Ultra-Low Power VLSI Circuit Design Demystified and Explained: A Tutorial

Massimo Alioto, *Senior Member, IEEE*

(Invited Paper)

***Abstract*—In this paper, the state of the art in ultra-low power (ULP) VLSI design is presented within a unitary framework for the first time. A few general principles are first introduced to gain an insight into the design issues and the approaches that are specific to ULP systems, as well as to better understand the challenges that have to be faced in the foreseeable future. Intuitive understanding is accompanied by rigorous analysis for each key concept. The analysis ranges from the circuit to the micro-architectural level, and reference is given to process, physical and system levels when necessary. Among the main goals of this paper, it is shown that many paradigms and approaches borrowed from traditional above-threshold low-power VLSI design are actually incorrect. Accordingly, common misconceptions in the ULP domain are debunked and replaced with technically sound explanations.**

***Index Terms*—Energy-autonomous systems, self-powered nodes, standard CMOS logic, subthreshold, ultra-low power, ultra-low voltage, very large scale integration (VLSI).**

List of Abbreviations:

DIBL	Drain induced barrier lowering.
FBB	Forward body biasing.
NBTI	Negative bias temperature instability.
MCML	MOS current-mode logic.
MEP	Minimum energy point.
PDN	Pull-up network.
PUN	Pull-down network.
PVT	Process, voltage, temperature
RBB	Reverse body biasing.
RDF	Random dopant fluctuation.
RNCE	Reverse narrow channel effect.
RSCE	Reverse short channel effect.
ULP	Ultra low power.
ULV	Ultra low voltage.

I. Introduction

Ultra-low power VLSI circuits are gaining considerable interest from the scientific community and more recently the market. Indeed, many recent and prospective applications explicitly rely on the availability of sensor nodes that are energy autonomous and extremely small sized. A few examples of such applications are wireless sensor networks, biomedical and implantable devices/networks, ambient intelligence, wearable computing, smart grids, pollution monitoring, plant monitoring, smart warehouses [1]–[5].

In the context of these applications and the related technologies, the main drivers are lifetime and size [6]–[8]. Battery lifetime in the order of several years or decades would be highly desirable, although currently it is hardly within reach. Millimeter size or less is also a target for future nodes [1], [2], and some prototype is now available [9]. Both lifetime and size are tightly constrained by the energy storage/scavenging device (they typically set the size of the whole node) and the node consumption. Since the battery technology has evolved much slower than CMOS technology, a considerable research effort has been devoted in aggressively reducing the consumption of nodes, which today can be well below the microwatt for the above applications. Innovation will be constantly required to continue the historical 10–100× reduction in computers' size every ten years according to the Bell's law [6], [7], while ensuring energy-autonomous operation.

Voltage scaling is certainly a very effective lever to reduce power and energy consumption, hence ultra-low power (ULP) design translates into ultra-low voltage (ULV) design. For this reason, energy-autonomous nodes work at voltages that are below the MOS threshold voltage. ULV operation offers great opportunities in terms of low consumption, and poses very serious challenges that need to be solved before ULP technologies go into mass production. Low chip cost (i.e., design effort, yield) is another required feature for many of the above mentioned applications (with few exceptions, e.g., implantable devices).

This paper is focused on VLSI digital circuits, which are a key component in every ULP system. Other than being used for obvious processing purposes, digital VLSI circuits also intrude into the analog and RF realm, where digitally-assisted circuits are employed to improve the features of purely analog/RF blocks and benefit from Moore's law [10], [11]. At the same time, enhanced digital processing capability permits to reduce the amount of wirelessly transmitted data. Since RF communications are energy costly, ULP systems greatly benefit from

Manuscript received March 14th, 2009; revised October 16, 2011; accepted November 06, 2011. Date of publication January 02, 2012; date of current version January 11, 2012. This paper was recommended by Editor G. Manganaro.

The author is with the DII (Dipartimento di Ingegneria dell'Informazione), Università di Siena, 53100 Siena, Italy, and also with the Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109-2122 USA (e-mail: malioto@dii.unisi.it; alioto@umich.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCSI.2011.2177004

moving computation inside the nodes, as well as from exploiting the advantages brought by technology scaling.

The paper is organized as follows. Energy and power requirements in ULP systems in view of the real applications are discussed in Section II. ULV MOS models and properties are introduced in Section III, and are used in Section IV to identify the general issues and principles that guide the design of ULP circuits and systems. General considerations on variations are introduced in Section V. Voltage scaling limits are then addressed in Section VI, and voltage optimization for energy/power minimization (or tradeoff with performance) is discussed in Section VII. The impact of nonidealities on energy/power minimization is separately analyzed in Section VIII. The above general design principles are then applied to two important classes of circuits: standard-cell circuits (Section IX) and memory subsystem (Section X). Standard circuit techniques to reduce leakage are put in the context of ULP design in Section XI. Micro-architectural tradeoffs and basic techniques in ULV systems (pipelining, hardware replication) are analyzed in Section XII. Run-time techniques to deal with variations and occasional errors are then reviewed in Section XIII. Remarks and recent directions are reported in Section XIV, and conclusions are drawn in Section XV. In all these sections, the underlying message is that ULP design is very different from traditional above-threshold low-power design, and common misconceptions are systematically pointed out and clarified.

II. Ultra-Low Power VLSI Circuits: Design Constraints Imposed by Applications

A. Constraints in Battery-Operated and Energy-Scavenged Systems

In battery-operated systems, the lifetime of the battery depends on the average power consumed by the system according to

$$T_{\mathrm{battery}} \approx \frac{E_{\mathrm{battery}}}{P_{\mathrm{avg}}} \tag{1}$$

where E_{battery} is the energy that can be delivered by the battery over a reasonable range of voltages and including nonidealities like memory effects, which tend to degrade the effective capacity of the battery. For example, considering a 10-mAh button cell, a lifetime of one year (which is rather small for the intended applications) requires $P_{\mathrm{avg}} \sim 1\ \mu\mathrm{W}$. Using more compact printed batteries with capacity in the order of tens of $\mu\mathrm{Ah}$ and targeting a lifetime of decades, the required P_{avg} can easily be in the order of nWs or even less [9].

ULP systems typically perform a repetitive short task at a given wakeup period T_{wkup} imposed by the application, hence their power consumption can be drastically reduced through duty cycling. As shown in Fig. 1, the blocks performing the repetitive task periodically wake up and turn to active mode (for just 0.1%–1% of the period or less), whereas they stay in sleep mode for most of the time. ULP systems also contain some very simple always-on circuitry that periodically triggers the wakeup mode at the desired rate and stores information between consecutive tasks. Hence, always-on blocks are typically slow timers, simple control circuits, registers or retentive SRAM memories [12].

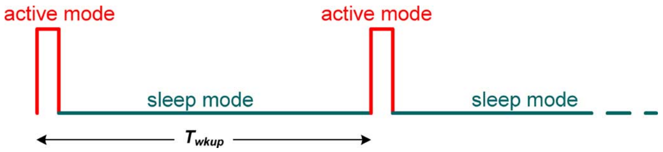


Fig. 1. Operation of duty-cycled ULP systems.

In duty-cycled ULP systems, the average power consumption is equal to

$$P_{\mathrm{avg}} = P_{\mathrm{always-on}} + P_{\mathrm{sleep}} + \frac{E_{\mathrm{active}}}{T_{\mathrm{wkup}}} \tag{2}$$

where $P_{\mathrm{always-on}}$ is the average power consumed by the always-on blocks, P_{sleep} is the power consumed by duty cycled blocks in sleep mode and $E_{\mathrm{active}}/T_{\mathrm{wkup}}$ is the average power in active mode [1] (i.e., the energy E_{active} spent in active mode averaged over T_{wkup}). In some cases, in sleep mode the clock is stopped and the supply is power gated, hence P_{sleep} is only due to the leakage (reduced by power gating) of the duty cycled block, and easily exceeds $P_{\mathrm{always-on}}$ since this block is far more complex than always-on circuitry. On the other hand, in ULP systems with extremely tight power budget, the voltage regulator powering the duty-cycled block is shut down in sleep mode to completely eliminate leakage, thereby making $P_{\mathrm{sleep}} = 0$ in (2). This assumption will be always implied in the following sections (anyway, if $P_{\mathrm{sleep}} \neq 0$, it can be included in $P_{\mathrm{always-on}}$).

Very similar considerations hold for energy-scavenged systems, which are typically designed for perpetual operation. In this case, the battery is included to store the energy that is delivered when no energy is momentarily scavenged. In addition, perpetual operation requires the average power dissipated by the system to be lower than (or equal to) the average power delivered by the energy scavenging device. The only difference with respect to the purely battery-operated case is that P_{avg} is constrained by the energy scavenging device, rather than the battery capacity.

Finally, another typical scenario is represented by batteryless systems powered through pure energy scavenging. In this case, the instantaneous power consumption is the parameter of interest, rather than the average power.

B. Role of Energy and Power

From (2), the average power can be reduced by minimizing E_{active} and $P_{\mathrm{always-on}}$, assuming that $P_{\mathrm{sleep}} = 0$. In applications with low or moderate wakeup period (a fraction of a second or less), $P_{\mathrm{avg}} \approx E_{\mathrm{active}}/T_{\mathrm{wkup}}$ is dominated by duty cycled blocks. In this case, P_{avg} is minimized by minimizing the energy per operation E_{active} in the duty cycled blocks, which justifies the extensive research in the last decade on VLSI circuits with minimum energy per operation. On the other hand, in applications with larger T_{wkup} (e.g., seconds or more), $P_{\mathrm{always-on}}$ plays an important role in determining P_{avg}.

From the above considerations, in applications requiring ultra-low power consumption, one has to minimize both energy and power. To be more specific, always-on circuits have to be

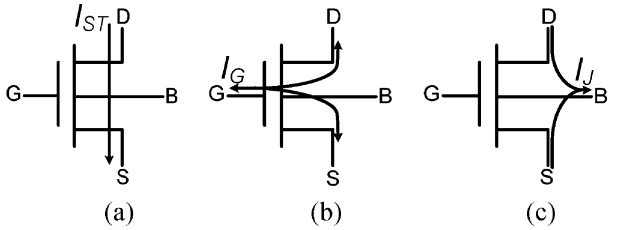


Fig. 2. NMOS transistor current contributions in subthreshold. (a) Subthreshold current. (b) Gate current. (c) Junction current.

TABLE I
NMOS/PMOS Transistor Parameters (65-nm, Std-V_{TH}) [17]

	n	I_0 [A]	V_{TH0} [V]	λ_{DS}	λ_{BS}
NMOS	1.39	6.65E-5	0.598	9.0E-2	9.9E-2
PMOS	1.27	5.95E-6	0.532	8.0E-2	1.1E-1

designed minimizing their power, whereas duty-cycled blocks must be designed minimizing their energy.

III. Transistor Models for ULP/ULV Operation

In this section, simplified transistor models are presented to better understand the specific features of MOS transistors at ultra-low voltages. All considerations are made for the NMOS transistor and are immediately extended to PMOS.

A. *MOS* I-V *Characteristics at Ultra-Low Voltages*

An NMOS transistor operating in subthreshold (i.e., $V_{GS} < V_{TH}$, where V_{TH} is the transistor threshold voltage) experiences the three current contributions in Fig. 2(a)–(c): the subthreshold current I_{ST} (it originates from the diffusion of minority carriers between drain and source [13]), the gate current I_G (due to tunneling through the dielectric) and the junction current I_J (mainly due to band-to-band-tunneling current across the thin depletion regions) [14]. Due to the much stronger dependence on the gate voltage, I_G tends to be much lower than I_{ST} at low voltages, and the same holds for I_J [14]. Hence, the NMOS current at ULV is dominated by the subhtreshold contribution I_{ST} in Fig. 2(a), which is usually written in the following form [13], [15]:

$$I \approx I_{ST} = I_0 \frac{W}{L} e^{(V_{GS}-V_{TH})/n \cdot v_t}(1 - e^{-V_{DS}/v_t}). \quad (3)$$

Where I_0 is the technology-dependent subthreshold current extrapolated for $V_{GS} = V_{TH}$, $v_t = kT/q$ is the thermal voltage, W/L is the aspect ratio and n is the subthreshold factor [13]. In (3), the threshold voltage V_{TH} also depends on the drain-source voltage V_{DS} (through the drain induced barrier lowering (DIBL) effect) and the bulk-source voltage V_{BS} (through the body effect) according to

$$V_{TH} = V_{TH0} - \lambda_{DS} V_{DS} - \lambda_{BS} V_{BS} \quad (4)$$

where $\lambda_{DS} > 0$ is the DIBL coefficient and $\lambda_{BS} > 0$ is the body effect coefficient [16]. As an example, their value is reported in Table I for 65-nm std-V_{TH} transistors.

From a design standpoint, it is convenient to rewrite (3)–(4) according to [17]:

$$I = \beta \cdot e^{V_{GS}/nv_t} \cdot \left[e^{\lambda_{DS} V_{DS}/nv_t}\left(1 - e^{-V_{DS}/v_t}\right)\right] \quad (5a)$$

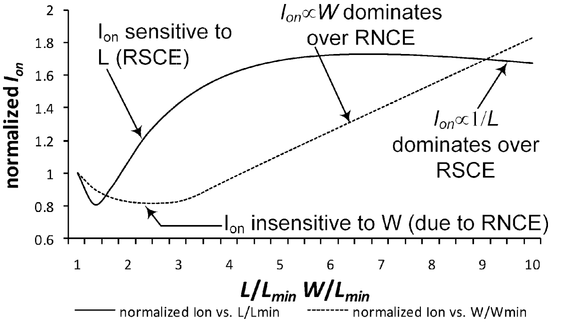


Fig. 3. NMOS strength normalized to the case with $W = W_{\min}$ ($L = L_{\min}$) versus $W/W_{\min}$ ($L/L_{\min}$).

$$\beta = I_0 \frac{W}{L} e^{-(V_{TH0} - \lambda_{BS} V_{BS})/nv_t} \quad (5b)$$

where (5a) highlights the exponential dependence on V_{GS} and the dependence on V_{DS} (in square brackets), whereas all other terms are grouped in the parameter β, which represents the transistor strength [17].

B. *Knobs Available for Tuning the Transistor Strength*

From (5b) the transistor strength tuning can be performed:

- by tuning the aspect ratio W/L;
- by selecting the zero-bias threshold V_{TH0} among the low/std/high values available in the adopted technology;
- by statically or dynamically tuning the bulk voltage V_{BB} (i.e., body biasing).

Regarding W/L, the explicit linear dependence of β in (5b) adds to the implicit dependence of V_{TH} on W/L, which is significant for narrow or short channels (i.e., with $W \sim W_{\min}$ and $L \sim L_{\min}$). For this reason, sizing guidelines in subthreshold are completely different from above threshold.

More specifically, an increase in W leads to a threshold increase due to the reverse narrow channel effect (RNCE) [13], [15], which overcompensates the linear strength increase expected from (5b). As a result, β is rather insensitive to W for small sizes, hence W is not an effective knob to increase the strength. This can be seen in Fig. 3, which shows that the NMOS strength is relatively insensitive to W for $W \leq 4W_{\min}$. The strength is increased more effectively by connecting minimum-sized transistors in parallel, although at the expense of some area overhead. Conversely, β can be increased by increasing L thanks to the reverse short channel effect (RSCE), which determines a threshold decrease when L is increased [18]. As shown in Fig. 3, an increase in L in the range $L_{\min} - 3L_{\min}$ leads to an appreciable strength increase, which decreases again for larger L due to the $1/L$ dependence. Hence, transistor sizing at ULV is very different from above threshold and is strongly technology dependent, thereby requiring a deeper knowledge of process and device.

A much stronger knob to tune the strength is the threshold voltage, thanks to the exponential dependence of β on V_{TH0}. As a numerical example, assuming the typical case where low-V_{TH} transistors offer a threshold voltage reduction by 100 mV compared to std-V_{TH} transistors, from (5b) and Table I, the strength of the former transistors is about 18 times that of std-V_{TH} transistors with the same W/L.

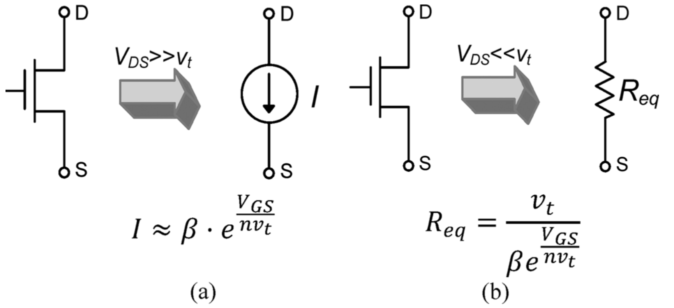

D
S
$V_{DS}>>v_t$
I
D
S
$I \approx \beta \cdot e^{\frac{V_{GS}}{nv_t}}$
(a)
D
S
$V_{DS}<<v_t$
R_{eq}
D
S
$R_{eq} = \frac{v_t}{\beta e^{\frac{V_{GS}}{nv_t}}}$
(b)

Fig. 4. Large-signal MOS models: (a) for high V_{DS} ($> 3 - 4v_t \sim 100\ \mathrm{mV}$); (b) for low V_{DS} ($< v_t \sim 25\ \mathrm{mV}$).

In regard to body biasing, it is also effective in tuning the strength thanks to the exponential dependence of β on the threshold voltage. This is significantly different from traditional above-threshold circuits, in which body biasing is considered an ineffective knob (because in that case the strength has a much weaker dependence on the body voltage). As an example, using the data in Table I, applying a typical Forward Body Bias (FBB) voltage $V_{BB} = 300\ \mathrm{mV}$ to reduce V_{TH}, the strength in (5b) is increased by 2.3× compared to the zero-bias condition $V_{BB} = 0$. On the other hand, the adoption of a Reverse Body Bias (RBB) voltage $V_{BB} = -300\ \mathrm{mV}$ to increase V_{TH}, the strength in (5b) is reduced by the same factor.

C. Design-Oriented Large-Signal Transistor Models

Let us approximate (5) under high and low values of V_{DS}. When $V_{DS} > 3 - 4v_t \sim 100\ \mathrm{mV}$ at room temperature, we get $(1 - e^{-V_{DS}/v_t}) \approx 1$ in (5a), hence

$$I \approx \beta \cdot e^{V_{GS}/nv_t} \qquad \text{if } V_{DS} \gg v_t \tag{6}$$

where the relatively small dependence of term $e^{\lambda_{DS} V_{DS}/nv_t}$ on V_{DS} was ignored (as $\lambda_{DS} \ll 1$ and V_{DS} is low). From (6), the MOS transistor in subthreshold with "high" V_{DS} (i.e., 100 mV or above) is equivalent to a voltage-controlled current source, as shown in Fig. 4(a).

On the other hand, under "low" values of $V_{DS} < v_t \sim 25\ \mathrm{mV}$ at room temperature, $(1 - e^{-V_{DS}/v_t})$ in (5a) can be expanded in Taylor series and truncated to first order, thereby yielding

$$I \approx \beta \cdot e^{V_{GS}/nv_t} \cdot \frac{V_{DS}}{v_t} \qquad \text{if } V_{DS} \ll v_t. \tag{7}$$

From (7), for low V_{DS} the MOS transistor is equivalent to a resistance R_{eq} given by $R_{eq} = V_{DS}/I = v_t/\beta e^{V_{GS}/(nv_t)}$, as depicted in Fig. 4(b).

Summarizing, the MOS transistor in subthreshold region can be modeled either as a current source or as a resistor, as in above-threshold transistors [1], [17]. The only difference with respect to the latter case is that the transition from current source to resistor behavior occurs abruptly (i.e., over a range of voltages of just a few tens of mVs) because of the much stronger dependence of I on V_{DS} due to the term $(1 - e^{-V_{DS}/v_t})$ in (5a).

D. Small-Signal Transistor Models

The MOS low-frequency small-signal model in subthreshold in Fig. 5 is defined by the transconductance $g_m = \partial I/\partial v_{GS}$ and the output resistance $r_0 = [\partial I/\partial v_{DS}]^{-1}$, which from (5a) turn out to have the following expression [1], [17]:

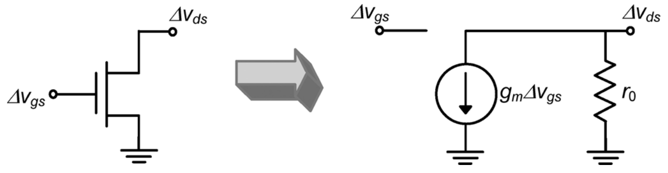


Fig. 5. Small-signal MOS model in subthreshold.

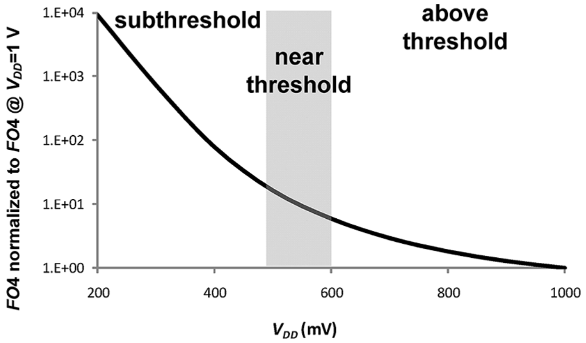


Fig. 6. $FO4$ normalized to its value at $V_{DD,\max} = 1\ \mathrm{V}$ vs. V_{DD}.

$$g_m = \frac{I}{n \cdot v_t} \tag{8a}$$

$$r_0 = \frac{v_t}{I} \cdot \frac{1}{\frac{\lambda_{DS}}{n} + \frac{1}{e^{-V_{DS}/v_t} - 1}}. \tag{8b}$$

The output resistance in (8b) accounts for two different physical effects associated with the two addends in its denominator. Indeed, (8b) includes only the first term λ_{DS}/n in the denominator if we consider only the dependence of I on V_{DS} due to DIBL in (5a) (i.e., only the term $e^{\lambda_{DS} V_{DS}/nv_t}$). On the other hand, the denominator of (8b) includes the second term when we consider the dependence through the carrier diffusion $(1 - e^{-V_{DS}/v_t})$.

From (8b), at ultra-low voltages the second term clearly dominates over the first, hence the intrinsic voltage gain of the MOS transistor $g_m r_0$ tends to degrade dramatically. On the other hand, when V_{DD} is increased to values close to V_{TH}, the first term dominates and $g_m r_0$ asymptotically tends to $1/\lambda_{DS}$ [17].

IV. General Analysis of Issues Arising in Ultra-Low Voltage VLSI Systems at Nominal Conditions

In this section, issues and design considerations are discussed by resorting to the simple models in Section III. Variations are herein neglected for the sake of simplicity, and will be discussed in the following section.

A. Impact of ULV Operation on Current and Performance

From (5a), the MOS subthreshold on-current is given by

$$\begin{aligned} I_{\text{on}} &= \beta \cdot e^{V_{DD}/n \cdot v_t} \cdot \left[e^{\lambda_{DS} V_{DD}/n \cdot v_t} \left(1 - e^{-V_{DD}/v_t} \right) \right] \\ &\approx \beta \cdot e^{V_{DD}/n \cdot v_t} \end{aligned} \tag{9}$$

where $V_{DD} \gg v_t$ is assumed. Clearly, in the subthreshold regime, the reduction in V_{DD} determines an exponential reduction in I_{on}, which reflects into an almost exponential degradation in the delay τ_D [19]–[21], as shown by its classical CV/I expression in (10):

$$\tau_D = \frac{C}{I_{\text{on}}}\frac{V_{DD}}{2} = \frac{C}{2\beta}\frac{V_{DD}}{e^{V_{DD}/n \cdot v_t}}. \tag{10}$$

As an example, Fig. 6 depicts the $FO4$ delay figure of merit normalized to the value at the maximum voltage $V_{DD,\max}$ versus V_{DD}, for a 65-nm technology. The $FO4$ trend in subthreshold is approximately exponential, as expected from (10).

Expressing the clock cycle as a multiple (set by the microarchitecture) of $FO4$, Fig. 6 provides an estimate of the clock cycle scaling versus V_{DD} for a given architecture [22]. From this figure, the performance penalty in subthreshold ranges from 2 to 4–5 orders of magnitude, compared to the case with $V_{DD} = V_{DD,\max}$ under a given microarchitecture. In addition, current subthreshold microarchitectures usually have a 2×–5× higher logic depth compared to traditional energy-efficient designs [23] (see Section XII), hence their clock cycle is further increased by the same factor. As expected, ultra-low power operation comes at a high performance cost.

As discussed in the Introduction, the above speed penalty is not an issue in many ULP applications. On the other hand, in applications that adopt ultra-wide dynamic voltage scaling [24], this speed penalty can be acceptable during normal low-power operation (i.e., when V_{DD} is a few hundreds mV), whereas the maximum speed is still available when peak performance is occasionally needed (i.e., when $V_{DD} = V_{DD,\max}$). In this case, achieving energy-efficient designs over a wide range of voltages is a hard challenge, due to the strong impact of V_{DD} on the circuit and architecture optimization.

Remarks on the MOS Capacitances at Ultra-Low Voltages

Adoption of ultra-low supply voltages also affects the parasitic transistor capacitances, although to a lesser extent. Indeed, the gate capacitance in subthreshold is smaller than above threshold [25]. In a 65-nm technology, this reduction was found to be about 20%. On the other hand, the drain-bulk junction capacitance in subthreshold is less reverse biased compared to above-threshold, hence it increases [13]. In 65-nm technology, this junction capacitance increase is about 30%.

In practical cases where both gate and junction capacitances contribute to the capacitive load at a given node, the above two opposite effects tend to compensate each other. In those rare cases where the gate or junction capacitance dominates (e.g., SRAM word line or bit line), the node capacitance experiences a change by 20–30% at most. Hence, the impact of ULV operation on capacitances is rather marginal, compared to above threshold [1].

Finally, it is useful to recall the impact of body biasing on the transistor junction capacitances (gate capacitance is unaffected by V_{BB} [13]). Under RBB, the source-bulk/drain-bulk junctions are more reverse biased compared to the case without body biasing, hence their capacitances are slightly smaller than the latter case (typically, by a few percentage points). On the other hand, under FBB, the junctions are forward biased and their capacitances are significantly higher (typically by up to 30–40%, for moderate values of V_{BB}) than the case without body biasing [13].

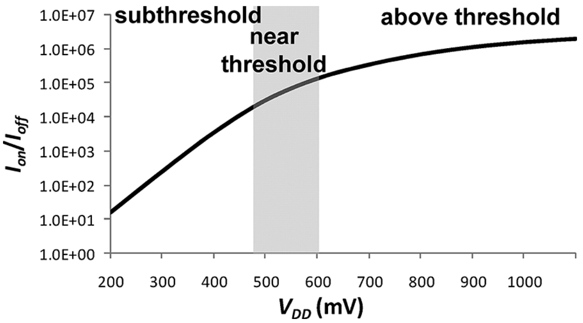


Fig. 7. $I_{\text{on}}/I_{\text{off}}$ vs. V_{DD}.

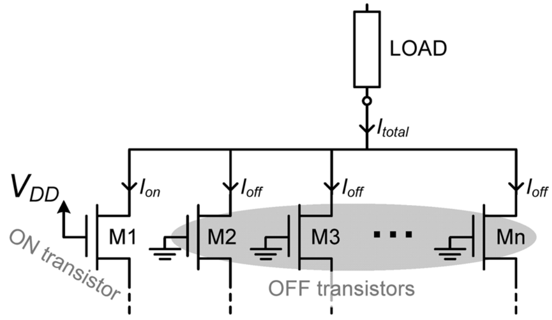


Fig. 8. Topology with n transistors connected to the same node.

B. Impact of ULV Operation on $I_{\text{on}}/I_{\text{off}}$: Design Considerations

Another side effect of ULV operation is the reduction in the ratio between the on and the off current $I_{\text{on}}/I_{\text{off}}$. Indeed, from (5) the off current is

$$I_{\text{off}} = \beta e^{\lambda_{DS} V_{DD}/n \cdot v_t}\left(1 - e^{-V_{DD}/v_t}\right) \approx \beta \tag{11}$$

hence from (9) and (11) the $I_{\text{on}}/I_{\text{off}}$ ratio results [21]

$$\frac{I_{\text{on}}}{I_{\text{off}}} = e^{V_{DD}/n \cdot v_t} \tag{12}$$

which exponentially depends on the supply voltage and also on technology (through slope factor n). For typical values of $n \sim 1.3 - 1.5$, from (12) the $I_{\text{on}}/I_{\text{off}}$ ratio reduces by 15×–20× for a 100-mV voltage reduction. As an example, $I_{\text{on}}/I_{\text{off}}$ for a 65-nm std-V_{TH} transistor is plotted versus V_{DD} in Fig. 7. Apparently, $I_{\text{on}}/I_{\text{off}}$ reduces by various orders of magnitude compared to above-threshold voltages [26], varies by up to three orders of magnitude in subthreshold, and is only few tens at $V_{DD} \sim 200$ mV.

The dramatic reduction in $I_{\text{on}}/I_{\text{off}}$ at ultra-low voltages has many consequences from a design point of view. As a first obvious consequence, the current of OFF transistors becomes significant compared to that of ON transistors, hence leakage has clearly a much stronger impact on power, compared to above-threshold designs (this aspect will be addressed in Sections VII and XI).

As a second important consequence of the $I_{\text{on}}/I_{\text{off}}$ degradation, circuit topologies consisting of many equal transistors connected to the same node as in Fig. 8 suffer from severe robustness degradation [1]. Indeed, correct operation requires that the on-current of the only transistor in the ON state (M1) dominates over the overall off-currents of all other transistors $(n-1)I_{\text{off}}$, in order to distinguish the high and low level of the overall current I_{total}. To this aim, the number n of transistors connected at a common node must be kept well below $I_{\text{on}}/I_{\text{off}}$ (say, by one-two orders of magnitude). Hence, from (12) the maximum number of connected transistors exponentially decreases when reducing V_{DD}. For example, $I_{\text{on}}/I_{\text{off}}$ of a few thousands is available at 300 mV, hence the maximum number of connected transistors is in the order of tens or a few hundreds. When V_{DD} is reduced to 200 mV, $I_{\text{on}}/I_{\text{off}}$ is a few hundreds, so the maximum number of transistor is reduced to multiple units or a few tens. This clearly imposes design constraints on practical circuits having a high fan-in or in memory arrays [1]. Practical design strategies to deal with these limitations will be presented in Section IX for logic, and Section X for memories.

C. Impact of ULV Operation on NMOS/PMOS Imbalance: Design Considerations

Another important issue observed in ULV circuits is the imbalance between the NMOS and PMOS strength. In general, NMOS/PMOS strength must be comparable to ensure adequate noise margin and reasonably symmetric rise-fall transitions [27].

At above-threshold voltages, the imbalance between NMOS and PMOS is not an issue, as the NMOS strength is typically twice that of PMOS transistor at same size. At ultra-low voltages, the NMOS/PMOS imbalance is typically much higher, thereby degrading the noise margin [28]. This can be understood by evaluating the "imbalance factor" IF in (13), which is defined as the strength ratio between the stronger and the weaker transistor, regardless of whether the stronger one is the PMOS or the NMOS [17]

$$IF = \max\left(\frac{\beta_p}{\beta_n}, \frac{\beta_n}{\beta_p}\right) \geq 1. \tag{13}$$

Depending on the specific adopted technology, the ratio β_p/β_n in (13) can be either greater or lower than 1, as opposed to above-threshold operation. Indeed, from (5b) the transistor strength is very sensitive to V_{TH}, hence a small (positive or negative) difference in the PMOS and NMOS threshold voltage can easily lead to a large difference in terms of strengths [26], [29], [30]. For the same reason, NMOS and PMOS transistors in subthreshold suffer from a strong imbalance (i.e., $IF \gg 1$). Conversely, matching their strength requires a considerable increase (more precisely, by a factor IF) in the strength of the weaker transistor.

In the specific case of the adopted 65-nm technology, the imbalance factor is $IF \approx 7$ since the NMOS strength is larger than PMOS by the same factor. As expected, this imbalance factor is much greater than that above threshold (which is found to be 1.8). Hence, a perfect NMOS/PMOS balance is obtained by increasing the PMOS strength by $IF \approx 7$, which can be achieved through the strategies discussed in Section III-B. According to Section III-B, applying the maximum amount[1] of FBB on PMOS and no body biasing on NMOS (i.e., connecting both body terminals to ground), the NMOS/PMOS imbalance is reduced to 3.5. The latter can be further reduced by increasing the PMOS strength through sizing, at the expense of larger capacitances and consumption.

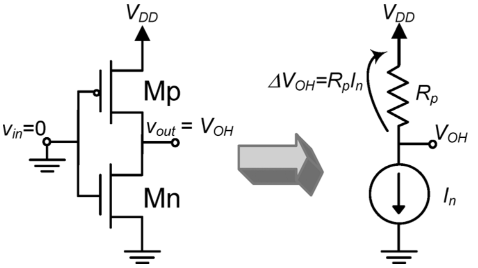


Fig. 9. Equivalent dc circuit of an inverter gate with low input.

The large PMOS/NMOS imbalance at ultra-low voltages has important consequences on the dc behavior of CMOS logic. This can be understood by considering the CMOS inverter gate in Fig. 9, where a low input and equally sized transistors are assumed. As the output voltage is high (i.e., close to V_{DD}), Mn has a large V_{DS} and is thus equivalent to a current source from Fig. 4(a), whereas Mp is equivalent to a resistance from Fig. 4(b). From the resulting equivalent model of the inverter in Fig. 9, the high output voltage V_{OH} is lower than V_{DD} by the voltage drop ΔV_{OH} across M2 [17], [21]:

$$V_{OH} = V_{DD} - \Delta V_{OH} \tag{14a}$$

$$\Delta V_{OH} = R_p I_n = v_t \frac{\beta_n}{\beta_p} e^{-V_{DD}/n_p v_t}. \tag{14b}$$

Similarly, swapping NMOS and PMOS in Fig. 9, the low output level V_{OL} is greater than the ground voltage by

$$V_{OL} = \Delta V_{OL} = v_t \frac{\beta_p}{\beta_n} e^{-V_{DD}/n_n v_t}. \tag{15}$$

From (14)–(15), the output levels are exponentially degraded at low voltages, and their values depend on the NMOS/PMOS strength ratio. In other words, standard CMOS logic at ultra-low voltages actually behaves like a "ratioed" logic style [27], as opposed to above-threshold standard CMOS logic. Clearly, this is an undesirable property of CMOS logic at ultra-low voltages, since it determines a degradation in the output logic levels and hence on the voltage swing [17]. As was shown in [17], the output level degradation for the same technology becomes significant (about 18%) for $V_{DD} \approx 100$ mV.

As a side effect, the degradation of the output levels due to the NMOS/PMOS imbalance determines an increase in the leakage power consumption of the subsequent logic gate. For example, a degradation ΔV_{OL} in the low output level of a given logic gate determines an equal increase in the gate-source voltage of the OFF NMOS transistor in the next logic gate. This translates into an exponential increase in its leakage current. For example, the previously mentioned 18% voltage level degradation at $V_{DD} \approx 100$ mV is found to determine a 40% leakage increase in the next logic gate.

V. PVT Variations and Impact on ULP Circuits

The above discussed issues are more critical when variations are considered. Variations can be classified into process

[1] Reverse body biasing would require the generation of voltages that are below ground or above V_{DD}. This would require additional boosting circuits like charge pumps and a higher design effort [55], which are typically impractical in ULP chips with tight constraints on the cost.

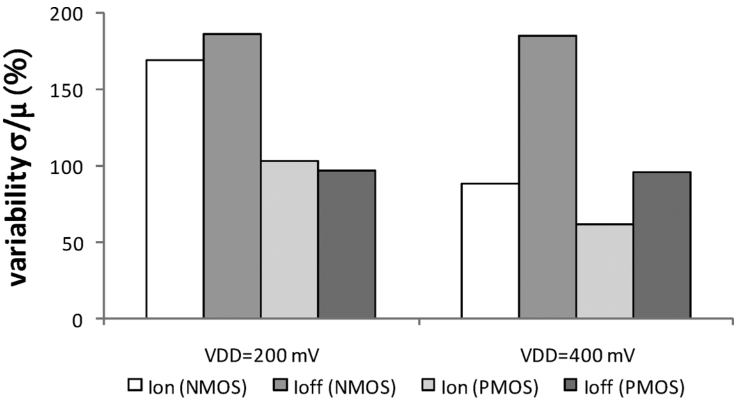


Fig. 10. Percentage variability $(\sigma/\mu) \cdot 100$ of NMOS/PMOS on- and off-current at $V_{DD} = 200\ \mathrm{mV}$ and $V_{DD} = 400\ \mathrm{mV}$.

(P), voltage (V) and temperature (T) variations [31]. In the following, variations in transistor parasitic capacitances will be neglected, since they are much smaller than those in the transistor current [31], [32].

A. *Review of Dominant Variability Sources in ULV Circuits*

Process variations mainly affect the transistor current through the variations in the threshold voltage V_{TH}, due to its exponential dependence in (5a). In turn, V_{TH} variations in subthreshold are mainly due to random dopant fluctuations (RDF) and are purely random (V_{TH} variations of different transistors are completely incorrelated) [31], [33], [34]. The latter fact is clearly undesirable, since existing feedback techniques to counteract fully correlated variations (see Section XIII) are ineffective in this case.

Voltage variations in ULP systems are due to only the fluctuations in the external supply, since the voltage drops across the on-chip supply distribution network are negligible (as I_{on} is orders of magnitude lower than above threshold). In battery-powered systems, the supply voltage is relatively constant and the resulting voltage variations are small and are further reduced through voltage regulation. On the other hand, battery-less systems suffer from much more pronounced supply variations, which are necessarily smoothed through the voltage regulation as well. As opposed to above-threshold systems, voltage variations of different subthreshold circuits on the same chip are fully correlated since the supply network is equi-potential. This enables the implementation of feedback schemes in Section VII to dynamically adjust the voltage [1].

Temperature variations in ULP systems are set by the room temperature, as there is no self-heating. Hence, practical chip temperatures are within reasonably narrow ranges. For example, implantable chips work at the almost constant body temperature, wearable sensors work at a relatively narrow temperature range. Only in some specific application, such as plant monitoring, sensor nodes might experience much larger temperature variations. As a peculiar property of ULV circuits, both I_{on} and I_{off} increase when increasing the temperature, as opposed to transistors operating above threshold.

Summarizing, process variations are dominated by purely random variations in the threshold voltage, and are the most harmful variability source. Voltage and temperature variations are instead fully correlated and can be further reduced through voltage regulation and feedback schemes (see Sections VII and XIII).

B. *Impact of Variations on Transistor Current and NMOS/PMOS Imbalance*

The above mentioned threshold voltage variations reflect into large variations in the strength β, due to its exponential dependence in (5b). For this reason, variations are a very critical issue in subthreshold (even more than above-threshold circuits) [1].

In general, the variations in β lead to variations in the leakage current I_{off} in (11) and the on-current I_{on} in (9). As an example in the reference 65-nm technology, Fig. 10 depicts the variability of the leakage current (i.e., the ratio between the standard deviation σ and the mean value μ, expressed as a percentage). From Fig. 10, the off-current variations are basically independent of V_{DD} as expected from (11). Observe that the large leakage current variations (up to more than 150%) are dominated by intradie purely random variations (i.e., RDF), whereas interdie variations were found to contribute by at most 30% in the reference technology. On the other hand, the on-current variability tends to be lower when increasing V_{DD} since threshold voltage variations in the low side lead to operation in near threshold rather than in subthreshold, thereby mitigating the increase in I_{on} [1].

As a further consequence of β variations, the imbalance factor IF in (13) is also subject to process variations. In the considered technology, the probability density function (PDF) of the imbalance factor from Monte Carlo simulations is plotted in Fig. 11. The shape of the PDF is similar to a lognormal distribution [1] (as one might expect from the exponential dependence of (5b) on V_{TH}), and has a mean value μ_{IF} of 13.9 and a standard deviation σ_{IF} of 25.6. Hence, the nominal imbalance factor $IF \sim 7$ (see Section IV-D) underestimates the mean value by a factor of two, which is explained by the asymmetric trend of the PDF in Fig. 11. Actually, the nominal IF underestimates its typical values by a much larger factor, due to its very large variability $\sigma_{IF}/\mu_{IF} \sim 185\%$ as well as the long tail at the right-hand side of the PDF in Fig. 11. More specifically, as shown by the cumulative distribution function (CDF) of IF in Fig. 12, the worst-case IF is 33 under a 90% level of confidence, and increases by up two orders of magnitude under typical levels of confidence, as shown in Table II.

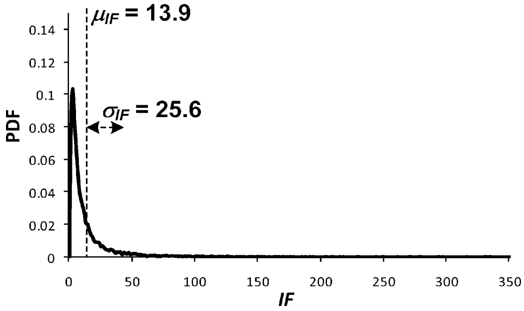


Fig. 11. Probability density function of the imbalance factor IF (10,000 Monte Carlo simulations under intradie and interdie variations).

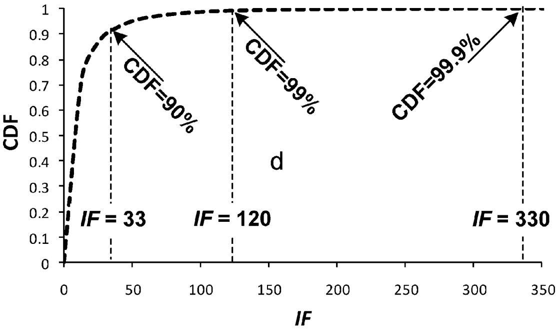


Fig. 12. Cumulative distribution function of the imbalance factor IF (10,000 Monte Carlo simulations under intradie and interdie variations).

TABLE II
WORST-CASE IMBALANCE FACTOR UNDER VARIATIONS FOR A GIVEN LEVEL OF CONFIDENCE (65-NM)

level of confidence	worst-case IF	increase w.r.t. nominal case
90%	33	5X
99%	120	17X
99.9%	330	47X
99.99%	600	86X
99.999%	1,200	171X
99.9999%	2,160	308X
99.99999%	3,730	533X
99.999999%	6,170	881X

(values for 99.99% or greater are extrapolated assuming a perfectly lognormal distribution for IF, due to the limited number of Monte Carlo runs)

From the above considerations, the NMOS/PMOS imbalance can be very high due to variations. From a circuit design point of view, this means that circuits and logic styles that explicitly rely on current contention (i.e., the relative strength of NMOS and PMOS) are very sensitive to variations and are thus unsuited for ULP/ULV applications [1]. For example, logic styles such as pseudo-NMOS ratioed logic are not a realistic option in ULV systems, unless under dramatic restrictions are introduced (e.g., only NAND gates with short-circuited gate/bulk) [35]. As another example, flip-flop topologies based on the current contention between the forward and the feedback path cannot be used in ULV applications [21]. Also, the robustness issue in multiple transistors connected to the same node (see Fig. 8) is exacerbated by variations. Indeed, the maximum number of connected transistors n must be further reduced to ensure that adequate robustness is achieved also in the worst case where I_{on} is reduced and I_{off} is increased, due to variations.

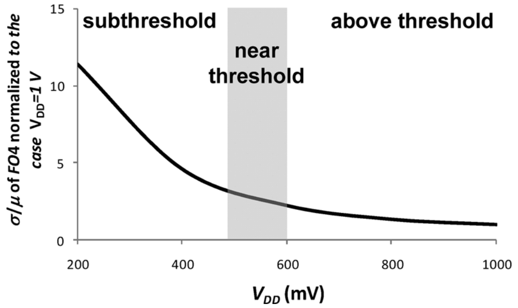


Fig. 13. $FO4$ variability (σ/μ) normalized to the case $V_{DD} = 1$ V vs. V_{DD}.

C. Impact of Variations on Performance

Process variations strongly affect also the gate delay, due to the variations in I_{on} and parasitic capacitances from (10). The gate delay variations are certainly dominated by I_{on} variations, as transistor parasitic capacitances have a much smaller variability. Accordingly, the variability of $FO4$ was found to be close to that of I_{on} in Fig. 10, and to be dominated by intradie variations (accounting for 90% of the overall variation).

The resulting gate delay variability is typically worse than above threshold by an order of magnitude [1]. Again, this is because of the exponential dependence of I_{on} on V_{TH}, as was discussed in Section V-A. This is shown in Fig. 13, which depicts the variability of $FO4$ normalized to the value at $V_{DD,\text{max}}$ versus V_{DD} in the adopted 65-nm technology. This performance degradation clearly adds up to the speed penalty that is already observed because of the exponential reduction of I_{on} at low voltages, as was discussed in Fig. 6.

VI. DC CHARACTERISTICS OF CMOS LOGIC AND PHYSICAL LIMITS TO ULTRA-LOW VOLTAGE OPERATION

The degradation of $I_{\text{on}}/I_{\text{off}}$, the NMOS/PMOS imbalance and the variations degrade the dc robustness of ULV CMOS logic. In the following, this degradation is analyzed under nominal conditions (Section VI-A) and including variations (subsequent subsections).

A. DC Robustness of ULV CMOS Logic: Inverter Gate

The dc characteristics of ULV CMOS logic gates severely degrades at ultra-low voltages [17], as was partially shown in Section IV-D by referring to the logic swing. For similar reasons, all other parameters of interest in the dc characteristics are considerably degraded at ultra-low voltages.

A V_{DD} reduction and stronger NMOS/PMOS imbalance both lead to a worse symmetry and wider transition region due to the reduced voltage gain, other than a degraded logic swing, as qualitatively summarized in Fig. 14. In regard to the symmetry, the analysis in [17] shows that the logic threshold of a CMOS inverter gate is

$$V_{LT} \approx \frac{V_{DD}}{2} \pm \frac{n}{2} v_t \ln(IF) \qquad (16)$$

where the plus (minus) sign holds when the PMOS (NMOS) is stronger than the NMOS (PMOS). As an example, the deviation in (16) from the ideal value $V_{DD}/2$ is in the order of $1.3 v_t \sim 30 - 35$ mV for typical values of $n \sim 1.3$ and $IF \sim 7$ found

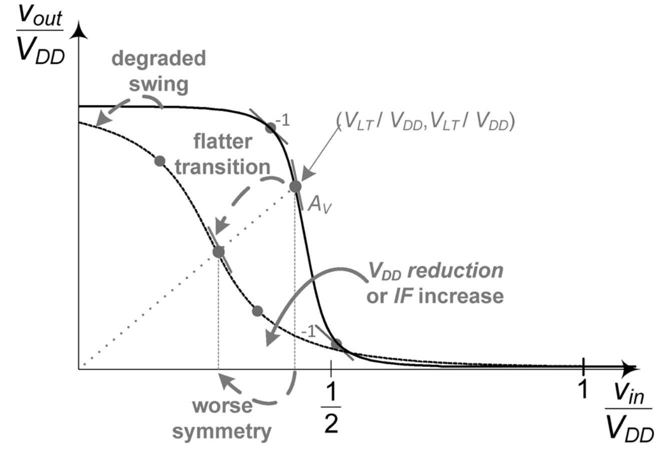


Fig. 14. Sketch of the degradation of the dc characteristics of standard CMOS gates at low voltages.

in the adopted technology. This shift in V_{LT} turns into a significant asymmetry in the dc characteristics, when the above deviation becomes a substantial fraction of $V_{DD}/2$. For example, at V_{DD} equal to 200, 150 and 100 mV this deviation degrades V_{LT} by 34%, 45% and 70%, respectively. In other words, the dc characteristics becomes intolerably asymmetric at ultra-low voltages and under strong NMOS/PMOS imbalance. Hence, to keep asymmetry within reasonable limits, a larger imbalance must be compensated by increasing V_{DD}, i.e., the imbalance sets the minimum operating voltage.

Regarding the steepness of the dc characteristics in the transition region, the magnitude of the voltage gain A_V around the logic threshold is [17]

$$A_{V1} \approx \frac{1}{\lambda + \frac{\frac{n}{2}}{\sqrt{\frac{1+e^{V_{DD}/(n \cdot v_t)}}{IF} - \frac{1}{2}}}} \tag{17}$$

where λ is the average NMOS/PMOS DIBL coefficient [17]. Again, A_v in (17) degrades at low V_{DD} and stronger imbalance. Since $A_V > 1$ must be ensured, a larger imbalance must be compensated with a higher V_{DD}, hence the imbalance sets again the minimum operating voltage.

As a result of the degradation of all the above dc parameters, the noise margin NM is severely impaired at low V_{DD} and large imbalance [17]. This is clear from the resulting expression of NM from the models in Section III [1], [17]

$$NM \approx \frac{V_{DD}}{2} - v_t - v_t \frac{n}{2} \ln(IF). \tag{18}$$

From (18), the noise margin of subthreshold CMOS logic gates consists of three contributions. The first term $V_{DD}/2$ represents the maximum possible noise margin which would be achievable under a full logic swing and a perfectly symmetric dc characteristics with an infinite voltage gain A_V. The actual NM in (18) is reduced by a thermal voltage, which is a fixed price that standard CMOS logic always has to pay for, when operating in subthreshold. This degradation might be mitigated only by resorting to alternative logic styles, as briefly discussed in Section XIV. Also, NM is further reduced by the third term, which is proportional to the logarithm of the imbalance factor IF: as expected, a larger imbalance leads to a worse noise margin. For example, for the adopted technology in nominal conditions, the noise margin suffers from a degradation $v_t + v_t(n/2)\ln(IF) \sim 60\,\mathrm{mV}$, compared to the ideal value $V_{DD}/2$. Observe that the NM degradation is dominated by the imbalance term, as expected.

TABLE III
NOISE MARGIN REDUCTION AND $V_{DD,\min}$ INCREASE DUE TO VARIATIONS VERSUS GATE COUNT (65-NM)

gate count $N_{gate} \rightarrow$	10 kgate	100 kgate	1 Mgate
level of confidence* $1 - \frac{1}{100 N_{gate}}$	99.9999%	99.99999%	99.999999%
worst-case IF (from Table II)	2,160	3,730	6,170
NM reduction w.r.t. nominal due to variations (mV)	197	215	233
$V_{DD,min}$ (room temperature)	$8v_t \approx$ $\approx 312\,mV$	$8.5v_t \approx$ $\approx 330\,mV$	$9v_t \approx$ $\approx 348\,mV$

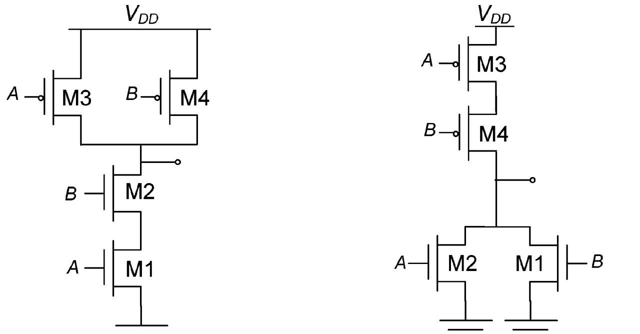
* based on assumption that only one gate is allowed to fail (i.e., it has higher IF than in the worst-case value in the table) within one chip out of 100

Summarizing, a larger imbalance must be necessarily compensated by increasing V_{DD} to keep all dc parameters within an acceptable range. Hence, the minimum possible supply voltage is set by the amount of imbalance, as will be discussed in detail in Section VI-D.

B. DC Robustness Degradation Under Variations: Inverter Gate

Process variations can be incorporated in the above analysis by simply considering the worst-case IF under a desired level of confidence (see Section V-B), instead of the nominal value [1]. For simplicity, let us analyze the simple but representative case of a chip consisting of inverter gates only, whose number is N_{gate}, as was proposed and implemented in a test chip in [36], [37] for the same purposes. As was discussed in Section V-B and Table II, the worst-case IF is set by the level of confidence, which has to be kept high in real designs where yield must be close enough to 100% for cost reasons. In particular, let us assume that IF is allowed to be greater than the worst-case value (i.e., the dc degradation is allowed to be greater than the maximum acceptable) only in one (malfunctioning) gate out of N_{gate} within the same chip, and that this is acceptable only in one chip out of 100. Hence, a level of confidence of $[1 - 1/(100 \cdot N_{\mathrm{gate}})]$ is required. The results for a 10-kgate, 100-kgate and 1-Mgate chip are reported in Table III. From this table, IF can be a couple of thousands due to variations, which translates into a further noise margin degradation in the order of 200 mV or more [1].

From the above observations, the imbalance factor is mainly set by process variations, whose effect strongly dominates over the intrinsic imbalance at nominal conditions. This clearly shows that variations are a very critical issue in ULP/ULV standard CMOS circuits since they dramatically impair the dc robustness of standard CMOS logic (i.e., the functional



A	B	PUN strength β_{PUN}	PDN strength β_{PDN}
0	0	-	$2\beta_p$
0	1	-	β_p
1	0	-	β_p
1	1	β_n/X_{stack}	-

(a)

A	B	PUN strength β_{PUN}	PDN strength β_{PDN}
0	0	-	β_p/X_{stack}
0	1	β_n	-
1	0	β_n	-
1	1	$2\beta_n$	-

(b)

Fig. 15. PUN/PDN strength versus input value in: (a) NAND2 gate; (b) NOR2 gate ($\beta_n = \text{strength of a single NMOS}$, $\beta_p = \text{strength of a single PMOS}$).

yield). This phenomenon is distinctive of ULV operation and adds to the variability issues in performance and energy (i.e., parametric yield).

C. Extension to Arbitrary Logic Gates

Arbitrary gates differ from the reference inverter only by the fact that the strength of the Pull-Up-Network (PUN) and Pull-Down-Network (PDN) depends on the inputs. When transistors are connected in parallel, the strength approximately ranges from the value pertaining to one transistor (when one is ON and the others are OFF) to the sum of all their strengths (when all of them are ON). When transistors are in series, their overall strength is lower than that of a single transistor by the well-known stacking factor X_{stack} [14]. All NMOS (PMOS) transistors are herein assumed to be equally sized, since adopting different sizes in stacked transistors as in [52] brings very limited and questionable advantages. As an example, the PUN/PDN strength for a NAND2 and NOR2 gate is depicted versus the input in Fig. 15.

At ultra-low voltages, the stacking factor for the on (off) current is greater (lower) than above threshold [17]. As an example, X_{stack} is plotted in Fig. 16 versus V_{DD} for two and three stacked transistors. From this figure, X_{stack} for the on (off) current is higher (lower) than in above threshold by about 1.2× (1.5–2×). As a result, the $I_{\text{on}}/I_{\text{off}}$ ratio in stacked transistors is further degraded compared to a single transistor. Similar considerations hold for parallel transistors, since in the worst case the I_{on} is the same as a single transistor, but the leakage can be up to the sum of the leakage of all parallel transistors [1].

In complex gates, the NMOS/PMOS imbalance problem translates into a PUN/PDN imbalance, whose negative impact (see Section IV-D) is more pronounced when the fan-in is increased. This is easily understood in logic gates having series transistors of the weak type and parallel transistors of the strong type. In our technology, this case corresponds to the series connection of PMOS transistors and parallel NMOS transistors, as in the case of NOR2 gate in Fig. 15. In this case, PMOS stacking (NMOS parallel connection) makes the PUN even weaker (PDN even stronger) than an inverter with same sizes, thereby emphasizing the PUN/PDN imbalance problem. Clearly, this issue is more critical in large fan-in gates with a larger number of stacked (parallel) transistors in the PUN (PDN) [17]. From a design perspective, this means that logic gates with stacked transistors of the weak type should be avoided, especially if they have a large fan-in.

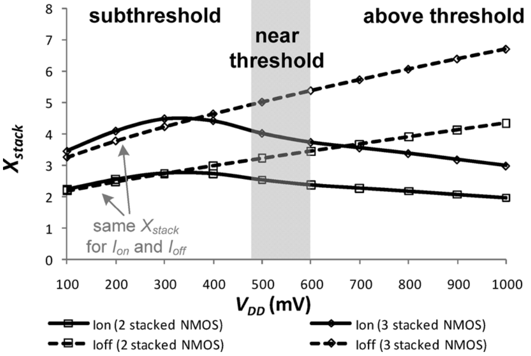


Fig. 16. Stacking factor for on and off current for 2 and 3 stacked NMOS transistors.

Summarizing, the issues related to the degradation in $I_{\text{on}}/I_{\text{off}}$ and imbalance are exacerbated in large fan-in gates [17], due to the presence of series and parallel transistors. This justifies why only low fan-in cells are usually kept in ULP designs, as was discussed for example in [21], [38], where the maximum adopted fan-in was only two. Analysis of these issues will be presented in the representative cases of CMOS standard cells (Section IX) and memories (Section X).

D. Physical Limits to Ultra-Low Voltage Operation

The degradation of the dc characteristics of CMOS gates at ultra-low voltages sets a practical lower bound $V_{DD,\min}$ to the supply voltage [1]. Practical circuits should never work at $NM < 0$, hence the very minimum supply voltage $V_{DD,\min}$ is that leading to $NM = 0$ (practical voltages must be higher to ensure a desired positive NM). Accordingly, by inverting (18) and setting $NM = 1$, $V_{DD,\min}$ results to [17]

$$V_{DD,\min} = nv_t \left[\ln\left(\frac{2}{n}\right) + 1 \right] + nv_t \ln(IF). \tag{19}$$

The previous empirical $V_{DD,\min}$ definition in [21] (i.e., the voltage that leads to 10% voltage swing degradation) provides similar numerical results as (19) [17]. In other words, practical logic gates with positive noise margin have a lower than 10% swing degradation.

From (19), the ideal $V_{DD,\min}$ under perfect balance is about $2v_t$, which agrees with the most recent predictions [29], [39]. Under nonzero imbalance, this limit can be approached by tuning the threshold voltages through adaptive body biasing [30]. Previous predictions about $V_{DD,\min}$ in the literature converged to the above value through progressively lower values, starting from $8v_t$ in [26] and going through $3 - 3.5v_t$ in [40], as a few examples. A historical survey of such predictions can be found in [21].

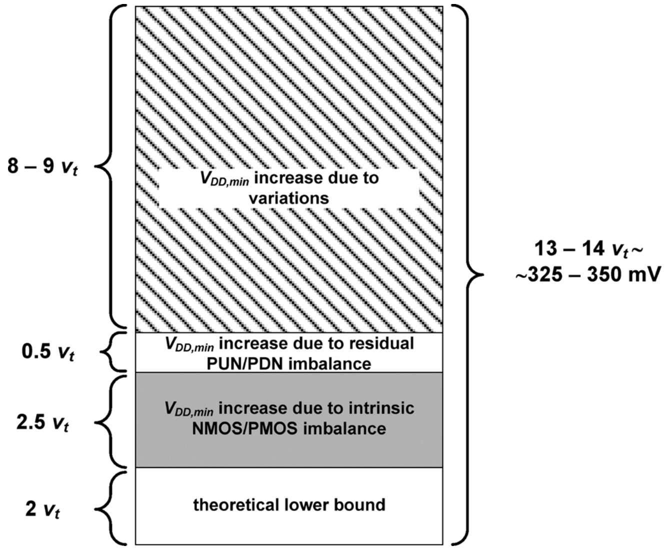

8 – 9 v_t
$V_{DD,min}$ increase due to variations
0.5 v_t
$V_{DD,min}$ increase due to residual PUN/PDN imbalance
2.5 v_t
$V_{DD,min}$ increase due to intrinsic NMOS/PMOS imbalance
2 v_t
theoretical lower bound
13 – 14 v_t~
~325 – 350 mV

Fig. 17. Breakdown of $V_{DD,\min}$ expressed in thermal voltages.

TABLE IV
IDEAL ENERGY REDUCTION OBTAINED THROUGH V_{DD} REDUCTION W.R.T. $V_{DD,\max} = 1$ V

V_{DD}	energy reduction w.r.t. $V_{DD,max}$
125 mV	64X
250 mV	16X
375 mV	7X

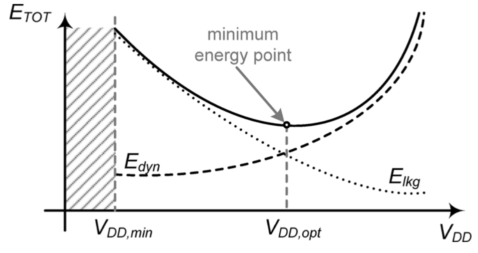


Fig. 18. Trend of energy contributions and overall energy vs. V_{DD}.

In practical circuits where imbalance takes place, $V_{DD,\min}$ significantly increases due to the dependence on IF from (19). When variations are not considered, $V_{DD,\min}$ for an inverter gate results to $4.5v_t \sim 115$ mV for $IF \sim 10$. Variations lead to a further increase in $V_{DD,\min}$, due to the increase in the worst-case value of IF [41]. For example, including the noise margin degradation in Table III, from (19) variations determine a $V_{DD,\min}$ increase by $8 - 9v_t \sim 220$ mV for a fairly wide range of gate counts. This $V_{DD,\min}$ increase is slightly more pronounced under higher gate count (e.g., $8v_t$ for 10 kgates, $9v_t$ for 1 Mgates). Intuitively, this is because the minimum supply of a system is set by the cell having the highest $V_{DD,\min}$, hence it is more likely that some gate experiences a very large IF (and hence $V_{DD,\min}$) increase if the gate count is increased [1].

Fig. 17 summarizes all contributions to $V_{DD,\min}$ according to (19) and previous discussion, including the small contribution due to the PUN/PDN imbalance due to the dependence of their strength on the input (as discussed in Section IV-C – see Section IX for its explicit evaluation). Fig. 17 clearly shows that variations are the dominant contribution to $V_{DD,\min}$. In other words, the voltage limit in ULV circuits is basically set by variations. This justifies the experimental results in [36], [37] based on long chains of inverters connected as a ring oscillator, whose measurements on multiple chips agree very well with predictions in Table III.

From the above considerations, variations play a major role in determining the real $V_{DD,\min}$ of a system. Hence, $V_{DD,\min}$ obtained in a single prototype (as in most of the work published so far) is not representative of the typical $V_{DD,\min}$ encountered in mass-production chips. This means that the functional yield at ultra-low voltages is actually rather low, and on average chips need a rather high $V_{DD,\min}$ in the order of 300–350 mV from Fig. 17. Considerations on yield will be given in Section VIII.

Finally, it is interesting to observe that the cells that are expected to set the V_{DD} lower bound in a VLSI system are those which either intrinsically have a large imbalance, or are particularly sensitive to such imbalance. In regard to the former, in Sections IV and VI it was shown that imbalance tends to be larger when increasing the fan-in, hence combinational gates with high fan-in tend to have a larger $V_{DD,\min}$. On the other hand, a strong sensitivity to imbalance is observed in cells that inherently involve current contention, as in the case of topologies based on feedback, i.e., different circuit sections driving the same node (e.g., flip-flops, SRAM memory cell). In other words, $V_{DD,\min}$ of VLSI digital systems is typically limited by storage elements as well as combinational gate with high fan-in. Hence, these classes of cells tend to further raise $V_{DD,\min}$ of VLSI systems, compared to the values that were found with inverters only (as in Fig. 17) [1]. Hence, to keep this $V_{DD,\min}$ increase within reasonable limits, the composition of ULP standard cell libraries must be carefully chosen, as discussed in Section IX.

VII. ENERGY/POWER OPTIMIZATION AND TRADEOFF WITH PERFORMANCE

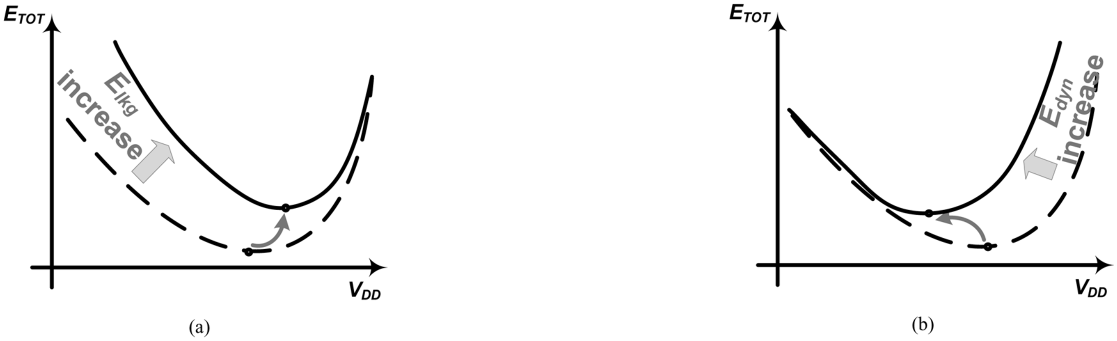
In this section, energy/power dependence on process and design parameters is discussed under nominal conditions.

A. Minimum Energy per Operation

The overall energy per clock cycle of a VLSI system consists of two main contributions, the dynamic and the leakage energy. The energy contribution due to short-circuit currents is usually negligible, due to the steep (exponential) MOS I-V characteristics. The dynamic energy per clock cycle of a VLSI system, module or logic gate is

$$E_{\mathrm{dyn}} = C_{\mathrm{eff}} V_{DD}^2 = \alpha_{sw} C_{\mathrm{TOT}} V_{DD}^2 \tag{20}$$

where C_{eff} is the effective capacitance, C_{TOT} is the total physical capacitance and α_{sw} is the activity factor [27]. As usual, the reduction in V_{DD} leads to a squared reduction in E_{dyn}, and an additional energy reduction (10–20%) is obtained from the reduction of C_{TOT}, as discussed in Section IV-B. Compared to the case with $V_{DD,\max} = 1$ V, operation at ultra-low voltages enables a dynamic energy reduction by up to two orders of magnitude, as reported in Table IV.



(a) (b)

Fig. 19. Qualitative description of the MEP shift due to an increase in (a) leakage energy and (b) dynamic energy.

The energy per clock cycle due to the leakage of a VLSI (synchronous) system, module or logic gate is given by

$$E_{\rm lkg} = V_{DD} I_{\rm off} T_{CK} = V_{DD} I_{\rm off} \tau_D \frac{LD}{X_{\rm stack}} \qquad (21)$$

where the clock cycle T_{CK} was expressed as the product of the average gate delay τ_D within the critical path and the number of cascaded logic gates LD (i.e., the logic depth, which is set by the adopted microarchitecture, as discussed in Section XII.). Also, the single-transistor leakage is reduced by the average stacking factor $X_{\rm stack}$ across the cells used to implement the circuit under analysis.

Focusing on the first three factors in (21), a reduction in thesupply voltage leads to a linear reduction in the first factor, a slight reduction in $I_{\rm off}$ thanks to the DIBL effect (see (11)), and an almost exponential increase in the gate delay from (10). As a consequence, the leakage energy per cycle in (21) tends to increase almost exponentially when reducing V_{DD}, i.e., it has an opposite trend compared to the dynamic energy, as shown in Fig. 18. This justifies the presence of a minimum-energy point (MEP) that arises from an optimum balance between dynamic and leakage energy. As these trends are observed with $V_{DD} < V_{TH}$, the minimum energy point certainly occurs in subthreshold. In published prototypes, $V_{DD,opt}$ was shown to be in the order of 300–400 mV for a wide range of circuits (see, e.g., [42]–[51]).

Due to the above reported experimental results, it is generally believed that operation at voltages lower than 300–350 mV is not of interest in practical ULP applications, and that lowering $V_{DD,\rm min}$ below these values is not a concern in the design. This common belief is actually incorrect for various reasons, which will be justified in detail in the next sections. In short: $V_{DD,\rm min}$ at nominal conditions has to be significantly lower than the above mentioned values to 1) ensure an adequate functional yield at the MEP in the presence of variations (see Section VIII-A), 2) obtain the very minimum energy point in systems with optimized microarchitecture (see Section XII), 3) enable ultra-low power operation in blocks that are always on (see Section VII-D).

B. Dependence of MEP on Dynamic and Leakage Energy

To better understand the main properties of the MEP, let us consider the case where either the dynamic or the leakage energy is increased, while keeping the other constant. When $E_{\rm lkg}$ is significantly increased, the minimum energy clearly increases, and hence the MEP moves up. At the same time, the optimum supply voltage moves to the right, as a V_{DD} increase enables the reduction in the dominant leakage energy (see Fig. 18). On the other hand, if dynamic energy significantly increases, the MEP shifts up again, but moves to the left since reducing V_{DD} leads to a reduction in the dominant dynamic energy. These considerations are summarized in Fig. 19.

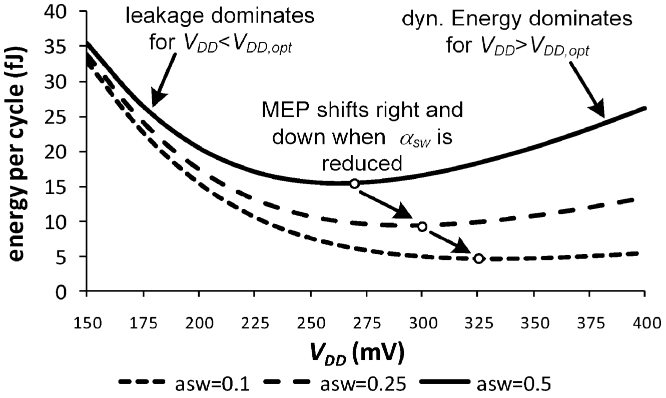


Fig. 20. Energy per cycle vs. V_{DD} in a combinational block with $120FO4$ logic depth and activity factor ranging from 0.1 to 0.5.

As a simple but representative circuit example, let us consider a circuit consisting of cascaded inverters, each of which has a fan-out of 4 (so that the number of cascaded inverters defines the logic depth LD in terms of $FO4$). The resulting energy per operation is plotted versus V_{DD} in Fig. 20 for $LD = 120FO4$ and α_{sw} equal to 0.1, 0.25, and 0.5. Fig. 20 shows that the energy curve versus V_{DD} has typically a rather flat minimum, hence energy can be kept close to the MEP also in the presence of V_{DD} uncertainties. Also, the MEP in a given circuit significantly depends on the activity factor. For example, the minimum energy reduces by 3× if α_{sw} reduces from 0.5 to 0.1, and the optimum supply voltage $V_{DD,opt}$ increases from 275 to 325 mV. This is because a reduction in α_{sw} determines a reduction in $E_{\rm dyn}$ from (20), which in turn leads to a right-down shift of the MEP from Fig. 19(b) (consider the opposite arrow direction). The energy curve versus V_{DD} has a flatter minimum when dynamic energy is reduced, since energy starts increasing very slowly at the right of the MEP.

According to the above considerations, $V_{DD,opt}$ strongly depends on the balance between leakage and dynamic energy, which in turn is strongly affected by changes in the activity factor, workload and environmental conditions (e.g., temperature) over time. Hence, operation at a given energy point (e.g., the MEP) requires appropriate feedback schemes that keep the

TABLE V
IMPACT OF VOLTAGE SCALING AND BODY BIASING ON PERFORMANCE AND ENERGY IN TYPICAL DESIGN SCENARIOS (FROM (10), (20) AND (21))

		design scenario	
		$E_{dyn} \gg E_{lkg}$	$E_{lkg} \gg E_{dyn}$
impact of voltage scaling (fixed V_{BB})	performance $1/\tau_D$ vs. V_{DD}	$\propto e^{\frac{V_{DD}}{n \cdot v_t}}$	$\propto e^{\frac{V_{DD}}{n \cdot v_t}}$
	energy E vs. V_{DD}	$E_{dyn} \propto V_{DD}^2$	$E_{lkg} \propto e^{-V_{DD}/n \cdot v_t}$
impact of body biasing (fixed V_{DD})	performance $1/\tau_D$ vs. V_{BB}	$\propto e^{\frac{\lambda_{BS} V_{BB}}{n \cdot v_t}}$	$\propto e^{\frac{\lambda_{BS} V_{BB}}{n \cdot v_t}}$
	energy E vs. V_{BB}	$E_{dyn} \propto C$ (see note below*)	~independent of V_{BB}

* Under RBB (FBB) $E_{dyn} \propto C$ is independent of V_{BB} (increases when increasing V_{BB}).

TABLE VI
SUMMARY OF RUN-TIME STRATEGIES FOR ENERGY-EFFICIENT PERFORMANCE TUNING (FROM TABLE V)

	design scenario	
	$E_{dyn} \gg E_{lkg}$	$E_{lkg} \gg E_{dyn}$
increase performance (min. energy penalty)	NMOS: increase V_{BB} PMOS: reduce V_{BB}	increase V_{DD}
reduce performance (max. energy reduction)	reduce V_{DD}	NMOS: reduce V_{BB} PMOS: increase V_{BB}

system at the desired energy point despite these changes, as will be discussed in Section XIII.

It is worth noting that leakage plays a fundamental role in ULP circuits. Indeed, leakage is responsible for the energy increase at low voltages (i.e., at the left of the MEP) and hence determines the minimum energy point. If leakage were negligible, from (20) the energy could be monotonically reduced by reducing V_{DD}, and the minimum energy would be essentially set by $V_{DD,\text{min}}$. In other words, leakage sets the energy lower bound and is the dominating energy contribution at ultra-low voltages (i.e., at the left of the MEP), as depicted in Fig. 20. A review of techniques to counteract leakage is presented in Section XI.

C. Knobs to Tune Performance at High Energy Efficiency

Previous considerations were derived assuming that energy has to be minimized at all costs, regardless of performance. In many applications, performance is constrained and its requirement dynamically varies over time. In these cases, performance has to be raised (reduced) at run time with minimal energy penalty (maximum energy saving), compared to a given operating point.

As discussed in Sections III-B and IV-A, the most powerful knobs to tune the performance at run time are the supply voltage (dynamic voltage scaling) and the body bias voltage (dynamic body biasing). The impact of the two techniques on performance and energy is summarized in Table V, where their expressions are derived by approximating[2] (10), (20) and (21) in the two typical scenarios where either the dynamic or the leakage energy dominates the overall consumption (i.e., for inputs respectively leading to high or low activity in (20)). The trends summarized in this table provide simple guidelines on how to tune performance at run time in the most energy-efficient way, when moving from a previous operating point, as discussed in the following.

When $E_{\text{lkg}} \gg E_{\text{dyn}}$ (i.e., low activity), performance should be increased by increasing V_{DD},, since this leads to both an exponential performance increase (proportional to $e^{V_{DD}/n \cdot v_t}$) and an exponential decrease in energy (since it is approximately $E_{\text{lkg}} \propto e^{-V_{DD}/n \cdot v_t}$). On the other hand, excess of performance should be reduced by reducing V_{BB} in NMOS transistors, since this has an exponential impact on performance (proportional to $e^{\lambda_{BS} V_{BB}/n \cdot v_t}$) and no energy penalty at first order (dual considerations hold for PMOS transistors). Interestingly, in this case a reduction in V_{DD} would have degraded both performance and energy, since the dominating leakage energy contribution tends to rapidly increase when reducing V_{DD}. This is significantly different from traditional above-threshold designs, and confirms that the choice of the run-time knob to tune performance is critical in terms of energy efficiency.

In the opposite scenario where $E_{\text{dyn}} \gg E_{\text{lkg}}$ (i.e., high activity), performance should be increased by increasing V_{BB} in NMOS transistors, since this leads to an exponential performance increase (proportional to $e^{\lambda_{BS} V_{BB}/n \cdot v_t}$) and a fairly small energy increase. More specifically, the energy is dominated by the dynamic energy in (20), which is only moderately affected by V_{BB} through the transistor junction capacitance C. Indeed, the latter approximately independent of V_{BB} in the case of RBB, and can typically increase by up to 30–40% in the case of FBB, as discussed in Section IV-B. This explains the experimental results in [38], where a circuits with 30 cascaded inverters was found to experience a 10% reduction in energy under RBB, and a 40% energy increase under FBB. Dual considerations hold when performance is reduced to save energy.

The above discussed design guidelines are summarized in Table VI, which fully justify the experimental results in [38]. Also, this shows that ULP systems with time-varying performance requirement actually require the run-time tuning of both V_{DD} and V_{BB} [1].

D. Considerations on the Power Consumption

As discussed in Section II-A, most of the blocks within an ULP VLSI system are duty cycled, hence they should be designed to minimize energy. However, in ULP applications with moderate to high wakeup period, the average power of always-on is also a serious concern.

The dynamic power of always-on circuits is reduced by adopting very low V_{DD}, which also makes it easier to keep the always-on clocks slow (at higher voltages, oscillators are faster and clock has to be divided, which entails some power penalty). Reduction in V_{DD} also leads to a decrease in the leakage power $P_{\text{lkg}} = V_{DD} I_{\text{off}} \propto V_{DD} \cdot \beta \cdot e^{\lambda_{DS} V_{DD}/(n \cdot v_t)}$ from (11). For very low voltages such that $V_{DD} \ll n \cdot v_t/\lambda_{DS}$ (i.e., in the order of 150 mV or lower), from (11) the leakage power of an always-on logic gate is

$$P_{\text{lkg}} \approx V_{DD} \cdot \beta. \tag{22}$$

[2] Indeed, from (10) we get $1/\tau_D \propto e^{V_{DD}/n \cdot v_t}/V_{DD} \sim e^{V_{DD}/n \cdot v_t}$. Analogously, from (21) we get $E_{\text{lkg}} \propto V_{DD}^2/e^{V_{DD}/n \cdot v_t} \sim e^{-(V_{DD}/n \cdot v_t)}$, considering that the stacking factor is roughly independent of V_{DD} (see Fig. 16 and Section XI) and I_{off} is essentially independent of V_{DD} at low voltages from (11).

Since always-on circuits usually operate at very low frequencies, their overall power is approximately the sum of the leakage contributions in (22) for all always-on logic gates. From (22), V_{DD} has to be kept as small as possible, hence the lower limit is set by $V_{DD,\min}$. At the same time, the strength β has to be kept as small as possible by adopting a high V_{TH} and avoiding FBB (RBB could also be an option). In regard to the transistor size, narrow transistors should be used in always-on blocks, while keeping the channel length significantly greater than $L_{\min}$ to reduce the subthreshold slope (i.e., leakage) and variations, as discussed in Section IX.

From the above considerations, always-on circuits have to operate at voltages much lower than $V_{DD,opt}$ (i.e., well at the left of the MEP in Fig. 18), and the potential power reduction is limited by the minimum supply voltage. Hence, minimizing $V_{DD,\min}$ is key in always-on blocks, as was anticipated in Section VII-A (this issue is often overlooked in the literature). This means that keeping a good NMOS/PMOS balance in always-on circuits is extremely important, according to Section VI-D. Design criteria to achieve this important goal will be discussed in Sections IX–X for logic and memories.

VIII. Impact of Variations and NMOS/PMOS Imbalance on Energy/Power

In this section, the considerations presented in the previous section are extended by including process variations (Sections VIII-A–VIII-B), voltage/temperature variations (Section VIII-C) and imbalance (Section VIII-D).

A. Impact of Process Variations on Energy and Power Consumption (Parametric Yield)

Practical ULP systems have to meet the required average power despite of variations. From (2), process variations impact P_{avg} through both E_{active} and $P_{\text{always-on}}$ (T_{wkup} is set by the application and might need to be tuned).

Regarding E_{active}, let us separately analyze the dynamic and leakage energy. E_{dyn} in (20) is affected by process variations through parasitic transistor capacitances, whose intradie and interdie variations are negligible as discussed in Section V. On the other hand, the leakage energy contribution of E_{active} is insensitive to interdie variations and is fairly sensitive to intradie variations [1]. More specifically, from (10)–(11) interdie variations in V_{TH} do not significantly affect E_{lkg} in (21) since they determine an increase in the delay and the same decrease in the leakage of all cells[3]. Intra-die variations affect E_{lkg} through two different mechanisms: one is through the leakage variations, the other is indirectly due to the variation gate delay variations. The former effect is negligible since these variations are averaged over the large number of cells within the circuit (roughly, the overall leakage variability is proportional to $1/\sqrt{N_{\text{gate}}} \ll 1$). Regarding the latter effect, intradie gate delay variations require that T_{CK} is increased to include some margin to account for variations (see Section XIII for details), thereby increasing E_{lkg} from (21). For this reason, as will be discussed in Section XIII, appropriate adaptive feedback techniques should be adopted in ULP systems to keep variations under control and ensure an adequate parametric yield. As opposed to traditional high-performance designs, adaptive schemes are usually adopted to minimize leakage energy, rather than speed variations.

Finally, $P_{\text{always-on}}$ is significantly affected only by interdie variations, as intradie variations average out over a large number of cells. Hence, from (22) the variability of $P_{\text{always-on}}$ of always-on circuits is basically the same as the variability in I_{off} of a single device ($\sim 80\%$ in the adopted technology). Accordingly, in (2) the highest variability is experienced by $P_{\text{always-on}}$, provided that delay variations are mitigated through an appropriate microarchitecture and adaptive run-time techniques, as discussed in Section XII.

B. Relationship Between $V_{DD,\min}$ and $V_{DD,opt}$ (Functional Yield)

As discussed in Section VI-D, it is generally believed that minimization of $V_{DD,\min}$ at nominal conditions makes no sense, since $V_{DD,opt}$ is expected to be higher anyway. In the same section, it was clarified that this is not true in many real designs, hence $V_{DD,\min}$ minimization is actually an important objective in ULP design. In particular, $V_{DD,opt}$ and $V_{DD,\min}$ are shown in the following to be related by yield considerations.

To better understand the relationship between $V_{DD,\min}$ at nominal conditions and $V_{DD,opt}$, we should observe that the usual definition of the MEP focuses on energy only. However, yield should be incorporated in the definition of MEP in mass-produced circuits. Thus, a more appropriate definition of the MEP is the minimum-energy point that is actually reachable by an assigned percentage of manufactured chips. This definition goes beyond the previous attempts to evaluate the MEP through the occasional characterization of single prototypes.

As was discussed in Section IV, ULV operation degrades the ability of logic gates to fully switch, especially under variations that strongly degrade the NMOS/PMOS imbalance. According to Section VI, from a statistical point of view, the percentage of logic gates that correctly switch (which sets the functional yield) increases when increasing V_{DD}. As qualitatively shown in Fig. 21, the minimum supply voltage $V_{DD,yld}$ that meets a given yield target is typically higher than V_{DD}, opt, hence the MEP is actually reachable only by a minority of manufactured chips. This is even truer when considering designs with particularly low $V_{DD,opt}$ (see Section VI-D), since yield further degrades when reducing the voltage. This significant degradation in the functional yield due to the large impact of variations on the dc behavior of CMOS logic gates is a distinctive feature of ULP VLSI systems, and easily dominates over the traditional contribution of defects. This requires the specific development of yield-aware design methodologies for ULP applications [1], which is currently a very open research field.

The impact of $V_{DD,\min}$ on functional yield can be intuitively grasped from Fig. 21. Indeed, to make the MEP reachable at the given yield target, $V_{DD,yld}$ in Fig. 21 has to be moved to the left. Intuitively, this can be done by moving the yield curve in Fig. 21 to the left, which in turn requires the reduction of $V_{DD,\min}$ at nominal conditions (see arrow in Fig. 21). In other words, a lower $V_{DD,\min}$ gives more room to tolerate variations around the MEP, thereby providing an improvement in the functional yield at the MEP. This also suggests that nominal $V_{DD,\min}$ and

[3] It is reasonably assumed that the clock cycle scales like the gate delay. This is certainly true if the clock is generated on chip through ring oscillators, so that the clock oscillator period tracks the gate delay.

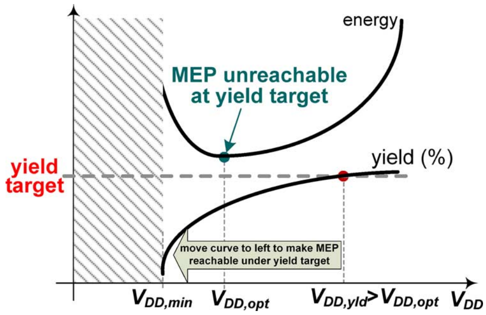


Fig. 21. Qualitative trend of yield vs. V_{DD} showing that the yield target point is typically at the right of the MEP (i.e., MEP is not reachable at targeted yield).

$V_{DD,opt}$ are actually strongly related, as opposed to the common belief, since the distance between the two voltages is actually the available margin to accommodate for variations and avoid functional failure at the MEP [1] (or any other energy target). Further investigation is needed to gain a more quantitative understanding of this usually neglected relationship between $V_{DD,\min}$ and $V_{DD,opt}$, as well as to better understand the underlying tradeoffs between functional yield and energy.

C. Impact of Voltage/Temperature Variations on Energy and Power Consumption

Around the MEP, voltage variations have a negligible effect on the energy of duty-cycled blocks, due to the flat minimum. Around different energy points, energy sensitivity to V_{DD} is still rather small (typically 1–2). Analogously, the sensitivity of $P_{\rm lkg}$ is approximately equal to one ($P_{\rm lkg} \propto V_{DD}$ from (22)), hence voltage variations are not an issue in terms of $P_{\rm lkg}$ variability.

Temperature variations have a negligible effect on the dynamic energy from (20). This also holds for the leakage energy, since an increase in temperature determines a decrease in the threshold voltage, and hence an increase in the transistor strength β. In turn, this determines an exponential increase in the leakage current $I_{\rm off}$ from (5b) and (11), and an equal decrease in the gate delay τ_D (and hence T_{CK}). Accordingly, the leakage current and the clock cycle variations due to temperature shifts compensate each other, so that the leakage energy in (21) is basically independent of the temperature. On the other hand, the leakage power is strongly influenced by temperature variations, due to the exponential increase of β due to a temperature increase. Typically, $P_{\rm lkg}$ varies by three orders of magnitude when the temperature ranges from -40 to 125 °C. Hence, achieving ultra-low power in applications that require operation at high temperature is extremely difficult.

Summarizing, the energy of duty-cycled blocks is rather insensitive to voltage and temperature variations. The power consumption of always-on blocks in (22) is relatively insensitive to voltage variations, whereas it is very sensitive to temperature variations. This is acceptable in applications where temperature is relatively constant (like implantable chips or indoor sensors), whereas some feedback compensation scheme might be needed in other cases.

D. Impact of NMOS/PMOS Imbalance on Leakage Energy

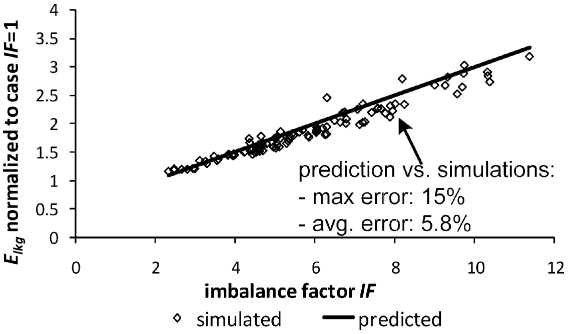


Fig. 22. Leakage energy $E_{\rm lkg}$ normalized to the case $IF = 1$ vs. IF (reference circuit: logic depth $120FO4$, $\alpha_{sw} = 0.1$).

The leakage energy per cycle significantly depends on the NMOS/PMOS imbalance and is minimum under perfect balance [38]. In the latter reference, this was shown by referring to the simple but representative case of equal cascaded inverters, whose leakage energy depends on the transistor strength according to

$$E_{\rm lkg} = V_{DD} I_{\rm off} \tau_D \frac{LD}{X_{\rm stack}} \propto (\beta_n + \beta_p) \cdot \left(\frac{1}{\beta_n} + \frac{1}{\beta_p}\right) \quad (23)$$

where it was observed that $I_{\rm off} \propto (\beta_n + \beta_p)/2$ (assuming the same probability for high and low input in (11)). To correctly interpret the results in [38], let us observe that in (23) the gate delay in (10) was implied to depend only on β_n and β_p, assuming a constant load C (i.e., for an assigned transistor aspect ratio[4]). In other words, (23) holds if β_n and β_p are tuned without changing the transistor size (e.g., through body biasing or V_{TH} selection, according to Section III-B). Under this assumption, from (23) $E_{\rm lkg}$ is minimum when $\beta_n = \beta_p$, i.e., when NMOS and PMOS are perfectly balanced.

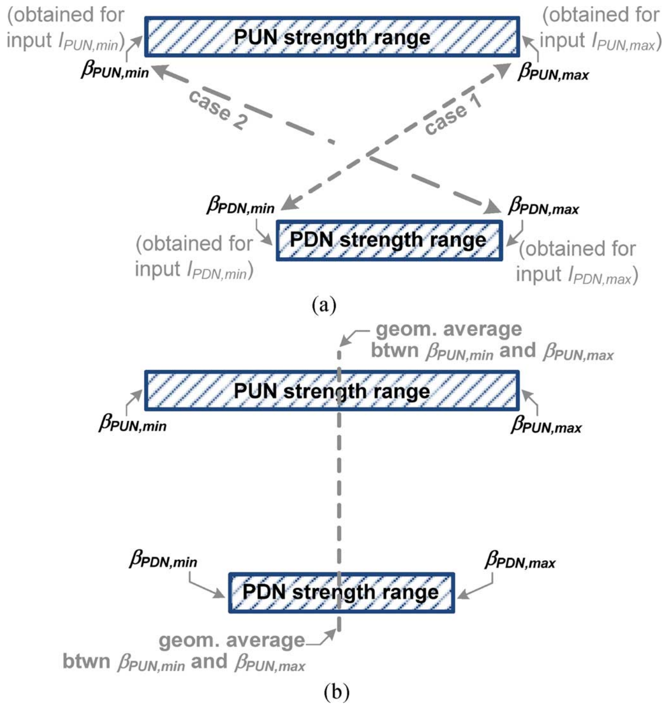
In practical cases where NMOS and PMOS are not perfectly balanced (i.e., $IF > 1$), from (23) the leakage energy increase with respect to the ideal case $IF = 0$ results to [1]

$$\frac{E_{\rm lkg}|_{IF>1}}{E_{\rm lkg}|_{IF=1}} \propto \frac{2 + IF + \frac{1}{IF}}{4} \approx \frac{1}{2} + \frac{IF}{4} \quad (24)$$

where the right-hand side holds if $IF \gg 1$. Equation (24) is a useful tool to predict the impact of imbalance at module or system level. As an example, the normalized leakage energy in (24) is plotted versus IF in Fig. 22 for the reference circuit introduced in Section VII-B (IF was widely varied through Monte Carlo simulations). The approximately linear dependence on IF in Fig. 22 and (24) is explained by the fact that a large imbalance tends to increase the delay associated with the weaker transistor (e.g., rising transition if the PMOS is weaker), compared to the stronger one [1]. At the same time, a larger imbalance also determines a larger leakage, as the stronger transistor has an increased leakage, compared to the weaker one.

Hence, IF is again a key design parameter in the design of ULP logic circuits, as it also impacts leakage energy, in addition to its effect on the dc characteristics degradation discussed in Section IV-D. To keep imbalance within bounds, the adaptive feedback schemes discussed in Section XIII must be employed to compensate such imbalance [19], [38].

[4] Instead, increasing transistor size of the weaker transistor (e.g., PMOS) to reduce IF in a given logic gate may be counterproductive in terms of $E_{\rm lkg}$. Indeed, this leads to an increase in the logic gate leakage, as well as the delay of the driving logic gate.


(obtained for input $I_{PUN,min}$)
PUN strength range
(obtained for input $I_{PUN,max}$)
$\beta_{PUN,min}$
$\beta_{PUN,max}$
case 2
case 1
$\beta_{PDN,min}$
$\beta_{PDN,max}$
(obtained for input $I_{PDN,min}$)
PDN strength range
(obtained for input $I_{PDN,max}$)
geom. average btwn $\beta_{PUN,min}$ and $\beta_{PUN,max}$
PUN strength range
$\beta_{PUN,min}$
$\beta_{PUN,max}$
$\beta_{PDN,min}$
$\beta_{PDN,max}$
PDN strength range
geom. average btwn $\beta_{PUN,min}$ and $\beta_{PUN,max}$

(a)

(b)

Fig. 23. PUN/PDN strength range and notation: (a) sizing for nonminimum imbalance; (b) sizing for minimum imbalance according to criterion in (26).

IX. Standard-Cell Design: From Transistor-Level Strategies to Design Flows

In the following, various aspects related to standard-cell VLSI design of ULP systems are discussed.

A. *Transistor Sizing Strategies in ULV Standard Cells*

In cell libraries for nearly-minimum energy, minimum-sized transistors should be used as much as possible. Larger than minimum transistors should be used to make critical path faster (to reduce T_{CK} and hence $E_{\rm lkg}$ from (21)) and ensure an adequate PUN/PDN balance. The former case simply requires the adoption of cells with larger strength, and is discussed in Section IX-C. The latter case is discussed in the following by targeting minimum-strength balanced cells.

In the simple case of an inverter, a good balance (i.e., $IF \approx 1$) is obtained by sizing the stronger transistor at minimum, and over-sizing the weaker transistor. In the adopted technology, the intrinsic imbalance (i.e., for minimum-sized nominal transistors) is $\beta_n/\beta_p = 7$. Hence, a balanced inverter gate is built with minimum-sized NMOS and increasing β_p by $7\times$. To avoid a large area penalty, such large imbalance should be first reduced by applying full-voltage FBB to PMOS (i.e., $V_{BB,p} = 0$), and then further reduced through sizing. In the adopted technology, FBB enables a β_p increase by about 2×, whereas the residual imbalance can be approximately compensated by increasing the PMOS strength by 2×–3×.

In more complex gates, the above criteria cannot be applied straightforwardly since actually the PUN/PDN strength varies according to the input value (see examples in Fig. 15). Hence, it is unclear how to match the PUN strength $\beta_{\rm PUN}$ and the PDN strength $\beta_{\rm PDN}$. Indeed, with referral to Fig. 23(a), the PUN strength ranges from its minimum and maximum values $\beta_{\rm PUN,min}$ and $\beta_{\rm PUN,max}$, which are respectively obtained for a certain input value $I_{\rm PUN,min}$ and $I_{\rm PUN,max}$ (similar notation is adopted for PDN). For example, the NAND2 gate in Fig. 15(a) has $\beta_{\rm PUN,max} = 2\beta_p$ (obtained for $I_{\rm PUN,max} = 00$), $\beta_{\rm PUN,min} = \beta_p$ (obtained for $I_{\rm PUN,min} = 01$ or 10), $\beta_{\rm PDN,max} = \beta_{\rm PDN,max} = \beta_n/X_{\rm stack}$ (obtained for $I_{\rm PDN,min} = I_{\rm PDN,max} = 11$).

Since $\beta_{\rm PUN}$ and $\beta_{\rm PDN}$ cannot be perfectly matched due to their dependence on the input, some residual imbalance must be accounted for anyway [1]. Accordingly, CMOS cells must be sized to minimize the worst-case IF, which from (13) may occur when $\beta_{\rm PUN}/\beta_{\rm PDN}$ is either maximum (case 1 in Fig. 23(a), with $\beta_{\rm PUN} = \beta_{\rm PUN,max}$, $\beta_{\rm PDN} = \beta_{\rm PDN,min}$) or minimum (case 2 in Fig. 23(a), with $\beta_{\rm PUN} = \beta_{\rm PUN,min}$ and $\beta_{\rm PDN} = \beta_{\rm PDN,max}$). Hence, the worst-case IF is [1]

$$IF = \max\left(\frac{\beta_{\rm PUN,max}}{\beta_{\rm PDN,min}}, \frac{\beta_{\rm PDN,max}}{\beta_{\rm PUN,min}}\right). \qquad (25)$$

From the circuit designer point of view, transistors must be sized to minimize (25), i.e., by strengthening the weakest between PUN and PDN. For example, if the designer strengthens the PUN transistors by a given factor, $\beta_{\rm PUN,max}$ and $\beta_{\rm PUN,min}$ will increase by the same factor, hence $\beta_{\rm PUN,max}/\beta_{\rm PDN,min}$ increases and $\beta_{\rm PDN,max}/\beta_{\rm PUN,min}$ decreases by the same factor. Clearly, the worst-case IF is minimum when these two terms are equal, i.e., when [1]

$$\beta_{\rm PUN,max}\beta_{\rm PUN,min} = \beta_{\rm PDN,max}\beta_{\rm PDN,min} \qquad (26)$$

which is the general sizing criterion for CMOS standard cells with minimum worst-case imbalance. As a simple interpretation of (26), transistors in CMOS standard cells must be sized so that the geometric average between the maximum and the minimum strength is the same for both PUN and PDN, as summarized in Fig. 23(b). Under the optimum sizing criterion in (26), from (25) the minimum worst-case IF results to $\beta_{\rm PUN,max}/\beta_{\rm PDN,min}$ (or equivalently $\beta_{\rm PDN,max}/\beta_{\rm PUN,min}$), where $\beta_{\rm PUN,max}$ and $\beta_{\rm PDN,min}$ are now evaluated with the optimum sizes.

B. *Design Examples*

From the previous subsection, the minimum-strength version of a CMOS standard cell with minimum imbalance is obtained by adopting minimum transistors in the strongest between PUN and PDN, and up-sizing the others to satisfy criterion in (26).

In the case of NAND2 gate in Fig. 15(a), from (26) and using the numbers indicated in Section IX-A, transistors have to be sized such that $\beta_n/\beta_p = \sqrt{2}X_{\rm stack} \approx 3.5$ ($X_{\rm stack} \approx 2.5$ for two stacked transistors from Fig. 16). The latter is lower than the intrinsic imbalance $\beta_n/\beta_p = 7$ (obtained with minimum-sized transistors) by a factor of 2. Hence, M1-M2 must be minimum sized and the strength of M3-M4 must be increased by a factor of about 2 compared to minimum-sized PMOS (by using the strategies discussed in Section III-B). The resulting worst-case IF in (25) after the above optimization is $\beta_{\rm PUN,max}/\beta_{\rm PDN,min} = 2\beta_p/(\beta_n/X_{\rm stack}) = \sqrt{2}$.

Similarly, the NOR2 gate in Fig. 15(b) has $\beta_{\rm PUN,max} = \beta_{\rm PUN,min} = \beta_p/X_{\rm stack}$, $\beta_{\rm PDN,max} = 2\beta_n$ and $\beta_{\rm PDN,min} = \beta_n$. From (26), transistors must sized such that $\beta_n/\beta_p =$

TABLE VII
PMOS Transistor Sizing for Various Logic Gates: Numerical Examples (NMOS Transistors are Always Minimum Sized)

	design for minimum worst-case *IF*		*IF* for all minimum transistors
	PMOS strength*	worst-case *IF*	
NAND2	2X	$\sqrt{2}$	2.8
NOR2	24X	$\sqrt{2}$	35
NAND3	~1X (min. sized)	$\sqrt{3}$	1.75
NOR3	50X	$\sqrt{3}$	84

* normalized to the minimum-sized PMOS strength

$1/\sqrt{2}X_{\text{stack}} \approx 0.29$, i.e., 24 times lower than the intrinsic ratio at minimum size. Hence NMOS transistors must be minimum sized and the PMOS strength should theoretically be increased by 24× (!) compared to its minimum-size value, and the resulting worst-case IF is again $\sqrt{2}$. All these data are summarized in Table VII, along with the results obtained for NAND3 and NOR3 gate (considering that $X_{\text{stack}} \approx 4$ for three stacked transistors from Fig. 16). The imbalance in the case with all minimum transistors is also included in the last column.

From Table VII, the sizing procedure in the previous subsection leads to a small worst-case IF (basically, it is the square root of the fan-in). This is achieved with a reasonable sizing in NAND gates, whereas it requires an unfeasibly large PMOS strength increase in NOR gates (24× for NOR2, 50× for NOR3). As discussed in detail in Section VI-C, this is because PMOS transistor is much weaker than NMOS in the considered technology, hence their series connection in NOR gates makes PUN/PDN balancing even harder. The only design option that might mitigate this problem is the appropriate mix of transistors with different transistor flavors (e.g., std-V$_{TH}$ PMOS and high-V$_{TH}$ NMOS), in order to widely adjust strength under reasonable size.

In general, the worst-case IF when all transistors are minimum sized represents a fair measure of the intrinsic imbalance of a logic gate. From the last column in Table VII, the resulting worst-case IF for minimum-sized NAND gates is of a few units, whereas it is in the order of many tens for minimum-sized NOR gates (it basically doubles each time the fan-in is increased by one). As a negative consequence of the worse NOR gate intrinsic imbalance, from (19) $V_{DD,\min}$ of NOR2 (NOR3) gate is higher than NAND2 by 106 mV (130 mV). It is interesting to observe that, compared to the minimum-size case, IF cannot be dramatically improved in general. Indeed, to avoid an excessive area penalty, the transistor strength can typically be increased at most by 2×–3×, which from (19) brings a limited reduction in $V_{DD,\min}$ by $n \cdot v_t \ln(3) \sim 30 - 40$ mV. In other words, minimum-sized designs give an idea of the intrinsic cell voltage limitations.

Finally, the impact of variations can be easily incorporated into the above transistor-level design strategy by observing that they simply tend to widen the strength range in Fig. 23, due to the additional uncertainty in the transistor strength. Clearly, the design methodology discussed above still applies, although the resulting imbalance is increased as discussed in Section V-B. Unfortunately, in ULP cells it is not possible to take advantage of the variability reduction offered by transistor stacking (i.e., averaging of intradie variations), since the allowed fan-in is typically limited to 2, as discussed in the next subsection.

C. ULP Standard Cell Libraries

The design of ULP standard cell libraries is very different from the traditional case of above-threshold libraries in terms of composition and physical design, other than the transistor-level strategies discussed in the previous subsections.

Regarding the composition, cells with fan-in larger than two or three must be certainly avoided, especially those with stacked transistors of the weak type, since they significantly increase the minimum supply voltage, thereby making ULV operation unfeasible [21]. At the same time, the cell strength required by the synthesis of ULP systems is usually small, hence very few versions of each cell with small or moderate strength are actually required (e.g., 1× and 4×). A larger strength might be needed only in critical paths, as they determine the clock cycle and hence the energy leakage of the entire system from (21). For those few exceptions, larger strengths can be achieved anyway by putting low-strength cells in parallel (the resulting overhead is small, since this case is infrequent).

From the above considerations, ULP libraries should consist of few cells that are implemented in very few different strengths, hence the cell count can be kept in the order of few tens. This is very different from above-threshold libraries (whose cell count can easily be in the order of a thousand), and makes the design of ULP libraries reasonable from a design cost point of view. Libraries specifically designed for ULP make even more sense when considering that they usually contain small transistors, hence the cell height can be reduced by 1.5×–2× compared to above-threshold cells. This clearly reduces the average length of interconnects, thereby reducing the wire parasitics and the driving transistor strength, thereby significantly reducing both E_{dyn} and E_{lkg}. These considerations have been usually ignored in previous work, and have lead to a diffused and simplistic conclusion that above-threshold libraries are adequate for ULP applications [53]. This common belief is also due to the lack of understanding of the important role played by $V_{DD,\min}$ (see Section VIII-B), whose reduction requires ULP-specific sizing strategies, as discussed previously. Further investigation is required to quantify the benefits of ULP-specific cell libraries.

D. Considerations on ULP Design Flows

Commercial tools and above-threshold design flows do not explicitly deal with many of the above discussed issues arising at ultra-low voltages, thereby requiring careful selection of synthesis directives [54]. On the other hand, from some point of view, ULP design flows are simpler than traditional ones. For example, supply network voltage drops are not an issue in ULP systems (I_{on} is orders of magnitude lower than above threshold), hence minimum-sized supply rails can be used and decoupling capacitors are not needed. As another example, wire parasitics like inductance (and often resistance as well) can be neglected because of the low performance, thereby simplifying automated placement and routing, as well timing closure. Thermal issues like hot spots do not arise either in ULP systems. Dealing with leakage is also relatively simple in ULP design flows, since the dominance of subthreshold current makes gate and junction leakage modeling superfluous. Automated signal integrity check is also relatively easy, as current spikes drawn from the supply are small, signals are much slower and minimum-width interconnects are used most of the time, thereby making vertical cross-talk and substrate coupling small. Interaction with the package is also very predictable, since bonding self and mutual inductances have negligible effect.

In regard to timing analysis, purely random variations due to RDF tend to dominate over systematic contributions. On one hand, this means that interconnect variability can be generally ignored, which simplifies timing analysis. On the other hand, the characterization of gate delay variability requires a higher computational effort compared to above threshold, due to the larger number of simulations required to approximate the gate delay log-normal distribution in timing look-up tables [50]. Also, the dominance of purely random variations makes corner-based timing analysis overly pessimistic when searching for setup and hold time violations, thereby leading to unnecessary margining and energy/performance penalty. This can be avoided by performing Monte Carlo circuit simulations to include averaging of delay variations among cascaded logic gates, at the expense of a significantly larger computational effort. This computational effort can be reduced through appropriate timing analysis design flows that resort to preliminary corner analysis and variability estimation, and then restricting Monte Carlo simulations only to paths that are found to be likely to violate hold time [50].

Currently available CAD tools do not deal well with functional/parametric yield issues at ULV. Hence, yield-aware design methodologies and flows would be highly desirable to systematically deal with variations. Also, voltage scaling and tuning is an inherent feature of ULV systems, which requires the characterization of cell libraries at multiple supply/body bias voltages. Hence, efficient strategies to characterize cells with a limited effort would be highly desirable.

Finally, clock tree synthesis is different from above threshold, as the wire delay plays a very limited role, and the skew is dominated by clock buffers. This clearly has a strong impact on the way the clock hierarchy should be designed. Further investigation is certainly required to enable energy/variation-aware clock tree synthesis for ULP applications. Also, for the reasons discussed in Section XII-A, alternative clocking styles (e.g., 2-phase latch clocking) are an interesting option in the ULP domain, although commercial tools are usually less effective in minimizing area and energy, compared to flip-flop clocking.

X. Ultra-Low Voltage Memory Design

SRAM memory is the only realistic option for the implementation of the memory subsystem in ULP applications. Indeed, dynamic storage of information as in DRAMs would be too sensitive to leakage and its variations, which are particularly critical in subthreshold as previously discussed.

A. *General Issues in ULP SRAM Arrays*

SRAM arrays are either employed for providing temporary information storage for processing purposes, or for retaining previous information (e.g., acquired or generated in previous wakeup cycles) for future use. In the first case, the memory array is duty cycled and it contributes to the system power budget in (2) through its energy per access. On the other hand, in (2) the consumption of the retaining memory is set by its leakage. Hence, the two subarrays are typically designed in a different way to minimize their impact on the overall consumption.

The energy per access of the duty-cycled SRAM subarray can be minimized by using the wide range of low-power techniques that are currently adopted in above-threshold low-power caches [55]. On the other hand, the design of the retaining subarray is very different from traditional caches, since the dominating leakage power must lowered by accepting some compromise in terms of dynamic energy and area. As an example, at the algorithmic level the leakage power can be lowered through data compression [12], [56]. This significant leakage reduction is obtained at the expense of a slightly increased dynamic energy to perform the compression. At the circuit level, leakage power is reduced by aggressively reducing the operating voltage, which in turn requires careful circuit design to enable reliable operation [57], as discussed in the following.

Designing robust SRAM arrays for ULV operation is challenging. On top of the voltage reduction, ULV SRAM cell margins are further degraded due to the large NMOS/PMOS imbalance within each cell (worsened by the intrinsically high sensitivity to variations of SRAM cells – see Sections IV-D and V-B). The $I_{\mathrm{on}}/I_{\mathrm{off}}$ degradation further reduces robustness in read mode due to the connection of a large number of cells to the same bitline (see Section IV-C), as shown in Fig. 24(a) (which is conceptually equivalent to Fig. 8). These effects tend to degrade the array density: cells must be larger to ensure robustness, and less dense architectures must be used to limit the number of cells per bitline.

A first attempt to mitigate the $I_{\mathrm{on}}/I_{\mathrm{off}}$ degradation and its impact on the bitline leakage was presented in [43], where MUX-tree based decoders were adopted to limit the number of cells connected to the same node to two, instead of connecting all the cells directly to the bitlines. This clearly comes at the expense of a significant area overhead. On the other hand, standard decoders can be used if the SRAM cell is designed to reduce I_{off} of unaccessed cells in Fig. 24(a). For example, this can be done by using more complex cells (typically 8T or 10T) that are connected to a dedicated read bitline through a gated read buffer as in Fig. 24(b), which enables full-swing single-ended read in 8T [58] and 10T cells [59]. Through this approach, the number of cells per bitline can be increased to many hundreds [59], and can be further extended to a few thousands by avoiding bitline leakage data dependency through appropriate read buffer topologies [44].

B. *ULP SRAM Cells*

In ULP applications, an adequate noise margin must be ensured during hold, read and write mode despite of the reduced supply voltage. In 6T cells, it is well known that read stability is always worse than hold stability, and also improvements in read stability deteriorate the write margin of SRAM cells. Hence, traditional 6T cells are designed by managing the tradeoff between read stability and write-ability in the presence of variations. Typically, their robustness is not adequate for voltages lower than 500–600 mV, since their is actually above threshold [58]. Hence, to reliably operate at ultra-low voltages, other SRAM cell topologies must be adopted that either intrinsically improve

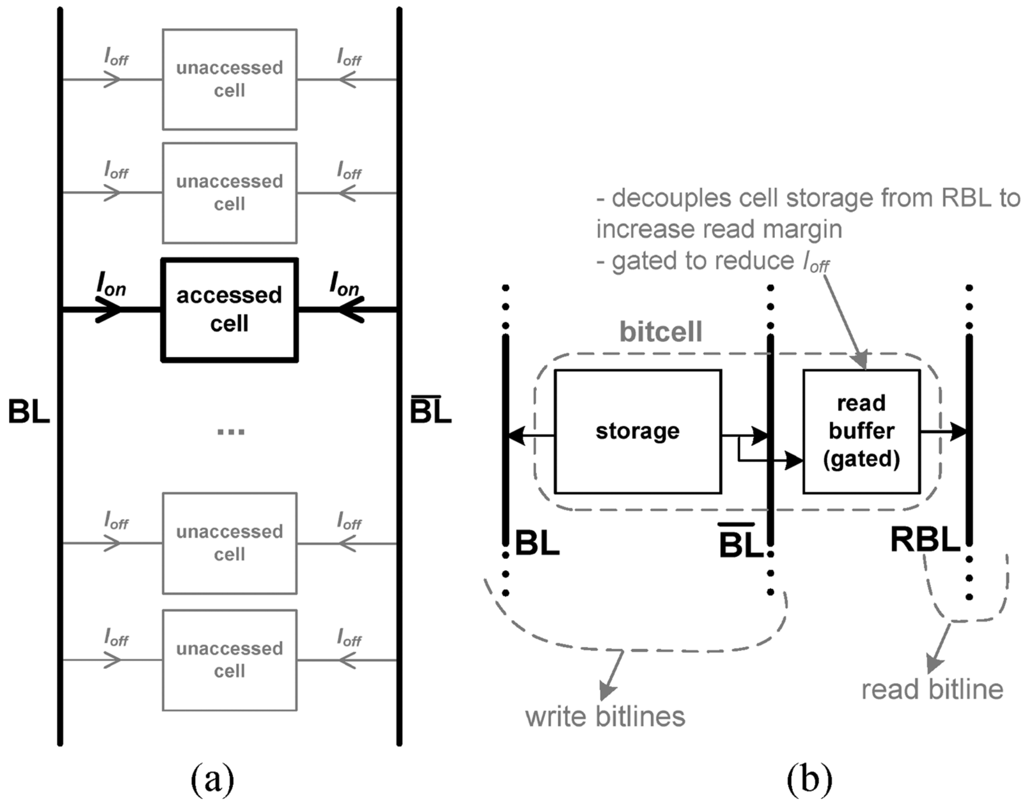


Fig. 24. SRAM memory subsystem: (a) general problem of $I_{\rm on}/I_{\rm off}$ degradation, (b) cell architecture with gated read buffer.

read stability or break this read/write margin tradeoff. As an example within the first category, [60] proposed the introduction of hysteresis within the cell feedback loop through a 10T topology.

On the other hand, most of the proposed solutions belong to the second category, in which the introduction of a separate read buffer to improve the read stability and/or some write assist mechanism to improve the write-ability are introduced. The above discussed read buffer permits to decouple the write and read margins, and the read margin becomes essentially equal to the hold margin [44], [58], [59]. On the other hand, write assist improves the write-ability by weakening the SRAM cell feedback path when the forward path coming from the bitline is forcing (writing) a new value. This limits current contention, which was shown to dramatically increase the sensitivity to variations in Section VI-D. As a practical example, the 10T cell in [43] has a gated feedback inverter, which is disabled to break the feedback path during the write phase. As another way to weaken the feedback path in write mode, the 10T cells in [59] are power gated, so that they are loosely connected to the supply in write phase. Hence, cell over-writing is easier and does not conflict with read stability requirements. To reduce the area overhead, the additional power gating transistor is shared among neighboring cells within the same row, hence all these cells must be written at once. This means that neighboring cells are associated with the same word, which makes the array more prone to multibit soft errors, as only one bit per word at a time can be corrected in typical SRAMs [61]. The 10T cell in [61] solves this problem by adding a write wordline distributed by columns, which adds to the row wordline, thereby over-writing only the cell at the intersection between the two wordlines (instead of all cells within a row at a time). This hence permits to intersperse the bits of different words, thereby making the array robust against multibit soft errors.

Another approach to alter the tradeoff between write and read margins is the selective adoption of voltage boosting/reduction to temporarily alter the transistor current to enhance the margin that currently matters [55], [58]. Indeed, thanks to the exponential dependence of the current on V_{GS} in (9), voltage boosting in subthreshold is very effective, as small voltage increase (many tens of mV) leads to significant current increase. Other bitcells with eight to ten transistors were reported in the literature. For example, in the 9T cell in [62], one transistor is taken off from the topology in [59] to reduce area.

Topologies with fewer transistors ranging from 5T to 7T were also proposed to reduce the area penalty imposed by noise margin constraints. A 7T cell was proposed in [63] that enables operation at 440 mV and yet high performance (20 ns access time at 0.5 V). A modified 6T cell was introduced in [45] that has power gating on the feedback inverter as write assist, as well as full-swing single-ended access through a transmission gate instead of a pass transistor. This relaxes the sense amplifier design and enables 200-mV operation. Cell designs with fewer transistors (5T) are not well suited for ULV operation (see, e.g., [64]).

Robustness issues also arise in the design of the sense amplifier (SA). Various techniques were proposed to maintain adequate immunity to noise and variations in the SA. For example, SA based on simple inverters with tunable threshold was adopted in [44], whereas SA redundancy was exploited in[58]. Offset voltage specification was also relaxed in [45] thanks to the adoption of full-swing bitline signaling.

TABLE VIII
SUMMARY OF BASIC FEATURES OF EXPERIMENTAL ULP SRAM PROTOTYPES

Design	cell	cells per BL	$V_{DD,min}$ (mV)	performance	consumption	technology (nm)
[59]	10T	256	380	400 kHz	3 μW	65
[44]	10T	1,024	200	100 kHz	2 μW	130
[60]	10T	256	160	620 kHz @400 mV	0.15 μW	130
[43]	10T	hierarchical (2:1)	180	164 Hz	tens of nW	180
[58]	8T	128	350	100 MHz	2.6 mW	65
[63]	7T	8	440	20 ns /access @ 0.5 V	N/A	90
[45]	6T	16	210	21.5 kHz	0.78 pJ /access @300 mV	130

The main features of the above cited ULP SRAM prototypes are summarized in Table VIII. In typical ULP systems, one should also consider that a limited amount of memory is actually needed (down to a few kb, in very simple systems). In turn, this relaxed requirement on the memory capacity can be leveraged to relax the array density constraint, i.e., trading area for higher robustness to a certain extent. The typically low memory capacity also permits to reduce the statistical margin that must be added to ensure a given cell yield, thereby relaxing the read/write/hold margin constraint.

XI. EFFECTIVENESS OF STANDARD TECHNIQUES TO REDUCE LEAKAGE

As was discussed in Section II-B, the power P_{sleep} consumed by duty cycled blocks in sleep mode in (2) can be eliminated altogether by shutting down the on-chip regulator powering the processing units. This might not be an option in systems with low power delivery flexibility, in which the supply voltage V_{DD} of the duty cycled blocks is the same in both active and sleep mode. In these cases, P_{sleep} must be reduced through techniques such as power gating, forced stacking and body biasing.

Power gating is extensively used in above-threshold designs to keep leakage under control. Unfortunately, in subthreshold power gating is less effective than above threshold for various reasons, as discussed in the following assuming that a header transistor is used [14]. First, the reduction in virtual supply VV_{DD} due to the voltage drop across the sleep transistor leads to an exponential delay increase from (10), which turns into an exponential increase in E_{lkg} during active mode (see Section VII-A). This increase in E_{lkg} is not appreciable in above threshold circuits due to the lower delay sensitivity on V_{DD}. In subthreshold, the increase in E_{lkg} can be limited by avoiding an excessive reduction in VV_{DD}, i.e., by increasing V_{DD} and increasing the sleep transistor size (to reduce its voltage drop). In turn, the increase in V_{DD} and sleep transistor size leads to an increase in P_{sleep}, due to the higher voltage and the higher leakage of the sleep transistor. Hence, V_{DD} and the sleep transistor size have to be optimized jointly to minimize the overall consumption in (2) [65], i.e., to optimally balance active leakage, sleep leakage and dynamic energy. In this optimization, robustness considerations should also be taken into account [66]. Typically, the leakage reduction in sleep mode offered by power gating is in the order of many various tens, and be as high as two orders of magnitude for systems with long wakeup period.

Another popular technique to reduce leakage in both active and sleep region is the "stack forcing", where a single transistor is replaced by two equal transistors in series with halved size [14]. From Fig. 16, the stacking factor in subthreshold is the same for both I_{on} and I_{off}, as opposed to transistors operating above threshold. Hence, in subthreshold the ratio $I_{\text{on}}/I_{\text{off}}$ does not improve with stacking, as opposed to above-threshold designs where transistor stacking reduces I_{off} much more than I_{on}. Thus, from (21) the leakage energy E_{lkg} does not benefit from stacking, since E_{lkg} is inversely proportional to $I_{\text{on}}/I_{\text{off}}$ from (10) and (21) (since $\tau_D \propto 1/I_{\text{on}}$). At the same time, the reduction in I_{on} due to stacking degrades performance, which can be compensated through V_{DD} increase, which in turn translates into an increase in E_{dyn}. In other words, forced stacking has to be avoided in subthreshold, since it does not bring significant advantage in terms of leakage energy, and can degrade the dynamic energy.

Body biasing is another option to keep P_{sleep} low, especially when the body voltage V_{BB} is adjusted dynamically to have higher threshold in sleep mode to reduce leakage. As an example, from (5b) and (11), the leakage current (and hence P_{sleep}) can be reduced by a factor $e^{\lambda_{BS} V_{BB}/nv_t}$, which for $V_{BB} = -0.5$ V is about 5× from data in Table I for the 65-nm NMOS transistor (more negative voltages are typically unfeasible, since other leakage mechanisms become dominant).

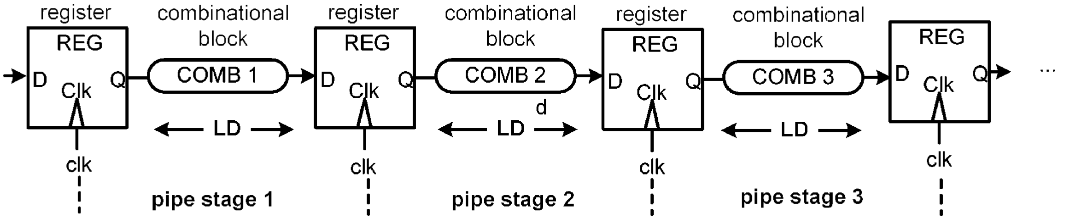


Fig. 25. General scheme of pipelined systems.

Technologies with more pronounced short channel effects (i.e., with larger λ_{BS}/n) can benefit more from body biasing, typically up to 10× [66].

From the above numbers, leakage reduction offered by existing techniques is not large enough to make P_{sleep} negligible in (2). Hence, dynamic voltage scaling including the extreme case where the dc-dc converter is shut down (i.e., $V_{DD} = 0$ V in sleep mode) remains the only realistic option to make P_{sleep} negligible.

XII. Microarchitectures for ULP Systems

The architecture and the micro-architecture have a dramatic impact on the energy per operation. According to the considerations in Section II-B, and considering that the bit width can widely vary among different architectures, a fair figure of merit for energy efficiency is the average energy per instruction divided by the bit width. The resulting energy per instruction per bit typically ranges from a few pJ to many tens of pJ for ULP-specific architectures [47]–[51], [67]–[71]. In general, architectures with larger instruction set and greater bit widths tend to have higher energy per instruction per bit, due to the more-than-linear complexity increase and routing distances. Architectures for ULP are still a very open research field, as it is not trivial to incorporate energy optimality and heavy nonidealities arising in subthreshold (e.g., leakage, variations) into high-level design explorations.

In regard to micro-architectures, in the following pipelining and hardware replication are reviewed from the perspective of energy efficiency in ULP applications [1].

A. Impact of Pipelining on Energy/Power

With reference to the general pipelining scheme in Fig. 25, shallow pipelining has been the preferred choice in most of the work on ULP circuits reported in the literature. This is because the adoption of a large logic depth per stage LD (i.e., amount of combinational logic per pipeline stage) greatly simplifies the design, since the impact of random process variations is averaged out across a large number of logic gates. In previous works, logic depth ranges from $50FO4$ to $200FO4$ [68].

Actually, the logic depth LD is an effective knob to minimize the energy per operation in pipelined VLSI circuits. Intuitively, a lower LD reduces the time in which each logic gate is kept idle, thereby reducing the leakage energy per operation of the combinational logic. At the same time, lower LD increases the number of registers and the operating frequency, thereby increasing the dynamic energy due to clocking. Hence, the logic depth controls the balance between leakage and dynamic energy, whose tradeoff ultimately determines the MEP (see Section VII-B).

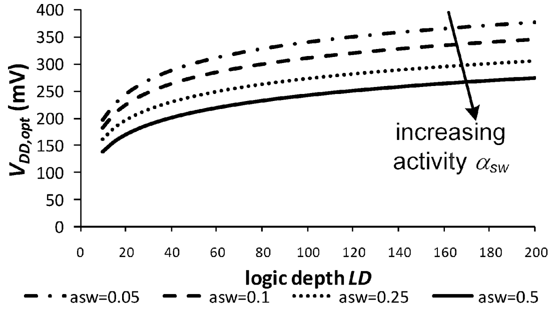


Fig. 26. Optimum supply voltage vs. logic depth LD in reference combinational block (activity factor ranging from 0.05 to 0.5).

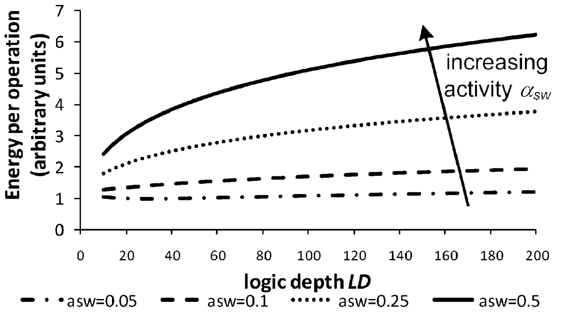


Fig. 27. Energy per operation vs. logic depth LD in reference combinational block (activity factor ranging from 0.05 to 0.5).

From the above considerations, adopting a low logic depth permits to reduce leakage at the expense of a larger dynamic energy. In turn, the latter can be reduced by further reducing V_{DD}, hence the joint optimization of logic depth and V_{DD} is very effective in reducing the energy per operation in ULP micro-architectures. More specifically, it makes sense to adopt very low logic depth to make the circuit faster and suppress leakage, and then trade off the excess of speed for lower energy [1]. From this perspective, high-performance microarchitectures with deep pipelining (i.e., very low V_{DD}) are well suited for ULP systems, if supply is also jointly optimized. Deep pipelining makes even more sense than in above threshold since the clocking overhead is a smaller fraction of the clock cycle[5].

As an example, Figs. 26–27 show the optimum supply and the energy per operation (in arbitrary units) of the reference circuit introduced in Section VII-B with LD varying from $20FO4$ to $200FO4$. As expected, optimization of LD can provide significant benefits compared to the typically large adopted values. In particular, the benefit is more pronounced at higher activity factor, since dynamic energy dominates and hence the overall energy greatly benefits from voltage

[5] Indeed, the RC wire delay and its variations within the clock network are small, as the wire delay is a much smaller fraction of the gate delay, compared to above-threshold designs.

scaling. For example, α_{sw} equal to 0.5 (0.05), the energy is reduced by a factor of 2 (10–20%), compared to the case with $LD \sim 100 - 200FO4$. The optimum logic depth for α_{sw} equal to 0.5 (0.05) turns out to be below $20FO4$ (around $30FO4$). Clearly, logic depths lower than $20FO4$ are generally difficult to implement due to the impact of variations and clocking overhead. Anyway, adoption of logic depths in the order of $20 - 30FO4$ generally provide significant energy savings, especially at high α_{sw}. This is in line with [33], which showed that ULP microarchitectures should have less than $20FO4$ per cycle, as well as the experimental results in [68], where a logic depth of $17FO4$ and $V_{DD,opt} = 270$ mV is adopted.

The adoption of the above mentioned deeply pipelined microarchitectures with reduced V_{DD} clearly emphasizes the importance of keeping $V_{DD,\min}$ as low as possible, as was anticipated in Section VII-A. Once again, the common belief that $V_{DD,\min}$ reduction is of secondary importance proves to be incorrect, since $V_{DD,opt}$ is rather low (easily in the order of 200 mV) when micro-architecture is optimized for minimum energy [1], as clearly shown in Fig. 26.

As a major downside of deeply pipelined circuits operating at ultra-low voltages, process variations have a strong impact on T_{CK}, and hence on E_{lkg}, as discussed above. In other words, the joint adoption of deep pipelining and low supply can significantly degrade robustness and adds significant leakage energy overhead due to variations. Hence, deep pipelining is actually applicable only if variations are kept within acceptable limits through appropriate techniques (discussed in Section XIII), as well as clocking schemes allowing time borrowing to average out delay variations among adjacent stages (e.g., two-phase latch clocking). Observe that hold time violations are intrinsically difficult to fix in such deeply pipelined systems, as they usually involve a limited number of gates (hence little averaging takes places) and there is not much room for buffer insertion because of the small allowed logic depth. To deal with hold time violations, flip-flop topologies that are intrinsically robust against hold variations should be used. Also, appropriate strategies for statistical timing analysis should be adopted [50].

The above considerations were derived for the energy, which is relevant to duty-cycled blocks. On the other hand, the power consumption of always-on circuits does not benefit from pipelining because power actually increases when reducing LD.

B. Impact of Hardware Replication on Energy/Power

Another popular micro-architectural technique is hardware replication [72], in which a given block is replicated N_{replica} times. Inputs are sequentially applied to different replicas, so that N_{replica} input data are concurrently processed as in Fig. 28. Thanks to the increased level of parallelism, performance (throughput) is increased by a factor N_{replica}, then this excess of performance is given back by reducing V_{DD} and consumption. In above-threshold designs, the resulting energy saving is enabled by the fact that E_{dyn} has a stronger dependence on V_{DD} compared to performance (i.e., a reduction in the performance in excess translates into a greater power reduction). Timing, energy and area overhead of the additional control circuits to manage the replicas is typically negligible, as assumed in the following.

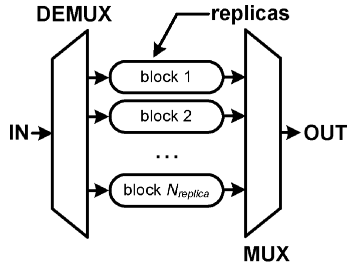


Fig. 28. General scheme for hardware replications.

The same concept can be applied in subthreshold. Let $\text{throughput}_{\min} \propto 1/\tau_D$ be the throughput that is achieved by the considered block with $N_{\text{replica}} = 1$ and $V_{DD} = V_{DD,opt}$. In other words, $\text{throughput}_{\min}$ is the throughput delivered by the single block (without replicas) at the MEP under the given technology. Clearly, all other practical designs operating at the right of the MEP have higher throughput. To express the targeted throughput in a technology-independent way, let us normalize it to , as in (27) [1]

$$\text{throughput}_{\text{norm}} = \frac{\text{throughput}}{\text{throughput}_{\min}} \geq 1. \tag{27}$$

When a single block is used (i.e., $N_{\text{replica}} = 1$), the only knob to obtain the targeted $\text{throughput}_{\text{norm}}$ is the supply voltage $V_{DD}|_{N_{\text{replica}}=1}$ (which is easily found by considering that $\text{throughput}_{\text{norm}} \propto 1/\tau_D$ depends on V_{DD} according to (10)). When hardware replication is applied (i.e., $N_{\text{replica}} > 1$), the targeted $\text{throughput}_{\text{norm}}$ is obtained by jointly optimizing N_{replica} and V_{DD} (which are clearly interdependent due to the throughput constraint). The resulting optimum number of replicas is easily found to be[6]

$$N_{\text{replica,opt}} = \text{throughput}_{\text{norm}}. \tag{28}$$

From (27)–(28), the theoretical optimum number of replicas for minimum energy is equal to the ratio $\text{throughput}/\text{throughput}_{\min}$ [1]. Actually, the number of replicas is limited by practical considerations on area, and reasonable values are in the order of a few units. Hence the design criterion in (28) can be rigorously applied only if the required throughput is relatively close to its minimum $\text{throughput}_{\min}$. Instead, when $\text{throughput}_{\text{norm}}$ is large, N_{replica} is set to the maximum reasonable value (e.g., 2–4) and the resulting energy saving is smaller.

As an example, let us consider the reference circuit introduced in Section XII-A and apply hardware replication, assuming $\alpha_{sw} = 0.1$, $LD = 120FO4$ and single pipeline stage. The resulting energy normalized to the case of single replica is plotted versus N_{replica} in Fig. 29 for three different throughput targets ($\text{throughput}_{\text{norm}} = 5$, 10 and 20). As

[6] Energy per operation under hardware replication is exactly the same as the energy of a single block at a given V_{DD} (since power and number of concurrent operations are both increased by a factor N_{replica}). Hence, under hardware replication, minimum energy is obtained when each replica operates at its optimum voltage $V_{DD,opt}$, exactly like the single stage. From (27), throughput of each replica at $V_{DD,opt}$ is $\text{throughput}_{\min}$, hence the overall throughput in (27) is $N_{\text{replica}} \cdot \text{throughput}_{\min}$, which justifies (28).

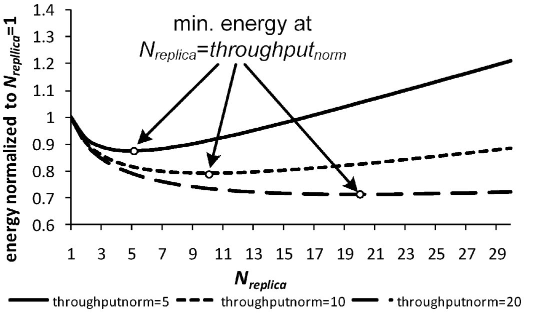


Fig. 29. Energy under hardware replication normalized to the case of single stage vs. N_{replica} (same circuit as in Section XII-A with one pipeline stage, $\alpha_{sw} = 0,1, LD = 120FO4$).

expected from (28), the energy per operation is minimized when the number of replicas is equal to 5, 10 and 20 respectively. From Fig. 29, for the above throughput targets, energy is respectively reduced by 13%, 20% and 30%. An appreciable energy reduction is still obtained when lower (non-optimum) values of N_{replica} are used. For example, for $N_{\mathrm{replica}} = 4$ the energy saving compared to the single stage is 13%, 17% and 19%.

In general, hardware replication offers larger energy savings for higher values of α_{sw} and lower logic depth LD, since in these cases the dynamic energy dominates over leakage, hence the energy benefits brought by the voltage reduction are more pronounced. For the same reason, larger energy savings are obtained when the targeted throughput is higher. In extreme cases (e.g., $\alpha_{sw} = 0.5$, $LD = 20FO4$ and $\mathrm{throughput}_{\mathrm{norm}}$ in the order of 100), the energy saving offered by hardware replication under single stage can be up to 50–60% under optimum N_{replica}, and up to 30% for the realistic choice $N_{\mathrm{replica}} = 4$. The resulting energy reduction is lower than that allowed by pipelining by one order of magnitude (see previous subsection). This is because E_{lkg} per operation is not reduced by increasing N_{replica}, hence voltage and energy cannot be scaled down as much as in pipelining. In addition, performance in subthreshold is actually much more sensitive to V_{DD} (exponentially) than energy (square law), thereby making hardware replication less energy efficient than above threshold [1]. The same considerations apply also to memory circuits, where a higher degree of parallelism can be achieved through bank interleaving [55].

Finally, regarding the power reduction of always-on circuits, hardware replication is ineffective for the same reasons discussed for pipelining.

XIII. Techniques to Mitigate Variations and Errors at Run Time

The large variations experienced by ULP systems make it very hard to constrain their performance and energy in a narrow region of the energy-performance space, i.e., to obtain an adequate parametric yield. For this reason, ULP systems typically require postsilicon feedback techniques, as discussed in the following.

Regarding energy, in Section VII the MEP (or any other operating point) was shown to strongly depend on the input activity and environmental conditions, which change over time. Hence, operation at a desired energy point actually requires V_{DD} tuning at run time through a feedback loop that tracks the desired energy consumption (e.g., the MEP). Clearly, open-loop control of the voltage/frequency pair according to the instantaneous performance requirement is far from energy optimal, as changes in environmental variations and activity cannot be tracked. Instead, a closed-loop control based on energy measurements and subsequent V_{DD} optimization can track very well the desired operating point [42]. Such feedback schemes dramatically reduce the deviation of the actual energy point from the target, thereby mitigating energy variations and improving parametric yield. Compared to open-loop control, 50–100% energy reduction over typical range of conditions was found in [42].

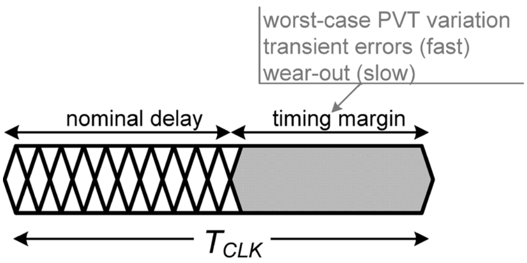


Fig. 30. Clock cycle increase due to timing margining.

Imbalance variations can be kept within bounds through closed-loop feedback control of threshold voltage, in order to ensure adequate functional yield (see Sections VI-D and VIII-B), as well as low E_{lkg} variations and hence adequate parametric yield (see Section VIII-D). Adaptive body biasing has been used for example to separately tune the threshold voltage in NMOS and PMOS [46], as well as to compensate their imbalance [30], [67]. Since these schemes close the loop across a reference circuit but apply the same voltages to the whole system, intradie variations cannot be compensated.

In ULP systems, delay variations also have to be kept within bounds to ensure adequate parametric yield in the presence of large variations. The latter can be mitigated by keeping the logic depth large enough to exploit delay averaging between cascaded gates (typically, $15FO4$ and above should be feasible with appropriate design techniques), adopting latch clocking to enable time borrowing between adjacent stages and sizing transistors carefully [33]. Some adaptive body biasing technique has also been proposed to statically compensate variations and operate at a desired performance target [19], [38].

In regard to the mitigation of timing errors, design margining is the most common approach to ensure always-correct operation under the worst case, as shown in Fig. 30. The timing margin must accommodate for PVT variations and fast transient errors (e.g., soft errors, occasional delay increase due to transient V_{DD} drops, capacitive coupling), and leads to a considerable increase in T_{CK} and E_{lkg} from (21). To avoid such energy/performance penalty, many postsilicon techniques have been proposed to correct occasional errors based on the scheme in Fig. 31. As summarized in this figure, error detection is typically performed through *in-situ* monitoring through appropriate latch/flip-flop topologies or transition detectors (e.g., Razor I [75], Razor II [76], TBTD and DSTB [77]). On the other hand, error prediction is typically performed by identifying potential timing violations through replica circuits mimicking the timing

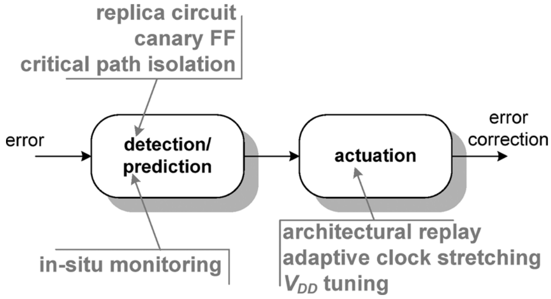


Fig. 31. General scheme for run-time error correction.

of the critical path [19], using "canary flip-flops" that detect signals that switch too close to the next clock edge [78], or isolating the critical path according to the current input [79]. Compared to error detection, the prediction generally involves less overhead but requires some design margin, although significantly smaller than always-correct designs.

Once an error is detected or predicted, its correction is typically performed through architectural replay (i.e., the pipeline is flushed [76] or brought back to the previous state [75]), by stretching operation by one cycle [79], slowly tuning V_{DD} or frequency to control the average BER [75], [80], or performing dynamic retiming [81]. Depending on the application, correction is performed to either recover from all detected errors (if error-free operation is required) or keep the error rate within bounds (if the application has some intrinsic resiliency at the algorithm level [73], [74]). In practical designs, the selection among the above techniques is made depending on the desired tradeoff between energy, robustness and recovery time.

XIV. Remarks on Other Logic Styles and Technology

In the above analysis, CMOS logic has shown its own limits for ULV operation. For this reason, various alternative logic styles have been proposed in the literature. Among them, MOS Current-Mode Logic (MCML) was recently demonstrated to exhibit significant advantages over CMOS logic for ULP applications. Indeed, MCML circuits can remarkably reduce power consumption by two orders of magnitude compared to CMOS logic, as was shown in [82], [83]. This is because the minimum power in CMOS logic is set by the supply voltage lower bound $V_{DD,\min}$ (i.e., by robustness issues), whereas the power of MOS Current-Mode Logic can be further scaled down by reducing its tail current instead of reducing V_{DD}. In other words, MCML circuits break the tradeoff between power and robustness, thereby enabling pW/gate operation [82], [83], which makes it well suited for ULP circuits (e.g., always-on block). As a further benefit, MCML has a significantly lower sensitivity to variations and hence higher yield, compared to CMOS logic [84]. This is a significant advantage, considering the yield issues discussed in the previous sections.

MCML logic style has the other advantage that the power-performance tradeoff can be widely adjusted through the tail current, whereas V_{DD} has a negligible effect on this tradeoff as well as robustness (as long as it is above pAs). This means that large variations in V_{DD} do not have any impact on the operation of MCML ULP systems, hence no or extremely simple voltage regulation is needed (even in purely energy-scavenged systems). As a drawback of MCML logic, $V_{DD,\min}$ is typically in the order of 500 mV or more [85], [86]. However, simple circuit techniques like FBB are available to reduce $V_{DD,\min}$ to about 200 mV [87], which basically fill the voltage gap between MCML and CMOS logic. In other words, MCML systems can work at very low voltages as CMOS logic, which makes system integration of digital blocks easier.

In regard to the process, the most important issues in CMOS logic (e.g., leakage, variability, imbalance) are expected to be even more critical in future technology generations [17], [88]. This is also because the process development is driven by mainstream technologies and does not account for the specific issues arising at ULV, and the assumption that the process can be tailored to ULV requirements seems rather unrealistic (see, e.g., [89], [90]).

As another important aspect related to the technology, the choice of the appropriate generation for a given design is an important and nontrivial decision in ULP applications. Indeed, the minimum energy is not achieved necessarily at the technology generation with smallest minimum feature size, due to its higher leakage. Typically, the optimum technology ranges from 90 nm to 180 nm, depending on the circuit activity [91].

XV. Conclusion

In this paper, an overview of the state of the art in ULP VLSI design has been presented in a unitary framework for the first time. Other than exploring design tradeoffs at various levels of abstractions, it has been shown that ULP design is very different from traditional low-power design: indeed, many paradigms and assumptions for above-threshold systems do not apply to subthreshold at all. A number of common misconceptions have been clearly identified and debunked.

Summarizing, from this tutorial, the reader should retain at least the following fundamental concepts and design principles:

— energy/power are equally important in ULP applications: power (energy) should minimized in always-on (duty-cycled) blocks

— ULP circuit design requires a good knowledge of the specific device used (no "thinking inside the box")

— transistor sizing strategies are completely different from above threshold (L is more effective than W)

— body biasing and threshold selection are more powerful circuit knobs than transistor sizing

— MOS transistor at ULV can be modeled as current source or resistance as above threshold, but the behavior changes much more rapidly

— PUN/PDN imbalance is a key design parameter that strongly influences the robustness, the leakage energy and minimum supply voltage; imbalance can be very large (hundreds or a few thousands) under variations

— avoid circuit topologies where a large number of transistors are connected to the same node (due to $I_{\rm on}/I_{\rm off}$ degradation)

— at ULV, CMOS logic acts like a ratioed logic style: all dc characteristics are severely degraded

— minimum energy is set by an optimal balance between dynamic and leakage energy

— energy reduction by up to 2 orders of magnitude at ULV is obtained with a speed penalty by 2–5 orders of magnitude

— trading off energy and performance at ULV must be done by using the right knob (supply or body biasing, depending on the specific case), otherwise energy efficiency is severely degraded

— the dominating variations at ULV are very difficult to compensate: they are purely random (RDF) and their impact is 10× larger than above threshold

— nominal $V_{DD,\min}$ at nominal conditions has a key role in the design of ULP systems to ensure adequate functional yield, and can be much higher (300–350 mV) than commonly believed due to variations

— operation at $V_{DD,\min}$ can actually be required in many practical cases (always-on block, optimized micro-architectures)

— $V_{DD,\min}$ and $V_{DD,opt}$ are actually strongly related: their distance determines the tolerance to variations and functional yield

— aggressive pipelining is a powerful approach to further reduce the energy per operation ("$\text{low energy} = \text{high performance}$")

— parallelism is rather ineffective in reducing energy in subthreshold

— leakage limits voltage and energy scaling, hence it must be aggressively reduced (traditional low-leakage techniques are rather ineffective in ULP systems)

— ULP standard cell libraries count few cells (tens) in few different versions with low strength, and

— must avoid logic gates with fan-in greater than 2-3 as well as topologies based on current contention

— ULV memory arrays have poor density due to robustness issues, and their read/write margins must be carefully traded off

— most critical consequence of timing uncertainty is the increase in leakage energy, rather than performance

— techniques to mitigate variations and errors are needed to improve functional and parametric yield, as well as energy efficiency.

REFERENCES

[1] M. Alioto, *Ultra Low-Power/Low-Voltage VLSI Design: From bASICs to Design*. Hoboken, NJ: Wiley, to be published.

[2] B. Warneke, M. Last, B. Liebowitz, and K. Pister, "Smart dust: Communicating with a cubic-millimeter computer," *Computer*, vol. 34, no. 1, pp. 44–51, Jan. 2001.

[3] R. Sarpeshkar, *Ultra Low Power Bioelectronics: Fundamentals, Biomedical Applications, and Bio-Inspired Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2010.

[4] *Ambient Intelligence*, W. Weber, J. Rabaey, and E. Aarts, Eds. New York: Springer, 2005.

[5] S. Massoud Amin and B. F. Wollenberg, "Toward a smart grid: Power delivery for the 21st century," *IEEE Power Energy Mag.*, vol. 3, no. 5, pp. 34–41, Sep.-Oct. 2005.

[6] G. Bell, "Bell's law for the birth and death of computer classes," *IEEE Solid-State Circuits Soc. News*, vol. 13, no. 4, 2008.

[7] Y. Lee, G. Chen, S. Hanson, D. Sylvester, and D. Blaauw, "Ultra-low power circuit techniques for a new class of sub-mm3 sensor nodes," in *Proc. CICC*, San Josè, CA, Sep. 2010, pp. 1–8.

[8] D. Blaauw, "Keynote speech," in *Proc. Int. Conf. Microelectron. (ICM)*, Cairo, Egypt, Dec. 2010.

[9] G. Chen, H. Ghaed, R.-U. Haque, M. Wieckowski, Y. Kim, G. Kim, D. Fick, D. Kim, M. Seok, K. Wise, D. Blaauw, and D. Sylvester, "A 1 cubic millimeter energy-autonomous wireless intraocular pressure monitor," in *Proc. ISSCC*, Feb. 2011.

[10] B. Murmann and B. Boser, "Digitally assisted analog integrated circuits," *ACM Queue*, vol. 2, no. 1, pp. 64–71, Mar. 2004.

[11] M. Ismail and D. Rodríguez de Llera González, *Radio Design in Nanometer Technologies*. New York: Springer, 2006.

[12] M. Seok, S. Hanson, Y.-S. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw, "The phoenix processor: A 30 pW platform for sensor applications," in *Proc. IEEE Symp. VLSI Circuits*, Jun. 2008.

[13] Y. Tsividis, *Operation and Modeling of the MOS Transistor*, 2nd ed. New York: McGraw-Hill, 1999.

[14] S. G. Narendra and A. Chandrakasan, *Leakage in Nanometer CMOS Technologies*. New York: Springer, 2006.

[15] Y. Cheng and C. Hu, *MOSFET Modeling & BSIM3 User's Guide*. Boston, MA: Kluwer Academic, 1999.

[16] S. Narendra, V. De, S. Borkar, D. Antoniadis, and A. Chandrakasan, "Full-chip subthreshold leakage power prediction and reduction techniques for sub-0.18-μm CMOS," *IEEE J. Solid-State Circuits*, vol. 39, no. 3, pp. 501–510, Mar. 2004.

[17] M. Alioto, "Understanding DC behavior of subthreshold CMOS logic through closed-form analysis," *IEEE Trans. Circuits Syst. I, Reg. Papers*, vol. 57, no. 7, pp. 1597–1607, Jul. 2010.

[18] T. Kim, H. Eom, J. Keane, and C. Kim, "Utilizing reverse short channel effect for optimal subthreshold circuit design," in *Proc. ISLPED 2006*, pp. 127–130.

[19] J. Kao, J. M. Miyazaki, and A. P. Chandrakasan, "A 175 mV multiply-accumulate DSP core using an adaptive supply voltage and body bias (ASB) architecture," *IEEE J. Solid-State Circuits*, vol. 37, no. 11, pp. 1545–1554, Nov. 2002.

[20] E. Vittoz, "Weak inversion for ultimate low-power logic," in *Low-Power Electronics Design*, C. Piguet, Ed. Boca Raton, FL: CRC, 2005.

[21] A. Wang, B. Calhoun, and A. Chandrakasan, *Sub-Threshold Design for Ultra Low-Power Systems*. New York: Springer, 2006.

[22] D. Harris, R. Ho, G.-Y. Wei, and M. Horowitz, "The fanout-of-4 inverter delay metric" [Online]. Available: http://www3.hmc.edu/~harris/research/FO4.pdf, unpublished

[23] M. Alioto, E. Consoli, and G. Palumbo, "From energy-delay metrics to constraints on the design of digital circuits," *Int. J. Circuit Theory Appl.*, 2011, to be published.

[24] A. Chandrakasan, D. C. Daly, D. F. Finchelstein, J. Kwong, Y. K. Ramadass, M. E. Sinangil, V. Sze, and N. Verma, "Technologies for ultradynamic voltage scaling," *Proc. IEEE*, vol. 98, no. 2, pp. 191–214, Feb. 2010.

[25] M. Alioto, "A simple and accurate model of input capacitance for power estimation in CMOS logic," in *Proc. ICECS*, Marrakech, Morocco, Dec. 2007, pp. 431–434.

[26] R. M. Swanson and J. D. Meindl, "Lon-implanted complementary MOS transistors in low-voltage circuits," *IEEE J. Solid-State Circuits*, vol. 7, no. SSC-2, pp. 146–153, Apr. 1972.

[27] N. Weste and D. Harris, *CMOS VLSI Design*. Reading, MA: Addison-Wesley, 2004.

[28] Y. Pu, J. Pineda-De Gyvez, H. Corporaal, and Y. Ha, "Vt balancing and device sizing towards high yield of sub-threshold static logic gates," in *Proc. ISLPED'07*, pp. 355–358.

[29] A. Bryant, J. Brown, P. Cottrell, M. Ketchen, J. Ellis-Monaghan, and J. Nowak, "Low-power CMOS at $\mathrm{Vdd} = 4\mathrm{kT}/\mathrm{q}$," in *Proc. Device Res. Conf.*, Jun. 2001, pp. 22–23.

[30] G. Ono and M. Miyazaki, "Threshold-voltage balance for minimum supply operation," *IEEE J. Solid-State Circuits*, vol. 38, no. 5, pp. 830–833, May 2003.

[31] M. Orshansky, S. Nassif, and D. Boning, *Design for Manufacturability and Statistical Design*. New York: Springer, 2008.

[32] M. Alioto, G. Palumbo, and M. Pennisi, "Understanding the effect of process variations on the delay of static and domino logic," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 18, no. 5, pp. 697–710, May 2010.

[33] B. Zhai *et al.*, "Analysis and mitigation of variability in subthreshold design," in *Proc. ISLPED*, Aug. 2005, pp. 20–25.

[34] K. Bernstein *et al., High-Speed Logic Styles*. Boston, MA: Kluwer Academic, 1999.

[35] H. Soeleman, K. Roy, and B. C. Paul, "Robust subthreshold logic for ultra-low power operation," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 9, no. 1, pp. 90–99, Jan. 2001.

[36] T. Niiyama, Z. Piao, K. Ishida, M. Murakata, M. Takamiya, and T. Sakurai, "Increasing minimum operating voltage (VDDmin) with number of CMOS logic gates and experimental verification with up to 1mega-stage ring oscillators," in *Proc. ISLPED 2008*.

[37] T. Niiyama, P. Zhe, K. Ishida, M. Murakata, M. Takamiya, and T. Sakurai, “Dependence of minimum operating voltage (VDDmin) on block size of 90-nm CMOS ring oscillators and its implications in low power DFM,” in *Proc. ISQED 2008*.

[38] S. Hanson, B. Zhai, M. Seok, B. Cline, K. Zhou, M. Singhal, M. Minuth, J. Olson, L. Nazhandali, T. Austin, D. Sylvester, and D. Blaauw, “Exploring variability and performance in a sub-200-mV processor,” *IEEE J. Solid-State Circuits*, vol. 43, no. 4, Apr. 2008.

[39] S. Gupta, A. Raychowdhury, and K. Roy, “Compact models considering incomplete voltage swing in complementary metal oxide semiconductor circuits at ultralow voltages: A circuit perspective on limits of switching energy,” *J. Appl. Phys.*, vol. 105, p. 094901, 2009.

[40] G. Schrom and S. Selberherr, “Ultra-low-power CMOS technologies,” in *Proc. Int. Semicond. Conf. (CAS) Dig. Tech. Papers*, 1996, pp. 237–246.

[41] M. Alioto, “Impact of NMOS/PMOS imbalance in ultra-low voltage CMOS standard cells,” in *Proc. ECCTD*, Linkoping, Sweden, Aug. 2011, pp. 557–561.

[42] Y. Ramadass and A. Chandrakasan, “Minimum energy tracking loop with embedded DC-DC converter enabling ultra-low-voltage operation down to 250 mV in 65 nm CMOS,” *IEEE J. Solid-State Circuits*, vol. 43, no. 1, Jan. 2008.

[43] A. Wang and A. Chandrakasan, “A 180-mV subthreshold FFT processor using a minimum energy design methodology,” *IEEE J. Solid-State Circuits*, vol. 40, no. 1, pp. 310–319, Jan. 2005.

[44] T.-H. Kim, J. Liu, J. Keane, and C. H. Kim, “A high-density subthreshold SRAM with data-independent bitline leakage and virtual ground replica scheme,” in *Proc. ISSCC 2007*, pp. 330–331.

[45] B. Zhai, S. Hanson, D. Blaauw, and D. Sylvester, “A variation-tolerant sub-200 mV 6-T subthreshold SRAM,” *IEEE J. Solid-State Circuits*, vol. 43, no. 10, pp. 2338–2347, Oct. 2008.

[46] C. H.-I. Kim, H. Soeleman, and K. Roy, “Ultra-low-power DLMS adaptive filter for hearing aid applications,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 11, no. 6, pp. 1058–1067, Dec. 2003.

[47] M. Ashouei, J. Hulzink, M. Konijnenburg, J. Zhou, F. Duarte, A. Breeschoten, J. Huisken, J. Stuyt, H. de Groot, F. Barat, J. David, and J. Van Ginderdeuren, “A voltage-scalable biomedical signal processor running ECG using 13 pJ/cycle at 1 MHz and 0.4 V,” in *Proc. ISSCC 2011*, San Francisco, CA, pp. 332–333.

[48] S. R. Sridhara, M. DiRenzo, S. Lingam, S.-J. Lee, R. Blazquez, J. Maxey, S. Ghanem, Y.-H. Lee, R. Abdallah, P. Singh, and M. Goe, “Microwatt embedded processor platform for medical system-on-chip applications,” in *Proc. Symp. VLSI*, 2010.

[49] M. Zwerg, A. Baumann, R. Kuhn, M. Arnold, R. Nerlich, M. Herzog, R. Ledwa, C. Sichert, V. Rzehak, P. Thanigai, and B. O. Eversmann, “An 82 $\mu\mathrm{A}/\mathrm{MHz}$ microcontroller with embedded FeRAM for energy-harvesting applications,” in *Proc. ISSCC 2011*, San Francisco, CA, pp. 334–335.

[50] J. Kwong, Y. K. Ramadass, N. Verma, and A. Chandrakasan, “A 65 nm sub-Vt microcontroller with integrated SRAM and switched capacitor DC-DC converter,” *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 115–126, Jan. 2009.

[51] S. C. Jocke, J. F. Bolus, S. N. Wooters, A. D. Jurik, A. C. Weaver, T. N. Blalock, and B. H. Calhoun, “A 2.6-$\mu\mathrm{W}$ sub-threshold mixed-signal ECG SoC,” in *Proc. VLSI Circits*, 2009.

[52] J. Keane, H. Eom, T.-H. Kim, S. Sapatnekar, and C. Kim, “Subthreshold logical effort: A systematic framework for optimal subthreshold device sizing,” in *Proc. DAC 2006*, pp. 425–428.

[53] B. H. Calhoun, A. Wang, and A. Chandrakasan, “Device sizing for minimum energy operation in subthreshold circuits,” in *Proc. CICC*, Oct. 2004.

[54] G. Gammie, N. Ickes, M. E. Sinangil, R. Rithe, J. Gu, A. Wang, H. Mair, S. Datla, B. Rong, S. Honnavara-Prasad, L. Ho, G. Baldwin, D. Buss, A. P. Chandrakasan, and U. Ko, “A 28 nm 0.6 V low-power DSP for mobile applications,” in *Proc. ISSCC 2011*, San Francisco, CA, pp. 132–133.

[55] K. Itoh, M. Horiguchi, and H. Tanaka*, Ultra-Low Voltage Nano-Scale Memories*. New York: Springer, 2007.

[56] K. Tanaka and T. Kawahara, “Leakage energy reduction in cache memory by data compression,” *ACM Sigarch Comput. Archit. News*, vol. 35, no. 5, pp. 17–24, 2007.

[57] A. Raychowdhury, S. Mukhopadhyay, and K. Roy, “A feasibility study of subthreshold SRAM across technology generations,” in *Proc. ICCD 2005*, pp. 417–422.

[58] N. Verma and A. Chandrakasan, “A 256 kb 65 nm 8T subthreshold SRAM employing sense-amplifier redundancy,” *IEEE J. Solid-State Circuits*, vol. 43, no. 1, pp. 141–149, Jan. 2008.

[59] B. Calhoun and A. Chandrakasan, “A 256-kb 65-nm sub-threshold SRAM design for ultra-low-voltage operation,” *IEEE J. Solid-State Circuits*, vol. 42, no. 3, pp. 680–688, Mar. 2007.

[60] J. P. Kulkarni, K. Kim, and K. Roy, “A 160 mV robust schmitt trigger based subthreshold SRAM,” *IEEE J. Solid-State Circuits*, vol. 42, no. 10, pp. 2303–2313, Oct. 2007.

[61] I. J. Chang, J.-J. Kim, S. P. Park, and K. Roy, “A 32 kb 10T sub-threshold SRAM array with bit-interleaving and differential read scheme in 90 nm CMOS,” *IEEE J. Solid-State Circuits*, vol. 44, no. 2, pp. 650–658, Feb. 2009.

[62] S. A. Verkila, S. K. Bondada, and B. S. Amrutur, “A 100 MHz to 1 GHz, 0.35 V to 1.5 V supply 256 × 64 SRAM block using symmetrized 9T SRAM cell with controlled read,” in *Proc. Int. Conf. VLSI Design*, 2008.

[63] K. Takeda, Y. Hagihara, Y. Aimoto, M. Nomura, Y. Nakazawa, T. Ishii, and H. Kobatake, “A read-static-noise-margin-free SRAM cell for low-VDD and high-speed applications,” *IEEE J. Solid-State Circuits*, vol. 41, no. 1, pp. 113–121, Jan. 2006.

[64] S. Nalam and B. H. Calhoun, “Asymmetric sizing in a 45 nm 5T SRAM to improve read stability over 6T,” in *Proc. CICC '09*, pp. 709–712.

[65] M. Seok, S. Hanson, D. Sylvester, and D. Blaauw, “Analysis and optimization of sleep modes in subthreshold circuit design,” in *Proc. DAC 2007*, pp. 694–699.

[66] D. Bol, C. Hocquet, D. Flandre, and J.-D. Legat, “Robustness-aware sleep transistor engineering for power-gated nanometer subthreshold circuits,” in *Proc. ISCAS 2010*, pp. 1484–1487.

[67] Y. Pu, J. Pineda-De Gyvez, H. Corporaal, and Y. Ha, “An ultra-low-energy multi-standard JPEG co-processor in 65 nm CMOS with sub/near threshold supply voltage,” *IEEE J. Solid-State Circuits*, vol. 45, no. 3, pp. 668–680, Mar. 2010.

[68] M. Seok, D. Jeon, C. Chakrabarti, D. Blaauw, and D. Sylvester, “A 0.27 V 30 MHz 17.7 nJ/transform 1024-pt complex FFT core with super-pipelining,” in *Proc. ISSCC 2011*, San Francisco, CA, pp. 342–343.

[69] W. Bracke, R. Puers, and C. Van Hoof*, Ultra Low Power Capacitive Sensor Interfaces*. New York: Springer, 2007.

[70] S. Mingoo, S. Hanson, Y.-S.. Lin, Z. Foo, D. Kim, Y. Lee, N. Liu, D. Sylvester, and D. Blaauw, “Phoenix processor: A 30 pW platform for sensor applications,” in *Proc. VLSI Circuits 2008*.

[71] B. Zhai, S. Pant, L. Nazhandali, S. Hanson, J. Olson, A. Reeves, M. Minuth, R. Helfand, T. Austin, D. Sylvester, and D. Blaauw, “Energy-efficient subthreshold processor design,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 8, pp. 1127–1137, Aug. 2009.

[72] A. P. Chandrakasan, S. Sheng, and R. W. Brodersen, “Low-power CMOS digital design,” *IEEE J. Solid State Circuits*, vol. 27, no. 4, pp. 473–484, Apr. 1992.

[73] R. Hegde and N. R. Shanbhag, “Soft digital signal processing,” *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 9, no. 6, pp. 813–823, Dec. 2001.

[74] R. Hegde and N. R. Shanbhag, “A voltage overscaled low-power digital filter IC,” *IEEE J. Solid-State Circuits*, vol. 39, no. 2, pp. 388–391, Feb. 2004.

[75] D. Ernst, S. Das, S. Lee, D. Blaauw, T. Austin, T. Mudge, N. S. Kim, and K. Flautner, “Razor: Circuit-level correction of timing errors for low-power operation,” *IEEE Micro*, vol. 24, no. 6, pp. 10–20, Nov. –Dec. 2004.

[76] S. Das, C. Tokunaga, S. Pant, W. H. Ma, S. Kalaiselvan, K. Lai, D. M. Bull, and D. Blaauw, “RazorII: In situ error detection and correction for PVT and SER tolerance,” *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 32–48, Jan. 2009.

[77] A. Bowman, J. W. Tschanz, N. Kim, J. C. Lee, C. B. Wilkerson, S.-L. Lu, T. Karnik, and V. De, “Energy-efficient and metastability-immune resilient circuits for dynamic variation tolerance,” *IEEE J. Solid-State Circuits*, vol. 44, no. 1, pp. 49–63, Jan. 2009.

[78] H. Fuketa, M. Hashimoto, Y. Mitsuyama, and T. Onoye, “Adaptive performance compensation with in-situ timing error prediction for subthreshold circuits,” in *Proc. CICC '09*, pp. 215–218.

[79] S. Ghosh, S. Bhunia, and K. Roy, “CRISTA: A new paradigm for low-power, variation-tolerant, and adaptive circuit synthesis using critical path isolation,” *IEEE Trans. Comput.-Aided Design Integr. Circuits Syst.*, vol. 26, no. 11, pp. 1947–1956, Nov. 2007.

[80] D. Bull, S. Das, K. Shivashankar, G. S. Dasika, K. Flautner, and D. Blaauw, “A power-efficient 32 bit ARM processor using timing-error detection and correction for transient-error tolerance and adaptation to PVT variation,” *IEEE J. Solid-State Circuits*, vol. 46, no. 1, pp. 18–31, Jan. 2011.

[81] S. Lee, S. Das, T. Pham, T. Austin, D. Blaauw, and T. Mudge, "Reducing pipeline energy demands with local DVS and dynamic retiming," in *Proc. ISLPED 2004*, pp. 319–324.

[82] A. Tajalli, E. J. Brauer, Y. Leblebici, and E. Vittoz, "Subthreshold source-coupled logic circuits for ultra-low-power applications," *IEEE J. Solid-State Circuits*, vol. 43, no. 7, pp. 1699–1710, Jul. 2008.

[83] A. Tajalli, M. Alioto, and Y. Leblebici, "Improving power-delay performance of ultralow-power subthreshold SCL circuits," *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 56, no. 2, pp. 127–131, Feb. 2009.

[84] M. Alioto and Y. Leblebici, "Analysis and design of ultra-low power subthreshold MCML gates," in *Proc. ISCAS*, Taipei, Taiwan, May 2009, pp. 2557–2560.

[85] A. Tajalli, M. Alioto, E. J. Brauer, and Y. Leblebici, "Design of high performance subthreshold source-coupled logic circuits," in *Proc. PATMOS*, Lisbon, Portugal, Sep. 2008, pp. 21–30.

[86] A. Tajalli, F. K. Gurkaynak, Y. Leblebici, M. Alioto, and E. J. Brauer, "Improving the power-delay product in SCL circuits using source follower output stage," in *Proc. ISCAS*, Seattle, WA, May 2008, pp. 145–148.

[87] M. Alioto and Y. Leblebici, "Circuit techniques to reduce the supply voltage limit of subthreshold MCML circuits," in *Proc. VLSI-SoC*, Rhodes Island, Greece, Oct. 2008, pp. 239–244, INVITED.

[88] D. Bol, R. Ambroise, D. Flandre, and J.-D. Legat, "Interests and limitations of technology scaling for subthreshold logic," *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.*, vol. 17, no. 10, pp. 1508–1519, Oct. 2009.

[89] S. Hanson, M. Seok, D. Sylvester, and D. Blaauw, "Nanometer device scaling in subthreshold logic and SRAM," *IEEE Trans. Electron Devices*, vol. 55, no. 1, pp. 175–185, Jan. 2008.

[90] B. Paul, A. Raychowdhury, and K. Roy, "Device optimization for digital subthreshold logic operation," *IEEE Trans. Electron Devices*, vol. 52, no. 2, pp. 237–247, Feb. 2005.

[91] M. Seok, D. Sylvester, and D. Blaauw, "Optimal technology selection for minimizing energy and variability in low voltage applications," in *Proc. ISLPED 2008*.

Massimo Alioto (M'01–SM'07) was born in Brescia, Italy, in 1972. He received the Laurea degree in electronics engineering and the Ph.D. degree in electrical engineering from the University of Catania, Italy, in 1997 and 2001, respectively.

In 2002, he joined the Department of Information Engineering of the University of Siena as a Research Associate and in the same year as an Assistant Professor. In 2005 he was appointed Associate Professor of Electronics. In the summer of 2007, he was a Visiting Professor at EPFL—Lausanne, Switzerland. In 2009–2011, he held a Visiting Professor position at BWRC—University of California, Berkeley, investigating next-generation ultra-low power circuits and wireless nodes. In 2011–2012, he is also Visiting Professor at University of Michigan, Ann Arbor, investigating on active techniques for resiliency in near-threshold processors, error-aware VLSI design for wide energy scalability, ultra-low power circuits and energy scavenging. He has authored or coauthored 170 publications on journals (60, mostly IEEE Transactions) and conference proceedings. Two of them are among the most downloaded IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATED (VLSI) SYSTEMS papers in 2007 (respectively 10th and 13th). He is coauthor of the book *Model and Design of Bipolar and MOS Current-Mode Logic: CML, ECL and SCL Digital Circuits* (Springer, 2005). His primary research interests include ultra-low power VLSI circuits and wireless nodes, near-threshold circuits for green computing, error-aware and widely energy-scalable VLSI circuits, circuit techniques for emerging technologies. He is the director of the Electronics Lab at University of Siena (site of Arezzo).

Prof. Alioto is a member of the HiPEAC Network of Excellence and the MuSyc FCRP center. He is the Chair of the "VLSI Systems and Applications" Technical Committee of the IEEE Circuits and Systems Society, for which he was also Distinguished Lecturer in 2009–2010 and member of the DLP Coordinating Committee in 2011–2012. He serves or has served as a Track Chair in a number of conferences (ISCAS, ICCD, ICECS, APCCAD, ICM). He was Technical Program Chair of the ICM 2010 and NEWCAS 2012 conferences. He serves as Associate Editor of the IEEE TRANSACTIONS ON VERY LARGE SCALE INTEGRATED (VLSI) SYSTEMS, the *ACM Transactions on Design Automation of Electronic Systems*, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS—PART I, the *Microelectronics Journal*, the *Integration—The VLSI Journal*, the *Journal of Circuits, Systems, and Computers*, the *Journal of Low Power Electronics*, as well as the *Journal of Low Power Electronics and Applications*. He was also Guest Editor of the Special Issue "Advances in Oscillator Analysis and Design" of the *Journal of Circuits, Systems, and Computers*.