New techniques for low power caches

Massimo Alioto, Sandro Bartolini, Paolo Bennati, Roberto Giorgi

Dept. Ingegneria dell'Informazione, University of Siena, Via Roma 56, 53100 Siena, Italy

ABSTRACT

Power consumption is one of the major concerns in modern embedded computing systems. On-chip caches represent a sizeable fractions of the total power consumption of microprocessors. Its reduction is becoming fundamental to develop both low-power and high-performance systems. Although large caches can improve performance, they equally increase the power consumption. The operation frequency and the transistor size are other important factors of the power consumption.

Basically, cache power consumption is mainly due to two factors: dynamic switching power (charging and discharging capacitors) and static power (short-circuit currents). Static power is increasing in importance in newer CMOS technologies (like e.g. 0.65 μm technology) and it is surpassing dynamic power. Recently, many studies describe new techniques for the reduction of both static and dynamic power consumption. A new proposal for the reuse of the charge potentially lost during the discharging of unused cells is presented. Two ideas are illustrated. The first uses the residual charge from cell put into drowsy-mode to charge the neighbors cell. The second and less complex idea is based on the use of the residual charge to drive the nearest cell drowsy bit, through a adequate network and circuitry.

KEYWORDS: poster session; cache; low power; drowsy, gated-vdd; charge reuse drowsy cache

1 Introduction

In modern microprocessor the power consumption of the cache memories represent a large fraction of the total power consumption; recent studies have demonstrated that it accounts for about 50% of the total power consumed in embedded computing systems [1]. Unfortunately high-performance and low-power are usually in contrast. In addition the distance between the characteristic of batteries and the demand of power will be always greater, because while batteries' performance grows 10%-15% a year [2], the power consumption grows much faster.

The fraction of the total power consumption due to on-chip caches is not negligible. Its reduction is fundamental for new low-power high-performance systems. The causes of power consumption in cache memories are mainly due to two factors: *dynamic power* and *static power*. Dynamic switching power is due to the charging and discharging of parasitic and/or input capacitors. Until very recently the scientific community interest focused only on the fraction of total power consumption due to dynamic power. Indeed static power is increasing in importance in newer CMOS technologies and it currently accounts for about 15%-20% of the total power on chips implemented in high-speed processes; with 0.65 μm

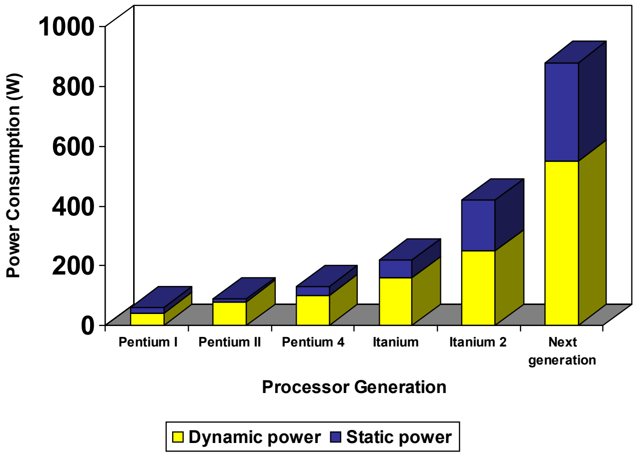


Figure 1: Dynamic and static power trend. The fraction of total power due by static power will surpass the fraction of dynamic power with 0.65 μm technology.

technology, static power will surpass dynamic power. The increase of static power rate is illustrated in Figure 1.

Since the majority of leakage comes from the largest processor components and the cache memories account a large percentage of the overall microprocessor die area, most of recent researches have developed techniques for their power consumption reduction.

In [3] the main result is that the fastest implementation is not always the most beneficial from the memory energy stand-point. A large decrease of leakage in caches can be achieved by a reduction of the power supplied. In literature the recent proposals can be broken down into two main categories: *state preserving* and *state non-preserving*. The difference between them depends on if the cache lines disabled lose data (the first) or if they maintain the information stored (the second).

The main methods of those two categories are gated-VDD [4] for the first and drowsy caches [5] for the second. The gated-VDD technique introduces an extra transistor to gate the supply of the cache SRAM cells. A dramatically reduction of the leakage current is achieved because the lines are powered off, but data stored in those lines are lost. An access to a switched off line involve a restore of the data from the low level of memory, with a consequent loss in performance and dynamic power dissipation increase. Instead, drowsy caches decrease leakage by reducing the power supply but not gating all lines: that avoids losing information. Lines supplied with the reduced voltage are called drowsy lines, and when there is an access on them they only need to be awakened prior to accessing the data (figure 2). The main advantage consist on no additional access to lower memory level during an access into a drowsy line; nevertheless, since the supply is only reduced but not eliminated, the leakage reduction is smaller than in Gated-VDD. Many others techniques are presented for power consumption reduction in cache memories such as [6], [7]and [8].

We propose a technique to reduce the wake-up latency and save some of the dynamic power too.

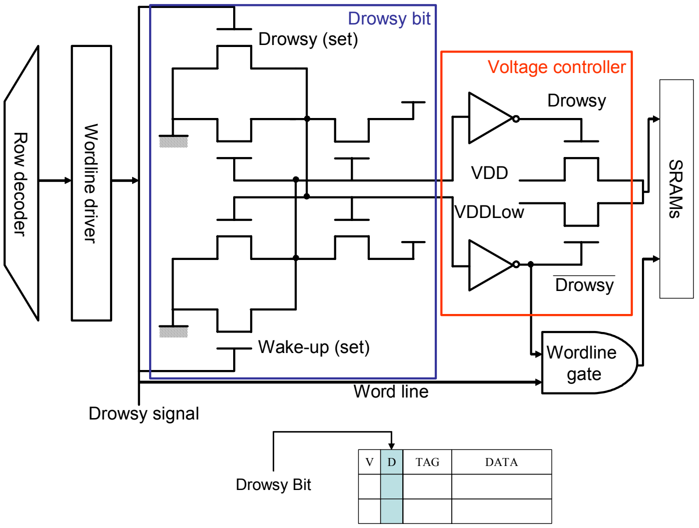
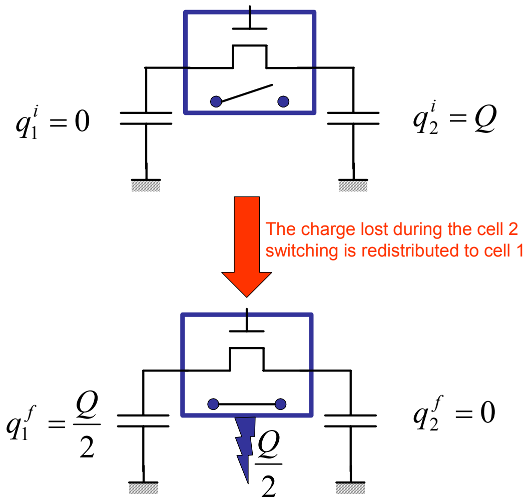


Figure 2: Drowsy cache implementation. Some addition are required.

2 Charge reuse drowsy cache

In the drowsy cache model [5] each time a memory cell is switched into a low-power state, a large fraction of the charge is lost. During the transition between high-state to low-state about half of total charge in the cell is dissipated and the other half is lost to ground. During the execution of a program there are many transition between high and low power states, and it increases with the use of more sophisticated policies. This factor limits the total power consumption reduction, because it increases the dynamic power.



The charge lost during the cell 2 switching is redistributed to cell 1

Figure 3: Hypothetical perfect redistribution scheme. Half of the total power of cell 2 goes to cell 1. In the case of drowsy caches, the saving charge will be a fraction of that.

Although it is impossible to reuse all the charge of a cell to charge another cell, the reuse of the fraction lost to ground is reasonable.

Let us consider the situation of figure 3. We have two cells connected through a transistor operating as a switch. At the beginning the cell 1 is OFF ($q_1^i = 0$), the cell 2 is ON ($q_2^i = Q$) and the switch-transistor is closed. Suppose that we have to switch-off the cell 2. If during its discharging the switch is closed, half of the initial charge of cell 2 is redistributed into the cell 1. The final situation is that the cell 1, that probably will be accessed soon, is just pre-charged with half of the total charge ($q_1^f = \frac{Q}{2}$) without any additional power and performance cost. To have the cell 1 ON, we have to provide only half of the total necessary charge.

A simpler and less costly technique could be to use the residual charge to drive the drowsy bit of the cache. We can assume an additional network with its circuitry connecting the drowsy bit controller and the word-line of two neighbours cells (figure 4). Imagine that at the beginning, all the cells are in drowsy. Let us suppose to access to cell 1. We have to wake-up the cell and wait the fully availability of the data a number of cycles such as the latency of common drowsy cache (e. g. 2 cycles). After some cycles, the policy could say to put the cell 1 into drowsy mode. During this operation there a lost of charge. In agreement with the programs' locality, probably the next cell will be accessed very soon.

The key idea is to exploit the residual reusable charge loosing to ground during the discharge of the cell 1 to set-up the drowsy bit for the cell 2. Thus, when it is accessed, there will be an hit with a smaller latency (e.g. 1 cycle).

3 Conclusion and future works

The proposals presented are based on two different way to reuse the charge lost during the decay of a wake-up cell into the drowsy state. The key idea is derived by the observation that the charge lost during the cells discharging is not negligible. Two ideas are presented: the first and more ambitious is to reuse the saved charge to switch-on the neighbours cell, the second is to use the charge to drive the nearest cell drowsy bit (or with consideration of some locality function). Those techniques shouldn't involve a big overhead, they should increase the performance of the cache memories with a parallel reduction of the power consumption, being particularly powerful for embedded system designs.

Those ideas will have to be validated from a circuital point of view and, after constructing a strong model, standard benchmarks will have to show the net power savings.

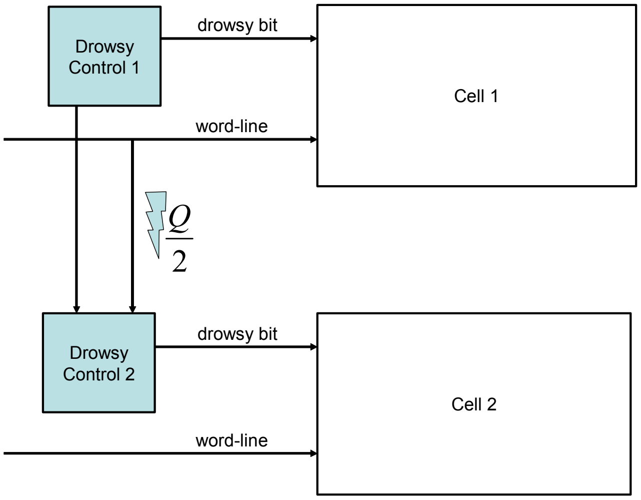


Figure 4: Possible scheme for the activation of the drowsy bit of the next cell through the reuse of residual charge.

References

[1] Malik, B. Moyer and D. Cermak, "A Low Power Unified Cache Architecture Providing Power and Performance Flexibility," Int. Symp. on Low Power Electronics and Design, June 2000.

[2] M. J. Irwin and V. Narayanan, "Low Power Design: From Soup to Nuts", ISCA Tutorial: Low Power Design 2000

[3] L. Benini, A. Macii, E. Macii and M. Poncino, M., "Analysis of energy dissipation in the memory hierarchy of embedded systems: a case study", in Proc. MELECON 2000, 2000, pp. 236 - 239 vol.1

[4] M. Powell, S. H. Yang, B. Falsafi, K. Roy, and T. N. Vijaykumar, "Gated-Vdd: A Circuit Technique to Reduce Leakage in Deep-Submicron Cache Memories", in ISPLED '00, 2000, pp. 90 – 95

[5] K. Flautner, N. S. Kim, S. Martin, D. Blaauw and T. Mudge, "Drowsy caches: simple techniques for reducing leakage power", in Proc. ISCA '02, pp. 148 – 157

[6] M. J. Geiger, S. A. McKee and G. S. Tyson, "Drowsy Region-Based Caches: Minimizing Both Dynamic and Static Power Dissipation", *CF '05*, Ischia, Italy, 2005

[7] N. Mohyuddin, R. Bhatti and M. Dubois, "Controlling Leakage Power with the Replacement Policy in Slumberous Caches" , *CF '05*, Ischia, Italy, 2005

[8] S. Petit, J. Sahuquillo, J. M. Such and D. Kaeli, "Exploiting Temporal Locality in Drowsy Cache Policies", presented at CF '05, Ischia, Italy, 2005