Preface

The integrated circuit has evolved tremendously in recent years as Moore's Law has enabled exponentially more devices and functionality to be packed onto a single piece of silicon. In some ways however, these highly integrated circuits, of which microprocessors are the flagship example, have become victims of their own success. Despite dramatic reductions in the switching energy of the transistors, these reductions have kept pace neither with the increased integration levels nor with the higher switching frequencies. In addition, the atomic dimensions being utilized by these highly integrated processors have given rise to much higher levels of random and systematic variation which undercut the gains from process scaling that would otherwise be realized. So these factors—the increasing impact of variation and the struggle to control power consumption—have given rise to a tremendous amount of innovation in the area of adaptive techniques for dynamic processor optimization.

The fundamental premise behind adaptive processor design is the recognition that variations in manufacturing and environment cause a statically configured operating point to be far too inefficient. Inefficient designs waste power and performance and will quickly be surpassed by more adaptive designs, just as it happens in the biological realm. Organisms must adapt to survive, and a similar trend is seen with processors – those that are enabled to adapt to their environment, will be far more competitive. The adaptive processor needs to be made aware of its environment and operating conditions through the use of various sensors. It must then have some ability to usefully respond to the sensor stimulus. The focus of this book is not so much on a static configuration of each manufactured part that may be unique, but on *dynamic* adaptation, where the part optimizes itself on the fly.

Many different responses and adaptive approaches have been explored in recent years. These range from circuits that make voltage changes and set body biases to those that generate clock frequency adjustments on logic. New circuit techniques are needed to address the special challenges created by scaling embedded memories. Finally, system level techniques rely on self-correction in the processor logic or asynchronous techniques which remove the reliance on clocks. Each approach has unique challenges

and benefits, and it adds value in particular situations, but regardless of the method, the challenge of reliably testing these adaptive approaches looms as one of the largest. Hence the subtitle the book: Theory and Practice. Ideas (not necessarily good ones) on adaptive designs are easy to come by, but putting these in working silicon that demonstrates the benefits is much harder. The final level of achievement is actually productizing the capability in a high-volume manufacturing flow.

In order for the book to do justice to such a broad and relatively new topic, we invited authors who have already been pioneers in this area to present data on the approaches they have explored. Many of the authors presented at ISSCC2007, either in the Microprocessor Forum, or in the conference sessions. We are humbled to have collected contributions from such an impressive group of experts on the subject, many of whom have been pioneers in the field and produced results that will be impacting the processor design world for years to come. We believe this topic of adaptive design will continue to be a fertile area for research and integrated circuit improvements for the foreseeable future.

Alice Wang — *Texas Instruments, Inc.*
Samuel Naffziger — *Advanced Micro Devices, Inc.*

Chapter 2 Technological Boundaries of Voltage and Frequency Scaling for Power Performance Tuning

Maurice Meijer[1], José Pineda de Gyvez[1,2]

[1] NXP Semiconductors, [2] Eindhoven University of Technology

In this chapter, we concentrate on technological quantitative pointers for adaptive voltage scaling (AVS) and adaptive body biasing (ABB) in modern CMOS digital designs. In particular, we will present the power savings that can be expected, the power-delay trade-offs that can be made, and the implications of these techniques on present semiconductor technologies. Furthermore, we will show to which extent process-dependent performance compensation can be used. Our presentation is a result of extensive analyses based on test-circuits fabricated in the state-of-the-art CMOS processes. Experimental results have been obtained for both 90nm and 65nm CMOS technology nodes.

2.1 Adaptive Power Performance Tuning of ICs

The integration density of Integrated Circuits is doubling every 18 months. Soon, advanced process generations will integrate 1 billion transistors on a single chip. Such chips are the heart of a new generation of devices that are changing our daily life fundamentally. Power consumption of conventional electronic devices is a major concern because the dense devices produce a significant amount of heat imposing constraints on circuit performance and IC packaging. The case for portable devices is obvious, e.g. the goal is to maximize battery time. Designing ICs for low power will be a key practical and competitive advantage in the coming decade.

From a technological standpoint, power consumption can be reduced by downscaling transistor dimensions. CMOS transistor scaling consists of

A. Wang, S. Naffziger (eds.), *Adaptive Techniques for Dynamic Processor Optimization*,
DOI: 10.1007/978-0-387-76472-6_2,

reducing all dimensions by a factor k (≈1.4), enabling higher integration density [1]. In the constant-field scaling scenario, the circuit speed increases, theoretically, with the amount of scaling k. Constant-field scaling has known benefits such as lower power per circuit, constant power density, and power-delay product that increases by k^3. However, for CMOS technology, over the last 10 years, it has been impossible to scale power supply voltage (V_{DD}) while maintaining speed because of the constraints on the threshold voltage (V_{th}) [2]. Due to increasing leakage current in scaled devices, V_{th} is not lowered to avoid significant static power consumption. Therefore, the electrical field is rising in proportion to k resulting now in almost constant circuit power despite scaling, increased power density by k^2, and power-delay product improvement by a factor of k only. In essence, the limits of a scaling process are caused by physical effects that do not scale properly, among them are quantum-mechanical tunneling, discrete carrier doping, and other voltage-related effects such as the subthreshold swing, and built-in voltage and minimum voltage swings.

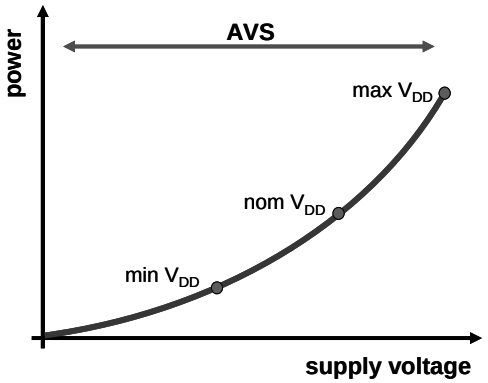

Figure 2.1 Power trends as a function of the supply voltage.

Besides technology scaling, one of the most effective ways to reduce active power consumption is by lowering V_{DD}. Ideally, quadratic power savings are observed as displayed in Figure 2.1. V_{DD} reduction can be applied to a complete chip, but it is most effective when it is applied to local voltage domains with own performance requirements. A common approach is to perform dynamic supply scaling, which exploits the temporal domain to optimize V_{DD} at run-time. This technique dynamically varies both operating frequency and supply voltage in response to workload demands. In this way, a processing unit always operates at the desired performance level while consuming the minimal amount of power. Two basic flavors exist, namely dynamic voltage scaling (DVS) and adaptive voltage scaling (AVS). DVS is

an open-loop approach, and it is based on the selection of operating points from a predefined {f,V} table. Alternatively, AVS is a closed-loop approach, and its operating points are based only on the frequency. Software decides on the performance required for the existing workload and selects a target frequency. The voltage is then automatically adjusted to support this frequency. AVS is considered as the most effective technique for achieving power savings through V_{DD} scaling.

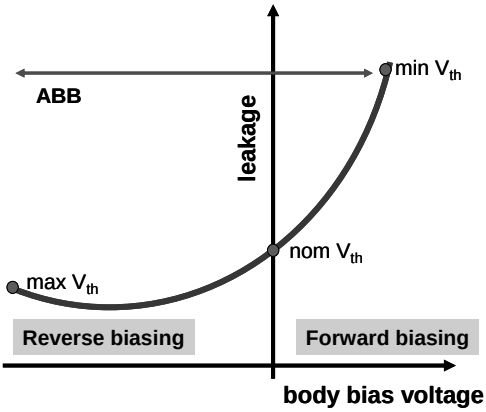

Figure 2.2 Leakage trends as a function of body biasing.

Yet another, but complementary, approach is to adapt to the threshold voltage of MOS devices using transistor body biasing. For NMOS, the V_{th} is increased when its body–source voltage is biased to be negative. This is referred to as reverse body biasing (RBB). Alternatively, the V_{th} is reduced when the body–source voltage is biased to be positive. This is referred to as forward body biasing (FBB). Figure 2.2 illustrates the behavior of leakage as a function of body biasing in modern nanometer technologies. Body biasing can effectively reduce the leakage power of the design, by improving its run-time performance. It is most effective when it is used in conjunction with V_{DD} scaling. Typically, body biasing is done in open-loop to calibrate circuit frequency or leakage for setting a desired mode of operation. Adaptive body biasing (ABB) refers to closed-loop control in which circuit parameters, e.g. speed, are monitored, compared, and controlled against desired values.

Not surprisingly, in recent years, the application of adaptive circuit techniques to control either or both V_{DD} and V_{th} has gained increased attention. This stems from the fact that modern electronics are hampered by the variation of fundamental process and performance parameters such as threshold voltage and power consumption. Design technologies such as

AMD's PowerNow! [3], Transmeta's LongRun [4], Intel's Enhanced SpeedStep [5], are vivid examples of commercial ICs that use power management based on V_{DD} scaling. In addition to these commercial accomplishments, chip demonstrators with V_{DD} and V_{th} scaling capabilities have also been reported in the literature archival [6–8]. Other reported uses of V_{DD} and V_{th} scaling, besides power management in processors, are in testing [9], product binning [10], and yield tuning [11].

2.2 AVS- and ABB-Scaling Operations

As the benefits of V_{DD} and V_{th} scaling are known, we concentrate on quantitative pointers for using such know-how in deep submicron technologies. For this purpose, we have evaluated various process technologies to determine technological boundaries for AVS and ABB when applied to digital logic circuits. Our evaluation is based on an extensive analysis of test-circuits fabricated in 90nm general-purpose (GP), 90nm low-power (LP), and 65nm low-power (LP) triple-well CMOS processes.

For all three CMOS processes, we have designed a clock generator unit (CGU) that consists of multiple independent ring-oscillators and corresponding selection circuitry. We use these CGU designs to determine power-performance trade-offs and leakage reduction factors with AVS and ABB. Each ring-oscillator uses minimum-sized standard-cell inverters as delay elements and a nand-2 gate for enabling control. The power supply of the clock generator can be controlled externally. Body biasing is enabled for N-well and P-well independently through triple-well isolation. The exact same clock generator was laid out in 90nm GP and LP-CMOS using a commercial place-and-route tool with constrained area-routing features. The 65nm LP-CMOS clock generator was designed full-custom using digital standard cells. Our second test-chip is a circular shift-register, which has only been laid out in 90nm LP-CMOS. The design contains 8K flip-flops and 50K logic gates. The logic gates are connected as delay lines between two consecutive flip-flop stages, which have an average logic depth of six cells. One can emulate the activity of any digital core with this circular shift register by shifting in a sequence of zeros and ones. Like the CGU, it has independent bias control over supply voltage, N-well and P-well biasing. The CGU provides the clock to the shift-register. The shift-register is used to perform correlated measurements against the CGU for validation purposes. All measurements have been performed using a Verigy 93K SoC test system in a controlled temperature environment. The temperature is controlled by a Temptronic Thermostream.

Devices in 90nm GP-CMOS operate at a nominal V_{DD} of 1V; their counterparts in LP-CMOS operate at 1.2V. GP-CMOS devices exhibit a lower V_{th} than LP-CMOS devices. On average, the nominal V_{th} is about 0.27V, 0.37V, and 0.43V for 90nm GP, 90nm LP, and 65nm LP-CMOS, respectively. Since ABB enables adaptation of these nominal V_{th} values, we will show the range over which V_{th} can be tuned for one of the considered process technologies. Figure 2.3 puts into perspective V_{th} versus body biasing for 65nm LP-CMOS devices as obtained from circuit simulations. Observe that the actual value of V_{th} and its sensitivity to body bias strongly depend on the process corner: fast, typical, or slow. For the typical NMOS device, body biasing from 0.4V (FBB) down to –1.2V (RBB) spans over a V_{th} range of about 135mV. This range is somewhat larger for PMOS devices (~180mV). Since RBB has a direct impact on leakage reduction, it will become evident that this technique is not very effective because the sensitivity of V_{th} to V_{BS} is small. In the next sections, we quantify the impact of these V_{th} ranges on circuit power-performance tuning.

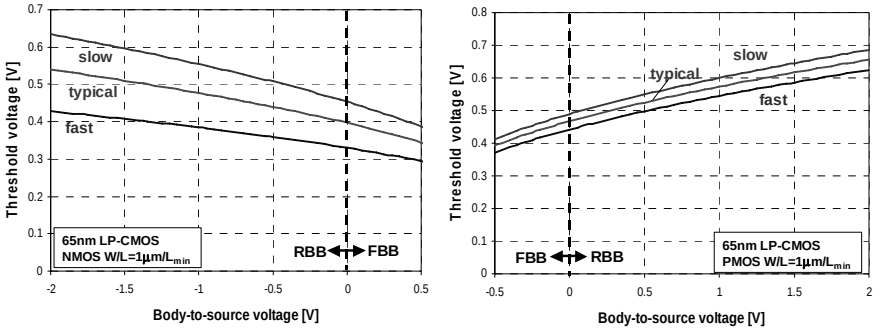

Figure 2.3 V_{th} adaptation through body biasing in 65nm LP-CMOS.

Let us now briefly introduce the conventions used for the AVS and ABB schemes. Figure 2.4 shows a graph of frequency versus power as a function of either or both AVS and ABB. The thick line shows the nominal trend when the supply voltage is varied from its maximum to its minimum value. The AVS operation consists of sweeping the supply voltage while maintaining a nominal constant body bias. The ABB is essentially the contrary approach: the supply voltage is kept constant and the body bias is swept. Here, it holds that frequency and power have an almost linear negative dependence on the threshold voltage. The result is a “cloud” of frequency–power points for a given supply voltage. Finally, AVS+ABB corresponds to the case when both supply voltage and body biasing are swept.

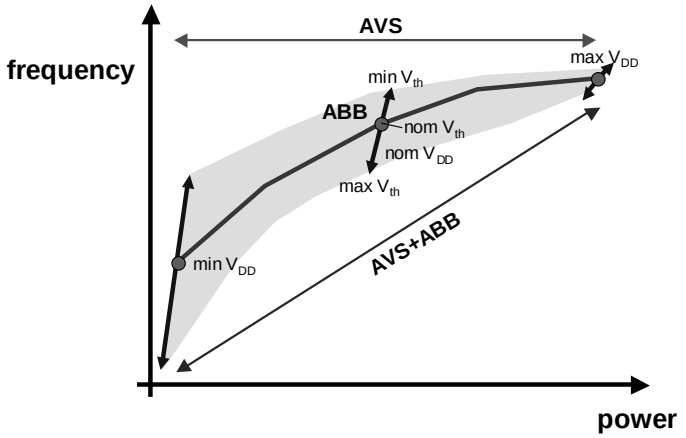

Figure 2.4 AVS and ABB operations.

Table 2.1 presents the voltage ranges that we employed during our measurements. Observe that the wells were forward biased for at most 0.4V and reverse biased by 1V (GP) or 1.2V (LP). Forward biasing is constrained by the turn-on voltage of the transistors' body–source junction diode. Essentially, reverse biasing is unconstrained, but high reverse biasing voltages result in increased gate-induced drain leakage.

Table 2.1 Voltage conventions for scaling operations.

		90nm GP	**90nm/65nm LP**
AVS	V_{DD}	[0.5,1.0]V	[0.6,1.2]V
ABB	V_{nwell}	[V_{DD}–0.4,V_{DD}+1.0]V	[V_{DD}–0.4,V_{DD}+1.2]V
	V_{pwell}	[–1.0,0.4]V	[–1.2,0.4]V
AVS+ABB	V_{DD}	[0.5,1.0]V	[0.6,1.2]V
	V_{nwell}	[V_{DD}–0.4,V_{DD}+1.0]V	[V_{DD}–0.4,V_{DD}+1.2]V
	V_{pwell}	[–1.0,0.4]V	[–1.2,0.4]V

In the next sections, we will illustrate how these techniques can be used to alter the power performance of integrated circuits. Please note that in the next sections, we will use the term ringo to refer to the ring oscillators in the CGU.

2.3 Frequency Scaling and Tuning

In most applications, there is not always a need for peak performance. In those cases, AVS can be used to lower the supply voltage and to slow down the core's computing power. In fact, operating frequency and supply voltage for a circuit design are coupled. This relationship can be expressed by Sakurai's alpha-power model [12]:

$$f \approx K \cdot \frac{(V_{DD} - V_{th})^{\alpha}}{V_{DD}} \tag{2.1}$$

where f is the operating frequency, K is a proportionality factor, and α is a process-dependent parameter that models velocity saturation. In the case of velocity-saturated devices, α is close to 1 and the frequency scales almost linearly with V_{DD}.

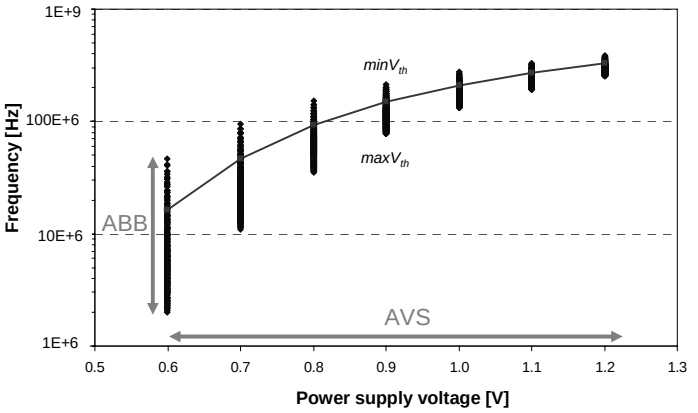

Figure 2.5 Frequency scaling and tuning for the 65nm LP-CMOS ringo.

Let us now investigate the frequency-scaling and tuning ranges offered by AVS and ABB in 65nm LP-CMOS. For this purpose, we determined the dynamic range of a 101-stage ringo that is part of the CGU test-chip. Figure 2.5 shows the ringo frequency as a function of power supply. Each cloud of dots is associated to a unique supply voltage. Each dot in a cloud corresponds to a unique N-well and P-well bias combination, and the line joining the clouds indicates the nominal trend. The ringo frequency at nominal supply (V_{DD}=1.2V) is 327MHz, and 16.2MHz at minimum supply (V_{DD}=0.6V). This results in an AVS tuning range of about 310MHz. Recall

that the V_{th} is about 0.43V on average for this technology at nominal V_{DD}. When operating at reduced V_{DD}, the V_{th} increases due to of drain-induced barrier lowering (DIBL). At V_{DD}=0.6V, the V_{th} increases by about 100mV. The large frequency reduction with AVS is because the supply voltage becomes close to the V_{th}. For those low V_{DD}s, the transistors are no longer velocity saturated (α=2). For the applied range, AVS renders an approximate 20× frequency reduction. If the lower bound of AVS would be set to 0.7V, the frequency reduces by about 7×.

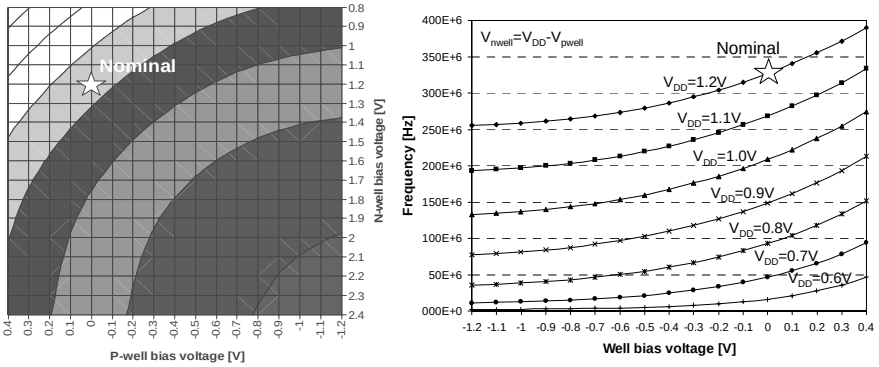

Figure 2.6 Frequency dependence on body-bias voltages; (a) Independent well biasing and V_{DD}=1.2V, (b) Symmetrical well biasing and various V_{DD} voltages.

We can now analyze the impact of ABB as a frequency-tuning mechanism at each V_{DD} point. Notice that the relative-tuning range is not the same for all V_{DD} values. In particular, we measured frequency spans of approximately –87% to +188% at V_{DD}=0.6V and approximately ±20% at V_{DD}=1.2V with respect to their nominal frequencies. The larger tuning range of ABB at reduced supply voltages can be explained by the fact that the threshold voltage is a larger portion of the gate drive of the transistors. At such low gate drive, the frequency becomes very sensitive to changes in V_{th}. Notice that a tuning range of –87% at V_{DD}=0.6V implies an 8.1× lower frequency for RBB. In fact, at V_{DD}=0.6V, the circuit operates in the subthreshold region for strong reverse body-biasing conditions. In this case, the current is exponentially related to the gate drive voltage, and the frequency is much lower than in case of nominal body biasing. For the measured silicon, ABB gives an absolute tuning range of 135MHz for the chosen N-well and P-well voltages when operating at V_{DD}=1.2V. At V_{DD}=0.6V, this tuning range is around 45MHz. Figure 2.6a shows a contour plot of the ABB-scaling operation at V_{DD}=1.2V. The contours are at 20MHz intervals, and the nominal frequency is at 327MHz. Notice that

it is possible to change the V_{th} of the PMOS and NMOS transistors independently and still attain the same frequency. Obviously, the choice of V_{th} has a significant impact on leakage power consumption as we will show later in this chapter. Figure 2.6b shows the frequency tuning for the ABB-scaling operation as function of a symmetrical well bias ($V_{nwell}=V_{DD}-V_{pwell}$) and various supply voltages. Notice that the frequency saturates for strong, reverse body biasing due to its limited V_{th} control range.

The same analysis has been performed for ringos in 90nm CMOS. A summary of the measured frequency-scaling and tuning ranges is given in Table 2.2. Notice the large frequency-scaling range for 65nm LP-CMOS as well as the large frequency-tuning range at reduced V_{DD}. For severe reverse body biasing, the threshold voltage saturates yielding as a result an asymptotic limit on the lowest possible operating frequency. Observe that GP-CMOS shows a lower dependence on V_{DD} and V_{th} as compared to LP-CMOS primarily because the threshold voltage of the former technology is lower.

Table 2.2 Frequency-scaling and tuning ranges for 90nm/65nm CMOS.

		90nm GP	**90nm LP**	**65nm LP**
AVS		3.4×	5.9×	20.1×
ABB	$V_{DD}/2$	[–29,24]%	[–81,76]%	[–87,188]%
	V_{DD}	[–8,6]%	[–27,15]%	[–22,19]%
AVS+ABB		5.1×	34.9×	194.1×

2.4 Power and Frequency Tuning

The ultimate use of the AVS and ABB schemes is for performance tuning with performance being the optimal combination of frequency and power, i.e. the lowest power for a given frequency. To investigate the available power–frequency-tuning range offered by AVS and ABB in 65nm LP-CMOS, we consider the same ring oscillator as before. Figure 2.7 presents a plot of the ringo frequency as function of the total power of the CGU, e.g. both CGU-static and dynamic power consumption of the ringo. In our experiments, static power takes into account all sources of leakage, e.g. subthreshold leakage, gate-oxide leakage, etc.

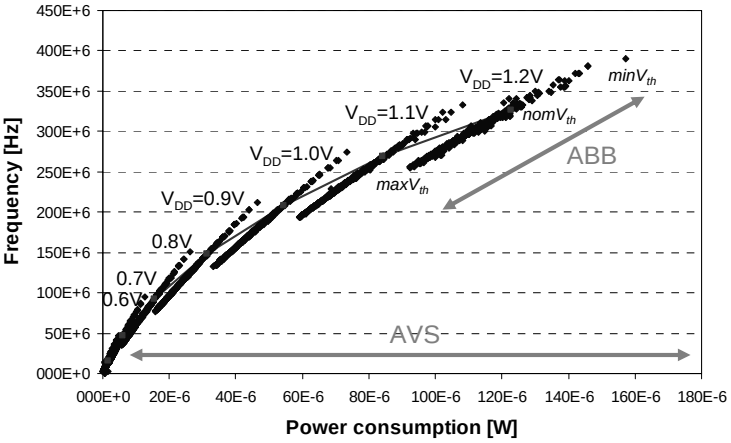

Figure 2.7 Frequency versus total power.

The plot of Figure 2.7 allows us to evaluate power savings and tuning-range control of AVS and ABB. Measurement results indicate 82× power savings by 20.1× frequency downscaling, using AVS when downscaling V_{DD} from 1.2V to 0.6V. The use of ABB at V_{DD} = 1.2V results in ±22% power and ±20% frequency tuning with respect to the nominal operating point. At V_{DD} = 0.6V, we observe a power-tuning range that spans from 78% to +217% and a frequency-tuning range from –87% to +188% with respect to no ABB. The combination of AVS and ABB yields ~790× power savings with ~194× frequency scaling from the highest possible frequency (minimum V_{th}) to the lowest one (maximum V_{th}). These results show the strength of the combined use of AVS and ABB.

Let us now explore possible power-performance tradeoffs by using AVS and ABB. Figure 2.8a shows a zoom-in of Figure 2.7 at V_{DD} =1.2V. If AVS and ABB are applied such that the nominal V_{DD} becomes 1.1V

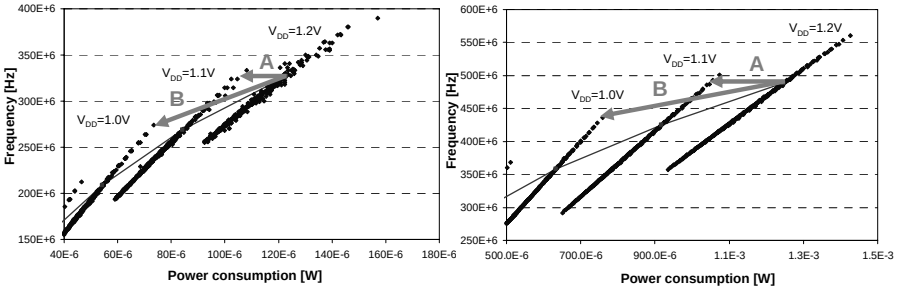

Figure 2.8 Frequency versus total power trade-off; (a) 65nm LP-CMOS, (b) 90nm LP-CMOS.

instead of 1.2V, and the V_{th}s are pulled to a smaller value as indicated by arrow A in Figure 2.8a, we see that it is possible to achieve ~14% power savings with no frequency penalty. A more aggressive V_{DD} downscaling to 1.0V, while pulling the V_{th}s to their minimum value, results in 40% power savings at about 16% frequency penalty as indicated by arrow B. Similar results have been found for 90nm LP-CMOS as shown in Figure 2.8b. In this case, the index factors are 16% power savings with no frequency penalty at V_{DD}=1.1V and 39% power savings with 11% frequency penalty at V_{DD}=1.0V. The benefits of combined AVS+ABB are not found to be technology-node dependent for the considered LP-CMOS process technologies. For 90nm GP-CMOS, however, a slightly larger voltage dependency of performance was observed. Downscaling from its nominal V_{DD} of 1.0V–0.9V, and lowering the V_{th}s a minimum, results in ~23% power savings with ~6% frequency penalty. At V_{DD}=0.8V and minimum V_{th}s, ~48% power savings are achieved with ~18% frequency penalty only. This indicates that there exists a lower frequency-tuning range with ABB for GP-CMOS.

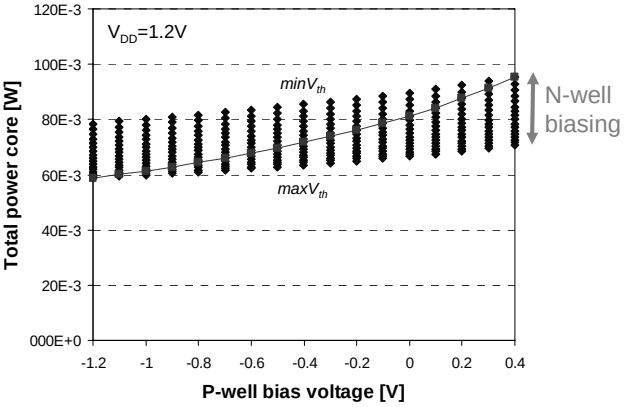

Figure 2.9 Power of 90nm LP-CMOS core as a function of well biasing.

Next we will investigate the properties of ABB in 90nm LP-CMOS on the shift register. Figure 2.9 shows the core's total power for a given circuit activity and V_{DD}=1.2V. Each dot in the clouds is associated to an N-well biasing condition. The line joining the clouds indicates the case when symmetric well biasing is applied. Observe that the well biasing allows a total power-tuning range of about 36mW; this represents about 40% of the nominal power consumption.

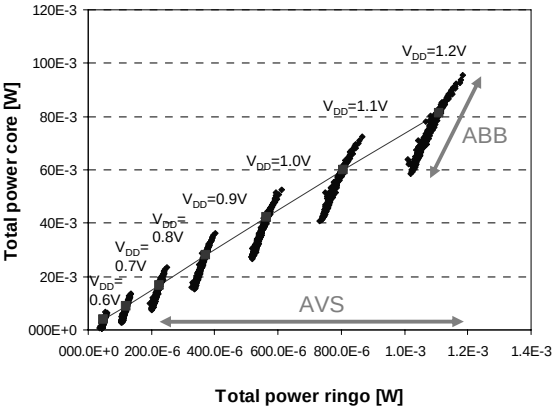

Figure 2.10 Total power correlation for the shift register and the ringo for different V_{DD} values.

Figure 2.10 shows the power consumption correlation between the shift register and the ringo for different V_{DD} values. In this plot, we have used the same conventions as before, i.e. each cloud is associated to a unique V_{DD} value and each point in the cloud corresponds to a unique N-well and P-well bias combination. The shift register operates at the same V_{DD} as the CGU, while its operating frequency is provided by the CGU. The circuit activity of the shift register is kept constant. The dynamic power dominates the total power in both circuit blocks, and therefore, their total power can be estimated by $P \approx aC \cdot V_{DD}^2 \cdot f$, where aC represents the switching circuit capacitance. Since both circuit blocks operate at the same supply voltage and frequency, their power consumption is linearly related by a ratio determined by the switching circuit capacitance. This can be observed in Figure 2.10, where the power consumption of the circuit blocks remains linearly correlated while applying AVS and/or ABB.

Table 2.3 puts into perspective the power–frequency ranges for the ringos in the considered process technologies. Notice that there exist large power–frequency ranges for each process technology. For the cases of AVS only, or AVS+ABB, the ratio of power and frequency shows a factor of 4× energy savings when scaling for the nominal V_{DD} to half of its value. This indicates that the total ringo power is dominated by dynamic power consumption. Furthermore, observe that LP-CMOS offers a larger power- and frequency-tuning range than GP-CMOS when utilizing ABB alone. The frequency-tuning range of GP-CMOS is about 3× lower.

Table 2.3 Power–frequency-tuning ranges for 90nm and 65nm CMOS.

			90nm GP	90nm LP	65nm LP
AVS		Power savings +	13.7×	23.6×	82.0×
		frequency penalty	3.4×	5.9×	20.1×
ABB	$V_{DD}/2$	Power tuning	[–29,29]%	[–77,65]%	[–78,217]%
		Frequency tuning	[–29,24]%	[–81,76]%	[–87,188]%
	V_{DD}	Power tuning	[–9,10]%	[–25,14]%	[–25,28]%
		Frequency tuning	[–8,6]%	[–27,15]%	[–22,19]%
AVS+ABB		Power savings +	21.2×	117.1×	790.5×
		frequency penalty	5.1×	34.9×	194.1×

2.5 Leakage Power Control

Leakage power is one of the main concerns in deep submicron technologies. In fact, AVS and ABB are often used for leakage reduction purposes. For older process technologies, leakage current is dominated by subthreshold conduction. Subthreshold leakage for a given device strongly depends on threshold voltage choice, process condition, supply voltage, and temperature. For sub-100nm CMOS, other leakage components have become increasingly important [13]. The most prominent ones are direct tunneling currents through the thin gate-oxide and gate-induced drain leakage (GIDL). Both leakage components are strongly V_{DD} dependent. Figure 2.11 puts into perspective leakage current as a function of power supply and temperature for a high-V_{th} NMOS device in 65nm LP-CMOS technology. These results are obtained through circuit simulations for a typical process condition. Observe in Figure 2.11a that subthreshold leakage, gate-oxide tunneling, and GIDL currents are of the same order of magnitude at nominal process–voltage–temperature conditions. Both Figure 2.11a,b show that the dominant leakage component in the total leakage depends on the operating condition.

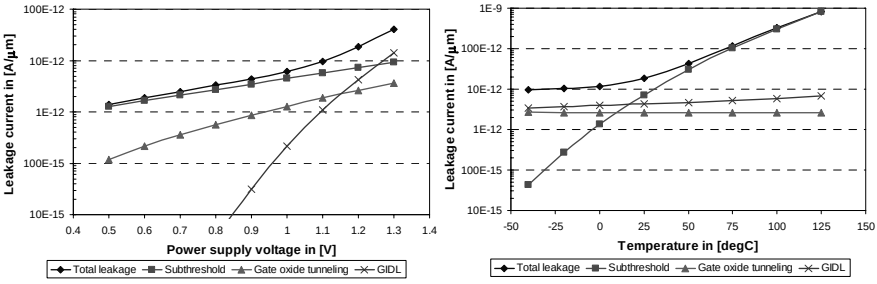

Figure 2.11 Leakage current trends for a 65nm LP-CMOS high-V_{th} NMOS device; (a) V_{DD} dependency at 25°C, (b) temperature dependency at V_{DD}=1.2V.

Figure 2.12 shows the impact of AVS and ABB on the leakage current for our CGU in 65nm LP-CMOS at 25°C. The plot shows measured leakage current versus body bias for three distinct values of power supply. Body biasing is applied symmetrically for N-well and P-well, respectively. The forward and reverse body-biasing ranges are indicated. Clearly, it is shown in Figure 2.12 that the leakage current grows exponentially when applying forward body biasing; this is because of the increased subthreshold leakage when lowering the V_{th}s. In reverse body-biasing operation, the leakage current achieves a minimum value around 500mV RBB. For stronger reverse body biasing, GIDL dominates the leakage current eliminating the ability of ABB to reduce leakage. Observe in Figure 2.12 that applying RBB of 300mV at V_{DD}=1.2V is as effective as lowering V_{DD} by that same amount. For larger RBB at V_{DD}=1.2V, AVS becomes more effective to reduce leakage. This is because GIDL and gate-oxide leakage are strongly reduced for lower V_{DD} operation.

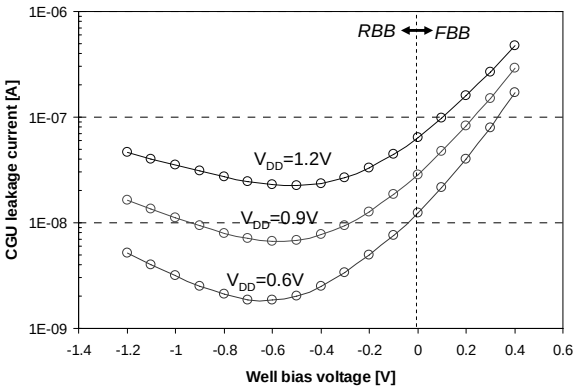

Figure 2.12 Leakage reduction in 65nm LP-CMOS using AVS and ABB.

For the measured die sample, leakage reduces by 5.1× when V_{DD} is scaled down from 1.2V to 0.6V. When using ABB alone at V_{DD} = 1.2V, leakage decreases only by 2.9×. This low impact of ABB is because of a high level of GIDL as explained before. When using ABB alone at V_{DD}=0.6V, leakage decreases by 6.8×. The combination of AVS with ABB renders a leakage reduction of 34.6×. Forward body biasing by 0.4V at V_{DD}=1.2V, 0.9V, or 0.6V increases the leakage current by 7.4×, 10.2×, or 13.7×, respectively.

The actual leakage savings utilizing AVS and ABB are impacted by temperature. At elevated temperatures, the V_{th}s become lower causing subthreshold leakage to become a bigger part of the total leakage current.

GIDL depends only weakly on temperature, and gate-oxide leakage is not temperature dependent. We have also measured temperature dependence of leakage current for various die samples to quantify its impact on the potential of AVS and ABB, to reduce leakage. Figure 2.13 shows experimental results for leakage reduction versus temperature for the same die sample as before. Observe that AVS becomes less effective to reduce leakage with increasing temperature, since the related leakage increase is supply voltage independent. However, the leakage increase is threshold voltage dependent, and therefore, ABB can reduce leakage slightly more effectively when temperature increases. At very high temperatures, i.e. the case of 100°C, the V_{th} is lowered so much that ABB cannot further reduce leakage because of the constrained ABB range we used in our experiments. The trend of AVS+ABB shows the collective effect of reducing leakage by AVS and ABB. In this case, leakage savings are about constant for temperatures up to 75°C.

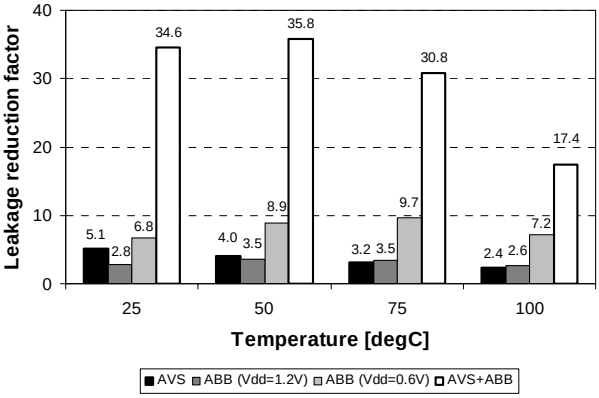

Figure 2.13 Temperature-dependent leakage reduction in 65nm LP-CMOS.

The actual leakage savings achieved by AVS and ABB are also impacted by process parameter variations. Subthreshold leakage strongly depends on process state, while gate-oxide leakage and GIDL are only weakly dependent. Leakage current of the CGU has been measured for 40 die samples from the same silicon wafer at 25°C. We have observed a leakage current ranging from 17.3nA to 322.6nA, depending on the die sample. This corresponds to leakage current variations of about 18.7×.

Table 2.4 shows the average leakage current savings for 65nm LP-CMOS obtained for the measured 40 die samples. The reduction factors for 90nm GP- and LP-CMOS technologies are also shown in this Table. The product of leakage savings with AVS ($V_{DD}/2$) and ABB yields substantial benefits as indicated in row AVS+ABB.

Table 2.4 Leakage current reduction for 90nm and 65nm CMOS at 25°C operation.

		90nm GP	90nm LP	65nm LP
AVS		5.3×	3.3×	5.6×
ABB	$V_{DD}/2$	4.1×	6.6×	4.5×
	V_{DD}	1.2×	3.5×	2.5×
AVS+ABB		21.6×	21.5×	24.8×

2.6 Performance Compensation

Understanding the trade-offs in performance and power is not sufficient to ensure a successful outcome of the IC. The basic problem is that failure of deep submicron process technologies to continue with constant process tolerances opens avenues for new challenging low-power process options and emerging design technologies. Basically, the assimilation of distinct high-performance, low operating power, and low standby power devices requires circuits and systems that concurrently exploit many degrees of freedom in both fabrication and design technologies.

Figure 2.14 shows the impact of process variability on performance spread of a single inverter for various technology nodes. A proportional inverter sizing was done across technology nodes for comparison

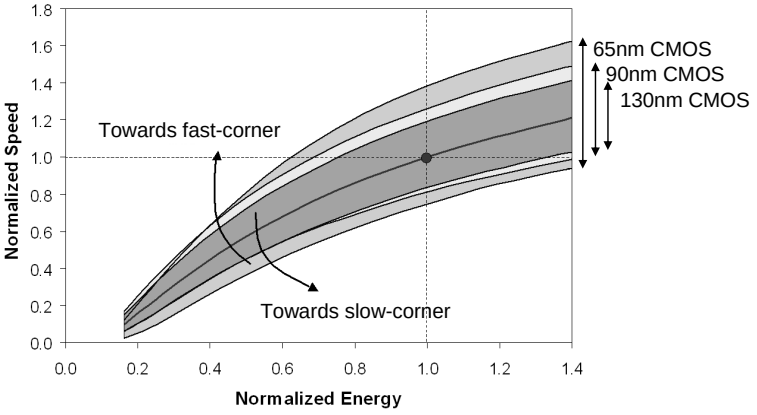

Figure 2.14 Energy spread across various technology nodes.

purposes. The inverter has further a fan-out of four gates. The vertical axis basically shows the spread of speed over three process corners, e.g. typical–slow–fast. The horizontal axis shows the normalized energy per operation. Notice that the performance window spread for 130nm, 90nm, and 65nm CMOS is about 40%, 50%, and 70%, with respect to the nominal operating conditions, respectively. What this graph also shows is that for a constant throughput, the wider the performance spread, the better the opportunities for energy savings are if voltage scaling is applied. For instance, in 65nm CMOS, the normalized speed of "1" can be achieved at an energy of "0.6" instead of at an energy of "1" if the power supply is scaled down. Today's design practices advocate a worst-case design style to ensure a target speed. This brings as implications overhead in area and power as shown in Figure 2.14. Basically, a worst-case design requires stronger cells, which are bigger in area and are also bigger power consumers, to meet timing closure of designs that fall beyond the 3σ due to process variability.

Figure 2.15 shows the impact of process variability on leakage power of the same inverter. One can see that leakage power spread at nominal supply voltage can span over 7×, 9×, and 11× for 130nm, 90nm, and 65nm CMOS, respectively. This spread can be detrimental in ultra low-power designs.

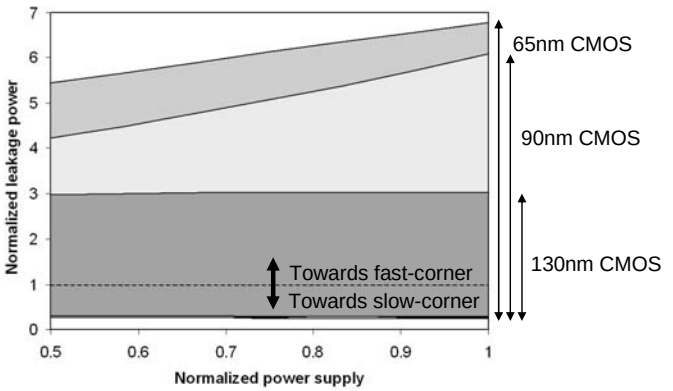

Figure 2.15 Leakage spread across various technology nodes.

As the variation of fundamental parameters such as channel length, threshold voltage, thin oxide thickness, and interconnect dimensions goes well beyond acceptable limits, "on-the-fly" performance compensation is becoming necessary. The influence of process parameter spread on circuit

behavior becomes higher and higher. For instance, in older technologies greater than 0.18μm, a V_{th} spread of say 50mV on a nominal V_{th} of 450mV was not that crucial; in nanometer technologies with a nominal V_{th} of 250mV, this variation can make circuit operation quite difficult.

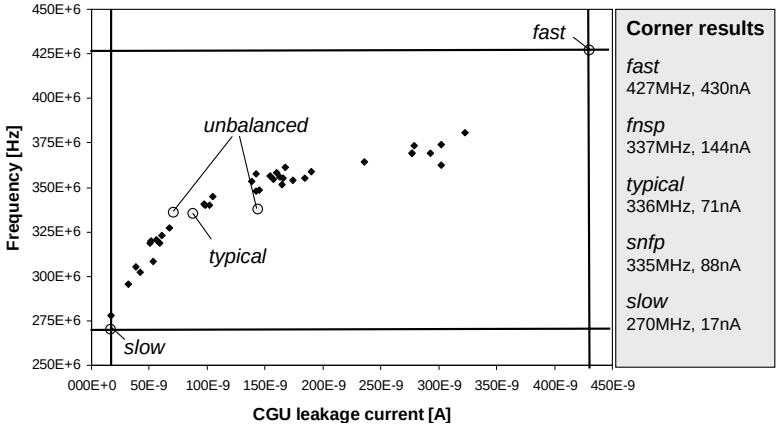

Figure 2.16 Frequency and leakage spread for 40 die samples of the same 65nm LP-CMOS wafer.

Figure 2.16 shows an example of frequency and leakage spread in which ringo frequency versus CGU leakage current is plotted at nominal V_{DD} for 40 die samples coming from the same 65nm LP-CMOS wafer. The five corner specifications for ringo frequency versus CGU leakage, as determined from circuit simulations, are also indicated in Figure 2.16. The total frequency and the leakage spread of the measured die samples are about 100MHz and 305nA, respectively. This translates into a relative frequency spread of ~36% and a relative leakage spread of ~18.7×. Note that we consider the samples with frequencies below "typical" as yield losses, while samples above "typical" are consuming unnecessary extra power. Moreover, the leakage current for a "fast" corner sample is about ~6.1× higher as compared to the "typical" reference, while the leakage current for a "slow" corner sample is about ~4.2× lower.

Next, we will discuss three strategies for compensating the undesired process-dependent frequency and leakage spread by means of post-silicon tuning. A first strategy is to perform post-silicon tuning with ABB only. From experiments, we have determined the tuning ranges for "fast" and "slow" samples. Figure 2.17 shows the potential of ABB to compensate performance for the same die samples as shown before. A 21% frequency increment from the slow corner renders a target frequency of 327MHz, and

likewise, a 14% adjustment from the fast corner results in a target frequency of 366MHz. At the same time, the leakage current increases by ~9.8× (from 17nA to 170nA) for a "slow" corner sample, and reduces by ~2.5× (from 430nA to 177nA) for a "fast" corner sample. Observe that in both cases, that is, from slow to typical and from fast to typical, the leakage current of the tuned device is approximately 2.4× higher than the "typical" reference. For the available die sample set, we showed that the application of ABB gives basically a 100% parametric yield improvement. In addition, the leakage spread can be reduced to a factor of ~3.8× as indicated in Figure 2.17 by the dotted line at a typical frequency of 336MHz.

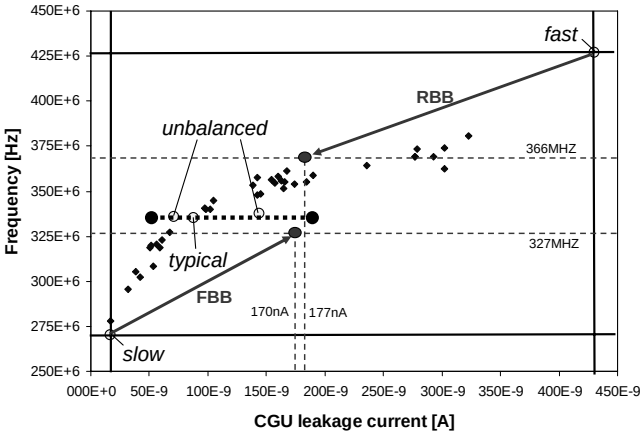

Figure 2.17 Process-dependent performance compensation with ABB.

A second strategy for compensating frequency and leakage spread is based on using ABB and AVS independently. ABB is used to increase the performance of "slow" samples as explained before. AVS is not used in this case because it would require a higher supply voltage than nominal, which may lead to reliability issues for the silicon. Therefore, AVS is only used to reduce the frequency and total power for "fast" samples. This approach is more power-efficient than when using ABB alone because now both dynamic and leakage power are reduced. For a "fast" corner sample, AVS can lower V_{DD} by about 124mV which reduces its switching energy by ~19.6% while still being able to meet the typical frequency specifications. Leakage current reduces less than when using ABB alone; the leakage reduces by ~1.1× (from 430nA to 386nA) for a "fast" corner sample. Consequently, the leakage current of the tuned device is about ~5.44× higher as compared to the "typical" reference.

A third and last strategy consists of setting AVS+ABB jointly. Again, ABB alone is used to increase the performance of "slow" samples. "Fast" samples are biased using AVS+ABB to meet typical frequency specifications while saving power. ABB is used to reduce V_{th} (FBB) such that AVS can reduce V_{DD} more than the case with no FBB, thereby, enabling further overall power savings. Combined AVS+ABB for a "fast" corner sample can lower V_{DD} by about 219mV, which reduces switching energy by about 33.3%. However, this comes at a penalty of increased leakage current. For a "fast" corner sample with 0.4V FBB, the leakage increases by about 3.7× (it becomes 1600nA) as compared to the "fast" corner with no FBB. When comparing against the "typical" reference, the leakage current is about 22.54× higher.

Figure 2.18 puts into perspective the previous results for compensating process-dependent frequency and leakage spread. The values for frequency, power supply voltage, and leakage current are plotted for reference and tuned process corners. The indicated numbers are normalized to the "typical" corner reference. Notice that ABB can effectively reduce frequency and leakage spread, while AVS can trade off higher operating frequency for improved power efficiency. Further total power savings can be achieved with AVS+ABB at the expense of increased leakage.

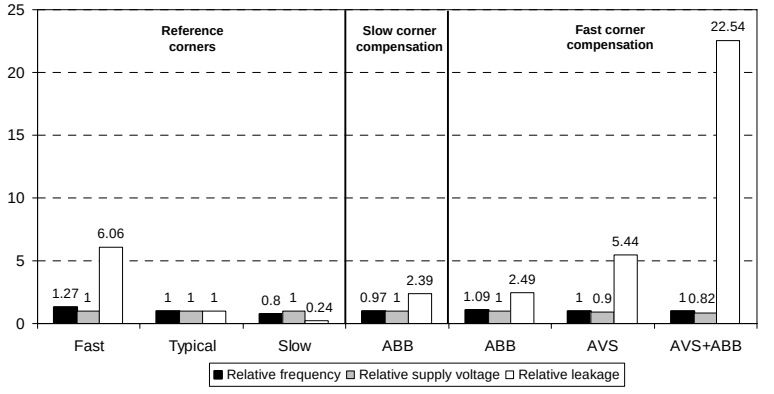

Figure 2.18 Performance compensation in 65nm LP-CMOS.

2.7 Conclusion

The race for low-power devices and the impediments of attaining low power through technology scaling only have opened avenues for design techniques

based on voltage and frequency scaling. We presented measurement results that show the extent to which adaptive voltage scaling and adaptive body bias are useful for power and delay tuning in the state-of-the-art CMOS technologies. We observe the benefits of AVS primarily for low power and of ABB for performance tuning. For instance, for a 65nm LP-CMOS, the state-of-the-art technology power savings are in the order of 82× through 20× frequency downscaling. Contrary to the belief that high V_{th} has a considerable impact on leakage power reduction, we observed that reverse-bias ABB alone reduces leakage only by 2.5× at V_{DD}=1.2V. At lower supply voltage (V_{DD}=0.6V), we observed a larger leakage reduction of 6.8×. However, combined AVS and ABB yield ~25× leakage reduction.

With the increased impact of process variability on circuit design, ABB turns out to be a good design technology to keep parametric yield under control. In particular, we observe the means to tune devices with characteristics in the slow or fast process corners to performance specifications of a typical process corner. While at V_{DD}=1.2V, a ±20% frequency and a ±22% power-tuning range of ABB may look limited, the frequency-tuning range proves to be effective for process-dependent performance compensation. In fact, we observed a continuous frequency tuning despite the wide frequency spread. These tuning indices show that the combined use of AVS and ABB offers significant performance control. Of course, this tuning comes at the price of increased static power consumption. In our results, this static power increase is in the order of 2.4× to meet the required specs.

AVS and ABB design technologies have been reported in the technical literature archival as point solutions, usually through custom-based designs. However, the main impact on circuits-and-systems design will show off only when these techniques are methodologically applied. Along with AVS/ABB design techniques come challenges such as the design of supply and well grids, signal integrity at low voltages, voltage-domain crossing, etc. Fortunately, the electronic design automation (EDA) industry is picking up these concepts. Major EDA companies already offer tools for voltage-domain partitioning, multiple static voltage choices, power gating, and leakage control. Yet the dynamic voltage and frequency-scaling techniques have not been totally automated, partly because these techniques are also application dependent. The use of body biasing is slowly making its way into modern designs, yet automation is lacking behind. It is not unusual to see a wrong perception that ABB is used for leakage control only. We also showed in this chapter that in an era where poor V_{th} to V_{SB} sensitivity is evident, the best benefits of ABB design techniques are on parametric yield, i.e. on performance compensation.

References

[1] W. Haensch, et al., “Silicon CMOS devices beyond Scaling”, IBM Journal of Research and Development, July/September 2006, Vol. 50, No. 4/5, pp. 339–361

[2] D.J. Frank, “Power constrained CMOS scaling limits”, IBM Journal of Research and Development, March/May 2002, Vol. 46, No. 23, pp. 235–244

[3] AMD PowerNOW! Technology, AMD white paper, November 2000, http://www.amd.com

[4] M. Fleishman, “Longrun power management; Dynamic power management for crusoe processor”, Transmeta white paper, January 2001, http://www.transmeta.com

[5] S. Gochman, et al., “The Intel Pentium M processors: Microarchitecture and performance”, Intel Technology Journal, May 2003, Vol. 7, No. 2, pp. 22–36

[6] T. Kuroda, et al., “Variable supply-voltage scheme for low-power high-speed CMOS digital design”, IEEE Journal of Solid-State Circuits, March 1998, Vol. 33, No. 3, pp. 454–462

[7] K. Nowka, et al., “A 32-bit PowerPC system-on-a-chip with support for dynamic voltage scaling and dynamic frequency scaling”, IEEE Journal of Solid-State Circuits, November 2002, Vol. 37, No. 11, pp. 1441–1447

[8] V. Gutnik and A. Chandrakasan, “Embedded power supply for low-power DSP”, IEEE Transactions on Very Large Scale Integration (VLSI) Systems, December 1997, Vol. 5, No. 4, pp.425–435

[9] T. Miyake, et al., “Design methodology of high performance microprocessor using ultra-low threshold voltage CMOS”, Proceedings of IEEE Custom Integrated Circuits Conference, 2001, pp. 275–278

[10] J. Tschanz, J. Kao, S. Narendra, R. Nair, D. Antoniadis, A. Chandrakasan, and Vivek De, “Adaptive body bias for reducing impacts of die-to-die and within-die parameter variations on microprocessor frequency and leakage”, IEEE Solid-State Circuits Conference, February 2002, Vol. 1, pp. 422–478

[11] T. Chen and S. Naffziger, “Comparison of Adaptive Body Bias (ABB) and Adaptive Supply Voltage (ASV) for improving delay and leakage under the presence of process variation”, IEEE Transactions on VLSI Systems, October 2003, Vol. 11, No. 5, pp. 888–899

[12] T. Sakurai and R. Newton, “Alpha-power law MOSFET model and its applications to CMOS inverter delay and other formulas”, IEEE Journal of Solid-State Circuits, April 1990, Vol. 25, No. 2, pp. 584–593

[13] K.Roy, S. Mukhopadhyay, and H. Mahmoodi-Meimand, ”Leakage current mechanisms and leakage reduction techniques in deep-submicrometer CMOS circuits ”, Proceedings of the IEEE, February 2003, Vol. 91, No. 2 pp. 305–327

[14] M. Meijer, F. Pessolano, and J. Pineda de Gyvez, “Technology exploration for adaptive power and frequency scaling in 90nm CMOS”, Proceedings of International Symposium on Low Power Electronic Design, August 2004, pp.14–19

[15] M. Meijer, F. Pessolano, and J. Pineda de Gyvez, “Limits to performance spread tuning using adaptive voltage and body biasing”, Proceedings of International Symposium on Circuits and Systems, May 2005, pp.23–26