Low Power Digital Circuit Design

Takayasu Sakurai
Center for Collaborative Research, and Institute of Industrial Science, University of Tokyo
4-6-1 Komaba, Meguro-ku, Tokyo, Japan, 153-8505
Phone: +81-3-5452-6251, Fax: +81-3-5452-6252, E-mail: tsakurai@iis.u-tokyo.ac.jp

Abstract
The paper describes approaches for achieving low power digital circuits. The approaches are classified from the standpoint of spatial granularity, temporal granularity and variable granularity. The trend is moving from coarse-grain to the finer grain to save more power with the higher engineering cost. The newer approach includes dynamic adaptive control of VDD and VTH at a block level. The paper also touches on low-power applications.

Trend of Power Consumption
The power consumption of processors has been increasing for these 20 years as shown in Fig.1. It can be shown that the increase in power density comes directly from the scaling law and thus inevitable [1,2]. In the figure, the future trend of the power is also added assuming that no countermeasures are taken.

The main message here is that the leakage component of the power consumption will be increasing drastically. This is due to the lower threshold voltage (VTH). It may be argued that the higher supply voltage (VDD) is used to eliminate the lower VTH but then the dynamic power component increases dramatically as was in 1980's when VDD was fixed. The constant voltage scaling is basically against the reliability of MOSFET’s anyway. Two orders of magnitude reduction of power will be necessary in ten years to meet the requirement set by the ITRS (International Technology Roadmap for Semiconductors).

General Approaches for Low Power Design
The power of CMOS digital circuit is expressed by the following expression.

\[
\text{Power} = a \cdot f \cdot C(W) \cdot V_s \cdot V_{dd} + I_s \cdot W \cdot 10^{\frac{V_{th}}{V_s}} \cdot V_{dd} \ (1)
\]

The power and delay dependence on VDD and VTH are shown in Fig. 4 using typical parameters for 65nm generation. It is seen that the region where the power is small is also the region where the delay is large. Thus, trade-off should be taken between power and delay. For a long time, we have been using just one static VDD and VTH for a chip but the situation will be changing for power-aware designs.

General approaches to take the trade-off are shown in Fig.5. Variables which affects the power are VDD, VTH, gate size (W), frequency, temperature, workload, activity, etc. When we change the variables to take the trade-off, the approaches can be classified by three granularities: spatial granularity, temporal granularity and variable granularity. Most of the proposals for low power design are categorized by this classification.

The spatial granularity starts from a chip level, a block level, a gate level, to a transistor level. Temporal granularity, that is a transition time, ranges from infinity, that is, static, milli-second, and down to nano-second order. The variable granularity, that is the number of steps of a variable, can be dual, discrete, or continuous. Usually the granularity starts from coarser and moving toward finer. The finer granularity means more power saving but with more engineering cost.
Static Approaches for Low Power

When the temporal granularity is infinitely large, the approach can be called static, more specifically an optimized static assignment of variables as shown in Fig.6, since the variables do not change in time. Historically speaking, gate sizing is introduced first followed by dual $V_{TH}$ and dual $V_{DD}$. The larger $W$, lower $V_{TH}$ and higher $V_{DD}$ are applied to critical paths while the smaller $W$, higher $V_{TH}$ and lower $V_{DD}$ are applied to non-critical paths.

The spatial granularity of $W$ and dual-$V_{TH}$ is now at a gate level and will be moving toward a transistor level. $V_{DD}$ assignment is now at a block level as is found in some literatures [3,4] but will be more minute level in the future. Other than the general trend from the coarser to the finer in granularity, combined optimization using multiple variables such as sizing and dual-$V_{TH}$ is the new trend.

Dynamic Approaches for Low Power

When variables such as frequency, $V_{DD}$ and $V_{TH}$ are to be changed in time, adaptive approaches are usually adopted, where some target quantity is monitored and some variable are controlled to achieve the goal. Quantity to monitor ranges from leakage current, speed, logical threshold, temperature, workload, activity to reliability. These are monitored using replica circuit or actual circuit. The variable to control includes frequency, $V_{DD}$, and forward/reverse body bias. The method how to control the variable can be either through analog circuit, digital circuit, or software as shown in Fig.7 [5,6].

In the dynamic approaches, one of the popular methods is a variable $V_{DD}$ control. Figure 8 and 9 show $V_{DD}$-hopping where depending on the workload, $V_{DD}$ and frequency are changed to achieve lower power. It has been shown that this kind of control can be effective in reducing the power of real-time multimedia processing system working under multi-tasking O/S [7,8,9]. The number of control steps can be reduced down to two to achieve 2/3 of power reduction. The software-circuit cooperative control has been adopted. Although at present the spatial granularity is a chip level, the block-level $V_{DD}$-hopping can be used in the future.

Since lowering $V_{DD}$ is effective to reduce the gate leakage, the junction leakage and the subthreshold leakage as shown in Fig.10, the $V_{DD}$-hopping is also effective in reducing the leakage power as well as the dynamic power. The subthreshold leakage is reduced because $V_{TH}$ increases in lower $V_{DD}$ region due to the Drain Induced Barrier Lowering (DIBL) effect.

Power Gating

Power gating is another popular method in dynamic approaches for low power. The basic idea is to cut off power line when a circuit is not in use. There are several derivatives in the power gating scheme as are shown in Fig.11 [10].

There are a couple of issues to implement the power gating. One is that the information stored in an F/F will be lost when the switch is cut off and special cares should be taken for the storage elements in a circuit as shown in Fig.12.

One more issue is in the power switch sizing. The switch size is better be small for area consideration but is better be large for circuit speed consideration. Even if we fix the size of the switch, the power gating may degrade circuit speed unpredictably, since the speed degradation is a complex function of circuit activity as shown in Fig.13. A rule of thumb is to make the switch width 10% of the total W of the block. This leads to 5% speed degradation but this is just a rule of thumb. In order to eliminate this difficulty in performance characterization, one way is to insert power-gating switch in selected gates [3]. This is called a selective power gating which is usually used in conjunction with dual-$V_{TH}$ scheme. Here the performance characterization is done at gate level.

Currently clock gating is the most effective way to reduce dynamic power as is shown in Fig.14 but the clock gating loses its effectiveness in the leakage dominant environments. If we would like to substitute clock gating with power gating, one more issue arises, which is a slow wake-up time. Zigzag CMOS scheme as is shown in Fig.15 [11], however, can reduce the wake-up time to within a couple of tenth of a cycle time, whose detailed measured results will be published in the forthcoming CICC [12].

It is to be noted that inserting the power switch is not essential in reducing the leakage. Rather decreasing the power supply voltage to the extreme is
essential to cut off the leakage. If we can incorporate the voltage down function in a power delivery system, power gating can be realized in a more natural way.

**Adaptive Body Bias**

Dynamic $V_{TH}$ approaches are attracting much attention lately. There are two reasons for this. One is the increasing need for compensation against the $V_{TH}$ fluctuation, which is getting severer in low $V_{DD}$ region as shown in Figs. 16 and 17. The other reason is that in achieving lower leakage in the lower performance modes, it is more efficient to change $V_{TH}$ and $V_{DD}$ at the same time rather than changing $V_{DD}$ alone. Changing $V_{TH}$ through a body bias in a standby mode has been adopted by SH and StrongArm microprocessors. The new trend is to control $V_{TH}$ adaptively through the body bias.

Figure 18 shows a historical proposal of an adaptive $V_{TH}$ scheme [13]. In this scheme, monitoring leakage, an analog circuit adjusts the reverse body bias to compensate the as-processed $V_{TH}$ fluctuation. The effectiveness is shown in Fig.19. The newer proposal from Intel on adaptive body bias is shown in Fig.20. In this scheme, monitoring speed of a replica circuit, a digital circuit adjusts the forward body bias to compensate the as-processed $V_{TH}$ fluctuation. The granularity of the control is changed from a chip level to a block level. The chip is divided into 21 sub-sites and separately controlled. Thus in ten years, the granularity gets finer.

The new trend in adaptive body bias is the use of software control. Figures 21 and 22 show such an example, where $V_{TH}$ is controlled by software depending on the required speed. It is interesting to note that changing $V_{TH}$ in time is more efficient than changing $V_{TH}$ in space as shown in Fig.21 [14]. This means that more waste in power does exist in time than in space in multimedia jobs.

In this example, only $V_{TH}$ is controlled to achieve the required performance, but it is more effective to change $V_{DD}$ and $V_{TH}$ at the same time. In order to show this, a simple analysis is carried out. Expression (1) is valid but sometimes it is difficult to find $a$, $C$, $f$, and $I_0$. More tractable expression is derived here. Let’s define $η_L$ which is a leakage power component ratio in the total power at the maximum leakage condition, which occurs at the highest temperature, the highest supply voltage, $V_{DDL}$, and the lowest as-processed $V_{TH}$ at the highest temperature, $V_{THL}$.

\[
P_{\text{TOTAL, } H} = P_{\text{DYNAMIC, } H} + P_{\text{LEAK, } H} = (1 - η_L) P_{\text{TOTAL, } H} + η_L P_{\text{TOTAL, } H} = a f C V_{DD, H}^2 + I_0 10^{-\frac{V_{THL}}{8}} V_{DD, H}.
\]

Then, $a$, $C$, $f$, and $I_0$ can be expressed by using the single quantity, $η_L$.

\[
a f C = (1 - η_L) \frac{1}{V_{DD, H}}
\]

\[
I_0 = η_L P_{\text{TOTAL, } H} / 10^{-\frac{V_{THL}}{8}} V_{DD, H}.
\]

The total power at any $V_{DD}$ and $V_{TH}$, $P_{\text{TOTAL}}$, is expressed as follows using $η_L$ as a design parameter.

\[
P_{\text{TOTAL}} = (1 - η_L) \left( \frac{V_{DD}}{V_{DD, H}} \right)^2 + η_L 10^{-\frac{1}{8} |V_{TH} - V_{THL}| + \delta |(V_{DD, H} - V_{DD})|} \left( \frac{V_{DD}}{V_{DD, H}} \right).
\]

$δ$ is the DIBL coefficient. On the other hand, delay is expressed as follows.

\[
\text{Delay} = k_D \frac{CV_{DD}}{I_D} \propto \frac{V_{DD}}{(V_{DD} - V_{TH})^\alpha} (\alpha = 1.3).
\]

Using these formulas, equi-delay and equi-power curves are plotted on a $V_{DD}$-$V_{TH}$ plane in Fig.23 with solid lines and dotted lines, respectively. There are two solid circles on the plot, both of which show the same delay. It is seen that by changing $V_{DD}$ and $V_{TH}$ simultaneously, it is possible to reduce power consumption more than the case where $V_{TH}$ is fixed.

$V_{TH}$ controllability through the body bias depends on device characteristics including Gate Induced Drain Leakage (GIDL) and body bias coefficient, $γ$, as shown in Fig.24. One may argue that the device performance will be degraded some percentage to improve the GIDL and the $γ$ property but it will be worthwhile to modify the device because the degraded performance at the device level can be compensated by the improvement of performance at the circuit and system level as is shown in Fig.25. Thus the collaboration of device, circuit and EDA disciplines are necessary as is shown in Fig.26.
Future Perspectives for Low-Power Design

The final solution to the leakage power problem is yet to be discovered. However, since the leakage is expressed by the formula shown in Fig.27, a double gate structure [15] will be effective in controlling $V_{TH}$ to compensate the $V_{TH}$ fluctuation, to cut off leakage at standby and to reduce leakage in the lower performance modes.

One important aspect of low-power design is power delivery. IR drop in the power delivery grid as is shown in Fig.28 becomes an issue. Assuming the effective sheet resistance of the grid to be $R$ and the total current consumed in the area to be $I$, the maximum voltage droop is calculated to be $0.074RI$.

In order to mitigate the voltage droop and also mitigate the EMI problems by reducing current, it is beneficial to use higher voltage like 10V to distribute supply voltage over the chip. The high voltage is then transformed to the lower voltages which are used in the blocks of the chip as is shown in Fig. 29. In this context, on-chip voltage converters are of importance [16]. The $V_{DD}$ and $V_{TH}$ of each block on the chip will be changed in ns order. The processors will be parallel in architecture.

System-in-a-Package (SiP) approach will be used to implement electronic systems, since it will reduce the power consumed by I/O’s and also reduce non-recurring engineering cost (NRE) which is getting humongous in recent deep sub-micron VLSI’s.

From Virtual to Physical: Ubiquitous Electronics

Lastly, I would touch on applications of low-power electronics. Recent VLSI’s are mostly used in information technology related products, such as PC’s, communication tools, and digital consumer equipments. It can be said that the current LSI products are targeting mostly to virtual applications. Since most of the important applications become mobile, the low-power design is getting more critical to meet the needs of the market.

The future applications and important technologies are summarized in Fig.30. Wireless sensor network, medical electronics and robots are oriented more toward the physical electronics. The future ubiquitous electronics needs gigantic number of chips as shown in Fig.31 and thus each of the chips should be implemented at the lower power level. Otherwise, the LSI’s eat up the energy of our society.

In the ubiquitous electronics, large-area sensors will become an important component. Since silicon devices are getting smaller and smaller, they are not suitable for direct human interfaces. Integrated circuits based on organic transistors as shown in Fig.32 may give low-power solution in the ubiquitous electronics especially for large-area sensors. Low-power feature is realized by activating a part of a circuit using an organic transistor array [17].

Conclusions

In coping with leakage power increase, both of $V_{DD}$ and $V_{TH}$ are controlled in time as well as in space. Granularity of the control becomes finer as the power problem is getting severer as is shown in Fig.33. In order to fully exploit the effectiveness of low-power circuit design, collaboration with device, EDA and software engineers is necessary. A part of the paper is based on the research supported by STARC.

References


Fig.1 Trend of power consumption

Fig. 2 MOSFET’s go leaky

Fig. 3 Leakage is a real headache

Fig. 4 Power and delay on VDD-VTH plane

Fig. 5 Methods for achieving low-power

Fig. 6 Multiple W, VTH, VDD (static assignment)
Fig. 7 Dynamic control of variables for low power

Fig. 8 Control mechanism in VDD hopping. Application slicing and software feedback are the key to achieve real-time feature.

Fig. 9 VDD hopping achieves more than 70% power saving

Fig. 10 Reducing VDD is effective for leakage reduction

Multiple VDD is not essential. Making effective VDD small is essential to reduce leakage.

Block level power-gating is new.

Fig. 11 Power-gating derivatives

Fig. 12 Multi-Threshold CMOS Circuit

Fig. 13 Power gating MOS width optimization.

Fig. 14 Clock-gating efficiently reduces power, NOW

Conventional cut-off switch several clock cycles to wake-up

K. Min et al, ISSCC 2003, Paper #22.8

Fig. 15 Zigzag cut-off CMOS: Clock-gating substitute in leakage dominant era

Fig. 16 VTH fluctuation is a real headache in low VDD region.
### Fig. 17 Delay fluctuation gets worse in lower $V_{DD}$

- Low $V_{th}$ → large leakage → SSB ON → deepVBB to high $V_{th}$
- High $V_{th}$ → little leakage → SSB OFF → shallow VBB to low $V_{th}$
  - Control $V_{th}$ to adjust leakage current
  - Compensate $V_{th}$ fluctuation

### Fig. 18 Variable Threshold CMOS Scheme

- Leakage, Replica, Reverse body bias, Analog, Chip level

### Fig. 19 Body bias control to minimize fluctuation

- Speed, Replica, Body bias, Digital, Block level

### Fig. 20 Digital control in adaptive body bias

- Leakage power
- Dynamic power
- $V_{th}=0.9V$
- $94\%$ Vs operation

### Fig. 21 Software control of adaptive body bias

- Workload, Actual, Backward body bias, Software, Chip level

### Fig. 22 Schematic of $V_{TH}$-hopping

- Equi-power
- Equi-delay

### Fig. 23 Equi-power and equi-delay curves on $V_{DD}$-$V_{TH}$ plane

- Gate Induced Drain Leakage (GIDL)

### Fig. 24 $V_{TH}$ controllability

- Normalized $I_{GIDL}$ vs $V_{BS}$

### Fig. 25 Performance increase with body-bias control

- $V_{TH}$ process $= \pm 0.1V$
- $V_{TH}$ process $= \pm 0.05V$

**Condition:** keep leakage power at 100°C with lowest $V_{TH} (V_{TH}=0.15V_{DD})$ constant
Fig. 26 Collaboration among levels is needed to cope with leakage power issue

\[
\begin{align*}
\text{LEAK} &\propto 10^{(V_{\text{GS}} - V_{\text{TH}}) / \lambda} \\
V_{\text{GS}} &= V_{\text{TH}} - \delta (V_{\text{GS}} - \lambda V_{\text{GS}}) \\
\text{Negative } V_{\text{GS}} &\text{ Body bias} \\
\text{DIBL} &\text{ Double gate}
\end{align*}
\]

Main gate \( V_{\text{GS}} \)

Source Body Drain

2nd gate \( V_{\text{DS}} \)

Double gate transistor

Fig. 27 Adaptive leakage control for \( V_{\text{TH}} \) variability

\[
4V_{\text{GS}}(V_{\text{GS}} + V_{\text{GS}} V_{\text{GS}} + V_{\text{GS}} V_{\text{GS}}) - R_{\text{IS}}
\]

\( V_{\text{GS}} \) is approximated as

\[
\begin{align*}
\frac{V}{R} &\approx V - R \Rightarrow V = R \left( 1 - \frac{1}{R} \right)
\end{align*}
\]

\[
V_{\text{min}} = \frac{1}{16} \frac{1}{R} x R
\]

\( \sum \text{Sheet resistance} = R \)

Take IR as unity voltage drop

Numerical simulation

IR Drop

Fig. 28 IR Drop

Fig. 29 Leakage-aware LSI depends advanced power delivery

Fig. 30 Ubiquitous electronics and key technologies

Fig. 31 Number of processors per person is increasing.

Artificial skin may add low-power delicate touch to robots

\[
\begin{align*}
\text{Variables} &\quad V_{\text{DD}}, V_{\text{TH}}, W, \text{ (temp., workload, …)} \\
\text{Spatial granularity} &\quad \text{Chip-level, block, gate, transistor} \\
\text{Temporal granularity (transition time)} &\quad \text{Infinity (static), milli-second, nano-second} \\
\text{Variable granularity (# of steps)} &\quad \text{Dual, discrete, continuous} \\
\text{Mode of control} &\quad \text{Analog, digital, software}
\end{align*}
\]

Fig. 32 Future directions of low-power designs