1L−8

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 ーシステムアーキテクチャー

加藤光幾　松平直樹　新家正総　陣崎　明
(株)富士通研究所

はじめに

我々は富士通のフォールトトレラントコンピュータSURE SYSTEM2000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SURE-SXを研究試作した。SXはネットワーク仮想記憶方式(NET-VMS)を用いた分散型共有メモリでメッセージ通信を実現する。本稿ではSXのシステムアーキテクチャ、ハードウェア構成について述べる。

SUREによる大規模分散システムの構築

SURE SYSTEM2000はPM(プロセッサモジュール)を複数搭載する疎結合型マルチプロセッサである。PM間の独立性を高め障害の伝播を最小限に抑えるために、PM相互間の通信にメッセージ通信を用いている。

SUREは搭載するPM数を増やして段階的にシステム性能を増大できるが、SURE一台に搭載できるPM数はPM間を結合するバスの駆動能力により制限される。メッセージ通信を用いてより大規模分散システムを構築するために、多数のSURE装置間をネットワークで結合し、SURE内のバスがSURE装置間に延長されたようにメッセージ通信を行う、メッセージ通信システムSURE-SXを試作した。

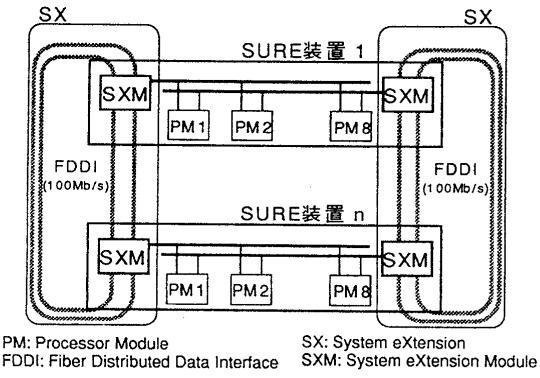

図1　SURE-SXシステム構成

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

分散型共有メモリとメッセージ通信

SXは各SURE装置に搭載するモジュールSXMとSXM間を接続するFDDIネットワーク(ISO9314、100Mbit/秒)からなる(図1)。SXはSURE間でのメッセージ通信を、SXが持つ共有メモリを用いて行う。図2にその概念図を示す。他SURE装置のPM宛のメッセージをSXMがそのままの形で共有メモリに格納する。他SUREへのメッセージ転送はSXが自動的に行うので、PMはSXを意識する必要がない。

分散型共有メモリは、従来から提案しているネットワーク仮想記憶方式(NET-VMS)[1]を用いて実現した。NET-VMSは各SXMが持つ仮想記憶システムをブロードキャストネットワークで結合し、全体を一個の共有仮想記憶システムとして構成する(図3)。NET-VMSを用いると通信制御の大部分をハードウェア化できるので高速にメッセージ通信を行える。

SXの通信能力

SURE装置内のPMはコモンバス2本で結合されており、各コモンバスは最大4KByteのメッセージを2500メッセージ/秒の速度で通信できるチャネル4本からなる。SXで接続PM数に比例するシステム性能を得るには、少なくともSURE装置内チャネルと同等のメッセージ通信性能が必要である。そのためSXがメッセージ処理能力2500メッセージ/秒以上、通信遅延2.5ms以下を達成するように設計した[3]。SXをSURE装置に複数実装したときは、実装数に比例した量のメッセージを通信できる。

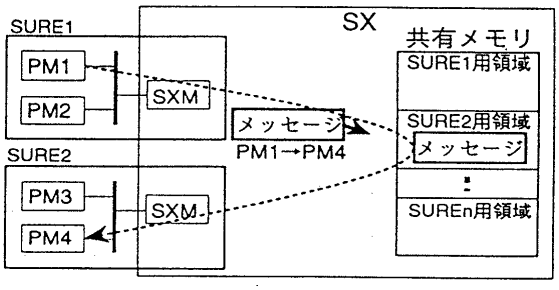

図2　共有メモリを用いたメッセージ通信

SXハードウェア構成

SXMの構成を図4に示す。仮想記憶用制御メモリはアドレス変換メモリと通信制御タグからなる。MBCはコモンバスと実メモリのインタフェース機能を持つ。制御CPUはG_{MICRO}100を用いている。仮想メモリをページ単位に管理し、メッセージ(最大4KByte)を1ページに納められるようにした。今回の試作システムではSUREを64台(PMが512台)まで接続できる。

SXは以下のように動作する: (1)MBCは他SURE装置宛のメッセージを検出すると、実メモリにDMA転送し、制御CPUに通知する。(2)制御CPUは(1)のページを、宛先SURE装置に対応させた仮想アドレスに割りつける。(3)そのページデータを宛先SURE装置のSXMに転送する。(4)宛先SURE装置のSXMはページデータが転送されたことを検出すると、MBCを起動して共有メモリのデータ(メッセージ)をバスに送信する(図3)。

SXの高速化技術

SXは次の2点により高速の通信制御を実現した: (a)送信フレームの一周回で、データを送信し、受信側から応答を受ける。(b)制御CPUはハードウェアを起動した後、直ちに次の処理に移れるので、ハードウェアと制御ソフトウェアが並列に動作できる。

(a)は以下のように実現されている。前述の(3)のページデータの転送は制御CPUがネットワーク回路を起動することにより行われ、そのページを含むFDDIフレーム(図5)が送出される。他のSXMで、FDDIアダプタを通過する送信フレームのSXヘッダ部まで受信すると、ネットワーク回路が制御メモリをアクセスしコマンド(ページデータ送信、要求など)を実行できるかをギャップが通過する間に(0.64 μs)調べる。結果をon the flyで応答部に書き込むと共に、可能ならコマンドを実行する。応答部を書き換えられたフレームは次のSXMに送られる。送信フレームがネットワークを一周して送信SXMに戻って来ると、ネットワーク回路が応答部とCRCを検査し、受信終了と伝送エラーの有無を制御CPUに通知する。

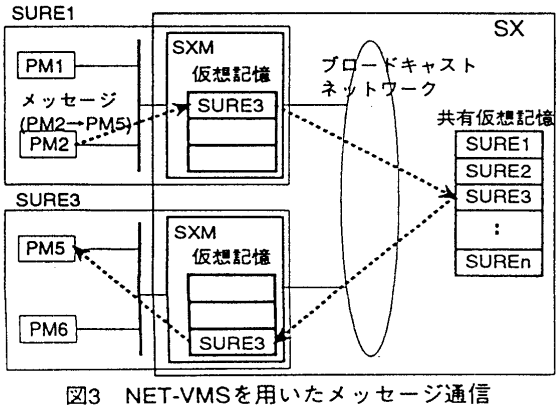

図3　NET-VMSを用いたメッセージ通信

終わりに

SXのシステムアーキテクチャについて述べ、分散型共有メモリに基づくメッセージ通信技術の概念とハードウェア化による高速処理技術について説明した。SXの通信制御方式については[2]、測定結果については[3]に示したので参照されたい。

参考文献

[1]陣崎など: 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、情処研報89-OS44-9(1989-9)など

[2]松平など: 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　ー通信制御方式ー、本大会予稿(1992-10)

[3]新家など: 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　ー性能評価ー、本大会予稿(1992-10)

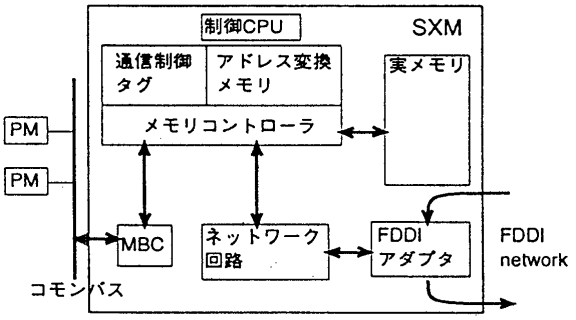

図4　SXMブロックダイアグラム

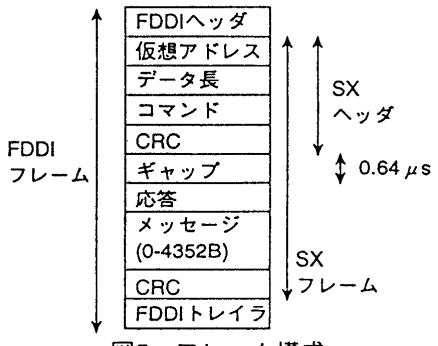

図5　フレーム構成