Leakage Power Proliferation in Short Channel Cache Memories

N. Mohamed and N. Botros
Department of Electrical and Computer Engineering
Southern Illinois University, Carbondale, IL 62901-6603

***Abstract*-This work investigates the escalation of leakage power loss in today's low power processors as their feature size shrinks rapidly. The problem seems to exacerbate as technology scaling dives steadily into Very Deep Submicron (VDSM). A quantitatively analysis has been carried out in several cache systems and has shown that the 1-way set associativity optimizes this power component across various generations.**

I. INTRODUCTION

Up till recently, the leakage dissipation in most microprocessors-the one that used to be negligible- has increasingly become a dominant factor. With well above 40% of the StrongARM power being lost in cache system [1], cache dissipation is widely considered to be reflexive of microprocessor power loss. This work focuses on the leakage component of that loss. StrongARM, a commercially well-known processor, has been selected to represent a class of low power embedded microprocessors. It has been classified into several variants with each meant to represent an alien of a scaled down generation. For each variant, the impact of the leakage current contribution to cache total power loss has been studied quantitatively. The effect of cache organizations in each generation has also been scrutinized. The resulting performance degradation associated with that has been investigated as well. The rest of this paper is organized as follows: section II presents the basic building block of standard SRAM and the models that are used in simulation. Section III studies the impact of the reduced cache feature size on leakage current dissipation. Here, we show that cache organization can be used to boost the associated performance loss. In section IV, we present our experimental results and conclude in section V.

II. LEAKAGE POWER MODELS

Ideally, SRAM arrays dissipate no static power. Yet, as these arrays start to downsize to meet the growing demand for low-power processors, they have increasingly become leaky. The main reason for that is the reduction in CMOS inverter threshold voltages- known as threshold roll off- that is associated with the supply voltage reduction. This threshold roll off marks a trade off between power and speed at the device level of abstraction. For while reducing device supply voltage does drastically reduce its power dissipation, as well as its speed, the latter effect is normally offset by cutting on threshold voltage [2]. Hence the device becomes more vulnerable to standby current when it is presumed turned-off. This emerging current component does not seem to show any sign of abating as technology dives steadily into VDSM. In contrary, it has become the dominant factor [2]. The major component of leakage current is known as subthreshold current; I_s . It flows from drain to source due to the diffusion of minority carries when the NMOS device operates in the weak inversion region as shown in figure 1. This is in contrast to the drift current which dominates when the device operates in the strong inversion region.

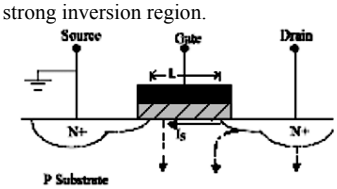

Figure 1: Leakage current components in NMOS transistor

The magnitude of I_s is given by [4]:

$$I_{sn} = (W/L)\,\mu\, V_{th}^2\, C_{sth} \exp[(V_{gs} - V_T + \eta V_{ds})/n V_t][(1-\exp(V_{ds}/V_{th}))] \quad (1)$$

Where W and L are the device width and length, μ is the carrier mobility, $V_{th} = kT/q$ is the thermal voltage at temperature T, C_{sth} summation of depletion region capacitance and interface trap capacitance per unit area of the gate, η is the drain-induced barrier lowering(DIBL) coefficient and n is the subthreshold slope shape factor given by:

$n = 1 + C_{sth} / C_{ox}$ where C_{ox} is the gate input capacitance per unit area of the gate. The same relationships apply for the PMOS transistors as well.

For SRAM-based caches that are built using such transistors, like that of figure 2, a model is required to calculate the cell total

T. Sobh (ed.), Innovations and Advanced Techniques in Computer and Information Sciences and Engineering, 349–352.

leakage. In terms of the Data Retention Voltage (DRV), Qin el al. [5] provides an accurate relationship where the overall leakage current I_s is calculated as:

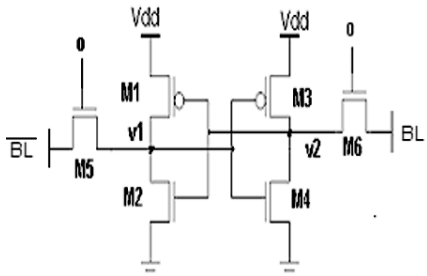

Figure 2: Six-transistor CMOS SRAM cell

$$I_s = I_{s2} + I_{s4} \quad (2)$$

$= (W/L)_2\ i_2 \exp[v_2\, q/n_2 kT)].[\ 1 - \exp(-v_1 q/kT)] + (W/L)_4\ i_4 \exp[v_1 q/n_1 KT].[1\text{-}\exp((v_2\text{-}DRV)q/kT)]$

where

$$i = I_0 \exp[-V_{th}/(n\ k\ T/q)] \quad (3)$$

and I_0 is given by:

$$I_0 = \mu_o\ C_{ox}(W/L)\ V_{th} \exp(1.8) \quad (4)$$

with μ_o being the zero bias mobility.

$v_1 = (k\ T\ (A1 + A_5)/q\ A_2) \exp[-q\ DRV_1/n_2\ k\ T]$ $v_2 = DRV_1 - kT\ .A_4/q.A_3.\ \exp[-q\ DRV_1/n_3\ kT\ ]$ and

$A_i = I_0\ (W/L)_i \exp(-q\ V_{th}/k\ T\ n_i)$

These models reveal an asymmetry in the basic cache cell structure, as well as in transistor characteristics and physical geometry. Integrating these models into the power estimator, have led to more accountable results. For various power estimations, this work employed a modified version of cacti [6], and Wattch [7].

III. SHORT CHANNEL LENGTH CACHE IMPACT

The equations of the precedent section clearly show that the leakage power dissipation in CMOS devices is a strong function of transistor channel length (L). While shrinking (L) decreases the gate capacitance of the device and hence its dynamic power consumption [3], it stimulates, on the other hand, the standby current component leading to an increase in leakage dissipation. Moreover, such an increase exacerbates as the threshold voltage is scaled down along with supply voltage to yield some performance boost. Voltage scaling is perhaps the most effective method of saving power due to the square law dependency of digital circuit active power on supply voltage. This trend is rampant in most modern process scaling. The associated leakage escalation is said to be due Short Channel Effect (SCE).For analytical purposes, we propose four models of the original StrongARM processor. Namely: Strong-1, Stong-2, Strong-3 and Strong-4 with Strong-1 being the baseline-fabricated with 0.18um process. These models represent the past, today and feature technologies. The leakage losses in each of these processor caches are investigated and analyzed independently.

TABLE-I
StrongARM VARIANTS AND CHARACTERIZATION

CPU Alias	Strong-1	Strong-2	Strong-3	Strong-4
Generation (um)	0.18	0.13	0.1	0.07
Supply Voltage(V)	2	1.1	1.2	1
Threshold Voltage(V)	0.2	0.1	0.09	0.001

The unwanted effect of shortening device channel length comes at the expense of great performance degradation. The problem has been tackled at a higher level of design abstraction by some research. Khouri et. al. [3] has dealt with the problem at the gate level and provided great optimization. We approached the problem at higher level of abstraction exploring the cache microarchitecture. The cache system organization is targeted. The impact of cache associativity and cache-line size in the magnitude of the subthreshold current is examined. The instruction and data caches of the various StrongARM clones are modeled and run under different associativities. The cache sizes for all processors are kept the same as that of the baseline. The behavior of the corresponding subthreshold currents due cache reorganization is closely monitored. The results are presented thoroughly in the next section.

IV. EXPERIMENTAL RESULTS

All results have shown consistent reduction of total cache power when migrating from one generation to another as technology downsizes. The leakage loss echoed the same trend

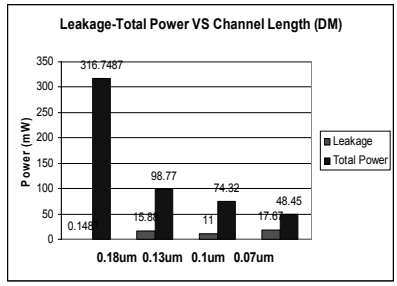

Figure 3: Cache System Profiles of Storng-1, Strong-2, Strong-3 and Strong-4 (when all directly mapped)