スーパーテクニカルサーバSR8000のノード間入出力高速化方式

B-43

鵜飼 敏之† 清水 正明† 森 利明‡
Toshiyuki Ukai Masaaki Shimizu Toshiaki Mori

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 を使用して開発した、分散メモリ型並列コンピュータ向きのノード間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、UNIX ファイルシステム(UFS)のセマンティクスに従ったファイルアクセスを実現しつつ、分散バッファキャッシュにより入出力装置接続ノードの負荷集中回避を可能としたことである。あるノードで入出力されたファイルのデータは、そのノード内メモリの分散バッファキャッシュに保持される。これにより、再利用性の高いデータや小規模データの入出力は、ノード間データ転送が効率化され、入出力性能が向上する。この分散バッファキャッシュは、トークンにより管理されており、ファイルの一貫性を確保している。

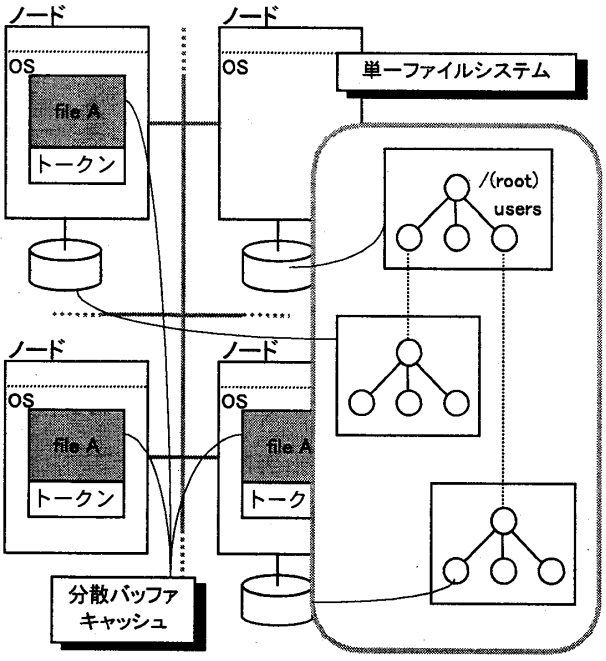

図1 SR8000 のファイルシステム概要

3．入出力高速化方式の概要

3.1 分散バッファキャッシュ

分散バッファキャッシュにより、通常の多くのケースで入出力処理性能は向上する。しかし、アプリケーションの入出力特性によっては、その性能が引き出せない場合もある。

例えば、大規模ファイルを逐次アクセスする場合、データの再利用性が小さいことに起因して、分散バッファキャッシュの管理コストのみが増加する恐れがある。

このようなケースではメモリ内のバッファキャッシュへのバッファリングをスキップする直接入出力が有効である。従って、SR8000 の入出力高速化方式として直接入出力を適用することを考え、実装方式の検討および試作を行った。

3.2 直接入出力適用の課題

直接入出力の適用に当たっては次の二点を課題として抽出した。

(A) ユーザ指定による入出力モード切り換え
(B) ノード間データ転送の低オーバーヘッド化

(A)は運用面における課題である。あるファイルの入出力モード(バッファ入出力と直接入出力)は、ユーザのモード指定に応じて、切り換えられることが望ましい。このためには、分散バッファキャッシュを実現するトークン制御をこれに適応させる必要がある。

本試作では、これを実現するため、直接入出力モードのときには、ファイルオープン時にサーバノード側でトーク

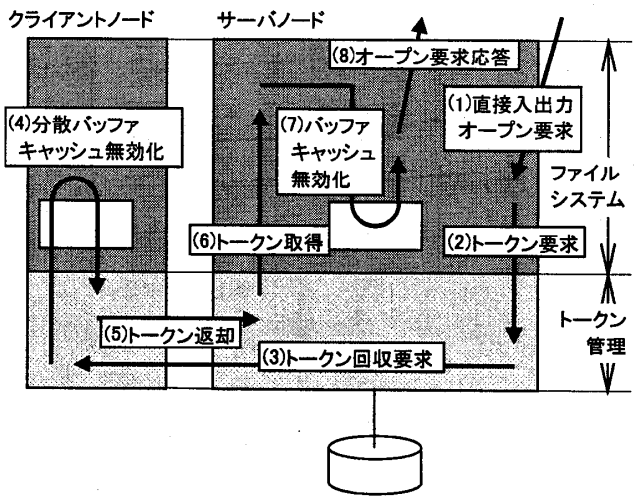

図2 直接入出力モードにおけるオープン処理の概要

†(株)日立製作所 中央研究所
‡(株)日立製作所 ソフトウェア事業部

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。直接入出力モードのオープン処理の流れを図2に示す。

(B)は性能面での課題である。これについては次節で詳細に述べる。

4. ノード間データ転送の低オーバーヘッド化

4.1 課題

HI-UX/MPP の標準のノード間 IPC(プロセス間通信)は、汎用性を重視して設計されている。このため、標準 IPC は柔軟性が高いという利点がある反面、メモリ管理に起因するオーバーヘッドが大きくなりがちである。

このオーバーヘッドの主な要因は、仮想メモリの動的な割り当て/解放が発生することや、大規模データにもかかわらずページ単位の処理が発生することである。

4.2 処理方式

上記課題を解決するため、直接入出力向けに、高速ノード間通信方式を試作した。この方式は次の二つの技術を利用している。

・オーバーライト型 IPC
・リモート DMA (RDMA) 通信

一つ目のオーバーライト型 IPC は、受信領域をあらかじめ確保し、その領域を指定してデータ転送を行う技術である。標準 IPC に比較して柔軟性は低くなるが、メモリの割り当て/解放に伴うオーバヘッドの削減が可能である。

二つ目の RDMA 通信は、SR8000 のハードウェア機能である RDMA を直接利用する通信である。RDMA 通信では専用に管理されたメモリを使用する必要があるが、ハードウェアによる通信データの保証や、低レイテンシ/高スループットのデータ転送が可能である。

本高速ノード間通信では、通信に伴う複雑な制御メッセージのやり取りをオーバーライト型 IPC で行い、実際のデータ転送を RDMA 通信で行っている。

これらを利用した高速ノード間通信処理の概要を図 3 および図 4 に示す。

いずれも、受信側でデータ転送用のバッファを確保してオーバーライト型 IPC を利用できるようにして、送信側で

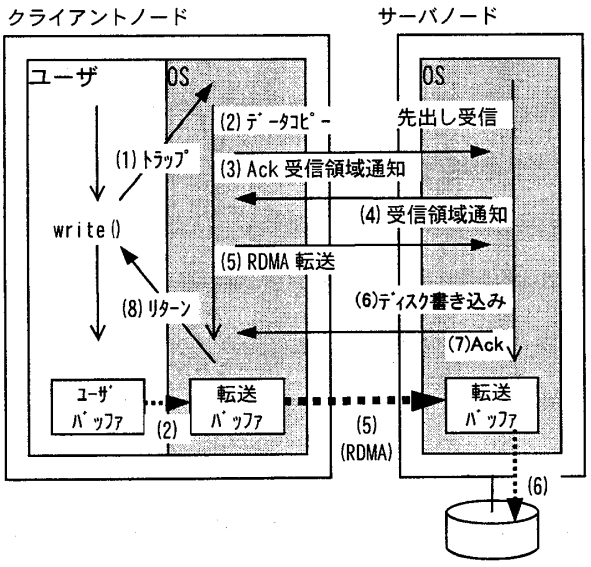

図 3　通信処理の概要 (書き込み)

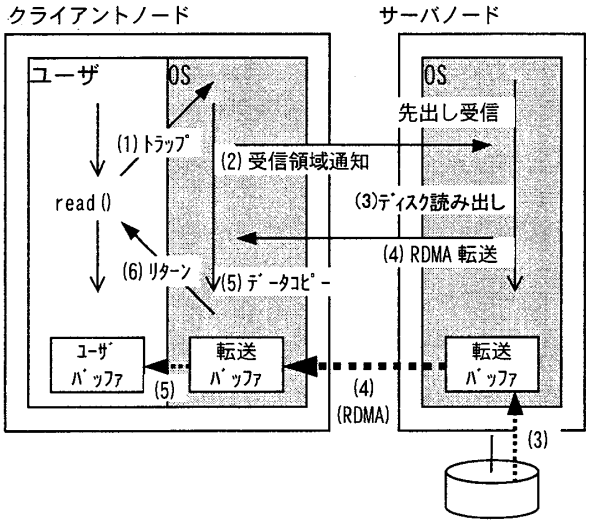

図 4　通信処理の概要 (読み出し)

そのバッファを指定した RDMA 通信を起動している。

5. 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 5 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、CPU-ディスク装置の逐次動作に起因する処理時間増などが原因で、直接入出力の性能が劣っている。しかし、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 倍の性能を達成しており、開発方式の有効性が確認できた。

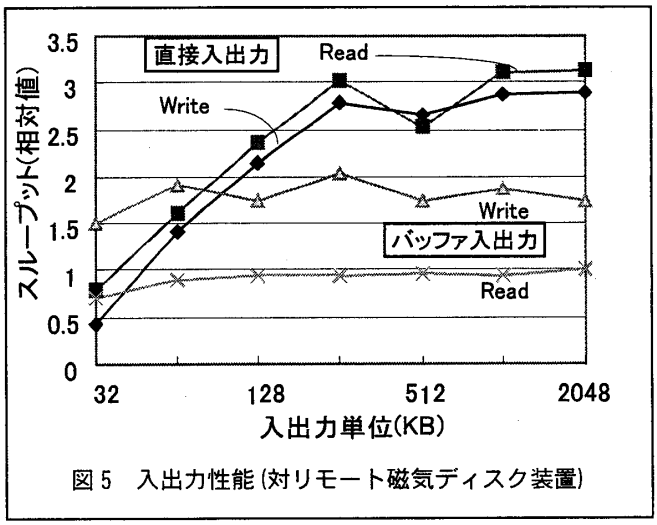

図 5　入出力性能 (対リモート磁気ディスク装置)

6. おわりに

分散メモリ型並列コンピュータである SR8000 において、直接入出力を実現するためのノード間入出力高速化方式の検討および試作を行った。その特徴はオーバーライト型 IPC と RDMA 通信を利用したことである。評価では大規模ファイル入出力で多用される長大データ入出力において、従来比 2〜3 倍の性能向上を確認した。