クラスタ型スーパスカラ・プロセッサにおける 分散投機メモリフォワーディング手法の提案

入江 英嗣
情報理工学系研究科　電子情報学専攻

1　はじめに

情報処理の中核となるマイクロプロセッサには常に性能向上が期待されている。マイクロプロセッサは、デバイス微細化、スーパパイプライニング技術、並列実行技術によって性能を向上させてきた。しかし、現行のスーパスカラ方式は、肥大したデータパスに配線遅延が大きく影響することが予想され、性能向上の限界が指摘されている。

この問題に対し、次世代プロセッサの選択肢として注目されている方式がプロセッサのクラスタ化である。クラスタ型プロセッサでは実行コアを複数のクラスタに分割し、発行キュー、レジスタ、データパス等を局所化する。処理が複数クラスタに分散することによりオーバヘッドが生ずるが、タイミング・クリティカルパスが縮まり、スーパパイプライニング効果と並列実行のより良いバランスが期待できる。一方、局所化の難しいメモリ参照処理の遅延は相対的に大きくなることが予想される。

本研究では高クロック指向のクラスタ型スーパスカラ・プロセッサについて、メモリ参照処理に注目し、“分散投機メモリフォワーディング”を提案する。提案手法では、メモリ依存予測と動的ステアリングを利用して、メモリ参照をクラスタ内に局所化する。

2　メモリ参照のオーバヘッド

キャッシュ参照遅延は、ロード命令の値を利用する命令の発行を妨げ、実行性能に悪影響を与える。実行クラスタとキャッシュの間のフロアプラン的な通信遅延や、大きなキャッシュアレイの参照等、配線遅延に耐性のない要素が含まれるため、クラスタ化

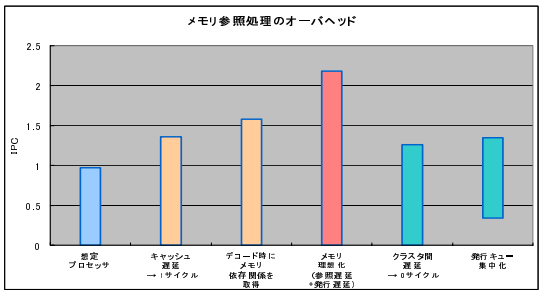

図 1: オーバヘッド要素の評価

実行コアに比べて相対的に長い遅延を生ずることが予想される。また、クラスタ型スーパスカラ・プロセッサでは、同時に処理される命令数が多く、曖昧なメモリ依存による発行遅延の影響が大きくなる。更に、メモリ命令同士が異なるクラスタに割り当てられている場合には先行するメモリ命令の発行から次のメモリ命令の発行までにクラスタ間通信遅延が加わり、更に長い遅延を生ずる。

図 1 に高クロック指向クラスタ型スーパスカラ・プロセッサをベースラインとして、各オーバヘッド要素を理想化したときの IPC を示す。ベースラインプロセッサは 1 実行幅のクラスタ 8 個によって構成されており、各クラスタは少ない遅延で稼動する。一方、フロアプラン的な遅延は相対的に大きく設定されている。メモリ参照処理オーバヘッドの影響は他の主要なクラスタ化オーバヘッドと同様に大きい。また参照遅延、発行遅延双方が大きなオーバヘッドとなっている。

3　分散投機メモリフォワーディング

クラスタ型スーパスカラ・プロセッサでは、メモリ命令の参照遅延、発行遅延共に影響が大きいため、本研究では双方を改善する投機メモリフォワーディ

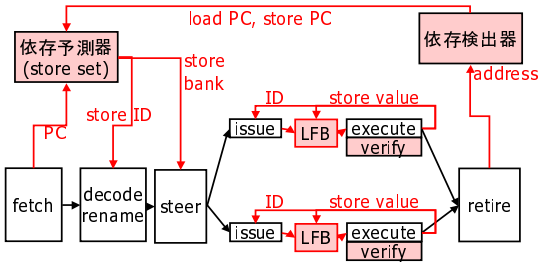

図 2: 分散投機メモリフォワーディング

ング技術 [2] に注目し、クラスタ型スーパスカラ・プロセッサへの適用を考える。

従来の投機メモリフォワーディング手法の様に、フォワーディング用のバッファを集中構成とした場合、キャッシュと同様の参照遅延を生じてしまい、ゲインが軽減されてしまう。提案手法では、各クラスタに小容量のフォワーディング用バッファ、ローカルフォワードバッファを分散配置する。更に、メモリ依存予測に基づいたステアリングを導入し、フォワーディングを局所化する。

また、先行研究におけるフォワーディングは、リザベーションステーションを想定した、データ駆動的なものであり、発行タグで制御される高クロックアーキテクチャでは新たに機構を具体化する必要がある。実行コアの高速性を保つため、提案手法では、フロントエンド処理に解析を追加し、ストア命令からコンシューマ命令へのフォワードを実現する。また、発行キューに拡張を施し、フォワードのタイミングを制御する。

提案手法の概念図を図 2 に示す。

4 提案手法の評価

分散投機メモリフォワーディングをベースライン・モデルに実装し、シミュレータによる評価を行った。提案手法は小容量の追加ハードウェアにより、全ロード命令の 35%以上に適用可能であった。図 3 にキャッシュ参照遅延を 1 サイクルから 8 サイクルまで変化させたときの IPC の様子を示す。曖昧な依存関係のみを軽減するモデルとして、ストア・セット予測 [1] を適用した場合の IPC を点線で示し、比較した。

遅延が大きくなるほど、フォワーディングの効果が大きくなっており、8 サイクルの遅延では提案手法に対して 20%以上の性能向上が得られている。キャッシュ参照遅延に対するセンシティビティに注目すると、投機を用いないベースラインでは 1 サイクル増加する毎に 3.96%の性能低下となっており、ストア・セットのみでも同様の値である。提案手法では 1 サイクルあたり 2.91%となっており、耐性が高まっている。また、参照遅延が少ない設定でも提案手法は 10%以上の性能向上を得ており、ストア・セットのみのモデルよりも高性能となっている。これは、メモリ依存に基づくステアリングにより、発行チェインの解決が早まっていること、ロードをバイパスして直接コンシューマ命令を発行していることの影響である。

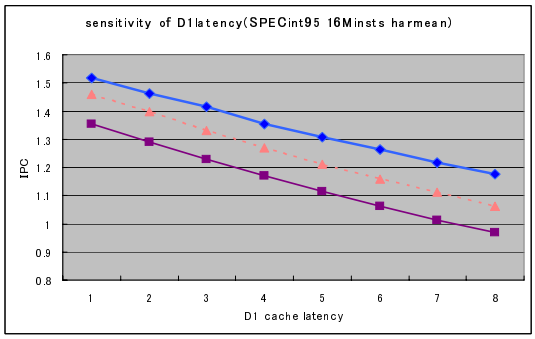

図 3: D1 キャッシュレイテンシと提案手法の効果

5 おわりに

本論文では、高速な実行コアと長い通信遅延を持つクラスタ型スーパスカラ・プロセッサを仮定し、メモリ参照処理のオーバヘッドを調べた。また、投機メモリフォワーディングの考え方を利用し、ロード命令発行の早期化と参照の局所化を実現する、分散投機メモリフォワーディングを提案した。

参考文献

[1] G.Z.Chrysos、J.S.Emer, “Memory Dependence Prediction using Store Sets”, 25th Int.Symp.on Computer Architecture、pp.142-153、Jul 1998.

[2] A.Moshovos、G.S.Sohi、“Speculative Memory Cloaking and Bypassing”、International Journal of Parallel Programming、Oct 1999。