1H-05　スーパーテクニカルサーバ SR8000 向け入出力高速化方式

鵜飼敏之†　森利明†　清水正明†　山崎康雄†　熊崎裕之‡
†(株)日立製作所 中央研究所　　‡(株)日立製作所 ソフトウェア開発本部

1. はじめに

近年コンピュータシステムで実行するジョブの規模は増大し、扱うデータ量は増加の一途をたどっている。ジョブの実行性能を高めるためには、これら大量のデータを高速に入出力することが不可欠である。

本稿では、スーパーテクニカルサーバ SR8000 向きに開発した入出力高速化方式について報告する。

2. SR8000 のファイルシステムの概要

スーパーテクニカルサーバ SR8000 は分散メモリ型の並列コンピュータである。各ノードは協調型マイクロプロセッサ機構を有する複数のプロセッサから成り、これらノードを高速ネットワークで接続する。

SR8000 用 OS である HI-UX/MPP for SR8000(以下 HI-UX/MPP)は、マイクロカーネル技術を採用し、分散メモリ型システムに対しても高性能かつ柔軟な単一システム運用を実現している。

図 1に SR8000 システムの構成ならびに HI-UX/MPP のファイルシステムの概要を示す。このファイルシステムの特徴は、分散バッファキャッシュにより入出力装置接続ノードにの負荷集中回避を可能としたことである。

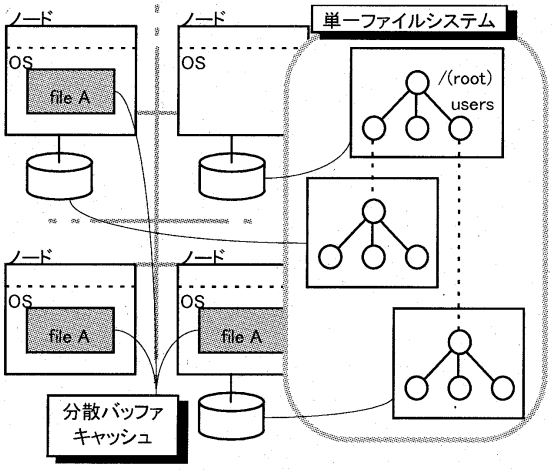

図 1 SR8000 のファイルシステム概要

High-Speed Input/Output Method on Super Technical Server Hitachi SR8000,
Toshiyuki UKAI, Toshiaki MORI, Masaaki SHIMIZU,
Yasuo YAMASAKI, Hiroyuki KUMAZAKI,
HITACHI, Ltd.

3. 直接入出力方式

通常使用では効果的な分散バッファキャッシュであるが、入出力バウンドなジョブでは、余分なデータコピーによって CPU オーバヘッドを増大させる場合もある。このため、新たに以下の 2 方式を開発した。

- **分散バッファキャッシュ管理処理削減**
- **データ量依存処理削減**

これによりCPUオーバヘッド削減を図り、特に大規模な入出力に対し効率的な処理の実現を目指した。

3.1 分散バッファキャッシュ管理処理削減

HI-UX/MPP では、分散バッファキャッシュを様々な要求処理ごとにトークンで制御する。この管理オーバヘッド削減のため、本方式ではファイルオープン時にサーバノード側でトークンを取得した後、クローズまでトークンをサーバノードで管理することにした。

オープン処理の概要を図 2に示す。

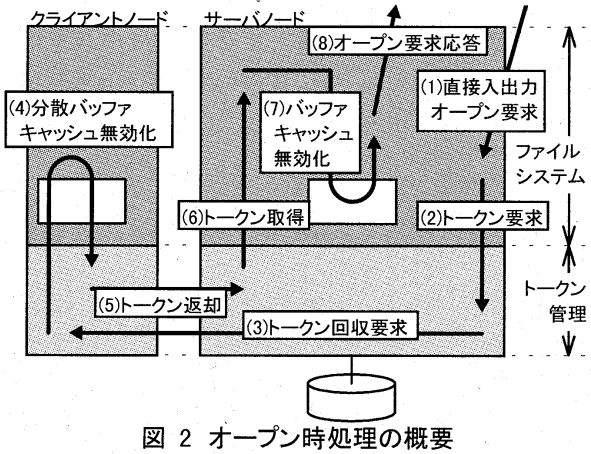

図 2 オープン時処理の概要

これにより、各入出力要求処理でトークン制御を行う必要をなくした。

3.2 データ量依存処理削減

分散バッファキャッシュ管理処理削減により、各クライアントノード上の入出力要求はサーバノードに集中する。従って、サーバノードではデータコピー等データ量依存処理を削減することにより、CPU オーバヘッドの軽減を図った。この概要を図 3に示す。

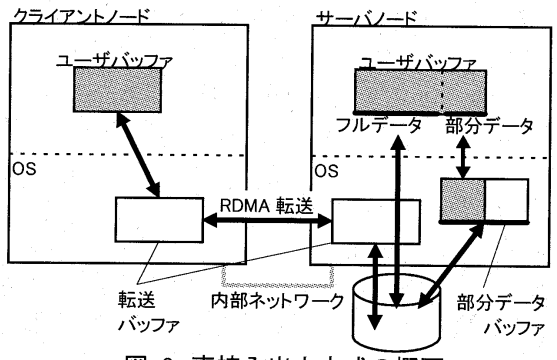

図 3 直接入出力方式の概要

入出力要求処理時に、サーバノード側で、ユーザ要求データがファイルシステム管理単位の境界条件を満たす場合(フルデータと呼ぶ)には、ユーザメモリ-装置間で直接データ転送を行う。境界条件を満たさない場合(部分データと呼ぶ)には、通常のUNIX入出力インターフェースと互換を図るため、部分データバッファを経由した入出力を行う。

また、ノードを渡る入出力(リモート入出力)では、データ送り先ノードのメモリを直接指定して転送可能なリモートDMA(RDMA)転送を直接利用し、データの高速転送を行う。

このように、従来の UNIX 入出力インタフェースとの互換性を保ちつつ、バッファキャッシュをスキップすることにより入出力処理の高速化を図った。

4. 評価

評価は、通常のバッファ入出力と直接入出力のそれぞれについて、1GB のファイルに対する逐次入出力性能を測定して行った。

図 4に磁気ディスク装置上のファイルに対するリモート入出力性能(相対値)を示す。

入出力単位が小さいときには、部分データバッファリングの介在と、CPU-ディスク装置の逐次動作に起因する処理時間増により、直接入出力の性能が劣るが、256KB を超える長大データに対しては、Read 時従来比 3 倍、Write 時でも約 2 倍の性能を達成した。

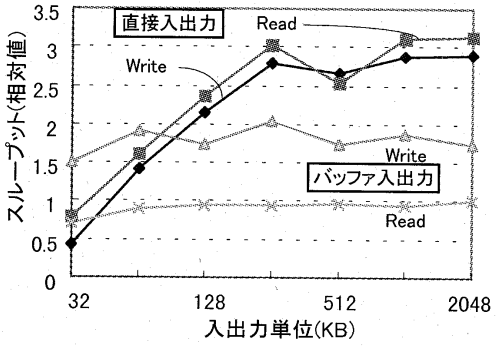

図 4 入出力性能(対リモート磁気ディスク装置)

図 4とは別に、特にバッファキャッシュ管理処理削減分の効果を明確にするための測定も行った。この結果を図 5に示す。これはメモリディスク上のファイルに対する入出力性能(相対値)で、入出力要求元と装置接続ノードが同じ場合(ローカル入出力)である。

十分に高速な装置に対しては、直接入出力ではバッファキャッシュ管理処理削減の効果により高い入出力性能が確保できていることがわかる。

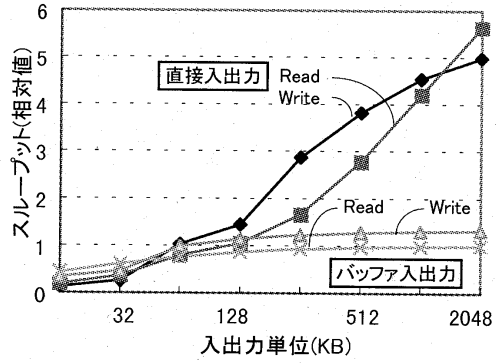

図 5 入出力性能(対ローカルメモリディスク)

5. おわりに

分散メモリ型並列コンピュータ SR8000 において OS の入出力バッファをスキップし、ユーザメモリとディスク装置との間で直接入出力する方式を開発して効果を確認した。