CONTENTS

Full Paper
Experimental Analysis on the Effect of Initial Concentration in the Direction Solidification of SN-PB Alloy .... Long-Sun Chao, Yu-Ru Chen, Hsien-Chang Peng and Wu-Shyang Li 729
Fire Resistance Tests of a Glass Pane with Down-Flowing Water Film .... Chan-Wei Wu, Ta-Hui Lin, Ming-Yuan Lei, Tien-Hsu Chung, Chien-Chang Huang and Wei-Tang Chiang 737
Automated Structural Optimization System for Integrated Topology and Shape Optimization .................................................. Chyi-Yeu Lin and Yu-Hsin Chou 745
Friction Stir Welding of Al 6061-T6 Thick Plates: Part I – Experimental Analyses of Thermal and Mechanical Phenomena ................................ Yuh J. Chao, Shu Liu and Chi-Hui Chien 757
NARX Neural Networks for Nonlinear Analysis of Structures in Frequency Domain .................. Ching-Yun Kao and Chin-Hsiung Loh 791
Design of a Low-Power Way-Predicting Cache Using Valid-Bit Pre-Decision Strategy ......... Hsin-Chuan Chen 805
Design of Intra-Bs Handoff in Multi-Hop IEEE 802.16 Broadband Wireless Networks .... Chien-Chung Chen, Jean-Lien C. Wu, Chia-Hu Chen and Chun-Yen Hsu 815
Numerical Simulations of Non-Ductile RC Frames with Infilled Brick Panel Under Cyclic Loading .... Chung-Yue Wang, Haishang-Yung Ho, Ren-Zuo Wang, and Hsu-Huei Huang 827
Shallow-Water Gravity Anomalies from Satellite Altimetry: Case Studies in the East China Sea and Taiwan Strait .................................. Cheinway Hwang and Hsin-Ying Hsu 841
Modeling Urban Landscapes with Integrated Datasets ................................................. Yi-Hsing Tseng, Pang-Wei Liu, and Fu-Chen Chou 853

Short Paper
Establishing an Invar Leveling Calibration System ........................................ Chun-Sung Chen, Chien-Ting Wu, Ming-Wei Chang and Wei-Chen Chang 861
Efficient Erasure Recovery Algorithm for Double Disk Failure Correction .... Chih-Shing Tau 867
Linear Relationship of the Control Points of Bezier Curve for Degree Reduction: A Tridiagonal Matrix Approach ........................................... Jyh-Jeng Deng and Tei-Hui Kao 873
Effect of a Magnetic Field on Vortex Instability of Natural Convection Flow over a Horizontal Plate with Blowing and Suction ........................................... Jin-Sheng Leu and Jin-Yuh Jang 879
Approach to Accurately Estimating Low-Contrast Inclusion with A Priori Guidance in Near-infrared Diffuse Optical Tomography .......... Min-Cheng Pan, Chien-Hung Chen, Min-Chun Pan and Yi-Ming Shyr 885
Thermodynamic Equilibrium Prediction for Natural Gas Dry Reforming in Thermal Plasma Reformer ................................................. Huan-Liang Tsai and Chi-Sheng Wang 891
DESIGN OF A LOW-POWER WAY-PREDICTING CACHE USING VALID-BIT PRE-DECISION STRATEGY

Hsin-Chuan Chen

ABSTRACT

Way-predicting cache is one of the set-associative caches that can effectively reduce power consumption, which only speculatively selects an MRU (most recently used) way before starting a normal cache access. Focusing on the way-predicting cache using sub-block placement, we propose a new cache scheme that uses valid bits from data memory to decide the disabled tag-subarrays and data-subarrays. By using valid-bit pre-decision, the proposed scheme has a significant improvement in average energy saving over the conventional way-predicting cache, especially for large associativity and small sub-block size. Moreover, because those original second accessed ways are first accessed during the first cycle when the valid bit of the MRU way does not exist, the proposed cache also can reduce the average access time.

Key Words: way-predicting cache, sub-block placement, average energy, valid-bit pre-decision.

I. INTRODUCTION

In computer architectures, a cache memory plays an important role in reducing the speed gap between the processor and the main memory, because processors access the cache memory very frequently, and several studies have shown the cache memory consumes about 25%~50% of the total power in many microprocessor systems (Kamble and Ghose, 1997). Therefore, to reduce the overall power consumption of computer systems, low-power caches have become important for modern computer architecture. Due to a lower miss rate, set-associative caches are usually used in modern computer systems to improve system performance. To increase associativity of a set-associative cache is good for reducing the probability of block contention (Hill, 1988). However, in that case, larger energy dissipation is 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.

Several approaches have been proposed to reduce the energy dissipation of a set-associative cache, such as subbanking (Su and Despain, 1995) 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 filter cache (Kin et al., 1997), a small low-energy cache, is placed in front of the L1 cache. If the filter cache can serve many L1 cache accesses, the energy efficiency of the memory system can be improved. The selective-way cache (Albonesi, 1999) 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 the cache to save cache power. Both adjustable-way cache (Chen and Chiang, 2005) and configurable-way cache (Chen and Chiang, 2003) can provide flexibilities to adjust their associativity according to different program behaviors; they therefore can reduce the average power consumption due to enabling fewer tag ways and data ways, where the configurable-way cache also can eliminate unnecessary energy dissipation for intra-block accesses. Alternate cache schemes, way memorization cache (Ma et al., 2001) and the span cache (Witchel and Asanović, 2001), only enable one data-subarray for accessing an intra-block cache data without performing tag checking to save power. Due to high prediction-hit rate of the MRU way, the way-predicting cache proposed by K. Inoue (Inoue et al., 2000) can improve the average energy dissipation...
compared with the conventional set-associative cache because only one way with tag-subarray and data-subarray is enabled for most cache references.

Combining the sub-block placement, we propose a new way-predicting cache that uses the valid bits from data memory to decide to disable the unnecessary tag-subarrays and data-subarrays with valid bits being "0" in advance. For a cache with large associativity and small sub-block size, the proposed way-predicting cache can produce a significant improvement in average energy saving over the conventional way-predicting cache by using the valid-bit pre-decision. In this paper, we focus on the proposed way-predicting cache architecture and its energy saving. Section II describes the concepts of the conventional way-predicting cache, and Section III explicates the sub-block placement and the impacts of the sub-block size. Section IV introduces the proposed way-predicting cache, including its architecture, operations, overheads, and performance evaluation. In Section V, we provide simulation results for the average energy dissipation and average access time of our proposed way-predicting cache compared with the conventional way-predicting cache. Finally, Section VI presents our conclusions.

II. WAY-PREDICTING CACHE

Due to the high prediction-hit rate of the MRU way, the way-predicting cache is one of the set-associative caches that can effectively reduce power consumption. In this Section, the architecture and operation of the way-predicting cache (WPD cache) will be described in detail.

1. Architecture and Operation

The basic architecture of the way-predicting cache (for example, 4-way) is shown in Fig. 1 (Inoue et al., 2000), where the way-predictor is a main control circuit that can maintain the status of the MRU table and send the enabled signals to all tag-subarrays and data-subarrays. Like the MRU cache (Chang et al., 1987; Chen et al., 2001), the MRU table records the most recently used block bits (i.e. MRU way) for each set in a cache. While a set of the cache is referred to, the way-predicting cache always directly uses the MRU block bits stored in the MRU table to only enable the MRU way with tag-subarray and data-subarray at the first cycle. If the prediction of the MRU way misses but the cache hits, then the current MRU way has to be disabled and another cycle is required to enable the other ways, and the MRU table can be updated by the tag-comparison results. If the cache misses, the way-predictor uses the refilled way determined by the LRU replacement circuit to update the MRU table. Consequently, the way-predicting cache can effectively reduce energy dissipation because only one way with tag-subarray and data-subarray is enabled for most cache references.

2. Energy and Access Time Models

According to the CACTI model proposed by Wilton (Wilton and Jouppi, 1996) and the analytical energy dissipation model proposed by Kamble (Kamble and Ghose, 1997), in a conventional set-associative cache, the major components that dissipate energy include row address decoder, memory cells, comparators, data multiplexor, sense amplifiers, and input/output drivers. However, tag RAM cells (i.e. tag-subarrays) and data RAM cells (i.e. data-subarrays) account for most of the energy dissipation in a cache, which includes energy dissipation to pre-charge bit lines, select word lines, and activate sense amplifiers of the RAM cells (Kamble and Ghose, 1997). Therefore, the total energy dissipation ($E_{\text{cache}}$) can be simplified by the following equation (Su and Despain, 1995):

$$E_{\text{cache}} = E_{\text{cell}} + E_{\text{decode}} + E_{\text{I/O}}$$

(1)

In Eq. (1), $E_{\text{decode}}$ is the energy dissipation to drive the address bus and decode memory address; $E_{\text{I/O}}$ is the energy dissipation to drive external I/O pins for the replacement operation when a cache miss occurs, and $E_{\text{cell}}$ is the energy dissipation to access all tag-subarrays and data-subarrays. From the results simulated by Kin (Kin et al., 1997), we found that the energy consumption of the decoders is about three orders of magnitude smaller than that of the other components, and thus $E_{\text{decode}}$ can be negligible compared to the other components (Bahar et al., 1998). For an $n$-way conventional set-associative cache with high hit rate, which means the cache miss penalty is ignored (i.e. $E_{\text{I/O}}$ is
also ignored), and thus the total energy dissipation \( E_{\text{cache}} \) can be simplified as the following equation:

\[
E_{\text{cache}} = E_{\text{cell}} = n \times (E_{\text{tag}} + E_{\text{data}}),
\]

where \( E_{\text{tag}} \) and \( E_{\text{data}} \) denote the energy dissipated in a tag-subarray and data-subarray, respectively. Usually, the partitioned subarray number of tag memory and data memory is equal to the cache associativity, and \( E_{\text{data}} \) is larger than \( E_{\text{tag}} \) because the data memory has a larger size than that of the tag memory in a cache, where \( E_{\text{tag}} = [(\text{tag bits})/(\text{block size} \times 8)] \times E_{\text{data}} \) (Inoue et al., 2000). Consequently, the average access time \( (T_{\text{AS}}) \) of a set-associative cache is:

\[
T_{\text{AS}} = 1 \text{ Cycle}.
\]

Because the way-predicting cache always enables only one MRU way before starting a normal cache access, and disables the other ways with tag-subarrays and data-subarrays, which means no pre-charging of bit lines, no selection of word lines, and no activation of sense amplifiers for these disabled tag-subarrays and data-subarrays. If a prediction miss occurs, then it enables the other ways at the second cycle. Therefore, considering the cache misses, the average energy \( E_{\text{C(WPD)}} \) and average access time \( (T_{\text{AS(WPD)}}) \) for the way-predicting cache based on the sub-block placement can be expressed by (Inoue et al., 2000; Zhu and Zhang, 2002):

\[
E_{\text{C(WPD)}} = H \times PR \times (E_{\text{tag}} + E_{\text{data}}) + H \times (1 - PR) \\
\times n \times (E_{\text{tag}} + E_{\text{data}}) + (1 - H) \times (n + 1) \\
\times (E_{\text{tag}} + E_{\text{data}}),
\]

\[
T_{\text{AS(WPD)}} = H \times PR + H \times (1 - PR) \times 2 \\
+ (1 - H) \times (P + 2) \text{ Cycles},
\]

where \( H \) is the cache hit rate, \( PR \) is the prediction-hit rate, and \( P \) is the miss penalty cycles depending on the sub-block size. If \( PR \) is high, the WPD cache significantly improves the average energy over the conventional set-associative cache; however, the extra access time overhead is required on way-prediction misses. If the overhead energy dissipated in the MRU table is also considered, then Eq. (4) would be:

\[
E_{\text{C(WPD)}} = H \times PR \times (E_{\text{tag}} + E_{\text{data}} + E_{\text{MRU}}) \\
+ H \times (1 - PR) \times [n \times (E_{\text{tag}} + E_{\text{data}}) \\
+ 2E_{\text{MRU}}] + (1 - H) \times [(n + 1) \\
\times (E_{\text{tag}} + E_{\text{data}}) + 2E_{\text{MRU}}]
\]

III. SUB-BLOCK PLACEMENT

Increasing block size will reduce the tag memory size for an on-chip cache design; however, a large miss penalty is incurred due to large block size. Usually, the sub-block placement (Hennessy and Patterson, 1997), which only refills a part of the entire block into the cache when the cache miss occurs, is an appropriate approach to reduce miss penalty. In this cache scheme (shown in Fig. 2), each data block in each data-subarray 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. For example, a data block with 32 bytes can be divided into 8 sub-blocks plus 8 valid bits when the sub-block size is 4 bytes. Therefore, for a set-associative cache with sub-block placement, when the cache is accessed, in addition to tag checking of all ways, the corresponding valid bits of all ways must be checked together.

The sub-block size mainly affects the cache miss penalty, and thus smaller sub-block size can achieve lower miss penalty (Diwan et al., 1995). However, the sub-block size, similar to the cache block size (Zhang et al., 2003), also affects the cache miss rate depending on the locality of executed program. For a program with good spatial locality, we find that the miss rate increases as the sub-block size decreases. Furthermore, these two factors will directly impact the total average access time. On the other hand, because the rate of sub-blocks with valid bits “1” for each way decreases as the sub-block size decreases, more power savings can be achieved if we only enable the necessary ways whose valid bits of their sub-blocks are “1” for each cache access. Therefore, the sub-block size should be decided from the tradeoff among miss penalty, miss rate, and energy dissipation according to the program locality.
IV. PROPOSED LOW-POWER WAY-PREDICTING CACHE

Focusing on the way-predicting cache with sub-block placement, fortunately, the valid bits from the data memory (as in Section III) can be used to pre-eliminate the probes of unnecessary memory cells at each cache access (Chen and Chiang, 2004). Furthermore, it can make the other original enabled ways, at the second cycle, become first-cycle accesses if the valid bit of the MRU way does not exist. Based on this idea, we had proposed a way-predicting cache with configurable sub-block size (Chen and Chiang, 2004), which focuses on how to organize the sub-block bank for various sub-block sizes by adding a sub-block size register (SB register) into the replacement circuit, where the SB register stores the number of data items in one sub-block configured by a special operating system. This way-predicting cache using valid-bit pre-decision (WPD-V cache) is mainly extended from the cache architecture that we proposed at the 2005 IEEE AINA Conference (Chen and Chiang, 2005), in addition to progressively reducing energy dissipation, the energy overhead caused by valid-bit pre-decision is also evaluated for completeness and realism in this paper. Among recent low-power cache examples, the PTC-V cache (Peng et al., 2006) proposed at TENCON 2006, cited the valid-bit pre-decision concept proposed by my paper (Chen and Chiang, 2005) to apply to the cache with partial tag comparison. Basically, the PTC-V cache is a partial tag comparison cache rather than being a way-predicting cache, and it uses the valid-bit pre-decision to reduce the unnecessary power of the bit-lines in the data memory during the time waiting for the result of the tag comparison.

1. Architecture of WPD-V Cache

Basically, the architecture of the proposed way-predicting cache shown in Fig. 3 is similar to the conventional way-predicting cache. However, the valid bits of all sub-blocks are separated from the data memory subarrays, and they are organized as a single n-bit valid-bit bank for an n-way way-predicting cache, where each bit represents one valid bit of the accessed sub-block for each way, and the size of this single valid-bit bank is \((2^s \times m) \times n\), where \(s\) denotes the set bits of the cache, \(w\) denotes the word offset bits of one block, and \(m\) is the sub-block size. Therefore, the valid-bit bank can be concurrently accessed while the MRU table is fetched before enabling memory subarrays. Additionally, the MRU table is maintained when the way-prediction misses or the cache misses. The way-predictor also sends \(n\) enabled signals to activate their own ways with tag-subarrays and data-subarrays according to the valid bits, and updates the status of the valid-bit bank when the cache misses.

2. Operations of WPD-V Cache

The main difference in operation between a conventional WPD cache and a proposed WPD-V cache is their prediction strategies, which means the WPD-V cache performs its way-prediction based on valid bits and the original MRU information. The operations of the WPD-V cache are described as follows:

1. While a set of the cache is referred to, 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 when 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 from one of the \(n\) data-subarrays 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 other 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 other ways with valid bits “1” should be enabled at the first cycle instead of the second cycle. Of course, no tag-subarrays and data-subarrays need to be enabled if these valid bits of the other ways are also “0”.
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 the valid-bit bank will be maintained.
3. Overheads

There mainly exist access time overhead and energy consumption overhead when the MRU table and the valid-bit bank are accessed, additionally, the hardware cost also needs to be taken into account. We know that the basic memory cells of SRAM are constructed of flip-flops; however, the power consumption of SRAM also includes energy dissipation to pre-charge bit lines, select word lines, and activate sense amplifiers of RAM cells. If the valid-bit bank is an SRAM memory table, then the extra energy dissipation caused by the valid-bit bank will be incurred at each cache access, and this energy dissipation $E_{vld}$ approaches $(n/m) \times (E_{data}/8)$ for an $n$-way WPD-V cache with sub-block size $= m$. Therefore, the valid-bit bank, like the MRU table, is also implemented by directly using flip-flops to store these valid bits rather than SRAM, so as to reduce the access time overhead and energy overhead (Inoue et al., 2000). Naturally, this structure requires a decoder and some flip-flops of size $\left(2^{\log_2(m)}/m \right) \times n$ to construct the valid-bit bank, which increases a lot of hardware cost. Due to concurrent accesses of the MRU table and valid-bit bank, the proposed WPD-V cache adds almost no extra access time compared with the conventional WPD cache. Since the main energy dissipated in flip-flops occurs at the clock transiting time, deactivating the clock of flip-flops for read operations can reduce the power consumption. Therefore, the energy consumption overhead to read the MRU table and valid-bit bank will be insignificant compared with the energy dissipated in the SRAM memory cells. On the other hand, the update number of the MRU table and valid-bit bank is very small due to the high prediction-hit rate and low miss rate, and thus the energy dissipation to update the MRU table and valid-bit bank also can be ignored. Besides, the way-predictor needs some combinational logic gates to decide which enabled signals should be sent according to MRU bits and valid bits, which may cause a little time delay. However, for the circuit implementation using logic gate arrays such as CPLD devices, this delay time can be neglected compared with the time to access SRAM memory cells.

4. Performance and Energy Evaluation

Based on the operations of the proposed WPD-V cache, and ignoring the access time and energy overheads caused by the MRU table and valid-bit bank, the average energy dissipation ($E_{C(WPD-V)}$) and average access time ($T_{AS(WPD-V)}$) for an $n$-way WPD-V cache can be respectively modified from Eqs. (4) and (5) in Section II as the following equations (Chen and Chiang, 2005):

$$E_{C(WPD-V)} = H \times PR \times (E_{tag} + E_{data})$$
$$+ H \times (1 - PR) \times [VR_1 + VR_2]$$
$$\times (n - 1)] \times (E_{tag} + E_{data}) + (1 - H)$$
$$\times [VR_{m1} + VR_{m2} \times (n - 1) + 1]$$
$$\times (E_{tag} + E_{data}),$$

(7)

$$T_{AS(WPD-V)} = [H \times PR + H \times (1 - PR) \times (1 - VR_1)]$$
$$+ [H \times (1 - PR) \times VR_1] \times 2 + (1 - H)$$
$$\times (P + 1 + VR_{m1})$$
$$= H + (1 - PR) \times VR_1 + (1 - H)$$
$$\times (P + 1 + VR_{m1}) \text{ Cycles},$$

(8)

where $VR_1$ and $VR_2$ are the valid-bit presence rates of the MRU way and other ways when the way-predicting misses but the cache hits, respectively, and $VR_{m1}$ and $VR_{m2}$ are the valid-bit presence rates of the MRU way and other ways when the cache misses, respectively. If the extra energy dissipated in the MRU table and valid-bit bank is also considered, then Eq. (7) would be modified as:

$$E_{C(WPD-V)} = H \times PR \times (E_{tag} + E_{data} + E_{MRU} + E_{vld})$$
$$+ H \times (1 - PR) \times [(VR_1 + VR_2 \times (n - 1))]$$
$$\times (E_{tag} + E_{data}) + 2 \times (E_{MRU} + E_{vld}) + (1 - H)$$
$$\times [(VR_{m1} + VR_{m2} \times (n - 1) + 1) \times (E_{tag} + E_{data})$$
$$+ 2 \times (E_{MRU} + E_{vld})].$$

(9)

When the sub-block size of a cache decreases, all valid-bit presence rates also decrease, which will help in improving average energy and average access time, especially for a cache with large associativity.

V. SIMULATION RESULTS

To verify 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 including the conventional WPD cache and proposed WPD-V cache, where 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 remodeling Dinero to trace various trace programs (Sites and Agarwal, 1988), where some trace programs belong to SPEC benchmark suites such as SPICE, GCC, and XLISP. Here, we use ‘cycle’ as a basic access time unit and $E_{data}$ as a basic energy unit.
respectively, and thus $E_{\text{tag}} = \frac{[\text{tag bits}]}{(\text{block size} \times 8)} \times E_{\text{data}}$. For example: if the cache block size is 32 bytes and the tag address bits are 20, then $E_{\text{tag}} = 0.078 E_{\text{data}}$. In our simulation, we measure some basic parameters of the way-predicting cache such as miss rate and prediction-hit rate etc., and then change the range of the sub-block size and associativity separately to observe the average energy dissipation and average access time for the conventional WPD cache and proposed WPD-V cache, respectively.

For realistic circuit analysis, we try to use a CPLD synthesis tool (MAX+plus II) employing MAX7000B circuit technology to design the way-predictor within our proposed 4-way WPD-V cache by VHDL codes, and practically verify the logic function and evaluate its access time. Due to the fact that the way-predictor circuit is mostly composed of combinational logical gates, we find that the propagation delay time from the inputs of MRU bits and valid bits to the enabled outputs is only about 10.8 ns.

### 1. Various Rates for Way-predicting Cache

The impact of sub-block size on miss penalty is well known. Here, we try to investigate the variety of various rates for the way-predicting cache with sub-block placement at different sub-block sizes. Tables 1 and 2 show the miss rate/ prediction-hit rate (MR, PR) and two valid-bit presence rates ($VR_1$, $VR_2$) of various benchmarks for different sub-block sizes at the fixed associativity (32-way), respectively. For all benchmarks, the miss rate (prediction-hit rate) will increase (decrease) as the sub-block size decreases; only the decrement of the prediction-hit rate is not obvious. For most benchmarks, two valid-bit presence rates significantly decrease as the sub-block size decreases, and the average number of the charged ways depends on $VR_1$ and $VR_2$. Consequently, using a valid-bit pre-decision approach will help in improving energy dissipation if cache ways are activated according to the status of valid bits. For example,

<table>
<thead>
<tr>
<th>Sub-block size</th>
<th>Parameters</th>
<th>FORA</th>
<th>GCC</th>
<th>IVEX</th>
<th>XLISP</th>
<th>SPICE</th>
<th>UE02</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Byte</td>
<td>MR</td>
<td>0.072</td>
<td>0.070</td>
<td>0.145</td>
<td>0.021</td>
<td>0.037</td>
<td>0.172</td>
</tr>
<tr>
<td></td>
<td>PR</td>
<td>0.754</td>
<td>0.769</td>
<td>0.732</td>
<td>0.762</td>
<td>0.829</td>
<td>0.681</td>
</tr>
<tr>
<td>2 Bytes</td>
<td>MR</td>
<td>0.068</td>
<td>0.068</td>
<td>0.141</td>
<td>0.020</td>
<td>0.036</td>
<td>0.165</td>
</tr>
<tr>
<td></td>
<td>PR</td>
<td>0.757</td>
<td>0.770</td>
<td>0.734</td>
<td>0.762</td>
<td>0.829</td>
<td>0.687</td>
</tr>
<tr>
<td>4 Bytes</td>
<td>MR</td>
<td>0.060</td>
<td>0.067</td>
<td>0.125</td>
<td>0.020</td>
<td>0.032</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>PR</td>
<td>0.764</td>
<td>0.771</td>
<td>0.747</td>
<td>0.763</td>
<td>0.833</td>
<td>0.707</td>
</tr>
<tr>
<td>8 Bytes</td>
<td>MR</td>
<td>0.040</td>
<td>0.037</td>
<td>0.078</td>
<td>0.011</td>
<td>0.021</td>
<td>0.093</td>
</tr>
<tr>
<td></td>
<td>PR</td>
<td>0.781</td>
<td>0.798</td>
<td>0.790</td>
<td>0.770</td>
<td>0.842</td>
<td>0.752</td>
</tr>
<tr>
<td>16 Bytes</td>
<td>MR</td>
<td>0.026</td>
<td>0.022</td>
<td>0.047</td>
<td>0.006</td>
<td>0.013</td>
<td>0.072</td>
</tr>
<tr>
<td></td>
<td>PR</td>
<td>0.793</td>
<td>0.812</td>
<td>0.817</td>
<td>0.773</td>
<td>0.849</td>
<td>0.782</td>
</tr>
<tr>
<td>32 Bytes</td>
<td>MR</td>
<td>0.017</td>
<td>0.013</td>
<td>0.030</td>
<td>0.004</td>
<td>0.008</td>
<td>0.068</td>
</tr>
<tr>
<td></td>
<td>PR</td>
<td>0.799</td>
<td>0.819</td>
<td>0.831</td>
<td>0.774</td>
<td>0.853</td>
<td>0.798</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Sub-block size</th>
<th>Parameters</th>
<th>FORA</th>
<th>GCC</th>
<th>IVEX</th>
<th>XLISP</th>
<th>SPICE</th>
<th>UE02</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Byte</td>
<td>$VR_1$</td>
<td>0.249</td>
<td>0.712</td>
<td>0.272</td>
<td>0.242</td>
<td>0.309</td>
<td>0.237</td>
</tr>
<tr>
<td></td>
<td>$VR_2$</td>
<td>0.075</td>
<td>0.488</td>
<td>0.084</td>
<td>0.095</td>
<td>0.098</td>
<td>0.058</td>
</tr>
<tr>
<td>2 Bytes</td>
<td>$VR_1$</td>
<td>0.383</td>
<td>0.715</td>
<td>0.378</td>
<td>0.325</td>
<td>0.417</td>
<td>0.358</td>
</tr>
<tr>
<td></td>
<td>$VR_2$</td>
<td>0.149</td>
<td>0.491</td>
<td>0.148</td>
<td>0.135</td>
<td>0.162</td>
<td>0.118</td>
</tr>
<tr>
<td>4 Bytes</td>
<td>$VR_1$</td>
<td>0.599</td>
<td>0.727</td>
<td>0.578</td>
<td>0.495</td>
<td>0.584</td>
<td>0.547</td>
</tr>
<tr>
<td></td>
<td>$VR_2$</td>
<td>0.306</td>
<td>0.500</td>
<td>0.313</td>
<td>0.234</td>
<td>0.297</td>
<td>0.260</td>
</tr>
<tr>
<td>8 Bytes</td>
<td>$VR_1$</td>
<td>0.744</td>
<td>0.796</td>
<td>0.691</td>
<td>0.585</td>
<td>0.731</td>
<td>0.683</td>
</tr>
<tr>
<td></td>
<td>$VR_2$</td>
<td>0.481</td>
<td>0.798</td>
<td>0.456</td>
<td>0.310</td>
<td>0.477</td>
<td>0.414</td>
</tr>
<tr>
<td>16 Bytes</td>
<td>$VR_1$</td>
<td>0.861</td>
<td>0.888</td>
<td>0.826</td>
<td>0.732</td>
<td>0.841</td>
<td>0.827</td>
</tr>
<tr>
<td></td>
<td>$VR_2$</td>
<td>0.679</td>
<td>0.754</td>
<td>0.654</td>
<td>0.452</td>
<td>0.649</td>
<td>0.629</td>
</tr>
<tr>
<td>32 Bytes</td>
<td>$VR_1$</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td></td>
<td>$VR_2$</td>
<td>0.992</td>
<td>0.992</td>
<td>0.986</td>
<td>0.759</td>
<td>0.960</td>
<td>0.992</td>
</tr>
</tbody>
</table>
the GCC benchmark has a decrement of two valid-bit presence rates (especially $VR_1$) as the sub-block size decreases, and there are $0.5 \times (32 - 1) = 16$ ways charged on average at the sub-block size = 4 bytes when the valid bit of the MRU way is “0” according to Eq. (7). Therefore, the suitable sub-block size seems to be 8 bytes based on the tradeoff between the increment of miss rate and the energy reduction.

2. Average Energy Dissipation

For the WPD cache, the average energy dissipation of all benchmarks at the associativity $= 32$ shown in Fig. 4 has a small increment as the sub-block size decreases due to the decrement of the prediction-hit rate. Like IVEX and UE02 benchmarks with higher miss rates have a more significant increment than other benchmarks in energy when the sub-block size starts decreasing from 8 bytes. However, for most benchmarks, the average energy dissipation of the proposed WPD-V cache (shown in Fig. 4) 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. Because GCC has worse program locality, the variety of energy improvement is not obvious when the sub-block size starts decreasing from 8 bytes.

3. Average Access Time

Figure 5 shows the average access time of all benchmarks at the associativity $= 32$ for the WPD cache and WPD-V cache, respectively. By using the valid-bit pre-decision, some second accessed ways only need one cycle when the valid bit of the MRU way does not exist on a way-prediction miss; 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 the miss penalty; however, a higher miss rate is also incurred such that more access cycles are required when the sub-block size shrinks. Therefore, for the WPD-V cache, the reduction of average access time is not as significant as that of the average energy dissipation.

4. Energy-Delay Product

One merit found when evaluating the power consumption of a cache is the energy-delay product (ED product), where $ED = (average\ energy\ dissipation) \times (average\ access\ time)$ for each cache access. Fig. 6 shows the ED products of the WPD-V cache, and these values are normalized to that of the WPD cache. We find that the ED product of our proposed WPD-V cache can obviously reduce the ED product as the sub-block size decreases for all benchmarks.
except GCC. Especially, for large associativity with small sub-block size, the WPD-V cache can significantly improve the average ED product of all benchmarks by about 53% at the associativity = 32 and the sub-block size = 4 bytes compared with the WPD cache.

5. Energy and Access Time Improvement

Figures 7 and 8 are obtained by averaging the simulation results of all benchmarks for different associativities, indicating the improvement of the WPD-V cache over the WPD cache in average energy dissipation and average access time, respectively. The improved rate in terms of average energy dissipation ($IMR_{EC}$) is expressed by:

$$IMR_{EC} = \frac{E_{C(WPD)} - E_{C(WPD-V)}}{E_{C(WPD)}} \times 100\%$$

(10)

and the improved rate in terms of average access time ($IMR_{TAS}$) is:

$$IMR_{TAS} = \frac{T_{AS(WPD)} - T_{AS(WPD-V)}}{T_{AS(WPD)}} \times 100\%$$

(11)

For the WPD-V cache with large associativity and small sub-block size, the improved rate in terms of average energy dissipation has a significant increment. In our simulation, $IMR_{EC}$ can be high up to 47% at the associativity = 32 and the sub-block size = 4 bytes, and $IMR_{TAS}$ also obtains about 5.7% under the same condition. Even for the most used associativity = 4 and sub-block size = 4 bytes, our proposed WPD-V cache still shows about 18% improvement in average energy dissipation.

6. Energy Overhead from Valid-bit Bank

Because both way-predicting caches own their MRU tables with the same size, the energy overheads ($E_{MRU}$) from these MRU tables are the same. In this sub-section, we only consider the energy ($E_{vld}$) dissipated in the valid-bit bank to further evaluate how much energy impact can be incurred. If we assume that the valid-bit bank is implemented by SRAM, Table 3 shows the energy evaluation by averaging the simulation results of all benchmarks at high associativity = 32 under considering the energy from the valid-bit bank. We find that the average energy dissipation

<table>
<thead>
<tr>
<th>Cache types (32-way)</th>
<th>Sub-block size</th>
<th>1 byte</th>
<th>2 byte</th>
<th>4 byte</th>
<th>8 byte</th>
<th>16 byte</th>
<th>32 byte</th>
</tr>
</thead>
<tbody>
<tr>
<td>WPD-V (NO $E_{vld}$)</td>
<td></td>
<td>2.917</td>
<td>3.631</td>
<td>4.961</td>
<td>5.131</td>
<td>5.622</td>
<td>6.575</td>
</tr>
<tr>
<td>WPD-V (With $E_{vld}$)</td>
<td></td>
<td>7.841</td>
<td>6.299</td>
<td>6.177</td>
<td>6.012</td>
<td>6.271</td>
<td>7.133</td>
</tr>
<tr>
<td>Overhead</td>
<td></td>
<td>4.924</td>
<td>2.668</td>
<td>1.486</td>
<td>0.881</td>
<td>0.649</td>
<td>0.55</td>
</tr>
</tbody>
</table>

* Energy unit: $E_{data}$

Table 3  Energy evaluation for different cache types
including $E_{\text{vld}}$ of the WPD-V cache is still less than that of the conventional WPD cache at any sub-block size, and this WPD-V cache also can effectively improve the WPD cache on the average energy dissipation. For example, the proposed WPD-V cache has about 31% improvement over the WPD cache at sub-block size = 4 bytes. Moreover, the energy overhead of the valid-bit bank is inconspicuous only if the sub-block size does not become too small. Actually, a small sub-block size such as 1 byte is not realistic for most applications. If the valid-bit bank can be implemented by flip-flops rather than SRAM, the energy overhead of the valid-bit bank will be further reduced. Therefore, the energy impact from the valid-bit bank is insignificant compared with the energy dissipated in the SRAM memory cells.

VI. CONCLUSIONS

In this paper, focusing on the way-predicting cache with sub-block placement, a new way-predicting cache using valid-bit pre-decision is proposed to progressively improve the average energy dissipation and average access time of the conventional way-predicting cache. We also provide various performance and energy indices of the conventional and our proposed way-predicting caches as considerations in choosing sub-block size. Because the valid-bit presence rates will decrease as the sub-block size decreases, especially for large associativities, many unnecessary probed ways can be eliminated. In our simulation, our proposed WPD-V cache can achieve an about 37% average improvement in energy for all sub-block sizes at 32-way. Even at 4-way, its average improvement in energy still is about 14%. If the energy overhead from the valid-bit bank is also involved, the proposed WPD-V cache has about a 31% improvement over the WPD cache at sub-block size = 4 bytes. Obviously, the main benefit of the proposed WPD-V cache is to achieve a significant improvement in energy saving over the conventional WPD cache, with a reduced access time also. Therefore, this proposed WPD-V cache has a better ED product than that of the conventional WPD cache, and it can be applied to low-power cache designs with large associativity.

NOMENCLATURE

- $E_{\text{bit}}$: energy dissipated in bit lines
- $E_{\text{cache}}$: average energy dissipation of a set-associative cache
- $E_{\text{cell}}$: energy dissipation to access all tag RAM cells and data RAM cells
- $E_{\text{C(WPD)}}$: average energy dissipation of WPD cache
- $E_{\text{C(WPD-V)}}$: average energy dissipation of WPD-V cache
- $E_{\text{data}}$: energy dissipated in a data-subarray
- $E_{\text{decode}}$: energy dissipation to drive address bus and to decode memory address
- $E_{\text{EIO}}$: energy dissipation to drive external I/O pins
- $E_{\text{MRU}}$: energy dissipated in the MRU table
- $E_{\text{tag}}$: energy dissipated in a tag-subarray
- $E_{\text{vld}}$: energy dissipated in the valid-bit bank
- $E_{\text{word}}$: energy dissipated in word lines and activation of sense amplifiers
- $E_D$: product of energy and access time
- $H$: hit rate of cache
- $IMR_{\text{EC}}$: improved rate in average energy dissipation
- $IMR_{\text{TAS}}$: improved rate in average access time
- $MR$: miss rate of a way-predicting cache
- $PR$: prediction-hit rate of a way-predicting cache
- $m$: sub-block size
- $n$: cache associativity
- $P$: miss penalty cycles of cache
- $s$: set bits of a reference address
- $t$: tag bits of a reference address
- $T_{\text{AS}}$: average access time of a set-associative cache
- $T_{\text{AS(WPD)}}$: average access time of WPD cache
- $T_{\text{AS(WPD-V)}}$: average access time of WPD-V cache
- $VR_1$: valid-bit presence rate of the MRU way on a cache hit
- $VR_2$: valid-bit presence rate of the other ways on a cache hit
- $VR_{m1}$: valid-bit presence rate of the MRU way on other ways on a cache miss
- $VR_{m2}$: valid-bit presence rate of the other ways on other ways on a cache miss
- $w$: word bits of a reference address

REFERENCES


*Manuscript Received: Apr. 24, 2007
Revision Received: Apr. 07, 2008
and Accepted: May 07, 2008*