RWCPにおけるクラスタ開発記

石川　裕・堀　敦史・手塚　宏史
[新情報処理開発機構]

概　要

1995年より新情報処理開発機構で開発されてきたクラスタシステム上のシステムソフトウェアであるPM通信ライブラリやオペレーティングシステムSCore-Dについては情報処理8月号[1]で解説した．今回は，クラスタシステムの開発経緯を通して，クラスタシステムの開発状況を紹介する．また，現在のクラスタシステムにおいて，どの程度の性能が出るのか，新情報で開発したPCクラスタおよびAlphaクラスタを例に示す．

Myricom社を見つける

我々は並列計算機上の基本システムソフトウェアとして，並列プログラミング言語MPC++および並列システム上でのマルチユーザ環境を実現するOSとしてSCoreシステムの研究開発を行ってきている．1994年頃より，並列システムのための基本システムソフトウェア研究のインフラとして，我々はワークステーションに接続可能な高速ネットワークを利用した分散メモリ型並列計算機の構築を考えていた．1990年初頭にはThinking Machines社CM-5やIntel社Paragonといった超並列計算機の開発が盛んであった．比較的予算が潤沢な研究所は超並列計算機を購入していた．一方で，10Mbps Ethernetと複数のワークステーションによるクラスタシステムの構築があったが，超並列計算機には比べ物にならないほど遅いものであった．

新情報処理開発機構でもCM-5とParagonが導入されていたが，システムソフトウェア，特にOS開発のためにこれらを利用することはできなかった．ネットワークの物理層において通信パケットの到着保証を行うネットワークを用いることができれば，並列計算機と遜色のない並列環境が実現でき，OSの開発が可能になると考えていた．

当時，ATM-LANが注目され導入事例が増えていた．「10Mbps Ethernetの次のLAN」，「WANとLANとのシームレスな通信が提供される」などの宣伝文句のもと注目されていた．しかし，並列計算機を構成するネットワークにATM-LANそのものを用いることはためらっていた．なぜならば，i）ATM-LANでは物理層においてパケット廃棄を許している，ii）物理層で定義されているATMセルパケットの大きさが53Bytesと小さい，の2点から上位プロトコル層において信頼性のある通信保証はコストが高くつき，また，低通信遅延，高バンド幅通信が実現できないと考えていたからである．我々はあくまでも超並列計算機の代替としてのクラスタシステム構築を考えていた．

ATM-LANそのものの導入は考えていなかったが，ATMのネットワークインタフェースカード（以降，NICと書く）には組込み型プロセッサを搭載したものがあり，NICに搭載されているプログラムを書き直し独自のプロトコルを開発することができるのではないかと考えていた．国内外でNICを製造している会社に営業経由で情報開示の打診をするが断られた．

そんな中，1995年春，米国Myricom社[4]のMyrinetネットワークはハードウェアレベルで通信パケットの到着の保証をし，またNICに搭載されているプロセッサ上のプログラムの開示ならびに開発環境を提供していることが分かった．当時，イリノイ大学のAndrew Chien教授のグループがMyrinet用にFM（Fast Messages）通信ライブラリを開発し，また，カリフォルニアバークレイ校のDavid Culler教授のグループもMyrinetにAM（Active Messages）通信ライブラリを移植していた．

ここで，Myrinetについて簡単に紹介する．Myrinetはギガビット級のネットワークで，1995年当時8ポートのスイッチが製品化されていた．同軸ケーブルで最大10メートル延ばすことができた．スイッチをつなぎ合わせることにより任意のネットワークトポロジを組むことができた．Myrinetのカードは米国では1,200ドル，8ポートスイッチが2,400ドルであり，1ドル100円計算できる時代であったときには非常に安価なネットワークであった．

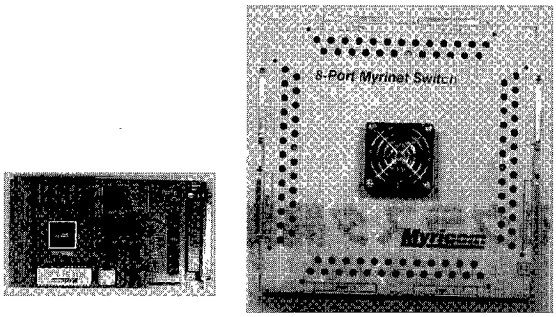


図-1　S-BUS NICと初期のMyrinet Switch

現在では16ポートスイッチも製品化され，ギガビット級のネットワークの中でこの規模のスイッチを製品化し，多くのサイトで使用されているのはMyricom社が最初であろう．また，Myricom社のWWWホームページ[4]を訪れると分かるが，多くの研究機関がMyrinetを利用して通信機能の研究を行っているのが分かる．米国のベンチャ企業における研究機関との交流と市場の形成を垣間見るような印象を持つ．

Myrinetとワークステーションによるクラスタ試作

1995年夏，MyrinetのNIC5枚（現在Myricom社が販売しているボードよりも数世代古い）と8ポートスイッチ1つが到着し，Sun社SPARCstation Model 20（以降SS20と呼ぶ）によるクラスタを構成した．図-1にSunのSバスにささるNICと当時のMyrinetスイッチの写真を示す．

Myrinetスイッチの中身を見た某国立大学の教授や日本の某一流メーカの研究者は，「こりゃあ大学で作るボード並みだな」，「こんな製品は日本のメーカでは出せない」と感想をもらした．初期のスイッチにはバグがあり通信時にエラーが生じることがあった．日本人の感覚では，このような未成熟な製品は販売されては困るわけである．しかし，米国ではこのような新しいネットワーク開発の萌芽期から大学の研究機関が使い，技術開発が進んでいくのである．現在，MyrinetのNICとスイッチは数回の改変が行われ，安定して稼働するようになっている．

RWCワークステーションクラスタ1号機上で，イリノイ大学で開発されたFMを入手してテストした．この当時イリノイ大学のグループが使用していたボードは我々が入手したMyrinetのボードよりも古く，動作しないということが生じ，新しいFMが利用できるようになるのを待っていた．やはり，自前で通信ライブラリを作るのがよいということになり手塚が開発を始めた．

手塚がMyrinet用に開発したライブラリの最初の版は9月頃に稼動していた．この頃，イリノイ大学のAndrew Chien教授が来訪したとき，手塚が開発したライブラリの性能が良いのに驚き，根掘り葉堀り質問してきた．当時，通信ライブラリに名前をつけていなかったが，すでに，AM，FMという通信ライブラリの名前があり，イリノイ大学のグループはAM変調よりFM変調の方が良いんだと冗談を言っていた．それではというので，我々の通信ライブラリはPMにしようということになり，この頃からPMという名前で呼ぶようになった．

PM通信ライブラリ[2]はMyrinet NICのファームウェアであるプログラムを書き直している．Myricom社のファームと異なり，ルーティング情報を静的にして処理をスリム化するとともに，注意深いプログラミングによってプロトコルオーバヘッドを極力小さくすることで，ユーザプロセスとユーザプロセス同士の通信時間（ラウンドトリップ時間）を46マイクロ秒に短縮した．ちなみに，Sバスに接続される現在のMyrinetのNICでは16マイクロ秒のラウンドトリップ時間が達成されている．PM通信ライブラリの作りについては，文献1）に解説記事を書いているので，参照して欲しい．

この当時，SS20とEthernetの組合せで，TCP/IPの通信遅延は数百マイクロ秒以上あった．また，イリノイ大学で開発されたFMはその当時片道通信で25マイクロ秒の遅延であると聞いていた．PMはFMよりもさらに低遅延を達成していた．

並列記述に必要な通信，同期機構などを提供するために，C++言語を拡張したMPC++ Version 1.0を開発していた．Sun上にMPC++実行時ライブラリおよびPMを用いた通信，同期機構を試作した．

RWCワークステーションクラスタ1号機上での実験システムソフトウェアにより，超並列計算機と同様の並列システムが実現できる可能性があることが分かり，マルチユーザ環境を実現するOSであるSCoreシステムの研究開発のために，台数規模の大きいクラスタを構築することにした．

ワークステーション上でのシステムソフトウェアSCore

ワークステーションクラスタ上でのシステムソフトウェアの研究として，プログラミング言語およびSCoreシステムの開発を次のように進めることにした．

MPC++コンパイラはゼロから開発を進めていたが，1995年秋までに拡張機能とテンプレート機能を除くC++機能をコンパイルできる状況であった．しかし，ユーザが利用するには不具合も多かった．一方，MPC++が定義している拡張構文は，C++が持つテンプレート機能で十分実現できるものであることが判明した．そこで，テンプレート機能による並列記述ライブラリとしてMulti Thread Template Library（MTTL）を開発し，MPC++ Version 2.0とした．

OS研究として，並列システムのためのOSカーネルをゼロから作るよりも，既存のOSで，どこまで並列環境が実現できるのかを追求するのが重要であると考えた．Sun社のOSカーネルのソース開示は困難であ

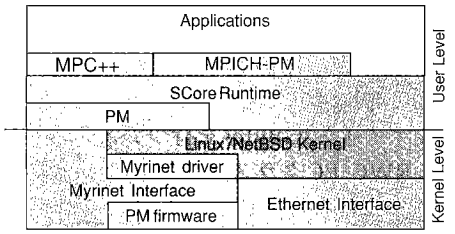


図-2 SCore System Software Architecture

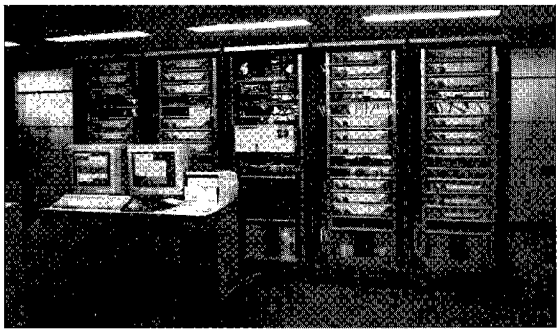


図-3 RWCワークステーションクラスタ2号機

ろうと判断し，OSカーネルには手を加えずにデバイスドライバとユーザレベルでできる機能を用いて実現することとなった．マルチユーザ環境を実現するOSとしてSCore-Dという名前のシステムを構築した．詳細は文献1）を参照して欲しい．

後に，PM，MPC++，SCore-Dを含めて全体のシステム名をSCore Systemと名付けた．図-2にSCore Systemの全体像を示す．フリーでソースコードが入手できるMPI通信ライブラリであるMPICHをPM上に実装した．

36台版ワークステーションクラスタ ーシステム開発にはやはりカーネルの ソースプログラムがないと...ー

32台規模のクラスタシステムを構築するのは初めてのことであり，不明な点が多かった．32台それぞれにローカルなディスクが存在していたら，ディスクが頻繁に故障して管理が大変なのではないか，かといって32台の計算機をディスクレスで運用したときに，サーバやネットワークの負荷が増大して飽和してしまうのではないかという懸念もあった．そこで，ローカルにディスクは持たせつつ8台ごとに1台のサーバが置けるような構成を考え，36台構成のクラスタシステムを構築した（図-3）．クラスタに対して全体で1つのサーバを持たせた．クラスタのコンソールはシリアルラインにして，それらをサーバに接続した．これにより，サーバはファイルサーバおよびコンソールサーバとして利用した．結局，システムはローカルディスクで立ち上げて運用した．

本システムは1996年2月の稼働時から2年間で，メモリなどのトラブルが2度ほど生じたが，それ以外は問題なく快適に使用することができた．

並列ベンチマークプログラムNAS並列ベンチマーク

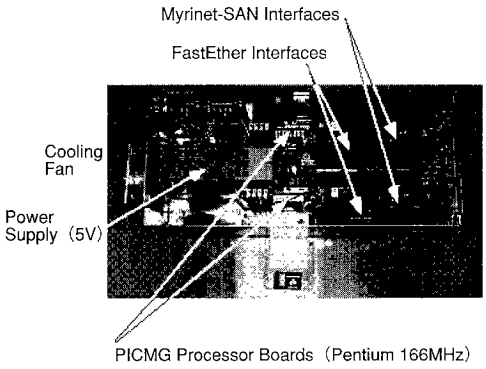


図-4 RWC PCクラスタ1号機のパーツ

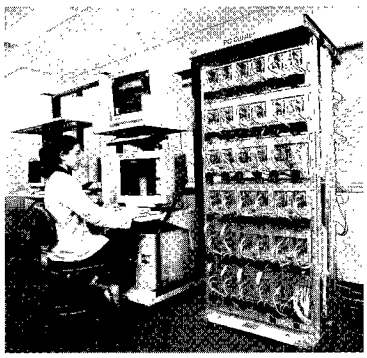


図-5 RWC PCクラスタ1号機の概観

の1つであるCG（共役勾配法による行列計算）をMPC++で記述した性能結果から，その当時現役であった超並列計算機であるCray社T3-Dに匹敵する性能が出ていることが判明した．SCore-Dの性能評価を行ったところ，実行プロセスの切り替え（コンテキストスイッチ）に50ミリ秒とばらつくことが判明した．しかし，なぜ，このようなばらつきがあるのかは分からなかった．この時，カーネルのソースコードがあれば分かるのにとつくづく思い知らされた次第である．

1996年年頭には，Myricom社はPCIバス対応のMyrinet NICを売り出し，パソコンメーカはIntel社Pentium 200MHzによるパソコンを市場に投入していた．ワークステーションの性能とPCの性能との差が縮み，PCを構成要素とするクラスタが十分並列計算機として利用できる時代になってきていた．そこで，PCによるクラスタ作りを1996年春より開始した．

PC32台にフリーOS， その上に自前ソフトウェア

当初，16台規模の小システム構築を検討していた．タワー型のPCを16台並べただけでもスペースをとり，また故障したときの保守性に難点があり問題となっていた．あるとき，堀がPICMGという工業規格ボードを見つけだし，これを用いてコンパクトかつ保守性を考慮したシステムができるだろうと検討しだした．

1996年夏，結局，180cmほどの高さのラックに32台のPICMG[7]によるPC（Pentium 166MHzに64MBytesメモリ）を収納できることが判明し，32台

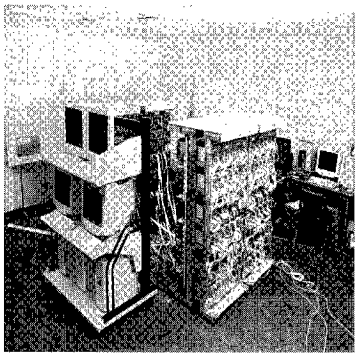


図-6　RWC PCクラスタ2号機の概観

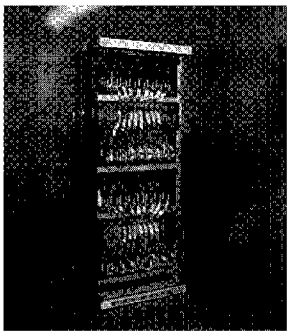


図-7　RWC Alphaクラスタ1号機の概観

規模のPCクラスタを作成することにした（**図-4**, **図-5**）．実装密度を優先するために，各PCにディスクを搭載せずに，ディスクレスとした．オペレーティングシステムにはNetBSDを採用した．RWCワークステーションクラスタにおいてOSカーネルのソースコードがないために，挙動不審の動きをしてもなぜそのような動きをするのか分からなかったという教訓もあり，ソースコードがフリーで入手でき安定して動いていたNetBSDを採用した．本PCクラスタ上でラウンドトリップ通信遅延が15マイクロ秒，通信バンド幅が100MBytes/sec以上の性能を達成した．

RWC PCクラスタ1号機のハードウェアアセンブリは9月末に終了し，その後，システムソフトウェアおよびデモプログラムとして実時間並列画像処理，並列マンデルブロー，並列nqueenなどを作成して，11月に開催された国際会議SC '96の研究展示で公開した．国際会議SC（Super Computing）は並列計算機関係では一番規模の大きい国際会議である．

展示場では，PCクラスタのコンパクトさ，マルチユーザ環境かつ高性能を実現したSCoreオペレーティングシステムと並列プログラミング言語MPC++によるトータルシステムソフトウェアが提供されている点が注目された．何人かの参加者から，Linuxオペレーティングシステム上で稼動しないかといった質問を受けた．米国ではLinuxとフリーに手に入るソフトウェアを組み合わせたクラスタシステムであるBeowulfプロジェクトが注目されつつあった．

なお，RWC PCクラスタ1号機は1996年10月の稼働時から約2年間で，Ethernetカードを2枚ほど交換した程度で安定して稼動している．

128台のPC，32台のDEC Alpha，OSはLinux

32台規模でのPCクラスタが実現された後，並列システムとしてスケーラビリティの検証を行うとともに並列応用プログラムを書くユーザに利用してもらえるシステムを開発することにした．スケーラビリティは128台規模を目指すことにしたが，一気に128台を製作するのは難しいと判断し，まず64台規模のRWC PCクラスタ2号機（**図-6**）を製作した．RWC PCクラスタ1号機同様PICMGを用い，プロセッサはPentium Pro 200MHz，主記憶256MBytes ECC付，ローカルディスク2GBytesを持つシステムとなった．

RWC PCクラスタ2号機は1997年10月に稼働し，11月に開催されたSC '97の研究展示に出展した．新情報処理開発機構の並列応用つくば研究室のグループがPCクラスタ上に開発した並列蛋白質情報解析システムPAPIA[6]などのデモを行い，注目を集めた．

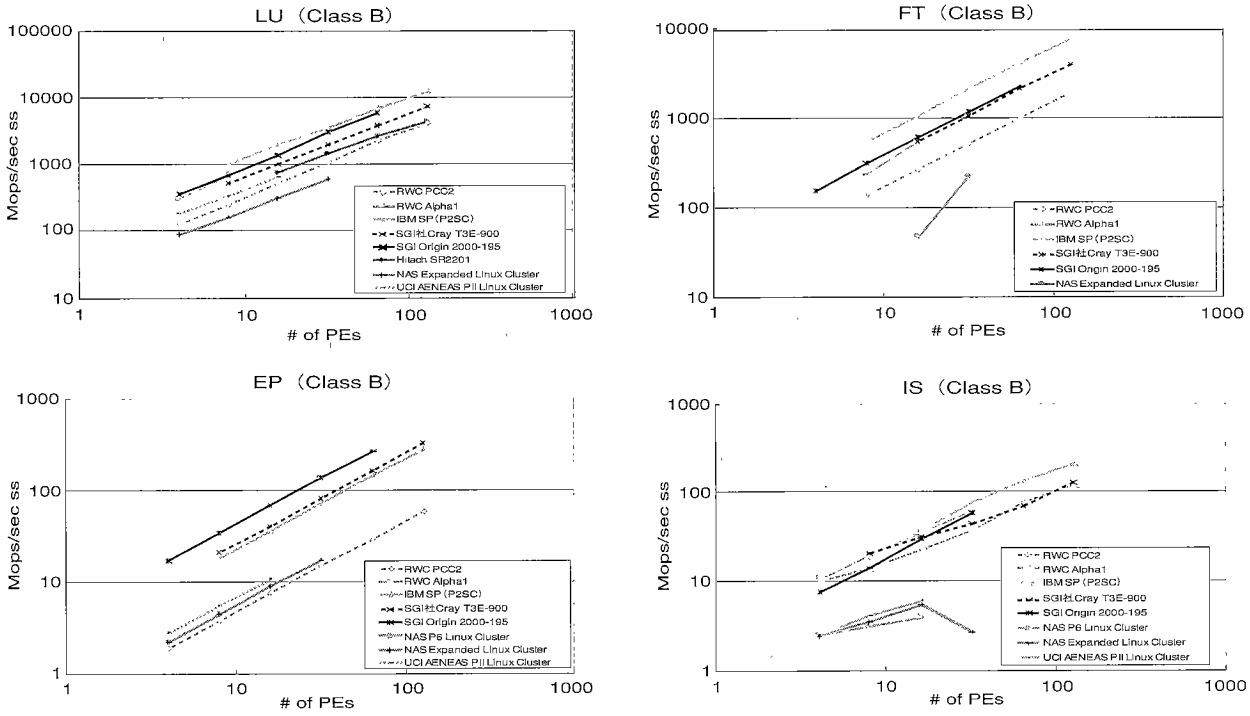
RWC PCクラスタ2号機は，1997年10月の稼働より約1年間で，ディスクが1つ故障した程度であり快適に利用することができている．その後，PAPIAシステム用にRWC PCクラスタ2号機がもう1台製作され，現在，Internet経由でPAPIAシステムが利用できるようになっている．

1998年夏，予定通り，RWC PCクラスタ2号機を128台規模に拡張した．さらに，今まで利用していたNetBSDカーネルからLinuxカーネルへと移行した．1996年以来米国ではLinuxカーネル上でクラスタシステムが構築されるのが主流になっていたため，我々のシステムが幅広く利用されるためには，Linuxカーネルに移行する必要があるという判断からであった．

RWC PCクラスタ2号機128台と他の並列計算機（商用並列計算機としてIBM社SP，SGI社Cray T3-E，SGI社Origin 2000，クラスタシステムとしてNASAが製作したPCを使ったクラスタシステム2つ）との性能比較として，NAS並列ベンチマーク[5]による結果を**図-8**に示す．NAS並列ベンチマークは，8つのベンチマークプログラムから構成されている．ここでは紙面の都合上，LU，FT，EP，ISの4つを掲載する．LUは一次方程式，FTはフーリエ変換，EPはモンテカルロシミュレーション，ISは整数ソートをそれぞれ解くプログラムである．なお，EPは浮動小数点演算によるルートとログ計算の性能が支配的なベンチマークである．

クラスタ技術の次のステップとして，新情報処理開発機構では，異なるアーキテクチャの計算機群の上で，その異機種性をユーザに意識することなくプログラミングを可能とするシームレス並列分散システムの研究開発を行っている．RWC PCクラスタ2号機はこのための研究インフラの1つとして使われる．異機種環境上で実験するためにはPC以外のアーキテクチャによるクラスタが必要である．そこで，DEC社Alpha 21164Aプロセッサを32台利用したRWC Alphaクラスタ（**図-7**）を製作している．図-8にはRWC Alphaクラスタ1号上での実行結果も載せた．

PCクラスタの価格は128台構成のPCクラスタの価格を128で割ると，およそ1台あたり100万円程度であり，Alphaクラスタは1台あたり200万円程度となる．PCクラスタにおけるプロセッサボード，メモリ（256MBytes ECC付），ネットワーク，筐体などにかかる費用の割合を**図-9**に示す．メモリとネットワークにその費用の大半が占められているのが分かる．



（他のシステムの結果はhttp://science.nas.nasa.gov/Software/NPB/NPB2Results/から引用）

図-8　RWC PCクラスタ2号機およびRWC Alphaクラスタ1号機の性能

DO IT YOURSELF クラスタシステム

新情報処理開発機構では，SCoreシステムを利用していただき，ユーザからのフィードバックによるさらなるシステムの改良を目的として，WEB経由で許諾書ベースにシステムソフトウェアの配布を行っている．

http://www.rwcp.or.jp/lab/pdslab/dist/

我々が開発したSCoreシステムはMyrinetネットワークによって接続されたクラスタシステム上でしか稼働しない．現在，Gigabit Ethernet上でも高性能通信を実現したSCoreシステムが動くように開発を進めている．

100 Base-Tなどの既存のネットワークおよびTCP/IPネットワークプロトコル上での通信ライブラリを利用してクラスタシステムを作りたいという人は，Red Hat社から配布しているExtreme Linux CD-ROM[8]を勧める．Extreme LinuxにはLinuxオペレーティングシステムと各種ネットワークドライバ，MPI，PVMなどの通信ライブラリが収録されている．また，既存ネットワークにExtreme Linuxの組合せでコンパクトにまとめた小規模PCクラスタの導入を検討している人は，次のURLを覗いてみることをお勧めする．

http://www.altatech.com/

Extreme Linuxは，現在，Red Hatと米国DOE（エネルギー省）傘下のLos Alamos国立研究所が中心になって次の版をまとめている[3]．Extreme Linuxの次の版に，新情報処理開発機構が開発したSCoreシステムの一部も収録されることになっている．

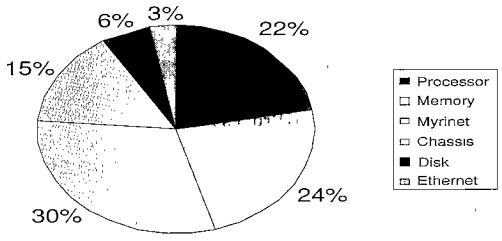


図-9　クラスタ1ノードにかかるコストの比率

コモディティハードウェアの組合せによるクラスタはいわばDIY（Do It Yourself）型の手作り並列計算機という色彩が現在は強い．計算科学を研究している人が，自ら，計算機システム屋の作っているようなクラスタを真似して作るのはハードルが高すぎる．ハードウェアの組合せだけでなく，ソフトウェアをどう組み合わせたらよいかという問題がある．Extreme Linuxや我々の研究アクティビティにより，誰もが使える並列計算機環境が実現できるようになることを期待している．

参考文献
1）石川: コモディティハードウェアを用いた並列処理技術，情報処理，Vol.39, No.8, pp.784-791（Aug. 1998）．
2）Tezuka, H., Hori, A., Ishikawa, Y. and Sato, M.: PM: An Operating System Coordinated High Performance Communication Library, In Sloot, P. and Hertzberger, B. editor, High-Performance Computing and Networking, Vol.1225 of Lecture Notes in Computer Science, Springer-Verlag, pp.708-717（Apr. 1997）．
3）http://www.extremelinux.org/
4）http://www.myri.com/
5）http://science.nas.nasa.gov/Software/NPB/
6）http://www.rwcp.or.jp/papia/
7）http://www.picmg.com/
8）http://www.redhat.com/extreme/

（平成10年9月21日受付）