様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成　３０　年　　８　月　２２　日現在

機関番号：　１２６０１
研究種目：　研究活動スタート支援
研究期間：　2016～2017
課題番号：　１６Ｈ０６６７７

研究課題名（和文）高電力効率なビッグデータ処理 のためのメモリシステム最適化

研究課題名（英文）Memory System Optimzation for Energy Efficient Big Data Processing

研究代表者
有間　英志（Arima, Eishi）

東京大学・情報基盤センター・特任助教

研究者番号：５０７８０６９９

交付決定額（研究期間全体）：（直接経費）　　2,300,000円

研究成果の概要（和文）：本研究では、大規模データ処理を高い電力当たり性能で行うことを目的とし、その際に性能・電力の両面でボトルネックとなるメモリシステムをハードウェア・ソフトウェアの両側から最適化することを目指した。具体的には、(1)アドレス変換を考慮したキャッシュのデータ配置最適化、(2)ストレージクラスメモリを有するシステム上での電力パラメータの最適化というアプローチによってこれを行った。これら手法により、最大で数10パーセントの電力効率の改善が可能であることを定量的に示した。

研究成果の概要（英文）：To improve the energy efficiency of big data processing, this work focused on optimizing memory systems, the major performance/power bottlenecks during executing big data applications, by hardware/software-side approaches. In particular, this work is based on two novel approaches: (1) address translation aware cache management and (2) storage class memory aware power management. Consequently, it is quantified that a few tens of percent of energy efficiency improvement can be achieved by applying those methods.

研究分野：　計算機システム

キーワード：　メモリシステム　ビッグデータ　高電力効率化　キャッシュ　ストレージクラスメモリ

１．研究開始当初の背景

大規模グラフ処理に代表されるビッグデータ処理を行うアプリケーションは、人間社会に様々な新しいサービス・知見をもたらすことが期待され、近年大きな注目を集めている。そのため、これらアプリケーションの高速化へのユーザの要求は高い。一方で、近年の計算機システム、例えば大型計算機では、その膨大な消費電力も問題となっている。従って、これらアプリケーションを実行する際にも、高い電力効率(電力当たり性能)で処理することが要求される。

しかし、その高電力効率化の妨げとなっているものが、メモリシステムである。即ち、性能の面からは(1)局所性の低い間接参照を多用するために、TLB・キャッシュ上でミスが頻発し、さらに電力の面からは(2)大規模データを格納するために大容量のメインメモリが必要となり、その電力も膨大となるという問題がある。

２．研究の目的

本研究では、大型計算機等においてビッグデータ処理を高い電力効率で行うことを目的とし、その際に性能・電力の両面でボトルネックとなるメモリシステムをハードウェア・ソフトウェアの両側から最適化する。 具体的には以下のアプローチによる。

- 提案 1：アドレス変換を考慮したキャッシュのデータ配置最適化
- 提案 2：ストレージクラスメモリを有するシステム上での電力パラメータの最適化

提案１は性能向上を目的として、メモリシステム内のキャッシュに焦点を当て、そこでのデータ配置をアドレス変換時の PTE(Page Table Entry)アクセスに着目して行う。提案 2 は高電力効率化を目的として、近年普及が進みつつあるストレージクラスメモリを有するメモリシステムに焦点を当て、その上での電力パラメータの最適化を行う。これら提案技術を合わせることで、上記処理における電力効率の飛躍的な向上を目指した。

３．研究の方法

提案 1:

本研究では、データアクセスの局所性が低く、広範囲なメモリ領域にアクセスするというこれらアプリケーションの性質に着目し、これらを効率良く実行するためのキャッシュ制御方式について検討を行った。具体的には、キャッシュアクセスを通常のデータアクセスと TLB ミス後に生じる PTE アクセスとに分類し、これらのメモリアクセス特性の違いに基づいて、キャッシュ配置の優先度を決めるというものである。より具体的には、PTE アクセスは通常のデータアクセスよりも高い局所性を持つが、キャッシュミスを引き起こす大量のデータアクセスによって再利用される前に追い出されるという観測に基づいており、PTE を保持するキャッシュラインが長時間キャッシュラインに残る様に、キャッシュリプレイスメントアルゴリズムを改変するというものである。

これを実現するために、既存のプロセッサにて広く用いられている LRU アルゴリズムに改変を加えて評価を行った。LRU に基づくアルゴリズムは、Eviction Policy、Insertion Policy、Promotion Policy からなる。Eviction Policy とは、キャッシュミスが生じた場合にどのキャッシュラインをキャッシュから追い出すかを決めるポリシーであり、Insertion Policy とは、ミスを起こした後に下位のメモリ階層から送られてきたキャッシュラインを、LRU スタック上のどの位置に配置するかを規定するものであり、Promotion Policy とは、キャッシュヒットが起きた場合に、ヒットしたキャッシュラインを LRU スタック上のどの位置に配置するのかを規定するものである。本アルゴリズムでは、Eviction Policy は通常の LRU アルゴリズムと同様に LRU スタック上で最下位のもの(LRU アルゴリズムでは最も長くアクセスされていないキャッシュライン)を選択する。一方、Insertion Policy では、キャッシュラインが PTE を保持するかどうかで挿入位置を変更する。具体的には PTE を保持する場合には、LRU スタック上の最上位に配置し(LRU アルゴリズムでは最近にアクセスされたキャッシュライン)、そうでない場合には、LRU スタック上で N 番目に挿入する。そうすることで、キャッシュミスを引き起こす可能性の高いデータをキャッシュから早く追い出すことができ、PTE を保持するキャッシュラインをより長時間キャッシュ上に配置し続けることができる。一方で、Promotion Policy では通常の LRU アルゴリズムと同様に PTE を保持するかどうかに関わらず、LRU スタック上の最上位に配置する。これは、1 度再利用されたキャッシュラインはその後何度もアクセスされる可能性が高いというデータアクセスの性質に基づいている。

この様に提案アルゴリズムはシンプルであるため、ハードウェアの変更量も少なく実用的である。具体的には、キャッシュタグに PTE を保持するかどうかを判別するビットを 1 ビット追加し、このビットに応じて挿入位値を変える様に、既存の Insertion Position 指定様のハードウェアに信号を入れるだけである。

提案 2:

近年、大容量、低消費電力、不揮発といった特性を持つストレージクラスメモリが注目を浴びている。例えば、2017 年には Intel 社より 3D Xpoint memory と呼ばれるストレージクラスメモリが既に実用化されている。これらメモリは上述の観点からは有望である反面、速度の面からは既存の DRAM メモリ

よりも劣るため、既存の DRAM メモリとの併用が必要となる。

そこで、本提案では、その様なストレージクラスメモリを有するシステムを対象とし、その上での各コンポーネントに対する電力キャップ値等の電力パラメータを最適化することで、電力効率の向上を図る。一般的に、その様な電力パラメータの最適化は、性能ボトルネックとなったコンポーネントに対して、より大きな電力を与えることによって行われる。本研究では、当該システムにおいては、データサイズ S によってボトルネックコンポーネントが変化するという現象に着目し、それに応じてコンポーネント間で電力をシフトする手法を提案している。具体的な問題設定は以下の通りである。

$$\max \ Perf(S, P_{cpu}, P_{mem})$$
$$\text{s.t. } P_{cpu} + P_{mem} \leq P_{total}$$

即ち、目的は性能($Perf$)を最大化することであり、これは、データサイズ S、CPU への電力割り当て P_{cpu}、DRAM メモリへの電力割り当て P_{mem} の関数となる。これを、ノードに割り当てられた電力制約 P_{total} を満たす様に、CPU、DRAM メモリへの電力割り当てを最適化することで行う。ただし、ストレージクラスメモリの消費電力は比較的小さいため、ここでは考えないものとする。これら電力パラメータの最適化問題を解くことで、電力効率を大幅に向上させることができる。

4．研究成果

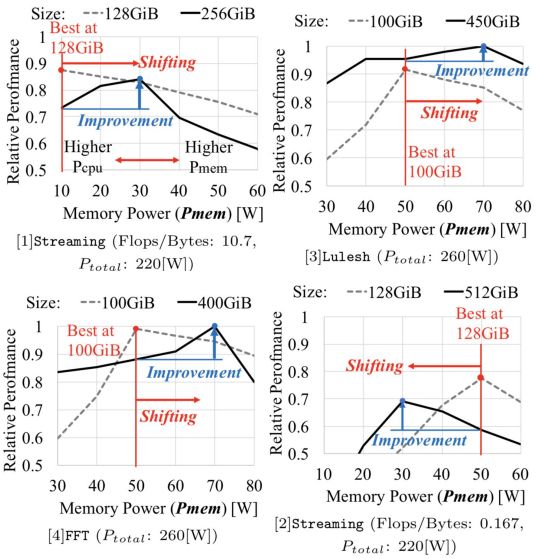
提案 1:

検討したアルゴリズムをオープンソースのプロセッサシミュレータ上に実装し、既存の LRU ベースのアルゴリズムとの比較を行った。評価の結果、最大で 40%程度、平均で 10%程度 PTE を保持するキャッシュラインの L1 データキャッシュ上でのヒット率を向上させることができた。一方、L2 キャッシュ上でも提案手法によって平均で 15%程度 PTE アクセスのヒット率を向上させることができた。一方で、通常のデータを保持するキャッシュラインに対するアクセスに関しても、L1 データキャッシュ、L2 キャッシュ上でのヒット率の低下はほとんど見られなかった。従って、再利用されるデータを犠牲にすることなく、PTE をキャッシュ上に優先的に配置できていることが分かる。結果として提案アルゴリズムを適用することで、数%の性能向上を確認している。

提案 2:

評価では、幾つかのワークロードに対して、与えられた総電力予算 P_{total} の下、$\{P_{cpu}, P_{mem}\}$ の組み合わせを変化させて性能を計測した。そのために、各コンポーネントに対する電力キャップ値を RAPL (Running Average Power Limit) インターフェースを利用して変更した。ワークロードとしては、Flops/Bytes 比を変更可能なシンセティックなストリーミングコードで評価し、さらに、FFT 、Lulesh といったアプリケーションも使って評価を行った。これらは、純粋なビッグデータアプリケーションではないものの、利用データサイズに応じて電力割り当てを最適化するという制御自体やここで得られた知見等はビッグデータ処理にも成り立つ上、展開することも容易である。

下図に評価結果を示す。図で横軸は P_{mem} を示しており、縦軸は相対性能を示している。ここで、相対性能とは、各々のデータサイズごとに、電力キャップなしで実行した性能を基準とし、電力キャップをかけた場合にその何割の性能が出るのかを示している。$P_{total} = P_{pcu} + P_{mem}$ と設定しているため、P_{mem} が大きくなればなるほど、P_{cpu} は小さくなる。



グラフに示す通り、最適な$\{P_{cpu}, P_{mem}\}$の割り当ては、データサイズに応じて、これら全てのワークロードで変化している。Streaming(Flops/Bytes:10.7)、FFT、Lulesh については、データサイズをスケールさせた際に、電力を CPU から DRAM メモリにシフトすることで、大幅な性能向上が見られる。これは、データサイズを大きくすることで、低速なストレージクラスメモリがより頻繁にアクセスされ、結果として CPU がボトルネックでなくなったからであり、従って、CPU は電力を必要とせず、メモリシステム側にシフトすることが得策となる。

一方、メモリインテンシブな Streaming (Flops/Bytes:0.167)はこれとは異なる振る舞いを示す。結果として、DRAM メモリは与えられた電力を効率的に性能へと変換できていないことが分かる。これは、データサイズが DRAM メモリよりも大きい場合には、頻繁にストレージクラスメモリがアクセスされるためである。それと比較して、データサイズが 128GiB の場合には、ストレージクラスメモリは滅多にアクセスされず、DRAM メモリ

がボトルネックとなるために、DRAM メモリにより多くの電力を与えるのが得策である。以上をまとめると、電力割り当てを決める際には、(1)CPU インテンシブなアプリケーションに対しては、データサイズに応じて CPU とメモリシステム側でボトルネックの移動が起きることに注意し、(2)メモリインテンシブアプリケーションに対しては、DRAM メモリとストレージクラスメモリのアクセス比率とデータサイズとの関係を考慮する必要があり、それを行うことで、最大で約 20%の電力効率の改善が可能であることが分かった。

５．主な発表論文等
（研究代表者は下線）

〔雑誌論文〕（計　0 件）

〔学会発表〕（計　4 件）

① Eishi Arima、 Toshihiro Hanawa、 Martin Schulz、“Toward Footprint-Aware Power Shifting for Hybrid Memory Based Systems” International Conference on Parallel Processing (ICPP)、 Poster Session、 2018
② 有間 英志、塙 敏博、ハイブリッドメモリを搭載するシステムにおけるデータサイズを考慮した電力制御、2018-HPC-164、No. 8、pp. 1-7、2018
③ Eishi Arima、 Hiroshi Nakamura、 "Page Table Walk Aware Cache Management for Efficient Big Data Processing" The eighth workshop on Big data benchmarks、 Performance Optimization、 and Emerging hardware (BPOE-8) (in conjunction with ASPLOS 2017)、 2017
④ Eishi Arima、"Near Memory Processing on Hybrid Memories" IPSJ SIG Notes 、 2017-ARC-224、 No.33、 pp.1-4、 2017

〔図書〕（計　0 件）

〔産業財産権〕

○出願状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
国内外の別：

○取得状況（計　0 件）

名称：
発明者：
権利者：
種類：
番号：
出願年月日：
取得年月日：
国内外の別：

〔その他〕

６．研究組織
(1)研究代表者

有間　英志　（ARIMA、 Eishi）
東京大学情報基盤センター・特任助教
研究者番号：50780699

(2)研究分担者
（　　）

研究者番号：

(3)研究協力者

Martin Schulz
Lawrence Livermore National Laboratory・Computer Scientist
(現 Technical University of Munich ・Professor)