

Xeon による PC Cluster の構築と運用・利用について

赤井光治, 久長穰, 刈谷丈治, 浜本義彦*

山口大学 メディア基盤センター, * 工学部

宇部市常盤台 2-16-1, 755-8611. E-mail: akaikoji@po.cc.yamaguchi-u.ac.jp

Construction and Management of a Xeon Clustering System

Koji Akai, Yutaka Hisanaga, Joji Kariya, Yoshihiko Hamamoto*

Media and Information Technology Center, *Faculty of Engineering,
Yamaguchi University, Tokiwadai 2-16-1, Ube, 755-8611.

概要

今年 3 月に山口大学メディア基盤センターに Xeon クラスター計算機が導入された。計算ノードが 64 台、128 個のプロセッサを持つ。ノード数が比較的多いことから Gigabit ネットワークに比べ高速な Myrinet により相互接続されている。このシステムの概要および性能評価の結果について述べる。

keywords:

クラスター計算機、Xeon、Myrinet2000、SCore

1 はじめに

近年パソコンの性能が急速に向上しており、1 CPU あたり数 GFLOPS の性能を持つようになってきている。このようなパソコンを多数、高速なネットワークでつなぐによりスーパーコンピュータ並みの計算性能を持つ計算機が安価に構築することができ、10 年ほど前からこのような並列計算機 (クラスター計算機) が注目されている。

山口大学メディア基盤センターでは計算サーバーとして前回の機器更新以後、ベクトルプロセッサを持つ NEC SX5s(2CPU) および並列プロセッサ SUN Enterprise 4500(8CPU) をメインサーバーとしてサービスを提供してきた。SX5s はメインメモリー 8GB、2CPU 構成で理論性能が 8GFLOPS である。Enterprise 4500 はメインメモリーが 8GB でありクロック数は 400MHz である。導入当時これらの計算機はパソコンに比べ遙かに高速であったが、既に並列計算機の 1 プロセッサあたりの性能は Pentium4 クラスの計算機に全く及ばなく



図 1: クラスター計算機外観。

なっている。しかし計算機が研究に及ぼす効果は益々高まっている。実験の条件出しや材料の性能を事前に見積もり、研究の方向付けを行うなど、複雑で自由度の大きな対象を扱うために勘と経験だけでは対応できなくなり、計算機の役割が重要になってきているためである。

そのような状況のもと、大学の研究・教育の中核的な役割を担うシステムの一つとして、今年3月に Xeon クラスター計算機がメディア基盤センターに導入された。クラスターの規模は64ノードと中規模であるが、super computer の世界ランキングを示す TOP500[1] にランクイン可能な性能を持つ。本予稿では Xeon クラスターの機器構成、性能および設定状況の紹介を行う。また、今後このようなクラスターリングシステムが増えることが期待されることから、どのようなシステムを構成すべきかについて話題を提供したい。

2 機器構成

クラスター計算機は 1) ファイルサーバー、2) クラスター管理サーバー、3) 計算ノード、4) ネットワーク、5) UPS で構成されている。ファイルサーバー、計算ノードの基本構成は以下の通りである。

- Xeon dual cpu 2.4GHz
- memory: DDR266 DIMM 512MBx4
- Ethernet ポート: 100Mbpsx1, 1Gbpsx1
- 高速ネットワーク: Myrinet 2000

ファイルサーバーは RAID カード (12 ポート) を 1 機搭載しハードディスクは 250GBx12 の構成になっている。本システムではファイルシステムへのアクセスを高速にし、ディスク障害に対応するため RAID 1+0 を採用し、ディスク容量 1.5TB を確保している。各計算ノードは上記ハード構成に加え、40GB の HD を供え各計算ノードでのスクラッチファイルを提供し、ファイルサーバーへアクセスが分散できるように設計されている。筐体については、1U タイプのクラスターも見受けられるようになってきたが、メモリー及び HD の熱対策のために 2U のラックマウント型筐体を選択した。クラスター管理サーバーは DNS、NIS、DHCP の各種サーバー機能

を持ち、計算部と管理システムの機能をできるだけ分離するように設計した。

図 2 はクラスター計算機のネットワーク構成の概念図を示している。クラスター計算機は学内 LAN とは別にサブネットワークを構成しており、ファイルサーバーおよび管理サーバーのみが学内 LAN との接点を持っている。クラスター内 LAN は管理用の 100base ネットワークと NFS 用 Gigabit ネットワークで構成される予定であるが、現時点では 100base ネットワークのみが利用でき NFS も兼用している。これらの ethernet ポートはオンボードになっており、Gigabit の network チップセットは Intel Pro/1000 MX 系である。このチップは同じ Intel の XT 系に較べ性能的に劣っており、今後どの程度の性能を示すかテストする必要がある。MX 系では理論性能の 3 割程度しか性能がでないとの情報もある。

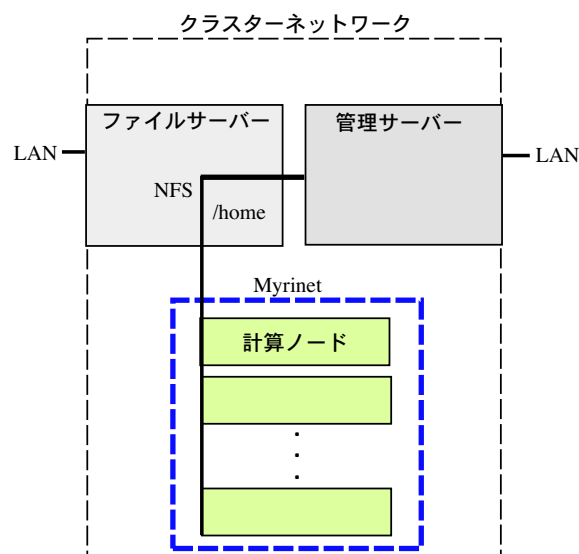


図 2: クラスターのネットワーク概念図.

Myrinet2000 は MyriCom 社が制作・販売している高速ネットワークであり、全 2 重通信で理論性能値は双方向 4Gbps とされている。クラスター計算機では計算ノードが増加するにつれ、ノード間通信が増加するため高速なネットワークを必要とする。更に、Myrinet2000 は Ethernet に較べレイテンシーが小さいことも特徴の一つである。このネットワーク性能を生かすためには Pentium4 系のチップセットが提供する 32bitx33MHz の I/O インターフェイスでは不十分であり、64bitx66MHz 以上の PCI-X バスを供えたチップセットが要求される。

導入時点でこれを満たすチップセットは Xeon dual cpu 用 E7500 のみであったため、Xeon が選択された。CPU-メインメモリー間のバス幅およびメモリーの性能の点では Pentium4+RDRAM の方が CPU あたりの性能は高い。図 2 に示されるように Myrinet ネットワークは計算ノード間で構築されており、各計算ノードが Myrinet エンクロージャーを経由して optical ケーブルで物理的に相互接続されている。これらのファイルサーバー、計算ノード、エンクロージャーは全て UPS に接続されており、電源の障害時に対応できるようになっている。また UPS を含め全ての機器が 6 本のラックにマウントされ、コンパクトなシステムとして構成されている。(図 1)

3 セットアップ

本クラスター計算機は SCore5.4.0+RedHat7.3 の構成でセットアップを行った。SCore については書籍 [2] およびホームページ [3] があり、詳しい情報が提供されている。SCore の設定等の詳細については参考文献を参照して頂くとして、ここでは主な特徴について説明する。

クラスター計算機の種類や構築法は様々であるが、最も良く耳にするのはベオウルフシステムである。MPI, PVM, Gloubus, LFS 等のクラスターリングツールをインストールし構築される。[4] 今回クラスター計算機のみドルウェアに SCore を選択した主な理由は 1) インストールが容易であること、2) ジョブのスケジューリングが優れていることの 2 点である。

今回のクラスター計算機は主に機器の購入のみであったため、ソフトウェアを含むクラスターの設定はメディア基盤側で行う必要があった。このため、ジョブ管理サーバー 1 台のみに OS および SCore をインストールすれば、残り数十台の計算ノードについては簡易インストーラーを使いほぼ自動的にインストールできる SCore が魅力的であった。SCore5.4.0 では計算ノードには RedHat7.3 がインストールされるようになっている。このとき、計算ノードには必要なパッケージのみがインストールされるよう設計されており、通常の Linux ディストリビューションで見られるように多くの不要なアプリケーションがインストールされ、数 GB のディスクスペースが浪費されることがない。今回のインストー

ルでは 500MB 程度のディスク容量で済み、節約できたスペースをチェックポイントファイルなど SCore の運用に必要な領域に割り当てた。このため利用者のワークスペースを少しでも多く確保できるうえ、不必要なデーモンの起動が容易に回避できる。また自動インストールにより、クラスター構築に必要な MPI, PVM, PBS 等も同時にインストールおよび設定が行われ、SCore のインストールだけで最低限利用できる環境が構築された。

MPI 等により投入されたジョブには必要な数のプロセッサが割り当てられ実行される。このとき複数のジョブがクラスター計算機で実行されていると、一般的なシステムでは各プロセッサで実行されるプロセスは非同期である。つまり同じジョブであっても、各プロセッサで同時に実行状態にあるとは限らない。このためプロセス間で通信が発生した場合、一方のプロセスは他方が実行状態になるまで待たされることになる。MPI 等の通信ライブラリーには同期型と非同期型の関数が用意されているため、非同期状態でも動作するようにプログラムは作成できるが、プログラマーはそのことを意識しコードを開発する必要がある。一方、SCore では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機能を持ち、個々のプロセスが同期を保ちつつ実行される。メディア基盤センターのように利用者のプログラミング能力が様々であり、全ての利用者に対応することが求められるシステムでは、この機能は利用者のプログラム開発支援の点から意味がある。また、プログラムの実行性能の点からも、受信側と送信側間で待ち時間が発生しにくいため、計算時間の短縮が期待できる。

4 ベンチマーク

Top500 で対象とされる Linpack に対する正確な性能評価はまだ行っていない。しかし、Basic Linear Algebra Subprogram ライブラリーの一種である ATLAS や Intel Math Kernal Library を用いた場合、単一ノードでの性能であるが 3GFLOPS 以上の性能を確認している。このことから、並列化に最適化されたコードでは 300GFLOPS 以上の性能が期待できる。実際に TOP500 にランクされているシステムの中で本クラスター計算機と同じ構成のシステムが 200 番台後半にランクされており、400GFLOPS 程度の最大性能が示されている。以

下ではクラスター計算機の性能を評価するため、幾つかのコードおよびアプリケーションについてベンチマークテストを行ったのでその結果について述べる。

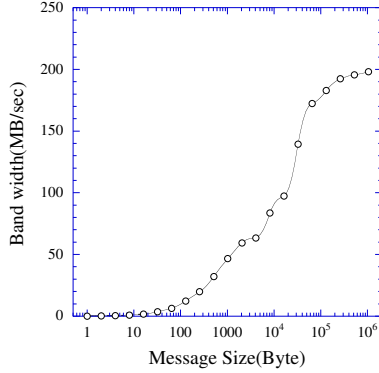


図 3: myrinet のバンド幅.

Myrinet のドライバーは Myricom 社がドライバーを提供しており、GM と呼ばれている。SCore では GM とは別に通信用ドライバー (PM) が提供されており、通常 PM ドライバーが使われる。図 3 は PM ドライバーを用いた Myrinet のバンド幅を示している。バンド幅のベンチマークは MPICH のパッケージに付属している、"systest.c" が用いられた。メッセージのサイズが 300KB 程度で転送速度がほぼ飽和し、200MB/sec 程度に達している。通常の Gigabit ネットワークに較べると高速なバンド幅を持っていることが分かる。理論性能が 250MB/sec であるためほぼ 80% の性能である。なお文献値では PM ドライバーを用いて 227MB/sec のバンド幅が報告されている。[2] 一方、GM ドライバーではバージョン 1.6.4 において 240MB/sec のバンド幅が達成されている。[5] このため、高い転送速度は持つがまだ改善の余地があると考えている。

次に、少数ノード数でのベンチマーク結果を示す。ベンチマークに用いるプログラムは Laplace 方程式を解く MPI コードを用いた。[6] このコードはノード内の計算量が多くネットワークよりむしろ CPU に負荷がかかるため、単体性能を調べるのに適当なコードである。Xeon クラスター性能をより詳しく評価するために Pentium4 クラスターのベンチマーク結果を並べて示す。表 1 は 1 つあるいは 2 つのノード数 (n) に対しプロセッサ数 (p) を変化させた場合の計算時間を示す。"Xeon" が今回導入されたクラスター計算機であり、"Pen4" は別

表 1: 計算時間の比較

プロセッサ	1n · 1p	2n · 2p	1n · 2p
Xeon	115 s	58 s	124 s
Pen4	73 s	37 s	—

の Pentium4 クラスターを表している。Pen4 クラスターのスペックは cpu: 2.53GHz, RIMM1066(1GB 搭載) である。なお、Pen4 クラスターのノード間は Gigabit 接続となっている。この結果が示すように単体では Pen4 計算機が Xeon 計算機より 1.5 倍程度高速である。クロック数の違いでは全く説明できず、メモリーと CPU 間のバス幅の違いによると理解される。この傾向は CPU 1 つのみで実行した場合と dual CPU で実行した場合とで顕著に現れている。ベンチマークからは dual の効果は見られずバス幅が計算速度を律則していることを示しており、dual にすることによりスレッドを増やすコスト分計算時間が遅くなっているのが分かる。ノード数が少ないためスケラビリティは Xeon および Pen4 共に良く成り立っており、Gigabit でも対応できるほど通信量が少ない計算と言える。

最後に実用性の点から、実際にクラスター計算機が用いられるアプリケーションに対して多数の計算ノードを用いたベンチマークの結果を示す。今回は第一原理電子構造計算ソフトを用いた。[7]

電子構造計算では電子の状態を数値計算により求めることが目的であり、個々の電子状態が電子の空間分布を表す関数で表現される。電子状態を表す関数は 2 つのパラメータ ($n, \mathbf{k}$) により特徴づけられ、ハミルトン作用素 $H(\mathbf{k}, \mathbf{r})$ を用いて $H(\mathbf{k}, \mathbf{r})\Psi_{n\mathbf{k}}(\mathbf{r}) = \varepsilon_{n\mathbf{k}}\Psi_{n\mathbf{k}}(\mathbf{r})$ と表せる。このとき n は整数であり、 $\mathbf{k}$ は 3 次元の実ベクトルである。この計算の特徴は並列計算に適したアルゴリズムを持っていることである。つまり、電子状態を求める方程式がパラメータ $\mathbf{k}$ ごとに分割され、系全体の状態を求めるためにそれぞれの $\mathbf{k}$ に対して状態を求める問題に置き換えられる。更に、電子状態 $\Psi_{n\mathbf{k}}(\mathbf{r})$ を適当な独立関数で展開すれば、結局電子状態を求める計算は行列の固有値問題に帰着し、計算時間の大部分を占めている行列の対角化を並列に実行することになる。ただしハミルトン行列は求めるべき全電子密度 $\sum_n \int d\mathbf{k} |\Psi_{n\mathbf{k}}(\mathbf{r})|^2$ に依存しているため、 $\mathbf{k}$ についての

積分を実行する必要がある。数値計算としては電子密度を求める積分は和で近似され、 k のサンプル数を増やすことにより精度を高めることができる。今回のベンチマークでは、 k を 20 点取った。20 プロセッサと 1 プ

表 2: プロセッサ数に対する計算時間の比較

$1n \cdot 1p$	$1n \cdot 2p$	$20n \cdot 20p$
45630s	34897s	2930s

ロセッサの速度比は 15.6 であり、かなり大きな粒度を持つコードであることを反映して高いスケーラビリティが得られていることが分かる。しかし、計算時間の 90%以上が独立したプロセッサ上で計算が行われていることを考えると、更に高いスケーラビリティが達成されるべきである。スケーラビリティを低下させている要因はネットワークの送信速度であると考えている。計算では数 GB のファイルが生成され、全てファイルサーバーに保存している。今回テストした環境では NFS に 100base のネットワークを使っており、今後 Gigabit に置き換えればある程度向上することが期待できる。最も有効な改善方法は各ノードのローカルディスク領域を用いることである。今回のベンチマークではこの場合のテストは実施できていない。

電子状態計算のベンチマークでも Laplace 方程式のベンチマーク同様、1 ノード 2 プロセッサでは性能が出ないことがわかる。このベンチマークでは 2 プロセッサと 1 プロセッサの速度比は 1.37 であり、ある程度 2 プロセッサの効果は見られている。Xeon の E7500 チップセットでは 3.2GB/s のバスを 2 プロセッサで共有していることから、2 プロセッサで性能を出すためにはまずバス幅を確保することが重要と言える。更に Xeon dual processor クラスターで性能を出すためには 512KB の L2 キャッシュのヒット率を高めるような、かなり踏み込んだコーディングの最適化を必要とするように見える。

5 終わりに

メディア基盤センタに導入された Xeon クラスター計算機の構成・設定およびベンチマーク結果を紹介した。

ベンチマークの結果、Myrinet による通信は 200MB/s のバンド幅を実現し、クラスター計算機として高い性能を有していることは確認できた。一方、Xeon のメモリー-CPU 間のバンド幅の制約から、2cpu 構成で十分な性能を得ることが難しいことが確認された。今後のクラスター計算機を構築する上で考慮すべき点である。

謝辞

今回の予稿に示された Xeon クラスターの性能評価において、Myrinet の通信性能および Laplace 方程式の MPI コードによるベンチマークは山口大学 河野貴久氏ならびに名古屋工業大学 尾形修司氏により行われたものです。ベンチマークの実施を快く引き受け、結果を提示して頂いたことに感謝します。

参考文献

- [1] 現在のランキングは TOP500 SUPERCOMPUTER SITES: <http://www.top500.org/>で確認できる。
- [2] 石川裕、佐藤三久、堀敦史、住元慎司、原田浩、高橋俊行、"Linux で並列処理しよう-Score で作るスーパーコンピュータ", 共立出版 (2002).
- [3] <http://www.pccluster.org>.
- [4] クラスター計算機の構築に関する参考文献は多数存在するが、例えば T.L. Sterling, J. Salmon, D.J. Becker, D.F. Savarses, "How to Built a Beowulf: A Guide to the Implementation and Application of PC Clusters." MIT Press(1999); (日本語訳), 北野宏明監訳, 奥乃博、諸橋峰雄、京田耕司、中壺一博訳, "PC クラスタ構築法 Linux によるベオウルフ・システム", 産業図書 (2001) がある。
- [5] <http://www.myricom.com/>
- [6] W. Gropp, E. Lusk, and A. Skjellum, "Using MPI", MIT Press (1994).
- [7] P. Blaha, K. Schwarz, G. Madsen, D. Kvasnicka and J. Luitz, program package WIEN2k, Technical University of Vienna (2001).