A Hybrid Memory Architecture for Low Power Embedded System Design

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

Ishitobi, Yuriko
Graduate School of Information Science and Electrical Engineering, Kyushu University

Ishihara, Tohru
System LSI Research Center, Kyushu University

Goudarzi, Maziar
System LSI Research Center, Kyushu University

他

出版情報：Proceedings of the Workshop on Synthesis And System Integration of Mixed Information Technologies, pp.56-62, 2007-10-15
バージョン：

A Hybrid Memory Architecture for Low Power Embedded System Design

Tadayuki Matsumura, Kyushu Univ., ISEE
Yuriko Ishitobi, Kyushu Univ., ISEE
Tohru Ishihara, Kyushu Univ., SLRC
Maziar Goudarzi, Kyushu Univ., SLRC
Hiroto Yasuura, Kyushu Univ., ISEE

ABSTRACT— On-chip memories are one of the most power hungry components of today's system on a chips (SoCs). The on-chip memories generally use higher supply (V_{dd}) and threshold (V_{th}) voltages than those of logic parts to suppress the static power consumption without increasing the access delay of the memories. This design policy, however, increases the dynamic power consumption since the dynamic power consumption is quadratically proportional to the V_{dd}. This paper proposes a hybrid memory architecture which consists of the following two regions; 1) a frequently accessed region which uses low V_{dd} and V_{th} and 2) a rarely accessed region which uses high V_{dd} and V_{th}. The key of our architecture is that the access delays for the two regions are equal to each other, which eases to integrate this memory into processors without major modifications of an internal processor architecture. This paper also proposes a technique for finding the sizes and the code allocation for the regions so as to minimize the total power consumption of the memory. Experimental results demonstrate that the total power consumption of the scratchpad memory can be reduced in every cases.

I. INTRODUCTION

The rapid growth of battery operated applications has motivated circuit designers to move toward low-power design. Power consumption is divided into two components, dynamic power consumption and static power consumption. Since the dynamic power is dominated by switching power in current technology node, we approximate the dynamic power as the switching power. The switching power is given by:

$$P_{dynamic} \simeq s \cdot f \cdot C \cdot V_{dd}^2 \tag{1}$$

where V_{dd} is supply voltage, s is switching activity, f is the clock frequency, C is the average switched capacitance of the circuit. As one can see from the equation (1) the V_{dd} has a strong impact on the dynamic power consumption due to its quadratic dependence. The CMOS circuit delay of ultra deep sub-micron short-channel transistors is given by [1]:

$$delay \propto \frac{V_{dd}}{(V_{dd} - V_{th})^{\alpha}} \tag{2}$$

where V_{th} is threshold voltage, and α is about 1.3 depending on the technology. From equation (2), decreasing 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 leakage power consumption. In the latest process technology, the leakage current is dominated by subthreshold leakage current and gate leakage[2][4][5]. In active mode, the subthreshold leakage power is main fraction of the total leakage power due to higher die temperature and exponential temperature dependence of subthreshold leakage current. The subthreshold leakage current is given by:

$$I_{Subthreshold} \propto \exp \frac{-V_{th}}{\theta \cdot V_T} \tag{3}$$

where V_T is thermal voltage, θ is subthreshold factor and its value is about from 1.4 to 1.65. Because of above reasons, designers have to decrease the dynamic power and static power in balance. It is important for designers to consider the dynamic-to-static power ratio, and to decide the V_{dd} and V_{th} carefully. 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 [7], and as a result, there is also deflection in the dynamic power consumption. More specifically, only a few addresses are frequently accessed, and the large percentage of the dynamic power is wasted in such addresses. On the other hand, most addresses are accessed infrequently, and the leakage power is wasted if low V_th is used. 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 low-dynamic-power region which uses low V_{dd} and V_{th} and 2) a low-leakage-power region which uses high V_{dd} and V_{th}. We can save the total power consumption by concentrating the memory accesses on the low-dynamic-power region. The key of our architecture is that the access delays for the two regions are equal to each other, which eases to integrate this memory into processors without major modifications of an internal processor architecture. In [3] a similar technique has already been proposed. However there are 2 different access delays and it needs more complicated hardware mechanisms than our hybrid memory architecture. The rest of the paper is organized as follows. In section2, we present a motivational example and our approach to reducing the power consumption of on-chip memories. The formal definition for problem to minimize the total power consumption of the memory is presented in Section3. Section 4 presents experimental results. The final Section concludes the paper.

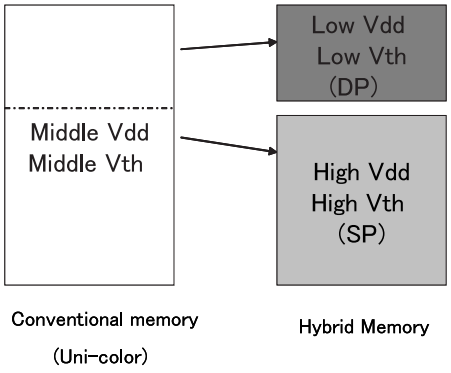

Fig. 1. Conventional memory and hybrid memory

TABLE I
SPECIFICATION OF CACHE MEMORIES

Processor	M32R-II	
Clock Frequency	400MHz	
Temperature	348K	
L1 (Instruction)	The number of sets	128
	The number of ways	2
	Line size	16 byte
	Total size	32 KB
L1 (Data)	The number of sets	128
	The number of ways	2
	Line size	16byte
	Total size	32 KB
L2 (Unified)	The number of sets	256
	The number of ways	4
	Line size	16 byte
	Total size	128 KB

II. CODE ALLOCATION FOR HYBRID MEMORY

A. Hybrid Memory

In memory circuits, the dynamic-to-static power ratio depends on several factors such as process technology, temperature, memory size, access frequency and clock frequency. Since there is reference locality in memory accesses, there is also deflection in the power consumption. More specifically, only a 10% region of memory dissipates 90% of the dynamic power and the other region consumes 90% of the static power. To balance the deflection, our hybrid memory architecture employs two different regions, that is 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 designed with high V_{dd} and high V_{th} to decrease static power consumption. The total power consumption can be reduced by allocating frequently accessed data to the DP region and infrequently accessed data to the SP region. The important point is that we can access to the DP region and the SP region with the same latency. Therefore, there is no need to modify an internal processor architecture in case of embedding our memory into the processor. Figure A. shows an example of the hybrid memory architecture. In this paper we apply this technique to Scratch Pad Memory (SPM).

B. Motivational Example

This section shows the importance of careful assignments of V_{dd} and V_{th} considering a dynamic-to-static power ratio for low power design. Suppose we have a processor with level-1 (L1) and level-2 (L2) on-chip caches. Generally, access frequencies of L1 and L2 caches are much different from each other. Figure 2 shows the number of accesses to cache memories in M32R-II processor. Our M32R-II design used in this example has L1-instruction, L1-data, and L2-unified caches. The numbers of accesses to the caches are measured by a cache simulator, and the energy consumption is estimated using, HSPICE, a commercial circuit simulator of SYNOPSYS. Experimental conditions are given in TABLE I. The number of accesses depends on application program. However in all applications, the number of L2 cache accesses are much less than that of L1 cache accesses. Therefore the dynamic-to-static power ratios are much different between L1 and L2 caches. Parameters for DP and SP used in this estimation are given in TABLE II. P.O., Delay, leak/cell, leak/SA, SW_{bit}, and SW_{word} represent process option, the read access latency, leakage current per SRAM cell, leakage current per sense amplifier, energy consumption per bit-line charge/discharge, and energy consumption per word-line charge/discharge, respectively. We obtain the values in TABLE II by using HSPICE. We use a commercial 90nm CMOS process technology. In this process technology, 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. Figure 3 and 4 show the estimated energy consumption breakdown of the DP-design and the SP-design for a 1 million instruction trace. Note that the SP-design uses high-V_{dd} and high-V_{th} for the entire memory region. Contrary the DP-design uses low-V_{dd} and low-V_{th}.

Figure 3 shows the L1-instruction cache is always minimized by applying the dynamic-power-conscious design (DP-design). In the L1-instruction cache, we can save energy consumption by 48% compared to the SP-design in the best case, and there is no performance degradation because the access latencies of the DP-design and the SP-design are almost the same from each other (see TABLE III). On the other hand, the power consumption of the L2 cache is always minimized by applying the static-power-conscious design (SP-design). In the L2 cache we can save energy consumption by 27% compared to the DP design in the best case. This difference is due to its deflection in the dynamic-to-static power ratio. The example demonstrates the importance of careful assignments of V_{th} and V_{dd} considering dynamic-to-static power ratio.

TABLE II
ACCESS DELAY AND POWER CONSUMPTION OF MEMORY MODULES

	Vdd [V]	P.O.	Delay [nsec]	leakage/cell [nW]	leakage/SA [nW]	SW_{bit} [J]	SW_{word} [J]
DP	0.75	HP	0.197	9.27	98.9	4.43e-14	3.38e-14
SP	1.2	MP	0.198	6.71	60.7	1.32e-13	1.40e-13

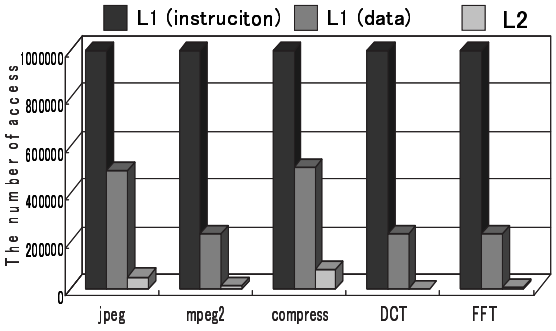

Fig. 2. The number of accesses to each cache

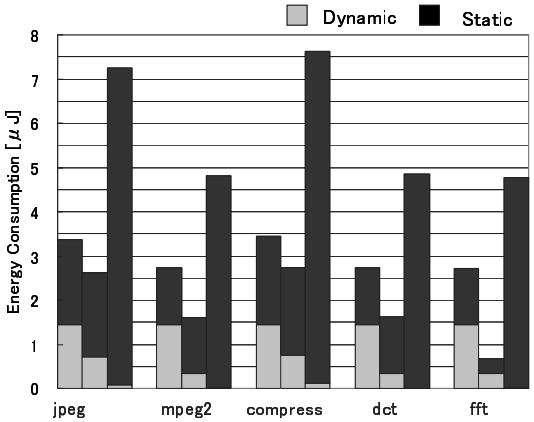

Fig. 3. Energy breakdown of DP-design

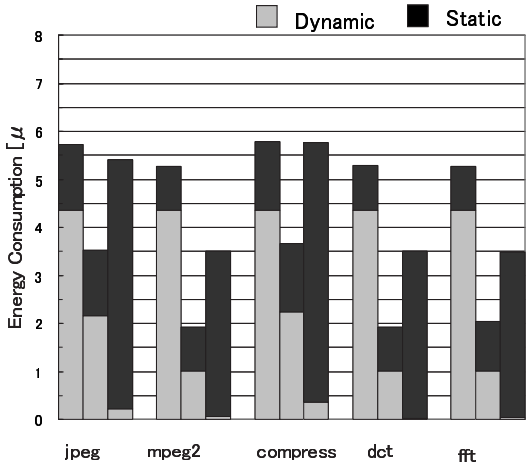

Fig. 4. Energy breakdown of SP-design

C. Related Work

In [3], non-uniform set-associative (NUSA) cache is proposed. The idea of NUSA cache is similar to our proposed hybrid memory. 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 cahce memory. However, the NUSA cache needs an extra mechanism for exchanging data between the fast and slow ways to gather frequently accessed data to the fast way. It causes a performance loss. In addition, 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. Unlike the NUSA cache, our hybrid memory does not suffer from such problems.

D. Our Approach

In this paper we apply the idea of our hybrid memory architecture to scratchpad memory (SPM). SPM is a small and high speed on-chip memory similar to cache memory. Both of cache memory and SPM consist of an SRAM. However SPM consumes lower power than that of cache memory, because SPM's addresses are assigned statically by software programmers, and as a result, there is no need to perform tag search operations which are necessary for cache memories. Because of above advantages, SPM is widely used for not only to improve performance but also to reduce the power consumption. There are several code allocation techniques proposed before. The typical code allocation technique finds basic blocks, procedures or functions which should be placed in SPM instead of a main memory so as to improve the performance or decrease the power consumption of processors in [8, 9]. The code allocation is done at the compilation phase. In this paper, we find functions and data 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 function and each data object for a given application program. We use an instruction set simulator for obtaining this information. By using a linker script, we can get the sizes of functions and data objects. 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

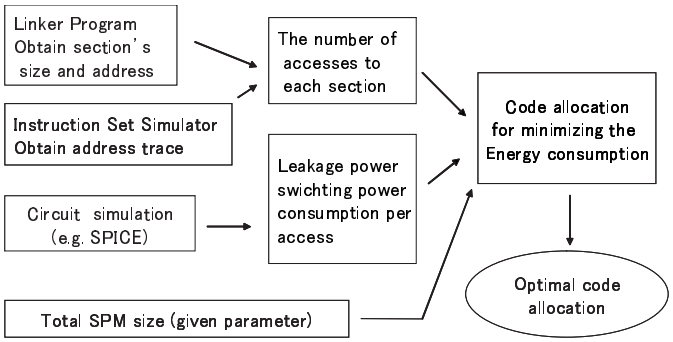

Fig. 5. Optimization flow

consumption of the memory. The optimization flow is described in Figure5.

III. PROBLEM DEFINITION

Formal problem definition is described in this section.

A. Notation

- A : The number of given application programs.
- $FS_{i,j}$: The size of the j^{th} function or data object in the i^{th} application program in byte.
- $X_{i,j}$: The number of accesses to the j^{th} function or data object in the i^{th} application program.
- NA_i : The number of all functions and data objects in the i^{th} application program.
- N_i : The number of functions and data objects in the i^{th} application program allocated on SPM
- ED_{DP},ED_{SP}: The dynamic energy consumption per access to DP and SP regions of SPM, respectively.
- ES_{DP},ES_{SP}: The static energy consumption per byte in DP and SP regions in SPM, respectively.
- MS: The total on-chip memory size in byte.
- s: the size of the DP region in byte.
- T: The total program execution time.
- $a_{i,j}$: 0-1 integer variable to be determined. If the j^{th} function or data object in the i^{th} application program is allocated on the DP, $a_{i,j}$=1. Otherwise $a_{i,j}$=0
- $b_{i,j}$: 0-1 integer variable to be determined. If the j^{th} function or data object in the i^{th} application program is allocated to SPM, $b_{i,j}$=1. Otherwise $b_{i,j}$=0

We obtain the ED_{DP}, ED_{SP}, ES_{DP}, ES_{SP} values from SPICE simulation. The MS value is given. We assume ED_{DP} and ED_{SP} are independent from the MS value in this paper. This assumption is impractical in some cases. Using more practical model is our future work.

B. Preliminary

At first, we have to find functions and data objects which should be allocated to SPM from the all functions and data objects in a given application program. There are several techniques proposed before to effectively use SPM [8][9]. In this paper, the functions and data objects allocated to SPM are found by maximizing the number of accesses to SPM. Namely, The functions which allocated to SPM is determined by solving the following optimization problem:

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

$$Maximize: \quad \sum_{j=1}^{N_k} X_{k,j} \cdot b_{k,j} \tag{4}$$

$$Subject\ to: \quad \sum_{i}^{NA_k} FS_{k,j} \cdot b_{k,j} \leq MS \tag{5}$$

The objective function and the constraint are given by (4) and (5), respectively. This problem is the typical knapsack problem.

C. Code Allocation Problem for Hybrid Memory

After finding functions and data objects allocated to the entire SPM, we find optimal s (i.e., DP-region size) and optimal allocation to DP and SP regions for minimizing the energy consumption. The objective function and the constraint are given by (6) and (7), respectively. By solving the following problem, we obtain the optimal value of s (i.e., the size of DP region) which minimizes the energy consumption of the memory for a set of given application programs.

$$\begin{aligned} Minimize: \quad & \sum_{i=1}^{A} \left\{ \sum_{j=1}^{N_i} ED_{DP} \cdot X_{i,j} \cdot a_{i,j} \right. \\ & \left. + \sum_{j=1}^{N_i} ED_{SP} \cdot X_{i,j} \cdot (1 - a_{i,j}) \right\} \\ & + A \cdot \{ES_{DP} \cdot s + ES_{SP} \cdot (MS - s)\} \cdot T \end{aligned} \tag{6}$$

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

$$\sum_{j=1}^{N_k} FS_{k,j} \cdot a_{k,j} \leq s \tag{7}$$

Finally, we find the optimal code allocation for predetermined s by solving the following optimization problem for each application program. This problem is also typical knapsack problem. The objective function and the constraint are given by (8) and (9), respectively.

$$\begin{aligned} Minimize: \quad & \sum_{j=1}^{N_i} ED_{DP} \cdot X_{i,j} \cdot a_{i,j} \\ & + \sum_{j=1}^{N_i} ED_{SP} \cdot X_{i,j} \cdot (1 - a_{i,j}) \\ & + (ES_{DP} \cdot s + ES_{SP} \cdot (MS - s) \cdot T \end{aligned} \tag{8}$$

$$\begin{aligned} Subject\ to: \quad & \sum_{j=1}^{N_i} FS_{i,j} \cdot a_{i,j} \leq s \\ & \sum_{i=1}^{N_i} FS_{i,j} \cdot (1 - a_{i,j}) \leq (MS - s) \end{aligned} \tag{9}$$

TABLE III
ACCESS DELAY AND POWER CONSUMPTION OF MEMORY MODULES

	Vdd [V]	P.O.	Delay [nsec]	leakage/cell [nW]	leakage/SA [nW]	SW_{bit} [J]	SW_{word} [J]
DP1	0.75	HP	0.197	9.27	98.9	4.43e-14	3.38e-14
SP1	1.2	MP	0.198	6.71	60.7	1.32e-13	1.40e-13
DP2	0.85	HP	0.151	10.5	122	6.07e-14	5.77e-14
SP2	1.0	MP	0.157	4.14	167	1.05e-13	8.13e-13

IV. EXPERIMENTS AND RESULTS

A. Experimental Setup

This section shows the experimental results of the hybrid SPM and demonstrates its effectiveness of the energy reduction. The target processor used in this experiment is SH3-DSP processor. The clock frequency of the processor is asuumed to be a 400 MHz. Three benchmark programs (i.e. jpeg, mpeg2, compress) and three different sizes of SPM is experimented. For each experiment, 2 type of DP and SP conditions are used as shown in TABLE III. Sense amplifier circuits used in DP2 and SP2 designs are same from each other. Contrary, sense amplifier circuits used in DP1 and SP1 design are same in terms of structure, but use different transistor sizes so as to match the access latencies of different regions. The DP and SP represent dynamic-power-conscious design and static-power-conscious design, respectively. The DP design uses low-V_{dd} and low-V_{th}. Contrary, the SP design uses high-V_{dd} and high-V_{th}. 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*.

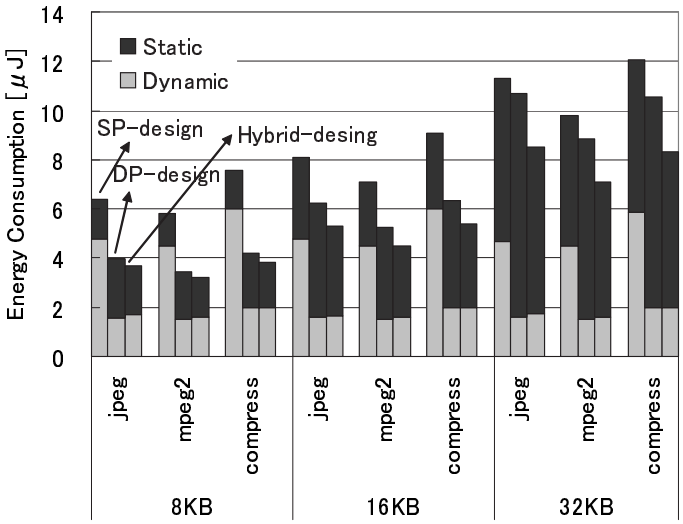

Fig. 6. Experimental result of DP1, SP1 condition

B. Results

TABLE IV, V and Figure 6, 7 show the experimental results for three different sizes of SPM. To compare the energy consumption of the hybrid memory architecture with the conventional memory architecture, the energy consumptions of SP-design and DP-design are also estimated. Note that the SP-design represent a SPM which uses high-V_{dd} and high-V_{th} for the entire memory region. Contrary the DP-design uses low-V_{dd} and low-V_{th}. E_hyb, E_sp, and E_dp represent energy consumption of SPM for hybrid-design, SP-design and DP-design respectively. Reduction_SP (or DP) represents the energy reduction achieved by hybrid memory compared to SP-design (or DP-design). In this experiments, the application domain is composed of 3 programs (i.e., jpeg, mpeg2, compress). Therefore, the size of DP region s is optimized for the three application programs.

The results show that our approach could saves energy consumption in all cases and in the best case, 49% energy consumption is saved compared to SP-design (compress 8KB). In the case of comparison with SP-design, the reduction is due to the reduction of the dynamic energy consumption. On the other hand, static power is increasing in all cases because the SP-design is the leakage optimal design in these experiments. However in total, the reduction of the

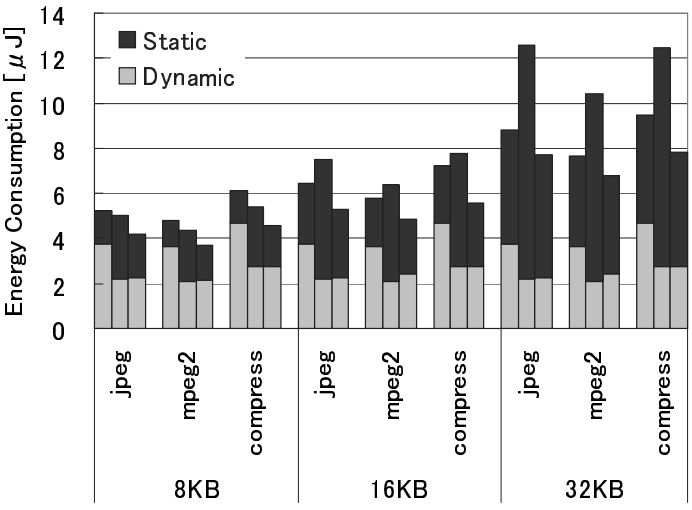

Fig. 7. Experimental result of DP2, SP2 condition

TABLE IV
THE EXPERIMENTAL RESULT FOR DP1 AND SP1 CONDITIONS

MS	s / MS	application	E_hyb [μ J]	E_sp [μ J]	Reduction (SP)	E_dp [μ J]	Reduction (DP)
8KB	0.347656	jpeg	3.67	6.38	42.5%	3.97	7.6%
		mpeg2	3.20	5.81	44.8%	3.44	7.0%
		compress	3.85	7.56	49.1%	4.23	9.0%
16KB	0.199219	jpeg	5.30	8.07	34.3%	6.22	14.8%
		mpeg2	4.51	7.10	36.5%	5.24	13.9%
		compress	5.39	9.09	40.7%	6.32	14.8%
32KB	0.123047	jpeg	8.53	11.29	24.4%	10.71	20.3%
		mpeg2	7.09	9.80	27.7%	8.82	19.6%
		compress	8.31	12.05	31.0%	10.54	21.1%

TABLE V
THE EXPERIMENTAL RESULT FOR DP2 AND SP2 CONDITIONS

MS	s / MS	application	E_hyb [μ J]	E_sp [μ J]	Reduction (SP)	E_dp [μ J]	Reduction (DP)
8KB	0.347656	jpeg	4.19	5.21	19.5%	4.99	16.0%
		mpeg2	3.71	4.80	22.6%	4.36	14.8%
		compress	4.58	6.09	24.9%	5.38	15.0%
16KB	0.130859	jpeg	5.31	6.43	17.4%	7.52	29.4%
		mpeg2	4.87	5.76	15.5%	6.38	23.6%
		compress	5.59	7.22	22.5%	7.75	27.8%
32KB	0.064453	jpeg	7.71	8.84	12.8%	12.57	38.7%
		mpeg2	6.79	7.68	11.6%	10.41	34.8%
		compress	7.83	9.46	17.2%	12.48	37.2%

dynamic energy consumption is larger than the increase of the static energy. As a result, the total energy consumption can be reduced in every cases. The most important point is that any performance loss is not involved. As shown in Table IV, the energy consumption of the hybrid memory is much smaller than that of the SP-design in DP1 and SP1 conditions. In DP1 and SP1, however, energy reduction of hybrid memory compared to DP-design is not very large, especially in 8KB SPM the energy reductions are less than 10%. This si because, if the dynamic power consumption is dominant, the DP-design is very power-efficient and there is a small chance to save more power. In our experiments, the dynamic power consumption is always dominant for 8KB SPMs. Another reason is that assigned V_{dd} of SP-design is much higher (1.2V) tahn that of DP-design. To match the memory access latency of each region, high V_{dd} should be assigned to SP region. However it causes an increase of static energy consumption despite the SP region is static power conscious region. Therefore the difference of static power consumption between DP-design and SP-design is not large compare to dynamic power consumption difference in DP1 and SP1 condition. From the above observation, it is also needed to assign the V_{dd} carefully according to the performance requirement.

V. CONCLUSION AND FUTURE WORK

Hybrid memory architecture is proposed for decreasing the on-chip memory energy consumption. The idea of the hybrid memory architecture is applied to SPM, and its effectiveness is demonstrated by simulations. The results show that our hybrid memory architecture can save the total energy consumption by 49% at the best case compared to the conventional SP-design memory. The hybrid memory architecture is applied to only SPM in the paper. However this idea can be applied to the other type of memories as well. Our future work will be devoted to extend our current approach for applying it to cache memories.

ACKNOWLEDGEMENTS

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. We are grateful for their supports.

REFERENCES

[1] T. Sakurai and A. Newton, "Alpha-Power Law MOSFET Model and Its Applications to CMOS Inverter Delay and other Formulas", IEEE JSSCC Vol. 25, No2. April 1990.

[2] H. Chang and Sachin S. Sapatnekar "Full-Chip Analysis of Leakage Power Under Process Variations, Includ-

ing Spatial Correlations", in Proc. of DAC, pp.523-528, June, 2005.

[3] 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.

[4] R. Gonzalez, Benjamin M. Gordon, and Mark A. Horowitz "Supply and Threshold Voltage Scaling for Low Power CMOS" IEEE JOURNAL OF SOLID-STATE CIRCUITS, Vol. 32, No. 8, august 1997

[5] 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

[6] E. Morifuji, T. Yoshida, M. Kanda, S. Matsuda, S. Yamada and F. Matsuoka, "Supply and Threshold-Voltage Trends for Scaled Logic and SRAM MOSFETs" IEEE TRANSACTIONS ON ELECTRON DEVICES, Vol. 53, No. 6, June 2006

[7] John L. Hennessy and David A. Patterson, "COMPUTER ARCHITECTURE A Quantitative Approach First Edition" MORGAN KAUFMAN PUBLISHERS, 1990

[8] 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.

[9] 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.