計算機アーキテクチャ 137－23
ハイパフォーマンス 80－23
コンピューティング
（２０００．３．３）

ローエンドPCを用いたPCクラスタの構築

増田景一，坂上仁志，高橋　豊
姫路工業大学工学部情報工学科

計算用途に使用されるPCクラスタは，PCの中では高性能なハイエンドPCによって構成されることが一般的である．しかし，計算用途以外のデータベースやWWWサーバとして使用する場合には，I/O性能に比べ計算性能はそれほど重要ではないためローエンドPCでも十分な場合も多い．また，並列プログラムの開発環境として，安価で小型なクラスタが必要になる場合もある．我々は価格を抑える事を追求した，ディスクレスのローエンドPCを100Base-TXのEthernetで結合させたPCクラスタを構築したので報告する．

Construction of a PC cluster with Low-end PCs

Keiichi Masuda, Hitoshi Sakagami and Yutaka Takahashi
Computer Engineering, Himeji Insitute of Technology

Generally, PC clusters for calculation are composed with High-end PCs. The performance of Low-end PC is lower than that of High-end PC, but low-end PC has adequate a performance in some cases such as database or WWW server because the calculation performance is less important than the I/O one for those cases. In addition, an inexpencive and compact clusters are sometimes needed as programming environments for parallel software developments. We have described the construction of a PC cluster with diskless Low-end PCs connected via the 100Base-TX Ethernet, looking for low prices as a primary goal.

1. はじめに

近年，パーソナルコンピュータ(PC)の性能は急速に高まっている．PCの演算能力，グラフィック性能やストレージとしての能力は，ワークステーション(WS)やその用途向け専用に開発されたコンピュータの能力に迫っている．最近ではWSも一部はPCの構成部品を用いて製作されており，PCとWSの違いは次第に無くなってきているが，ここではUltraSparcといったCPUを採用して業務用途に使用されるような高価なコンピュータをWSと呼ぶことにする．

PCクラスタは従来のWS等で構成されているクラスタを，PCを用いて構築したシステムである．PCは一般にWSに比べて価格が低く，さらにPC間を結合するネットワークに汎用的なものを用いることができるため，PCクラスタは比較的安価に構築できる．PCクラスタはWSで構成される高価なクラスタに劣らない高性能を実現しながら，かつ価格でも数分の一に抑えることが可能なシステムとして盛んに製作されている．

PCの構成部品の中で中心となる物は，CPUとそれに対応するマザーボード，ハードディスクドライブ(HDD)，メモリモジュールおよびグラフィックボードである．さらにサウンドボード，SCSIボードやイーサネットボードといったボードを目的に応じて搭載している．現在のPCは用いられるCPUやグラフィックボードの違いから，大きく分けてハイエンドPCとローエンドPCに別れている．Intel社はハイエンド用のCPUとし

てPentiumIII/Xeonを，ローエンド用のCPUとしてCeleronを位置づけている．またAdvaced Micro Devices(AMD)社はハイエンド用のCPUとしてAthronを，ローエンド用としてK6-2を位置づけている．マザーボードに関しても，マルチプロセッサ構成を可能とするマザーボードや，高帯域のPCIバスを採用しているマザーボードがハイエンドPCに用いられることがある．それに対しローエンド用のマザーボードは，最大メモリ搭載量が少ないといった制限があるものの，価格が抑えられている．グラフィックボードでは大容量の専用メモリと高速なグラフィックエンジンを搭載したボードが，ハイエンド用や特定用途向けとして用いられる．一方ローエンド用にはマザーボード上のノースブリッジ内に統合されたグラフィック機能を使用することで価格を抑えたシステムがある．ストレージには，高速性や信頼性を高めるためにRAID構成を用いたHDDをハイエンド用として使用する場合もある．その一方で，ローエンド用として用いられるIDEのHDDは非常に安価になっている．用途や価格に応じて，上記のような構成を組み合わせた様々なPCが存在している．そのため価格も様々であるが，一部の高価な部品を用いない限り，WSや専用コンピュータの価格を超えることはない．

従来から製作されているPCクラスタはPentiumProやPentiumII-XeonといったCPUをSMP構成にしたPCを用いている．よって，それらのPCはハイエンドPCに分類できる．我々はPCクラスタのノードとして，従来から用いられているハイエンドPCではなく，ローエンドPCを用いたPCクラスタを製作した．その目的はPCクラスタをできる限り低い価格で製作することである．

2. PCクラスタの構成

2.1 構成部品の検討

PCクラスタの製作にあたっては，ノードの構成に関して様々な選択肢が考えられる．CPUに何を用いるかは，PCの性能を左右する重要な選択である．CPUの選択では，整数演算性能と浮動小数点演算性能やキャッシュ性能といった要素の他にも様々な判断材料があるが，今回の製作にあたっては，できる限り低い価格のCPUを選択している．

マザーボードはCPUとメモリやPCIバス等を接続する重要な基板である．マザーボードはチップセットと呼ばれる一連のLSIを中心に構成されている．チップセットの中で中心となるのはノースブリッジと呼ばれるチップであり，メモリインタフェースやPCIバスを備え，CPUとそれらのリソースを接続する．マザーボードの中には，ノースブリッジ内にグラフィック機能を内蔵したボードが存在する．この機能はグラフィック用のメモリとして主メモリの一部を占有して使用する．そのため，メモリアクセスが増加し，CPU側からのメモリアクセス性能が低下する．PCクラスタの各ノードは，通常時にはグラフィック機能を必要としないので，この機能を持たないマザーボードを選ぶ．

HDDの性能はデータベース等の用途では重要な要素である．しかし，本クラスタでは低価格に抑えることを目的としているため，クラスタの各ノードはディスクレスシステムとする．HDDはブート用のブートサーバに1台だけ用意し，クライアントであるノードの各ディスク領域はサーバのHDD上に確保する．HDDが1台のみになることで，管理が簡単になり，クライアントの環境を均一にすることが容易になる．また，稼働部品が減ることにより，全体の故障率を下げることができ，さらにバックアップ等の作業も一度で済む等の利点がある．

ネットワークをどのように構成するかもクラスタの性能を決める重要な要素である．Gigabit-Ethernetシステムは極めて高い性能を実現できるが，ネットワークインタフェースカード(NIC)や特に集線装置であるスイッチングHubは高価である．100Base-TXを用いるFast-Ethernetと呼ばれるネットワークは，10Base-Tを用いるネットワークに比べて10倍高速である．そして，構築要する費用にほとんど差はない．また，ネットワーク環境が発展するに従って，スイッチングHubも安価になっている．しかし，多数のポートを持つスイッチングHubは現在でもやや高価である．

フロッピーディスクドライブ(FDD)やCD-ROMドライブは通常のPCには欠かせない．しかし，PCクラスタにおいて，FDDは緊急時に必要であるが常時必要ではない．また，CD-ROMドライブはOS等のインストール作業で必要となる場合もあるが，それ以外では全く必要ない．

メモリモジュールの種類にはいくつかあり，DRDRAMといった高速のメモリモジュールはかなり高価である．しかし，一般的なSDRAMとの性能差は大きくないと言われており，SDRAMを用いるのが現実的な選択である．メモリの総容量はクラスタの使用用途によって決定すべきである．計算時のスワップによるディスクアクセスは大幅な計算速度の低下を招く．その上，ディスクレスシステムでのスワップの処理はネットワーク経由で行わなくてはならず，さらに計算速度は低下する．よって，計算実行時の最大メモリ使用量やOSのカーネルが占有するメモリ使用量を考えて十分な容量のメモリを搭載すべきである．一方，最大メモリ容量はマザーボードとそのノースブリッジの種類によって決まる．一般に1GB以上のメモリ容量に対応するマザーボードは，一部のサーバ用途向けのみである．

2.2 ディスクレスシステムの構成方法

製作するPCクラスタではディスクレスシステムを用いる．Linux[1]の比較的新しいカーネルでは，NFSサーバ上のディスク領域をRootファイルシステムとしてマウントする機能をサポートしている．この機能を利用してディスクレスシステムが構築可能である．NICの設定に必要なIPアドレス等の情報は，一般にはRootファイルシステム下にある設定ファイルに書かれているため，NFSを介してディスク領域をマウントする前に，別の何らかの方法でIPアドレス等を設定しなければならない．Linuxでは，ブートストラッププロトコル(BOOTP)を用いて，BOOTPサーバに対し問い合わせを行い，NICのMACアドレスからIPアドレスとNFSサーバのIPアドレス等を得て設定するという方法に対応している．そこで問題となるのは，HDDを持たないディスクレスシステムでどのようにしてOSのカーネルイメージを読み込ませ，実行させるかということである．

まず，カーネルイメージを書き込んだフロッピーディスク(FD)とFDDを用いる方法が考えられる．この方法は完全なディスクレスシステムとは呼べないが，ディスクレスシステムの容易な構築方法として実際に用いられている．この方法の欠点として，クライアントがFDDを搭載していなければならないということがあげられる．さらにカーネルイメージに変更が加えられた場合に，全てのFDを書き直さなければならないという手間がある．また，FDの信頼性が比較的低いのも問題となる．

他の方法として，ブート用のROMをPC上に用意し，そこにプログラムを書き込む方法がある．通常，カーネルイメージのサイズは大きく，数Mbitの容量を持つROMを用意しなければならない．そこで，カーネルイメージをネットワーク経由で読み出すための簡単なプログラムをROMに書き込み，このプログラムがカーネルイメージを読み出すという手順が用いられる．この起動手順と通常のPCにおける起動手順との比較を図1に示す．通常のPCの起動手順では，マザーボード上のBIOSの処理が終わった後，HDD上のプログラムが実行される．Linuxではマスターブートレコード(MBR)に書かれたliloが実行され，liloがカーネルを読み出す．一方ブートROMを用いる方法では，BIOS処理の後にブートROM上のプログラムが実行される．そして，プ

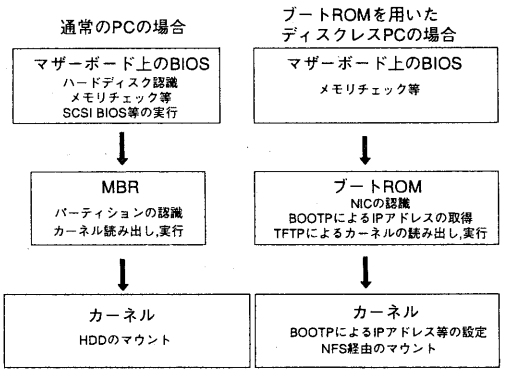

図1 起動手順の比較

ログラムがネットワーク経由でカーネルイメージを読み出す．カーネルイメージを読み出すための手段はいくつか考えられるが，FTPの簡易版であるTFTPを用いるのが一般的な方法である．この方法を用いると，TFTP自体もIPアドレスを必要とするため，BOOTPを用いてIPアドレス等を取得する．そのため，カーネル起動時を含めると2度BOOTPを使うことになる．この手法はネットワークコンピュータ等のディスクレスシステムの構成方法として用いられている方法である．この方法に対応した多くのNICが販売されており，それらのNICはROMを搭載するためのソケットをボード上に持っている．この方法には故障の可能性が低いことや，カーネルイメージをサーバ上で一元管理できる等の利点がある．

3. 製作

2.1で検討した結果，表1に示す構成部品を選択した．CPUには製作時期(1999年夏頃)に最も安価に入手できたAMD K6-2を選択し，メモリモジュールには一般的なSDRAMを用いた．また，ディスクレスシステムの構築方法について，前述のブートROMを用いる方法の方が利点が多く，より価格を抑えることができるため，製作するクラスタではこの方法を用いた．電源およびケースに用いたアルミ版やスイッチ，リレー等を含めた一台あたりの単価は約39,000円程度となった．

以上の構成部品を用いて，加工したアルミ板の上に5台のノードを取り付けたシステムを3セット製作した．マザーボードとアルミ版が接触しないようにプラスチック製のスペーサを挟ませた．各ノードにはそれぞれ電源を一個用意した，1個の電源を分岐して2ノードに供給する方法でも試作したが，定格外駆動になることと，ATX電源の電源管理用の回路を製作しなければならないため断念した．図2に1ノードの様子を示す．ブートサーバを除いてキーボードやマウスを取り付けていない．そのため，このままでは起動時にキーボードの検出エラーが発生し，起動しない．これを回避するため，BIOSの設定において検出エラーの無視を指定した．また，全15ノードの電源管理を一カ所で行うための電源スイッチパネルを製作した．完成したクラスタ全体の様子を図3に示す．1セットの大きさは，幅860mm高さ307mm奥行き400mmとなった．マザーボードであるSY-5EMMは，マイクロATX規格と呼ばれる比較的小型のマザーボードである．この小型のマザーボードを用いることで，本棚に収まる大きさのコンパクトなシステムが構築できた．15ノードのうち1ノードはブートサーバとしHDDを搭載させ，メモリ容量を128MBに増強した．OSにはLinux Kernel 2.2.14を用い，ブートROMに書き込むプログラムであるブートローダにはEtherboot 4.2.10[2]を使用した．プログラムのサイズは16KBとなり，ROMライタを用いてROMにプログラムを書き込みを行った．また，ブートサーバ上で1クライアントあたりに必要となるディスク領域を削減するために，ディレクトリの共通化を行った．ログファ

表1 製作したクラスタの主な構成部品

CPU	AMD K6-2 400MHz L1Cache 64KB
マザーボード	SOYO SY-5EMM L2Cache 512KB
メモリ	64MB PC100 SDRAM(8ns)
NIC	Realtek RTL8139A
ROM	Hitachi HN27C256AG 256Kbit
Hub	Planex FX-16N

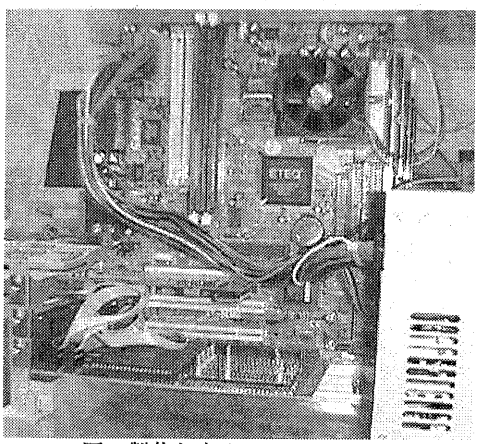

図2 製作したクラスタの1ノード

イル，ロックファイル等が存在するディレクトリや，ノードごとに異なる情報が存在するディレクトリ等を除いて共有させることが可能である．この作業により，共有領域は480MBとなり，個別に必要なディスク領域は1クライアントにつき40MB程度に抑えることができた．

製作当初，このクラスタは計算目的では無く，多数のノードを必要とするネットワーク実験を目的として製作した．そのための図2に見られるようなシリアルポート増設カードを搭載している．このシリアルポート同士をPPPを介して接続することによって，様々なトポロジーのIPネットワークが実現可能である．

4. 性能評価

製作したクラスタの簡単な性能評価を行う．浮動小数点演算性能の評価には並列版LINPACK[3]を用いる．整数演算性能の比較にはNAS Parallel Benchmarks2.3[4]内のISを用いる．並列計算用ライブラリとしてMPIの実装の一つであるMPICH-1.2.0[5]を用いている．

4.1 LINPACKによる性能評価

LINPACKはガウスの消去法を用いて連立一次方程式を解くプログラムである．計算量やオーバーヘッドに対する配慮がなされており，浮動小数点演算性能の指標としてよく用いられている．連立方程式の係数行列は，1000x1000のサイズを用いる．

1ノードで計算した性能を表2に示す．比較のためにUltra5(UltraSparcIIi 270MHz, 64MB)を用いて計算した性能も示す．CPUのアーキテクチャや環境が異なるため一概には比較できないが，動作周波数でかなりの差があるにもかかわらず，性能ではUltraSparcIIiが上回っている．これはK6-2の浮動小数点演算性能の低さが表れているためと思われる．

次に並列に計算した場合の測定結果を図4に示す．製作したクラスタは1ノード時では劣っていたにもかかわらず4ノード付近で性能が逆転し，14ノード時には10%程度高速になっている．この原因はいくつか考えられる．Hubのスイッチング

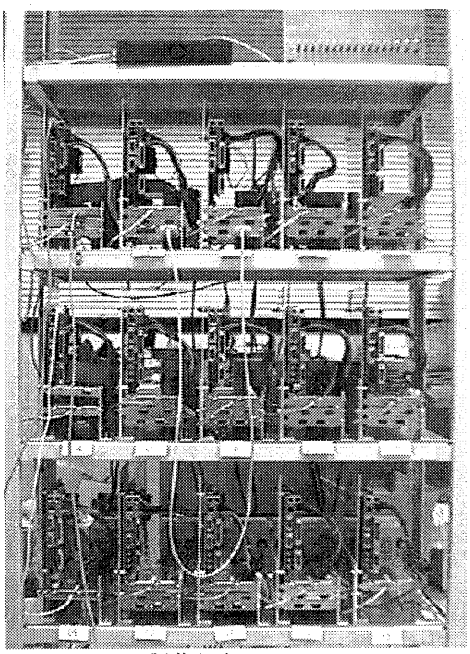

図3 製作したPCクラスタ

表2 LINPACKにおける単体での性能

	製作したクラスタ	Ultra5
CPU	K6-2 400MHz	UltraSparcIIi 270MHz
Performance (Mflops)	56.72	61.02

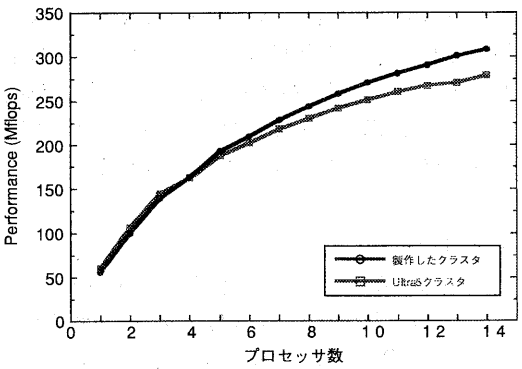

図4 LINPACKにおける計算性能

速度やNICの違いは，通信量が多い場合にしばしば大きな要素になる．メモリ性能に関しても，Ultra5にはやや低速なメモリが用いられている．また，並列化が進み浮動小数点演算が分散されるに従って，並列化処理に関係する整数演算の性能差が表面化した可能性もある．

4.2 ISによる性能評価

ISは大規模な整数ソートを行うプログラムで，整数演算性能を示す指標である．問題のサイズとしてCLASS-Wを用いる．

図5に測定結果を示す．結果から製作したクラスタの方が20%程度高速であることが解る．整数演算性能に関しては製作したクラスタの方が優

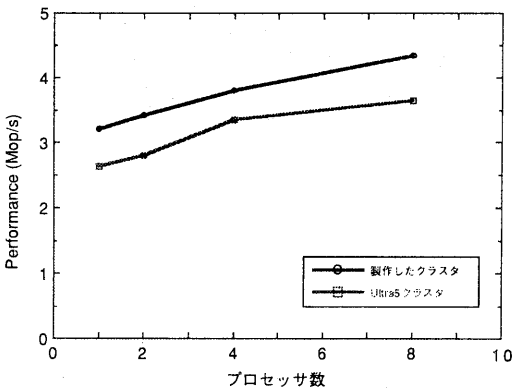

図5 ISにおける計算性能

れていると言える．

以上の測定結果から，メモリ性能や整数演算性能を含めた総合的な面で，市販のWSと同等の性能を持っていると判断できる．また，コストパフォーマンスで比較すると，Ultra5の標準価格は1台50万円以上であり，製作したPCクラスタのコストパフォーマンスは10倍以上優れている．

5. おわりに

製作に着手した1999年夏ごろ，500ドルPCと呼ばれる安価なPCや，条件付きながら200ドルという価格のPCが売り出されていた．そこで，これらの安価なPCを用いてPCクラスタを製作することにした．その結果コンパクトなPCクラスタを安価に構築できた．今回行ったベンチマークは限定された状況による測定ではあるが，市販のWSに迫る能力を持っていることが解った．それでもマルチプロセッサ構成等を用いたクラスタに比べて絶対的な計算性能で高いとは言えない．また，PC用のCPUの浮動小数点演算性能は他のWS用のCPUに対して大きく劣ると言われている．しかし，コストパフォーマンスでは優れている．計算性能をあまり必要としない，ネットワーク実験用としては最適といえる．また，並列ソフトウェアの開発環境を安価に構築したい場合や，整数演算性能やメモリ，HDDの性能が大きく影響するWWWサーバやデータベースといった用途にも用いることができるだろう．クラスタにローエンドPCを用いることは，コストパフォーマンスを追求する選択肢として，十分に有効である．

参考文献

1) http://www.linux.org/
2) http://www.slug.org.au/etherboot/
3) http://www.netlib.org/lapack/index.html
4) http://www.nas.nasa.gov/Software/NPB/
5) http://www-unix.mcs.anl.gov/mpi/mpich/