サイエンティフィック・システム研究会
科学技術計算分科会

2001 年 10 月 31 日

大規模 PC クラスタ構築経験

新情報処理開発機構
つくば研究センタ
石川 裕、住元 真司、原田 浩

1 はじめに

新情報処理開発機構では、1995 年以来、スパコン並の性能を引き出すクラスタのためのシステムソフトウエアである SCore の開発を行なっている。最近、クラスタ構築が花盛りになったが、大規模クラスタを構築した時に期待通りの性能が引き出せるのか、システムが安定して稼働するのか、など未知の世界であった。ユーザに対して、大規模クラスタ実現の先鞭をつけると共に今後の並列処理市場を牽引することを目的に、本年春、1,024 プロセッサを有するクラスタである SCore III を開発した。

本発表では、SCore ソフトウエアの紹介の後、SCore III の設計と実装の解説を通して、大規模 PC クラスタ構築時の注意点について述べる。

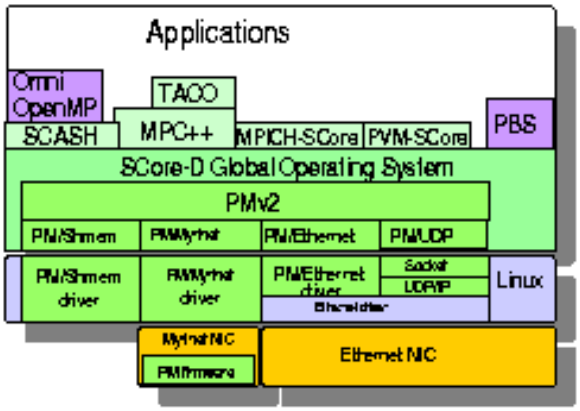


図 1: SCore ソフトウエアアーキテクチャ

2 SCore の概要

SCore ソフトウエアアーキテクチャを図 1 に示す。

2.1 PMv2 通信ライブラリ

PMv2 通信ライブラリは、クラスタ コンピューティング用低レベル通信ライブラリである。PMv2 API (Application Program Interface) は、クラスタにおける複数種類のネットワークや共有メモリに同一の方法でアクセスできるように設計されている。

- PM/Myrinet
  Myricom 社 Myrinet ネットワーク用の PM 通信レイヤである。PM/Myrinet は、Myrinet NIC 上のプログラム、Linux カーネルドライバ、ユーザレベルライブラリから構成されている。PM/Myrinet では、ユーザレベル通信およびゼロコピー通信と呼ばれる手法を用いて低遅延、高バンド幅ネットワーク通信を実現している。
- PM/Ethernet
  PM/Ethernet は、Ethernet 上における PM 通信プロトコルを実現している。Ethernet デバイスドライバの上に構築されたカーネル内プロトコル処理ルーチンとユーザレベルライブラリから構成される。PM/Ethernet は、TCP/IP と比べて軽量なプロトコル処理で済むように設計されている。

  Ethernet パケットタイプに PM/Ethernet タイプを追加し、PM/Ethernet 用のパケットと TCP/IP のような従来の通信パケットも共存できるようにしてある。すなわち、TCP/IP のアプリケーションと PM/Ethernet を使用したアプリケーションを同時に動かすことが可能である。

  PM/Ethernet は、複数の Ethernet リンクを束ねて通信バンド幅を向上させるネットワークトランキング機能を実現している。最近の PC サーバにはオンボードで 2 つの Ethernet の口を持っていることが多い。Ethernet スイッチを追加するだけで性能があがる。

- PM/Shmem
  PM/Shmem はオペレーティングシステムの共有メモリ機構を利用して実現されている。PM/Shmem により同一コンピュータ内で複数のプロセスが PM 通信 API で通信が可能となる。

2.2 MPICH-SCore

MPICH-SCore は、MPI 通信ライブラリを実装したフリーソフトウエアである MPICH を PM 通信ライブラリ上で稼動するようにしたソフトウエアである。MPICH-SCore では、単一プロセッサから構成されるクラスタだけでなく、共有メモリ型並列コンピュータから構成されるクラスタ上でも効率良く稼動するように工夫している。

2.3 PVM-SCore

米国 Tennessee 大学、Oak Ridge 国立研究所、Emory 大学が開発した PVM を SCore 用に移植している。これを PVM-SCore と呼んでいる。

PVM は異機種環境で並列環境を提供し、動的な計算ノードの追加が可能である。しかし、PVM-SCore では、SCore 環境下の計算ノード上でしか実行でない。また、計算ノードの最大数は、PVM 実行に必要なデーモンプロセス起動時に決まる。

2.4 SCore-D

SCore-D は、複数のユーザが同時にクラスタを利用するときに効率良くコンピュータ資源を管理する機能を提供するグローバルオペレーティングシステムである。SCore-D は、カーネルを変更することなくデーモンプロセス群で実現されている。ギャングスケジューリングと呼ばれるスケジューリング手法を用いて、並列アプリケーションを時分割スケジューリングしている。今まで、ギャングスケジューリングは特殊なハードウエアがないと効率良く実現できないと言われてきた。SCore-D により専用ハードウエアがなくても効率良くギャングスケジューリングが実現できることを初めて実証した。さらに SCore-D では次のような機能を提供している。

- 実時間ロードモニタ
- デッドロック検出
- チェックポイント・リスタート機能
- 対話型デバッガ起動

2.5 Omni OpenMP on SCASH

SCASH は、カーネルを変更することなく、PM 通信ライブラリを用いてユーザレベルで実現したソフトウエア分散共有メモリシステムです。分散共有メモリによりマルチスレッドプログラミングが可能となる。

Omni OpenMP は SCASH ソフトウエア分散共有メモリシステムを使い、OpenMP のプログラムを変更しなくてもクラスタ上で動く。OpenMP は、共有メモリシステム上でのマルチスレッドプログラミングを支援する仕様である。共有メモリを使ったマルチスレッドプログラミングでは、共有メモリ上でプロセッサによる複数の実行の流れを制御するプログラムを書く。OpenMP では、この制御をコンパイラに対する指示文と実行時ライブラリで行なう。

商用 OpenMP コンパイラで、クラスタ上で稼働する OpenMP として宣伝されている場合がある。この場合、計算ホスト内でのマルチスレッドプログラミングを OpenMP で記述し、計算ホスト間の通信を MPI などの通信ライブラリを使って記述する、と言ったハイブリッドなプログラミングをしなければクラスタ上では稼働しない。

2.6 MPC++

MPC++はオブジェクト指向言語 C++を基にした並列プログラミング言語である。MPC++には、言語機能を拡張せずに C++が持つ class、template 機能を用いて並列処理記述プリミティブを提供している。

2.7 PBS

バッチシステム環境を提供している PBS(Portable Batch System) を SCore 環境に移植している。

2.8 TACO

トポロジと通信を抽象化した C++テンプレート機能として TACO がある。

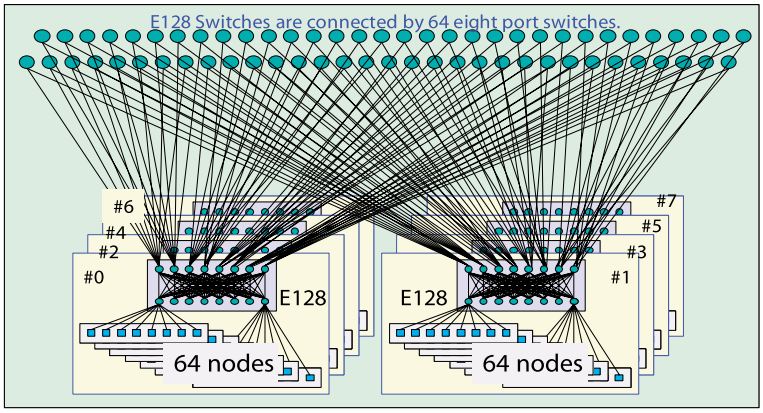


図 2: Myrinet ネットワークトポロジ

表 1: SCore III の仕様

	仕様	台数
計算ホスト	NEC Express Server 5800 120Ra-1 (Pentium III 933 MHz x 2 512 Mbytes 主記憶, 9.1 Gbytes SCSI ディスク x 2)	512
Server	NEC Express Server 5800 120Rc-2	4
Network	Myrinet-2000 100Mbps Ethernet	1 2

3 1,024 台 PC クラスタの設計

2000 年より、高さ 1 U （4.4cm）のラックマウント型 PC サーバが市販されるようになった。メーカにより多少の差はあるが、概ね、以下のような特徴を持つ。

1. 2 つの CPU が搭載された SMP 構成が可能
2. ボード上に 2 つの Ethernet が搭載されている
3. 2 つの PCI バススロットがある
4. 2 つのリムーバブル SCSI ディスクが搭載可能

このような PC サーバと Myricom 社 Myrinet-2000 を用いて、コンパクトなクラスタを構築できるようになった。新情報処理開発機構が製作した 1,024 台 CPU 構成の PC クラスタである SCore III のハードウェア仕様を表 1 に示す。

3.1 Myrinet ネットワークトポロジ

Myrinet ネットワークでは、構成要素として 16 ポートクロスバスイッチを結合してネットワークトポロジを組む。クロスバスイッチはパケットをスイッチ内部で蓄積することなく他のポートに直接送出する。Myrinet のスイッチではクロスバスイッチにワームホールルーティングと呼ばれる方式を採用することによりスイッチ処理を高速化している。

SCore III の Myrinet ネットワークトポロジは、512 台のコンピュータ接続においてバイセクションバンド幅が最大になるように設計した。バイセクションバンド幅とは、ネットワークに接続されているコンピュータを 2 分割した時に、2 分割間の通信バンド幅を意味する。

図 2 に Myrinet ネットワークトポロジを示す。図中、四角の線で囲われた部分が SCore III の 1 モジュールで、64 台のコンピュータと Myrinet Clos128 スイッチ（図中 E128）から構成される。E128 は、128 台のコンピュータを接続することが出来るが、このうちの半分を他のスイッチとの接続に使用している。図中、上部にある丸はスイッチを表現しているが、分かりやすくするために、Point to Point のつながりとして示した。一つの丸が一つのスイッチを表現しているわけではない。

このように、512 台のコンピュータにおけるバイセクションバンド幅は、64 x 4 x 2 x 2 Gps（full duplex すなわち送受信のバンド幅を考慮）、すなわち、1 Tera bps となる。これは、512 台のコンピュータを Myrinet を使って接続した場合のフルバイセクションバンド幅である。

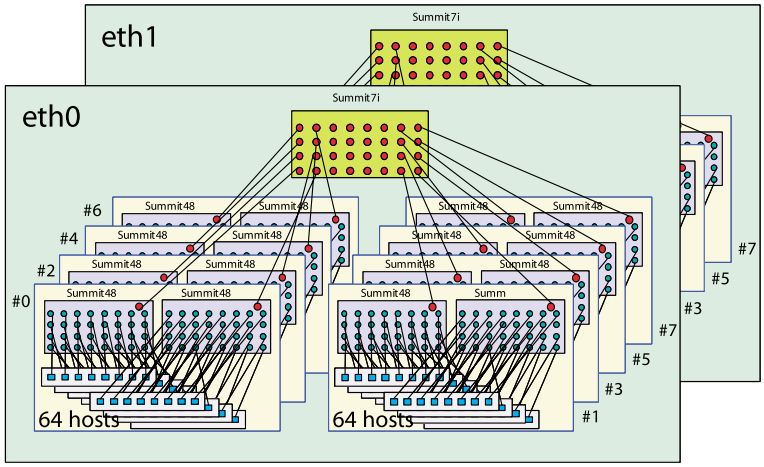


図 3: Ethernet ネットワークトポロジ

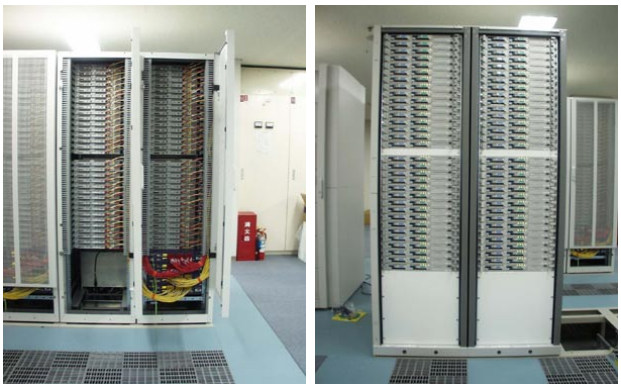


図 4: SCore III の 1 モジュール

図 5: サーバ・スイッチモジュール

3.2 Ethernet ネットワークトポロジ

PC オンボードに搭載されている 2 つの Ethernet NIC を用いて、2 系統の独立した Ethernet リンクを構築している。1 系統は IP アドレスを持ち TCP/IP による通信が可能である。もう一系統は PM/Ethernet が提供するネットワークトランキング機能を使ったときに使用される。

図3に示す通り、各リンクでは、32台の PCが一台の 100 Mbps Ethernet スイッチに接続され、16 台の 100 Mbps Ethernet スイッチは 1 台の 1 Gbps Ethernet スイッチに接続されている。すなわち、512 台のコンピュータにおけるバイセクションバンド幅は 16Gbps であり、100Mbps Ethernet を使ったときのフルバイセクションバンド幅 51.2Gbps の 1/3 でしかない。

Ethernet 系でフルバイセクションバンド幅を提供しなかったのは、予算の関係と Myrinet ネットワークを主に使うことを念頭においていたためである。

3.3 ラックの構成

PC サーバ、Myrinet スイッチおよび 2 系統の Ethernet スイッチを搭載するために 2 つの 19 インチラック（高さは 44 U）で 1 モジュールとなるような配置にした（図

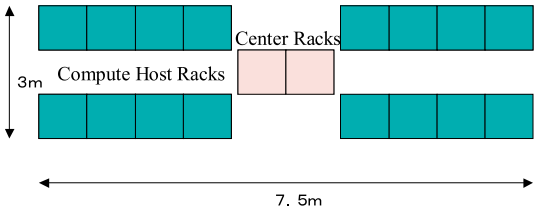


図 7: SCore III のレイアウト

4)。すなわち、32 台の PC サーバと 128 ポート Myrinet スイッチを、もう一つのラックには 32 台の PC サーバと 2 台の Ethernet スイッチが搭載されている。図 4 の左の写真は、組み立て中のモジュールで、左側のラックにはまだ Myrinet スイッチが設置されていない。右側ラックには Ethernet スイッチが 2 つ設置されている。512 台の PC サーバを格納するために 8 モジュール構成となっている。

各モジュールの Myrinet および Ethernet スイッチを接続するためのスイッチとクラスタ用サーバは、2 本のラックに収められている（図 5)。図 5 において、右側のラックには Myrinet スイッチだけが設置されている。左側のラックには、サーバおよび Gigabit Ethernet スイッチ、Myrinet スイッチが設置されている。

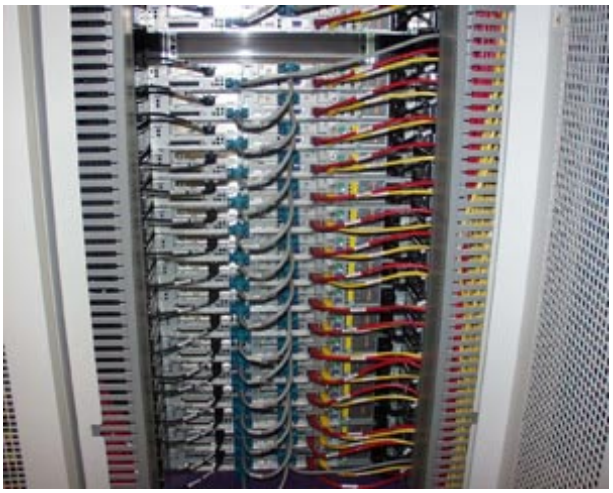
図 7 に示す通り、SCore III は合計 18 本の 19 インチラックから構成され、設置面積は 3 メートル x 7.5 メートルとなっている。図 6 は、正面左側から撮影した SCore III である。

3.4 配線

ラック内の配線の様子を図 8 に、ラック間の Myrinet 配線の様子を図 9 に示す。ラック内のケーブル類は 19 インチラックの両脇に収めるようにした。これは、ケーブルによる空調の流れの遮断をなくすようにするとともに、コンピュータ本体を取出すときに、ケーブルの着脱を容易にすることが目的である。

3.5 コンソールモニタ

クラスタを構築するとき、PC のコンソールをどうするか悩ましい問題である。大抵は、一つのディスプレイ・キーボード・マウスだけで操作できるようディスプレイ・キーボード・マウス スイッチを使うだろう。SCore III を構築する前に 128CPU 構成のクラスタである SCore II を



左側に延びているケーブルは、Myrinet-2000 の serial line、真中の ケーブルは PC サーバ間のシリアルリンク、右側のケーブルが 2 本の Ethernet リンクと電源ケーブル。

図 8: ラック裏側の配線の様子

図 9: Myrinet の配線

図 6: SCore III の外観

製作したとき、一つのラックに 16 台の PC サーバとディスプレイ・キーボード・マウス スイッチを載せた。しかし、太いケーブルがラックの裏でとぐろを巻き、保守性を著しく低下させたので、取り外してしまった。

このような経験から、SCore III では、図 8 の写真に示す通り、PC サーバ間をシリアルラインで数珠繋ぎしている。これは、ある PC サーバがネットワーク経由でログイン出来ない状況になった時に、隣の PC サーバにログインして、シリアルライン経由でコンピュータの状況を把握することを仮定している。

初期不良洗いだし時には、反応しなくなったコンピュータが発生したら、ディスプレイ・キーボード・マウスを載せた台を持っていき、直接つなげて調べていた。

一度、安定稼働すれば、コンソールモニタは必要ないので、全ての PC サーバに接続するためにディスプレイ・キーボード・マウス スイッチを用意するのは得策ではないだろう。

4 1,024 台 PC クラスタの実装

4.1 ハードウェアの設置

次のような手順で進められた。

- 筐体を設置するための床補強
- 512 台の PC サーバが収められる 16 ラック内の Ethernet ケーブル配線
  ケーブル長を合わせて配線を行う。このために 4 人の専門家が 2 日間要しました。
- 4 ラック毎にサーバの設置および配線
  このために 5 人が 4 日間かけて行いました。

4.2 ソフトウェアのインストール

SCore は EIT と呼ばれるネットワーク経由でインストールするツールを提供しています。新情報処理開発機構では、64 台規模までのクラスタ設置において、EIT の使用実績があります。大規模クラスタにおいては、ネットワークが高負荷状態になり、インストールに支障を来たすと判断し、以下に述べるディスクブート方式を採用しました。

以下の内容からなるディスクイメージを作成しました。

- ブートイメージ
- インストールに必要な binary RPM ファイル
- 改良版 anaconda インストレーションツール
- インストレーションスクリプトファイル群

このイメージを各 PC サーバのディスクにコピーします。新情報処理開発機構が 2000 年に製作した 64 台 PC

サーバから構成される SCore II クラスタは、SCore III で使用した PC サーバと同じシリーズの PC サーバを使用しています。PC サーバのディスクはリムーバブルディスクなので、SCore III のディスクを取出し、SCore II クラスタに装着し、コピーしました。すなわち、SCore II クラスタをコピーマシンとして使用しました。

コピー時間は次の通りでした。SCore II は研究開発に使われていたので、半分の 32 台をコピーのために使用しました。

- 32 台のホストにディスクイメージをコピーするのに 2 分間
  Myrinet ネットワーク経由でイメージをコピーしました。

- 128 台のディスクにコピーするのに 6 分間
  一台の PC サーバに 4 台のディスクを装着できます。

コピーしたディスクを実機に装着して電源を入れると以下の手順でインストールが行われます。

- 第一フェーズ
  ディスクの最初のパーティション以外をパーティショニングしてフォーマットします。ディスクイメージを作成するときに、ディスクの geometry 情報が得られないため、最初のフェーズで、パーティショニングします。

- 第二フェーズ

  1. ファイルシステムを作成します。
  2. サーバから IP アドレスを取得します。サーバ上ではあらかじめ DHCP デーモンを立ち上げておきます。DHCP デーモンのコンフィギュレーションファイルに MAC アドレスと IP アドレスの対を定義して、IP アドレスを固定しておきます。
  3. anaconda を起動し、ローカルディスク上の RPM ファイルを使ってソフトウェアのインストールを行います。

4.3 テスト

以下の手順でテストを行った。

- 接続テスト
  ラック内コンピュータ、モジュール内コンピュータ、全てのコンピュータ、の順に接続テストを行った。

- SCore が提供する rcstest コマンドによる全体全通信テスト
  rcstest はランダムに通信を行うテストプログラムである。一昼夜動かすことによってネットワークの初期不良を検出することが可能である。

  一般に、Ethernet の場合、ifconfig 等のコマンドを使って、エラーやコリジョンの発生数を確認する。もしも、エラーの数が大きくなっていた場合には、ケーブルの接続不良や NIC の装着不良あるいは故障、スイッチの故障を疑う。今回、このようなことは生じなかった。

  Myrinet の場合、/proc/pm/myrinet/info/0 の内容を見て、CRC エラーが生じているかどうか確認する。一時間に何十回と CRC エラーが生じている場合には、ケーブルの接続不良、NIC の装着不良あるいは故障、スイッチの故障等を疑う。

- キラーアプリケーションによるストレステスト
  rcstest でも見つからない不良がある。Myrinet ネットワークの場合には、大量のパケットを送受信することにより NIC 上の LANai プロセッサおよびメモリにストレスをかける必要がある。また、PC 側のメモリやバスにもストレスをかける必要がある。このようなストレスをかけるためのアプリケーションを実行する必要がある。新情報処理開発機構では、そのようなアプリケーションを持っていないので、NAS 並列ベンチマークや Linpack のようなベンチマークプログラムを走らせることにより、初期不良ハードウェアの検出を行った。

5 初期不良の要因

SCore III 構築時に生じた初期不良と、その対処について述べる。

- 現場で追加したメモリカードや Myrinet NIC の装着不良
  工場出荷前にエージングを行っている製品でも、現場でメモリカードや Myrinet NIC などのハードウェアを追加すると、装着不良による問題が発生する。初

期不良を減らすためには、カスタマイズされた状態で工場でエージングが行われる必要がある。

- ケーブル装着不良
  今回使用した Myrinet は Serial Cable である。Serial Cable の方が Fibre Cable よりも理論的エラー発生率が低いと言われたが、コネクタ部分の装着不良によるエラーが多発した。Myricom 社は、現在、Fibre Cable を推奨している。

- ハードウェアの初期不良
  エージングしていても 512 台になると何台かは故障して動かない場合がある。現調時、予備パーツとして数台用意しておくと、システムの調整がスムーズに行くだろう。

- Myrinet スイッチ
  Myrinet スイッチは、Line カード、Spine カードという 2 種類のカードを組み合わせて、バックプレーンがついた Enclosure に装着する。この装着不良による故障も生じた。なお、Myrinet スイッチには、SNMP プロトコルおよび HTTP プロトコルでスイッチの状態を監視できる。

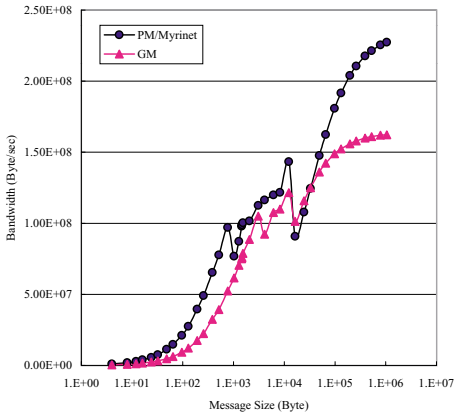


図 10: Myrinet における MPI の通信バンド幅による比較

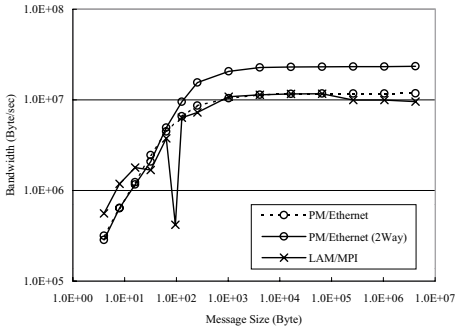


図 11: Ethernet における MPI の通信バンド幅による比較

6 1,024 台 PC クラスタの評価

MPICH-SCore の性能を他の通信ライブラリと比較します。比較対象として、Myrinet の場合には、Myricom 社が提供している MPICH-1.2.1 を基にした MPI-GM を、Ethernet の場合には米国ノートルダム大学が開発した LAM/MPI 6.5.1 を、それぞれ使用しました。LAM/MPI は、TCP/IP プロトコルを使用しています。

6.1 基本通信性能

図 10 および図 11 は、Myrinet および Ethernet における MPI レベルでの通信バンド幅を他の実装と比較したグラフです。図 10 において、PM/Myrinet が MPICH-SCore で、GM が MPI-GM の評価結果です。Myrinet においては、MPICH-SCore が最大 227 MByte/sec 出ているのに対して、MPI-GM では 162 MByte/sec しか出ていません。

図 11 において、PM/Ethernet が 100 Mbps Ethernet を 1 リンク使用した時の MICH-SCore の性能で、PM/Ethernet (2Way) が 100 Mbps Ethernet を 2 リンク使用した時の MPICH-SCore の性能です。TCP/IP(LAM) は、LAM/MPI の性能です。MPICH-SCore も LAM/MPI も最大 12MByte/sec とほぼ同じ性能が出ています。また、2 リンク使用時の MPICH-SCore の性能は、最大 24MByte/sec と 2 倍の性能がでています。

4 バイトメッセージにおける MPI の通信遅延を表 2 に掲載します。Myrinet での通信遅延は約 12 マイクロ秒であるのに対して、100 Mbps Ethernet では約 56 マイクロ秒で、4 倍以上の差があります。

6.2 アプリケーション起動時間

表 3 にジョブの起動時間を示す。これは、シングルユーザモードで、ユーザが scout 環境でアプリケーションを実行したときの起動時間である。マルチユーザモードでは、512 ホストにおけるジョブ起動時間は 4 秒程度となる。シングルユーザモードでは、アプリケーション起動時に各プロセッサが scoreboard データベースサーバとの通信が生

表 2: MPI の通信遅延による比較

RTT/2: 1/2 往復時間（マイクロ秒）

低レベルライブラリ	RTT/2
PM/Myrinet	12.3
GM	12.8
PM/Ethernet	55.6
PM/Ethernet (2Way)	55.6
TCP/IP (LAM)	77.5

表 3: アプリケーション起動時間

単位 秒

ノード数	Myrinet	Ethernet x 1	Ethernet x 2
16	2.42	2.38	5.48
32	2.71	2.72	5.72
64	3.47	3.41	6.28
128	4.39	4.50	7.02
256	7.31	7.60	9.40
512	13.31	14.16	14.17

じるため。このために、プロセッサ数が増えると起動時間が長くなる。マルチユーザモードの場合には、この処理がないために起動時間が早くなる。

6.3 姫野ベンチマークの結果

図 12 に姫野ベンチマークの結果を示す。Large (512x256x256) サイズでの結果である。コンパイラは PGI 社のコンパイラで最適化オプションは -O4 とした。なお、256 台までしか計測していないのは、Large サイズの大きさでは、256 台以上では正しく実行されないためである。

6.4 Linpack ベンチマークの結果

世界中のスーパコンピュータを Linpack と呼ばれるベンチマークプログラムを使った性能値でランク付けしている TOP500 と呼ばれるサイトがある (http://www.top500.org/)。2001 年 6 月の TOP500 では、SCore III は 547.90Gflops の性能で 36 位だった。この時は、一部ハードウェアが故障しており、全てのプロセッサが利用できなかった。2001 年 8 月には、1,012 台のプロセッサを使って、618.3 Gflops の性能を達成した。この時点で、2001 年 6 月の TOP500 リスト中のクラスタの中では一番の性能だった。

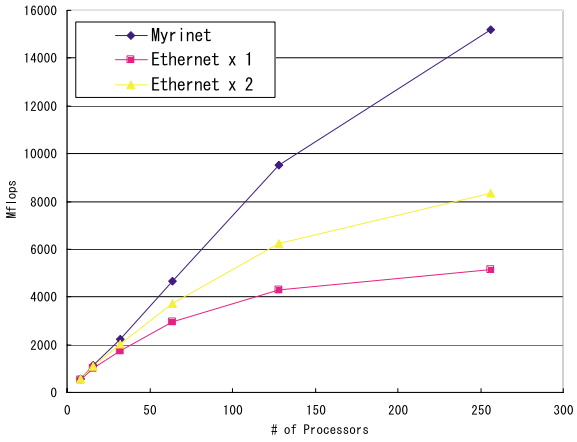


図 12: 姫野ベンチマークの結果

7 終わりに

コモデティハードウェアの組合わせによる 1,204 台プロセッサ構成の PC クラスタを安定して実現できることを実証した。システムソフトウエアにおいて、スケーラビリティの問題もないことを確認した。しかし、アプリケーションレベルでのスケーラビリティ検証については、今後の課題である。

ここで紹介しなかったが、NAS 並列ベンチマーク集のクラス C における性能評価も行っている。プロセッサ台数に制限のあるベンチマークがあり、1,024 台を使って NAS 並列ベンチマーク集全ての結果を得ていない。また、1,024 台規模のベンチマークでは、クラス C の問題サイズ以上が必要であろう。今後、小規模から大規模クラスタ用のベンチマークプログラムが必要と考えている。

PC クラスタは、誰でもが秋葉原で PC のパーツを買ってくれば容易に並列環境が構築できると考えがちである。しかし、多くの落とし穴が潜んでいる。様々なチップセット、プロセッサのリビジョン、Ethernet カード、ネットワークスイッチがあり、場合によっては本来出るべき並列処理性能が出ない場合がある。大規模クラスタの場合は言うまでもないが、小規模クラスタ構築においても、ハードウェアとネットワーク知識のないユーザに対して、技術的

表 4: 2001 年 6 月 TOP500 リスト中のクラスタ

順位	コンピュータ	性能 (GFlops)	プロセッサ数	開発元	国
	SCore III (P-III 933)	618.3	1,024	新情報処理開発機構	日本
30	Netfinity Cluster (P-III 1G)	594.00	1,024	IBM	アメリカ
31	Netfinity Cluster (P-III 1G)	594.00	1,024	IBM	アメリカ
42	CPlant (A-21264 466)	512.40	1,000	Sandia 国立研究所	アメリカ
43	AlphaServer SC ES40/EV67	507.6	512	Compaq	アメリカ
44	AlphaServer SC ES40/EV67	507.6	512	Compaq	アメリカ
57	HPC 4500 400MHz	420.44	896	Sun	アメリカ

備考：コンピュータ欄の括弧内はプロセッサタイプとクロック（G がついているのは GHz でそれ以外は MHz)。

P-3 は Intel 社 Pentium III。A-21264 は Compaq 社 Alpha 21264。Ath は AMD 社 Athlon。

にしっかりとサポートできる企業の育成が急務と考える。

新情報処理開発機構での SCore の開発は 2001 年 11 月末日で終了する。SCore システムソフトウエア開発の継続と、PC クラスタ市場の健全な発展を目的に、2001 年 10 月 4 日、21 社の賛同を得て、PC クラスタコンソーシアム (www.pccluster.org) を立ち上げた。また、SCore III はプロジェクト終了後も PC クラスタコンソーシアムを中心に主にアプリケーション性能の検証に利用される予定である。