様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ６ 月 １１ 日現在

機関番号：１２６０１
研究種目：挑戦的萌芽研究
研究期間：2013～2014
課題番号：２５５４００１８

研究課題名（和文）時空間上のデータ制御実行モデルの研究

研究課題名（英文）Computation Model of Spatiotemporal Data Control

研究代表者
中村　宏（NAKAMURA, HIROSHI）

東京大学・情報理工学（系）研究科・教授

研究者番号：２０２１２１０２

交付決定額（研究期間全体）：（直接経費）　　2,900,000円

研究成果の概要（和文）：コンピューティングの高性能化と低消費電力化を妨げる要因は演算や処理を行う部分ではなく、演算処理部と記憶部の間のデータ転送、および記憶部へのデータアクセスにある。この問題を解決すべく、データ移動と処理の「タイミング」、およびデータの「物理的な場所」を陽に制御する、新しい実行モデル「時空間上のデータ制御実行モデル」を提案し、このモデルに基づく実行最適化手法を研究した。多様なコンピューティングシステムに対する有効性を検討するために、提案する手法を、３次元積層VLSIシステム、高性能サーバシステム、ならびにセンサネットワークシステムに対して適用し、その有効性を確認した。

研究成果の概要（英文）：Bottlenecks of both performance and power exist not in the arithmetic calculation or computation but in the data transfer between memory and arithmetic unit or in memory access. To overcome this problem, new computation model of Spatiotemporal Data Control is proposed. This model can explicitly specify both the timing of data movement and computation and the physical location of data. This research also investigated how to optimize execution by using this new model. The proposed method is applied to wide variety of computing systems, including three dimensional integrated VLSIs, high performance server systems, and sensor network systems. The preliminary experimental results reveal that the proposed method can successfully improve performance or reduce power consumption.

研究分野：計算機アーキテクチャ

キーワード：コンピューティング　実行モデル　アーキテクチャ　低消費電力化

１．研究開始当初の背景

今日、VLSI 内部での実行から広域分散環境での実行に至るあらゆるコンピューティングにおいてその高性能化と低消費電力化を妨げる大きな壁は、演算や処理を行う部分ではなく、演算処理部へのデータ供給、すなわち演算処理部と記憶部の間のデータ転送と記憶部へのデータアクセスにこそ存在している。この問題は Memory Wall Problem あるいは Power Wall Problem として広く認識されている。たとえば、シリコンチップ上で考えた場合、今後集積度が上がりさらに多くの演算処理装置が搭載可能となっても、その半分以上は動作させることができず、シリコンチップが暗く(Dark Silicon)なってしまうと、文献（H.Esmaeilzadeh, et.al.:Dark Silicon and the End of Multicore Scaling, ISCA-38, 2011）では指摘されている。また、将来の高性能スーパーコンピュータにおいては、図１に示すように、FPU(演算装置)が消費する電力は全体の約 1/3 に過ぎず、Memory および Interconnect の消費する電力が全体の 2/3 に達すると予測されている。

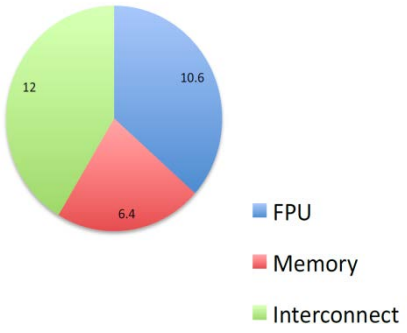


図１：将来のスーパーコンピュータの消費電力の内訳（文献：DOE Report, “Scientific Grand Challenges: Architectures and Technology for Extreme Scale Computing”, DoE, Dec. 2009, の Fig. 3.2 より引用）

これらの問題に対する従来の解決法は、遅くて遠いメモリからデータを供給するには長い時間と大きな電力を必要とするため、小容量高性能なキャッシュメモリをチップ上に搭載するというものである。しかし、ソフトウェアからはデータの「論理的な場所」しか指定できないため、当該データがキャッシュメモリに存在することを直接的には保証できない。低消費電力化技術としては、動作不要部の電源遮断を行う技術が有望であるが、電源遮断自体がエネルギーを消費するため、電源遮断期間を長く、しかも電源遮断頻度を減少させることが重要である。しかし、ソフトウェアからは処理の「順序」は制御できるが処理の「タイミング」は制御できないため、この問題を本質的には解決できない。センサネットワーク等に代表される広域分散環境においても、中継ノードは他ノードからの転送要求の「タイミング」が制御できないため、転送処理自体は短時間で終了するにもかかわらず常に転送要求を監視する必要があり、電源遮断ができない。また「物理的な場所」を陽に制御できないため消費電力の大きい遠距離通信が発生し、電力消費の増大を招いている。このように、ソフトウェアから処理の「順序」と「論理的な場所」しか見えない従来の実行モデルでは、上記問題を解決する直接的な手段を持ち合わせていない。

２．研究の目的

本研究は、背景で述べた問題を解決し VLSI システムから広域分散環境下に至るまで、社会が必要としているあらゆるコンピューティングの高性能化と低消費電力化を実現することを目的する。その解決方法として、データ移動と処理の「タイミング」、およびデータの「物理的な場所」をソフトウェアから陽に制御することが可能な新しい実行モデルとして「時空間上のデータ制御実行モデル」を提案し、このモデルに基づく実行最適化手法により当該問題を解決することを目指す。多様なコンピューティングシステムに提案手法を適用した場合の性能面と消費電力面での有効性を示すことも目的とする。

３．研究の方法

前述の目的を達成するために以下の２つを主たる検討項目として研究を行う。

・実行モデル

まず、コンピューティングの高性能化と低消費電力化を時空間上のデータ制御で実現するハードウェア機構を検討する。単一システムの実行では、性能に大きな影響を与えるメモリ階層を中心に検討する。従来のメモリはデータを忘れない素子として扱われていたが、低消費電力化を目指し、消費電力とデータ記憶期間をトレードさせることができるメモリ素子などの提案もされており、データを記憶する場所と時間を制御する必要がある。このような制御を実現するハードウェア機構を検討する。また、処理を空間的に移動させる場合には必要供給電力に変化は生じないが、処理を時間的に移動させる場合には必要供給電力も変わるので、この点も考慮する。広域分散処理ではデータ転送自身が性能上も消費電力上もボトルネックとなっているため、システム内のメモリ階層だけではなくデータ転送を時空間上で最適化可能なハードウェア機構も検討する。広域分散処理において、センサネットワークシステムにおいては、通信経路は静的に決定しているが、通信の頻度が動的に決定するものもある。これはつまり、通信の空間位置は不変だが通信の時間位置が変動することを意味する。そのような場合についても検討を加える。

次に、検討したハードウェア機構を十分に活用するために必要となるソフトウェアからの指示についての列挙とその体系化を行う。この体系化においては、データ制御の時間粒度と空間粒度と、ソフトウェアのどのレイヤがその指示を与えるべきか、の対応関係の体系化も行う。これにより、実現すべき時空間上のデータ制御に対し、ハードウェアが責任を持つべき事項とソフトウェアが責任を持つべき事項が整理され、ハードウェアとソフトウェアのインタフェースレイヤが規定される。

・実行最適化手法の検討と有効性検討：

実行モデルの検討で明らかになる実行モデル、および規定されるハードウェアとソフトウェアのインタフェースレイヤを活用することで、システムの高性能化・低消費電力化を実現する実行最適化手法を開発し、その評価を行う。実行モデルの考え方は多様なコンピューティングシステムに共通のものであり、有効性も多様なコンピューティングシステムに対して適用して明らかにすべきである。しかし、あらゆるコンピューティングシステム全てに対する適用実験は困難である。そこで、３次元積層 VLSI システム、高性能サーバシステム、ならびにセンサネットワークシステムを取り上げ、それらに提案手法を適用し、有効性を評価する。これらは空間的なスケールとしては、cm, m, km とオーダーが違うものであり、提案する手法の有効性を一般的に論じる際に適していると考えた。評価実験は、実システムを構築することは難しいため、評価シミュレータ・エミュレータを構築して行った。

４．研究成果

(1) ３次元積層 VLSI システム

TSV(Through Silicon Via)などの 3 次元積層技術の進歩により、複数のプロセッサコアと大容量のキャッシュメモリが搭載される VLSI の利用が可能となりつつある。特に LLC(Last Level Cachce)と呼ばれる最下層のキャッシュメモリを複数のコアが共有する VLSI システム構成が有望視されている。しかし、仮想記憶を管理するページテーブルのキャッシュの役割を果たすコピーを保持する TLB に関して、その TLB ミスが性能に与える影響については、これまで十分に考慮されてこなかった。TLB サイズの大幅な増加は今後期待できないとされている。これは、TLB はコンテクストスイッチの度にフラッシュする必要があり、巨大な TLB を用意してもそれを実行中のプロセスが使い切ることが難しいためである。そのため、LLC の大容量化が進むにつれて、LLC 上にはデータはあるのに、当該ページアドレスが TLB 上に存在しない（TLB ミス）ものの割合は増大する。

TLB ミス時には、当該ページテーブルエントリを LLC もしくはメインメモリから取得し、TLB 上にこれを格納する。必要なページテーブルエントリをより上位の階層に配置することできれば、TLB ミスによる性能ペナルティを削減することができ、性能低下を抑えられる。特に、メインメモリアクセスのオーバーヘッドは大きいため、LLC におけるページテーブルエントリのヒット率を向上させることは重要である。LLC は今後も大容量化が進むと期待されており、その大容量性をページテーブルエントリヒット率の向上に転嫁できる技術は重要となる。

そこで、大容量 LLC 上の全ラインに対し、ページテーブルエントリを保持するラインの存在割合を最適化することによって、TLB ミスペナルティを削減し性能向上を目指す。この制御を行うときの目的関数は（式 1）で与えられる。

$$\mathrm{Max}(Nd(\alpha)+Np(\alpha)) \quad \text{（式 1）}$$

ただし、$\alpha, Nd(\alpha), Np(\alpha)$，はそれぞれ、LLC におけるページテーブルエントリを保持するラインの存在割合、ページテーブルエントリ以外の通常のデータの LLC におけるヒット回数、ページテーブルエントリの LLC におけるヒット回数、を表す。αを大きくすれば、$Nd(\alpha)$は小さくなり、 $Np(\alpha)$は大きくなる。すなわちこの目的関数はこれら 2 種類のデータを包括する LLC 上の全てのデータのヒット回数の最大化を目的とする。

この様子を図 2 に示す。横軸はα（ページテーブルエントリを保持するラインの存在割合）を示し，縦軸は LLC におけるヒット回数を表わす。したがって、この最適化問題は、αが図 2 におけるα_{opt}になるように LLC のデータ配置の制御を行うことである。

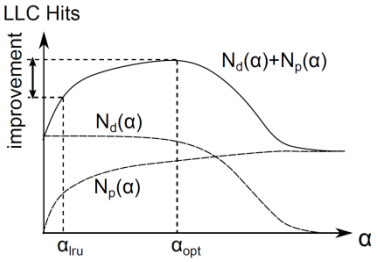


図２：ページテーブルの存在割合と目的関数との関係

以上が実行時のモデルであるが、このモデルに基づく時空間的最適化手法として、LLC 上でのページテーブルエントリとそれ以外のデータの配置場所を最適化する手法を提案する。具体的には、あるラインが新たに LLC へ登録挿入される際の位置（i: insertion）および、登録後に再度アクセスされた場合に配置する位置(p: promotion) の LRU ポジションを、そのラインがページテーブルエントリを保持しているかどうかで変更するというものである。ページテーブルエントリを保持するラインに対し、挿入時とアクセス時の LRU ポジションをそれぞれ ip, pp とし、それ以外の通常のラインについては挿入直後，

アクセス直後の LRU ポジションを id、pd とする。この値が小さいほどそのラインが LLC で将来にわたって存在する確率を大きくすることができる。最適化を行わない場合には ip=pp=id=pd=1 である。これに対し、この値を変更することでαをα_{opt} に近づけることができる。この提案手法を適用した時に評価結果を図３に示す。

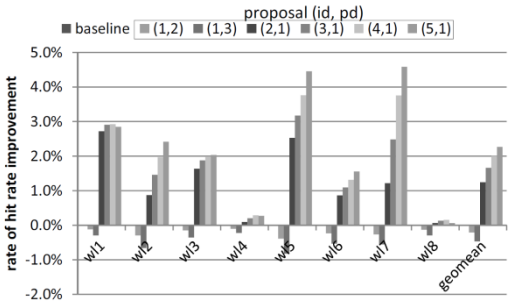


図３：提案手法による LLC のヒット率の改善

図３の評価は２コア、LLC は 8way 2MB の構成である。最適化を行わない ip=pp=id=pd=1 に対し、ip と pd を変更した場合の LLC ヒット率の向上を示している。横軸は、２つのコアに与えるプログラムセットの組み合わせを示し、最も右の値は全プログラムに対する幾何平均を表す。この結果から分かるように、多くの場合にヒット率を改善できることが分かった。

(2)高性能サーバシステム
スーパーコンピュータに代表される高性能サーバシステムは、性能向上と大規模化に伴い、電力供給系や冷却系の建設コスト、および、電気代といった運用コストの増大が深刻になっている。こうしたコストを削減する目的で、電力供給系をオーバープロビジョニングすることで電力供給系の使用効率を高める設計手法が提案されている。オーバープロビジョニングするとは、従来電力供給系の設計に用いられてきたシステムフル稼働時の電力（ピーク電力）ではなく、実運用時の平均的な消費電力に合わせて、電力供給系を設計することである。実運用時の平均的な消費電力は仕様上のピーク電力よりも大幅に低いことが多いため、このような設計手法を用いることで、不必要に大規模な電力供給系を用意する必要がなくなることが知られている。またこのような設計手法を用いる際には、電力超過の可能性を除去するために、運用時に計算機資源に電力制約を課す電力管理手法が必要となる。この手法としては、プロセッサの周波数やタスク量を調節することで、計算資源に電力制約を課すパワーキャッピング手法が研究されてきた。これは、いわば、計算機資源の間で電力を融通、すなわち空間的なパワーシフティングと考えることができる。しかしこれらの手法は単純に電力制約を守ることを目的として設計されているため、大きな性能低下を引き起こすことがある。そこで、UPS（無停電電源装置）内の蓄電池を用いて時間方向に電力を融通する、時間的なパワーシフティング手法を提案する。提案手法では，停電時のために設置されている UPS（無停電電源装置）内の蓄電池と周波数制御を併用し、電力を投入しても性能が上がりにくいフェーズ（プログラム実行の一部の区間）から、電力を投入することで性能が大きく上がるフェーズへ電力を融通することで高い性能を達成する。
この問題は、（式２）として定式化できる。ただし、p_{max}は与えられる電力制約、Δp_iは当該フェーズ i において蓄電池から供給される電力、T_iはそのフェーズの実行時間を表す。制約式の左辺はアプリケーション全体を通じて蓄電池から供給されるエネルギーを意味し、この制約式はエネルギー保存を表す。

$$\min \quad \sum_{i=1}^{n} T_i(p_{max} + \Delta p_i) \qquad \text{（式２）}$$

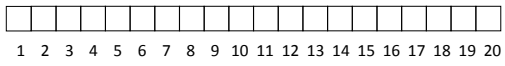
$$\text{s.t. } \sum_{i=1}^{n} \Delta p_i T_i(p_{max} + \Delta p_i) \leq 0$$

現実のプロセッサが動作する周波数は有限個しかないため、Δp_iの組み合わせも有限となる。そのため、この式は離散最適化問題として説くことが可能であり、周波数・充放電計画を決定することができる。
初期評価として、提案手法を 4CPU から構成されるシステムと 2GPU から構成されるシステム適用した場合の効果をシミュレーションで評価した。その結果、CPU 上で実行されるアプリケーションで平均 4.5%、GPU 上で実行される並列アプリケーションで平均 17.1%の実行時間の短縮ができることがわかった。

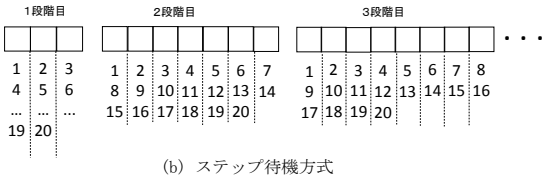
(3)センサネットワークシステム
センサネットワーク全体で消費エネルギーを削減するためには、センサでのセンシングやタスクの処理に加えて個々の通信も考慮した構成を考える必要がある。通信部分での消費エネルギーを削減するには、送信回数を減らしデータをまとめて送信することなどが考えられる。たとえば環境モニタリングの場合、観測値の変化がきわめて小さい、あるいは変化率がきわめて小さい、という場合が良くあり、これらの場合には単純な線形圧縮を用いるだけで送信回数を大きく減らすことが可能となる。このような場合、送受信をしないこととし、受信側も省略された送信に応じて適切な挙動をすることで通信での消費エネルギーを削減することができる。
センサネットワークシステムでは一対多通信が主となるので，通信の衝突を避ける通信方式を採用する必要があり、通信はスロットと呼ばれる細かい時区間に分割して行う時分割多重通信方式がよく採用される。各タイムスロットで高々1 つのデバイスしか送信せず、受信側がそのタイムスロットで受信待機

していれば、この送受信は成功する。しかし、同じスロット内で複数のデバイスからの送信が存在すると衝突が生じるため通信は失敗する。しかしこの時、受信側は無効なデータを受信するので衝突が生じたことは把握でき、受信完了信号を送付しないので、送信側も、送信に失敗したことは把握できる。

環境モニタリングのような場合、センサノードの数は変わらないため、受信側は送信デバイスの数があらかじめ分かっている。この場合、全ての送信デバイスが送信すると考え、送信デバイスの数（以降、NodeNum とする）と同じ数のスロットで受信待機すれば、確実に送受信できる。これを全待機方式と呼び、この時の待機スロットを図４(a)に示す。これは確実に送受信できるものの、送信しないデバイスがある場合には、受信側が無駄な電力を消費する。なぜならば、たとえ受信しなくても受信のためのスロットを必要とし、待機電力を消費するからである。



(a) 全待機方式



(b) ステップ待機方式

図４　全待機方式とステップ待機方式

そこで、図４(b)に示すステップ待機方式を提案する。基本的な考え方は、送信デバイスの送信確率は低いと想定し、送信デバイスの数よりも少ない受信スロットを用意する。その場合には、衝突が発生する可能性もあるため、衝突が発生した場合には、次により多くの受信スロットを用意し再度送受信を行う、というものである。図４(b)を用いて説明する。まず 1 段階目として，m_1 個（図中では m_1=3）のスロットを確保し、各送信デバイスを、デバイス ID を m_1 で割った剰余の番号のスロットに割り当てる。1 段階目での衝突を検出した場合、2 段階目として $m_2 = \lceil NodeNum / m_1 \rceil$ 個のスロット（$\lceil x \rceil$ は x の天井関数）を確保し，各送信デバイスを、デバイス ID を m_2 で割った剰余の番号のスロットに割り当てる。2 段階目においても衝突を検出した場合は、新たに 3 段階目として $m_3 = m_2 + 1$ 個のスロットを確保し、各送信デバイスを、デバイス ID を m_2 で割った剰余の番号のスロットに割り当てる。以下同様に、k 段階目における衝突を検出した場合は，新たに k+1 段階目として $m_{k+1} = m_k + 1$ 個のスロットを確保する。この場合、もし送信デバイス数が多い場合には、全待機方式の方が結局必要となる受信待機スロットが少なくなるので消費エネルギーも小さくなる。一方、送信デバイスの送信確率が小さいときには、明らかに全待機方式の方が不利になる。このため、その優劣は、送信デバイス側の送信確率に強く依存する。例えば、送信デバイスが 20 個の時、全待機方式では送信確率に依存せず必ず 20 個のスロットを必要とするが、送信確率が 0.1 の時、ステップ待機方式では平均的に 7.27 個のスロットしか必要としない。図５に、送信デバイス数を 50 と仮定し、モンテカルロシミュレーションにより２つの方法を比較した結果を示す。横軸は送信確率を示す。この図で、全待機方式において送信確率が高くなるとエネルギーが低下する理由は、スロット内で実際に送信が行われる場合には受信側は受信完了後にスロット内で sleep モードに入るのに対し、送信がなされない場合にはスロット内で常に受信待機のために sleep モードに入れないためである。この評価結果から、送信確率が小さいときは提案手法が大きく消費電力を削減でき、例えば送信確率が 0.1 の時、約 73%の消費電力削減に成功している。

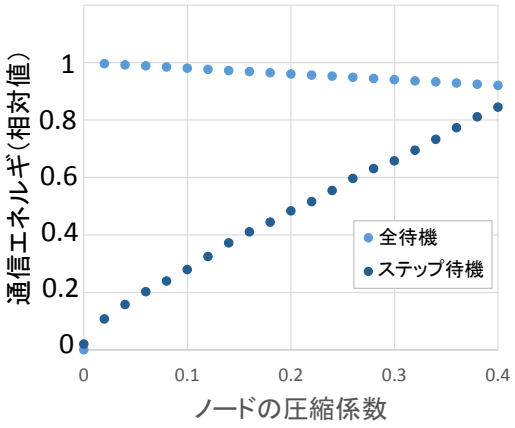


図５　送信デバイス数が 50 の時全待機方式とステップ待機方式の比較

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計０件）

〔学会発表〕（計５件）

①米澤 亮太，會田 翔，三輪 忍，中村 宏 ：物理メモリの増減による電力制約下での HPC システムの性能向上， 情報処理学会研究報告 2014-HPC-143，No. 24，pp. 1-8 (2014)．（石川県七尾市）

②酒井 崇至，薦田 登志矢，三輪 忍，中村宏 ：電力制約下における蓄電池を用いた HPC システムの性能向上， 情報処理学会研究報告 2014-HPC-143，No.25，pp.1-6（2014）（石川県七尾市）

③有間 英志，三輪 忍，中田 尚，中村宏 ：TLB ミスペナルティ削減のための大容量 LLC の利用法に関する初期検討，情報処理学会研究報告 2015-ARC-214 No.7， pp.1-6（2015）（神奈川県横浜市）

④田中 維人，中田 尚，中村 宏 ： 通信タイミング最適化によるセンサネットワークシステム消費エネルギ削減の検討，情報処理学会研究報告 2015-EMB-36 No.21,pp. 1-6（2015）（鹿児島県奄美市）

⑤富井 潤，近藤 正章，中村 宏 ：不揮発性メモリを用いたニューロチップに関する検討，情報処理学会研究報告 2015-EMB-36 No.25,pp. 1-6（2015） （鹿児島県奄美市）

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）
○取得状況（計０件）

〔その他〕

６．研究組織
(1)研究代表者
中村宏（NAKAMURA HIROSHI）
東京大学・大学院情報理工学系研究科・教授
研究者番号：20212102

(2)研究分担者
なし

(3)連携研究者
三輪忍（MIWA SHINOBU）
東京大学・大学院情報理工学系研究科・助教
研究者番号：90402940