分散型共有メモリを用いた高速メッセージ通信システムSURE-SXの研究試作 ―通信制御方式―

1L−9

松平直樹　加藤光幾　新家正総　陣崎　明
(株)富士通研究所

1．はじめに

富士通のフォールトトレラントコンピュータSURE SYSTEM 2000をFDDIで結合し、装置間で2700メッセージ/秒以上(11MByte/秒以上)の性能を実現するメッセージ通信システムSXを試作した。本論文では、分散型共有メモリを用いてメッセージ通信を実現する通信制御方式、及び制御ソフトウェア構成について述べる。

2．SXの通信制御のアプローチ

一般に、確実な通信を保証する通信制御処理は主としてソフトウェアにより実現されてきた。ところが、高速な伝送路を用いた通信では、このソフトウェア処理がボトルネックとなり、伝送路速度に見合った通信性能の達成が困難となる。我々が行ったアプローチは、従来提案してきたネットワーク仮想記憶方式(NET-VMS)[1]をメッセージ通信に適用する事により通信制御をハードウェア化し[2]、さらにソフトオーバヘッドを削減できるように通信制御を簡略化することである。

3．通信制御ハードウェアの動作

SXではパケット通信を共有仮想記憶空間におけるページアクセスで実現可能とした(図1)。共有仮想記憶空間は制御メモリにより実現され、共有仮想記憶空間のアドレス毎に、ページ状態を格納する通信制御タグと、メッセージを格納するバッファのポインタから構成される。パケットは、共有仮想記憶空間のアドレス、コマンド、gap、応答、メッセージ、CRCから構成した。

パケットを受信すると、通信制御ハードはパケットを中継しながら格納されている宛先で制御メモリを参照する。宛先は、装置アドレスに対応させた共有仮想記憶空間上のアドレスである。制御メモリ参照により出力されるページ状態と、パケット内のコマンドを用いて、データ受信等の動作を決定する。この演算はフレームのgapが通過する時間(約0.6μs)で実現し、その演算結果をフレームの応答領域に上書きして設定する。

SURE-SX: The High-speed Message Communication System using Distributed Shared Memory -Communication Protocol-
Naoki Matsuhira, Koki Kato, Tadafusa Niinomi, Akira Jinzaki
Fujitsu Laboratories Ltd.

4．ページ状態、コマンドの定義

ページ状態は以下の7状態とした。

NA:ページにバッファが割り付けられていない。
VP:ページにデータ未設定のバッファが割り付けられている。
SD:ページに送信データが設定されたバッファが割り付けられている。
SW:SD状態のページを送信しようとしたが、受信側がバッファビジーのため送信待ち。
SC:SD状態のページを送信完了。
RD:ページに、受信データが設定された実メモリが割り付けられている。
RW:RD状態のときに別の送信が行なわれた。送信側は、SW状態となる。

またコマンドは、SEND(SD状態のページを送信する)、READY(SW状態のページに対して受信準備完了を通知する)などとした。

5．ページ状態遷移

ページ状態の遷移は、制御ソフトが行う場合、送信したコマンドに対する応答により通信制御ハードが行う場合、及び、受信したコマンドとコマンド受信時のページ状態により通信制御ハードが行う場合の3通りがある。ページ状態遷移は、エラー制御やフロー制御も含めた通信を高速に実現できるように決定した。ページ状態の送信側の遷移を図2に、受信側の遷移を図3に示す。

通常の通信は、以下の制御で実現した。(0)自装置アドレスに対応する仮想ページに受信バッファを割り付け、ページ状態をNAからVPに変更する。

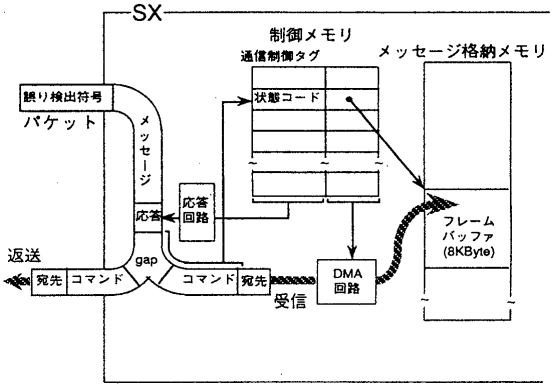

図1　SXの通信制御ハードウェア構成

(1)送信側SXの制御ソフトは、送信メッセージを格納したバッファをメッセージの宛先に対応するページに割り付け、ページ状態をNAからSDに書き換え、通信制御ハードに送信指示する。(2)受信側SXの通信制御ハードは、パケットを受信する。自装置に対応するページ状態がVPのとき、メッセージを割り付けられているバッファにDMAを開始し、成功の受信応答を行う。メッセージにエラーが無ければページ状態をVPからRDに書き換え、制御ソフトウェアに受信通知する。(3)伝送路を一周したパケットを受信した送信側SXの通信制御ハードは、応答領域が成功であるため、ページ状態をSDからSCに変更する。その後、制御ソフトに送信終了通知する。

6．エラー制御

SXのエラー制御は、以下の制御で実現した。(番号は、5節に対応)

(2)'パケットを受信した受信側SXの通信制御ハードは、データエラーを検出すると、ページ状態を書き換えず、制御ソフトへの通知も行わない。従って、制御ソフトからは何も受信しなかった様に見える。(3)'伝送路を一周したパケットを受信した送信側SXの通信制御ハードもデータエラーを検出する為、ページ状態をSDのままにし制御ソフトに送信エラーを通知する。制御ソフトは、(1)の制御に戻り再送する。

7．フロー制御

SXのフロー制御は、バッファを最大パケット長以上のサイズで管理して、バッファオーバフローが起こり得なくている。手順を以下に示す。

(2)"受信側SXの通信制御ハードがパケットを受信し、自装置のページ状態がRDのとき受信出来ないため、受信不可の受信応答を行いページ状態をRWに更新し、受信不可を制御ソフトに通知する。制御ソフトは、受信不可を記録する。(3)"伝送路を一周したパケットを受信した送信側SXの通信制御ハードは、受信応答が受信不可のため、ページ状態をSWに更新し、制御ソフトに送信待ちを通知する。制御ソフトは、記録だけを行う。(4)受信側SXの制御ファームは、受信可能なバッファができた時点で、受信バッファを割り付けページ状態をVPに書き換える。更に、受信要求の記録を基に、受信可能を送信側に知らせる為にREADYコマンドを送信する。(5)READYコマンドを受信した、送信側のSXの通信制御ハードは、ページ状態をSWからSDに変更し、制御ソフトに通知する。制御ソフトは、送信指示を行う。

8．制御ソフトウェア構成

上記で示した制御ソフトはC言語で記述した。また、ボトルネックとなる要因を排除するために以下の構成とした。第一に、オーバヘッドの大きい割り込みを用いず、全てポーリングを用いた。第二に、全体の処理をポーリングによるイベント検出処理と、イベント駆動処理に分割した。第三に、待ち行列等の検索を不要とするようなデータ構造を用いた。

9．まとめ

SXの通信制御方式について述べ、通信制御を簡略化したことによるメッセージ通信の高速化、高信頼化技術を明かにした。文献3は、本システムの性能評価を述べている。

参考文献

[1]陣崎他:オブジェクト共有型分散オペレーティングシステムの構想、情処技報89-OS44-9(1989-9)等
[2]加藤他:分散型共有メモリを用いた高速メッセージ通信システムSURE-SXの研究試作-システムアーキテクチャ-、本大会予稿
[3]新家他:分散型共有メモリを用いた高速メッセージ通信システムSURE-SXの研究試作-性能評価-、本大会予稿

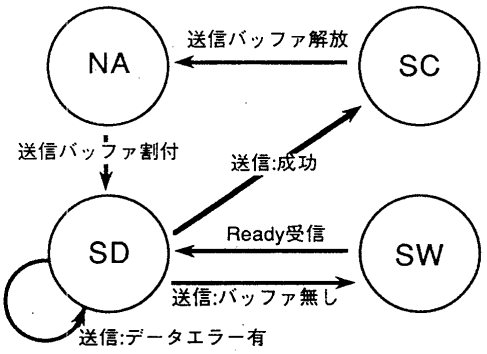

図2　送信側のページ状態遷移

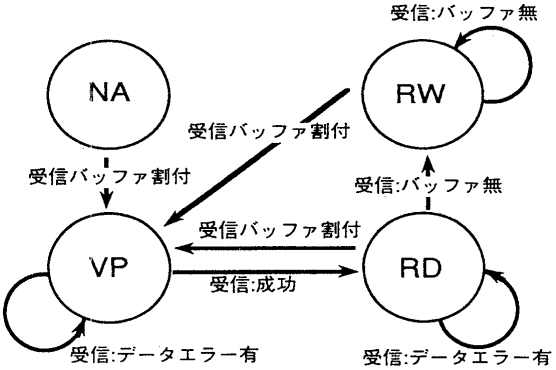

図3　受信側のページ状態遷移