2009-SLDM-138 (14)
2009 / 1 / 29

テスト用 Linux クラスタシステムの試作

北野 皓一* 寺本晃司** 堀田 忠義*
*職業能力開発総合大学校 〒229-1196 神奈川県相模原市橋本台 4-1-1
**雇用・能力開発機構 〒231-8333 神奈川県横浜市中区桜木町 1-1-8

あらまし 商用のクラスタシステムは，高価だが OS や開発環境のバージョンが古いことが多く，また高価ゆえに他の研究室と共用での運用となるため，OS や開発環境のバージョンアップの実施は困難である．さらに，多ノードのシステムは広い物理スペースを占有し，かつ多大な電力が必要なため，特別な電源工事や冷房設備工事などが必要となる．

本研究では，玄人志向社製の複数台の「玄箱」から成るクラスタシステムを提案する．このクラスタシステムは，計算能力の観点からは低スペックではあるが，多ノードのシステムを安価で低消費電力，かつ小さい物理サイズで構築できる．これを，新しいバージョンの OS や並列計算環境のテストを含めた並列化プログラムのテスト用のシステムとして提案する．

キーワード クラスタシステム，MPI，玄箱，姫野ベンチマーク

abstract Most computer cluster, which is produced by a company, is so expensive compared to a PC, and the versions of its operating system and parallel programming development tools are old. But in most cases, a user cannot install and test the higher version OSs and tools, because most cluster is shared by multiple laboratories or sections inside universities, institutes, or companies, because of its high cost. In addition, a cluster which consists of a lot of computers repuires huge physical spaces, a special construction of the electric power, and suitable air conditioners, and these also repuires additional expensive costs.

In this paper, the computer cluster, which consists of multiple KURO-BOXes of Kuroutoshikou Inc., called "KURO-BOX cluster", is proposed. Its performance is very low, but the cluster is very low cost, very low electric power, and very small physical size, comared to most cluster by a company. We propose this cluster as a tester of parallel computer programs, including operating systems and parallel programming development tools.

Keyword Computer cluster, MPI, KURO-BOX, Himeno benchmark

1. はじめに

クラスタシステムとは複数のコンピュータを用いて処理を行う分散メモリ型の並列計算機の 1 形態であり，1 台で処理を行うよりも高速に処理できるという利点がある．特に PC クラスタシステムは，そのコストパフォーマンスの良さから，WEB 検索エンジン，データマイニング，ゲノム情報処理，高度医療情報処理，自然現象シミュレーションなど，巨大な計算能力を必要とする様々な分野で広く利用されている．またクラスタシステムを扱う企業も既に多く存在する．そのような企業によって構築された，いわゆる商用のクラスタシステムは，安定性を重視するために，OS や開発環境のバージョンが古いことが多い．また高性能なものは非常に高額であるため，個人や研究室のような少人数のグループ専用として運用することは難しい．従って，リース品として納入されることも多く，かつ他の研究室や部署と共用での運用となるため，OS や開発環境のバージョンアップにより性能が向上したり，新しい機能が使用可能になることが期待できるにも関わらず，そのようなバージョンアップ作業の実施およびそのテストは非常に困難である．加えて，高性能かつ多ノードのシステムは広い物理スペースを占有し，かつ多大な電力が必要なため，特別な電源工事や冷房設備工事などが必要となる．

本研究では，玄人志向社製の複数台の「玄箱」から成るクラスタシステムを提案する．このクラスタシステムは，計算能力の観点からは低スペックあるが，多ノードのシステムを安価，低消費電力，かつ小さい物理サイズで構築できる．これを，新しいバージョンの OS や並列計算環境のテストを含めた並列化プログラムのテスト用のシステムとして提案する．OS に Fedora Core 6，並列計算環境に openMPI を選択した場合の，システムの構築方法，消費電力，および姫野ベンチマークの実行結果を示す．

2. 構成要素について

2.1 玄箱

玄箱とは，玄人志向社より発売されている NAS（Network Attached Storages）組み立てキッドである．内部に CPU，メモリ，ネットワークインタフェース，USB ポートを持ち，Linux マシンとして扱うことができる．また，Linux マシンとして扱えるがゆえに幅広い拡張性を持ち，本来の用途であるファイルサーバとして利用するだけでなく，Web サーバ，メールサーバ，プリンタサーバ，DNS サーバとしても利用可能である．

玄箱には 3 種類があり，それぞれ CPU の性能，メモリの容量，ネットワークインタフェースについて違いがある．本研究では，表 1 に示すスペックを持つ KURO-BOX/HG を使用する．

表 1：KURO-BOX/HG のスペック

CPU	Power PC 266MHz
メモリ	128MB
記録媒体	3.5'HDD
物理サイズ	60x173.5x185mm
ネットワークインタフェース	1000Base-T/100Base-T/10Base-T
価格（HDD なし）	11970 円（amazon.co.jp にて，2008/12/10 現在）

低スペック，小サイズ，低消費電力，かつ低価格な Linux クラスタシステムを構築しようとする場合，KURO-BOX/HG による以外にも，低価格なノート PC による方法や，他の NAS デバイスによる方法，あるいは，組み込み Linux 用のデバイスによる方法，などが考えられる．しかし，それらの大半は KURO-BOX/HG と比較して高額

であり，ネットワークインタフェースの速度が 100Mbps 以下であり，かつ消費電力が大きい．加えて KURO-BOX/HG においては，出荷時にプレインストールされている Linux の他に，比較的簡単な設定作業によって，PowerPC 版の Fedora が動作する．以下に挙げる理由により，この Fedora が実行可能という意義は大きいと思われる．

Fedora は，他の Linux ディストリビューションの中では，最新の技術を取り入れるスピードが早く，その為の OS のバージョンアップも早いペースで頻繁に行われている．また，そのような OS であるため，MPI を含めた並列計算環境の移植およびバージョンアップも頻繁に行われており，トラブル解決された上で実行形式にコンパイルされたそれらのファイルパッケージ(rpm)がネット上に配布されているため，Fedora がインストールされた KURO-BOX/HG への並列計算環境のインストールおよびバージョンアップは容易である．さらに Fedora は，多くの開発者に注目されている OS であるため，web 上あるいは書籍として問題解決のドキュメントが多く提示されている．前述の他のデバイスでは Linux は動作するとしても，Fedora が動作しないか，動作させるために多大な努力が必要なものがほとんどである．ただし，KURO-BOX/HG においては，RAM の増設ができないことや，CPU の交換ができないことなど，拡張性に乏しいことが欠点として挙げられる．

2.2 MPI

MPI とは，Message Passing Interface の略称で，分散メモリ型の並列計算機を利用し，複数のプロセス間でデータ，メッセージの通信を行う並列計算用通信ライブラリである．MPI は専用の関数を C，C++，Fortran などのプログラミング言語を用いて利用する．本研究では，C 言語を利用した openMPI version 1.1-7 と呼ばれる MPI パッケージを利用する．

3. クラスタの構築方法

Fedora は，http://www.shinkr-webpj.jp/ を参照してインストールする．その後，以下の rpm パッケージ群を，サーバおよび計算ホストそれぞれの場合についてインストールする．

- ssh
  - openssh-4.3p2-10
  - openssh-server-4.3p2-10
  - openssh-clients-4.3p2-10
- NIS
  - ypbind-1.19-5
- NFS
  - nfs-utils-1.0.9-8.fc6
  - nfs-utils-lib-1.0.8-7.2
  - nfs-utils-lib-devel-1.0.8-7.2
- MPI
  - openmpi-libs-1.1-7.fc6
  - openmpi-devel-1.1-7.fc6
  - openmpi-1.1-7.fc6
- rsh
  - rsh-0.17-36.ppc.rpm
  - rsh-server-0.17-36.ppc.rpm
- gcc
  - gcc-4.1.1-30
  - libgcc-4.1.1-30
- yum 関連
  - libxml2-2.6.26-2.1.1.ppc.rpm
  - rpm-python-4.4.2-32.ppc.rpm
  - nmap-4.11-1.1.ppc.rpm
  - python-elementtree-1.2.6-5.ppc.rpm
  - python-sqlite-1.1.7-1.2.1.ppc.rpm
  - python-urlgrabber-2.9.9-2.noarch.rpm
  - yum-3.0-6.noarch.rpm
  - yum-metadata-parser-1.0-8.fc6.ppc.rpm

これらのパッケージ群は，(http://download.fedora.redhat.com/pub/fedora/linux/core/6/ppc/os/Fedora/RPMS/)より入手可能である．

上記のパッケージ群をまず 1 台にインストールし，その記憶媒体をコピー元として，Century 社製の"これ do 台"を用いて別の記憶媒体にコピーし，別の玄箱に接続した上でネットワーク設定等を行うことを繰り返すことにより，計 16 台からなる玄箱クラスタを構築した．玄箱クラスタの構成図を図 1 に示す．

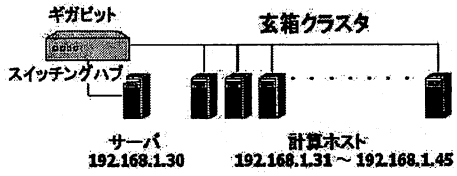

図 1：玄箱クラスタの構成図

16 台の玄箱が，24 ポートのギガビットスイッチングハブを経由して接続されている．IP アドレスについては，192.168.1.30～192.168.1.45 をそれぞれに割り当てている．サーバには 192.168.1.30 が割り当てられており，かつサーバは計算ホストの 1 台としても機能する．また，玄箱クラスタの外観を図 2 に示す．

図 2：玄箱クラスタの外観

4. 性能評価

試作したクラスタの処理速度を測定するために，C 言語と MPI ベースの姫野ベンチマークを使用する．姫野ベンチマークとは，ポアッソン方程式解法をヤコビの反復法で解く場合に主要なループの処理速度を測るベンチマークプログラムであり，商用 PC クラスタの性能評価などの目的で広く使用されている．比較のために，表 2 のスペックを持つ PC4 台から成る PC クラスタのデータも示す．

表 2：PC1 台のスペック

CPU	Athlon X2 4200+(2.2GHz)
メモリ	1GB
OS	Fedora Core 6
C コンパイラ	gcc-4.1.1-30
MPI	openmpi-1.1-7.fc6

図 3 に，PC と玄箱のそれぞれのクラスタにおけるプロセス数と処理速度の関係を示す．

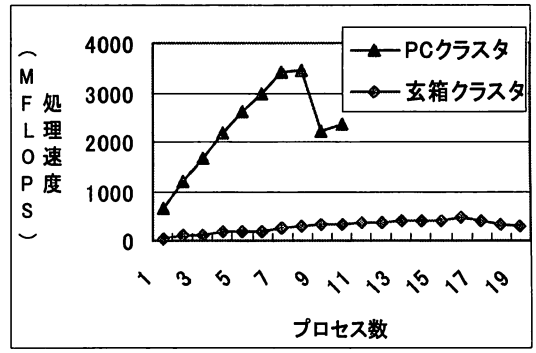

図 3：PC クラスタと玄箱クラスタのプロセス数と処理速度の関係

図 3 において，"PC クラスタ"は，姫野ベンチマークのデータサイズを M とした場合のデータである．"玄箱クラスタ"は同サイズを S とした場合のデータである．

図 3 から，両クラスタにおいて，CPU 数以上のプロセス数で実行すると処理速度が落ちている．また，PC クラスタに比べて，玄箱クラスタの処理速度は非常に低スペックであり，玄箱クラスタでのプロセス数 16 台のスペックが，PC1 台のそれとほぼ等しい．図 4 に，サイズが M および S の場合の

玄箱クラスタのプロセス数と処理速度の関係を示す.

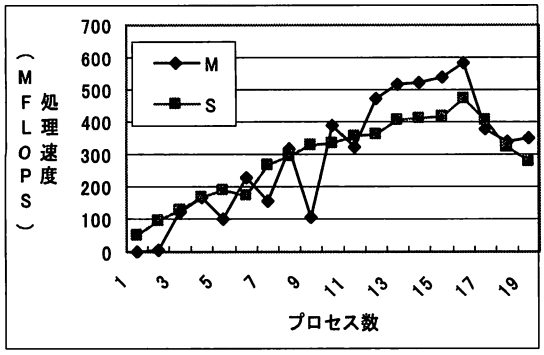

図 4：玄箱クラスタの処理速度とプロセス数の関係

図 4 から, サイズが M の場合のデータは, S の場合に比べて, 処理速度が高い場合も見られるが, プロセス数の増加に対する処理速度の増加の傾向が堅調でなくなっている. これは, 玄箱の RAM 容量不足と考えられる.

それぞれのクラスタの各計算ホスト 1 台における消費電力を, アイドル時と姫野ベンチマーク実行時の場合について, システムアートウェア社の SHW3A を用いて測定した. なお, 玄箱で CF カードを使用する場合に用いた IF デバイスは, MTG 社製の MTG4617 である. その結果を表 3 に示す. 1 ノード当たりで見れば, 玄箱クラスタは PC クラスタに比べて, 非常に低消費電力であることが分かる.

表 3：消費電力の比較

	アイドル時 (Watt.)	実行時 (Watt.)
玄箱(HDD)	14.8	16.6
玄箱 (CF カード)	9.7	11.4
PC	62.9	116

5. まとめ

新しいバージョンの OS や並列計算環境のテストを含めた並列化プログラムのテスト用のクラスタシステムとして, 玄箱クラスタを提案した. このクラスタシステムは, 計算能力の観点からは低スペックではあるが, 多ノードのシステムを安価, 低消費電力, かつ小さい物理サイズで構築できることを示した. また, OS に Fedora Core 6, 並列計算環境に openMPI を選択した場合の姫野ベンチマークの実行結果を示した.

参考文献

- できる！玄箱 Fedora 化！！(F-7 対応), http://www.shinkr-webpj.jp/main.html
- 姫野ベンチマーク紹介ページ, http://w3cic.riken.go.jp/HRC/HimenoBMT/index.html