A Study on Cache Memory Architecture Based on Data Compression for High/Performance Processors

岡, 慶太郎

（別紙様式 2）

氏　　名　：岡　慶太郎

論 文 名　：A Study on Cache Memory Architecture Based on Data Compression for High-Performance Processors
（プロセッサの高性能化を目的としたデータ圧縮に基づくキャッシュメモリアーキテクチャに関する研究）

区　　分　：甲

論　文　内　容　の　要　旨

現在，計算機サーバやラップトップ PC，携帯電話などの様々な電子機器システムにプロセッサが搭載されている．1970 年代初頭に世界初となるワンチップ・プロセッサが発明されて以来，その性能は半導体微細化技術の進歩とともに飛躍的に向上してきた．1990 年代には動作周波数の向上や命令レベル並列性の活用により性能を改善し，2000 年以降は複数のプロセッサコアを搭載したマルチコア方式へと進化した．さらに，近年では数千ものコアを搭載しオンチップ超並列処理を可能にする GPU（Graphics Processing Unit）が実用化され，スーパーコンピュータに代表される高性能計算機システムのみならず，カーエレクトロニクスなどの様々な組込みシステムへとその応用が拡大している．しかしながら，プロセッサ（GPU も含む）性能の改善がそのままコンピュータシステムの性能向上へとつながる訳ではない．その原因として，プロセッサー主記憶間の性能差の拡大（いわゆるメモリウォール問題）の深刻化が挙げられる．主記憶として用いられる DRAM はプロセッサと比較して低速であるため，主記憶アクセス・レイテンシが増大する．また，半導体パッケージの I/O ピン数は物理的に制限されるため，プロセッサ性能の向上に伴い十分なメモリバンド幅を確保することがより難しくなる．その結果，主記憶アクセスが頻発するプログラムではメモリ性能がボトルネックとなり，実効性能が低下するといった問題が生じる．この問題を解決すべく現代のプロッセッサにはオンチップ・キャッシュ（以降，キャッシュと略す）が当然のように搭載されているが，依然としてメモリウォール問題は顕在化しており，その更なる性能向上が求められている．

この課題を解決すべく，本論文では，データ圧縮技術をキャッシュへと適用することでキャッシュヒット率を大幅に改善する新しいアーキテクチャを提案し，定量的評価によりその有効性を示している．本論文の第一の貢献は，汎用プロセッサに搭載されるキャッシュを対象とし，新しいデータ圧縮方式を提案した点にある．全く同じデータ値を有するキャッシュ・ラインが複数存在することに着目し，これらが単一のメモリスペースを共有するための機構を搭載する．理論的には，キャッシュ内の全キャッシュ・ラインが同一値を有する場合，キャッシュサイズ相当（例えば32 KB）のデータ量をキャッシュ・ラインのサイズ（例えば 32 B）にまで圧縮可能となる．評価の結果，従来型のキャッシュメモリに対して最大40ポイントの性能向上を得られることが明らかになった．第二の貢献は，GPUに搭載された最上位層キャッシュに適用可能なデータ圧縮手法を考案した点にある．汎用プロセッサ向けに提案された既存手法をGPU向けへと改良し，平均11ポイントの性能向上を達成することを示した．第三の貢献は，第二の貢献で提案したアーキテクチャの改善を目的に，GPUの実行モデルであるスレッドレベル並列処理

の特性に着目した新しいキャッシュデータ圧縮方式を提案した点である.GPU向けの複数データ圧縮方式を選択可能とし，アプリケーションの特性に応じて適切な圧縮方式を選択する．評価の結果，第二の貢献と比較して最大20ポイントの性能向上を達成することを示した.

本論文は 6 章から構成される．第 1 章は本研究の背景と目的を述べ，第 2 章にて関連研究を整理することで本研究の位置付けを明確にする．第 3 章では汎用 CPU 向けの提案キャッシュ圧縮について述べる．第 4 章では GPU 向けの圧縮・復元レイテンシ削減手法について論じ，第 5 章では GPU 向けの圧縮キャッシュを提案する．最後に第６章で論文をまとめるとともに今後の研究の方向性を展望する.