九州大学学術情報リポジトリ
Kyushu University Institutional Repository

# Simultaneous optimization of memory configuration and code allocation for low power embedded systems

Matsumura, Tadayuki
Graduate School of Information Science and Electric Engineering, Kyushu University

Ishihara, Tohru
System LSI Research Center, Kyushu University

Yasuura, Hiroto
Faculty of Information Science and Electric Engineering, Kyushu University

https://hdl.handle.net/2324/10569

出版情報：Proceedings of the 18th ACM Great Lakes symposium on VLSI, pp.403-406, 2008-05. Association for Computing Machinery
バージョン：
権利関係：

# Simultaneous Optimization of Memory Configuration and Code Allocation for Low Power Embedded Systems

Tadayuki Matsumura[1]  Tohru Ishihara[2]  Hiroto Yasuura[3]

[1]Graduate School of Information Science and Electric Engineering, Kyushu University
[2] System LSI Research Center, Kyushu University
[3] Faculty of Information Science and Electric Engineering, Kyushu University
{matsumura,yasuura}@c.csce.kyushu-u.ac.jp  ishihara@slrc.kyushu-u.ac.jp

## ABSTRACT

This paper proposes a hybrid memory architecture which consists of the following two regions; 1) a dynamic power conscious region which uses low $V_{dd}$ and $V_{th}$ and 2) a static power conscious region which uses high $V_{dd}$ and $V_{th}$. This paper also proposes an optimization problem for finding the optimal memory division ratio, the code allocation, $\beta ratio$ and $V_{dd}$ so as to minimize the total power consumption of the memory under constraints of static noise margin (SNM), memory access delay and area overhead. Experimental results demonstrate that the total power consumption can be reduced by 50.8% with 7.7% memory array area overhead without degradations of SNM and access delay.

## Categories and Subject Descriptors

B.3 [MEMORY STRUCTURES]: General

## General Terms

Experimentation

## Keywords

Low power design, On-chip memory, code allocation

## 1. INTRODUCTION

Low power design is one of the most important criteria for today's circuit designers. Power consumption is divided into two components, dynamic power consumption and static power consumption. Since the dynamic power consumption depends on $V_{dd}$ quadratically, the dynamic power consumption can be reduced drastically by lowering $V_{dd}$. However, lowering the $V_{dd}$ causes an increase of the delay which degrades the entire synchronous processor performance. To keep the processor performance, designers have to lower the $V_{th}$ as well. However in deep sub-micron technology, this causes an exponential increase in subthreshold leakage power consumption [1]. Because of above reasons, it is important for designers to consider the dynamic-to-static power ratio, and to decide the $V_{dd}$ and $V_{th}$ carefully. On-chip memory is one of the most power hungry components of today's microprocessors [2]. In general, memory is designed by using high $V_{dd}$ and high $V_{th}$ due to its low activities and a static power dominant characteristic. It is common observation that there is reference locality in memory systems, and as a result, there is also deflection in the dynamic power consumption. We exploit the memory reference locality for reducing the total power consumption using a hybrid memory architecture. The hybrid memory architecture consists of the following two regions; 1) a dynamic power conscious region which uses low $V_{dd}$ and low $V_{th}$ and 2) a static power conscious region which uses high $V_{dd}$ and high $V_{th}$. The total power consumption can be saved by concentrating memory accesses on the dynamic power conscious region. The key of our architecture is that the access delays for the two regions are equal to each other by assigning low $V_{th}$ to the dynamic power conscious region, which eases to integrate this memory into processors without major modifications of an internal processor architecture. Our proposed power reduction technique does not degrade performance and static noise margin (SNM) with a slight area overhead.

The rest of the paper is organized as follows. In section 2, our approach and related works are presented. an optimization problem for minimizing the total memory power consumption under constraints of SNM and area overhead is formally defined. Section 4 presents experimental results. The final Section conclude the paper.

## 2. CODE ALLOCATION FOR HYBRID MEMORY ARCHITECTURE

### 2.1 Hybrid Memory Architecture

Since there is reference locality in memory accesses, a large amount of dynamic power is consumed in those frequently accessed addresses.To exploit this reference locality, our hybrid memory architecture employs the following two regions; a dynamic-power-conscious region (we refer to this region as DP) and a static-power-conscious region (we refer to this region as SP). The DP region is designed with low $V_{dd}$ and low $V_{th}$ to decrease dynamic power consumption without increasing the access delay. The SP region is the same as a conventional memory which uses higher $V_{dd}$ and higher $V_{th}$ than those of DP region. The total power consumption can be reduced by concentrating memory accesses on

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
*GLSVLSI'08,* May 4–6,2008, Orlando, Florida, USA
Copyright 2008 ACM 978-1-59593-999-9/08/05 ...$5.00.

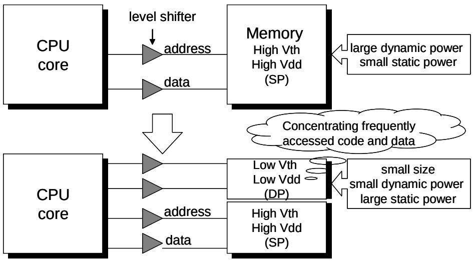

Figure 1: Target System and Proposed System

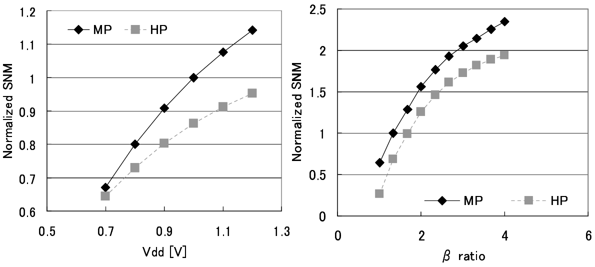

Figure 2: SNM vs $V_{dd}$, $\beta_{ratio}$

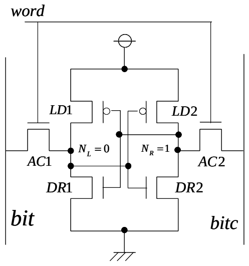

Figure 3: SRAM cell schematic

DP region. The key of our hybrid memory architecture is that the $V_{th}$ of the DP region is lowered to compensate an increase of access latency. Because of this, there is no performance degradation. Moreover it makes easy to embedding our memory into the processor since there is no difference between our hybrid memory and conventional memory with regard to memory access delay. Figure 1 shows a simplified image of conventional system and our proposed system.

## 2.2 Our Approach

In this paper we apply the idea of hybrid memory architecture to scratchpad memory (SPM). SPM is a small and high speed on-chip memory which consists of an SRAM. At first, we find the optimal DP-to-SP region ratio, supply voltage and SRAM cell size for a given application domain. The application domain represents a set of target applications. Once the optimal memory configuration is found, the next step is to find a code allocation to DP and SP regions for each application. In past, there are several code allocation techniques proposed for SPM to improve performance and reduce power consumption [3, 4]. The code allocation is done at the compilation phase. In this paper, we find functions and data objects ( these are referred as memory objects ) which should be allocated into the two regions of the hybrid memory for minimizing the total power consumption of the memory. The data objects include global variables and constants. For finding the optimal code allocation, we need to measure the number of accesses to each memory object for a given application program. We use an instruction set simulator for obtaining this information. The dynamic and static power consumptions of memory modules can be obtained through SPICE simulation. For such given base data, we find the optimal code allocation to minimize the total power consumption of the memory by solving optimization problem described in section 3. The optimization problem also finds the optimal memory division ratio, the optimal $V_{dd}$ and $\beta ratio$ of the DP region for minimizing the total power consumption with some area overhead without access delay and SNM degradations.

## 2.3 SRAM Stability Model

The SRAM cell stability is one of the most important criteria for SRAM circuit design. The static noise margin (SNM) is a widely used criterion for representing a SRAM cell stability. The SNM is defined as the minimum DC noise voltage necessary to flip the state of a cell [8]. The SNM is degraded with a decrease of $V_{dd}$ and $V_{th}$. In proposed hybrid memory architecture, low $V_{dd}$ and low $V_{th}$ are assigned to the DP region. Figure 2 shows the relationship between the SNM and $V_{dd}$. The results shown in Figure 2 are obtained by SPICE simulation. In this paper, a commercial 90nm CMOS process technology SPICE model is used consistently. In this model, 2 process options; HP and MP are provided. The HP model library is a performance oriented model, and its $V_{th}$ and $T_{ox}$ (gate oxide thickness) are chosen for increasing performance of the circuit. On the other hand, the parameters of the MP model library are chosen for low power design. In this paper, it is assumed that DP region and SP region are designed using HP model and MP model, respectively. Figure 2 shows that the SNM of the DP region is much less than that of the SP region at the same $V_{dd}$. To compensate for the SNM degradation, a larger $\beta$ ratio ($=\beta_{DR}/\beta_{AC}$) SRAM cell is used for the DP region. $\beta_{DR}$ and $\beta_{AC}$ are the transconductance factors of the transistor DR and AC respectively (see Figure 3). $\beta$ is given by $W/L$ where L and W indicate the transistor length and width, respectively. Figure 2 shows the relationship between the SNM and $\beta$ ratio. The SNM increases by using a large $\beta$ ratio SRAM cell though it causes undesirable area overhead. However, DP region is expected to be small due to a reference locality and, as a result the area overhead is also expected to be very small.

## 2.4 Related Work

In [5], non-uniform set-associative (NUSA) cache is proposed. The NUSA cache consists of one fast cache-way and several slow cache-ways. Frequently accessed data are gathered to the fast cache-way and infrequently accessed data are placed to the slow cache-ways. The slow ways use high $V_{th}$ to suppress the leakage power. This technique drastically reduces the leakage power of cache memory. However, since the access latencies for the fast and slow ways are different from each other, the NUSA cache needs a complicated pipeline structure which makes it difficult to integrate this cache into off-the-shelf processor IPs. In [6], Biased Partitioning (BP) configuration is proposed. BP divide the on-chip memory into 2 regions so as to reduce the dynamic power consumption. By dividing the memory into biased 2 regions, one region's load capacitance of the bit line gets smaller and by concentrating the access on this small load capacitance region, the dynamic power consumption can be reduced. However, in [6], the same $V_{dd}$ and $V_{th}$ are assigned to the two divided regions, and static power is not discussed. In [7], a technique exploiting a small subprogram memory whose $V_{dd}$ and $V_{th}$ are lower than those of conventional memory is proposed. An optimization flow to find the optimal $V_{dd}$, $V_{th}$ and code allocation for the subprogram to minimize the total power consumption is also proposed. However, this technique needs to insert extra jump instructions at compiling phase and memory stability issue is not taken into account nevertheless the subprogram memory is designed using lower $V_{dd}$ and $V_{th}$.

# 3. OPTIMIZATION PROBLEM FOR MINIMING POWER CONSUMPTION

## 3.1 Power and Delay Models

The access delay, the dynamic and static power consumption for each memory region are obtained by SPICE simulation. These parameters depend on $V_{dd}$ and $\beta ratio$. These dependencies are stored in a look-up table so that they can be use in our optimization problem. Moreover, the dynamic energy consumption per memory access and the access delay depends on the memory division ratio (i.e. memory size). To consider this dependency, the energy consumption and the access delay to each memory region are expressed as functions of the number of SRAM cells connected to each bit line. Figure 4 shows the access delay and the dynamic energy consumption per memory access for the different number of SRAM cells connected to each bit line. Figure 4 indicates that the access delay and the dynamic energy consumption can be approximated as a linear function of the number of SRAM cells connected to each bit line. These approximation coefficients are also stored in the look-up table for each $V_{DD}$ and $\beta ratio$.

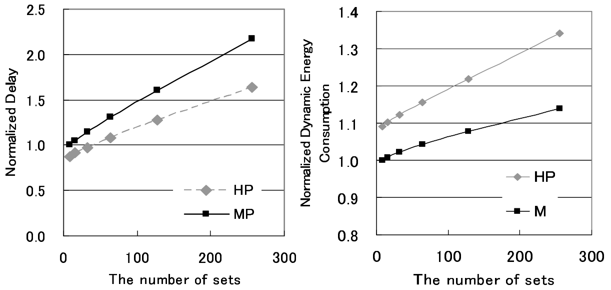

Figure 4: Access Delay and Dynamic energy consumption vs The number of SRAM cells connected to each bit line

## 3.2 Notation

- $A$ : The number of given application programs.
- $N_i$ : The number of memory objects
- $MS$: The total SPM size in byte.
- $s$: The size of the DP region in byte.
- $T_i$: The total program execution time
- $FS_{i,j}$: The size of the memory object
- $ACC_{i,j}$: The number of accesses
- $VDD_{DP,SP}$: The supply voltages
- $\beta r_{DP,SP}$: The beta ratios
- $ED_{DP,SP}$:The dynamic energy consumptions per access
- $PS_{DP,SP}$:The static power consumptions per byte
- $D_{DP,SP}$:The access delays
- $SNM_{DP,SP}$: The static noise margins
- $a_{i,j}$: 0-1 integer variable to be determined.

The subscripts $i$ and $j$ indicate that "$i^{th}$ application program" and "$j^{th}$ memory object", respectively. The $A$, $MS$ $VDD_{SP}$, and $\beta r_{SP}$ are given parameters.

## 3.3 Problem Description

At first, we find memory objects which should be allocated to SPM from the all memory objects in a given application program. In this paper, memory objects allocated to SPM are found by maximizing the number of accesses to SPM. After finding memory objects allocated to the entire SPM, we find optimal $VDD_{DP}$, $s$ (i.e., DP region size), $\beta r$ and optimal code allocation to DP and SP regions for minimizing the power consumption under constraints of a memory access delay, a static noise margin and an area overhead. The objective function and constraints are given by (1)-(6). $T_{Delay}$ (target delay) and $T_{SNM}$ (target SNM) are the original access delay and SNM of the memory which is designed as SP region entirely. $T_{AO}$ (target area overhead) is an input parameter. As described in previous section, $ED$ (dynamic energy consumption), and $D$ (access delay) of each memory region are functions of $VDD$, $\beta r$ and $s$. $PS$ (static power consumption) is also a function of $VDD$ and $\beta R$. The ED, PS and D are calculated by SPICE simulation and are stored in look-up table so that we can use them in our optimization problem. The optimal $VDD_{DP}$, $\beta r_{DP}$ and $s$ are decided for given application domain which consists of several application programs, and the optimal code allocations are found for each application program.

$Minimize:$

$$\sum_{i=1}^{A} \frac{1}{T_i} \left\{ \sum_{j=1}^{N_i} ED_{DP}(VDD_{DP}, \beta r_{DP}, s) \cdot ACC_{i,j} \cdot a_{i,j} \right. \\ \left. + \sum_{j=1}^{N_i} ED_{SP}(VDD_{SP}, \beta R_{SP}, s) \cdot ACC_{i,j} \cdot (1 - a_{i,j}) \right\} \\ + \{PS_{DP}(VDD_{DP}, \beta r_{DP}) \cdot s + PS_{SP}(VDD_{SP}, \beta r_{SP}) \cdot (MS - s)\} \tag{1}$$

$For \quad each \quad k = 1 \cdots A$

$$\sum_{j=1}^{N_k} FS_{k,j} \cdot a_{k,j} \leq s \tag{2}$$

$$\sum_{j=1}^{N_k} FS_{k,j} \cdot (1 - a_{k,j)} \leq (MS - s) \tag{3}$$

$$D_{DP}(VDD_{DP}, \beta r_{DP}, s) \leq T_{Delay} \tag{4}$$

$$SNM_{DP}(VDD_{DP}, \beta r_{DP}) \leq T_{SNM} \tag{5}$$

$$area\ overhead \leq T_{AO} \tag{6}$$

Table 2: The Experimental result

| MS | application | s/MS | $VDD_{DP}$ | $\beta ratio_{DP}$ | P_org [mW] | P_hyb [mW] | Reduction | Area Overhead |
|---|---|---|---|---|---|---|---|---|
| 8KB | jpeg | 0.258 | 0.7 | 3.66 | 1.15 | 0.66 | 43.2% | 7.72% |
| | mpeg | | | | 1.32 | 0.82 | 38.2% | |
| | compress | | | | 1.42 | 0.70 | 50.8% | |
| 16KB | jpeg | 0.172 | 0.7 | 3.66 | 1.71 | 1.05 | 38.2% | 5.15% |
| | mpeg | | | | 1.91 | 1.09 | 42.8% | |
| | compress | | | | 2.04 | 1.10 | 46.3% | |
| 32KB | jpeg | 0.0857 | 0.7 | 3.66 | 2.82 | 1.80 | 36.1% | 2.56% |
| | mpeg | | | | 3.10 | 1.85 | 40.4% | |
| | compress | | | | 3.27 | 1.83 | 44.0% | |

Table 1: Cell Area overhead vs $\beta ratio$

| $\beta ratio$ | 1.0 | 1.33 | 1.66 | 2.0 | 2.33 |
|---|---|---|---|---|---|
| $Areaoverhead$ [%] | -4.27 | 0 | 4.27 | 8.55 | 12.8 |
| $\beta ratio$ | 2.66 | 3.0 | 3.33 | 3.66 | 4.0 |
| $Areaoverhead$ [%] | 17.1 | 21.4 | 25.6 | 29.9 | 34.2 |

## 4. EVALUATION RESULT

This section shows the evaluation results of our proposed technique and demonstrates its effectiveness for power reduction. The optimization problem defined at previous section is solved using heuristic algorithm. The heuristic algorithm is based on the facts that $V_{DD}$ has stronger impact on the access delay and energy consumption than $\beta_{ratio}$. The target processor used in this experiment is SH3-DSP processor. The clock frequency of the processor is assumed to be a 400 MHz. The temperature of the chip is assumed to be a 75 ℃ for accurately estimating the active leakage current of the memory instead of the stand-by leakage current. Application domain is composed of three benchmark programs (i.e. jpeg, mpeg2, compress). Three different sizes of SPM are experimented. Input parameters $VDD_{SP}$, $\beta ratio_{SP}$ and $T_{AO}$ (target area overhead) are assumed to be 1.2V, 1.33, and 8% respectively. $ACC_i$ and $T_i$ are calculated from an instruction trace obtained by an instruction set simulator of SH3-DSP processor. The length of the trace is 1 million instruction long. The relationship between the $\beta ratio$ and SRAM cell area overhead is shown in TABLE 1. TABLE 1 is obtained by our 90nm SRAM cell design.

Table 2 shows the evaluation results. In every cases, the dynamic power consumption decreases while the static power consumption increases. However, the total power consumption is reduced in all cases since the reduction of dynamic power consumption is larger than the increase of the static power consumption. The reason why the total power consumption can be drastically reduced is that the optimal DP region size are small due to reference locality. Therefore, even low $V_{dd}$ can satisfy the delay constraint (see Figure 4) and it suppress the increase of the static power consumption. Although the large power saving can be obtained by applying our proposed technique, as described in section 2, DP region require large $\beta ratio$ cell to satisfy the SNM constraint and it enlarges the memory array area. However, the DP region of the optimal memory configuration is much smaller than the SP region. Therefore, area overhead of entire memory array area is tolerable. Especially for the 32KB SPM, the area overhead is only 2.56% The most important point is that our technique does not involve any performance and SNM degradations.

## 5. CONCLUSION AND FUTURE WORK

Hybrid memory architecture and code allocation problem for the hybrid memory are proposed for minimizing the on-chip memory power consumption without performance and SNM degradation. The proposed technique is applied to SPM, and its effectiveness is demonstrated by simulations. The results show that our proposed technique can save the total power consumption by 50.8% at the best case compared to the conventional memory with 7.72% memory array area overhead. In this paper, only memory array overhead is discussed and any other peripheral circuits are not discussed. However, dividing the bit line into two lines causes extra area overheads. Evaluating these overheads is our future work.

## Acknowledgement

This work is supported by VDEC, the Univ. of Tokyo with the collaboration of Renesas Technology, STARC, Panasonic, NEC Electronics, Toshiba, Synopsys, Inc., Cadence Design Systems, Inc., and Mentor Graphics, Inc. This work is also supported by CREST program of JST.

## 6. REFERENCES

[1] H. Chang and Sachin S. Sapatnekar “Full-Chip Analysis of Leakage Power Under Process Variations, Including Spatial Correlations”, in Proc. of DAC, pp.523-528, June, 2005.

[2] S. Segars, “Low Power Design Techniques for Microprocessors”, ISSCC Tutorial note, Feb. 2001.

[3] S. Stenke, L. Whmeryer, B. Lee and P. Marwedl, “Assigning Program and Data Objects to Scratched for Energy Reduction” in Proc of Design, Automation and Test in Europe Conference and Exhibition, pp.409-415, 2002.

[4] R. Banakar, S. Steinke, B. Lee, M. Balakrishnan and P. Marwedel, “ Scratchpad Memory:A Design Alternative for Cache On-chip memory in Embedded Systems ”in Prof of Design space exploration and architectural design of HW/SW system,pp73-78,2002.

[5] A. Sakanaka, S.Fujii and T. Sato, “A Leakage-Energy-Reduction Technique for Highly-Associative Cache in Embedded Systems” ACM SIGARCH Computer Architecture News Vol. 32, No. 3, June 2004.

[6] Naoyuki Kawabe and Kimiyoshi Usami, “Low-Power Technique for On-Chip Memory Using Biased Partitioning and Access Concentration” IEEE Custom Integrated Circuits Conference, pp. 275-278, May. 2000.

[7] T. Ishihara and K. Asada, “A System Level Power Optimization Technique Using Multiple Supply and Threshold Voltages”, in Proc. of ASP-DAC, pp.456-461, 2001

[8] E. Seevinck, F. List, and J. Lohstoh, “Static-Noise margin analysis of MOS SRAM cells” IEEE J. Solid-State Circuits, vol. SC-22, pp.748-754, 1987.