スーパーテクニカルサーバSR8000におけるネットワーク高速化方式

3ZB−05

清水 正明　鵜飼 敏之　佐川 暢俊
(株)日立製作所 中央研究所

1. はじめに

近年のスーパーコンピュータへの性能要求では、演算性能の向上ばかりではなく、演算データの入出力を行なう外部装置および外部計算機とのネットワーク性能が重要度を増している。特に外部汎用ネットワークであるEthernetの性能は重要とされている。

本稿では、マイクロカーネル構成のOSを採用するスーパーテクニカルサーバSR8000におけるEthernet性能高速化方式について報告する。

2. スーパーテクニカルサーバSR8000の概要

SR8000のハードウェアは、演算と入出力を行なう単位であるノードと、ノードを相互接続するハイパー・クロスバ・ネットワーク(HXB)で構成されている。各ノードは複数の共有メモリ型マルチプロセッサとローカルメモリ、ネットワークインタフェースアダプタおよび外部I/Oを制御するI/Oアダプタ(IOA)より成る。IOAにはSCSIアダプタ、100MbitおよびGigabit Ethernetアダプタ等の入出力アダプタを接続可能である。

オペレーティングシステムHI-UX/MPP for SR8000は基本スケジューリングと仮想記憶、デバイスドライバをMachマイクロカーネルが受け持ち、標準UNIXとしての機能はUNIXサーバが担当するマイクロカーネル構成である。

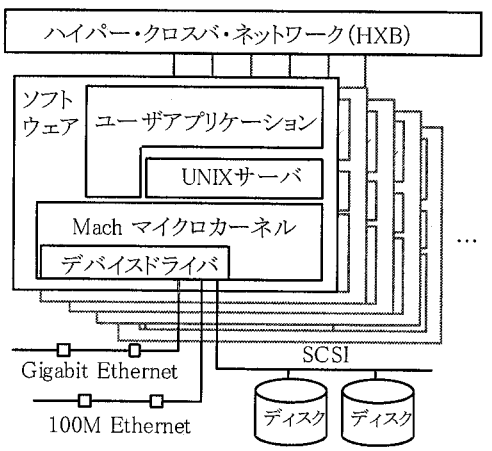

図1 SR8000のソフトウェア構成

High-speed Network I/O Method on SR8000
Masaaki Shimizu,Toshiyuki Ukai, Nobutoshi Sagawa
Central Research Laboratory HITACHI,Ltd.

3. SR8000のネットワーク処理

HI-UX/MPPではMachマイクロカーネルがEthernetアダプタのデバイスドライバを持ち、Ethernetパケット入出力のみの低レベル入出力を提供する。また、Machマイクロカーネル上のUNIXサーバがTCP/IP等のプロトコル処理を提供する。UNIXサーバはユーザからの入出力要求をUNIXのシステムコールで受理し、さらにデバイス入出力をMachのシステムコールとしてデバイスドライバに要求する。デバイスドライバは入出力要求を処理して結果をUNIXサーバに戻す。このとき、Ethernetデバイス入出力は非同期で行ない、受信パケットはMachマイクロカーネルからUNIXサーバへのメッセージとして渡される。

4. ネットワーク入出力高速化方式

一般にマイクロカーネル構成のOSにおいてはOSモジュール間の呼び出しがシステムコールまたはメッセージになり、デバイスドライバ等で呼び出し回数が多い場合には性能を出しにくい。そこで、マイクロカーネル構成のOSにおいてEthernet性能を高める次の高速化方式を開発した。

(1) システムコール一括化（システムコール回数削減）

HI-UX/MPPはマイクロカーネル構成を採用しているのでデバイスドライバを呼び出す毎にMachシステムコールが必要である。そこで、システムコール回数を削減するために送受信するパケットをチェインして一括処理要求するMachシステムコールを新設した。一括で処理する数は要求性能と一括処理で発生するレイテンシとのトレードオフから43個(1.5Kbyte×43=64KByte)とした。この方式でシステムコール回数を最大1/43に削減可能である。

(2) ハードウェア割り込みの抑制（割り込み回数の削減）

割り込みの回数を減らすことは割り込み処理回数、コンテキストスイッチ回数を減らすことになり、性能向上に有効である。パケットを連続で送受信している時にはパケット毎にアダプタからの送受信割り込みを受けることを止めて、一定時間毎に受信パケットおよび送信完了割り込みの処理を行なうことにした。一定時間毎にパケ

ットをチェックする方式の場合、時間間隔の初期に受信したパケットに対する処理遅延が発生するが、チェック間隔を1～2msecに設定することで性能とレイテンシのバランスを取った。パケットの到着が疎な時にはパケット毎に割り込みを受けるモードに切り替えるようにした。

(3) バッファの一括管理（コピー回数の削減）

MachカーネルとUNIXサーバの間で送受信パケットを受け渡す際に、送受信データのコピーが発生している。また、Mach カーネルの内部においてもハードウェア送受信バッファとカーネルのメッセージバッファの間でコピーを行なっている。コピーが発生するとコピー時間以外にも、受け側ページに対するページフォールトが発生するためメモリ管理処理の時間も必要になる。そこで、Machカーネルと UNIX サーバそして送受信ハードウェアの三者でメモリを共有する機構を用意し、OS 内部でのデータコピー回数を削減した。

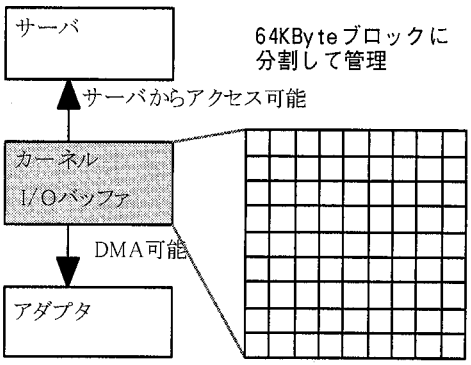

図2 共有ブロックの管理

(4) スレッド切り替えの削減

パケットを受信した場合、ファーストハンドラで受信パケットを一旦キューイングして受信スレッドを起動し、受信スレッドで受信パケットのフィルタ処理および UNIX サーバへのメッセージ転送を行なっている。しかし、複数のスレッドを使用することはスレッド切り替えを必要とし、スケジューリングおよびlock競合が発生する。そこで、割り込みを受けたカーネルスレッドの延長で受信処理まで行なう方式にすることで処理時間を削減した。この方式の場合割り込み禁止時間が延びる弊害があるが、実際にテストしたところ100Mbit Ethernetの限界速度で送受信しても一回の割り込み禁止時間が数百μsec であり許容できると判断した。

(5) メッセージ通信の排除

受信処理において受信パケットはMach カーネルからUNIX サーバへのメッセージとして渡される。しかし、メッセージ通信では、メッセージの生成、送信、受信、受信スレッドの起動等の処理が必要である。そこでUNIXサーバが受信パケットメッセージを待つ方式から、受信システムコールを発行したスレッドがブロックしてパケットを待つ方式に変更して、メッセージ送受信を排除した。

(6) パケットフィルタの軽量化

パケットのフィルタリングは、インタプリタ型のパケットフィルタを別スレッドで起動して使用している。このとき、パケットを一旦キューイングしてからパケットフィルタ処理を個々のパケットに対して行なっているので、使用しないパケットに対してもキューイング、スレッド起動等の処理を行なっている。そこで、パケットのキューイング前に静的なフィルタテーブルを使用してフィルタ処理を行なうことで処理の軽量化を行なった。

5. 性能評価

HI-UX/MPP for SR8000の100Mbit Ethernetドライバおよび Gigabit Ethernet ドライバに上記高速化方式を適用し、ドライバを呼び出すUNIXサーバのプロトコル処理部にも変更を加えたOSを試作した。

性能測定ベンチマークとして 1MByte のバッファを 100回TCP/IP インタフェースにて送受信するプログラムを使用した（ウィンドウサイズは128KByte）。

テストに使用した100Mbit Ethernetにおいてハードウェア限界の10MByte/secを達成し、Gigabit Ethernetにおいては50MByte/secを超える性能を確認した。

また、Gigabit Ethernet ドライバの処理性能は200MByte/sec 以上を達成しており、複数枚アダプタの処理にも対応できることを確認した。

6. おわりに

マイクロカーネル構成のOSを採用するSR8000において、Gigabit Ethernet ハードウェアの性能を引き出す高速化方式を開発した。パケット一括送受信、マイクロカーネル・UNIX サーバ・アダプタ間送受信領域共用によるコピーオーバーヘッド削減等により、Gigabit Ethernet TCP/IPにおいてマイクロカーネルOSとして世界最高水準の性能を達成した。