論文の内容の要旨

論文題目 Run-time Optimization for Computer Systems based on Statistical Modeling of Hardware Events

（プロセッサ内イベント情報の統計的モデリングに基づく実行時最適化に関する研究）

氏名　佐々木 広

近年，コンピュータシステムにおけるプログラムの実行に関して種々の最適化が強く求められてきているが，とりわけ実行時における最適化の重要性が増してきている．一般に実行時（動的）における最適化手法の，静的な最適化手法と比べた場合の利点は，データセットの違いや，ハードウェア構成などの違いからくる動的な振る舞いに対処することが可能であり，効率的な実行を提供できるという点にある．ここで，静的に予測できない挙動とは，例えば最適化の対象として適切な周波数・電源電圧の選択や，キャッシュや命令キューのサイズ，マルチコア・マルチスレッドプロセッサにおける実行スレッド数の選択などが挙げられる．アプローチの手段も様々であり，ハードウェア的なアプローチから，コンパイラ・OSなどのソフトウェア的なものまでと多岐にわたっている．

OSによる手法やハードウェアによる手法では，一般的に実行時の情報を用いるため，キャッシュミスなどの動的な振る舞いにも対応できる．しかし，フェーズの変化などのプログラムの情報を明示的に利用することはできないため，最適化が困難な場合がある．また，コンパイラによる手法は，主としてプロファイリングを用いオフラインで分析を行うというものである．これらの手法の問題点として，例えばデータセットが異なっていた場合に，プロファイリング時と実際のプログラム実行時における振る舞いが異なっていると有効に最適化が行えないことが挙げられる．このように，ハードウェアやOSなどによる動的な手法と，コンパイラによる静的な手法にはそれぞれ一長一短がある．近年ではお互いの良い点を用いるハイブリッドな手法が提案されており，より優れた実行時における最適化を行うことが可能であると考えられている．

このように実行時における最適化が重要になってきている一方で，最適化を行うためには計算機の振る舞いを解析・理解し，どのような場合にどういった最適化を行うかというモデルを立てることが必要である．しかし，計算機システムは年々加速的に複雑化してきている．例えば命令実行においては，スーパーパイプラインや高度な分岐予測によるアウト・オブ・オーダー実行などが行われている．また，SMTやCMPといったワンチップ上でリソースを共有しつつ複数のスレッドを実行可能なプロセッサも登場し，多階層のキャッシュメモリを有している．こういった事情から，計算機の振る舞いを定性的に理解することが非常に困難になってきている．今後，さらに複雑化していく計算機システムにおいて，定性的な解析によるモデルを用いた最適化手法を生み出していくための時間およびコ

ストは増大する一方である．

例えば従来の手法では，プロセッサ内のあるイベント情報を指標として性能モデルを作成し，周波数・電源電圧を制御する DVFS 手法が提案されているが，プラットフォームが異なり，当該イベント情報を取得することができない場合は，新たに性能モデルを作成し直す必要がある．また，取得可能な場合であっても，性能モデルに用いられている数式のパラメータなどは，プラットフォーム毎に異なるため，再探索する必要があり，定性的にそれらを構築する手法には限界があると考えられる．

そこで我々は，このような問題に対処しつつ種々の最適化を可能にするために，計算機システムの振る舞いを定量的に解析し最適化のためのモデルを構築し，実行時に得られたモデルに基づいて最適化を行う手法を提案する．具体的には，まず計算機の振る舞いを示す様々なハードウェアイベントの定量的な値から統計的な学習を行い最適化実行のためのモデル（予測式および，着目すべきハードウェアイベントの決定）を作成する．その上で，実行時にはそれらハードウェアイベントにおける値を指標として最適化を行う．ハードウェアイベントの定量的な測定には，最近の大半のプロセッサに搭載されているパフォーマンスカウンタと呼ばれる機構を用いる．パフォーマンスカウンタにより，キャッシュのヒット/ミス回数や，分岐予測ミス回数などがソフトウェアから計測可能となっている．

本論文では以下の二種類の最適化問題について，実機のプラットフォーム上で提案手法を用いて最適化のためのモデルを統計的に構築した上で，実行時に最適化を実現し，評価を行う．

1. 動的電源電圧制御による低消費電力化
2. CMP におけるリソース競合に着目したプロセススケジューリング

一つめの動的電源電圧制御を用いた低消費電力化は，実行時の情報をもとに電圧・周波数変更時の性能を予測し，性能低下を決められた範囲内に抑えた上でなるべく低周波数でプログラムを実行し消費電力を削減するものである．プロセッサの周波数を変更したときの性能変化の割合は，実行しているプログラムがどの程度メモリバウンドであるかによって大きく異なる．また，プログラムがどの程度メモリバウンドかというのは，データセットのサイズと実行しているプロセッサのキャッシュサイズとの兼ね合いなどによって動的に変化するため，静的な予測は困難であり，動的に性能を予測することが重要となる．

二つめの CMP におけるリソース競合に着目したプロセススケジューリングは，実行時の情報をもとに CMP 上で実行している複数のプロセス同士のリソース競合による性能低下の割合を予測し，性能低下の割合の公平性（Fairness）の改善や，トータルスループットの向上などの最適化を行うものである．あるプロセスがリソース競合によってどの程度性能低下を起こしているかを予測するためには，そのプロセスの共有リソースに対するアクセス頻度とともに，同時に実行している他のプロセスのアクセス頻度を観測する必要があるため，一つめの最適化問題と同じく動的な予測が重要になる．

これら二つの評価結果より，統計的な学習を用いた最適化モデルの構築による実行時最

適化手法が実機のプラットフォーム上で有効に働くことがわかった．これは，実行時における最適化が重要な問題に対して，最適化の指針を与えるための効果的なモデルが，プロセッサ内のハードウェア情報を用いることによって統計的に構築することができたことを意味している．これらの評価結果から，複雑化していく計算機システムにおいて，統計的な学習を用いたモデル化による実行時最適化手法が有効であると結論づけることができる．