社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2007－ARC－174 (20)
2007／8／2

キャッシュ階層動的切り替えによる低消費電力化

恩賀琢也† 佐々木敬泰†
大野和彦† 近藤利夫†

現在，プロセッサには高性能と低消費エネルギーの両立が求められている．特に，回路の微細化にともないキャッシュメモリ内でのリークエネルギーが年々増加しているため，これを削減する事が重要である．そこで本稿では，可変レベルキャッシュを提案する．可変レベルキャッシュは，動的にキャッシュの要求性能を判断し，あまり性能が必要ないと判断したときにキャッシュの半分をスリープモードに移行し１つ下位のレベルの Exclusive cache として動作する事で，消費電力の削減と性能の維持を両立させる手法である．本稿で行ったシミュレーション結果によると，可変レベルキャッシュは従来の DRI キャッシュと比較してエネルギー遅延積 (ED 積) で 12%程度性能向上することが分かった．

Reduce Power Dissipation with Variable Level Cache

Takuya Onga,† Takahiro Sasaki,† Kazuhiko Ohno† and Toshio Kondo†

Power dissipation is a major concern not only for mobile computing but also high performance computing, and achieving both low energy and high performance at the same time is required. Especially, it is important to reduce leakage energy consumed in a cache memory because power dissipation by leakage current is dominant factor in deep submicron technologies and a cache memory consists of a large number of transistors. This paper proposes Variable Level Cache to achieve both low energy consumption and high performance simultaneously. Variable Level Cache analyzes cache performance and if it detects that the running program does not need so large capacity of cache memory, half of the cache memory is put into standby mode, and is treated as a lower level exclusive cache. According to the simulation results, the performance of Variable Level Cache is about 12% superior to that of DRI cache in the energy-delay product.

1. はじめに

現在，ノートパソコン，PDA，携帯電話などのモバイル端末の高性能化にともない，消費電力が増大しバッテリによる駆動時間が短くなってきている．そこで，これらのモバイル端末の性能を落とすことなく低消費電力を実現する事が要求されている．

プロセッサで消費されるエネルギーは動的消費エネルギーと静的消費エネルギーに大別できる．動的消費エネルギーはトランジスタのスイッチングによって消費されるエネルギーである．一方，静的消費エネルギーはトランジスタの漏れ電流 (リーク電流) によって引き起こされ，スイッチングに関係なく消費されるエネルギーで，リークエネルギーとも言う．近年，回路の微細化にともなって，動的消費エネルギーが削減される反面，リーク電流およびリーク電流によって発生するリークエネルギーが増加する傾向にある．リークエネルギーはトランジスタ数に比例するため，プロセッサの高性能化に伴って増大したキャッシュシステムのリークエネルギーの削減が重要となっている．本研究では，キャッシュの性能を出来るだけ維持しつつ低消費エネルギー化を実現することを目的としている．

高性能と低消費電力の両立を目指すキャッシュシステムは様々なものが提案されているが，本研究では，DRI キャッシュ[1)2)] に着目する．DRI キャッシュは，キャッシュの要求性能を動的に判断しキャッシュ容量を変更する手法であるが，待機状態にするラインの電源を完全に落とす為，そのライン内のデータが失われ，ヒット率の低下，ひいてはプロセッサの性能の低下を招く．そこで，本稿では DRI キャッシュを改良した可変レベルキャッシュを提案する．可変レベルキャッシュは，動的にキャッシュの要求性能を判断し，あまり性能が必要ないと判断したときに，キャッシュの半分をスリープモードに移行し，擬似的に１つ下位のレベルの Exclusive cache として動作させる事で，消費電力の削減と性能の維持を両立させる手法である．

† 三重大学大学院工学研究科情報工学専攻
Graduate School of Engineering, Mie University

本稿の構成は以下の通りである．第2節では，DRIキャッシュを含む関連研究について述べる．次に，第3節では，動的にセット数を切り替えるリークエネルギー削減手法として可変レベルキャッシュを提案し，第4節では，性能と消費エネルギーの評価を行い，最後に第5節でまとめる．

2. 関連研究

これまでにキャッシュの様々なリークエネルギー削減手法が提案されてきた．これらの手法は通常状態と待機状態を切り替える単位で，大きく2つに分類する事ができる．

2.1 ライン単位の状態切り替え

1つ目は，ライン単位で通常状態と待機状態を切り替えるものである．[3)4)5)7)8)]

代表的な研究として，Drowsy Cache[4)5)]やウェイ予測キャッシュ[7)8)]が挙げられる．Drowsy Cacheは定期的に全てのキャッシュ・ラインへの電源電圧を下げ，アクセスが発生したラインに対してのみ電源電圧を回復する事でリークエネルギーを削減する手法である．ウェイ予測キャッシュは，アクセス開始前に最も最近アクセスされたウェイ情報 (MRU) を利用して，参照データが存在するウェイを予測し，選択的に活性化する事で低消費エネルギー化を行う手法である．

これらのようなライン単位で切り替える手法は，キャッシュの状態を細かく切り替えることが可能である一方，センスアンプ等のライン以外にかかる電力を削減できないといった問題がある．また，1つのセット内で通常状態のラインと待機状態のラインが混在する状態となる，すなわち連想度が落ちてしまうという問題がある．

2.2 バンク単位の状態切り替え

もう1つは，DRIキャッシュのようにキャッシュシステム内を複数のバンクに分割し，バンク単位で通常状態と待機状態を切り替えるものである．このタイプは，ライン単位で切り替える手法のように状態の細かな切り替えが出来ない反面，待機状態時にセンスアンプ等のライン以外の電力削減が可能となる．

バンク単位で切り替えを行う手法は，バンクを構成する際にセット単位，ウェイ単位のどちらでも可能であるが，本稿ではセット単位でバンクを構成する事を考える．ウェイ単位で切り替える事はライン単位で切り替えるのと同様に実効的に連想度を低下させてしまうが，セット単位の切り替えの場合，各セット内の連想度が通常状態と同じに保たれる利点があるからである．次項で説明するDRIキャッシュにおいても同様に扱う．

2.3 DRIキャッシュ

本項では，DRIキャッシュについて説明する．

図1にDRIキャッシュの概要図を示す．DRIキャッシュは，実行プログラムにおけるキャッシュへの要求性能を動的に検出し，それに伴いキャッシュ・サイズを変更することでリークエネルギーを削減する．具体的には，ある一定時間間隔でキャッシュミス数をカウントする．そして，ミス数がある閾値より小さい場合には，キャッシュ・サイズを縮小しても性能には大きな影響を与えないと判断する．一方，ミス数が閾値よりも大きい場合にはキャッシュ・サイズを増大して性能低下を防ぐ．キャッシュ・サイズを減らす場合はその時の容量の半分にし，逆にキャッシュ・サイズを増やす場合は倍にする．これを複数段階に分けて実装を行う事で，キャッシュ・サイズを必要に応じて変更する．例えば，動的に256KB，128KB，64KB，32KBのキャッシュ・サイズに変更することができる．キャッシュラインへのアクセスは，アドレスにマスクを掛ける事でキャッシュ・サイズの変化に対応させている．

また，DRIキャッシュは電源を切る部分のキャッシュを1つのバンクとして電源管理を行う事で，センスアンプやビット線などの，SRAMセル以外の回路の電源も切る事が出来るメリットがある．

DRIキャッシュでは，キャッシュ・サイズを縮小した場合，未使用領域のSRAMセルに対して電源電圧の供給を停止することでリークエネルギーを削減している．その際，縮小する部分への電力の供給を完全に停止し，データを破壊するため，その部分にあるデータを下位の記憶層へと書き戻す処理を行っている．また，キャッシュ・サイズによってデータの配置が異なるために，キャッシュ・サイズを増大させる際には，現在キャッシュに入っているデータを下位の記憶層へと書き戻している．DRIキャッシュはこれらの処理が性能へ悪影響を与えているという点で問題がある．

しかし，DRIキャッシュでは通常状態と待機状態を切り替える時に，そのライン内のデータを下位の記憶階層に書き戻しており，それが性能に影響を与えるという欠点がある．そこで，本研究ではDRIキャッシュを改良し，待機状態への切り替え時の書き戻しを抑制する可変レベルキャッシュを提案する．

3. 可変レベルキャッシュ

本節では，本研究で提案する可変レベルキャッシュについて述べる．

3.1 概　　要

可変レベルキャッシュは，DRIキャッシュと同様，ある一定時間間隔でキャッシュミス数をカウントする事で，キャッシュへの要求性能を動的に判断する．あまり

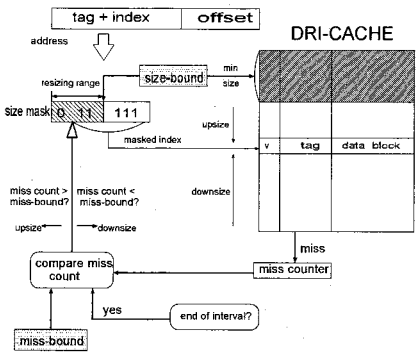

図 1　DRI の概要図

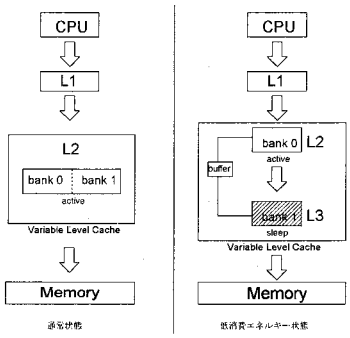

図 2　可変レベルキャッシュへのアクセス

性能（容量）が必要でないと判断されたときは，DRI キャッシュとは異なり可変レベルキャッシュは，キャッシュの半分を 1 つ下のレベルの Exclusive cache とし，その部分をスリープモードにする事で低消費電力を実現する手法である．図 2 に可変レベルキャッシュの動作の様子を示す．例えば，256KB の L2 キャッシュに可変レベルキャッシュを適用した場合，性能があまり必要でないと判断したときは，半分の 128KB は通常の L2 キャッシュとして，残りの半分の 128KB はスリープモードへと移行し L3 の Exclusive cache として動作する．すべてのラインがアクティブの時を通常モード，上記のようにキャッシュの半分をスリープモードにし，Exclusive cache として動作する時を低消費エネルギーモードとする．可変レベルキャッシュでは，低消費エネルギーモードから通常モードに切り替える時に，キャッシュ内のデータを下位層へと書き戻す点は DRI と同じだが，通常モードから低消費エネルギーモードに切り替える場合は，L3 となるライン内のデータを破壊しないので，下位層への書き戻しを行う必要がなくなるというメリットがある．Exclusive cache については次項で説明する．

スリープモードとは，電源の供給を完全に停止するのではなく，データの内容が破壊されない程度に電源電圧を下げた状態の事を言う．スリープモードにするとデータの内容が保存されるというメリットがある．また，電源供給を停止する場合よりはリークエネルギーの削減率が低くなるが，通常モードよりは大幅にリークエネルギーを削減出来るというメリットがある．しかし，スリープモードのラインへのアクセスがあった場合，通常モードに戻した後にアクセスを行うので，通常モードへのアクセスと比較してアクセス時間は長くなる．上述の通り，スリープモードで動作するとアクセス時間は増加してしまうが，それでも主記憶にアクセスするよりはアクセス時間が充分短い．そのため，主記憶の一つ上の記憶階層に可変レベルキャッシュを適用すれば，他の記憶階層に適用する場合より，性能の低下を抑える事が可能であると考えられる．

3.2 Exclusive Cache

Exclusive Cache は AMD の開発したキャッシュアーキテクチャで，L1 キャッシュと L2 キャッシュのデータを排他的にすることでキャッシュメモリを有効に利用する手法である．

以降の説明で，キャッシュの構成を L1 キャッシュは 128KB，L2 キャッシュは 256KB と想定する．

従来のキャッシュでは，すべてのデータはまず L2 キャッシュに格納されて，その後 L2 キャッシュから L1 キャッシュへとコピーされる．そのため，L1 と L2 の割り当てがあっても全体的なキャッシュ・サイズは L2 キャッシュに相当する 256K バイトであるといえる．それに対し，Exclusive Cache では，図 3 のように，L1 キャッシュから L2 キャッシュへと書き戻されるデータを一度バッファに移し，その後 L2 キャッシュから必要なデータをロードし，最後にバッファのデータを L2 キャッシュに書き戻している．このように，データを L1 と L2 との間で交換する事 (排他的に処理する事) で全体的なキャッシュ・サイズが 128+256=384K バイトとなり，キャッシュサイズを有効に活用できる．

4. 評　　価

4.1 評 価 方 法

可変レベルキャッシュの有効性を評価するため，DRI キャッシュ，可変レベルキャッシュの実行時間と消費エネルギーの評価を行った．また本研究では，性能の維持と低消費エネルギーの両立が目的であるため，実行時間と消費エネルギーの積 (ED 積) を求め，比較を行った．

まず，実行時間の評価については，それぞれの手法におけるプログラムの実行サイクル数を調べることで比較を行った．

次に，消費エネルギーについては，文献[6)9)] を参考

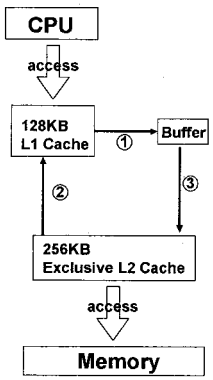

図 3　Exclusive Cache のアクセスの様子

に以下のように近似した．

まず，動的エネルギーの総和 DE_{total} はキャッシュアクセスで消費するエネルギーであるため，ライン当たりの動的エネルギー DE_{line} とアクセス回数 $Access$ の和で求められる．よって近似式は

$$DE_{total} = DE_{line} \times Access$$

となる．

次にリークエネルギー LE_{total} はライン当たりの平均リークエネルギー LE_{line} にキャッシュサイズ $CSize$ とプログラム実行サイクル CC の積で求まる．更に LE_{line} は，キャッシュ全体のうちスリープモードとなっているラインの割合 SR とスリープモードのライン当たりのリークエネルギー LE_{sline} の積に，通常モードのラインの割合 $(1-SR)$ と通常モードのライン当たりのリークエネルギー LE_{aline} の積を加えたものである．よって近似式は

$$LE_{total} = CC \times LE_{line} \times CSize$$

$$LE_{line} = SR \times LE_{sline} + (1-SR) \times LE_{aline}$$

となる．

キャッシュアクセスの際，キャッシュラインをスリープモードから通常モードに切り替えるエネルギーは

$$CE_{total} = CE_{line} \times BSize \times Access_{sline}$$

と表される．CE_{line} はキャッシュライン当たりのモード切替エネルギーであり，$BSize$ はモードを切替えるライン数 (バンクの大きさ)，$Access_{sline}$ は通常モードへの切替回数，つまり可変レベルキャッシュにおけるスリープモード時の L3 に当たる部分へのアクセスの回数である．

そして，これらの和を全体の消費エネルギー E_{total} とする．すなわち

$$E_{total} = DE_{total} + LE_{total} + CE_{total}$$

である．

尚，通常キャッシュや DRI キャッシュでは，スリープモードのラインは存在しないため，$CE_{total} = 0$ となる．CE_{total} は厳密には動的エネルギーに含まれるが，議論をしやすいように本稿では敢えて分けて扱っている．

表 1　評価に用いるエネルギー値 (単位：J)

DE_{line}	LE_{aline}	LE_{sline}	CE_{line}
1.51E-7	8.34E-13	1.32E-13	5.12E-11

表 2　SimpleScalar のキャッシュに関するパラメータ

キャッシュ容量	
L1 i-cache	32KB(64B/entry, 1way, 512entry)
L1 d-cache	32KB(64B/entry, 2way, 256entry)
L2 d-cache	512KB(64B/entry, 4way, 8192entry)
ヒット・レイテンシ	
L1 cache	1cycle
L2 d-cache	16cycle
主記憶	250cycle

キャッシュのエネルギーを評価する値は，文献[4)6)]を参考に，ラインサイズを 64B として表 1 のように定めた．

以上の式で求まった，実行サイクル数と消費エネルギーの積を求め比較を行う．

4.2 実験環境

可変レベルキャッシュに関して，性能と消費エネルギーの評価を行うため，学術・研究分野で広く利用されているマイクロプロセッサ・シミュレータである SimpleScalar[10)] を改造し，DRI キャッシュと提案手法である可変レベルキャッシュを実装した．ここで，DRI キャッシュ，可変レベルキャッシュは SimpleScalar の L2 の統合キャッシュに実装する事を想定し，プロセッサ構成は表 2 に示す．

本実験では，可変レベルキャッシュが，L2 が 512KB の通常モードと，L2 が 256KB，L3 が 256KB の Exclusive cache で動作する低消費エネルギーモードの 2 種類を動的に切り替えるのに対し，DRI キャッシュは 512KB の通常モードと，半分の電源を切り L2 を 256KB とする低消費エネルギーモードを動的に切り替えるものとする．また，DRI キャッシュ，可変レベルキャッシュ共に 100 万サイクル毎にキャッシュミスが 1000 回以上であれば通常モードに，1000 回未満であれば低消費エネルギーモードに切り替えるものとする．可変レベルキャッシュにおいて，スリープモードのラインへのアクセスがあり通常モードへと切り替える際のペナルティは 1cycle とする．

DRI キャッシュ，可変レベルキャッシュを組み込んだ SimpleScalar 上でベンチマークプログラムをアウト・オブ・オーダ実行し，それぞれの手法の性能を測定した．

ベンチマークプログラムは SPEC2000[11)] より，表 3 に示す 10 種類を使用した．

その際，プログラムの実行安定時の評価を行う為に，

表 3　使用するベンチマークプログラム

SPECint2000	164.gzip, 175.vpr, 176.gcc 181.mcf, 197.parser, 255.vortex, 256.bzip2
SPECfp2000	179.art, 183.equake, 188.ammp

プログラム実行開始より２０億命令実行後の２億命令を評価対象とした．

4.3 評価結果

実験によって得られたDRIキャッシュと可変レベルキャッシュの評価結果を図4，図5，図6に示す．図4は各ベンチマークでの実行時間の結果，図5は各ベンチマークでの消費エネルギーの結果，図6は図4と図5から得られた結果からED積を求めたものである．各図の横軸は使用したベンチマークを表しており，縦軸はそれぞれ，実行時間，消費エネルギー，ED積を標準のキャッシュの結果で正規化したものである．

上述の通り，DRIキャッシュ，可変レベルキャッシュの両方とも，キャッシュアクセスの局所性が高ければキャッシュヒット率が高くなり，低消費エネルギーモードで動作し，逆にキャッシュアクセスの局所性が低ければ，キャッシュミス率が高くなり，通常モードで動作する事が多くなる．本実験では，art，mcf，parser，vprにおいてはDRIキャッシュ，可変レベルキャッシュ共に終始通常モードで動作し，結果ED積は通常のキャッシュと同じになった．これらの場合においては，両手法とも通常のキャッシュと何ら変わりないため差は生まれない．その他の場合においては可変レベルの優位性が確認できた．その理由の１つとして，実行時間の差が挙げられる．DRIキャッシュは低消費エネルギーモードの時はキャッシュの容量の半分を使用しないのに対し，可変レベルキャッシュは容量の半分を低リークにしつつも有効に利用する事で，実行時間の延長を抑制できたと言える．また，equakeとvortexにおいて，可変レベルキャッシュは通常のキャッシュより実行時間が短くなるという結果が得られた．これは，通常のキャッシュの場合L2キャッシュでキャッシュミスをした場合，メモリにアクセスを行う．しかし，可変レベルキャッシュの低消費電力モードの場合，キャッシュ階層が増えた事で，L2でキャッシュミスをした場合でも，L3でヒットする場合があるためと考えられる．

可変レベルキャッシュが優れているもう１つの理由に，ヒット率の差が挙げられる．可変レベルキャッシュの方は，低消費エネルギーモードの時のキャッシュヒット率がDRIキャッシュと比較して高いため，低消費エネルギーモードで動作する確率も高くなる．また，DRIキャッシュは通常のキャッシュに比べて実行時間が長くなってしまい，リークエネルギーは実行時間に比例しているため，リークエネルギーが増大してしまう場合があった．

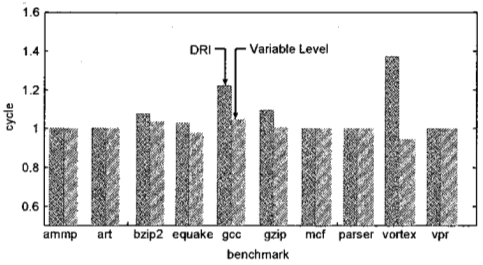

図 4　実行時間

可変レベルキャッシュとDRIキャッシュの差がもっとも現れたのはvortex，次いでgccであった．vortexについて，実行時間の内に低消費エネルギーモードの実行時間が占める割合は，DRIキャッシュで約25%，可変レベルキャッシュで約95%であった．また，可変レベルキャッシュでL3のラインへのアクセスの回数が他のベンチマークに比べて多かった．これはキャッシュミス回数がDRIキャッシュでは閾値より多少大きく，可変レベルキャッシュではヒット率が改善されて，閾値より小さくなったからと言える．これに伴ってDRIキャッシュではモードの切り替えが増えた結果，メモリへの書き戻しが増え，性能が低下した．またL2でのキャッシュミスが多発しているので，DRIはメモリアクセスが増える一方，可変レベルキャッシュはL3で性能低下を抑止できたと言える．gccについては，実行時間の内に低消費エネルギーモードの実行時間が占める割合はvortexに比べて差が大きくはならなかった．しかし，全てのベンチマークの中で，可変レベルキャッシュでのL3へのアクセス数は最も多い．つまり，DRIキャッシュではその分メモリアクセスが多かったので，その実行時間の差が結果に現れたと言える．しかしながら，可変レベルキャッシュのL3へのアクセスはL3のバンク全体を低消費エネルギーモードから通常モードに切り替える為，モード切替の消費エネルギーはとても大きい．そのため，通常のキャッシュに対して，DRIキャッシュ程ではないがED積が高くなってしまった．

5. おわりに

本論文では，新たなセット単位のキャッシュ・リーク削減手法として可変レベルキャッシュを提案した．可変レベルキャッシュは，キャッシュの要求性能を動的に判断し，高い容量を必要としない時は，キャッシュの半分の電源電圧をデータが壊れない程度に落とし，且つ，一つ下のレベルのExclusive Cacheとして利用する事で，低消費エネルギーと性能の維持を両立するものである．そして，実験によって従来のキャッシュ

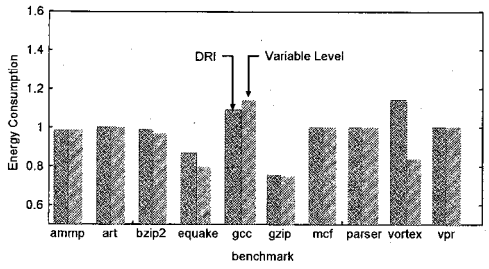

図 5 消費エネルギー

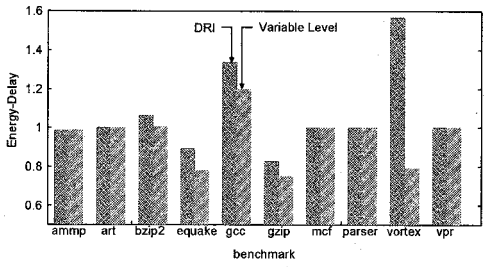

図 6 ED 積

と同等の性能を維持しつつ消費エネルギーの削減に成功し，従来のバンク単位 (セット単位) のキャッシュ・リーク削減手法である DRI キャッシュと ED 積で比較して優位性があることを示した．

今後は実際のハードウェアを設計し，詳細な評価を行っていく予定である．また，通常のキャッシュの方が消費エネルギーが少ない場合があるため，アプリケーションの特徴によって動的に可変レベルキャッシュを適用する手法の考案や，ウェイ予測キャッシュや Drowsy Cache のようなライン単位の切り替えを行う手法との比較評価を行っていきたい．

謝　辞

本研究の一部は科研費補助金 (19700042) の援助を受けている．

参 考 文 献

1) S.H.Yang, M.D.Powell, B.Falsafi, K.Roy, and T.N.Vijaykumar: "An Integrated Circuit / Architecture Approach to Reducing Leakage in Deep-Submicron High-Performance I-Caches.", Proc. of the 7th Int. Symp. on High-Performance Computer Architecture, pp.147-157, Feb.2001.

2) S.H.Yang, M.D.Powell, B.Falsafi, and T.N. Vijaykumar: "Exploiting Choice in Resizable Cache Design to Optimize Deep-Submicron Processor Energy-Delay.", Prc. of the 8th Int. Symp. on High-Performance Computer Architecture, pp.151-161, Feb.2002.

3) S.Kaxiras, Z.Hu, and M.Martonosi, "Cache Decay: Exploiting Generational Behavior to Reduce CacheLeakage Power", Proc. of the 28th Int. Symp. on Computer Architecture, pp.240-251, June 2001.

4) K.Flautner, N.S.Kim, S.Martin, D.Blaauw, and T.Mudge, "Drowsy Cache: Simple Techniques for Reducing Leakage Power", Proc. of the 29th Int. Symp on Computer Architecture, pp. 148-157, May 2002.

5) N.S.Kim, K.Flautner, D.Blaauw, and T.Mudge, "Drowsy Instruction Caches; Leakage Power Reduction using Dynamic Voltage Scaling and Cache Sub-bank Prediction", Proc. of the Int. Symp. on Microarchitecture, pp.219-230, Nov.2002.

6) 図子純平，冨山宏之，高田広章，井上弘士，"Drowsy キャッシュにおけるモード切替アルゴリズムの評価"，デザインガイア 2006，情報処理学会研究報告,2006-ARC-170, pp.37-41, 2006 年 12 月．

7) K.Inoue, T.Ishihara, and K.Murakami, "Way-Predicting Set-Associative Cache for High Performance and Low Energy Consumption", Proc. of the 1997 International Symposium on Low Power Design, PP.273-275, Aug. 1997.

8) 田中秀和，井上弘士，モシニャガ・ワシリー，"低消費電力を目的とした適応型ウェイ予測キャッシュとその評価"，電子情報通信学会技術報告，VLD2004-139, ICD2004-235(2003-3), pp.13-18, Mar. 2005.

9) 小宮礼子，井上弘士，モシニャガ・ワシリー，村上和彰，"キャッシュ・リーク電力削減アルゴリズムに関する定量的評価"，Proc. of the 17th Workshop on Circuit and Systems in Karuizawa, pp.235-240, Apr. 2004.

10) "SimpleScalar Simulation Tools for Microprocessor and System Evaluation", URL: http://www.simplescalar.org/.

11) "SPEC -Standard Performance Evaluation Corporation-", URL: http://www.spec.org/.