

# SCore 型クラスタ

新情報処理開発機構

石川 裕

ishikawa@rwcp.or.jp

---

## はじめに

---

技術研究組合 新情報処理開発機構 は、国内外のコンピュータ関連企業、研究所から構成される研究組合で、通商産業省が 1992 年から 10 年計画で実施しているリアルワールドコンピューティング (RWC) プロジェクトを受託している。本稿では、当研究組合が開発してきた RWC PC クラスタハードウェアならびに SCore クラスタシステムソフトウェア Version 3 の特長を述べる。さらに、RWPCPC クラスタ上で稼動している実用アプリケーションを紹介する。

---

## 背景

---

ネットワークに繋がったパソコンを束ねて並列コンピュータ環境を実現する、いわゆるクラスタ、が普及し出している。特に既存フリーソフトウェアの組み合わせで製作された Beowulf 型クラスタが注目を集めているが、次の 2 点に問題があり、その利用範囲は限られたものである。低い通信性能 100 メガビット秒程度のネットワークを使い、インターネットで使用されている TCP/IP ネットワークプロトコルを使用しているために専用並列コンピュータに比べて通信性能で 1/10 以下の性能しかでない。非効率なコンピュータ管理各パソコン上で独立したオペレーティングシステムが動いていて、全てのパソコンを取りまとめる機能がないために、並列プログラムを効率良く実行することがない。

RWC プロジェクトでは、1994 年よりクラスタ構築を検討し、1995 年以降、Sun ワークステーション、Intel Pentium、Compaq Alpha を使用したクラスタを制作している。これらクラスタ上で、クラスタに必要なシステムソフトウェアを開発してきた。本システムソフトウェアを SCore クラスタシステムソフトウェアと呼び、SCore クラスタシステムソフトウェアを搭載しているクラスタを SCore 型クラスタと呼んでいる。SCore により、Beowulf 型クラスタの欠点を取り除き、専用並列コンピュータと同等の性能を有するクラスタが実現できた。

---

## RWC クラスタ開発の歴史

---

1994 年、ワークステーションと高速ネットワークを使用したクラスタ開発を検討していた。その当時話題になっていた ATM/LAN を使用したクラスタ構築事例がいくつかあったが、我々は、ATM/LAN

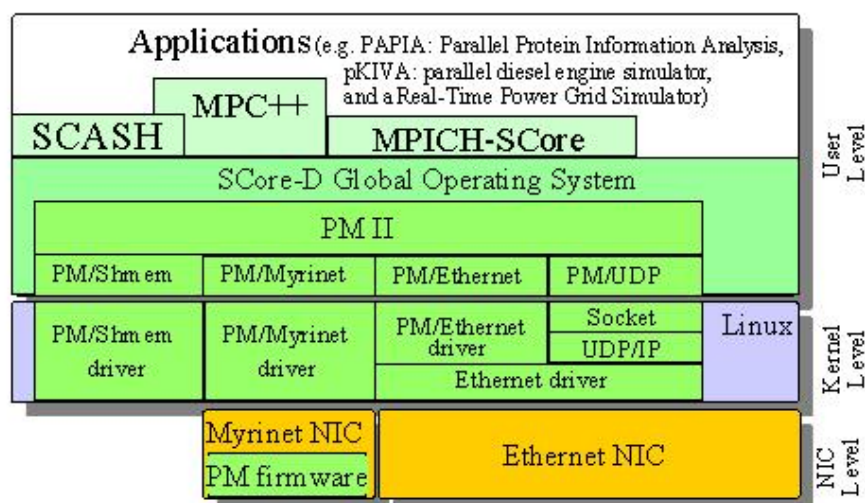


図 10.1: SCore クラスタシステムソフトウェア

以外のネットワークを探していた。なぜならば、ATM/LAN はパケット長が小さいだけでなく、物理層におけるパケットロストを許す仕様であったために、並列アプリケーションが要求する通信性能が提供出来ないと判断したからである。

1995 年春、Myrinet ネットワークを知った。Myrinet は、物理層でギガビット/秒の転送スピードを有し、ハードウェアレベルでパケット到着を保証する仕様である。また、Myrinet NIC に搭載されているプロセッサのプログラムを開発することにより、独自のネットワークプロトコルを実現することができるのも魅力的であった。

1995 年夏に Myrinet と Sun Sparc Workstation 20 を使用した小規模クラスタを構築し、idxPM 通信ライブラリ、ユーザレベルスレッドライブラリおよび MPC++、SCore-D を開発し、SCore クラスタシステムソフトウェアの原型を作った。1996 年初頭には、36 台の Sun Sparc Workstation 20 を使用したクラスタを構築した。この時、ベンチマークプログラムで Cray 社の Cray T3D に匹敵する性能が出ていることを確認している。

1996 年に 32 台の Intel 社製 Pentium プロセッサを使った RWC PC クラスタ 1 号機を開発し、1997、1998 年にかけて RWC PC クラスタ 2 号機を開発した。これら開発を通して、市販品のパソコンを用いても省スペースを実現出来ることを実証した。これら PC クラスタでは PICMG という工業規格ボードを用いている。

RWC PC クラスタ 2 号機では、Intel 社製 Pentium Pro 200MHz、256MBytes メモリ、4GBytes ディスクから構成されるコンピュータを使用し、全体で 128 台構成となっている。通信ハードウェアとして 100Mbps イーサネット以外に 1 ギガビット秒の通信性能をもつ Myricom 社 Myrinet を採用している。

1999 年には、Intel 社製 Pentium III を 2 台搭載した PC16 台および Compaq 社製 Alpha 21264 プロセッサを搭載した XP-1000 コンピュータ 16 台を、Myrinet、ギガビットイーサネットで接続したクラスタを制作した。

現在、Dual Pentium III 8000 MHz を使用した 128 プロセッサ数規模のクラスタである RWC PCCluster IV を構築中である。ネットワーク性能の違いによる並列アプリケーションの性能の違いを検証する目的で、Myrinet、Gigabit Ethernet および 3 系統の 100Mbps Ethernet が接続されている。

## SCore Cluster System Software

SCore は、ワークステーションや PC 等で稼働しているオペレーティングシステムである Linux 上に構築したトータルシステムソフトウェアである。最新版は、TurboLinux、Redhat、SuSE 上で稼働する SCore Version 3.1 である。現在、Version 3.2 の配布準備を進めており、9 月下旬には Version 3.2 を正式リリースする予定である。

図 10.1 に示すとおり、SCore クラスタシステムソフトウェア Version 3 は、

1. 高性能通信機能を実現した PM 2 通信ライブラリ<sup>3)</sup>、
2. PM 2 上に構築した MPI 通信ライブラリである MPICH-SCore<sup>4)</sup>、
3. コンピュータ資源を管理する SCore-D グローバルオペレーティングシステム<sup>2)</sup>、
4. 分散共有メモリシステム SCASH<sup>1)</sup>、
5. 並列プログラミング言語 MPC++<sup>1)</sup>、
6. Omni OpenMP コンパイラ<sup>6)</sup>、

から構成され、さらに、以下のツールを提供している。

- EIT イージーインストールツール
- SCOOP クラスタ監視ツール

以下、SCore の特長を紹介する。

### 高性能通信機構

通信ハードウェアの性能を引き出すために PM 2 通信ライブラリを開発した。PM2 は複数のネットワークハードウェアを支援できるよう、PM2 API レイヤ、PM2 デバイスレイヤの 2 つのレイヤから構成されている。現在までに、Myrinet および Ethernet 用ドライバ、共有メモリ型コンピュータ用に OS の共有メモリ機構を利用したドライバ、UDP プロトコルを利用したドライバ、が開発されている。これにより、次の 2 つの利用方法が可能となっている<sup>5)</sup>。

- 異なるネットワークが混在したクラスタを構築できる。たとえば、Myrinet を使った 2 つのクラスタをギガビットイーサネットでつなげたクラスタオブクラスタを構築することができる。
- PM2 上に構築された MPI 通信ライブラリ等のアプリケーションは、再コンパイルすることなく、シームレスにこれらネットワークハードウェアを利用できる。たとえば、100Mbps イーサネットを利用した小規模クラスタ上で稼働しているアプリケーションを、そのまま Myrinet を利用した大規模クラスタ上でも稼働する。

Myrinet 用 PM 2 では、Myrinet ネットワークインターフェイスカードのファームウェアを変更して、ユーザレベル通信およびゼロコピー通信機能を実現することにより、Pentium III(500MHz) を使った PC でラウンドトリップ 16 マイクロ秒(4 バイトメッセージ)、116Mbytes/sec(8 キロバイトメッセージ)のバンド幅を達成している。図 10.2 に MPI 通信ライブラリにおける 1 対 1 通信バンド幅の比較を示す。

SCore Version 3.2 からは、複数のネットワークを一つに束ねて通信する、ネットワークランキング機能、が実現されている。図 10.3 に、100Mbps Ethernet カード 5 枚挿しまでの性能を示す。440BX および 450MHz Pentium III と Tulip NIC の組合せでは、NIC を 3 台搭載するとほぼ 3 倍の性能が出ていることが分かる。

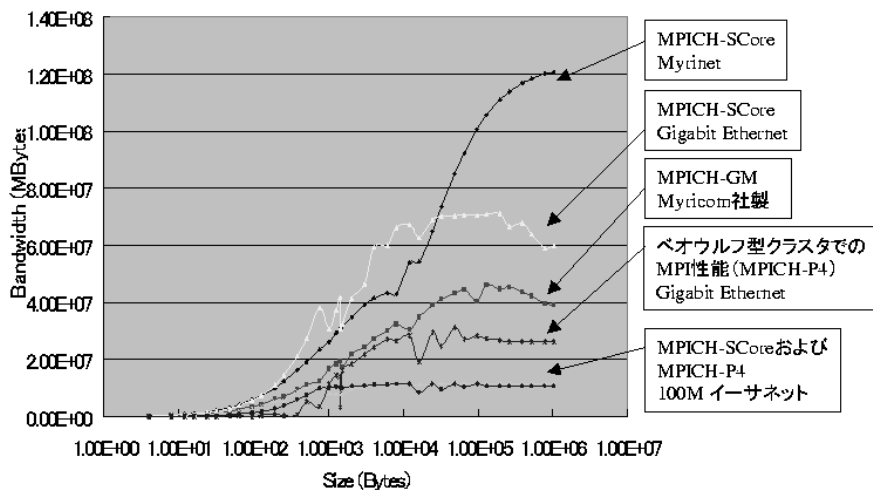
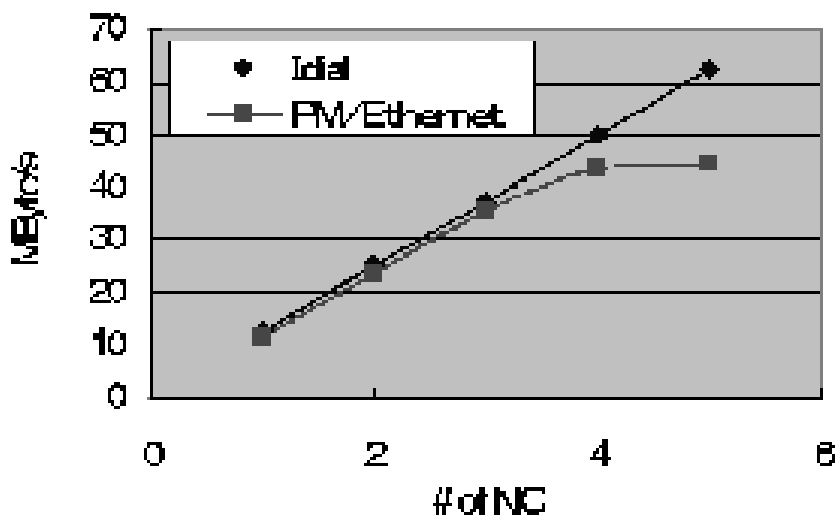


図 10.2: MPI 通信ライブラリにおける 1 対 1 通信バンド幅の比較



Pentium III 450 MHz 440BX , 5 つの Tulip NIC による一対一通信性能

図 10.3: PM/Ethernet ネットワークランキング性能

## 単一システムイメージ

商用並列コンピュータが提供しているシングルユーザモード，時分割・空間分割スケジューリングモード，バッチモードを提供している．アドバンスドランタイムその他，Score では，デッドロック検出機構，プログラムが異常終了した時に自動的にデバッガを立ち上げる機能，チェックポイント・リスタート機能を提供している．

SCore Version 3.2 から PBS バッチジョブシステムもサポートしている．

## 豊富なプログラミング支援環境

SCore では，並列アプリケーション記述で広く利用されている MPI 通信ライブラリだけでなく，マルチスレッド，リモートオブジェクト呼出し機能，グローバル演算を C++ で実現した MPC++ ，分散共有メモリによるプログラミングを支援する SCASH が提供されてきている．

さらに，SCore 3.2 では，クラスタ用 OpenMP として Omni/SCASH が提供されている．OpenMP は，SMP 用システムとして規格化され，多くのベンダがシステムを提供している．計算ノードが SMP 構成の SMP クラスタでは，ノード内の並列性記述には OpenMP を，ノード間の並列記述には，MPI 通信ライブラリを使用するというのが，一般的であった．

Omni/SCASH OpenMP システムにより，OpenMP で記述されたプログラムは，SMP の上でもクラスタの上でも稼働することが出来るようになった．

## 利用可能なコンパイラ

現在使用可能なコンパイラは次の通りである．

- Gnu コンパイラ
- Fujitsu コンパイラ
- Absoft コンパイラ
- KAI コンパイラ
- PGI コンパイラ

## アプリケーションの紹介

---

### pKIVA

pKIVA は，米ロスアラモス国立研究所で開発されたディーゼルエンジンシミュレーションプログラムである KIVA を，RWC プロジェクトで並列化したソフトウェアである．自動車工業界で実際に使われている KIVA の並列版である pKIVA により，今まで PC 一台で十数時間かかっていた処理が数十分で処理できるようになる．

### AMBER

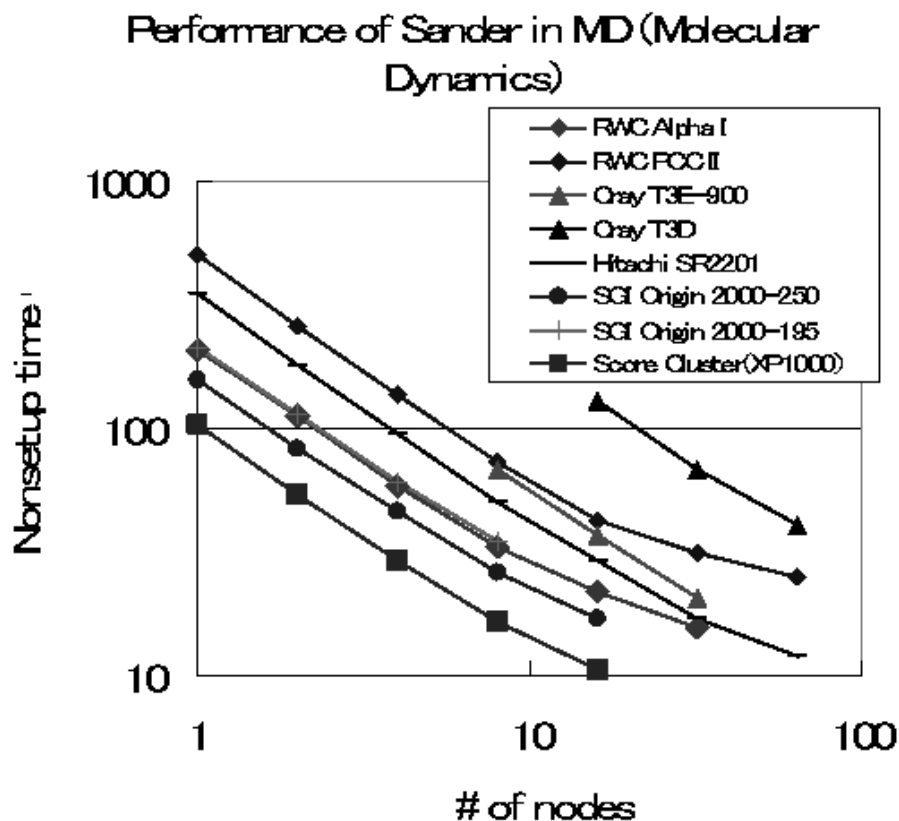


図 10.4: Amber の性能

分子動力学シミュレータで世界的に有名な AMBER というソフトウェアもクラスタ上で稼働している。図 10.4 に示す通り、我々のクラスタでは、今までのスーパーコンピュータである Cray T3E や日立 SR2201 以上の性能が出ている。

## PAPIA

PC クラスタの持つローカルディスクはデータベース検索エンジンにも最適である。その応用例として PAPIA システムを紹介する。PAPIA システムは、遺伝子の配列やタンパク質の立体構造からその意味を探りだすために、データベース検索を必要とするシステムである。PAPIA システムは 1998 年初頭より、世界中の研究者が利用できるように公開している (<http://www.rwcp.or.jp/papia/>)。

## ユーザ

1998 年秋より 300 サイト以上がインターネット経由でソフトウェアを入手している。国内では、三菱産業システム研究所が実時間電力シミュレータのシステムとして SCore が利用されている。米国では、ロスアラモス国立研究所で使用されている。ヨーロッパでもイギリス Oxford 大学計算センタ、ドイツ Bonn 大学応用数学科、フランスパリ南大学等で使われている。たとえば、ボン大学では、Parnass2 と呼ばれるクラスタを製作した。Parnas2 は、Myrinet ネットワークでつながれた Intel 社製 PentiumII 400 MHz をノードコンピュータに持つ 144 台構成のクラスタである。

---

## 終わりに

---

新情報処理開発機構では、クラスタおよび並列分散処理市場の拡大を支援すべく、昨年 10 月より、開発したソフトウェアをオープンソースとして提供している。これにより、現在、SuSE, Turbo-Labs/TurboLinux, Kondra は、SCore を使用したクラスタパッケージを開発中である。ソフトウェア配布に関しては <http://www.rwcp.or.jp/lab/pdslab/> を参照してほしい。

本年 10 月 3 日には、第一回 SCore Users Meeting がイギリス Oxford 大学で開催される。また、クラスタの普及を支援すべく、大学、研究機関、企業の研究所に新情報処理開発機構のクラスタを利用してもらおう機会を設けている。ご興味のある方は、[score-info@rwcp.or.jp](mailto:score-info@rwcp.or.jp) 宛てにメールしていただきたい。

Linux Distributor による SCore のサポート、ユーザーズミーティングによる know how の蓄積および共有を通して、クラスタ市場ならびに SCore 型クラスタが発展していくことを期待している。

---

## 参考文献

---

- 1) Y. Ishikawa, H. Tezuka, A. Hori, S. Sumimoto, T. Takahashi, F. O'Carroll, and H. Harada, "RWC PC Cluster II and SCore Cluster System Software - High Performance Linux Cluster," in Proc. of the 5th Annual Linux Expo, pp. 55-62, 1999.
- 2) A. Hori, H. Tezuka, and Y. Ishikawa, "Highly Efficient Gang Scheduling Implementation," in SC'98, 1998.
- 3) H. Tezuka, F. O'Carroll, A. Hori, and Y. Ishikawa, "Pin-down Cache: A Virtual Memory Management Technique for Zero-copy Communication," in IPPS/SPDP'98, pp. 308-314, IEEE, 1998.
- 4) T. Takahashi, F. O'Carroll, H. Tezuka, A. Hori, S. Sumimoto, H. Harada, Y. Ishikawa, and P. H. Beckman, "Implementation and Evaluation of MPI on an SMP Cluster," in Parallel and Distributed Processing - IPPS/SPDP'99 Workshops, Vol. 1586 of Lecture Notes in Computer Science, pp. 1178-1192, Springer-Verlag, 1999.
- 5) Toshiyuki Takahashi, Shinji Sumimoto, Atsushi Hori, Hiroshi Harada, Yutaka Ishikawa, "PM2: A High Performance Communication Middleware for Heterogeneous Network Environments," To appear in SC2000, 2000.
- 6) OpenMP, <http://www.rwcp.or.jp/lab/pdperf/> .