情報処理学会研究報告
IPSJ SIG Technical Report

Vol.2011-ARC-196 No.8
2011/7/27

温度を考慮した3次元積層LSI向け低消費エネルギーL2キャッシュの提案

阿部祐希[†1]　花田高彬[†1]
井上弘士[†2]　村上和彰[†2]

本稿では，温度を考慮した3次元積層L2キャッシュ向けバンク電源遮断による消費エネルギー削減手法について検討し，有効性評価を行う．3次元積層L2キャッシュは，垂直方向に隣接するコアの熱伝導のため，平面実装時のL2キャッシュと比較して高温となり，リーク消費電力が増大する．そこで我々は，積層L2キャッシュのバンク毎の温度分布の偏りに着目し，リーク消費電力を削減を実現する3次元積層キャッシュ向けのバンク電源遮断手法を検討している．本稿では，検討手法による消費エネルギー削減効果を評価した．ベンチマークプログラムを用いた評価の結果，最も効率的な電源遮断を実現できた場合で，エネルギー遅延積を28%削減可能であること示した．

1. はじめに

キャッシュメモリを大容量化するための手段の一つとして3次元積層の活用が注目されている[1)2)3)]．垂直方向にダイを積層し，TSV(Through-Silicon-Via) と呼ばれる層間金属柱などにより層間を直接接続することで，短い配線長を維持しつつ回路の大規模化が可能となる．また，製造プロセスが異なるダイを比較的容易に積層できるといった利点もある．たとえば，DRAMやMRAMといったロジックとは異なる製造プロセスを必要とする高集積メモリを大容量キャッシュメモリとして1チップに集積できる．

しかしながら，3次元積層キャッシュメモリの課題の一つとして熱によるリーク消費電力の増加が挙げられる．一般に，プロセッサコアなどのロジック部と比較して活性化頻度が低いL2キャッシュでは消費電力密度が低い．このため，平面実装時のL2キャッシュは比較的低温状態となる．これに対し，3次元積層キャッシュメモリでは，下層のプロセッサ・ダイの熱が伝導し，上層にあるキャッシュメモリが高温化する．リーク消費電力は温度に対して指数関数的に増加するため，特に3次元積層されたL2キャッシュではリーク消費電力の増加が深刻な問題となる[4)]．

これまでに様々な低リーク消費電力キャッシュが提案されてきた[5)6)]．これらの手法では，キャッシュ・メモリを部分的に停止する (電源供給を停止する) ことでリーク消費電力を削減する．前述したように，平面実装においてL2キャッシュは比較的低温である．したがって，多くの従来手法では時間的な電源制御に重きを置いたアプローチを採っている．しかしながら，3次元積層されたL2キャッシュにおいては，下層ダイにレイアウトされたプロセッサ・コアの温度が直接的に伝導する．一般に，コアの温度は実行対象プログラムの特性により変化するため，温度分布に偏りが生じる．そのため，3次元積層L2キャッシュにおいては，空間的な電源制御の適用が極めて重要になる．

そこで本稿では，3次元積層L2キャッシュのリーク消費電力削減を目的とし，温度分布の偏りを考慮した電源遮断方式を検討する．具体的には，4コアを搭載したマルチコア・プロセッサを前提とし，特性の異なるプログラムを同時実行した場合の温度分布を決定する．そして，最も効率的な電源遮断を実現できた場合の性能ならびに消費エネルギーを評価し，温度を考慮した電源遮断方式の重要性を明らかにする．

2. 3次元積層L2キャッシュとその問題点

2.1 3次元積層L2キャッシュ

近年，実行されるプログラムのワーキングセットが巨大化したため大容量のキャッシュメモリが必要とされている．キャッシュメモリを大容量化するためには，キャッシュメモリの面積を増加させなければならない．しかしながら，チップ面積は限られている．そこで，キャッシュメモリを大容量化するための手段の一つとして三次元積層による大容量キャッシュメモリの搭載が注目されている．三次元積層とは垂直方向にダイを積層させ，TSVと呼ばれる層間金属柱などで層間接続を行うことで，短い配線長を維持しつつ回路の大規模化が可能と

†1 九州大学大学院システム情報科学府
Graduate School of Information Science and Electrical Engineering, Kyushu University
†2 九州大学大学院システム情報科学研究院
Faculty of Information Science and Electrical Engineering, Kyushu University

なる．

2.2 3 次元積層 L2 キャッシュのリーク消費電力増加問題

一般にコアの温度分布はそのコアが割り当てられているプログラムによって異なる．したがって，マルチプログラム実行を考えた場合，コアひとつひとつの温度分布が異なる．このため，上層のキャッシュメモリの温度分布は，下層コアにおける実行プログラムに依存する．

ここで，L2 キャッシュメモリを積層した場合のリーク消費電力増加度について評価を行う．表 1 に本稿でのプロセッサ想定パラメータを纏める．大容量キャッシュメモリは，way ごとにバンク分割されているものとする．3 次元積層キャッシュメモリのバンクは 1 つの way を構成しているものとし，バンクの電源を 1 つ遮断すると 3 次元積層キャッシュメモリの容量が 128KB 減少し連想度が 1 低下するものとする．コアの温度分布は，文献[8] における温度評価をもとにする．ベンチマークプログラムはベンチマークセット SpecCPU2000[9] から 8 種類選んだ．文献[8] の温度評価はシングルプロセッサにおける温度評価であるため，本稿では隣り合わせにあるコア同士の熱伝導は無視し，文献[8] の温度分布が本評価でのコアの温度分布であるとする．そして，真上にあるバンクの温度もその温度に等しいとして評価を行う．本実験における下層コアと上層のバンクの位置関係を図 1 に示す．下層に 4 コアのマルチコアプロセッサが搭載されており，上層にキャッシュメモリが搭載されているとする．上層のキャッシュメモリは 16 のバンクに分けられているとする．下層のコアのブロックは文献[8] で評価をおこなっている粒度に等しく，上層のバンクの温度は，対応する下層コアのブロックにおける平均温度とする．その温度分布を用い，キャッシュメモリシミュレータ Cacti6.0[10] を用いてリーク消費電力を算出した．表 2 に本稿で用いた温度分布及び，そのコアの真上にあるバンクのリーク消費電力を示す．表 2 から，上層のバンクの温度分布は均一ではなく，それぞれのバンクによって，リーク消費電力が異なることが分かる．

図 3 はキャッシュメモリシミュレータ Cacti6.0[10] により実験した平面実装時と 3 次元積層時のリーク消費電力を示したグラフである．カッコ内は下層コアの実行プログラムを表す．平面実装時では，文献[8] より，プログラムによる温度依存性はほとんどなく，キャッシュメモリ内の温度はすべて 67(℃) として評価を行った．下層コアで実行しているプログラムは 175.vpr,179.art,188.ammp,301.apsi と仮定した．これらのプログラムは，コア内部の温度が平均的高温である．これらのプログラムで温度評価を行うことで，リーク電力の増大幅を見積もることが可能である．このような状況においては，キャッシュメモリを積層した場合，リーク消費電力が約 1.5 倍に増大すると評価する．

図 2 は，平面実装プロセッサである Sun microsystems の Niagara2 の消費電力内訳 (プロセステクノロジ 32nm) を示している[7]．図 2 からわかるように，L2 キャッシュメモリの消費電力はプロセッサ全体の 20%を占めているている．その中でも，L2 キャッシュメモリのリーク消費電力は全体の 11%も占めている．

したがって，平面実装時でもプロセッサ全体の消費電力のうち 11%も占めているため，3 次元積層した場合にリーク消費電力が 1.5 倍に増加することは問題である．そのため，3 次元積層キャッシュメモリのリーク消費電力を削減する方法が必要である..

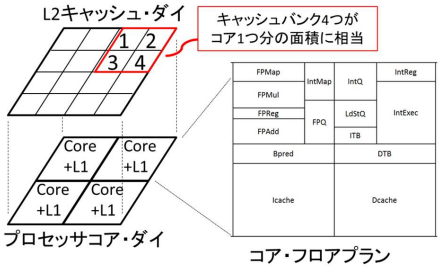


図 1　3 次元積層プロセッサ概略図

表 1　プロセッサ想定パラメータ

テクノロジサイズ	32nm
コア数	4
コアモデル	Alpha21364
動作周波数	1.2GHz
L1 キャッシュ	容量 32KByte，連想度 2 アクセスレイテンシ 1cc 単位アクセス消費エネルギー 0.39nJ
L2 キャッシュ	容量 2048KByte，連想度 16， アクセスレイテンシ 11cc，バンク数 16 単位アクセス消費エネルギー 1.46nJ
主記憶	容量 2GByte，バンク数 8 アクセスレイテンシ 191cc 単位アクセス消費エネルギー 7.92nJ

Vol.2011-ARC-196 No.8
2011/7/27

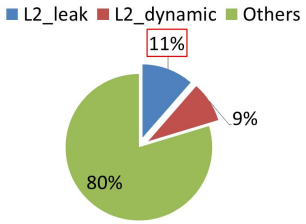


図 2　Niagara2 の消費電力内訳[7)]

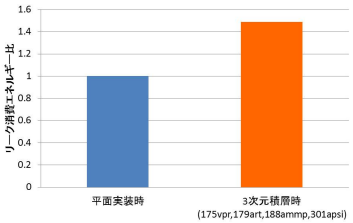


図 3　3 次元積層時と平面実装時のリーク消費電力の比較

表 2　L2 バンクの温度と消費電力

	バンク温度 (℃)[8)]				バンク消費電力 (mW)			
プログラム	1	2	3	4	1	2	3	4
164.gzip	64.8	80.3	71.4	67.8	49.6	96.6	60.6	52.9
171.swim	75.7	74.1	73.1	70.5	69.7	66.2	64.0	58.5
175.vpr	70.3	82.8	74.0	72.1	58.2	114	66.0	62.0
179.art	71.4	80.8	73.4	71.7	62.2	99.7	64.8	61.2
181.mcf	60.5	77.7	72.4	71.0	47.2	78.1	62.6	59.7
188.ammp	78.6	81.4	75.3	71.0	84.2	104	68.2	59.7
256.bzip2	64.9	81.1	70.6	71.9	49.7	102	58.9	61.5
301.apsi	79.4	82.0	76.1	74.0	89.9	108	70.4	66.1

3. 3 次元積層 L2 キャッシュ向けキャッシュサイジング

3.1 キャッシュサイジング

リーク消費電力を削減するための既存技術に，キャッシュメモリの一部の電源を遮断する手法がある[6)]．キャッシュメモリの電源の一部を遮断することで，キャッシュメモリの容量を動的に変更 (キャッシュサイジング) する．

この手法は，キャッシュメモリの一部の電源を遮断した際に増加する動的消費エネルギーと削減するリーク消費電力のトレードオフを考慮し，適切なキャッシュメモリの容量を選択する．このため，キャッシュメモリの消費エネルギー削減を行うことを可能としている．

3.2 3 次元積層キャッシュへのキャッシュサイジング適用

3 次元積層キャッシュメモリにキャッシュサイジングを適用しリーク消費電力の削減を行おうとした場合，問題になる点がある．それは，キャッシュメモリ内部でリーク消費電力が異なっていることである．第 2 節で述べたように，積層キャッシュメモリでは，温度分布が一様にならないからである．キャッシュサイジングによってリーク消費電力の削減効果を高めるため，電源遮断するバンクを選別する事は重要である．これは，電源遮断箇所のリーク消費電力が大きければ，キャッシュサイジングの効果は大きくなるためである．本稿では，3 次元積層キャッシュメモリを前提とした，下層コア温度を考慮したキャッシュサイジング適用による消費エネルギー削減効果について評価を行う．

4. 評　　価

4.1 消費エネルギーモデル

本稿にて仮定する 3 次元積層キャッシュメモリの有効性評価のため，メモリシステムの消費エネルギーモデルを構築する．

本稿にて仮定する 3 次元積層プロセッサのメモリシステム全体の消費エネルギー E_{total} は，式 (1) によって表わされる．一般に，消費エネルギーは動的な消費エネルギー $E_{dynamic}$ とリークな消費エネルギー E_{static} に分類できる．動的な消費エネルギー，ならびに，リークな消費エネルギーを詳細に分類したモデルを式 (2)，(3) に示す．動的成分，リーク成分それぞれにおいて，消費エネルギーは，L1 キャッシュメモリ ($E_{L1.d}$，$E_{L1.s}$)，L2 キャッシュメモリ ($E_{L2.d}$，$E_{L2.s}$)，主記憶の消費エネルギー ($E_{mem.d}$，$E_{mem.s}$) に分類できる．

$$E_{total} = E_{dynamic} + E_{static} \tag{1}$$

$$E_{dynamic} = E_{L1.d} + E_{L2.d} + E_{mem.d} \tag{2}$$

$$E_{static} = E_{L1.s} + E_{L2.s} + E_{mem.s} \tag{3}$$

次に，各メモリの動的消費エネルギーモデルを示す．式 (4)，(5)，(6) は，それぞれ，L1 キャッシュメモリ，L2 キャッシュメモリ，主記憶の動的消費エネルギーを示している．各メモリにおける動的消費エネルギーは，それぞれ，アクセス当たりの平均消費エネルギー ($E_{L1.access}$，$E_{L2.access}$，$E_{mem.access}$) と総アクセス回数 ($N_{L1.access}$，$N_{L2.access}$，$N_{mem.access}$) の積によって求められる．

$$E_{L1.d} = E_{L1.access} \times N_{L1.accesss} \tag{4}$$

$$E_{L2.d} = E_{L2.access} \times N_{L2.access} \tag{5}$$

$$E_{mem.d} = E_{mem.access} \times N_{mem.access} \tag{6}$$

続いて，各メモリのリーク消費エネルギーモデルを示す．式 (7)，(8)，(9) は，それぞれ，

L1 キャッシュメモリ，L2 キャッシュメモリ，主記憶のリーク消費エネルギーを示している．一般に，SRAM で構成される L1 キャッシュメモリ，ならびに，L2 キャッシュメモリのリーク消費エネルギーはリーク消費電力の時間積分によって表わすことができる．本稿では，モデル簡素化のため，各キャッシュメモリのリーク消費電力にはプログラム実行時の平均リーク消費電力を用いる．このため，L1 キャッシュメモリのリーク消費エネルギーは L1 キャッシュメモリ平均消費電力 $P_{L1.s}$ と実行時間の積によって表わすことができる．また，L2 キャッシュメモリのリーク消費エネルギーは L2 キャッシュメモリ・バンクの平均消費電力 $P_{i.bank.s}$ の総和と，実行時間の積によって算出できる．一方，DRAM で構成される主記憶のリーク消費エネルギーはリフレッシュに要するエネルギーと同義である．このため，リフレッシュ当たりの消費エネルギー $E_{mem.ref}$ と，主記憶のリフレッシュレート $T_{mem,ref}$ の積となる．

$$E_{L1.s} = P_{L1.s} \times T \tag{7}$$

$$E_{L2.s} = (\sum_{i=1}^{n} P_{i.bank.s}) \times T \tag{8}$$

$$E_{mem.s} = E_{mem.ref} \times \frac{T}{T_{mem.ref}} \tag{9}$$

4.2 消費エネルギーモデルによる定性的評価

L2 キャッシュバンクの電源遮断によるキャッシュサイジングは，L2 キャッシュメモリのリーク消費電力を削減を実現する一方で，L2 キャッシュメモリの容量が減少する．そのため，電源遮断されていない L2 キャッシュバンクへのアクセス数が増加し，また，L2 キャッシュミス率増加に伴い主記憶の動的消費エネルギー増加を引き起こす．さらに，L2 キャッシュミス率増加に伴い実行時間が増加するためリーク消費エネルギーが増加する可能性がある．これらを踏まえ，第節にて構築した消費エネルギーモデルを用い，3 次元積層キャッシュメモリにおける電源遮断によるキャッシュサイジングの消費エネルギー削減有効性評価を行う．評価対象は，L2 キャッシュメモリの動的，リーク消費エネルギー ($E_{L2.d}$，$E_{L2.s}$) と，主記憶の動的諸費エネルギー $E_{mem.d}$ の和である．

本評価において，実行時間 T は L2 キャッシュのミス率 MR_{L2} に依存する．実行時間 T を式 (10) によって表わす．本評価では，メモリアクセスが全体の実行時間の 20%を占めると仮定し評価を行う．なお，t_{L2} は L2 キャッシュメモリ・バンクへの平均アクセスレイテンシであり，t_{mem} は主記憶への平均アクセスレイテンシである．Memrate とは全体の実行時間のうちメモリアクセス時間の割合である．

表 3　L2 キャッシュの温度とリーク消費電力の関係[1)]

温度 (K)	340	350	360	370
リーク消費電力 (mW)	50.8	72.0	142	273

$$T = \frac{N_{L2.access}(t_{L2} + MR_{L2} \times t_{mem})}{Memrate} \tag{10}$$

本モデルによる評価では，L2 キャッシュメモリの温度がバンク毎に偏っているケースにおいて評価を実施する．バンクの温度とその時のリーク消費電力の値は文献[1)] にて示されている値を参考にした．結果，積層 L2 キャッシュの各バンクの温度とリーク消費電力を表 3 となった．なお，本評価では，積層 L2 キャッシュの総バンク数は 16 であり，バンク毎の温度の偏りは 4 バンク毎であると仮定した．即ち，340K，350K，360K，370K にて動作するバンク数がそれぞれ 4 バンクずつあると仮定する．

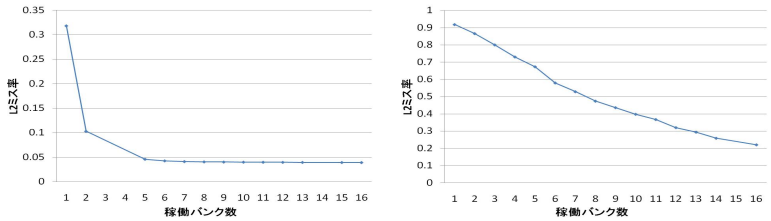
モデルによる消費エネルギー評価結果を図 5 に示す．x 軸の稼働バンク数を示している．稼動バンク数を減少させる際には、高温なバンクから電源遮断するものとする．

基準平面は消費エネルギー比 z=1 を表している．これは，あるプログラムを稼働バンク数 16 で動作さあせた場合の L2 キャッシュミス率が 0.20 である場合を基準とするためである．したがって，基準平面より下にあるケースでは，手法適用によって消費エネルギーが削減が可能である事を示している．

次に，図 5 に示しているプログラム特性曲線 1，2 の断面図を図 7，図 8 に示す．プログラム特性曲線とは，あるプログラムを実行した際の L2 ミス率と稼働バンク数の関係を仮定した曲線である．特性曲線 1 は図 4 の 164.gzip のように稼働バンクをある数まで減少させても L2 ミス率が増加しないプログラムを仮定する．特性曲線 2 は図 4 の 256.bzip2 のように稼働バンク数を減少させた場合に，L2 ミス率が上昇してしまうプログラムを仮定する．

図 7 は，特性曲線 1 を仮定した場合の断面図である，稼働バンク数を 13 まで減少させた際に消費エネルギーが最小になっている．これは，稼働バンク数を 13 まで減少させる際には，動的消費エネルギーの増加より，リーク消費電力削減効果が大きいため消費エネルギー削減効果が得られる．

図 8 は，特性曲線 2 を仮定した場合の断面図である．このようなプログラムを仮定した場合，稼働バンク数を減少させることによる消費エネルギー削減効果は得られない．この場合では，稼働バンク減少によるリーク消費電力削減効果よりも，実行時間の増加と L2 ミス



(a) 164.gzip (b) 256.bzip2

図 4 ベンチマーク毎のバンク数とミス率

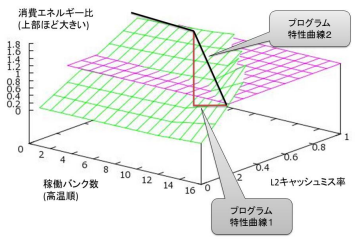


図 5 高温バンクから電源遮断した場合

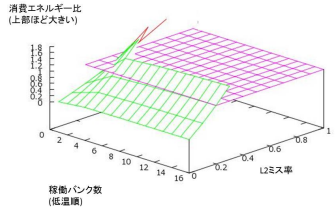


図 6 低温バンクから電源遮断した場合

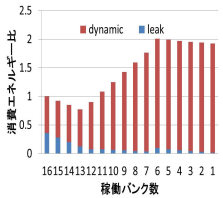


図 7 プログラム特性曲線 1 の断面

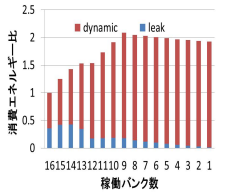


図 8 プログラム特性曲線 2 の断面

率増加による消費エネルギー増加が表れるためである．このようなプログラムにおいては，稼動バンク数を減少させず稼動する方が消費エネルギーの観点から望ましいことがわかる．図 7，図 8 より，実行するプログラムに応じて，消費エネルギーを最小化するバンク数を選択することが重要である事が示されている．

一方，稼働バンクを低温なバンクから電源を遮断した場合の消費エネルギー評価結果を表 6 に示す．稼働バンク数を減らした場合，基準平面より下にあるケースが少ない．これは，一部の電源を遮断した際に，電源遮断対象ではないバンクのリーク消費電力が大きいためである．これらのバンクのリーク消費電力は，少しばかりの実行時間増加による消費エネルギー増加が大きく，電源遮断によるリーク消費エネルギー削減効果を打消してしまう．したがって，図 5 と図 6 の比較より，電源遮断するバンクの優先度を決定するにあたり，温度を考慮することが重要である事が示されている．

4.3 ベンチマークプログラムを用いた有効性評価

本手法の消費エネルギー削減効果を見積もるため，ベンチマークプログラムを用いた定量的な評価実験を行う．本実験では，消費エネルギーモデルにシミュレータで得られた値を代入し，評価を行う．本評価における消費エネルギー評価対象は，メモリシステム (L1 キャッシュメモリ，積層 L2 キャッシュメモリ，主記憶) の消費エネルギーである．

評価に当たり，実行するベンチマークプログラム毎の実行時間 T，ならびに，各メモリへのアクセス回数 ($N_{L1.access}$，$N_{L2.access}$，$N_{mem.access}$) を求める必要がある．これらの値は，プロセッサシミュレーションより求める．マルチプロセッサシミュレータには M5[11] を用いた．また，メモリアクセス当たりの消費エネルギー ($E_{L1.access}$，$E_{L2.access}$，$E_{mem.access}$)，ならびに，キャッシュのリーク消費電力 P_{l1s}，$P_{i.bank.s}$ は，キャッシュメモリシミュレータ Cacti6.0[10] を用いて算出した．

本実験では，マルチプログラムを 1 億命令実行し，その実行時間を稼動区間と定義する．稼働区間内で最適な稼働バンク数，および高温なバンクが既知であるものとして評価を行う．また，稼働区間内では実行プログラムは変化しないものとする．比較対象は，全バンクが稼働状態 (即ち，稼動バンク数 16) とする．評価指標にはエネルギー遅延積 (ED 積) を用いる．本評価での ED 積とは，実行時間 T と消費エネルギー E_{total} の積をとったものである．この指標を用いることで，消費エネルギーだけでなく，実行時間に重みを置いた評価が可能である．各バンクの動作温度，ならびにリーク消費電力の仮定は，第 2.2 節と同様の方法で行い，表 2 に従うものとする．

本実験では，ベンチマークセット SpecCPU2000[9] から 8 種類選び，その中で実行する組合わせを選択した．プログラムの入力には train を用いた．プログラムによる L2 キャッシュのミス率とコア温度が本手法の消費エネルギー削減効果に大きな影響を及ぼすと考えられる．そこで，プログラムごとの L2 キャッシュの特性を評価を行いプログラムの分類を

Vol.2011-ARC-196 No.8
2011/7/27

表 4　評価実験に使用するプログラムの分類

	ワーキングセットサイズ	
	大	小
比較的温度が均一	179.art	175.vpr 188.ammp 301.apsi
局所的に高温	171.swim 181.mcf 256.bzip	164.gzip

表 5　実行プログラムの組み合わせ

プログラム	分類基準
175.vpr，179.art，188.ammp，301.apsi	比較的均一温度
164.gzip，171.swim，181.mcf，256.bzip2	局所的に高温
164.gzip，175.vpr，188.ammp，301.apsi	ワーキングセットサイズ小
171.swim，179.art，181.mcf，256.bzip	ワーキングセットサイズ大

行った．L2 キャッシュのミス率による分類のため M5 を用い，シングルコアでプログラムを実行した場合の稼働バンク数とミス率変化を解析した．その結果の一部を図 4 に示す．今回の分類では，稼働バンクを 8 バンクを減らした場合にキャッシュミス率が 25%を超えていないプログラムをワーキングセットサイズが小さいプログラムとして定義した．プログラムの分類を表 4 に示す．この分類を元に決定した，本実験で用いたプログラムの組合わせを表 5 に示す．

ベンチマークプログラムを用いた評価実験結果を図 9 に示す．横軸はプログラムセット，縦軸に全バンク稼働状態での ED 積を 1 として正規化した ED 積比を示している．本実験においては，縦軸の ED 積比が低いほど良い結果であるといえる．なお，図 9 中の提案手法バー直上の数字は，ED 積が最小となったバンク数を示している．

評価実験結果より，本実験で行ったプログラムの組み合わせにおいて，すべて ED 積削減効果が得られた．最大で 28%の削減を達成した．削減効果が大きかったプログラムの組合わせは，表 5 におけるワーキングセットサイズが大きいプログラムの組合わせと，温度が局所的なプログラムの組合せであった．前者において ED 積比が減少した理由は， 多くのバンクの電源を遮断しても L2 ミス率が増大せず，また実行時間の増加幅が小さかったためである．後者において ED 積比が減少した理由は，高消費エネルギーであるバンクの電源を遮断したことで，大幅な消費エネルギー削減効果が得られたためである．一方，

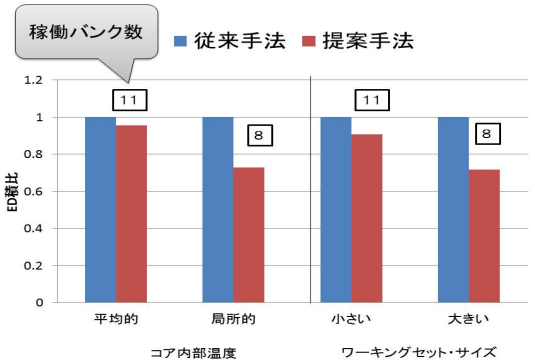


図 9　ED 積

ワーキングセットサイズが小さいプログラムの組合せでは，ED 積削減効果が小さかった．これは，稼働バンクを減少させた際に，L2 ミス率の増加が早い段階で現れ，実行時間が増加したためであると考えられる．評価実験の結果より，3 次元積層 L2 キャッシュにおいて，L2 キャッシュの温度を考慮したキャッシュサイジング手法は有効であると評価できた．今後，バンクの電源遮断アルゴリズムを考案する必要がある．なお，本評価では，バンク温度については下層コアと同温度であり，かつ時間的に変化しないと仮定している．温度はプログラム実行中に変化していくものであるため，温度について今後詳細に評価する予定である．

5. おわりに

3 次元積層技術を用いたキャッシュメモリを積層したプロセッサは，配線長を維持しつつキャッシュメモリの大容量化を実現できるとして，近年注目されている．しかしながら，垂直方向に隣接するコアの熱が伝導するため，3 次元積層キャッシュメモリは高温に伴うリーク消費エネルギーの増加が課題となる．

キャッシュメモリのリーク消費エネルギー削減を実現する既存手法として L2 ミス率を考慮したキャッシュサイジングが過去に研究されている．3 次元積層キャッシュメモリでは，温度によるリーク電流が支配的になる可能性が高い．このため，リーク電流が大きい高温な部分を優先的に電源遮断することで，より高いリーク消費エネルギー削減効果が期待できる．

そこで，我々は 3 次元積層キャッシュメモリを前提とした消費エネルギー削減を実現する電源遮断手法を提案する．本稿では，提案手法の有効性評価を実施した．初めに，モデルによる評価を実施し，3 次元積層キャッシュメモリにバンク温度の偏りを考慮した電源遮断を行うことで消費エネルギー削減が達成可能である事を示した．次に，ベンチマークプログラムを用いた評価を実施した．3 次元積層キャッシュメモリの温度分布はコアの実行プログラムによって異なる．実行プログラムに応じたバンク電源遮断によって最大 28%の ED 積削減が狙える事を示した．

今後の予定として，バンク電源遮断アルゴリズムの考案を念頭においた，実行プログラム毎の詳細な温度解析を実施する．

謝辞　日頃から御討論頂いております九州大学安浦・村上・松永・井上・アシル・杉原研究室ならびにシステム LSI 研究センターの諸氏に感謝します．本研究は主に九州大学情報基盤研究開発センターの研究用計算機システムを利用しました．なお，本研究は，独立行政法人新エネルギー・産業技術総合開発機構 (NEDO) 若手グラントの支援による．

参 考 文 献

1) Black, B., Annavaram, M., Brekelbaum, N., DeVale, J., Jiang, L., Loh, G., McCaule, D., Morrow, P., Nelson, D., Pantuso, D. et al.: “Die stacking (3d) microarchitecture” (2006).
2) Li, F., Nicopoulos, C., Richardson, T., Xie, Y., Narayanan, V. and Kandemir, M.: “Design and management of 3D chip multiprocessors using network-in-memory”, *Computer Architecture, 2006. ISCA’06. 33rd International Symposium on*, IEEE, pp.130–141 (2006).
3) Kim, J., Nicopoulos, C., Park, D., Das, R., Xie, Y., Narayanan, V., Yousif, M. and Das, C.: A novel dimensionally-decomposed router for on-chip communication in 3D architectures”, *ACM SIGARCH Computer Architecture News*, Vol.35, No.2, pp.138–149 (2007).
4) Li, P., Deng, Y. and Pileggi, L.: “Temperature-dependent optimization of cache leakage power dissipation”, *Computer Design: VLSI in Computers and Processors, 2005. ICCD 2005. Proceedings. 2005 IEEE International Conference on*, IEEE, pp. 7–12 (2005).
5) Kaxiras, S., Hu, Z. and Martonosi, M.: “Cache decay: exploiting generational behavior to reduce cache leakage power”, *ACM SIGARCH Computer Architecture News*, Vol.29, No.2, ACM, pp.240–251 (2001).
6) Yang, S., Powell, M., Falsafi, B. and Vijaykumar, T.: Exploiting choice in resizable cache design to optimize deep-submicron processor energy-delay, *High-Performance Computer Architecture, 2002. Proceedings. Eighth International Symposium on*, IEEE, pp.151–161 (2002).
7) Li, S., Ahn, J., Brockman, J. and Jouppi, N.: “McPAT 1.0: An Integrated Power, Area, and Timing Modeling Framework for Multicore Architecture”, *HP Labs* (2009).
8) Kong, J., John, J., Chung, E., Chung, S. and Hu, J.: “On the Thermal Attack in Instruction Caches”, *IEEE Transactions on Dependable and Secure Computing* (2009).
9) Henning, J.: “SPEC CPU2000: Measuring CPU performance in the new millennium”, *Computer*, Vol.33, No.7, pp.28–35 (2002).
10) Muralimanohart, N., Balasubramonian, R. and Jouppi, N.: “Optimizing nuca organizations and wiring alternatives for large caches with cacti 6.0”, *Microarchitecture, 2007. MICRO 2007. 40th Annual IEEE/ACM International Symposium on*, IEEE, pp.3–14 (2007).
11) Binkert, N., Dreslinski, R., Hsu, L., Lim, K., Saidi, A. and Reinhardt, S.: “The M5 simulator: Modeling networked systems”, *Micro, IEEE*, Vol.26, No.4, pp.52–60 (2006).