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　３１　年　　５　月　　６　日現在

機関番号：　１２６０８
研究種目：　挑戦的萌芽研究
研究期間：　2016～2018
課題番号：　１６Ｋ１２４０６

研究課題名（和文）ベンチマークの解剖学

研究課題名（英文）Anatomy of benchmarks

研究代表者
首藤　一幸（Shudo, Kazuyuki）

東京工業大学・情報理工学院・准教授

研究者番号：９０３０８２７１

交付決定額（研究期間全体）：（直接経費）　　2,600,000円

研究成果の概要（和文）：ベンチマークは、計算機システムの性能が目的に合致するかを評価し、比較するための欠かせない手段である。本研究では、アプリケーションベンチマークと構成要素・基本要素ごとの性能との間の関係を見出す統計的手法を提案した。当初予定していたCPUに加えて、ストレージシステムを対象とした。アプリケーションに対して、CPUのどの基本性能がどの程度寄与しているのかを調べることが可能となった。また、Storage Class Memoryといった現れつつあるデバイスを用いることでストレージシステムの性能がどうなるのか推定できるようになった。

研究成果の学術的意義や社会的意義
アプリケーションやアプリケーションを模したベンチマークの性能評価結果が一体何を表しているのかを調べられるようになった。また、ベンチマークスコアを吊り上げるチートと呼ばれる不正行為を検出できる可能性がある。

研究成果の概要（英文）：Benchmarking is an indispensable method to evaluate how much computer systems are suitable to our purpose by measuring and comparing them. We proposed a statistical technique to reveal relationships between application benchmarks and performance of system components. Storage systems were also our targets in addition to CPUs, that was planned at the beginning. Our technique enabled us to investigate which basic operations, such as add and memory access, contribute to the benchmarking results and how much they contribute. It is also enabled to estimate how a storage system performs well with emerging storage devices such as Storage Class Memories.

研究分野：　計算機性能評価

キーワード：　ベンチマーク　回帰分析

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９、ＣＫ－１９（共通）

１．研究開始当初の背景

サーバから組み込み機器まで、計算機システムの性能が目的に合致するかを評価し、比較するために、ベンチマークは欠かせない手段である。価格や消費電力なども重要だが、性能は依然もっとも重要な評価項目の１つである。我々は、どういったシステムを使用／購入するかを決めるために、ベンチマークプログラムを実行したり、手に入るスコアを参照する。それだけでなく、計算機システムの開発自体がベンチマークスコアの向上を目指して、つまり、ベンチマークに駆動されて行われている。CPU 開発は SPEC CPU が、データベースサーバ開発は TPC ベンチマーク群が駆動してきた。スマートフォン業界ではメーカが AnTuTu ベンチマークを意識するあまり不正を行っていると報道された。

ベンチマークプログラムが実アプリに近いほど、システムの構成要素や基本機能ごとの性能がスコアに与える影響は明らかではない。これが例えば以下の問題の大きな要因となっている。

- **構成要素・基本性能ごとの性能が 過剰な影響力を持っている**：
  スパコン性能ランキング TOP 500 で用いられる Linpack では、実アプリでは重要なメモリ帯域幅が狭くとも高いスコアは出せる。このように、限られた側面しか反映しないにも関わらず、行政も参照する重要なランキングとなっている。こうした一面的なスコアを冷静に解釈するためには、Linpack やそれを含む HPC Challenge ベンチマーク群のスコアと実アプリ性能（寄与率など）を明らかにすることが欠かせない。
- **何を表しているかまったく不明な総合スコアが 過剰な影響力を持っている**：
  SPEC CPU2006 は、ベンチマークプログラム 12 個の相乗平均を整数演算性能の総合スコアとする。SPEC JVM も相乗平均、AnTuTu は総和である。世を代表するアプリ 12 個の相乗平均が何を表すだろうか? ある実アプリの傾向でも、世の処理内容の平均でもない。しかしスコアの影響力は甚大で、開発・発展の方向性を決め、売り上げも大きく左右する。我々は、個別・総合スコアが何を表しているかを知らねばならない。構成要素・基本機能ごとの性能がスコアにどう影響しているかを明らかにする必要がある。

２．研究の目的

我々は、実アプリまたはそれに近いベンチマークのスコアと、構成要素・基本機能ごとの性能、つまりマイクロベンチマークの結果やシステム仕様との間の関係（例：寄与する要素や寄与率）を明らかにし、その手法を確立する（表１）.

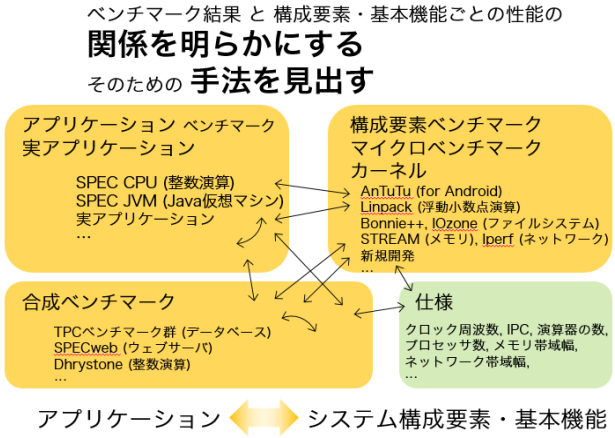

表１　研究の目的

３．研究の方法

(1) 実アプリケーションやそれに近いベンチマークプログラムを用いたベンチマークスコアと、(2) システムの構成要素・基本機能ごとの性能データ、この両者(1)・(2)を測定または入手し、統計的手法を用いて両者の関係を見出していく。

- スコアや性能データの入手：
  我々自身で測定するか、既存のものを入手する。
  (1) は、文字通り、実アプリケーションやアプリケーションベンチマーク（実アプリを基にしたベンチマークプログラム）を用いたベンチマークスコアであり、(2) としては構成要素ベンチマークやマイクロベンチマークのスコア、または、仕様上の数値を用いる。
- ベンチマークプログラムの選択：
  業界標準のもの、(1) については例えば SPEC や TPC が有力候補である。(2) は著名なものを試しつつ、必要に応じて自身でも開発していく。なぜなら、対象としたい基本機能やその粒度に応じた適切なベンチマークプログラムが存在するとは限らないからである。例

えば、整数演算性能という粒度で充分かもしれないし、インタプリタの命令解釈～実行ループやオブジェクト指向言語のプログラムなら、間接参照でのプロシージャ呼び出しといった細かい粒度での分析にも意味があるかもしれない。これは、近年、Intel 社のプロセッサで高速化された。
多くのシステムについてのスコアが入手可能であることが望ましい。例えば TPC-C V5 の結果は百数十システム分が入手可能であり、SPEC CPU2006 であれば 6,000 以上が入手可能である。

- 統計的手法を用いた関係の分析：
まずは、多変量解析を行う。具体的には、(1)、(2) の双方を基に関係を見出そうとするなら主成分分析を、(1) のみからスコアに影響を与える要素を見出そうとするなら因子分析を行うこととなる。

４．研究成果

当初、研究の対象としては CPU を考えていたところ、2 年目からはそれに加えて、ストレージシステムを研究対象とした。

①　CPU

CPU を対象として、アプリケーションベンチマーク SPEC CPU2006 と、自作マイクロベンチマークの関係を分析した。

SPEC CPU2006 からそれぞれ性質の異なる 3 つのベンチマーク 429.mcf、462.libquantum、470.lbm を対象とし、マイクロベンチマークは、ループ、整数の加算、乗算、除算、加算＆乗算、浮動小数点数の加算、乗算、除算、メモリ性能として整数のストア、ストア＆ロード、浮動小数点数のストア、ストア＆ロードを用意した。機材としては、Amazon Web Services のサービス Device Farm で利用できる 15 種類の Android OS 動作マシンを用いた。図１に内訳分析の結果を示す。

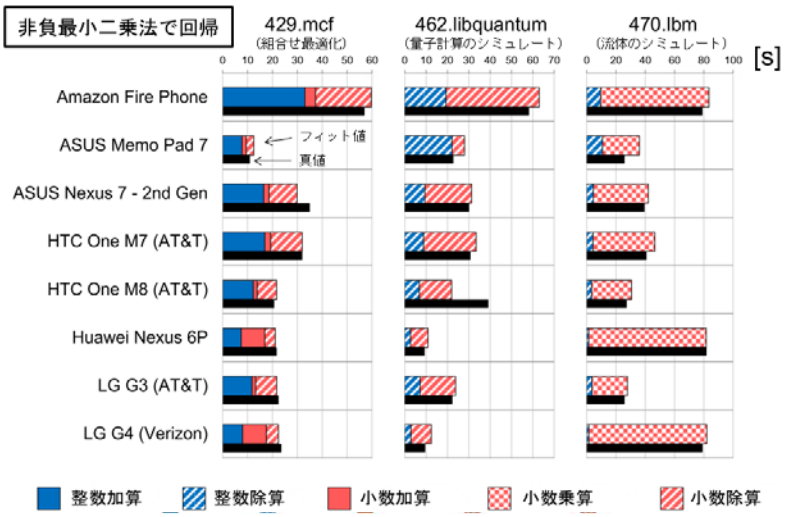

図１　SPEC CPU2006 の内訳分析の結果

一部でフィッティングの精度が悪い（例：HTC One M8 での 462.libquantum)。原因を調査している。

ベンチマークごとに異なる傾向が出ており、また、傾向は全マシンに共通であるため、ベンチマークそれぞれの特性をとらえることが出来ていると考えられる。

寄与率の高いマイクロベンチマークが明らかに誤っている個所がある。例えば、組み合わせ最適化である 429.mcf にて浮動小数点数の除算が大きく寄与しているとは考えにくい。これは、マイクロベンチマークどうしの傾向が近く、回帰によって識別し切れなかった結果であろう。

今後の課題は次の通りである：

- 精度が悪いケースの原因追求
- 適切な因子（今回はマイクロベンチマーク結果）の自動収集、組み合わせの自動選択
- サーバや高性能計算機への適用
- スコア吊り上げ（チート）検出の試み

②　ストレージシステム

データベース管理システム Apache Cassandra の改造版を対象として、アプリケーションベンチマーク Yahoo! Cloud Serving Benchmark（YCSB）と、マイクロベンチマーク fio の関係を分析した。ストレージデバイス 3 種類：NAND Flash、3D XPoint、DRAM それぞれを用いて、YCSB と fio で行った様々なベンチマーク結果を回帰分析した。

とてもよくフィットした条件の結果を図２に、あまりフィットしなかった条件の結果を図３に示す。グラフにプロットされた 3 点がそれぞれ、NAND Flash、3D XPoint、DRAM での結果を示す。

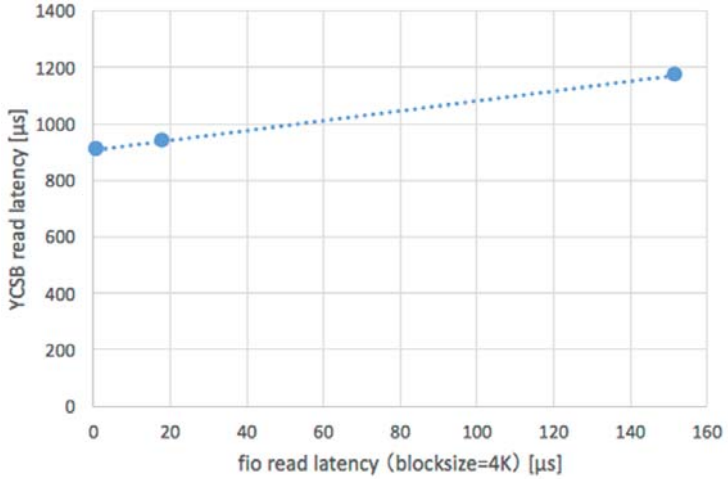

図２　YCSB のワークロード Write-Heavy、16GB の読み出し処理

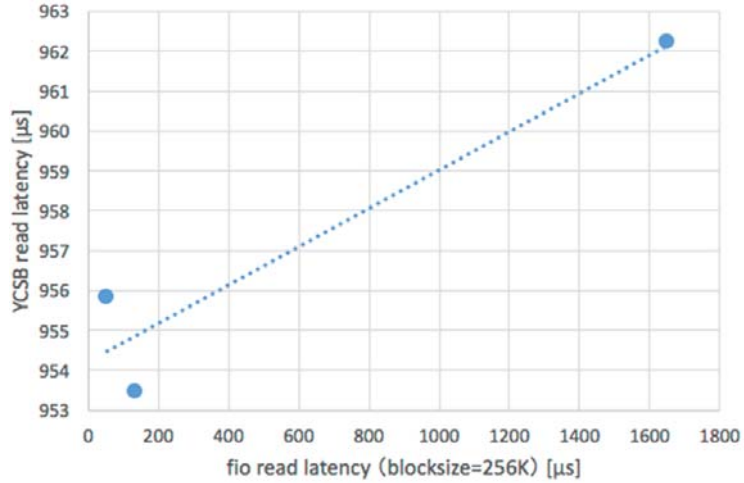

図３　YCSB のワークロード Read-Only、16GB の読み出し処理

フィットしていない条件があるからだめということではない。フィットする条件下であれば、新デバイス、例えば DRAM に迫る性能の Storage Class Memory (SCM) が現れた場合のアプリケーション (YCSB) 性能を精度良く推定できる、ということである。

５．主な発表論文等

〔学会発表〕(計２件)

① 曽我樹大、大西真晶、首藤一幸、回帰分析を用いたストレージシステムのベンチマーク結果推定、第 10 回データ工学と情報マネジメントに関するフォーラム (DEIM 2018)、2018 年 3 月 4 日～6 日

② Naoki Matagawa、Kazuyuki Shudo、Breakdown of a Benchmark Score Without Internal Analysis of Benchmarking Program、arXiv:1610.06307、2016 年 10 月 20 日

６．研究組織

(1)研究分担者
なし

(2)研究協力者
なし