ウェイ適応型キャッシュの高エネルギ効率化のための デッドブロック早期追いだしポリシ

東方雄亮[†] 佐藤雅之[†] 江川隆輔[††,†††]
滝沢寛之[†,†††] 小林広明[††,†††]

1. はじめに

近年，キャッシュメモリ（以下，キャッシュ）の低消費エネルギ化が求められている．低消費エネルギを実現するためのキャッシュとして，ウェイ適応型キャッシュ[1)]が提案されている．ウェイ適応型キャッシュは，スレッドの性能維持に必要のないキャッシュウェイへの電源供給を停止することによって，キャッシュの低消費エネルギ化を実現する．しかし，キャッシュのデータ管理に LRU 置換ポリシを用いているため，再利用されないブロックであるデッドブロックがキャッシュを占有する場合がある．ウェイ適応型キャッシュは，これらの性能向上に貢献しないデッドブロックも含めて，性能維持に必要なキャッシュ容量を見積もる．そのため，デッドブロックを保持するためのキャッシュ領域において電力が浪費され，ウェイ適応型キャッシュのエネルギ効率が著しく低下する．

本研究では，ウェイ適応型キャッシュのさらなるエネルギ効率向上を実現するために，デッドブロックを考慮したテータ管理ポリシを提案する．提案ポリシでは，デッドブロックの早期追い出しによって，キャッシュを占有するこれらのブロックを削減し，電源供給を停止できるウェイ数を増加させる．その結果，ウェイ適応型キャッシュにおいてさらなるエネルギ効率向上が期待できる．

2. デッドブロックのキャッシュ占有

ウェイ適応型キャッシュでは，データ管理ポリシとして LRU 置換ポリシが用いられている． LRU 置換ポリシでは，下位の階層へ書き戻されるブロックを選択するための優先度の順序列をキャッシュセット毎に考える．8 ウェイセットアソシアティブキャッシュの優先度の順序列を図 1 に示す．キャッシュミスの場合，優先度が最も高い位置 (MRU) にアクセスされたブロックが挿入され，最も低い位置 (LRU) のブロックが下位のメモリ階層に書き戻される．キャッシュヒットの場合，アクセスされたブロックは MRU に昇格される．このような動作によって，最も長い間アクセスされていないブロックを優先的に追い出し，時間的局所性に基づくデータアクセスを高速化できる．

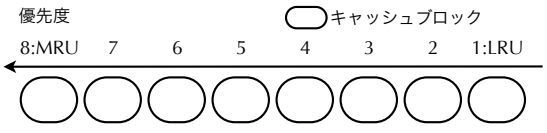

図 1 優先度の順序列

しかし，LRU 置換ポリシでは，将来再利用されないデッドブロックの場合でも，これらのブロックは MRU に置かれる．この結果，これらのブロックが追い出されるまでには長い期間が必要になるため，キャッシュ中におけるデッドブロックの割合が増大する．その結果，デッドブロックを保持するためのキャッシュ領域が浪費される．そのため，ウェイ適応型キャッシュにおいてエネルギ効率を高めるためには，再利用されるブロックが追い出されることを抑制しつつ，デッドブロックを早期に追い出す置換ポリシが必要である．

デッドブロックの中でも，挿入されてから一度も再利用されないデッドオンフィルブロックが特に多い．デッドオンフィルブロックを早期に追い出すポリシには，動的挿入ポリシ (以下，DIP)[2)] がある．DIP は，MRU と LRU に挿入位置を一定確率で切り替える二方向性挿入ポリシと LRU 置換ポリシを動的に選択することによってキャッシュミスを減らす置換ポリシである．しかし，いずれの置換ポリシが選択されても，デッドオンフィルブロックが MRU へ挿入される場合がある．また，DIP における昇格は LRU 置換ポリシと同様であるため，デッドブロックが MRU に昇格され，キャッシュに長時間保持される．以上より，DIP を適用したウェイ適応型キャッシュでは，デッドブロックのキャッシュ占有率をさらに削減する余地があると考えられる．

3. デッドブロック早期追い出しポリシ

デッドオンフィルブロックを早期に追い出すためには，挿入位置を LRU へ近づける必要がある．しかし，挿入位置を LRU へ過剰に近づけるとブロックが再利用される前に追い出される恐れが高まる．したがって，再利用されるブロックを保持しつつ，デッドオンフィルブロックを早期に追い出せるように優先度の順序列での適切な挿入位置を選択する必要がある．参照の局所性に基づき，ブロックが挿入された位置から LRU

† 東北大学大学院情報科学研究科
†† 東北大学サイバーサイエンスセンター
††† 科学技術振興機構戦略的創造研究推進事業

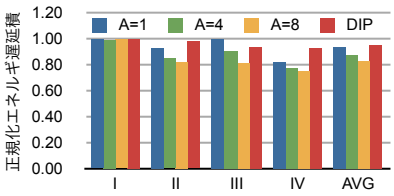

図 2 $R = \infty$ におけるエネルギ遅延積

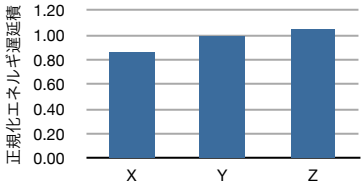

図 3 $A = 8, R = 1$ におけるエネルギ遅延積

に近づくにつれて，ブロックに対する再参照回数は少なくなる傾向にある．そのため，優先度の順序列のいかなる位置にブロックが挿入されても，初回再参照回数は LRU が最小となる．したがって，優先度の順序列における LRU への初回再参照回数を観測することによって，再利用されるブロックが最低優先度の位置まで降りてきていることを検知できる．そして，再利用されるブロックを保持しつつデッドオンフィルブロックを早期に追い出すために，最低優先度の位置以上に挿入位置を LRU に近づけないようにする．LRU における初回再参照回数を C，初回再参照回数の過不足を A とする時，LRU への初回再参照が多発している場合 ($C \geq A$) には，挿入位置を MRU 方向へ 1 つ移動する．一方で，LRU への初回再参照が全くない場合 ($C < A$) には，挿入位置を LRU 方向へ 1 つ移動する．以上のように適切な挿入位置の選択を行う．

また，昇格後に再利用されないデッドブロックをキャッシュから早期に追い出すためには，キャッシュヒットしたブロックを必要以上に昇格させないことが重要である．しかし，昇格を全くしないと再利用されるブロックも早期に追い出される恐れがある．したがって，昇格を停止する適切な再参照回数を選択する必要がある．本研究では，キャッシュヒットしたブロックを R 回の再参照まで MRU に昇格させ，$R+1$ 回目の再参照では昇格させないようにする．

4. 評価実験

本評価では，LRU 置換ポリシおよび DIP を用いたウェイ適応型キャッシュと，提案手法を用いたウェイ適応型キャッシュのエネルギ遅延積を比較する．なお，ベンチマークには SPEC CPU2006 を用いる．

まず，挿入方式を変更した場合におけるエネルギ遅延積を図 2 に示す．なお，各手法のエネルギ遅延積は LRU 置換ポリシを用いた場合のエネルギ遅延積によって正規化されている．また各ベンチマークは，再利用されるブロックの割合が多く，再参照の間隔が長い場合に I カテゴリへ，短い場合に II カテゴリへ分類されている．さらに，再利用されるブロックの割合が少なく，再参照の間隔が長い場合に III カテゴリへ，短い場合に IV カテゴリへ分類されている．図 2 より $A = 8$ の場合において，LRU 置換ポリシに対して 16.6 %，DIP に対して 12.4 %エネルギ遅延積が削減されていることがわかる．したがって，デッドオンフィルブロックの削減によって，ウェイ適応型キャッシュのエネルギ効率が向上することが明らかになった．なお，I カテゴリではエネルギ遅延積が全く削減されない．この原因として，デッドオンフィルブロックのキャッシュ占有率が低いことと，再参照の間隔が長いために挿入位置を LRU に近づけられないことが挙げられる．

昇格方式を変更した場合におけるエネルギ遅延積を図 3 に示す．なお，大部分のブロックは 1 回から 4 回までの再参照でデッドブロックになる．そのため時間的局所性より，1 回 MRU へ昇格されると，追い出されるまでに多くのブロックがデッドブロックになると考えられる．したがって，2 回目以降昇格をしない $R = 1$ の場合において昇格方式を変更した提案手法の評価を行う．エネルギ遅延積は $A = 8, R = \infty$ の場合における提案手法のエネルギ遅延積によって正規化されている．また，各ベンチマークは，エネルギ遅延積の削減効果に応じて 3 つのカテゴリに分類されている．X カテゴリはエネルギ遅延積が削減されるカテゴリであり，Y カテゴリは変化しないカテゴリである．また，Z カテゴリはエネルギ遅延積が増大するカテゴリである．図 3 より，昇格方式を変更することによって，X カテゴリでは 14.0 %エネルギ遅延積が削減されていることがわかる．しかし，Z カテゴリではエネルギ遅延積が 5.2 %増大する．昇格回数の制限が有効であるベンチマークの詳細な特徴分析と，動的な R の設定方法については今後の課題とする．

5. おわりに

ウェイ適応型キャッシュのエネルギ効率向上を目的として，デッドブロックをキャッシュから早期に追い出す置換ポリシを提案した．提案手法によってエネルギ遅延積を最大 50.6 %，平均 18.0 %削減できた．したがって，提案手法によるウェイ適応型キャッシュのエネルギ効率向上の可能性が示せた．

謝辞 本研究の一部は科学技術振興機構戦略的創造研究推進事業によるものである．

参考文献

1) Kobayashi, H., Kotera, I. and Takizawa, H.: Locality Analysis to Control Dynamically Way-Adaptable Caches, *ACM SIGARCH Computer Architecture News*, Vol. 33, No. 3, pp. 25–32 (2005).
2) Qureshi, M. K., Jaleel, A., Patt, Y. N., Steely, S. C. and Emer, J.: Adaptive Insertion Policies for High Performance Caching, *ACM SIGARCH Computer Architecture News*, Vol. 35, No. 2, p. 381 (2007).