{chip,WID}(t < T_{max}) = \left(F_{T_{cp,nom},WID}(T_{max})\right)^{N_{cp}}. \tag{3.14}$$

The chip's WID maximum critical path delay PDF is then calculated following [15] by taking the derivative of CDF with respect to T_{max}:

$$\begin{aligned} f_{chip,WID}(T_{max}) &= \frac{dF_{chip,WID}(T_{max})}{dT_{max}} \\ &= N_{cp} f_{T_{cp,nom},WID}(T_{max}) \left(F_{T_{cp,nom},WID}(T_{max})\right)^{N_{cp}-1}. \end{aligned} \tag{3.15}$$
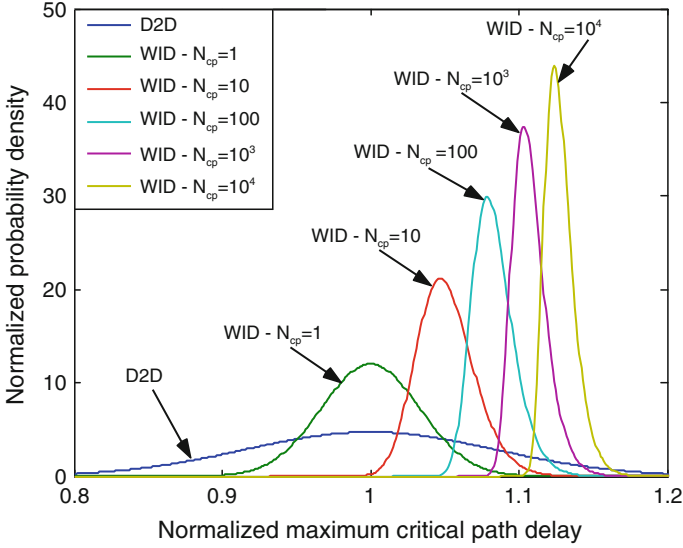
Fig. 3.12 Within-die (WID) maximum critical path delay distribution for different values of N_{cp} and die-to-die (D2D) critical path delay distribution

The chip's maximum critical path delay PDF (3.15) is illustrated in Fig. 3.12
By increasing N_{cp},

- The mean of $f_{chip,WID}$ increases, since the *slowest* critical path limits the chip's overall performance and the probability of a longer cycle time increases;
- The standard deviation of $f_{chip,WID}$ decreases and becomes relatively small compared to the standard deviation of $f_{T_{cp},nom,WID}$, e.g., for $N_{cp} = 10^4$, $\sigma_{chip,WID} = 0.3\sigma_{T_{cp},nom,WID}$, and $\sigma_{chip,WID} = 0.12\sigma_{T_{cp},nom,D2D}$;
- $f_{chip,WID}$ becomes less sensitive to further increases of N_{cp}; qualitatively, this means that an increase of N_{cp} from 1 to 10 has a greater effect on the mean and variance of the WID distribution than an increase from 10^3 to 10^4;
- The shape of $f_{chip,WID}$ becomes less symmetrical and more positively skewed.

As the number of transistors per chip increases and the number of average gate delays per critical path is reduced [28], N_{cp} is expected to increase for each technology generation and further reduce the sensitivity of maximum critical delay to N_{cp}.

Local Delay Variation Minimization Techniques

Redundancy can be used to reduce the delay variance of critical paths and thereby increase the overall circuit speed. The general principle consists in replicating R
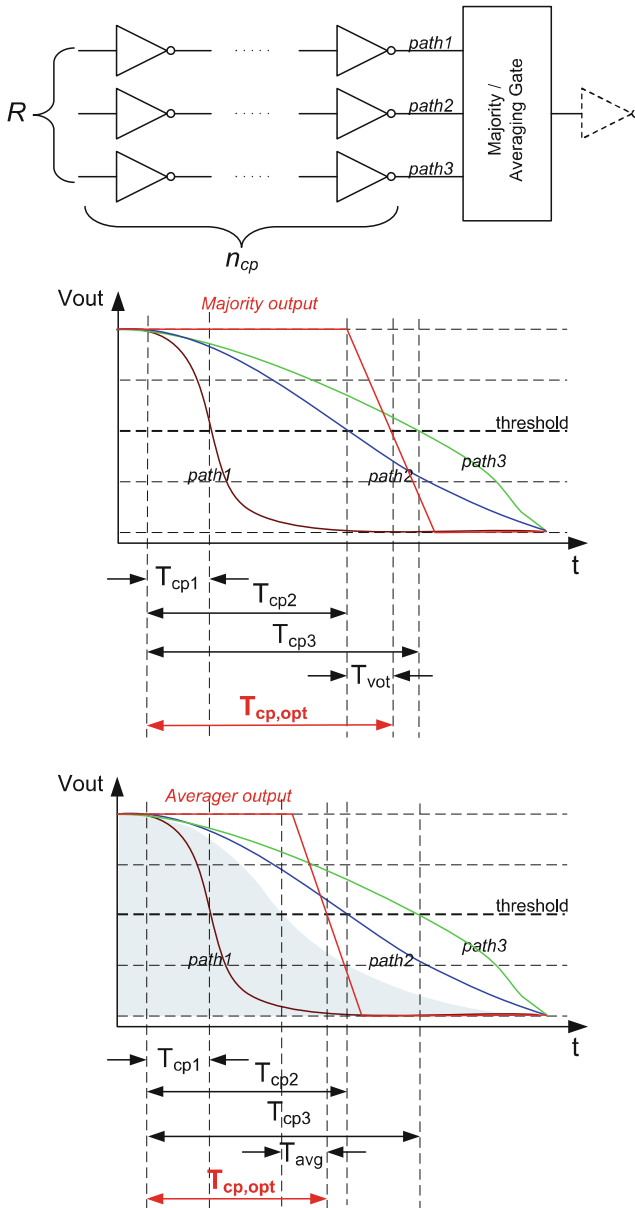
Fig. 3.13 Schematic of circuit for local delay minimization and signal timing diagrams for majority and averager gate realizations

times a critical path and evaluating critical path outputs using a function that delivers the output with reduced time delay variance. Functions that satisfy this condition, and that are considered in this paper, are realized using majority $(R + 1)/2$ out of R and averaging the voters. The principle of operation is illustrated in Fig. 3.13. For

a majority gate, a voter circuit switches when a path with a median value of time delay reaches a selected threshold voltage. For an averaging gate, a circuit switches when the average value of the outputs of all the paths reaches a selected threshold voltage.

The proposed technique can also be used to support recovering the correct operation, which is disrupted by other sources of variations and signal-integrity issues such as those caused by crosstalk aggressor signals.

Majority Gate Delay Variation Minimization

A majority gate performs a median function of its inputs. The median of a statistical distribution with CDF $D(x)$ is the value of x such that $D(x) = 1/2$. For a symmetric distribution, it is therefore equal to the mean. Having the order statistics $Y_1 = \min_j X_j$, Y_2, ..., Y_{R-1}, $Y_R = \max_j X_j$ gives the statistical median of the random sample (3.16) [22]:

$$\tilde{x} \equiv \begin{cases} Y_{(R+1)/2} & \text{if } R \text{ is odd} \\ \frac{1}{x}\left(Y_{R/2} + Y_{1+R/2}\right) & \text{if } R \text{ is even} \end{cases} \tag{3.16}$$

Only odd values of R are used since the majority gate can only support an *odd* number of inputs.

Taking into consideration that random variables X_1, ..., X_R follow a normal distribution with mean μ and standard deviation σ, $\tilde{x}$ also has a normal distribution with the mean (μ_{med}) equal to μ and standard deviation (σ_{med}) that has an asymptotic upper bound (for $R \to \infty$) equal to $\sqrt{\frac{\pi}{2}}\frac{\sigma}{\sqrt{R}}$ [29]. There are various estimations for the standard deviation [29–31]; nevertheless, we consider that a much simpler formula given in (3.17) provides the most appropriate estimation for all cases where a low value of R ($R < 10$) is considered. In Table 3.2 the exact values obtained from [29], approximation values from [30, 31], and the values obtained using our approximation (3.17) are given for $\sigma = 1$:

Table 3.2 Different estimations of standard deviation of median function for various redundancy factors (R)

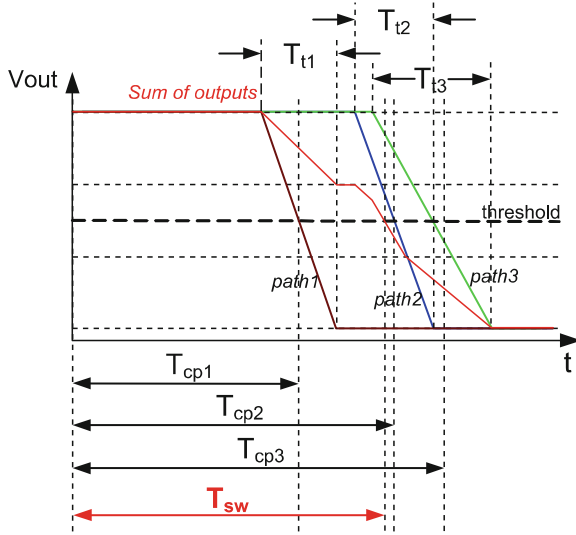|	Sample size				
	$R = 3$	$R = 5$	$R = 7$	$R = 9$	$R = 11$
Exact SD	0.6692	0.5356	0.4587	0.4075	0.3704
Our app. of SD	0.6699	0.5344	0.4576	0.4066	0.3696
SD from [30]	0.6202	0.5134	0.4466	0.4001	0.3654
SD from [31]	0.6637	0.5337	0.4580	0.4072	0.3701

Fig. 3.14 Linearized signals for critical path delays and a switching point delay

$$\sigma_{\mathrm{med}} = \sqrt{\frac{\pi}{2R+1}}\sigma. \tag{3.17}$$

We note that in a real implementation, the additional majority block may also add to the overall delay variation; however, this will not be taken into consideration.

Minimization of Averaging Gate Delay Variations

The averaging gate switches when the arithmetic average of its input levels reaches the threshold ($V_{\mathrm{DD}}/2$). The switching point depends on the path propagation delay ($T_{\mathrm{cp},i}$) and the last gate transition delay ($T_{\mathrm{t},i}$) for each path ($i = 1, \ldots, R$). The expression for the switching point delay for linearized signals (Fig. 3.14) is given in (3.18), where m is the number of inputs whose transitions are completed before the switching point and n is the number of inputs whose transitions start after the switching point, as illustrated in Fig. 3.14, where $m = 1, n = 0$, and $R = 3$:

$$T_{\mathrm{sw}} = \frac{\frac{(m-n)}{2} + \sum\limits_{i=n+1}^{R-m} \frac{T_{\mathrm{cp},i}}{T_{\mathrm{t},i}}}{\sum\limits_{i=n+1}^{R-m} \frac{1}{T_{\mathrm{t},i}}}. \tag{3.18}$$

Taking into consideration that $T_{\mathrm{cp},i}$ and $T_{\mathrm{t},i}$ for every $i = 1, \ldots, R$ are random variables following a normal distribution with mean $\mu_{T_{\mathrm{cp}}}$ and $\mu_{T_{\mathrm{t}}}$ and standard deviation $\sigma_{T_{\mathrm{cp}}}$ and $\sigma_{T_{\mathrm{t}}}$ respectively, T_{sw} also has a normal distribution with a mean (μ_{sw}) that

Table 3.3 Ratio of standard deviation over mean delay for different redundancy factors (R) and critical path lengths (n_{cp}) for majority and averaging gate realizations

%	$\dfrac{\sigma_{T_{cp}}}{\mu_{T_{cp}}}$	$R = 3$		$R = 5$		$R = 7$	
		$\dfrac{\sigma_{med}}{\mu_{med}}$	$\dfrac{\sigma_{sw}}{\mu_{sw}}$	$\dfrac{\sigma_{med}}{\mu_{med}}$	$\dfrac{\sigma_{sw}}{\mu_{sw}}$	$\dfrac{\sigma_{med}}{\mu_{med}}$	$\dfrac{\sigma_{sw}}{\mu_{sw}}$
$n_{cp} = 6$	5.44	3.64	3.17	2.91	2.46	2.49	2.09
$n_{cp} = 8$	4.78	3.20	2.81	2.56	2.19	2.19	1.85
$n_{cp} = 10$	4.44	2.97	2.63	2.37	2.05	2.03	1.73
$n_{cp} = 12$	4.08	2.73	2.44	2.18	1.89	1.87	1.60
$n_{cp} = 16$	3.62	2.43	2.21	1.93	1.71	1.66	1.45
$n_{cp} = 20$	3.32	2.22	2.05	1.78	1.59	1.52	1.34
$n_{cp} = 24$	2.98	1.99	1.85	1.59	1.44	1.36	1.22

is approximated with $\mu_{T_{cp}}$ and a standard deviation (σ_{sw}) that has an upper bound equal to

$$\sigma_{sw} = \frac{\sigma_{T_{cp}}}{\sqrt{R}}. \tag{3.19}$$

The exact value of the mean and standard deviations for a normal distribution of T_{sw} cannot be derived analytically. Therefore, values representing the ratio of the standard deviation to the mean delay obtained by Monte Carlo simulations are shown in Table 3.3 for different critical path lengths (n_{cp}) and different redundancy factors (R). The values related to the majority gate are also acquired from Monte Carlo simulations for the sake of comparison, and we observe that they comply with the approximation given in (3.17). The values related to the averaging gate are always close to the upper bound (3.19), thereby demonstrating that the averaging gate performs an optimal minimization of the standard deviation of critical paths delay. The mean and standard deviation values for a critical path and a transition time for the last gate in the path, $\mu_{T_{cp}}$, μ_{T_t}, $\sigma_{T_{cp}}$, and σ_{T_t} respectively, are given for a 65-nm fabrication technology.

The averaging function has two prominent advantages over the majority function, namely, it enables better minimization of the standard deviation of the output delay; moreover, the averaging function has no restriction on redundancy factor R, whereas the majority function demands odd values only.

Optimization Method Using the Proposed Techniques for Variation Minimization

The effect of reducing the standard deviation of critical paths by replicating a critical path and inserting a decision gate is exploitable only when the reduction in the standard deviation is large enough to compensate for the additional delay of the added decision gate. As previously shown, longer critical path lengths are beneficial

for the minimization of intradie delay variations thanks to the averaging effect over the gates in the critical path.

An architecture that has been optimized in terms of delay variations is considered in what follows as a multithreaded processor with $N_{cp} = 10^4$ critical paths. N_{cp} is estimated by assuming that the ratio between the number of independent critical paths and the number of transistors per chip remains approximately constant in different technologies [15]. The exact number is not significant since the sensitivity of the maximum critical path distribution to N_{cp} is small for large values of N_{cp} $(N_{cp} \sim 10^3–10^4)$, as demonstrated. As shown in [32], an optimal critical path length for power/performance in a multithreaded processor is in a range of 20 to 24 equivalent FO4 inverter delays. Therefore, cases where $n_{cp} = 20$ and 24 are taken into consideration in what follows. The distribution for a normalized maximum critical path delay is shown in Fig. 3.15, considering cases with and without optimization. The optimization is performed with *averaging* gates and a redundancy factor $R = 3$.

Figure 3.15 clearly shows that longer critical paths have a reduced mean value of maximum critical path delay and are therefore less sensitive to WID variations. A reduced standard deviation also reduces the mean value of the maximum critical path delay. However, adding a majority/averaging gate increases it. For the given technology, any path where $n_{cp} \geq 24$ has a smaller or equal normalized mean value of the maximum critical path delay when the proposed optimization technique is applied, compared to a default case (a case without any delay variations minimization technique) $\left(\mu_{T_{cp,max,opt}}/T_{cp,nom} = \mu_{T_{cp,max}}/T_{cp,nom} = 1.115\right)$. However, the standard deviation is smaller $\left(\sigma_{T_{cp,max,opt}}/T_{cp,nom} = 0.58\% < \sigma_{T_{cp,max}}/T_{cp,nom} = 0.91\%\right)$, which can give better yield. This is investigated in detail in the section "Adaptive
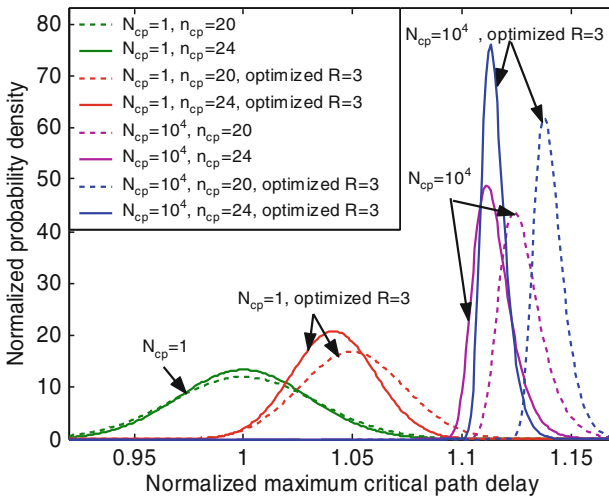
Fig. 3.15 Within-die (WID) maximum critical path delay distribution for $n_{cp} = 20$ and $n_{cp} = 24$ without and with optimization ($R = 3$)
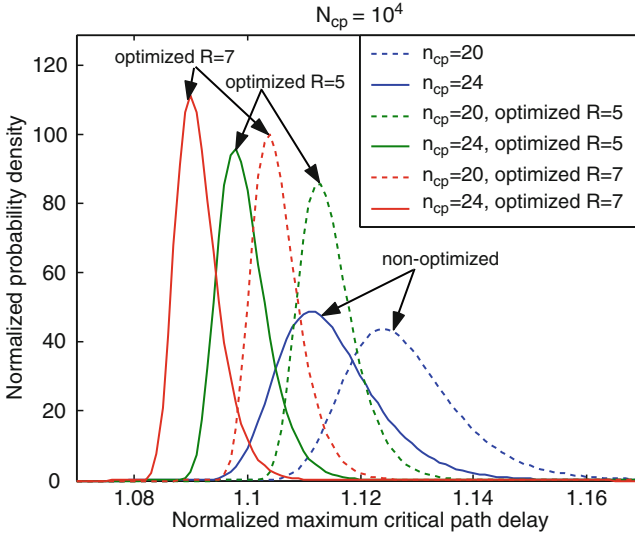
Fig. 3.16 Within-die (WID) maximum critical path delay distribution for $n_{cp} = 20$ and $n_{cp} = 24$ without and with optimization $\left(N_{cp} = 10,000, R = 5\right)$

V_{gs}: A Novel Technique for Controlling the Power and Delay of Logic Gates in a Sub-VT Regime," where D2D variations are also considered.

When $n_{cp} = 20$, the normalized mean value of the maximum critical path delay is larger if the proposed optimization technique is applied ($\mu_{T_{cp,max,opt}}/T_{cp,nom} = 1.131 > \mu_{T_{cp,max}}/T_{cp,nom} = 1.128$). In order to achieve a noticeable improvement in optimization performance, the redundancy factor of critical paths needs to be increased. The distribution of a normalized maximum critical path delay is shown in Fig. 3.16, which considers the cases where optimization uses redundancy factors of 5 and 7. The improvement in the mean value of maximum critical path delay is

- 1.5 and 1.35% compared to the default case (without optimization) for $n_{cp} = 24$ and $n_{cp} = 20$, respectively, when a redundancy factor of 5 is chosen;
- 2.3 and 2.2% for $n_{cp} = 24$ and $n_{cp} = 20$, respectively, when a redundancy factor of 7 is chosen.

Maximum Critical Path Delay Distribution with Combined Die-to-Die and Within-Die Variations

The maximum critical path delay distribution of the chip considering both D2D and WID types of variations can be obtained by combining the individual D2D and WID distributions, following an adapted version of the procedure presented in [15]. The maximum critical path delay is calculated as

$$T_{\text{cp,max}} = T_{\text{cp,nom}} + \Delta T_{\text{cp,WID}} + \Delta T_{\text{cp,D2D}} = T_{\text{cp,WID}} + \Delta T_{\text{cp,D2D}}, \qquad (3.20)$$

where $\Delta T_{\text{cp,D2D}}$ and $\Delta T_{\text{cp,WID}}$ are the deviations in the nominal critical path delay resulting from D2D and WID variations, respectively. The maximum critical path delay density function resulting from both D2D and WID variations is derived using convolution according to (3.20):

$$f_{T_{\text{cp,max}}} = f_{T_{\text{cp,nom}},\text{WID}} * f_{\Delta T_{\text{cp,max}},\text{D2D}}, \qquad (3.21)$$

where $f_{T_{\text{cp,nom}},\text{WID}}$ is as given in (3.10) and $f_{\Delta T_{\text{cp,max}},\text{D2D}}$ is a distribution resulting from shifting in the negative direction the D2D distribution (3.11) by $T_{\text{cp,nom}}$ expressed as

$$f_{T_{\text{cp,nom}},\text{D2D}} = N(0, \sigma^2_{T_{\text{cp,D2D}}}). \qquad (3.22)$$

Since the WID distribution has a significantly smaller standard deviation compared to the D2D distribution, which is further reduced when N_{cp} increases, the WID distribution can be legitimately approximated with an impulse function. As the D2D and WID distributions are statistically combined through (3.21), the resulting distribution has a mean equal to that of the WID distribution and a variance predominantly resulting from the D2D distribution. Thus, WID variations determine the mean of the maximum critical path delay distribution, and D2D variations determine the variance. With respect to this observation, any improvement in the mean value of the maximum critical path delay WID variations causes a direct improvement in the maximum critical path delay of any fabricated chip. A tradeoff can be considered between the redundancy level involved, which represents additional area/power, and a reduction of the maximum critical path delay, which actually means possible increased operating frequency. Assuming that each gate in the critical path consists of six transistors, $N_{\text{cp}} = 10^4$, and that a whole chip has 4×10^8 transistors according to ITRS [10], replicating each critical path 3, 5, and 7 times causes an overhead of 0.8, 1.5, and 2.3% respectively.

The maximum critical path delay distribution, including both WID and D2D variations, is analyzed for three technologies: one 65-nm commercial technology and two future technology nodes 45 and 32 nm. Both the 45- and 32-nm technologies are hypothetical, and the parameters have been selected to anticipate a realistic set [10, 18, 33–36].

Values for the ratio of the standard deviation to the mean delay (σ/μ) for WID, D2D variations, and for critical path length (n_{cp}) are given in Table 3.4. Standard deviation values are estimated by combining $\sigma_{V_{\text{th}}}$ and σ_L according to [18, 33–36], and n_{cp} is estimated as presented in [32]. $N_{\text{cp}} = 10^4$; finally, $R = 5$ in all calculations.
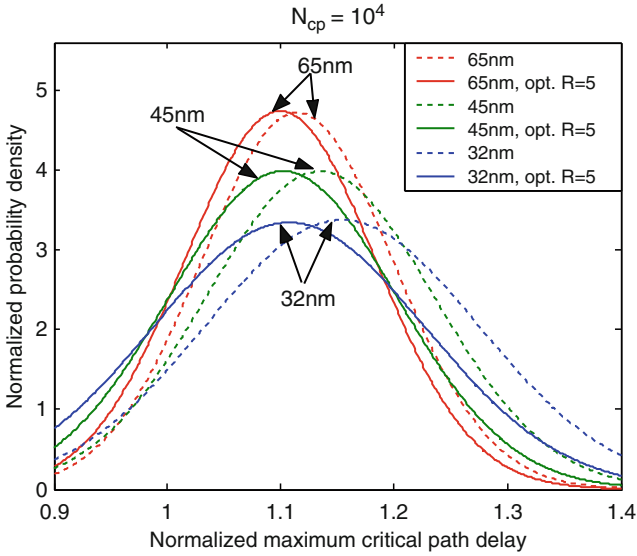
The maximum critical path delay distribution of the chip for each technology node is depicted in Fig. 3.17 and is calculated using the values from Table 3.4.

The improvement of the mean value of maximum critical path delay is equal to 1.5, 3.2, and 4.3% for the 65-, 45-, and 32-nm technologies, respectively. The larger improvement than overhead due to redundancy for future technologies suggests

Table 3.4 Values of σ/μ for gate and critical path WID and D2D variations and n_{cp} for different technologies

Technology (nm)	D2D (%): σ/μ	WID (%): σ_{gate}/μ_{gate}	WID (%): $\sigma_{T_{cp}}/\mu_{T_{cp}}$	n_{cp}
65	8.4	11.5	3	24
45	10	15	3.5	30
32	12	20	4	36

Fig. 3.17 Maximum critical path delay distribution for combined D2D and WID variations for different technologies ($N_{cp} = 10{,}000$, $R = 5$)

that there will be an optimal design point where this technique for delay variation minimization can be successfully applied.

Summary

This section studies the benefit of using the averaging technique to limit the impact of delay variations. Another popular fault-tolerant technique (majority voting) is used for the sake of comparison. The averaging technique has two prominent advantages over the majority technique, namely, it enables better minimization of the standard deviation of the output delay and, moreover, when using the averaging mechanism, the redundancy factor can be as low as 2. The technique is intended to reduce the effects of intradie variations using redundancy applied only on critical

segments (critical paths). This way, the proposed technique can be optimally used in the design of large synchronous digital systems.

The studies have shown that the technique can be already applied for a 65-nm CMOS technology process. However, the real benefit is expected for future nano-scale CMOS technologies such as 45- and 32-nm nodes where an optimal point has been shown to exist in the speed vs. area/power tradeoff.

Adaptive V_{gs}: A Novel Technique for Controlling the Power and Delay of Logic Gates in a Sub-VT Regime

The primary motivation for ultra-low-voltage operation is to reduce energy [37]. Analysis in [38] and chip measurements in [39] showed that minimum energy per operation occurs in the sub-VT region. An 8-T sub-VT SRAM in 65-nm CMOS is demonstrated in [40], and more complex sub-VT processors are appearing [41]. In a sub-VT region, with further supply voltage (VDD) scaling, gate delay and clock period increase exponentially, dynamic energy per operation decreases in a quadratic manner, but leakage power accumulates over the longer clock period, and finally leakage energy per operation exceeds the dynamic energy and causes the minimum energy point.

References [39–41] have used static CMOS gates. These gates continue to function in the sub-VT region and have a great potential for saving energy, but they face many challenges including temperature sensitivity, PV [42], and process imbalance [43].

There are four main sources of leakage current in digital circuits, i.e., reverse diode current (I_{diode}), sub-VT current (I_{SUB}), gate leakage (I_G), and GIDL current (I_{GIDL}). Usually I_{diode} is negligible. However, since the I_G and I_{GIDL} currents are exponential functions of supply voltage while I_{SUB} is a weak function of VDD (through DIBL effect), I_{SUB} is the dominant leakage term for the weak-inversion and sub-VT regions. There is no theoretical solution for I_{SUB} leakage. A commonly used expression for sub-VT current is given by [44]:

$$I_{sub} = \mu_0 C_{ox} \frac{W}{L}(n-1)V_{th}^2 \times e^{\frac{V_{gs}-V_{T0}-\gamma V_{sb}+\eta V_{ds}}{nV_{th}}} \times (1 - e^{\frac{-V_{ds}}{V_{th}}}), \qquad (3.23)$$

where V_{th} is the thermal voltage (kT/q), n the sub-VT slop factor, η DIBL coefficient, and γ the body effect coefficient.

In this work, simulations are done using the transistor model card provided by the foundry for a 65-nm low-power process. Our simulations in 65-nm and chip measurement in [45] in 180 nm shows that, approximately independent of technology, leakage current (and speed) increases $2 \times$ per 15°C temperature increase. One way to alleviate this strong temperature dependence is to change the frequency of operation as a function of temperature [45]. But this is not acceptable for most digital applications. Dynamic power is not a function of temperature. So I_{SUB} is the main source of power consumption at high temperatures.

Because of the exponential dependency of transistor current on the parameter variations, it is clear that any logic style designed for sub-VT operation should work sizing independently. Contention current between pull-up and pull-down networks (PUN and PDN) fails to work in the presence of intra-die variations.

Available Variation Compensating Techniques

In the sub-VT region, all kinds of parameter variations like TOX, channel length (L), V_T, and temperature variations show similar behavior. When they increase the I_{SUB}, they decrease the delay, and vice versa. So all of the circuit techniques that can change the power-delay tradeoff at runtime can be used for compensating all of these variations at the same time and independently of the variation source.

For over-100-nm technologies, adaptive body biasing (ABB) is a good technique for compensating the variations [46, 47]. Both forward body biasing (FBB) and reverse body biasing (RBB) can be used in a sub-VT regime. Since ABB changes the V_T value directly, it can control both leakage and delay. Also, the overhead of this technique is small. This technique is very good but has three important weaknesses. First, using ABB for compensating intradie variations of NMOS transistors need triple-well technology. Second, the increased SCE due to scaling decreases the body factor of bulk-CMOS drastically. According to the foundry data, with 65-nm technology, RBB can change the V_T value effectively to less than 60 mV. This amount causes an approx. fourfold change in delay and power in the sub-VT region, which is much less than PV and temperature effects. Also, much more than a 60-mV change in V_T is required for compensating slow and fast corners. And third, as we will discuss in the section "Minimizing Local Delay Variations for Nanoscale CMOS Technologies," the body factor is almost 0 in emerging multigate devices, which are promising candidates for future electronics.

Both power and gate delay strongly depend on VDD. So adjusting the VDD can be another technique for compensating variations. This can be done by using a variable supply voltage (Var-VDD). But this method does not provide a good control on I_{SUB} because this leakage current is a weak function of VDD through a second-order effect, i.e., DIBL.

Body Effect in Emerging Multigate Devices

For sub-50-nm technology, various nonplanar device structures have been explored for better SCE immunity and gate electrostatic control of the channel surface potential. Among the many approaches, double-gated FinFET, trigated, Π-gated, Ω-gated, NW body, and GAA MOSFETs have attracted much attention.

The GAA structure in which the gate oxide and the gate electrodes wrap around the channel region exhibit excellent electrostatic control, e.g., near ideal sub-VT

slope (S) (<63 mV/dec), very low DIBL (<10 mV/V), and with an I_{ON}/I_{OFF} ratio of <10^6 [48].

The body effect coefficient or body factor, γ, represents the dependency of the V_T on the back-gate bias. For any devices, one can write $\gamma = -d(V_T)/d(V_b)$, where V_b is the back-gate voltage. The body factor can be calculated using a simple capacitive equivalent circuit and the relationship $\gamma \propto C_{CH-B}/C_{G-CH}$, where C_{CH-B} is the capacitance between the channel surface and back -gate (or the substrate) and is the capacitance between the gate electrode and the channel [49].

In [49] and [50] it is shown that by increasing the gate electrode control in these devices from single-gate → double-gate → trigate → Π-gate → GAA, DIBL, V_T roll-off by channel-length, and S-factor decrease. All of these reductions are considered as good effect for logic circuits and cause better performance and less power consumption, but unfortunately γ also decreases in the same direction.

S-factor is very important for sub-VT operation. For example, in 65-nm planar-CMOS S-factor is <93 mV/dec, but in the GAA devices of [48] and multigate FETs of [51] it is <63 mV/dec. So in the sub-VT region for exactly equal device performance: I_{ON}, I_{OFF}, and $I_{ON}/I_{OFF} = 10^n$, we have $VDD_{GAA}/VDD_{PLANAR} \approx$ (63 mV × n) / (93 mV × n). At the minimum energy point, the ratio of $Energy_{Leakage}/Energy_{Dynamic} \approx 1/3$. In logic, circuits for given I_{OFF} and I_{ON}, delay/$Energy_{Leakage}/Energy_{Dynamic}$, are linear/linear/quadratic functions of voltage swing (VDD), respectively. As a result, in the sub-VT regime using GAA devices instead of planar-CMOS causes approx. 1.9× less energy per operation and approx. 1.5× better performance (speed). This means that GAAs are the best choice for weak inversion and sub-VT operation.

It is interesting to note that from the circuit design point of view, the main problem of FinFETs is the low mobility in the saturation region because in the sidewalls the crystal orientation is (110) [51]. But for sub-VT operation μ_0 is not important because the current is dominated by diffusion. Variation of μ_0 in (1) can be compensated by V_{T0} adjustment.

In summary, in multigate devices, the body factor is much smaller than in single-gate devices because of the enhanced coupling between gate and channel and because the lateral gates shield the device from the electric field from the back gate [49]. Measurements in [48] show that in GAA devices the body factor is exactly 0.

In addition, the study in [52] and measurements in [53] show that the leakage current in all of these multigate devices is very PV and temperature sensitive. So we need to find new compensation techniques as replacements for ABB.

Proposed Adaptive V_{gs} Technique

Today's fabrication technologies offer the feasibility of using V_T-low, V_T-nominal, and V_T-high on the same die. The difference between V_T values ($\Delta V_T = V_T$-nominal - V_T-low $\approx V_T$-high - V_T-nominal) is usually 70 to 100 mV. So for sub-VT

operations V_T-low devices are about ten times faster (and leakier) than V_T-nominal devices.

Among several available static logic styles, SCMOS and PTL are more popular. SCMOS is the simplest and most robust style and is especially good for designing simple functions like NAND and NOR. Designing PTL is more complicated but it has better performance than SCMOS in designing some complex functions like MUX and XOR. The best choice is using mixed PTL/SCMOS gates [54].

The AVGS technique requires four supply rails, two ground rails (ground and $0 + \Delta V_n$), and two power rails (VVDD and VVDD $- \Delta V_p$). VVDD and ground are the conventional rails. ΔV_n and ΔV_p are the new circuit parameters and on the order of a few tens of millivolts. In "Results and Discussions", we will propose a multioutput bulk-converter switching power supply that can generate all of these supply rails by using only one inductor.

AVGS–SCMOS

Figure 3.18a shows the proposed AVGS–SCMOS technique for a single-stage logic structure. All the input signals and the output signal F are full-swing ($0 \rightarrow$ VVDD). But the PDN and PUN are connected to $0 + \Delta V_n$ and VVDD $- \Delta V_p$, respectively. As a result, when the PDN is on, NMOS transistors have gate-source voltage $V_{gs} =$ VVDD$-\Delta V_n$, and when PDN is off $V_{gs} = -\Delta V_n$. Similarly for PUN, in the on state, the transistor drive voltage is $V_{sg} =$ VVDD$-\Delta V_p$, and in the off state $V_{sg} = -\Delta V_p$. In both on and off states, the transistor current changes exponentially according to (1). So $\Delta V_{n/p}$ can strongly control the leakage power and delay of the gate exactly the same as if we changed the transistor's threshold voltage.

But one needs a full-swing output signal for driving the next logic gate stages. The proposed 6-T level converting inverter (LCI) is also shown in Fig. 3.18a. This inverter generates the full-swing output signal F. When Do is high (VVDD$-\Delta V_p$), Mn0 and Mn1 are on, X_n and F are 0 V, and Mp2 is on, so $X_p =$ VVDD $- \Delta V_p$. Transistor Mp1 has $V_{gs} = 0$, so the leakage current through this transistor is the same as a conventional 2-T inverter at VDD = VVDD $- \Delta V_p$. If $\Delta V_p > 0$, then Mp0 has $V_{sg} = V_{sd} = \Delta V_p$, so it is in diode connected mode; otherwise Mp0 is completely off. The voltage across this V_T-low diode is tens of millivolts, so the DIBL effect is negligible, but its leakage current can be high. Mp2 redirects this current to the VVDD$-\Delta V_p$ supply rail. This means that the voltage drop across this leakage path is very small, so its leakage power is small. If $\Delta V_n > 0$, then Mn2 is off, but if $\Delta V_n < 0$, then it works like a diode between ground and the $0 + \Delta V_n$ line, and because ΔV_n is small, its leakage power consumption will be small. Similarly when Do is low, $F =$ VVDD, $X_n = \Delta V_n$, and V_{gs} (Mn1) $= 0$. If $\Delta V_n > 0$, then V_{ds} (Mn0) $= V_{gs}$ (Mn0) $= \Delta V_n$ and it works like a diode. Mp2 and Mn2 do not have any other functionality. They are always minimum-size transistors and only provide paths for diode currents.
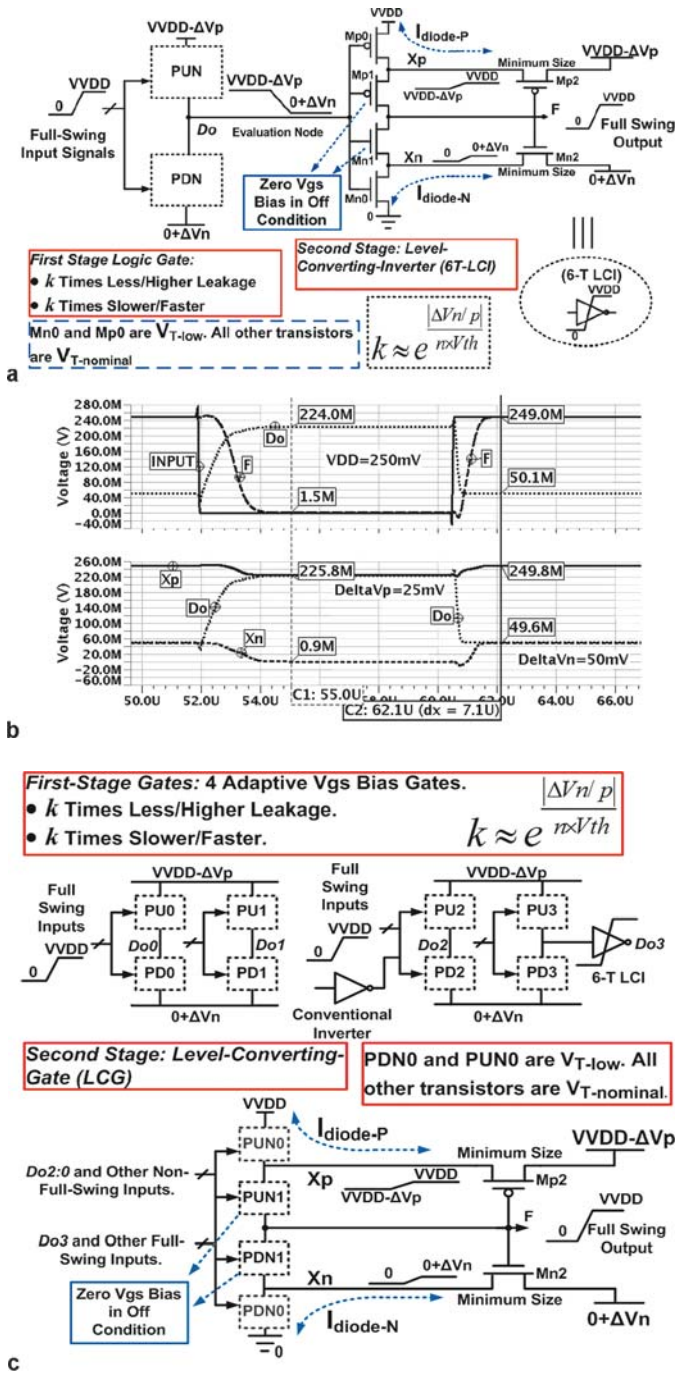
Fig. 3.18 Proposed AVGS style. **a** AVGS–SCMOS single-stage logic gate. **b** voltage waveforms of OR gate at VVDD = 250 mV, $\Delta V_n = 50$ mV, and $\Delta V_p = 25$ mV. **c** AVGS–SCMOS two-stage logic gate

Voltage waveforms of AVGS–SCMOS are shown in Fig. 3.18b. The voltage swings in X_n and X_p nodes are very small $(\Delta V_{n/p})$. So the effect of the parasitic capacitances of these nodes on the delay and dynamic power is very small. Mn0 and Mp0 are V_T-low devices and about ten times faster than Mn1 and Mp1, respectively. The W of these transistors can be smaller than Mn1 and Mp1. Because Mn0 and Mp0 contribute to only approx. 10% of 6-T LCI delay and their leakage power is small, the PV sensitivity of these devices is not important. Also, the effect of Mn2 and Mp2 on the PV sensitivity of timing and power is negligible because they only provide paths for leakage currents.

When we have several gate stages, we can decrease the overheads by using level converting gates (LCG) instead of LCIs. Figure 3.18c shows a general two-stage AVGS–SCMOS circuit. All of the first-stage gates are supplied by new power rails. LCG has exactly the same structure as LCI; PDN and PUN are duplicated using V_T-low devices. The LCG can have both full-swing and non-full-swing inputs. The voltage swing in X_n and X_p nodes is $\Delta V_{n/p}$. PDN0 and PUN0 are V_T-low, and PDN1 and PUN1 are V_T-nominal, so PDN0/PUN0 is much faster than PDN1/PUN1, respectively.

AVGS–PTL

Various PTL styles can be summarized in three general structures. The simplest way to implement a switch is using a transmission gate (TG). PTL implemented with TG is called PTL+ (also called CMOS+). This structure is very robust and provides full-swing input and output signaling and works sizing independent. Single-rail PTL and complementary PTL styles use NMOS switch networks and swing restoration is done by cross-coupled PMOS transistors (cross-coupled inverters in SRPL). Single-rail PTL and CPL are not suitable for the sub-VT region due to the sizing dependency and contention current between swing restorer and NMOS network.

One of the drawbacks of the PTL style is that some of the input signals are directly connected to source/drain junctions. This fact causes two problems. First, logic gate input capacitance is data dependent. In timing analysis we always have to consider worst-case delay, so this issue deteriorates the timing. Second, the charge sharing between interconnects and internal nodes as shown in Fig. 3.19a can increase the signal path delays several times. Also, it is difficult to model this effect in the CAD tools because slt1 and IN1 are two independent signals in two separate paths, path1 and path2. To eliminate both problems we can add inverters to buffer the input signals that drive the source/drains junctions. This solution is shown in Fig. 3.19b. Inputs that drive the transistor gates do not need to be buffered. As shown in Fig. 3.19b, it is also possible to have AVGS mixed SCMOS–PTL gates.

Figure 3.19c illustrates an example AVGS–PTL gate that calculates $F = (A + B)(C + D)$. As illustrated in Fig. 3.19d, it is quite easy to apply the AVGS technique to lookup tables (LUTs). Because charge sharing and crosstalk noise may change the SRAM internal data, we should add the buffers INV15:0. To the best of our
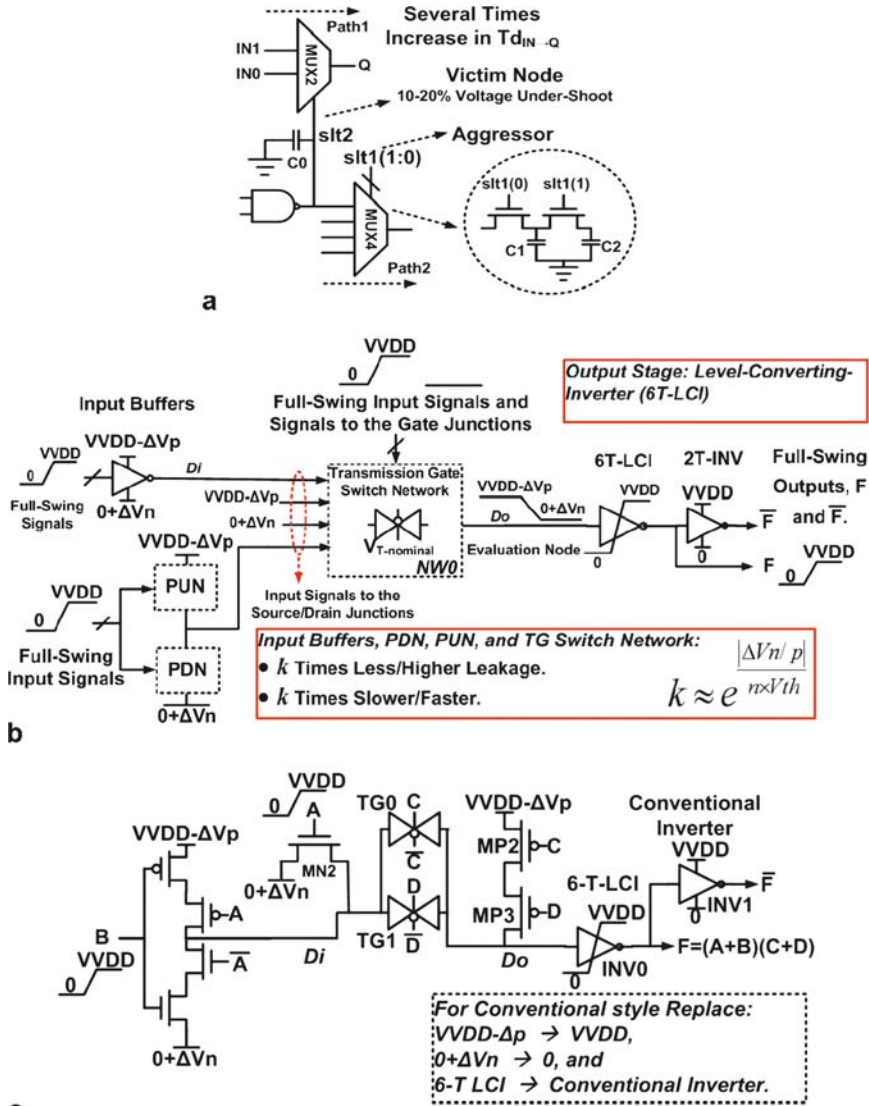
Fig. 3.19 Proposed AVGS style. **a** Charge-sharing and input-capacitance data-dependency problems of PTL. **b** Proposed AVGS–PTL logic style. **c** Function $F = (A + B)(C + D)$. **d** MUX-based sub-VT LUT4. **e** Voltage waveforms of LUT4 at VVDD = 300 mV, $\Delta V_n = 25$ mV, and $\Delta V_p = -25$ mV

knowledge, until now no work has been done on sub-VT FPGAs. Both FPGA and sub-VT circuits are considered as solutions for low-performance applications. Thus it is quite reasonable to have FPGAs with very low energy per operation. Example voltage waveforms are shown in Fig. 3.19e.
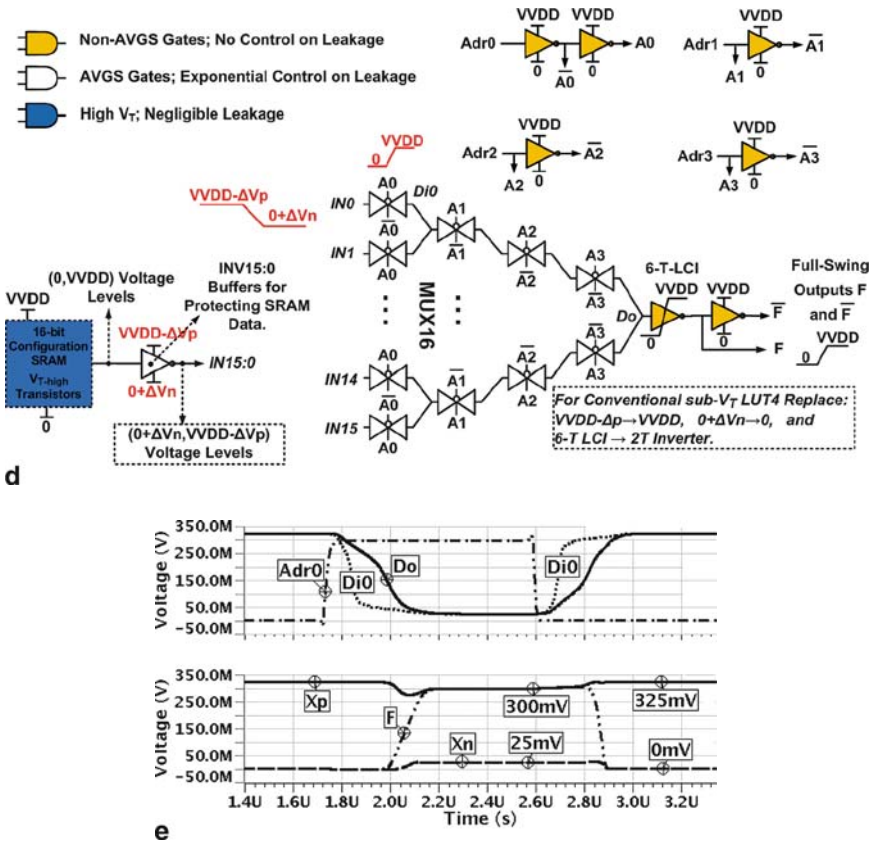
Fig. 3.19 Continued

Results and Discussions

Figure 3.20a–d shows the normalized power delay curves of AVGS–SCMOS, VAR-VDD, and ABB methods for an OR16 gate. SCMOS OR16 performance at VVDD= 300 mV is the reference. AVGS–SCMOS OR16 has four NOR4 in the first stage and level converter NAND4 in the second stage. In Fig. 3.20a, when $\Delta V_{n/p} = 0$, we see about 3% dynamic energy and 19% delay overheads because of ten extra transistors (25%). The dynamic energy overhead is small because switching activity in Do nodes is very small. By increasing $\Delta V_{(n/p)}$, delay increases and leakage decreases exponentially. It is not useful to increase $\Delta V_{n/p}$ more than $+80\,\text{mV}$ because the leakage curve will saturate to the leakage of LCG.

VAR-VDD and ABB are applied to conventional SCMOS OR16 gates. Figure 3.20b shows VAR-VDD behavior. Delay increases exponentially but power decreases in a quadratic form. As shown in Fig. 3.20c, RBB fails to save power at the TT corner. We have used an industrial 65-nm bulk-CMOS low-power process
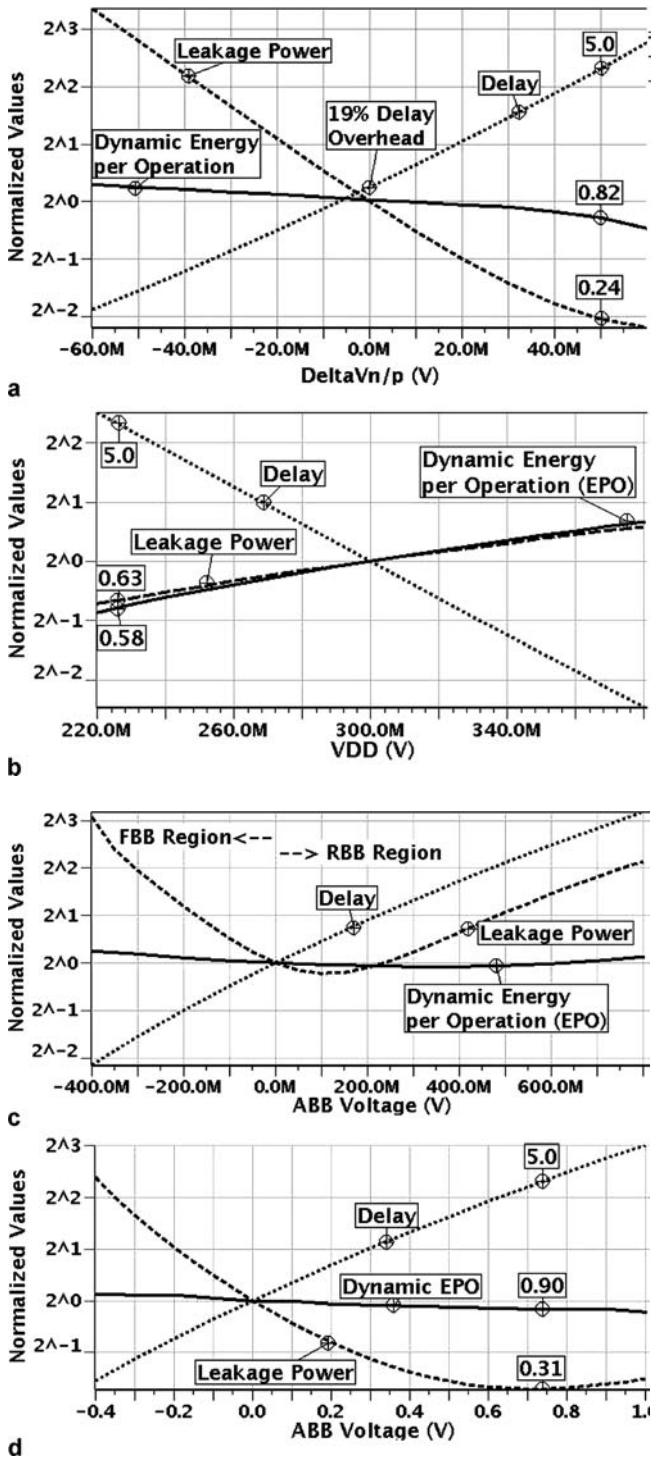
Fig. 3.20 Power-delay curves of OR16. **a** AVGS with $\Delta V_n = \Delta V_p$. **b** VAR-VDD. **c** ABB in TT corner. **d** ABB in FF corner

for simulations, and in this process I_{SUB} is comparable to other leakage terms. RBB increases V_T (and delay) but it also increases other leakage terms. But in the FF corner I_{SUB} is quite dominant and, as is shown in Fig. 3.20d, RBB can compensate this corner about 3.3×. AVGS and VAR-VDD work in all corners. The simulation results of LUT4 (Fig. 3.19d) are very similar to the power delay curves shown in Fig. 3.20a–d. So to save space, we omit them.

Figure 3.21 shows the ability of AVGS–PTL to compensate the process imbalance, FS, and SF corners. In this style, ΔV_n controls the NMOS transistors in NW0 and in input buffers (Fig. 3.19b), while ΔV_p controls the PMOS ones. So by increasing ΔV_n when ΔV_p is constant we can compensate the fast-NMOS slow-PMOS (FS) corner. Var-VDD cannot provide independent control on NMOS and PMOS transistors.

ABB and RVGS techniques can be combined together to provide more control over the power and delay. This is shown in Fig. 3.22. By using only an ABB/AVGS technique we can compensate the variations $3 \times /5.4\times$, respectively. But by using both techniques it is feasible to compensate 8.3× (RBB voltage = 300 mV for both NMOS and PMOS and $\Delta V_{n/p} = 50$ mV). In this technology NMOS and PMOS transistors have approximately equal body-effect coefficients.

Figure 3.23 shows the total energy (drawn from all power supplies) per operation of the LUT4 structures shown in Fig. 3.19d. For the sake of simplicity, transistors Mn0 and Mp0 of 6T-LCI shown in Fig. 3.18a are implemented by V_T-nominal devices. The energy shown in Fig. 3.23 is the average value for a long random input data pattern. The switching activity (the probability of data transition in each clock cycle) of input address signal (Adr3:0) is 10%. The LUT4 critical path delay is from Adr0 to F and is 20% of the clock period. The FS corner causes a 71% increase at the minimum energy point and 5.5× speed variation in this point. Both AVGS and ABB techniques can compensate these variations about two times, and both need two extra power supplies.

This method provides exponential control on 60 to 70% of transistors. Increasing $\Delta V_{n/p}$ decreases V_{ds} of Mn1/Mp1 and so causes less leakage, just as with VAR-VDD. Because AVGS increases the source voltage, it decreases I_{SUB} and gate leakage at the same time, but in the sub-VT region I_G is not important. RBB increases the drain-bulk reverse diode current and the GIDL current because it
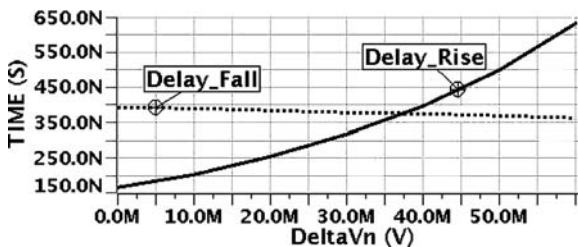

Fig. 3.21 Compensating FS corner by AVGS–PTL. Output F rise and fall delays of MUX8 gate
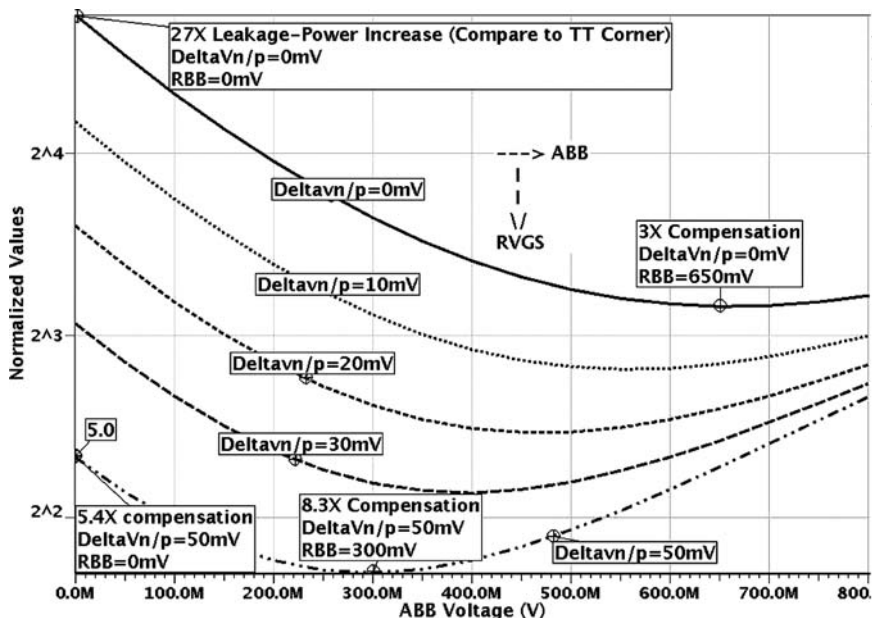
Fig. 3.22 Using AVGS in conjunction with ABB to compensate FF corner (reduced L, TOX, and VT)
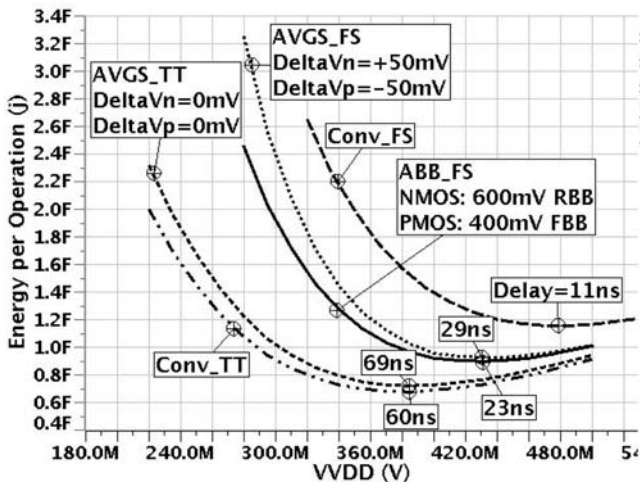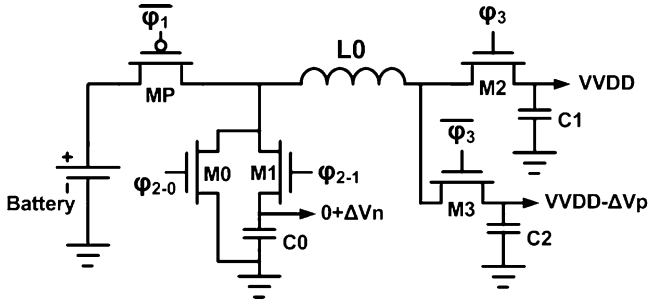

Fig. 3.23 Minimum energy point of LUT4 shown in Fig. 3.19d in TT and FS corners for both conventional (Conv) and AVGS–PTL styles. Given delay values are the critical path delay at minimum energy point

Fig. 3.24 Proposed bulk-down converter capable of generating all necessary power supplies

decreases the bulk voltage. It is important to note that every transistor is driven by $V_{gs} = \text{VVDD} - \Delta V_{n/p}$, and no transistor has $V_{gs} = \text{VVDD} - \Delta V_n - \Delta V_p$.

The area overhead is 2 to 4 transistors per gate. So in this logic style, using high fan-in gates is better. High fan-in gates also cause less leakage power due to the stack effect. In [51] it is shown that in future SOI-wafer multigate device technologies, delay increase vs. number of fan-in gates will be 41% less than bulk-CMOS. This means that high fan-in gates will work much better in these new technologies.

AVGS can be applied to bulk-CMOS, PD-SOI, FD-SOI, or FinFET. In all of these devices, increasing the source voltage when $V_g = 0$ increases the source-channel barrier height "seen" by electrons and so provides exponential control over leakage.

Figure 3.24 illustrates a modified single-inductor multioutput (SIMO) converter that can generate all of the necessary voltage levels. $\varphi1$ and $\varphi2$ are nonoverlapping. In $\varphi1$ MP is on. In $\varphi2$ one of the synchronous rectifiers, M0 or M1, can be turned on. $\varphi3$ selects which of the VVDD or VVDD$-\Delta V_p$ is to be charged. Logic gates charge C0 while L0 discharges it. With this converter it is possible to have both positive and negative $\Delta V_{n/p}$. A similar concept can be applied to switch-cap converters.

In summary, if γ-factor is sufficiently high, ABB is the best choice for compensating variations. But the technology trend shows a degradation of the body effect in all devices. In this work we proposed an AVGS method that can change the power and delay of digital gates by adjusting the V_{gs} voltage. AVGS does not need triple-well technology, works in all technology nodes, can be applied to any device, and can be used in conjunction with other conventional methods.

Conclusions

This chapter has shown some of the architectural solutions that are applicable to nanoelectronic circuits. Indeed, variability and power dissipation are two of the major hurdles that new technologies have to overcome. Crossbars realize circuits as a regular fabric and so minimize the spread of circuit delays. A new

method for reducing timing variability has also been shown. Finally, ultra-low-power nanosystems require specific design technologies; adaptive V_{gs} is a very promising approach.

References

1. Holmes JD et al (2000) Control of thickness and orientation of solution-grown silicon nanowires. Science 287(5457):1471–1473
2. Luo Y et al (2002) Two-dimensional molecular electronics circuits. Chem Phys Chem 3: 519–525
3. DeHon A (2005) Design of programmable interconnect for sublithographic programmable logic arrays. In: International symposium on field-programmable gate arrays, Monterey, CA, 2005, pp 127–137
4. Choi Y-K et al (2002) A spacer patterning technology for nanoscale CMOS. IEEE Trans Electron Devices 49(3):436–441
5. Cerofolini GF (2006) Search for realistic limits to computation. II. The technological side Appl Phys A 86(1):31–42
6. Moselund KE et al (2007) Cointegration of gate-all-around MOSFETs and local silicon-on-insulator optical waveguides on bulk silicon. IEEE Trans Nanotechnol 6(1):118–125
7. Abadir MS, Reghbati HK (1983) Functional testing of semiconductor random access memories. ACM Comput Surv 15(3):175–198
8. Ben Jamaa MH et al (2008) Variability-aware design of multi-level logic decoders for nanoscale crossbar memories. IEEE Trans Comput Aided Des 27(11):2053–2067
9. Adams RD (2003) High performance memory testing. Kluwer, Dordrecht
10. Ben Jamaa MH et al (2008) A stochastic perturbative approach to design a defect-aware thresholder in the sense amplifier of crossbar memories. To appear in ASP-DAC
11. International Technology Roadmap for Semiconductors (2006) ITRS URL: http://www.itrs.net/Links/2007ITRS/Home2007.html. Accessed 12 Nov 2008
12. Strojwas A et al (1996) Manufacturability of low power CMOS technology solutions. In: Proceedings of international symposium on low power electronics and design (ISLPED), San Diego, CA, pp. 225–232
13. Eisele M, Berthold J, Schmitt-Landsiedel D, Mahnkopf R (1996) The impact of intra-die device parameter variations on path delays and on the design for yield of low voltage digital circuits. In: Proceedings of international symposium on low power electronics and design (ISLPED), San Diego, CA, pp. 237–242
14. Bowman KA, Tang X, Eble JC, Meindl JD (2000) Impact of extrinsic and intrinsic parameter fluctuations on CMOS circuit performance. IEEE J Solid State Circuits 35:1186–1193
15. Bowman K et al (2002) Impact of die-to-die and within die parameter fluctuation on the maximum clock frequency distribution for gigascale integration. IEEE J Solid State Circuits 183–190
16. Suaris P, Kgil T, Bowman KA, De V, Mudge TN (2005) Total power-optimal pipelining and parallel processing under process variations in nanometer technology. In: Proceedings of the 2005 IEEE/ACM international conference on computer-aided design (ICCAD), San Diego, CA, pp 535–540
17. Bowman KA, Austin BL, Eble JC, Tang X, Meindl JD (1999) A physical alpha-power law MOSFET model. IEEE J Solid State Circuits 34:1410–1414
18. Cao Y, Clark LT (2005) Mapping statistical process variations toward circuit performance variability: an analytical modeling approach. In: Proceedings of DAC, pp 658–663
19. Bult K (2005) Scaling effects in analog design in deep sub-micron CMOS, Lecture Notes: Advanced CMOS Circuit Design, Mead Education, 2005
20. HSPICE User's Manual (1995) Meta-Software Inc

21. Ho R et al (2001) The future of wires. IEEE Proc 89(4):490–503
22. Papoulis A, Pillai SU (2002) Probability, random variables and stochastic processes, 4th edn. McGraw-Hill, New York
23. Friedberg P et al (2005) Modeling within-die spatial correlation effects for process-design co-optimization. In: Proceedings of international symposium on quality electronic design, pp 516–521
24. Stolk PA, Widdershoven FP, Klaassen DBM (1998) Modeling statistical dopant fluctuations in MOS transistors. IEEE Trans Electron Devices 45(9):1960–1971
25. Agarwal A, Blaauw D, Zolotov V (2003) Statistical timing analysis for intra-die process variations with spatial correlations. In: Proceedings of international conference on computer aided design, pp 900–907
26. Le J, Li X, Pileggi LT (2004) STAC:statistical timing analysis with correlation. In: Proceedings of design automation conference, pp 343–348
27. Frank DJ, Solomon P, Reynolds S, Shin J (1997) Supply and threshold voltage optimization for low power design. In: Proceedings of international symposium on low power electronics and design (ISLPED), San Diego, CA, pp 317–322
28. Gronowski PE, Bowhill WJ, Preston RP, Gowan MK, Allmon RL (1998) High-performance microprocessor design. IEEE J Solid State Circuits 33:676–686
29. Hojo T (1931) Distribution of the median, quartiles and interquartile distance in samples from a normal population. Biometrika 23:315–360
30. Keeping ES (1995) Introduction to statistical inference. Dover, New York
31. Moore PG (1956) The estimation of the mean of a censored normal distribution by ordered variables. Biometrika 43:482–485
32. Chishti Z, Vijaykumar TN (2008) Optimal power/performance pipeline depth for SMT in scaled technologies. IEEE Trans Comput 69–81
33. Asenov A (2007) Simulation of statistical variability in nano MOSFETs. In: Proceedings of IEEE symposium on VLSI technology, pp 86–87
34. Das A et al (2007) Mitigating the effects of process variations: architectural approaches for improving batch performance. In: Proceedings of 34th international symposium on computer architecture (ISCA)
35. Karnik T, Borkar S, De V (2004) Probabilistic and variation-tolerant design: key to continued moore's law. Invited talk in ACM/IEEE Int'l TAU Workshop on Timing Issues
36. Yamaoka M et al (2004) Low power SRAM menu for SOC application using Yin-Yang-feedback memory cell technology. In: Proceedings of symposium on VLSI circuits, pp 288–291
37. Vittoz E (2004) Weak inversion for ultimate low-power logic. In: Piguet C (ed) Chapter 16 in Low-power electronics design, CRC Press, Boca Raton
38. Calhoun BH et al (2005) Modeling and sizing for minimum energy operation in sub-threshold circuits. JSSC 40(9):1778–1786
39. Zhai B et al (2006) A 2.60pJ/Inst subthreshold sensor processor for optimal energy efficiency. VLSI Ckts Symp
40. Verma N, Chandrakasan AP (2007) A 65nm 8T sub-VT SRAM employing sense-amplifier redundancy. In: International solid state circuits conference (ISSCC), pp 328–329
41. Kwong J et al (2008) A 65nm sub-VT microcontroller with integrated SRAM and switched-capacitor DC–DC converter. In: International solid state circuits conference (ISSCC), pp 318–319
42. Zhai B et al (2005) Analysis and mitigation of variability in subthreshold design. In: Proceedings of international symposium on low power electronics and design (ISLPED05), San Diego, CA
43. Ryan JF et al (2007) Analyzing and modeling process balance for sub-threshold circuit design. In: GLSVLSI07 Proceedings of Great Lakes Symposium on VLSI Design
44. De V et al (2001) Techniques for leakage power reduction. In: Chandrakasan A, Bowhill W, Fox F (eds) Design of high-performance microprocessor circuits. IEEE Press, Piscataway, NJ, pp 46–62

45. Calhoun BH, Chandrakasan AP (2006) Ultra-dynamic voltage scaling (UDVS) using sub-threshold operation and local voltage dithering. JSSC06 41(1):238–245
46. Tschanz J et al (2002) Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage. JSSC IEEE Journal of Solid-State Circuits vol. 37
47. Jayakumar N, Khatri SP (2005) A variation-tolerant sub-threshold design approach. In: DAC, pp 716–719
48. Singh N et al (2006) High-performance fully depleted silicon nanowire (diameter $\leq$ 5 nm) gate-all-around CMOS devices. IEEE Electron Device Lett 27(5):383–386
49. Frei James et al (2004) Body effect in Tri- and Pi-Gate SOI MOSFETs. IEEE Electron Device Lett 25(12):813–815
50. Chaudhry A, Kumar MJ (2004) Controlling short-channel effects in deep-submicron SOI MOSFETs for improved reliability: a review. IEEE Trans Device Mater Reliability 4(1): 99–109
51. von Arnim K et al (2007) A low-power multi-gate FET CMOS technology with 13.9ps inverter delay, large-scale integrated high performance digital circuits and SRAM. In: Symposium on VLSI technology digest of technical papers, pp 106–107
52. Choi J-H, Murthy J, Roy K (2007) The effect of process variation on device temperature in FinFET circuits. In: Proceedings of the 2005 IEEE/ACM international conference on computer-aided design (ICCAD), San Diego, CA, pp 747–751
53. Cho KeunHwi et al (2007) Temperature-dependent characteristics of cylindrical gate-all-around twin silicon nanowire MOSFETs (TSNWFETs). IEEE Electron Device Lett 28(12):1129–1131
54. Cho GR, Chen T (2004) Synthesis of single/dual-rail mixed PTL/static logic for low-power applications. IEEE Trans CAD Integrated Circuits Syst 23(2):229–242