6D-01

~~3S-04~~ スーパーテクニカルサーバSR8000におけるノード間通信機能の実装と評価

松井 謙治　秋山 幸広　小林 耕三　菊池 正浩
(株)日立製作所 ソフトウェア開発本部

1 はじめに

スーパーテクニカルサーバSR8000におけるノード間通信機能として，ハードウェアが提供する送信主導型通信(PUT通信)により，ソフトウェアで実現した受信主導型通信(GET通信)の実装と評価を示す．

2 SR8000のノード間通信機能

大規模科学技術計算のニーズに応えるため，スケーラビリティの高い複数のノードから構成される分散メモリ型並列コンピュータシステムであるSR8000を提供している．SR8000では，物理メモリを送信,受信ノードの双方で固定的に割り当て，送信ノードのメモリ上のデータを受信ノード上のメモリに直接書き込むPUT通信とよぶ機能をハードウェアにより実装し，高速なノード間通信機能を実現している．PUT通信の動作の概念図を図1(1)に示す．PUT通信はSR8000のノード間通信の基本機能でありノード間通信を行うプログラム及びMPI(Message Passing Interface)をはじめとする通信ライブラリで利用する．

分散メモリ型並列コンピュータシステムでは，演算と通信をオーバーラップさせ，通信時間を演算時間に隠蔽することが，高いスケーラビリティを実現するために重要となる．

送信主導で行うPUT通信は，通信時間を演算時間に隠蔽することができ効率のよい実行が可能であるが，送信主導の通信にて通信時間を演算時間に隠蔽するためには，受信したデータの上書き防止のために通信毎に静的に受信領域を準備する必要があり，使用するメモリ量が増大する．使用するメモリ量を抑えるには，受信したデータの上書き防止の確認処理が必要となり，効率のよい実行ができなくなる．特に，通信ライブラリでは使用するメモリ量を抑える必要があるため本問題が顕著になる．

上記問題を解決するために，受信主導で行うGET通信とよぶ通信を提供する．GET通信の動作の概念図を図1(2)に示す．受信主導の通信により受信したデータの上書き防止のために通信毎に静的に受信領域を準備する必要がなくなるので，少ないメモリ量でも効率のよい実行が可能となる．

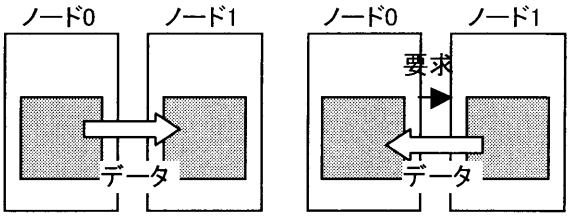


図1 PUT通信とGET通信

3 GET通信

3.1 GET通信の実現方式

GET通信は，データ受信ノードから送信ノードにデータ読み出しを要求し，送信ノードのメモリ上のデータを受信ノードのメモリに直接転送する通信である．SR8000では，GET通信を，GET通信専用に設けた割り込みと上述のPUT通信により実現している．受信ノードからPUT通信により読み出し要求の通信を行い，送信ノー

The implementation and evaluation of inter node communication facility on the Super Technical Server SR8000
Kenji Matsui, Yukihiro Akiyama, Kozo Kobayashi and Masahiro Kikuchi
Software Development Division, Hitachi, Ltd.
5030 Totsuka-cho, Totsuka-ku, Yokohama 244-8555, Japan

ドで割り込みを発生させ，割り込み処理で，送信ノードから受信ノードへPUT通信により要求されたデータ転送することでGET通信を可能とした．図2にGET通信の動作の流れを示す．

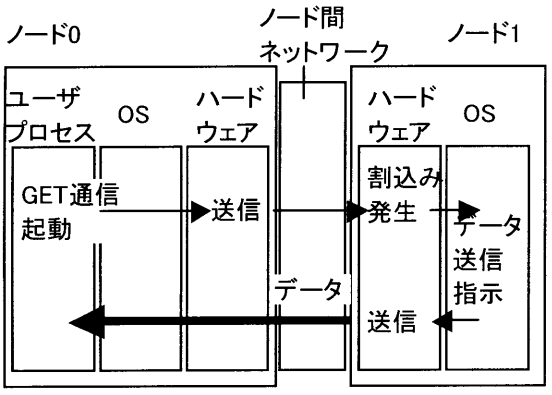


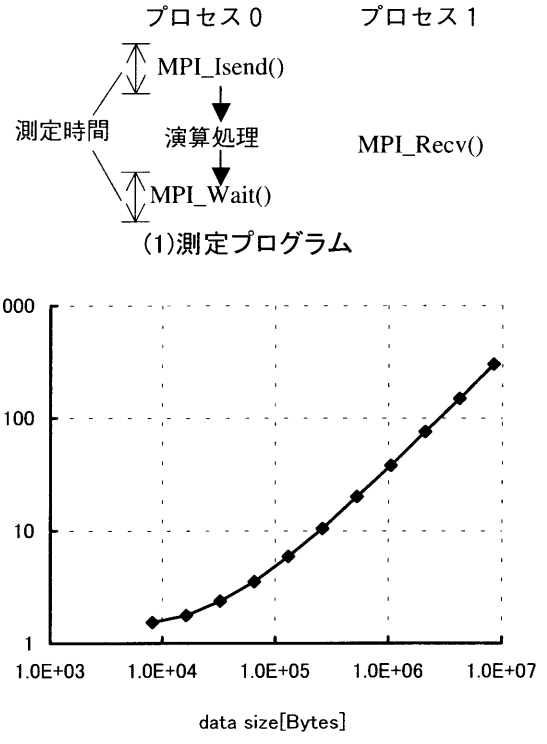
図2 GET通信の動作の流れ

3．2 GET通信の性能

GET通信はソフトウェアにより実現するため，1kBytes以下のデータサイズの通信では，通信時間に対するソフトウェア処理の占める割合が大きく，通信時間がソフトウェア処理で制約されるが，128kBytesを超えるデータサイズの通信では，ハードウェアの理論性能の90%以上を達成できる．

SR8000のMPI通信ライブラリでは，PUT通信で使用する受信バッファのメモリ量の増大を抑えるため，長大データサイズの通信にGET通信を適用する．この範囲では，GET通信は十分な性能が得られたと考える．

図3に通信と演算のオーバーラップを確認するプログラムの測定結果を示す．図3(1)は，MPIのノンブロッキング通信と演算を行う場合の通信関数の処理時間を測定するプログラムの概略図である．図3(2)は，図3(1)のプログラムの測定結果で，SR8000のMPI通信ライブラリへのGET通信適用前後の性能向上率を示している．図3(2)の横軸はMPIのノンブロッキング通信で送信するデータサイズ，縦軸はGET通信適用前に対するGET通信適用後の測定時間の比率を示している．図3(2)より通信時間の演算時間への隠蔽にGET通信が有効であるといえる．



(2)測定結果:GET通信適用前後の性能向上率

図3 通信と演算のオーバーラップの確認[1]

4 おわりに

本稿では，SR8000のノード間通信機能について示した．特に，ソフトウェアにより実現したGET通信の実装方式と性能について示した．

GET通信を適用したMPI通信ライブラリにより，大規模科学技術計算において演算と通信のオーバーラップが可能となり，プログラムの実行性能の向上を見込んでいる．

[1]測定はSR8000モデルF1(ノード間通信性能1GB/s)で行った．