計算機アーキテクチャ 136－6
（2000．1．7）

計算機クラスタにおける同期・通信コントローラの初期評価

早川 潔† 関口 智嗣†

計算機クラスタ上で効率良い並列処理を行うためには，同期および通信を低レイテンシかつ高スループットで行うことが重要である．本稿では，低レイテンシな同期・通信を実現ための新たなネットワーク（sync-comm network），同期機構および通信機構を持つ同期・通信コントローラについて述べる．本コントローラは，同期信号と通信信号を sync-comm network 上で融合させることにより，低レイテンシな同期・通信を可能にする．また，同期機構として，RBC（Reduced Barrier Code）同期機構を導入する．通信処理では，Collective 通信をハードウェアで行う．4 ノードの PC クラスタを用いてバリア同期のレイテンシを測定した結果，6.2 μs と MPI_Barrier に比べて，1／100 の結果を得た．

Preliminary Performance Evaluation on Sync-Comm controller

Kiyoshi Hayakawa †, Satoshi Sekiguchi †

Synchronization and communication via a network are integral parts of workstation/PC cluster computing. To execute parallel programs efficiently, workstation clusters need high-performance communication and synchronization support. This paper describes a new network design for low-latency synchronization and communication and descirbes the mechanism of synchronization on a synchronization and communication controller (SCC) that can be plugged into each workstation connected via a conventional LAN. Each SCC is connected by a new type of network (sync-comm network). Each node sends high-throughput data to the desired destination over a conventional LAN using TCP/IP, and synchronizes and sends low-latency data via the sync-comm network. In case of including software overhead, when SCC boards were plugged into each PC in a 4-node cluster, the barrier latency was only 6.2 μs.

1 はじめに

近年のマイクロプロセッサ技術の進歩により，クラスタコンピューティングよる効率良い並列処理が可能になっている．しかし，マイクロプロセッサ技術が進歩し，処理速度が向上するにつれて，同期や通信のオーバーヘッドが相対的に増大する．同期や通信のオーバーヘッドを削減させるため，高速なネットワークインターフェースが開発されている [3][4]．また，通信ライブラリにおいても，TCP/IP を使用した通信ライブラリを使用せずに独自の通信ライブラリを開発することにより，ソフトウェアを含むオーバーヘッドを削減している [5][6]．しかし，それらのネットワークインターフェースは高価であり，同期もパケット交換で処理するため，ある程度の同期レイテンシが生じてしまう．そこで，本稿では，バリア同期をパケット交換で行わず，電気信号を用いハードウェアで処理し，且つ，Collective 通信も行えることを可能にする同期信号と通信信号の融合ネットワーク方式を提案する．また，1 種類の同期コードでバリア領域が形成できる同期機構として，RBC 同期機構について述べる．さらに，同期・通信コントローラの初期評価として，MPI_Barrier と同期コントローラを用いたバリアとの性能評価について述べる．

†:電子技術総合研究所
Electrotechnical Laboratory

2 同期・通信コントローラの概要

2.1 ターゲットクラスタシステム

本コントローラを組み込むクラスタシステムは，分散メモリ型のマルチプロセッサシステムである．本コントローラが行う処理は，メッセージパッシング型の通信処理とバリアを拡張した同期処理である．分散共有メモリ型のマルチプロセッサシステムでの

処理（リモートメモリアクセスやセマフォなどの排他制御など）は行わない.

T３Eなどの分散メモリ型MPPシステム[1]は,同期や通信のオーバーヘッドを削減するため,高スループットなネットワークと低レイテンシなネットワークの２種類のネットワークを搭載している.本コントローラを用いることにより,高スループットを要求する処理は従来のネットワーク,低レイテンシが要求される同期や小量データパケットは sync-comm network を使用することが可能となり,より効率良い並列処理が実現できると考えている.

2.2 同期・通信コントロールボード

同期・通信コントロールボードは,PCIのプラグインボードであり,同期・通信コントロールLSIおよび sync-comm newtwork 用コネクタ（Link_A および Link_B）のみで構成されている.同期・通信コントロール LSI は,CPLD（アルテラ Flex10K50）上に同期・通信処理ハードウェアを組み込む形で実現される.sync-comm newtwork 用コネクタは,0.5mm ピッチの 40 ピンのパラレルケーブル用コネクタである.sync-comm newtwork 用コネクタの Link_A と Link_B をデージーチェーンで接続する（図 1参照）ことにより,ハブ等の外部接続装置を必要とせずに,sync-comm newtwork が構成可能である.

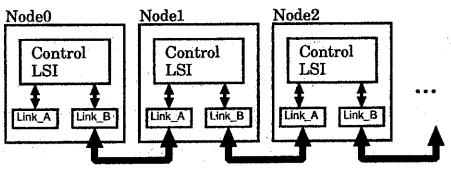

図 1: Sync-comm Network

2.3 同期・通信コントロール LSI

図 2に同期・通信コントロールLSIのブロック図を示す.同期・通信コントロールLSIは,PCIバス-内部バスブリッジ（PCI BUS-Internal BUS Bridge）,同期制御部（Synchronization Control）,通信制御部（Communication Control）,およびネットワーク制御部（Network Control）で構成される.内部バスは,ネットワーク制御部以外のブロックを接続する.同期制御部はネットワーク制御部と同期信号線で結合され,通信制御部はネットワーク制御部とパラレルデータ線で結合される.

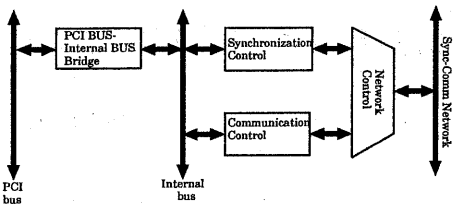

図 2: 同期・通信コントロール LSI

PCIバス・内部バスブリッジは,PCIバスプロトコルと内部バスプロトコルをインターフェースし,PCIのマスター／ターゲット機能もサポートする. 同期制御部は,sync-comm network を使用した同期処理アクセスを行い,任意参加バリア,Fuzzy バリアおよび RBC 同期機構の同期を処理する. 通信制御部は,パケットデータ処理を行う.パケット長は 64bytes で固定する.処理するトランザクションは,*Send-Receive, Reduce, Broadcast* および *Scan* である.

ネットワーク制御部は,同期信号と通信信号とを sync-comm network 上で融合させるための制御を行う.

3 同期信号と通信信号の融合

先に述べた T3E などの MPP システムは,低レイテンシな同期を行うため,低レイテンシ用ネットワークを使用し,同期パケットをハードウェアで処理している.これは,バリア同期の他に,Test & Add などの排他制御系の同期や小量データパケットにも対応できるためである.しかし,パケット転送で同期をハードウェアで処理する方法では,パケット転送のオーバーヘッドやルーティングロジックのコストがかかってしまう.

対象とする同期を拡張バリア同期のみに絞った場合,簡単にかつ高速に拡張バリア同期を処理できる.つまり,パラレルケーブルをバリア到達信号線とし,それらの信号の AND をとることにより,拡張バリア同期を処理できる.この方法により,高速でかつルーター等のハードウェアコストがかからずに同期を処理できる.

同期処理時の同期信号線の用途と通信処理時のパラレルデータ信号線の用途の両方の用途に sync-comm network を対応させるために,ネットワー

ク制御部で Link_A と Link_B との接続を切替える. ユーザープログラムでフラグをセット・リセットすることにより, 切替が行われる. データ通信時には,Link_A と Link_B は接続せず, 通信制御部に Link_A と Link_B の信号線それぞれを接続させる（図 3参照）. 同期時には,Link_A と Link_B とを接続させるとともに, 同期制御部にそれらの信号線を接続させる.

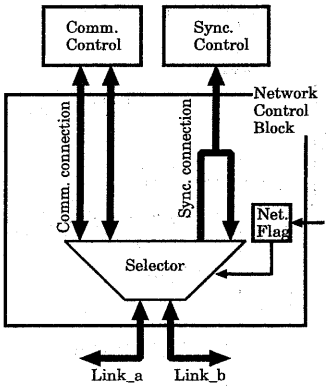

図 3: ネットワーク制御部の構成

3.1 sync-comm network での通信信号

通信時における sync-comm network の信号線は, 図 4に示すとおり, 双方向のパラレルリンクとして定義される. 各ノードは, 隣接ノード間の 6 ビットデータ送受信を同時に行うことができる. スループットは, 11.1MBytes/s である.

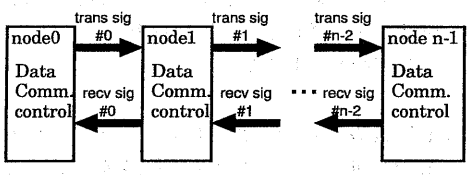

図 4: 通信時の信号定義

3.2 sync-comm network での同期信号

同期時における sync-comm network の信号線は, 図 5に示すとおり, バスとして定義される. 同期時の信号線は, 最大 16 台で構成されるマイクロクラスタ内での同期信号線とそのマイクロクラスタ間で同期を行うための同期信号線に分かれる.

同期成立の検出は, まず, マイクロクラスタ内で同期成立を検出し, その後にマイクロクラスタ間で同期成立を検出する. マイクロクラスタ内では, sync_n (n はマイクロクラスタ内でのノード番号) を使用し, マイクロクラスタを構成する各ノード間の同期成立を検出する.

sync_n 信号は, マイクロクラスタ内ノード番号 n のノードが同期ポイントに到達していることを示す信号である. この信号線は, マイクロクラスタ内のみのバス接続である.

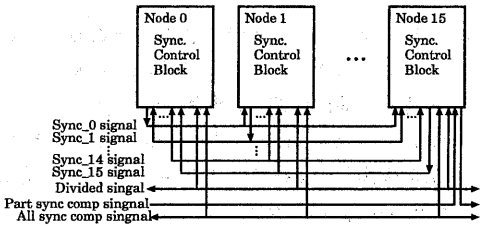

図 5: 同期時の信号定義

マイクロクラスタ間では,Divided 信号,Part-sync-comp 信号,All-sync-comp 信号を使用して, クラスタ間の同期成立を検出する. Divided 信号はターゲットクラスタシステムがマイクロクラスタに分割されていること（ノード数が 17 台以上であること）を示す信号,Part-sync-comp 信号はクラスタ内の同期が成立したことを示す信号,All-sync-comp 信号は, 全体の同期が成立したことを示す信号である.Divided 信号線,All-sync-comp 信号線は, クラスタ全体のバス接続である. Part-sync-comp 信号線は, デージチェーン接続である.

4 同期制御部の機能

同期制御部は, 同期成立検出回路 (synchronization completion detector) および R-BC 回路で構成される（図 6参照）. 同期制御部内のフラグ・レジスタアクセスや同期成立検出回路・後述する RBC 回路の制御は, メモリアクセスによって行われる.

同期成立検出回路には, マクロクラスタ内での任意参加バリアをサポートするために, 同期グループレジスタ（sync-group reg）(16 ビット) が用意されている. また, 同期の成立を知らせるために, 同期フラグ（sync. flag）が用意されている.

同期成立検出回路は, 以下の処理を行う.

- 同期フラグがセットされた時,sync_n 信号をアクティブにする.
- クラスタのノード数が 16 台以下の場合, 同期に参加しているノードの sync_n 信号が全てアクティブの場合, 同期フラグをリセットする.
- クラスタのノード数が 16 台以上の場合,All-sync-comp 信号がアクティブの場合, 同期フラグをリセットする.
- クラスタのノード数が 16 台以上の場合, マイクロクラスタ内でのノード番号が最大ノードの同期成立検出回路が, 前隣のマイクロクラスタの Part-sync-comp 信号と sync_n 信号で同期成立を検出し, 後隣のマイクロクラスタに Part-sync-comp 信号を送る. だだし, 一番後ろのマイクロクラスタ内でのノード番号が最大ノードの同期成立検出回路は, 前隣のマイクロクラスタの Part-sync-comp 信号と sync_n 信号で同期成立を検出し,All-sync-comp 信号をアクティブにする.

任意参加バリアおよび Fuzzy バリアの同期成立は, 同期成立検出回路のみで検出される.RBC 同期の成立は,RBC 回路と同期成立検出回路を使用して検出される.

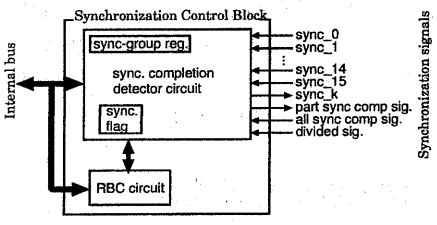

図 6: 同期制御部の構成

4.1 RBC 同期

RBC 同期は, 以下に示す規則に従い同期動作を行う.

「あるノードのある同期ポイントの実行順番を k とすると, その k 番目の同期ポイントを通過するには, その直前の $k-1$ 番目の同期ポイントと他ノードの $k-1$ 番目の同期ポイント全てが通過されていなければならない」

例えば, 図 7のように, ノード 1が実行する $(n-1)$ 番目の RBC 同期コード実行時刻より早い時刻に, ノード 2が n 番目の RBC 同期コードに到達した場合 (t1 の時), その RBC 同期コードの終了は, ノード 1 が実行する $(n-1)$ 番目の RBC 同期コード実行終了直後 (t2 の時) となる.

バリア領域は,$k-1$ 番目の同期ポイントと k 番目の同期ポイントとの間に形成される.

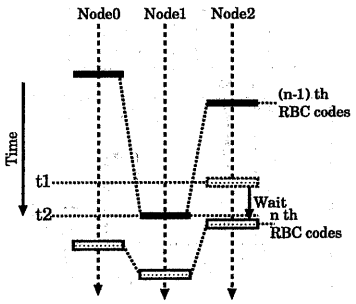

図 7: RBC 同期の例

RBC 同期は,RBC 回路内で以下に示す動作を行うことにより実現される.

1. ノードプロセッサが RBC 回路をアクセスすることにより,RBC 回路が同期成立検出回路内の同期フラグをチェックする.
2. 同期フラグがセットされている場合,RBC 回路はノードプロセッサに 1 を返す.
3. 同期フラグがセットされていない場合,RBC 回路はノードプロセッサに 0 を返とともに, 同期フラグをセットする.

RBC 同期ポイント到達時に, ノードプロセッサは,RBC 回路にアクセスし, 0 が返ってくるまで,spin-waiting をして待つ.

RBC 同期の特長として, 以下の点が挙げられる.

- 同期コード 1 種類でバリア領域を形成できる.
- Fuzzy バリアの出口コード直後に入口コードが挿入されている場合, それらのコードを RBC 同期コード 1 つに置き換えることができるので, 同期コード数を削減できる場合がある.

5 通信制御部の構成

両隣接ノード間のパケット送受信を同時に行うために, 通信制御部では,4 つメッセージバッファを用意する（図 8参照）. 各リンクには, 送信バッファ（trasmit buffer）と受信バッファ（receive）が用意されている. バッファサイズは,64Bytes である. パケット構成は, パケットヘッダ 8bytes, ペイロードが 56Bytes である.

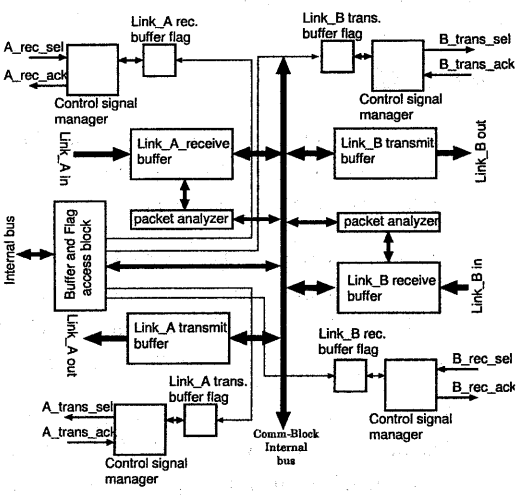

図 8: 通信制御部の構成

Link_A_in および Link_A_out は前のノードの Link_B_out および Link_B_in に, Link_B_in および Link_B_out は後のノードの Link_A_out および Link_A_in にそれぞれ接続されている.

パケット解析部（packet-analyzer）は, 受信パケットヘッダを解析し, 他のノードに送られるパケットと判断した時のみ, 送信バッファにそのパケットを転送する. ルーティングは, カットスルールーティングを採用し, ヘッダが届いたら直ちにヘッダ解析を行い, パケットの種類に従った処理を行う.

パラレルリンクでは,sel 信号および ack 信号を用いた 6 ビットの非同期パラレルデータ転送が行われる. 制御信号管理部（Control signal manager）がそれらの信号を制御する.

6 初期性能評価

この章では, バリア同期のハードウェアレイテンシ予測性能および PC クラスタを使用したソフトウェアレイテンシを含んだバリア同期のレイテンンについて述べる.

6.1 バリア同期におけるハードウェアレイテンシの予測性能

図 9にバリア同期時のハードウェアレイテンシの予測性能を示す. ハードウェアレイテンシの予測値は, アルテラ社の MAX+PLUSII で見積もった各回路の最大遅延時間を元に計算した. ここでいうハードウェアレイテンシは, 各ノードが同時に同期フラグをセットしてから, その同期フラグがリセットされるまでの時間である.

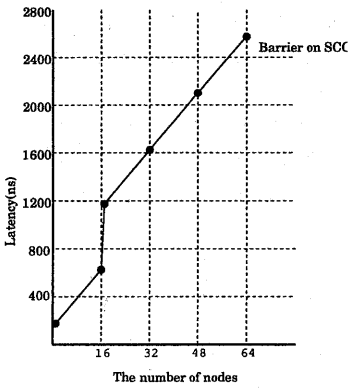

図 9: バリア同期におけるハードウェアレイテンシの予測性能

16 台では, ハードウェアレイテイシが約 600ns と非常に低レイテイシなバリア同期が可能になると予測される.17 台以上になるとマイクロクラスタ間の同期処理（540ns）が加算されるが,64 台でも 2.4μs と非常に低レイテイシな値が予測される.

6.2 PCクラスタを用いたバリア性能評価

本コントローラを 4 台構成の PC クラスタに実装し, ソフトウェアを含む本コントローラを使用したバリア同期の性能を計測した.PC クラスタの構成を表 1に示す. 比較対象として,MPICH（ver.1.2）を PC クラスタに実装し,100BASE-TX Ethernet を使用した MPI_Barrier のレイテンシを計測した.

表 1: Feature of PC cluster

CPU	Pentium II 333MHz
Memory	64Mbyte
Chipset	Intel 82440EX
NIC	Fast Ethernet PCI board
HUB	Fast Ethernet Switching HUB
OS	Redhat Linux5.2 Kernel Ver.2.0.36

図 10にノード数とバリア同期レイテンシの関係を示す. 本コントローラを使用したバリア同期のレイテンシは,4 ノードまではノード数に関係なく,3.2μs または,6.2 μs である. このシステムは, ほぼ 3μs に一回の割合で同期フラグにアクセスできる. しかし, 先に述べたハードウェア見積りに比べて非常に遅いので, 図 10に示す結果を得たと推測される. バリア同期のハードウェアレイテンシの見積りが 64 ノードでも 3μs を越えないため, 数十ノードのバリア同期でも,3.2μs または, 6.2 μs のレイテンシで処理できると推測される.

一方,MPI_Barrier は, ノード数が増えるにつれて直線的に増加し,4 ノードでは, 本同期コントローラのバリア同期レイテンシに比べほぼ 100 倍のレイテンシを得た.

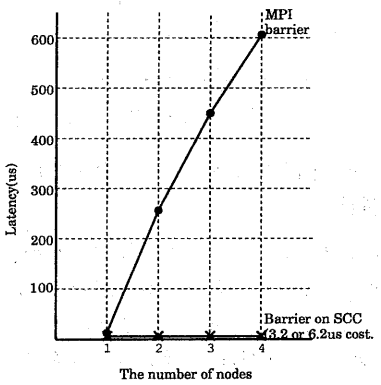

図 10: PC クラスタ上でのバリア同期レイテンシ

7 おわりに

バリア同期を電気信号を用いハードウェアで処理し, 且つ, Collective 通信も行えることを可能にする同期信号と通信信号の融合ネットワーク方式を提案した. また,1 種類の同期コードでバリア領域が形成できる同期機構として,RBC 同期機構について述べた. また, 同期・通信コントローラの初期評価として, MPI_Barrier と同期コントローラを用いたバリアとの性能評価について述べた. 同期・通信コントローラを用いたバリア同期のレイテンシは,100BASE-TX を使用した MPI_Barrier より最大で 100 倍の性能向上を得た.

今後の課題は, 通信も含んだ初期性能評価およびアプリケーションを使った性能評価を行うことである. また,BSP (Blus Synchronous Parallel)[7] のような本コントローラに適したコンピューティングモデルを提案していく予定である.

参考文献

[1] Steven L. Scott,"Synchronization and Communication in the T3E Multiprocessor", ASPLOS VII, pp.26-36, Oct.1996.

[2] S.Shang, K.Hwang,"Distributed Hardware Barrier Synchronization for Scalable Multiprocessor Clusters," IEEE, Trans, Parallel Distib. Syst., vol.6, pp591-605, June.1995.

[3] Nanette Boden, Danny Cohen, Robert Feldrman, Alan Kulawik, Chrle Seitz, Jakov Seizovic, and Wen-King Su, "Myrinet – A Gigabit per Second Local Area Network",*IEEE Micro*, vol.15, No.1, pp.29-36, Feb.1995.

[4] 西 宏章, 多昌 廣治, 工藤 知宏, 天野 英晴,"仮想チャネルキャッシュを持つネットワークルータの構成と性能", 並列処理シンポジウム *JSPP'99* , pp.71-78, Jun.1999.

[5] H.Tezuka, A.Hori, Y.Ishikawa, and M.Sato, " PM: An Operating System Coordinated High Performance Communication Library.",*High-Performance Computing and Networking*, vol.1225, pp.708-717, Apr.1997.

[6] S.Pakin, M.Lauria, and A.Chein, "High Performance Messaging on Workstations: Illinois Fast Messages (FM) for Myrinet",In Proc. of Supercomputing '95, 1995.

[7] http://www.bsp-worldwide.org/bswwact.htm