Low-power Way-predicting Cache Using Valid-bit Pre-decision for Parallel Architectures

Hsin-Chuan Chen[*#] and Jen-Shiun Chiang[*]
[*]*Department of Electrical Engineering,Tamkang University*
Tamsui, Taipei, 251,Taiwan
[#]*Department of Electronic Engineering, St. John's & St. Mary's Institute of Technology*
Tamsui, Taipei, 251, Taiwan
robin@mail.sjsmit.edu.tw

Abstract

Focusing on the way-predicting cache with sub-block placement, we propose a new cache scheme that uses the valid bits from data memory to pre-decide disabling the unnecessary tag-subarrays and data-subarrays. By valid-bit pre-decision, it significantly helps in improving the average energy saving of the conventional way-predicting cache without valid-bit pre-decision, especially for with large associativity and small sub-block size. Moreover, the proposed way-predicting cache can be applied to the parallel architecture systems to reduce the overall power consumption.

1. Introduction

In computer architectures, the cache memory plays an important role to reduce the speed gap between the processor and the main memory. Because processors access their own cache memories so frequently, for reducing the overall power consumption of parallel architecture systems such as multiprocessor systems, thus the low power caches become more important [1]. Due to lower miss rate, set-associative caches are usually used in modern computer systems to improve system performance. Increasing higher associativity of a set-associative cache will help in reducing the probability of block contention; however, the larger energy dissipation will be incurred. Therefore, how to maintain a low overall average access time and reduce the total average energy dissipation are important issues in the design of a set-associative cache.

In the past, several approaches have been proposed to reduce the energy dissipation of a set-associative cache, such as subbanking [1] in the data-subarray for each way, which only enables the desired subbank by word address decoding to save more power than the cache without subbanking. The selective-way cache [5] provides the ability to dynamically enable a subset of data ways on demand but all tag ways are checked together, and thus it reduces the switching activity of cache to save the cache power. The adjustable-way cache [6] can provide flexibilities to adjust its associativity according to different program behaviors; it therefore can reduce the average power consumption due to enabling fewer tag ways and data ways. One effective approach to reduce power is the way-predicting cache proposed by K. Inoue [3]. Due to the high hit rate of the MRU way prediction, the way-predicting cache can improve the average energy dissipation compared to the conventional set-associative cache because only one way's tag-subarray and data-subarray is enabled for most cache references. In this paper, we propose a new way-predicting cache based on sub-block placement, which uses the valid bits from data memory to pre-decide disabling the unnecessary tag-subarrays and data-subarrays. For a cache with large associativity and small sub-block size, we find that using valid-bit pre-decision indeed achieves a significant improvement in reducing average energy dissipation of the conventional way-predicting cache without valid-bit pre-decision.

2. Way-prediciting cache

In the architecture of the way-predicting cache, a way-predictor and an MRU table are required, where the MRU table records the information about the most recently used block for each set in a cache, and the way-predictor can maintain the status of the MRU table and send the enable signals to all tag-subarrays and data-subarrays. Like the MRU cache [7], when the first prediction of the way-predictor for each cache access is hit, it directly uses the information stored in the MRU table to only enable the MRU way's tag-subarray and data-subarray. If the prediction hit is not found, then it needs another cycle to enable the rest ways according to the results from comparators. Based on the CACTI timing

Proceedings of the 19th International Conference on Advanced Information Networking and Applications (AINA'05)
1550-445X/05 $20.00 © 2005 IEEE

model [2] and the analytical energy model [1], therefore, the average energy ($E_{C(WPD)}$) and average access time ($T_{AS(WPD)}$) of an n-way way-predicting cache (called WPD cache) can be expressed by [4]:

$$\begin{aligned} E_{C(WPD)} &= H \times PR \times (E_{tag} + E_{data}) \\ &+ H \times (1 - PR) \times n \times (E_{tag} + E_{data}) \\ &+ (1 - H) \times (n + 1) \times (E_{tag} + E_{data}) \end{aligned} \quad (1)$$

$$\begin{aligned} T_{AS(WPD)} &= H \times PR + H \times (1 - PR) \times 2 \\ &+ (1 - H) \times (P + 2) \quad Cycles \end{aligned} \quad (2)$$

where E_{tag} and E_{data} denote the energy dissipated in the tag-subarrays and data-subarrys, respectively, H is the cache hit rate, PR is the prediction-hit rate, and P is the miss penalty cycles depended on the sub-block size. From the above equations, if PR is high, the improvement over the conventional set-associative cache in average energy dissipation of the WPD cache will be significant. However, the extra overhead access time is required on way-prediction misses.

3. New low-power way-predicting cache

Usually, the sub-block placement [8], which only refills a part of the entire block into the cache when the miss occurs, is an appropriate approach to reduce miss penalty. In this cache scheme, each data block is divided into several sub-blocks, and each sub-block has a corresponding valid bit to indicate if this sub-block exists in the cache. Therefore, for a set-associative cache with sub-block placement, when the cache is accessed, in additional to tag checking of all ways, the corresponding valid bits of all ways must be checked together. In this paper, focusing on the way-predicting cache with sub-block placement, fortunately, the valid bits from the data-subarrays can be used to pre-eliminate the unnecessary probes at each cache access. Moreover, it can make the original rest enabled ways at the second cycle become the first-cycle access if the valid bit of the MRU way does not exist. Based on this idea, a new way-predicting cache using valid-bit pre-decision (called WPD-V cache) is proposed to progressively reduce the energy dissipation.

3.1. Architecture

The architecture of the proposed way-predicting cache shown in Figure 1 is similar to the conventional way-predicting cache. However, the valid bits of all sub-blocks are separated from the data memory, and they are organized as a single n-bit valid-bit bank implemented by flip-flops rather than SRAM, and each bit represents one valid bit of the accessed sub-block for each way. The way-predictor can decide sending the necessary enabled signals to the corresponding tag-subarrays and data-subarrays according to the MRU block bits and valid bits. In addition, the MRU table is maintained when the way-prediction misses or the cache misses, the way-predictor also updates the status of the valid-bit bank when the cache misses.

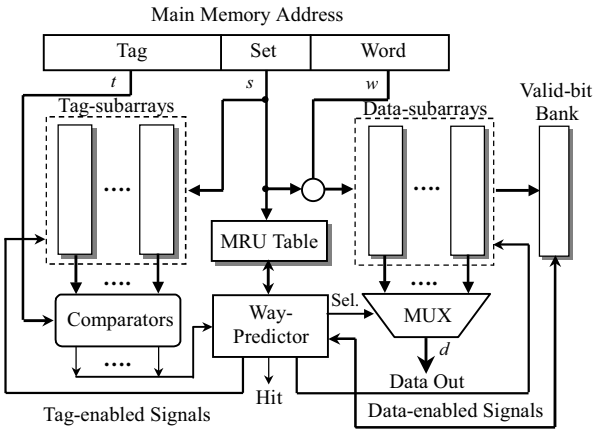

Figure 1. Architecture of WPD-V cache

3.2. Operations

The proposed WPD-V cache performs its way prediction based on the original MRU bits plus valid bits, which differs from the WPD cache only using MRU bits. The operations of the WPD-V cache are described as follows:

(1) While a set of the cache is referred, concurrently, the way-predictor fetches its MRU table and reads the valid bits from the valid-bit bank.
(2) The way-predictor decides which ways with tag-subarrays and data-subarrays should be disabled according as their corresponding valid bits are "0".
(3) When the valid bit of the MRU way is "1" and the way-prediction hits, only one MRU way with tag-subarray and data-subarray is enabled, and the desired data are speculatively read out in one cycle.
(4) When the valid bit of the MRU way is "1" but the way-prediction misses at the first cycle, the current MRU way is disabled and the rest ways with tag-subarrays and data-subarrays, whose valid bits are "1", are enabled at the second cycle.
(5) Once the valid bit of the MRU way is "0", the rest ways with valid bits "1" should be enabled at the first cycle instead of the second cycle.
(6) When a cache miss occurs, more cycles are required to refill a new sub-block from the lower-level memory during the replacement operation. Simultaneously, the status of the MRU table and valid-bit bank will be maintained.

3.3. Overheads

There mainly exist access time overhead and energy consumption overhead when the MRU table and valid-bit bank are accessed. The valid-bit bank, like the MRU table, is also implemented by flip-flops rather than SRAM to avoid the access time overhead and the energy overhead [3]. On the other hand, the update number of the MRU table and valid-bit bank are very few due to high prediction-hit rate and low miss rate, and thus the energy dissipation to update the MRU table and valid-bit bank can be ignored compared with the energy dissipated in the SRAM memory cells.

3.4. Energy/access time evaluation

The average energy dissipation ($E_{C(WPD\text{-}V)}$) and the average access time ($T_{AS(WPD\text{-}V)}$) are used to evaluate the performance and energy of the proposed WPD-V cache, respectively. From equations (1) and (2) in Section 2, $E_{C(WPD\text{-}V)}$ and $T_{AS(WPD\text{-}V)}$ of an n-way WPD-V cache can be respectively modified as the following equations:

$$\begin{aligned} E_{C(WPD-V)} &= H \times PR \times E_{way} \\ &+ H \times (1-PR) \times [VR_1 + VR_2 \times (n-1)] \times E_{way} \\ &+ (1-H) \times [VR_{m1} + VR_{m2} \times (n-1) + 1] \times E_{way} \end{aligned} \quad (3)$$

$$\begin{aligned} T_{AS(WPD-V)} &= H + H \times (1-PR) \times VR_1 \\ &+ (1-H) \times (P + 1 + VR_{m1}) \quad Cycles \end{aligned} \quad (4)$$

where $E_{way} = E_{tag} + E_{data}$, VR_1 and VR_2 are the valid-bit presence rates of the MRU way and rest ways when the way-prediction misses but cache hits, respectively, and VR_{m1} and VR_{m2} are the valid-bit presence rates of the MRU way and rest ways when cache misses, respectively. When the sub-block size of a cache decreases, the miss rate increases and the prediction-hit rate decreases. However, all valid-bit presence rates also decrease, which will help in improving its average energy and average access time, especially for the cache with large associativity.

4. Simulation results

For verifying the performance of the proposed way-predicting cache, we use a trace-driven cache simulator (Dinero) to simulate the access behaviors of two way-predicting caches, and both caches have the same cache size (= 32 KB), block size (= 32 Bytes), and replacement policy (LRU). The average access time and average energy dissipation of the way-predicting caches are evaluated by re-modeling Dinero to trace various trace programs. Here, we use 'Cycle' as a basic access time unit and 'E_{data}' as a basic energy unit, respectively, and thus E_{tag} = [(tag bits)/(block size × 8)] × E_{data} [3]. In our simulation, we change the range of the sub-block size and associativity separately to observe the average energy dissipation and average access time by averaging the simulation results of all benchmarks for the WPD cache and the WPD-V cache.

4.1. Average energy dissipation and access time

Table 1 shows the average energy and access time as the sub-block decreases at different associativities from 4 to 32. For large associativity, the average energy dissipation of the WPD cache significantly increases due to decrement of the prediction-hit rate. Opposite to the WPD cache, the average energy dissipation of the proposed WPD-V cache will dramatically decrease especially for large associativities because two valid-bit presence rates (VR_1 and VR_2) significantly decrease as the sub-block size decreases. By valid-bit pre-decision, we find that some second accessed ways only need one cycle when the valid bit of the MRU way does not exist; the WPD-V cache thus can achieve a lower access time than that of the WPD cache at the same associativity and sub-block size. Reducing sub-block size can reduce miss penalty, however, a higher miss rate is also incurred such that more access cycles are required when the sub-block size turns small. Therefore, for the WPD-V cache, the reduction of average access time is not as significant as that of the average energy dissipation.

Table 1. Average energy dissipation and access time for two way-predicting caches

Associativities			Sub-block Size (Bytes)					
			1	2	4	8	16	32
4-way	*WPD*	E_C	1.620	1.591	1.538	1.410	1.324	1.275
		T_{AS}	2.293	2.305	2.316	2.080	1.959	1.991
	WPD-V	E_C	1.172	1.210	1.260	1.252	1.250	1.260
		T_{AS}	2.162	2.192	2.224	2.026	1.935	1.991
8-way	*WPD*	E_C	2.373	2.319	2.219	1.980	1.820	1.730
		T_{AS}	2.257	2.266	2.274	2.047	1.931	1.959
	WPD-V	E_C	1.342	1.444	1.588	1.601	1.613	1.687
		T_{AS}	2.114	2.142	2.176	1.989	1.903	1.959
16-way	*WPD*	E_C	4.274	4.170	3.980	3.524	3.223	3.057
		T_{AS}	2.259	2.270	2.273	2.051	1.938	1.965
	WPD-V	E_C	1.760	2.032	2.423	2.522	2.637	2.912
		T_{AS}	2.095	2.125	2.164	1.985	1.906	1.965
32-way	*WPD*	E_C	9.341	9.148	8.785	7.915	7.348	7.038
		T_{AS}	2.299	2.309	2.313	2.094	1.982	2.009
	WPD-V	E_C	2.917	3.631	4.691	5.131	5.621	6.575
		T_{AS}	2.098	2.134	2.181	2.012	1.940	2.009

※ Energy unit is 'E_{data}' and access time unit is '*Cycles*'.

4.2. Improvement over WPD cache

Figures 2 and 3, respectively, indicate the improvement of the WPD-V cache over the WPD cache in average energy dissipation and average access time. For the WPD-V cache with large associativity and small sub-block size, the improved rate in average energy dissipation (IMR_{EC}) has a significant increment, and it can be high up to 68% at the associativity = 32 and the sub-block size =1 byte in our simulation. Besides, the improved rate in average access time (IMR_{TAS}) also obtains about 8.7% under the same condition. Even for the most used associativity = 4 and sub-block size = 4 bytes, our proposed WPD-V cache still has an about 18% improved rate in average energy dissipation.

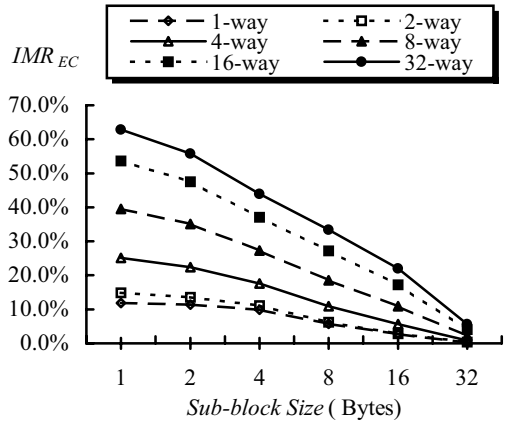

Figure 2. Improved rate in energy

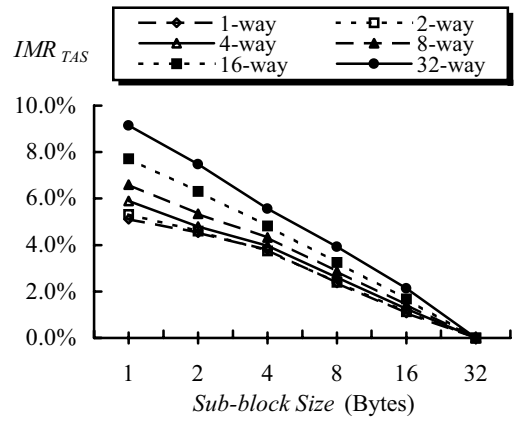

Figure 3. Improved rate in access time

5. Conclusions

In this paper, a new way-predicting cache using valid-bit pre-decision is proposed to progressively improve the energy dissipation and access time of the conventional way-predicting cache. Because the valid-bit presence rates will decrease as the sub-block size decreases, many unnecessary probed ways can be eliminated. Obviously, the proposed WPD-V cache indeed has a significant improvement in energy over the WPD cache, and the access time also further can be reduced. Therefore, this proposed WPD-V cache can be applied to the parallel architecture systems to reduce the overall power consumption.

6. References

[1] M. B Kamble and K. Ghose, "Analytical Energy Dissipation Models for Low Power Caches", *Proc. 1997 International Symposium on Low Power Electronics and Design*, Monterey, USA, Aug. 1997, pp. 143-148

.

[2] S. J. E. Wilton and N. P. Jouppi, "CACTI: An Enhancement Cache Access and Cycle Time Model", *IEEE Journal of Solid-State Circuits*, vol. 31, no. 5, New Jersey, USA, May 1996, pp. 677-688.

[3] K. Inoue, T. Ishihara, and K. Murakami, "A High-Performance and Low-Power Cache Architecture with Speculative Way-Selection", *IEICE Trans. on Electron*, vol. E83-C, no. 2, Tokyo, Japan, Feb. 2000, pp. 186-193.

[4] Z. Zhu and X. Zhang, "Access-Mode Predictions for Low-Power Cache Design", *IEEE Micro*, vol. 2, no. 2, Los Alamitos, USA, Mar. 2002, pp. 58-71.

[5] D. H. Albonesi, "Selective Cache Ways: On-Demand Cache Resource Allocation", *Proc. 32nd Annual International Symposium on Microarchitecture*, Haifa, Israel, Nov. 1999, pp. 248 –259.

[6] H. C. Chen and J. S. Chiang, "Design of An Adjustable-Way Set-Associative Cache", *Proc. 2001 IEEE Pacific Rim Conference on Communication, Computers and Signal Processing*, vol. 1, Victoria, Canada, Aug. 2001, pp. 315-318.

[7] H. C. Chen, J. S. Chiang and Y. S. Lin, "A Fast Sequential MRU Cache with Competitive Hardware Cost", *Proc. 2nd International Conference on Parallel and Distributed Computing, Application and Technologies*, Taipei, Taiwan, Jul. 2001, pp. 220-227.

[8] J. L. Hennessy and D. A. Patterson, *Computer Architecture: A Quantitative Approach*, Morgan Kaufman Publishers, Inc., 2nd ed., San Francisco, USA, 1997.