計算機アーキテクチャ 119－7
（1996．8．27）

PC とギガビット LAN による PC クラスタの構築

手塚宏史[†] 堀 敦史[†] 石川 裕[†]
曽田哲之[††] 原田 浩[††]
古田 敦[††] 山田 努[††]

我々は，PICMG 規格の産業用 PC と商用のギガビット LAN である Myrinet を用いた PC クラスタを構築中である．PICMG 規格の PC をノードプロセッサに用いることによって，Unix ワークステーションを用いたワークステーションクラスタよりも価格性能比が高く，大規模な並列処理システムを容易に構築することができ，CPU の性能向上に伴うシステムのアップグレードも容易になる．

A PC cluster using PC's and a Giga-bit LAN

HIROSHI TEZUKA,[†] ATSUSHI HORI,[†] YUTAKA ISHIKAWA,[†]
NORIYUKI SODA,[††] HIROSHI HARADA,[††] ATSUSHI FURUTA[††]
and TSUTOMU YAMADA[††]

We are developping a PC cluster using PICMG standard industrial PC's and Myrinet commercialized giga-bit LAN. The PC cluster has several advantages rather than workstation clusters using Unix workstations. It is more cost-effective, can consist of larger number of nodes, and can be upgraded to higher performance CPU.

1. はじめに

価格性能比の高い並列処理マシンを構成する方法として，市販の Unix ワークステーションを高速相互接続網によって結んだワークステーションクラスタが注目されている[1),2)]．相互接続技術の高速化によって，市販のワークステーションを用いても，専用の超並列機に比べて遜色ない性能が得られるようになった[3)]．我々は 36 台の SPARCstation 20/71 を商用のギガビット LAN の一つである Myrinet で接続したワークステーションクラスタ[4)]を開発し，マルチユーザの並列プログラミング環境を構築している[5),6)]．

ワークステーションクラスタには，市販の Unix ワークステーションを用いることで，開発期間が短い，その時点で最高性能の CPU が使える，Unix 環境をそのまま利用できるなどの大きな利点がある．しかし，その反面，ハイエンドの Unix ワークステーションは比較的高価であり，ワークステーションの匡体をそのまま用いるために実装密度が低く，大規模なクラスタシステムの構築は困難である．また，Unix ワークステーションの性能向上の速度は速いため，すぐシステムが陳腐化してしまう．

これらワークステーションクラスタの問題点は，産業用 PC をノードプロセッサに用いることによって解決することができる．我々は，現在のワークステーションクラスタよりも価格性能比の高い，大規模な並列処理システムを構築するために，産業用 PC とギガビット LAN を用いた PC クラスタを設計中である．

本論文では，第 2 節で PC をクラスタのノードプロセッサに用いる利点について述べ，第 3 節で我々が設計している PC クラスタについて述べる．次に第 4 節でプロトタイプによる性能の予備評価の結果を述べ，第 6 節でまとめと今後の予定を述べる．

2. PC クラスタ

2.1 ノードプロセッサとしての PC

近年の PC の高性能化，低価格化は，その巨大な市場を背景にして著しいものがあり，PC をノードプロセッサに用いることによって，より価格性能比の高いクラスタシステムを実現できるようになった．クラスタマシンのノードプロセッサに PC を用いることには次のような利点がある．

† 技術研究組合 新情報処理開発機構 つくば研究センタ
超並列ソフトウェア研究室
Massively Parallel Software Lab. TRC,
Real World Computing Partnership
http://www.rwcp.or.jp
†† 株式会社 SRA

開発期間が短い

市販の PC を用いることにより，ハードウェアの開発が不要であり，ワークステーションクラスタと同様に短期間でシステムを構築することができる．

低コスト

PC は業界標準部品によって作られているために，量産効果により低コストである．

業界標準アーキテクチャ

PC-AT アーキテクチャに準拠しているため，異なるメーカの PC でも同じソフトウェアが動作するので，ハードウェアの選択の幅が広い．

高速な I/O バス

PC の I/O バスに用いられている PCI[7] は，現時点で最高性能の I/O バスの一つであり，高い I/O 性能を期待できる．

豊富な I/O

PCI, ISA バス用の豊富な I/O カードを利用することが可能である．

フリーソフトウェア

NetBSD[8], FreeBSD[9], Linux[10] などのフリーな Unix ライクなオペレーティングシステムによって，Unix ワークステーションと同様なソフトウェア環境を構築できる．

しかし，通常の民生用 PC をそのままノードプロセッサに用いたのでは，マザーボードと I/O カードを垂直に実装する構造のために，実装密度を高めることができない．また，CPU その他の能動部品がマザーボード上に実装されているために，保守や CPU のアップグレードが容易とは言えない．我々は，次に述べる産業用 PC の CPU カードをノードプロセッサに用いることでこれらの問題を解決し，実装密度が高く保守の容易な PC クラスタを構築できると考えた．

2.2 産業用 PC

今回我々が用いた産業用 PC は PICMG(PCI Industrial Computer Manufacturers Group)[11] の策定した Passive Backplane PCI-ISA 規格☆と呼ばれるものである．PICMG 規格の CPU カードは，フルサイズの PCI カードの大きさの上に，CPU，キャッシュ，メモリ，ディスク I/F，シリアル/パラレル I/F など PC-AT 互換機のマザーボードの機能☆☆を搭載している．図 1 に PICMG 規格の CPU カードの例を示す．

この PICMG 規格の CPU カードをノードプロセッサに用いる利点は次の通りである．

☆ PICMG では，この他に，PCI バスのスロット数を拡張するための PCI-PCI Bridge や，ユーロカードを用いる Compact-PCI[12] などの規格を定めている．

☆☆ さらに，SCSI I/F やビデオ I/F などの PC-AT 互換機の基本機能すべてを搭載した CPU カードもある．

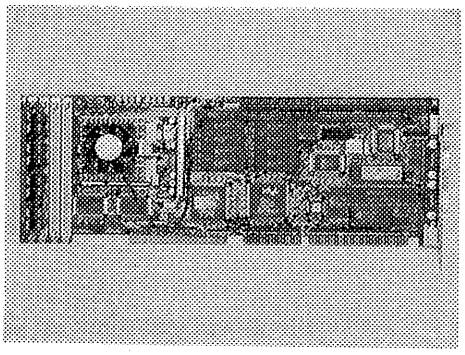

図 1　Passive Backplane CPU カード

実装密度が高い

PC-AT 互換機の機能が CPU カードに高密度実装されており，I/O カードと共に受動部品のみからなるバックプレーンのスロットに垂直に装着して実装するため，Unix ワークステーションや民生用の PC を用いた場合に比べて実装密度を高くできる．

保守が容易

能動部品はすべて着脱可能なカード上にあるため，故障時の交換などの保守が容易である．

アップグレードが容易

マイクロプロセッサの性能向上は著しいために，開発時点で最高性能の CPU を用いても短期間で陳腐化してしまう．産業用 PC では，CPU カードだけを交換することによって，低コストで最新の CPU にアップグレードできる．

我々はこれらの考察から，PICMG 規格の CPU カードを PC クラスタのノードプロセッサに用いることにした．プロセッサの性能の相違や，ディスクの有無などのために単純比較はできないが，産業用 PC を用いることによって，ノードプロセッサ当たりのコストはワークステーションクラスタの約半分，実装密度は約 4 倍になると見積られた．

2.3 Myrinet

我々がワークステーションクラスタおよび PC クラスタに用いた，Myrinet は Myricom[13] 社が開発，商品化したギガビット LAN[14] である．Myrinet のホストインタフェースには LANai と呼ばれる RISC プロセッサが搭載されており，Myrinet 上の通信プロトコルを制御する．ホストインタフェース間を接続するスイッチはカットスルー・ルーティングを行なうクロスバ型のスイッチであり，ネットワークトポロジを比較的自由に構成することができる．Myrinet は次のような特徴を持っている．

高性能

160M バイト/秒 × 双方向の非常に高いバンド幅を持ち，ハードウェアの通信レイテンシは数マイクロ秒以下と小さい．

プログラマブル

ホストインタフェースの技術情報は Myrinet のユーザに公開されており，LANai プロセッサのプログラム開発環境も提供されるため，独自の通信プロトコルを実装することができる．

SBus，PCI 用ホストインタフェース

ワークステーションの SBus，PC の PCI バス用のホストインタフェースが用意されている．

我々は，Myrinet が他の LAN に比べて高性能であること，オンボードプロセッサがプログラマブルであることに着目して，ノードプロセッサの相互接続網として Myrinet を採用した．

3. PC クラスタの構成

3.1 ハードウェア

我々が設計中の PC クラスタは，次に示すような**ノード PC** と **モニタ PC** の 2 種類の産業用 PC から構成されている．

ノード PC

CPU カード	PICMG
主記憶	64MB
高速ネットワーク	Myrinet
低速ネットワーク	100BASE-T
シリアル I/F	RS-232C × 2

モニタ PC

CPU カード	PICMG
主記憶	64MB
低速ネットワーク	100BASE-T
外部ネットワーク	10/100BASE-T
ディスク	2GB
シリアル I/F	RS-232C × 32

2 台のモニタ PC はノード PC のファイルサーバや外部ネットワークとのルータの機能を持つ．PC クラスタ内の高速ネットワーク (Myrinet) はノード PC 間の通信に，低速ネットワーク (100BASE-T) はノード PC のブートや標準入出力やファイルシステムなどの TCP/IP を用いたモニタ PC との通信に用いられる．また，図 2 には示していないが，ノード PC のコンソールやデバッグのために，各ノード PC とモニタ PC は 2 本の RS-232C で結ばれている．これら以外の I/O デバイス☆は別匡体に格納され，Myrinet によってノード PC と接続される予定である．

☆ 並列ディスク，ビデオ入出力などを検討している．

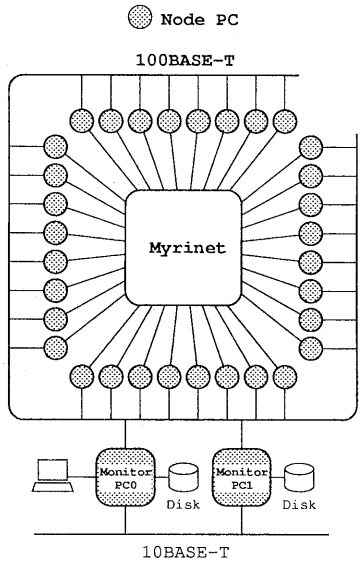

図 2　PC クラスタの構成

現在設計中の PC クラスタは 32 個のノード PC，2 個のモニタ PC，Myrinet スイッチ，および 100BASE-T の HUB が一つのラックに実装される．図 3 に PC クラスタの実装 (案) を示す．現在の我々が使用しているワークステーションクラスタでは一つのラック当たり 9 台のワークステーションを実装しているので，PC クラスタでは，ワークステーションクラスタの約 4 倍の実装密度が得られることになる．

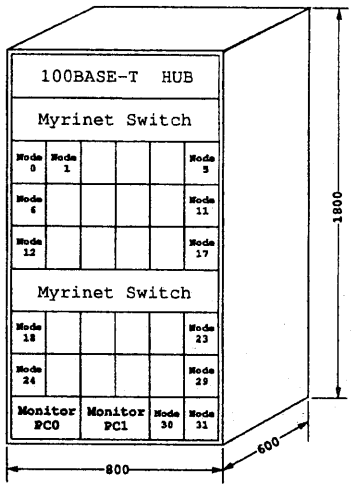

図 3　PC クラスタの実装 (案)

3.2 ソフトウェア

PC クラスタのノード PC およびモニタ PC 上では NetBSD[8]が動作する．ノード PC はディスクを持たないため，オペレーティングシステムは，我々の開発した、100BASE-T を介したブロードキャスト・ブート・プロトコルによって，モニタ PC から一斉にブートされ，以後，ノード PC はディスクレスシステムとして動作する．

Myrinet を介したノード PC 間の通信には PM[6]を用いる．PM は我々が開発した通信ライブラリで，ポーリングによる低レイテンシ，高バンド幅の通信，メッセージ配送の保証，マルチユーザをサポートするためのネットワークコンテキストスイッチなどの機能を持っている．各ノード PC では NetBSD 上のデーモンプロセス SCore-D[5] が並列プログラミング環境 SCore をサポートする．SCore-D は Unix のシグナルを用いてユーザプロセスの実行を制御し，TSSS(Time Space Sharing Scheduling)[15] に基づいたギャングスケジューリングを行なう．このようにワークステーションクラスタと同様な構成をとることによって，短期間で PC クラスタを稼働させることができる．将来は，現在開発中のオペレーティングシステム SCore-P がノード PC 上で動作する予定である．

4. 予備評価

現在 PC クラスタは設計段階であるため，2 台の PICMG 規格の CPU カード*を用いた PC を Myrinet でつないだノードプロセッサのプロトタイプ (以後 PICMG で示す) を用いて，メモリバンド幅，I/O バンド幅，PM による通信性能の予備評価を行なった．また，現在稼働している SPARCstation 20/71 クラスタのデータ (以後 SS20/71 で示す) についても計測して PICMG との比較を行なった．

CPU 性能については，すでに同様の CPU についての多数の評価結果が存在するために，今回は特に評価を行なわなかったが，ベンチマークの標準化団体 SPEC[16] の定めた SPEC95 によると，整数演算性能を示す SPECint95 の値が Pentium(166MHz) は 4.52，SS20/71 は 3.11 であり，浮動小数点演算性能演算性能を示す SPECfp95 の値は Pentium が 3.40 SS20/71 が 3.10 である．測定条件が異なるために我々の用いた CPU カードにはそのまま適用することはできないが，整数演算性能は Pentium が勝るが，浮動小数点演算性能はほぼ同等であると考えられる．

測定に用いたシステムの諸元を以下に示す．

PICMG

CPU	Pentium 166MHz
チップセット	Triton
二次キャッシュ	256KB
メモリ	32MB
I/O バス	PCI 33MHz
ノード間結合	Myrinet

SPARCstation20/71

CPU	SuperSPARC II 75MHz
二次キャッシュ	1MB
メモリ	32MB
I/O バス	SBus 25MHz
ノード間結合	Myrinet

なお，両方のシステムとも，実験に使用した Myrinet はプロトタイプのため，データ転送速度は 80M バイト/秒×双方向である．

4.1 メモリバンド幅

CPU が主記憶の連続領域をアクセスした場合の Read, Write, Copy のバンド幅について測定を行なった．キャッシュの影響を排除するために，二次キャッシュよりも大きな領域についてアクセスし，さらに，キャッシュラインの影響を見るために，アクセス開始位置をずらしながらバンド幅を測定した．Copy については，読み出し側のアドレスをを固定し，書き込み側のアドレスをずらしながら測定した．図 ?? には，測定結果の最小値と最大値のみを示した．

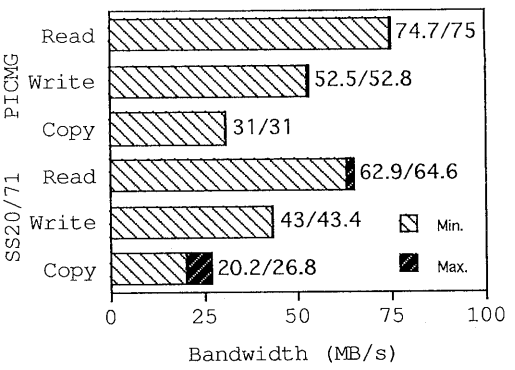

図4 メモリバンド幅

このように，今回測定した PICMG CPU カードのメモリバンド幅は SS20/71 よりも高い．また，SS20/71 では特定のアドレス位置の場合にバンド幅が低下しているのに対して，PICMG ではそのような傾向は見られなかった．

* Advantech 社製 PCA-6157.

4.2 I/O バンド幅

Myrinet のホストインタフェース上の DMA エンジンを利用して，主記憶と Myrinet のホストインタフェース上の SRAM 間で DMA 転送を行ない，I/O バスのバンド幅を測定した．比較のために，CPU のプログラム I/O によるバンド幅についても測定を行なった．DMA のバースト転送を有効にするために，1 度のデータ転送の長さは 128KB として，複数回のデータ転送を行なってバンド幅を求めた．

図 5 に測定結果を示す．ここで，CPU/R, CPU/W, DMA/R, DMA/W は，それぞれ，プログラムによる I/O データの読み出し (読み出したデータは捨てる)，プログラムによるダミーデータの I/O への書き込み，I/O から主記憶への DMA 転送，主記憶から I/O への DMA 転送についての測定結果である．

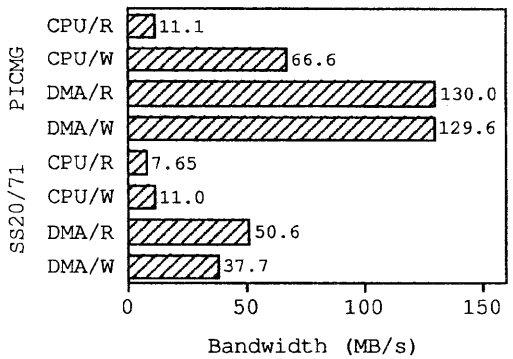

図 5 I/O バンド幅

図 5 のように，今回測定した PICMG の PCI バスは，測定項目すべてについて SS20/71 の SBus よりもかなり高いバンド幅を示している．CPU Read はほぼバスクロック比 (33MHz:25MHz) であるが，その他についてはバスクロックや，第 4.1 節のメモリバンド幅の違いを考慮してもそれ以上に違いが大きい．これは，SS20/71 の SBus のデータ転送のバースト長が 16 ワードなのに対して，PICMG の PCI バスのバースト長がより長い[17]ことによるものである．しかし，PCI バスのデータ転送速度は使われているチップセットによって大きく変わる[18]ことが知られているため，CPU カードの選択には注意が必要である．

4.3 ノード間通信性能

Myrinet 上に作成した通信ライブラリ PM を用いて，異なるノードプロセッサ間通信のラウンドトリップ時間とバンド幅を測定した．メッセージは送信ノードのユーザプロセス内の固定領域に割り当てられたバッファから，受信ノードのユーザプロセス内のバッファに転送される．測定はメッセージ長を 8 バイトから 8192 バイトまで変えながら行なった．ラウンドトリップ時間の測定結果を図 6 に示す．

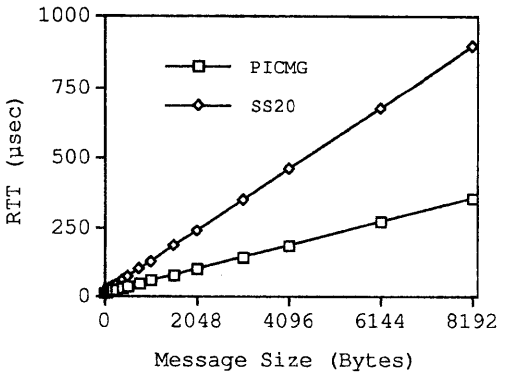

図 6 PM のラウンドトリップ時間

図 6 からは分かりにくいが，メッセージ長が 8 バイトの場合のラウンドトリップ時間は，PICMG で 15.5 マイクロ秒，SS20/71 で 19.5 マイクロ秒が得られた．このような差があるのは，Myrinet のホストインタフェース上の LANai プロセッサがバスクロックの速度で動作しているために，バスクロックのより速い PCI の方がプロトコル処理のオーバーヘッドが小さくなるためである．

また，メッセージサイズが大きい場合に PICMG の方が SS20/71 よりもラウンドトリップ時間が短いのは，次に述べるように PICMG の PM のバンド幅が SS20/71 に比べて高いからである．

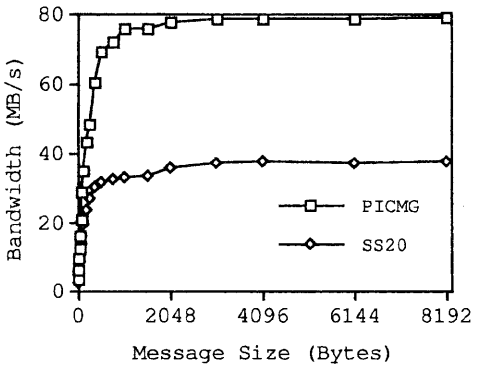

図 7 PM のバンド幅

PM のバンド幅の測定結果を図 7 に示す．SS20/71 のバンド幅が約 38M バイト/秒程度で頭打ちになっているのは，第 4.2 節で述べたように SBus の DMA 転送の速度が約 40M バイト/秒しかないためである．これに対して，PICMG の PCI バスのバンド幅は約 130M バイト/秒で Myrinet のバンド幅より高いため，ほぼ Myrinet の最大バンド幅いっぱいの約 78M バイト/ 秒が得られている．PCI バスの DMA 転送

の速度にはまだ余裕があるため，160M バイト/秒の Myrinet を用いれば，さらに高いバンド幅を得ることができると考えられる．

このように，メモリ，I/O，ネットワークなどの点で，産業用 PC はミッドレンジの Unix ワークステーションの性能を凌駕しており，ワークステーションクラスタに比べて高性能な PC クラスタの構築が可能である．

5. 関連研究

Mosix project[19] では 32 台の PC を Myrinet で接続し，BSD/OS[20] を拡張したオペレーティングシステムによって，資源の共有を実現している．Shrimp project[21] では，Intel Paragon のネットワークによって PC 接続し，マルチコンピュータシステムを実現しており，そのためのネットワークインタフェースなどを新しく設計している．UCB の NOW プロジェクトや，Illinois Univ. の CSAG でも，複数の Pentium プロセッサを Myrinet で接続したクラスタシステム[22),23] が計画されている．また，Beowoulf Parallel Workstation[24] は複数のディスクを持った Pentium を 100Mbps の Ethernet で結び，低コストで高いバンド幅のディスクシステムを実現している．

6. まとめと今後の予定

PICMG 規格の産業用 PC と Myrinet を用いることによって，価格性能比の高い，大規模な PC クラスタを短期間で開発することが可能である．我々は，今年中に 32 ノードの PC クラスタを完成し，NetBSD 上の SCore-D によるマルチユーザの並列プログラミング環境 SCore を実装する予定である．また，現在ノードプロセッサ上のオペレーティングシステムとして SCore-P の開発を行なっている．今後は，PC クラスタの並列 I/O システムとして，並列ディスク，ビデオ入出力装置などの開発を計画している．

参考文献

1) http://now.cs.berkeley.edu/.
2) http://www.cs.wisc.edu/ wwt/cow.html.
3) 田中良夫，久保田和人，佐藤三久，関口智嗣．並列アルゴリズムにおける Collective 通信の性能比較．情報処理学会研究会報告 HPC．情報処理学会，August 1996.
4) http://www.rwcp.or.jp/people/mpslab/score/wsc/wsc.html.
5) 堀敦史，手塚宏史，石川裕，曽田哲之，小中裕喜，前田宗則．並列プログラム実行環境のワークステーションクラスタ上での実装．並列処理シンポジウム JSPP'96．情報処理学会，June 1996.
6) 手塚宏史，堀敦史，石川裕．ワークステーションクラスタ用通信ライブラリ PM の設計と実装．並列処理シンポジウム JSPP'96．情報処理学会，June 1996.
7) http://www.intel.com/product/tech-briefs/pcibus.htm.
8) http://www.netbsd.org.
9) http://www.freebsd.org.
10) http://www.linux.org.
11) http://www.picmg.com.
12) 解説．"PCI バスの用途が多様化，産業機器やノート・パソコンに"．日経エレクトロニクス，pp. 117–124，7 月 15 日 1996.
13) http://www.myri.com.
14) N. J. Boden, D. Cohen, R. E. Felderman, A. E. Kulawik, C. L. Seitz, J. N. Seizovic and Wen-King Su. "Myrinet – A Gigabit-per-Second Local-Area Network". *IEEE MICRO*, Vol. 15, No. 1, pp. 29–36, February 1995.
15) A. Hori, Y. Ishikawa, H. Konaka, M. Maeda and T. Tomokiyo. "A Scalable Time- Sharing Scheduling for Partitionable, Distributed Memory Parallel Machines". In *Proc. 28th Annual Hawaii Int. Conf. on System Sciences*, pp. 173–182, 1995.
16) http://www.specbench.org.
17) http://www-cs.intel.com/oem_developer/chipsets/pci/general/pci001.htm.
18) http://www.myri.com:80/myrinet/performance/DMAperf.html.
19) http://www.cs.huji.ac.il/mosix/.
20) http://www.bsdi.com.
21) http://www.cs.princeton.edu/shrimp/.
22) http://http.cs.berkeley.edu/ wendyh/cs258/fault_now.html.
23) http://www-csag.cs.uiuc.edu/projects/clusters.html.
24) T. Sterling, D. J. Becker, D. Savarese, M. R. Berry, C Reschke. Achieving a balanced low-cost architecture for mass strage management through multiple fast ehternet channels on the beowulf parallel workstation. In *Proceedings of the 10th International Parallel Processing Symposium*, April 1996.