CORE
Metadata, citation and similar papers at core.ac.uk
Provided by Online Research Database In Technology

Techniques for Leakage Power Reduction in Nanoscale Circuits: A Survey[1]

Wei Liu
Department of Informatics and Mathematical Modeling
Technical University of Denmark

IMM Technical Report 2007 n. 4

Abstract

This report surveys progress in the field of designing low power especially low leakage CMOS circuits in deep submicron era. The leakage mechanism and various recently proposed run time leakage reduction techniques are presented. Two designs from Cadence and Sony respectively, which can represent current industrial application of these techniques, are also illustrated.

[1] The work behind this report is carried out as a 13-week special course with associate professor Alberto Nannarelli as supervisor during the autumn semester 2006 at the Department of Informatics and Mathematical Modeling, Technical University of Denmark.

I. Introduction

As technology scales down, the size of transistors has been shrinking. The number of transistors on chip has thus increased to improve the performance of circuits. The supply voltage, being one of the critical parameters, has also been reduced accordingly in order to maintain the characteristics of an MOS device. Therefore in order to maintain the transistor switching speed, the threshold voltage is also scaled down at the same rate as the supply voltage. As a result, leakage currents increase dramatically with each technology generation. Some researchers predict a 7.5-fold increase in the leakage current and a 5-fold increase in total energy dissipation for every new microprocessor chip generation. As the leakage current increases faster, it will become more and more proportional to the total power dissipation. Designers need to develop new low power techniques to reduce total leakage in nanoscale circuits, especially for chips that are used in power-constrained portable systems.

II. Leakage Mechanisms

Different leakage mechanisms contribute to the total leakage in a device. As Fig 1 shows, the three major types of leakage mechanisms are [1]

- Subthreshold
- Gate, and
- Reverse-biased, drain- and source-substrate junction band-to-band-tunneling (BTBT)

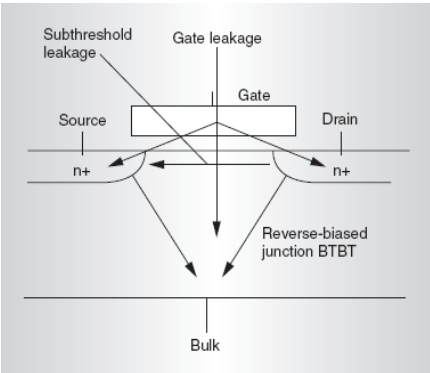

Fig 1 Major leakage components in a transistor

Currently, subthreshold leakage is still playing the main part in the three mechanisms. However, researchers believe that gate leakage and reverse-biased junction BTBT will be as important as subthreshold from 45 nm process downwards. In addition, with technology scaling, the gate oxide thickness will be reduced and the substrate doping densities will be increased. As a result other factors such as gate-induced drain leakage (GIDL) and drain-induced barrier lowering (DIBL) will also become more and more evident. Therefore, future effective low leakage design will need to target at several components since all of them play an important role in the total leakage consumption. Various techniques at process and circuit level exist to reduce leakage consumption, including modifying doping profile, oxide thickness and channel length. Forward or inverse body biasing is also one of them, which is a technique resulting in variable threshold CMOS. In the rest part of this report, we'll mainly focus on run time low leakage techniques.

III. Run time techniques

1. Low Power Bus

With technology scaling, designers are able to put more and more transistors onto the chip. In fact, the number of transistors grows so fast that much of the metal on chip are used to carry communication signals from one logic block to another. These wires consume a lot of power both leakage and dynamic. Thus minimizing wire leakage or designing low power bus will become an important research area in deep submicron technologies.

Busses usually contain a lot of buffers to maintain the signal strength at the receiving end. Using low leakage cells for these buffers could be an option at the first glance. But it is not always the case. Low leakage cells consume less leakage power comparing to high-speed cells but at the same time they are also slower. Thus for long wires to meet timing requirements more cells are needed. And for short wires to maintain signal driving strength, strong cells are required. From the simulation results of [8], it is shown that the use of LL cells can incur significant delay and dynamic energy. Since dynamic power dissipation still dominates, thus the total power consumption also increases. As a result, using LL cells in wire buffers becomes a poor design choice.

Another observation is leakage consumption is input dependent in CMOS gates. Thus for inverters on busses we can interleave large and small cells such that for half of the circumstances the large cells will be in its less leaky state. And by coding the signals transmitted over the bus, we can always guarantee that the bus will be in a low power state. The proposed method in [3] goes even further. The authors modified the threshold voltages of buffers to achieve both minimum leakage and delay as shown in Fig 2. As a result, the conducting paths in these buffers will always go through the low threshold voltage transistors to maintain speed and the non-conducting transistors have high threshold voltages to reduce leakage.

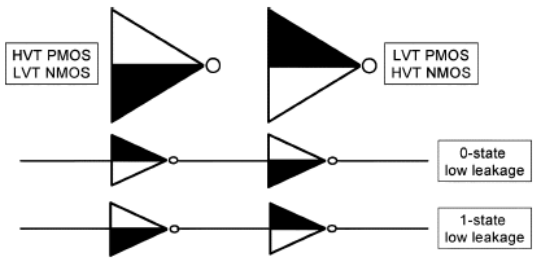

Fig 2 SVT buffers

It is easy to see that the design requires that signals on a wire should only be the values desired thus leakage aware encoding circuits are needed. The authors proposed an encoding algorithm that can generate the low leakage patterns and minimize crosstalk between adjacent wires at the same time. The encoding circuitry is simple enough not to compromise the total power gained.

Other microarchitectural methods also exist. For example, rearrange circuit topologies so that communicating parties are placed near to each other and local communications take place on local busses. Only signals sent to a far away destination are placed on the long bus, thus reducing the activities on the long bus to save power.

2. Input Vector Control

Many research have been done to model and estimate the nominal leakage current of a circuit. Due to the stacking effect, the leakage of a circuit depends on its input combinations [6]. The following table shows the leakage components for all input combinations of a 3-input NAND gate.

Input State	Subthreshold leakage (nA)	Gate leakage (nA)	Total Leakage (nA)
000	0.49	6.58	7.07
001	0.81	19.68	21.49
010	0.81	6.79	7.60
011	2.68	34.78	37.46
100	0.81	3.15	3.96
101	2.68	16.8	19.48
110	2.68	1.84	4.52
111	16.85	45.3	62.15

Table 1. Leakage current values for different input combinations of a 3-input NAND gate

Thus by finding a minimum leakage vector (MLV) and applying it to the circuit, we can guarantee the circuit turns into its low leakage state. In [6], the authors proposed a technique to identify the MLV and discussed several ways to apply the vector to the circuit. They first construct a Boolean network to compute the total leakage and then they use the SAT solver to find an input vector that results in the minimum leakage of the whole circuit. However, the effectiveness of the method does not rely on finding the MLV solely. Due to the limited access to internal nodes, it is very difficult to put the ideal value to every node especially for very complex circuits. Therefore the authors tried two ways to increase controllability of the internal nodes, namely adding multiplexers and modifying gates. Since both methods change the circuit to some extent, new Boolean networks need to be constructed. Although the authors found an efficient way to identify the MLV and figured out ways to increase controllability, we could see that the actual application of input vector control is still limited by the primary inputs. And to put a sleep circuit back to normal will require extra memory elements to store the original states, thus incurs both area and delay penalty. Another controllability increasing method could be found in [7].

3. Caches

Besides buses and datapath logic, there are also a lot of attentions paid to on chip caches. In microprocessors, large amounts of chip area are devoted to memory structures typically ranging from 30% to 60%. Since memory cells are relatively simple, they lack the inherent stacking effect and exhibit a larger leakage consumption comparing to datapath logic. It is predicted that in 70 nm processes, leakage will take up more than 60% of the total power consumption in level 1 caches. Therefore, a lot of research targeting at leakage reduction in caches could be found in literature. However, there are several differences from datapath logic that must be taken into account for

caches. One thing is that the preservation of cache states during standby mode is often desirable, which means it would be good if data stored in caches were not destroyed so that we won't need to access secondary memories on recovery. The other thing is that memory access time should not be greatly degraded, which means recovery time should be as small as possible otherwise it will severely compromise the system performance.

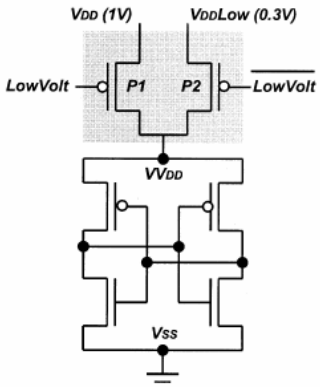

Fig 3 Supply voltage control mechanism

Two most widely cited methods are decay cache and drowsy cache. Decay cache utilizes the gated Vdd technique by using a high Vt transistor between virtual ground and the ground to put least recently accessed cache lines into power off state. This can dramatically reduce the power consumption but an obvious disadvantage is information stored in power off cells is totally lost. Drowsy cache as described in [10] provides a better solution. It also uses transistors to separate virtual Vdd from Vdd supply line but still supplies a very low voltage to the cell when it is turned into low power mode. The cell implementation is shown in Fig 3. According to the authors, the wake up latency is only a few clock cycles and thus does not have a major impact on the system performance. For data caches, all cache lines except the active one are put into drowsy mode every n clock cycles. n depicts the window size of how often should the cache be put into drowsy mode and they found 4000 is an adequate number for the benchmark they run on. Since programs typically only access a small portion of data after initiated, the drowsy cache method could gain a significant reduction in leakage consumption in the long run. For instruction caches, the situation is slightly different due to the instruction access characteristics. Therefore putting all cache lines into drowsy mode every n cycles does not work well for instruction caches. However, the spatial locality property can still be utilized. The authors of [10] proposed a low leakage instruction cache architecture based on the subbank method. The basic idea of subbank is to divide the cache into several subbanks and turn those inactive subbanks into low power mode. The proposed architecture extends the subbank method and adds Next Subbank Prediction Techniques to it. A prediction buffer keeps track of predicted subbank index and other information for the instruction fetched one cycle earlier. Thus if the instruction (e.g. a jump instruction) is going to access a subbank in drowsy mode, that subbank could be waken up one cycle earlier and thus enhance the performance. There are also other techniques for instruction caches such as the one described in [11]. The authors perceived that programs especially multimedia applications tend to spend most of their time in loops and execute only a sequence of instructions for most of their computations. Based on the observation, they propose a novel cache replacement policy for instruction caches, which forces instructions in a loop be placed in the same subbank and are not the first candidates

to be replaced into secondary memories when misses occur. In such a way, only one subbank will stay active and other subbanks can stay in the drowsy mode most of the time.

4. Dynamic Voltage Scaling in Dual Vt Cell Designs

It may be expected that dynamic voltage scaling will always reduce dynamic power dissipation in a long period of operating time since when the workload is reduced the supply voltage could also be reduced to save power. However, in deep submicron technologies, we need begin to take leakage power consumptions into account as well, especially when design with dual threshold voltage cells are becoming widely adopted. In [8], the author gave a comprehensive analysis of the consequences led to by applying Dynamic Voltage Scaling (DVS) to dual Vt cells design. A typical design scenario could be as follows. In the initial design, all cells used are low leakage (LL) cells to minimize power consumption. Then we could replace cells along critical paths by high-speed (HS) cells in order to shorten the delay. Thus we get a mix cells design. And now DVS could be used to further decrease dynamic power dissipation when a larger delay could be tolerated. When applying DVS, we only get power gains if total power consumption in the mixed cell design is less than that in the original single LL cell design. According to the author, subthreshold leakage is supply voltage dependent. Thereby we can establish mathematical equations modeling power consumptions in the two designs to verify that we did actually save power. From one example shown in [8], the results show that after applying DVS the dynamic power dissipation is indeed decreased however the leakage consumption is increased at the same time. It tells us that DVS could have a negative impact on leakage consumption under certain circumstances and thus careful analysis needs to be done before making design choices.

Another factor that may affect the power consumption characteristic is the behavior of synthesis tools. The threshold voltages of LL cells are higher than that of HS cells. Thus when reducing supply voltage, the delay on LL cells increases faster than HS cells. At this time, the synthesis tool may try to modify the design in some parts in order to make the design meet its timing constraints. And as a result, additional cells may be inserted and total power consumption increases.

5. Others

In [12], the authors propose a new approach to the joint optimization of leakage and delay using mathematical optimization methods. Conventional optimization problems are formulated as minimization or maximization of a single attribute objective (area, delay or power) with constraints on the remaining attributes [12]. The proposed method, based on the *utility theory*, takes into account the tradeoffs between leakage and delay and uses the energy-delay product (EDP) to express the relation as the optimization's objective function. The circuit is modeled as an acyclic directed graph in which nodes represent gates and edges represent connections between gates. The leakage and the delay models are expressed as functions of gate sizes. Designer's preferences on other attributes (e.g. delay and other physical constraints) can also be expressed in the form of *marginal disutility functions*. The whole process leads to a *convex function* of which the decision variables are sizes of gates. Thus the minimization problem is formulated as a gate-sizing problem. However, as we have discussed in previous sections, the leakage

characteristic in a transistor is determined by several parameters. Therefore the effectiveness of the proposed method is quite limited and cannot be used a general tool to guide designers to make optimal choices. But it could still be used as one step in a series of leakage reduction processes where optimization is dependent on a particular parameter like size of gates.

IV. Examples in industrial design

Most of the techniques discussed above are developed or proposed in academic research. Very few of them can be seen in industrial reports. This could be due to the fact that they are relatively new so that no mature CAD tools are supporting the automatic application of them yet. But another reason could be the effectiveness and applicability of these new techniques still need to be proved.

In this section, two industrial designs, one from Cadence Design Systems and the other from Sony Corporation, aimed at reducing both dynamic and leakage power consumptions are presented. The techniques involved are dual threshold voltage cell libraries, multiple supply voltage domains, dynamic voltage and frequency scaling. These techniques have been around for a few years and are more matured. However the application of these techniques are not trivial. The example from Cadence focuses on presenting the capabilities of their newly developed techniques in synthesis tools that can automatically optimize power consumptions using dual Vt libraries and multiple supply voltage domains. The example from Sony shows their design of a novel Dynamic Voltage and Frequency Management (DVFM) scheme with leakage compensation.

1. Cadence's 90-nm power optimization methodology

The work focused on the development of a general purpose power optimization methodology for synthesis-based digital designs [13]. The methodology which enables leakage current and clock rate optimization in a multiple supply voltage environment, was developed and applied to reduce power dissipation in a 355-MHz IC containing an ARM1136 JF-S microprocessor, which was realized in a 90-nm CMOS process. Leakage power has been reduced by 46% and the overall power dissipation has been reduced by about 40% in their design comparing to the single Vdd normal Vt version.

The methodology is composed of a suite of newly developed synthesis, routing and floor planning techniques. The general synthesis approach is based on creating the netlist with high Vt cells first to achieve timing constraints, utilizing low Vt cells only when necessary for critical paths [13]. Since leakage current of low Vt transistors can be six-fold greater than that of high Vt transistors, using high Vt cells for non-critical paths can significantly reduce leakage consumption. Other techniques aimed at both dynamic and leakage power are also applied in the synthesis algorithm, including buffer removal, logic resizing, pin swapping, buffered slow transition and logic restructuring.

One important aspect of the methodology is the selection of Vdd domains (0.8V and 1.0V) and the automatic placement of Voltage Level Shifting (VLS) cells. The on chip SRAM blocks as well as the main memory access paths are designed to operate in the 1.0-V Vdd domain to achieve

optimal performance. Logic cones are operated in one consistent Vdd domain. VLS cells function as interface between different voltage domains to isolate and translate signals going through. These VLS cells, conventionally inserted manually, can be automatically inserted, placed and routed in the new methodology. The algorithm will optimize the placement in such a way that continuous VLS cell abutment at the Vdd domains' share perimeter (Fig 4). As a result, these techniques eliminate the conventional constraint limiting the use of multiple Vdds and the synthesis tools can now do a better overall optimization.

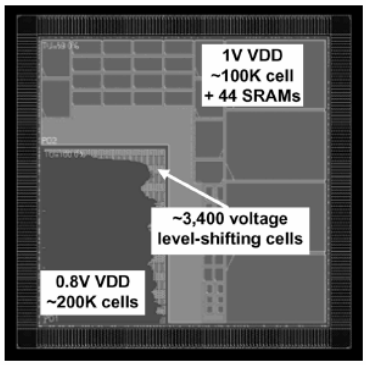

Fig 4 Vdd domains and functional usage

The work also includes timing and area optimizations, which are not described here. To summarize, the methodology incorporates techniques like dual Vt and multiple Vdd domains into various synthesis and optimization algorithms so that CAD tools can directly support them. The correctness and effectiveness of the methodology has been proved by verification results.

2. Sony's Dynamic Voltage and Frequency Management (DVFM) scheme

While the former example focuses on implementing algorithmic level support for dual Vt and multiple Vdd domains in synthesis tools, this example presents a fully manually designed scheme utilizing the advantage of dynamically adjust supply voltage and operating frequency at run time.

The microprocessor is designed for embedded systems and contains an ARM CPU, four 16-Mb DRAM macros, a 2-D graphics engine and a DSP core for audio and video processing purposes. The processor is realized in 0.18-μm processes. The DRAMs and PLL blocks are fixed to 1.6 V supply lines and the supply voltage of all other logic blocks is controlled by the DVFM. In this DVFM approach (Fig 5), clock frequency is autonomously and dynamically controlled from 8 to 123 MHz in steps of 0.5 MHz while voltage is adaptively controlled from 0.9 to 1.6 V in steps of 5 mV at the same time. A delay synthesizer in the Dynamic Voltage Control (DVC) circuit emulates and provides the circuit delay information while the Dynamic Frequency Control (DFC) circuit determines optimal operating frequency. [14]

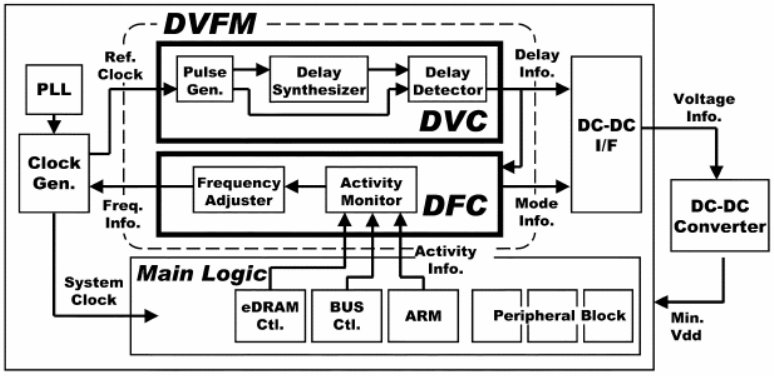

Fig 5 DVFM block diagram

The pulse generator in the DVC generates a detect pulse signal and a reference clock signal. The detect pulse signal goes through the delay synthesizer and is captured by the delay detector. The delay detector will determine whether to increase, decrease or maintain the supply voltage. The delay synthesizer is the most important component in DVC. It consists of three programmable delay components, gate delay, RC delay, and a rise/fall delay component [14]. It is configured to emulate the delay characteristic of the circuit and thus is application specific. Another novel aspect of the DVC is that it can detect the fluctuation of circuit delay due to process deviations and accordingly adjust the supply voltage to compensate leakage power. The DFC is composed of an activity monitor and a frequency adjuster. It gathers activity information from the CPU, bus and embedded DRAM and calculates the optimum operating frequency.

The design is reported to be able to reduce 82% of the power consumption when running Personal Information Management applications and 40% of power consumption when playing MPEG4 movies.

V. Conclusion

In this survey report, various run time and architectural techniques that are proposed in recent years aimed at reducing leakage consumption have been presented. Two designs from industry are also illustrated briefly. We could see that in deep submicron processes leakage consumption is playing a more and more important role in the total power consumption and more and more research both from academic and industry are being conducted in this area. Some methods like the staggered threshold voltage (SVT) buffers proposed in the leakage aware bus-encoding scheme are effective but requires modifications to the gate design and thus are only applicable in full custom design before standard cells are available. Methods like input vector control are limited by the controllability of internal nodes and thus may be difficult for complex circuits. Dynamic voltage scaling in dual threshold voltage cell designs does not result in reduction in leakage in all circumstances. The effectiveness of these techniques will depend on specific application and they themselves still need to be matured. But reducing leakage consumption is becoming more and more important. So we can predict that there will be more research in these and other techniques and they will need to be reevaluated in each process generation since the changing leakage mechanisms will dictate the effectiveness of these techniques.

References

[1] Amit Agarwal et al "Leakage Power Analysis and Reduction for Nanoscale Circuits" IEEE Computer Society, March-April 2004

[2] Yuh-Fang Tsai et al "Characterization and Modeling of Run-Time Techniques for Leakage Power Reduction" IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol.12, No.11, November 2004

[3] Rajeev R.Rao et al "Bus Encoding for Total Power Reduction Using a Leakage-Aware Buffer Configuration" IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol.13, No.12, December 2005

[4] Avnish R.Brahmbhatt et al "Low-Power Bus Encoding Using an Adaptive Hybrid Algorithm" Design Automation Conference, 2006 43rd ACM/IEEE

[5] Avnish R.Brahmbhatt et al "Adaptive Low-Power Bus Encoding Based on Weighted Code Mapping" IEEE International Symposium on Circuits and Systems, 2006

[6] Afshin Abdollahi et al "Leakage Current Reduction in CMOS VLSI Circuits by Input Vector Control" IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol.12, No.2, February 2004

[7] H.Rahman et al "An efficient Control Point Insertion Technique for Leakage Reduction of Scaled CMOS Circuits" IEEE Transactions on Circuits and Systems-II: Express Briefs, Vol.52, No.8, August 2005

[8] Martin Hans "Architectural Aspects of Design for Low Static Power Consumption" Master Thesis, IMM, DTU, 2004

[9] Michael Kristensen "Incorporating Leakage Current Considerations in Logic Synthesis" Master Thesis, IMM, DTU, 2004

[10] Nam Sung Kim et al "Circuit and Microarchitectural Techniques for Reducing Cache Leakage Power" IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol.12, No.2, February 2004

[11] Praveen Kalla et al "Distance-Based Recent Use (DRU): An Enhancement to Instruction Cache Replacement Policies for Transition Energy Reduction" IEEE Transactions on Very Large Scale Integration (VLSI) Systems, Vol.14, No.1, January 2006

[12] Sarvesh Bhardwaj et al "Formalizing Designer's Preferences for Multiattribute Optimization with Application to Leakage-Delay Tradeoffs" International Conference on Computer Aided Design, November 2005

[13] Aurang Khan et al "A 90-nm Power Optimization Methodology With Application to the ARM 1136JF-S Microprocessor" IEEE Journal of Solid-State Circuits, Vol.41, No.8, August 2006

[14] Masakatsu Nakai et al "Dynamic Voltage and Frequency Management for a Low-Power Embedded Microprocessor" IEEE Journal of Solid-State Circuits, Vol.40, No.1, January 2005