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

令和　２ 年　 ６ 月　 ７ 日現在

機関番号：１２６０１
研究種目：基盤研究(B)（一般）
研究期間：2017～2019
課題番号：１７Ｈ０１７０８

研究課題名（和文）不揮発メモリコンピューティング

研究課題名（英文）Non-Volatile Memory Computing

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

交付決定額（研究期間全体）：（直接経費）　　14,400,000円

研究成果の概要（和文）：不揮発メモリを想定し、データを退避することなく電源遮断できる場合に、性能と消費電力が異なる複数のプロセッサコアからなるヘテロジニアスマルチコアシステムにおいて、応答時間制約のある周期的な処理を低消費電力で実行可能なスケジューリング手法を提案した。また、不揮発メモリの大容量性を活かし、データセンタのサーバの主記憶に不揮発メモリを採用し、より多くの処理を少ない数の物理マシンに割り当てて実行することで省電力化を行う手法を提案し、不揮発メモリコンピューティングの有用性を明らかにした。

研究成果の学術的意義や社会的意義
あらゆるものがネットワークに接続されるIoT(Internet of Things)時代が到来し、物理世界と情報世界の間の高度なインタラクションに対する期待は大きい。その実現には、我々が存在する物理世界から大量に収集されるデータを高速かつ低消費電力で処理することが必要である。本課題は、高速かつ電源を切っても記憶を忘れない新しい原理に基づく不揮発メモリを高度に利用するコンピューティング技術を実現することで、この社会の要請にこたえるものである。

研究成果の概要（英文）：By using non-volatile memory, application can be suspended without storing data. For heterogeneous multi-core systems under this situation, a novel scheduling method to reduce power consumption is proposed for tasks which have their own deadline. Another advantage of non-volatile memory is large capacity. For systems in data centers, a new method of power reduction by consolidating many applications into smaller number of physical systems due to larger capacity of non-volatile memory. Evaluation reveals the advantage of non-volatile memory.

研究分野：コンピュータシステム

キーワード：計算機システム　不揮発メモリ

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

あらゆるものがネットワークに接続される IoT(Internet of Things)時代が到来し、物理世界と情報世界の間の高度なインタラクションに対する期待は大きい。その実現には、我々が存在する物理世界から大量に収集されるデータをエッジ系とサーバ系の両方のシステムで高速かつ低消費電力で処理することが必要となっている。この社会要請下で、技術的には電荷の蓄積ではなく抵抗値の変化でデータを記憶する新しい原理に基づく不揮発メモリ(MRAM, FeRAM など)が提案され、期待が高まっている。記憶の原理が異なるため、揮発メモリ(SRAM/DRAM)がスケーリング限界により更なる高集積化が難しいのに対し、不揮発メモリは更なる高集積化が可能で揮発メモリより大容量化が可能となっている。不揮発メモリは磁気ディスクよりは高速だが揮発メモリよりは遅いという欠点があった。このため、不揮発メモリは、磁気ディスクで構成される補助記憶の一部を置き換える SCM(Storage Class Memory)としての活用が検討されてきた。これは大規模データを扱い補助記憶へのアクセスが頻発するシステムには向いているが、主記憶が性能や電力上のボトルネックになるシステムには有効ではない。

これに対し、補助記憶ではなく、主記憶を上記の新原理に基づく不揮発メモリで構成する研究は盛んではなかった。

２．研究の目的

本研究は主記憶を大容量の不揮発メモリで構成し、データの永続性が保証された論理アドレス空間(persistent logical address space)を実現する「不揮発メモリコンピューティング」を提案する。これまでも、揮発メモリをこのような不揮発メモリで置き換える研究は存在したが、電力・速度・容量のトレードオフのみが議論されており「不揮発性」を積極的に活用する研究はなされていなかった。これに対し、本研究は、補助記憶上ではなく論理アドレス空間上でデータの永続性を実現するものであり、「不揮発性」を積極的に活用する新しいコンピューティング方式を検討する点が大きな特徴である。

この方式には２つの利点がある。１つは、補助記憶へのファイル操作を経ずともデータの永続性を保証できる点である。外部入出力を伴う処理、つまり外部からの入力データを保存した後に処理を加える場合や、処理結果を保存した後に外部へ出力する場合に、従来は、補助記憶上に OS 経由でファイルとして保存する必要があり長い時間を要した。これに対しアドレス空間が不揮発となる提案方式では、主記憶上でデータの永続性を保証できるため、OS 経由のファイル操作を必要とせず処理時間を短縮できる。これは、主記憶容量が性能上のボトルネックとなるサーバー系のシステムでは大きな利点となる。もう１つの利点は、アドレス空間上のデータを退避することなく電源を遮断することで、エネルギー消費を削減できる点である。従来は、電源遮断時には主記憶上のデータを補助記憶へ対する必要があったため、性能とエネルギーの両方を失っていた。これに対し、提案方式ではデータの退避・回復をすることなく瞬時にプロセスの中断・再開が可能となるため、非動作時の電源遮断による性能低下を招くことなくエネルギー消費を低減できる。外部入出力を伴い、間欠的に動作し非動作時間が占める割合が大きいシステム、例えば、モニタリングシステムや携帯端末など、環境や人間との高度なインタラクションを実現するこれからの IoT(Internet of Things)時代に飛躍的に増加するエッジ側のシステムにおいて、応答性と省エネルギー性の観点から、その利点は大きなものとなる。

本研究の目的は、これらの利点を真に享受可能な不揮発メモリコンピューティングの設計方法論を明らかにすることである。

３．研究の方法

物理世界と情報世界の間の高度なインタラクションを実現するべく、物理世界から大量に収集されるデータをエッジ系とサーバ系の両方のシステムで効率よく処理することが不揮発メモリコンピューティングで可能となることを示す。そのために、

(1) ヘテロジニアスマルチコア周期実行システムにおける省電力タスクスケジューリング

(2) 性能モデルを用いた不揮発主記憶サーバの省電力化の検討

の 2 つの研究を実施した。

(1)は、エッジ系のシステムを対象としたものである。従来エッジ系のシステムはセンサからの少量のデータ収集とそれらのデータに対する極めて軽い処理を行うものであったが、近年は、画像などの重たいデータも扱うため、システム構成も、処理能力は低いが消費電力も小さいプロセッサコア(small core)と、消費電力は大きいが処理能力も高いプロセッサコア(large core)を組み合わせるヘテロジニアスマルチコア(heterogeneous multi-core)が利用されるようになってきた。そのため、この構成を対象に、エッジ系のシステムでは典型的な、周期的に到着しデッドライン制約がある処理が複数種類存在する場合の省電力化を検討する。提案する省電力化手法は、ハードウェアの利点を有効に活用する、OS レベルのスケジューリング技術であり、アーキテクチャとシステムソフトウェアの協調による最適化手法である。

(2)は、データセンタのサーバ系システムを対象としたものである。これらのシステムでは主記憶が性能を律速することが多い。そのため、不揮発メモリの大容量性を活用することに着眼した。データセンタのサーバの典型的な利用形態としては、複数のウェブサービスが並列に多数実行されるが、セキュリティを確保するために、リソースを共用しつつ動作環境を分離し、これらのサービスは別々の VM(virtual machine:仮想マシン) に分離されて実行される。VM は実際には

複数の PM(physical machine:物理マシン) 上に配置され、負荷の分散や消費エネルギーの削減のために PM 間で VM を移動するマイグレーションが行われる。提案手法を端的に説明すると、不揮発メモリの大容量性を頼りに、PM の主記憶に不揮発メモリを搭載し大容量化することで、できるだけ多くの VM を少数の PM に集約することで必要となる PM の台数を減少し低電力化を目指す。しかしながら、表 1 に示すように、不揮発メモリ（図 1 では PCM）は、Flash メモリよりは高速であるものの、DRAM よりは低速であり、性能は低下する。そこで、主記憶には揮発メモリと不揮発メモリの両方を採用し、異なる種類の複数の VM を PM に集約した場合の性能をモデリングすることで、集約すべき VM を適切に選択することで、できるだけ性能低下を抑えながら消費電力を抑えることを目指す。性能モデリングにおいては、各 VM のハードウェア利用状況をモニタリングする必要があり、それはシステムソフトウェアの機能に依存する。したがって、この研究もアーキテクチャとシステムソフトウェアの協調による最適化手法である。本研究期間の後半では、実際に不揮発メモリである 3D Xpoint を主記憶として利用できる Intel Optane DCPM (Persistent Memory)が入手できるようになったので、提案手法の実機評価も含めて、その有効性を検証した。

表 1：主なメモリの特性

	記憶密度	レイテンシ (Read/Write)	エネルギー (Read/Write)
Flash	約 4 倍	2.0×10^4/ 2.0×10^5[ns]	1.1×10^2/7.9×10^2 [pJ/bit]
PCM	約 2 倍	50/150[ns]	3.0×10^2/1.6×10^3 [pJ/bit]
DRAM	1(基準)	30/30[ns]	15/16[pJ/bit]

４．研究成果

(1) ヘテロジニアスマルチコア周期実行システムにおける省電力タスクスケジューリング

図 1 に問題設定と提案手法の概要を示す。図 1(a)のように、複数の種類のジョブ(図中では Job A, JobB)が周期的に到来(周期は図中の Input Interval)し、それぞれ処理が終了するまでのデッドライン（図中では Deadline）を有していることを想定する。図中の各ジョブの大きさは、処理量の大きさを表している。デッドラインが周期より長いという過程ではマルチメディアストリーム処理では一般的であり、near real-time periodic task と呼ばれることもある。処理の軽いジョブと重たいジョブが存在し、ヘテロジニアスマルチコア構成の場合、スケジューリング（ジョブのプロセッサへの割り当てとその実行タイミングの両方を含む）を工夫することで低消費電力化が可能となる。

図 1(b)(c)(d) はこれまでに提案されているスケジューリングである。MCUlow は性能と消費電力が小さいコアを、MCUhigh 性能と消費電力が大きいコアをそれぞれ表す。 (b)は最も単純な ASAP(as soon as possible)と呼ばれるもので、到着順にできるだけ早くスケジューリングする

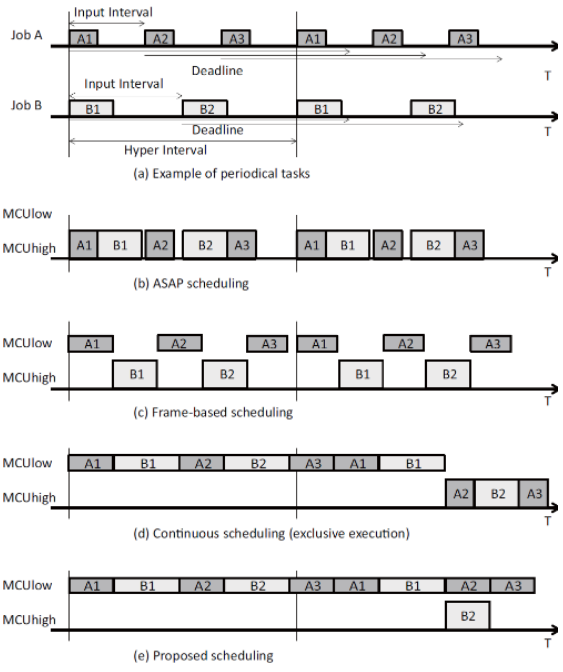

図 1 スケジューリング手法

Parameters	Task J_1, J_2, J_3	Task J_4, J_5, J_6
Task ID j	1, 2, 3	4, 5, 6
Input interval $I(j)$	100 ms	50 ms
Deadline period $d(j)$	250 ms	

(a) タスクミックス

	MCU1	MCU2	MCU3	MCU4
core ID c	c_1	c_2	c_3	c_4
Relative Performance $p(c)$	1.0	2.0	3.0	4.0
Power Consumption in				
Active state $P(c)$ [mW]	15.4	36.8	90.9	231
Sleep state $P_s(c)$ [μW]	0.69	2.07	6.20	18.6
Energy Overhead $E_{OV}(c)$[μJ]	51.0	124	253	430
Exec time of				
task J_1, J_2, J_3 [ms]	2.4	1.2	0.8	0.6
task J_4, J_5, J_6 [ms]	54.6	27.3	18.2	13.65

(b) ハードウェアの仮定

図 2 評価条件

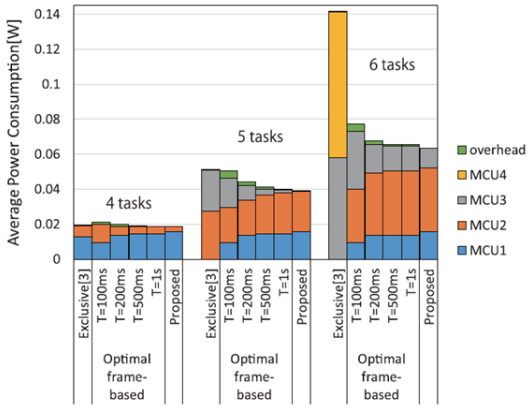

図 3 消費電力の評価結果

もので、各ジョブが割り当てられるコアが静的に固定され MCUlow を使うことができない。(c)(d)は、各ジョブが割り当てられるコアは動的に選択される（例えば JobB は MCUlow にも MCUhigh にも割り当てることがある）ものである。(c)は、Input Interval の最小公倍数で定義される hyper interval を導入し、hyper interval 毎にスケジューリングをするものである。これに対し、deadline が周期よりも長いことを積極的に利用し、できるだけ MCUlow にジョブを割り当てようとするスケジューリングが(d), (e) である。(d)では、MCUlow と MCUhigh が同時に動作することを許していないが、(e) はそれを許すもので、これが提案手法となる。
図 2 に示す複数のタスクとハードウェア構成を想定して評価した結果を図 3 に示す。図 3 では、実行するタスク数を 4~6 で変化させた場合の提案手法(図中では Proposed)の消費電力を、既存手法である exclusive execution と Optimal frame-based execution と比較して示している。Optimal frame-based の T は hyper interval の長さを表す。この図では、電力がどのコアで消費されたかの積み上げとして、消費電力を示している。性能はMCU1 < MCU2 < MCU3 < MCU4だが、消費電力効率はMCU1 > MCU2 > MCU3 > MCU4なので、できるだけ MCU1 を用いたほうが消費電力は下がる。が、この図評からわかるように、提案手法は、いずれの場合も他の手法に比べ消費電力を下げることに成功しており、提案手法の有効性がわかる。

(2) 性能モデルを用いた不揮発主記憶サーバの省電力化の検討
データセンタでは、さまざまなウェブサービスを並列に多数実行する。例えば、KVS(Key-Value Store)は Key とそれに対応する Value からなるデータベースであり、ネットワーク経由でデータを入れる(SET)ことや取り出す(GET)。HTTP はハイパーテキストなどの転送に用いられ、データを指定して Web サーバから取り出す(GET)動作が基本となる。機械学習もよく用いられる応用で、学習によって作成済みのモデルに新たなデータを当てはめる推論の際には行列積の計算に代表される積和演算の処理がその主要部となる。これらのサービスは、リソースを共用しつつ動作環境を分離し、セキュリティを確保するために別々の VM(仮想マシン)に分離されて実行される。VM は複数の PM(物理マシン)上に配置され、負荷の分散や消費エネルギーの削減のために PM 間で VM を移動するマイグレーションが行われる。
本研究では、これまでは DRAM のみを搭載した複数台の PM で動作していた VM を、主記憶に不揮発メモリを用いる NVMM (Non-Volatile Main Memory) を追加して少ない台数の PM 上に集約することで、データセンタの省電力化を目指す。つまり、図 4 に示すように、NVMM の大容量性を活かして NVMM を搭載した PM1 台あたりで動作する VM 数を増やし、動作不要となった PM の電源を遮断することで省電力化を目指す。適切に VM を PM に集約するためには、VM を PM 間で移動するときの干渉(interference)が小さくする必要がある。そこで、VM-ID=i の VM を PM-ID=d の PM へ集約するときの interference を図 5 に示す式でモデリングする。この式では、PM 上での干渉が、ネットワーク、CPU、DRAM、NVMM の 4 箇所のハードウェア資源で発生することから導出している。ξは主記憶のバンド幅を扱うもので、L3 キャッシュはメモリ階層中の LLC(last level cache)に当たり、L3 キャッシュミスが主記憶アクセスを発生させる。また、ηは、NVMM のバンド幅を扱うもので、データサイズが DRAM 容量を越えたときに NVMM との間で置き換えが発生することに起因する。
提案する Interference のモデルの妥当性を検討するために実機を用いた評価を行った。NVMM を搭載した PM 上で，複数のウェブサービスを稼働させて評価することとし、NVMM は Intel 社の Optane DC Persistent Memory (PCM 3D XPoint を搭載) を 256 GiB 搭載し、DRAM は 64 GiB 搭載し NVMM に対するキャッシュとして動作させた。CPU は Intel Xeon Gold 5215 (10 コア 20 スレッド)である。ウェブサービスのベンチマークとして表 2 のリクエストを他のマシンから送信した。2〜8 台の VM が動作中の PM に 2 台の VM を追加した際の、モデリングから得られる Interference の予測と実際の性能低下率の関係を図 6 に示す。この結果から、モデルによる Interference 予測と実際の性能低下率の最大値にはほぼ比例関係が見られる。一部、Interference が大きいにもかかわらず性能低下率が 100% ~ 150% に留まる場合も見られた。この場合、性能低下を過大に見積もることとなり積極的なマイグレーションを行わないが、その場合は VM 集約による省電力効果は低くなるものの実効性能に対する悪影響が発生しない。次に、省電力化効果の評価を示す。KVS サービス でメソッドが GET、リクエストサイズが 100kB の場合に、DRAM のみを搭載した PM と、NVMM を追加した PM の消費電力を図 7 に示す。このとき VM が 4 台以上の場合には DRAM のみの PM では主記憶容量が足りず、スワップが発生してサービスを継続できない。そのため、VM が 4 台以上の場合には、DRAM+NVMM ならば PM1 台，DRAM のみならば PM2 台が必要となる。図 7 より分かるように、DRAM のみ 1 台で実行できる場合には、NVMM の消費する電力がオーバヘッドとなり消費電力は大きくなる。しかし、VM 台数が増えると、DRAM のみならば PM が 2 台必要となるため NVMM を搭載する場合に比べて消費電力は大きくなる。一般に NVMM は DRAM よりも遅いため実行時間が長くなるケースがあり、その場合には電力は低下しても消費エネルギーの削減にはつながらない。しかし、KVS はメモリバンド幅がボトルネックとなるため、NVMM 搭載の PM でも実行速度は低下せず、消費エネルギーも削減できた。

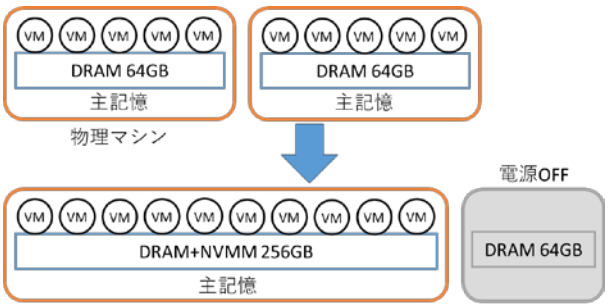

図 4 不揮発メモリを用いた省電力化

表 2 ベンチマーク

目的とするボトルネック	サービス	メソッド	サイズ	データ数
ネットワーク	HTTP	GET	100MB	1
ネットワーク	KVS	GET	100kB	200,000
ネットワーク	KVS	SET	100kB	200,000
DRAM	Matmul		15k	1
NVMM	Matmul		30k	1
NVMM	Seqread		16GB	1

$$\text{ネットワーク}: \gamma = \frac{\text{VMのネットワークバンド幅の和}}{\text{PMのネットワーク容量}}$$

$$\text{CPU}: \theta = \frac{\text{Demand: VMのCPU使用率の和}}{\text{Supply: PMのCPUコア数}}$$

$$\text{DRAM}: \xi = \frac{\text{VMの時間あたりL3キャッシュミス数の和}}{\text{PMが提供可能な時間あたりL3キャッシュミス数}}$$

$$\text{NVMM}: \eta = \begin{cases} \dfrac{\text{VMの主記憶からの読み出し速度の和}}{\text{NVMM使用時にPMが提供可能な主記憶からの読み出し速度}} & (\text{アクティブメモリ量} > \text{DRAM容量}) \\ 0 & (\text{アクティブメモリ量} \leq \text{DRAM容量}) \end{cases}$$

$$Interference_{i,d} = \exp\left(\max\left(\gamma_{i,d}, 1\right)\right) \times \frac{1}{1-\theta_{i,d}} \times \frac{1}{1-\xi_{i,d}} \times \frac{1}{1-\eta_{i,d}}$$

図 5 interference のモデリング

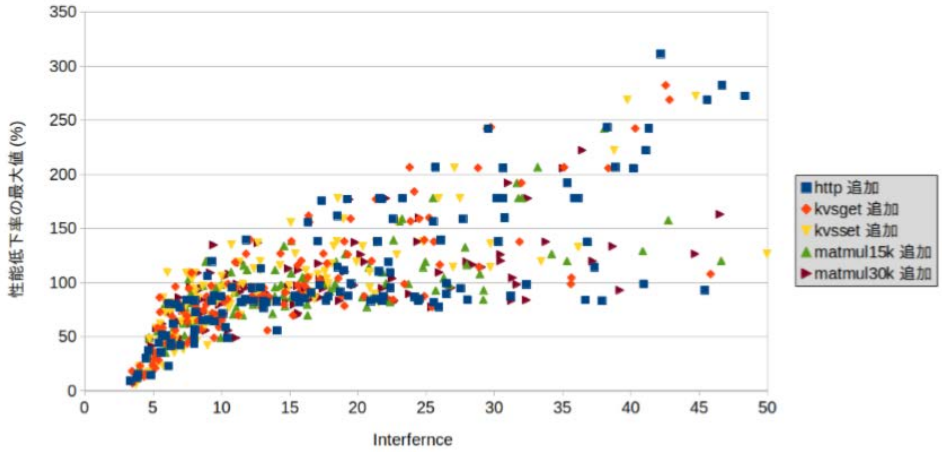

図 6 interference と性能低下率の関係

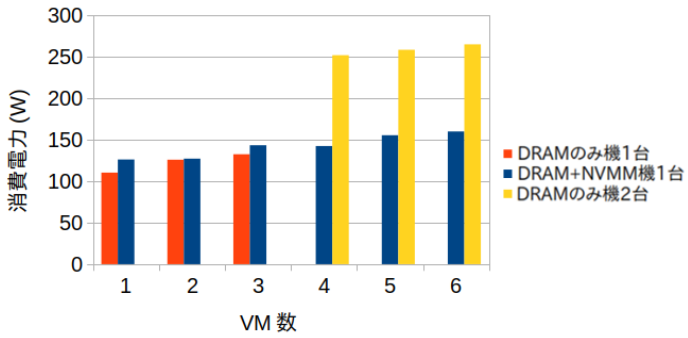

図 7 KVS を実行する場合の省電力効果

５．主な発表論文等

〔雑誌論文〕　計2件（うち査読付論文　2件／うち国際共著　0件／うちオープンアクセス　0件）

１．著者名 T. Nakada, H. Yanagihashi, H. Nakamura, K. Imai, H. Ueki, T. Tsuchiya, M. Hayashikoshi	４．巻 -
２．論文標題 Energy-aware Task Scheduling for Near Real-time Periodic Tasks on Heterogeneous Multicore Processors	５．発行年 2017年
３．雑誌名 Very Large Scale Integration, 2017 IFIP/IEEE International Conference on	６．最初と最後の頁 -
掲載論文のDOI（デジタルオブジェクト識別子） 10.1109/VLSI-SoC.2017.8203458	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 T. Nakada, H. Yanagihashi, K. Imai, H. Ueki, T. Tsuchiya, M. Hayashikoshi, H. Nakamura	４．巻 E103-D
２．論文標題 An Energy-Efficient Task Scheduling for Near Real-Time Systems on Heterogeneous Multicore Processors	５．発行年 2020年
３．雑誌名 IEICE TRANSACTIONS on Information and Systems	６．最初と最後の頁 329-338
掲載論文のDOI（デジタルオブジェクト識別子） 10.1587/transinf.2019EDP7101	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計10件（うち招待講演　0件／うち国際学会　2件）

１．発表者名 Kenji Oshiro, Shinsuke Hamada, Atsushi Koshiba, Mitaro Namiki
２．発表標題 Evaluation System for Low-power LSIs using SOTB Technology towards Software-based Body Bias Control
３．学会等名 IEEE Symposium on Low-Power and High-Speed Chips XXI（国際学会）
４．発表年 2018年

１．発表者名 Shinsuke Hamada, Soramichi Akiyama, Mitaro Namiki
２．発表標題 Reactive NaN Repair for Applying Approximate Memory to Numerical Applications
３．学会等名 The 8th Workshop on Systems for Multi-core and Heterogeneous Architectures（国際学会）
４．発表年 2018年

１．発表者名
大城 研治，小柴 篤史，濱田 慎亮，並木 美太郎

２．発表標題
ソフトウェアによるSOTBチップの動的ボディバイアス制御に向けた評価環境の構築

３．学会等名
情報処理学会第143回システムソフトウェアとオペレーティング・システム研究会

４．発表年
2018年

１．発表者名
濱田 槙亮、穐山 空道、並木 美太郎

２．発表標題
デバッグ情報とシグナルを用いたメモリエラーの修復

３．学会等名
情報処理学会　第30回コンピュータシステムシンポジウムComSys2018

４．発表年
2018年

１．発表者名
堀米亮汰，宇佐美公良

２．発表標題
3次元積層LSIの実チップ発熱・放熱時における温度の過渡解析と評価

３．学会等名
電子情報通信学会VLSI設計技術研究会(VLD2018-107)

４．発表年
2019年

１．発表者名
市川遼，並木美太郎

２．発表標題
準パススルー型仮想マシンモニタを用いたプロセス単位の解析イ ンターフェース

３．学会等名
情報処理学会第80回全国大会

４．発表年
2017年

１．発表者名
湯木大輝、坂本龍一、中村宏

２．発表標題
不揮発主記憶サーバの性能モデルと省電力化の検討

３．学会等名
情報処理学会コンピュータシステム・シンポジウムComSys2019

４．発表年
2019年

１．発表者名
市川遼、須崎有康、並木美太郎

２．発表標題
Staliの全バイナリを対象とした静的解析によるシステムコールの列挙

３．学会等名
暗号と情報セキュリティシンポジウムSCIS2020

４．発表年
2020年

１．発表者名
市川 遼、坂本龍一、中村宏、並木 美太郎

２．発表標題
ハイパーバイザーとNVMを用いたメモリトレイサーの検討

３．学会等名
情報処理学会第148回システムソフトウェア研究会

４．発表年
2020年

１．発表者名
秋葉爽輔、宇佐美公良

２．発表標題
2電源を用いた不揮発性フリップフロップの提案と評価

３．学会等名
電子情報通信学会VLD研究会

４．発表年
2020年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	並木　美太郎 (Namiki Mitaro) (10208077)	東京農工大学・工学（系）研究科（研究院）・教授 (12605)	
研究分担者	宇佐美　公良 (Usami Kimiyoshi) (20365547)	芝浦工業大学・工学部・教授 (32619)	