2004 年 8 月 6 日

サイエンティフィック・システム研究会
ｻｲｴﾝﾃｨﾌｨｯｸ・ｺﾝﾋﾟｭｰﾃｨﾝｸﾞ・ﾌｫｰﾗﾑ(科学技術計算分科会主催)資料

理研スーパー・コンバインド・クラスタ（RSCC）の紹介と運用事例

独立行政法人理化学研究所　情報基盤センター
重谷　隆之

理化学研究所では、今年 3 月に 2048CPU の Linux PC（Intel Xeon 3.07GHz, Dual CPU）を中核としたスーパーコンピュータシステム（理研スーパー・コンバインド・クラスタ：RSCC）を導入した。日本の計算機センターでは Linux クラスタの採用は初めてのことである。RSCC にはいくつかの新しい技術を導入している。本発表では、センターの計算機として Linux クラスタを採用した背景、システム設計における新しい試み、実際の運用形態と利用状況などを紹介する。

１． はじめに

理化学研究所・情報基盤センターでは、理研の研究支援を目的としてスーパーコンピュータ（スパコン）の導入・運用を行っている。これまでのスパコンとしては、1994 年からベクトル並列型計算機を採用してきたが、今年の 3 月にベクトル並列計算機単一システムではなく、1024 台の PC で構成した大規模 Linux クラスタと大容量メモリが利用可能な SX-7 という 2 種類の計算資源を備えたシステム：理研スーパー・コンバインド・クラスタ（RSCC）を導入した。

２． リプレースの背景

ベクトル並列計算機から大規模 Linux クラスタを中心としたシステムへ大きく変更した背景には、情報基盤センターでの Linux クラスタ・システムに関する調査・テストの結果、センターマシンとして十分な性能を発揮できると判断したことにある。また、旧システム（VPP700E/160）の利用状況、利用者からの要求などを踏まえた結果、出来るだけコスト性能比が良く、多くのフリーソフトが利用でき新規の研究分野の利用者を開拓できる環境を目指した。その結果、大規模 Linux クラスタを中心とし、さらに１CPU で大容量のメモリが利用できる環境として、大容量メモリ計算機（ SX-7）を合わせた複合システムとした。

３． システム構成

RSCC の Linux クラスタは、1024 台（2048CPU）の PC（計算ノード）で構成されている。これらを 1 つのクラスタとするのではなく、512 ノード（1024CPU）、128 ノード（256CPU）×4 という 5 つのサブ･システムに分割している。更に、理研で開発された分子動力学専用ボード（MDGRAPE-2）を 20 枚搭載した 64 ノード（64CPU）の Linux クラスタをあわせて、6 つのサブ･システムに分割している。これらの Linux クラスタでは、MPI などによる並列化されたプログラムを実行することを想定している。また、128 台のサブ･システムは占有利用も可能としている。ジョブ実行時間は、これまでの VPP700E/160 では最大でも 50 時間だったが、CPU 数が劇的に増加したため RSCC の Linux クラスタでは最大 1 週間のジョブクラスも用意している。RSCC には Linux クラスタ以外にアーキテクチャの異なる SX-7 も配置している。SX-7 は、共有メモリ型のベクトル計算機なので、

並列化できないプログラムで大きなメモリを必要とするプログラムやベクトル・チューニングされたプログラムの実行を想定している。また VPP700E で利用していたプログラムで、すぐに Linux クラスタに移行できない場合には、この SX-7 を“とりあえず”利用してもらい、将来的には Linux クラスタへ移行してもらう。つまり、VPP700E から Linux クラスタへの中継的な役割と位置づけている。

４． システムの特徴

複数のサブ･システムで構成した RSCC では、利用者の利便性やシステムの拡張性を考慮した設計を行い、新しい試みを行っている。

- シームレスな操作性

  複数に分割したサブ･システムをユーザーに意識させないために、RSCC ではフロントエンド計算機を配置している。このフロントエンド計算機から利用者のホームディレクトリをマウントし、利用者はプログラムの作成、コンパイル、バッチジョブの投入、結果の確認まで全ての作業をこのフロントエンド計算機で行えるようになっている。バッチ型ジョブを実行する際に、利用者は予め用意されたジョブのキュー名を指定するだけで、どのサブ･システムでジョブを実行するかを意識しなくても良いように工夫している。

- ファイル転送

  Linux クラスタの計算用 PC には、利用者のホームディレクトリをマウントしていない。中規模までの Linux クラスタであれば、NFS などによりネットワーク経由でホームディレクトリをマウントすることが考えられるが、RSCC ではノード数が多いため、ネットワーク経由でのファイルシステムのマウントは安定性に欠け、ファイル I/O の性能も低下することが予想された。そこで、プログラム実行時に、各計算用 PC のローカル HDD に必要なファイルを転送し、ローカル HDD でのファイル I/O を行うように設計した。これにより、ネットワーク経由に比べて高速なファイル I/O を実現することが可能である。また、計算用 PC が万が一壊れてファイルを壊しても、元のファイルはホームディレクトリにあるので、プログラムの再実行が容易であるという利点もある。さらに、このファイル転送の仕組みは OS、システム構成に依存しないため、今後 RSCC にサブ･システムを追加した場合でも容易に対応可能である。

- チェックポイント・リスタート、ジョブ凍結機能

  RSCC では、臨機応変な運用変更のためにチェックポイント・リスタート機能・ジョブ凍結機能を実装している。この機能を用いると、たとえば平日には 128CPU までの並列ジョブを実行させ、金曜日に実行中のジョブを凍結し、週末には別のジョブを実行させ、次の週には金曜日に凍結したジョブを途中から再実行させるということが可能になる。この機能は停電時などにも有効である。

- ポータルサイトの構築

  コマンドを用いた計算機の操作に不慣れな利用者のために、Web ブラウザを用いてより簡単にシステムを利用できるように、HPC ポータル・Bio ポータルを構築しサービスしている。

- 特定の研究分野による占有利用

  常に最新のバイオ関連 DB を Linux クラスタのローカル HDD に配置するために、256CPU を占有し、バイオ関連 DB ミラーサービスと連携を行っている。また、アメリカ・ブルックヘブン国立研究所との共同研究による、高エネルギー物理実験データ解析で用いている高速テープライブラリ装置（HPSS)を RSCC にも導入している。

- 実時間可視化環境の提供

  RVSLIB を導入し、計算実行中にデータをリアルタイム可視化することを可能としている。また可視化画像は、Web ブラウザ/可視化端末（AVS)で表示・変更を可能としている。

- XP Fortran
  Linux クラスタで VPP Fortran をそのまま実行出来るように、Linux 用 XP Fortran を導入している。
- ITBL プロジェクトとの連携
  RSCC では、GRID 技術の導入により、計算資源を仮想化し計算機を意識することなくジョブの実行を可能としている。将来的には ITBL プロジェクトにおける計算リソースとしての利用も予定している。

５． RSCC における問題点

実際に運用を開始して浮上してきた問題を２つ紹介する。１つは１ノードに２CPU 搭載した SMP の計算ノードにおいて、計算リソースが効率的に利用出来ない。つまり、並列化されていない１つのジョブが１ノードを占有して、1CPU ずつ利用することが出来ないという問題がある。Linux クラスタで実行するジョブは、主に並列ジョブであると想定していたが、実際には並列化されていないジョブを高スループットで実行したいという要望が予想以上に多かった。この問題はバッチ・ジョブ・スケジューラとして採用した富士通製 NQS の修正により解決の目処が立っている。

２つめは、ジョブ凍結機能の問題である。RSCC の Linux クラスタでのジョブ凍結は、SCore の機能を利用している。そのため、ジョブ凍結を可能とするためにはいくつかの制限がある。実際の運用でジョブ凍結に成功したのは、実行中のジョブの 1 割程度であった。ジョブ凍結できなかった原因については現在調査中であるが、将来的にはジョブ凍結の制限事項を減らす必要があると考えている。

６． システムの性能

システムの実効性能の例として、「姫野ベンチマーク」と「LINPACK ベンチマーク」を用いた結果を紹介する。姫野ベンチマークは、非圧縮性流体解析ソフトのカーネルを抜き出したベンチマーク・プログラムで、非常に簡単なソースコードとなっている。Web ページ（http://accc.riken.jp/）からは、オリジナルに以外にも、MPI や VPP Fortran(XP Fortran) により並列化したバージョンもダウンロードが可能である。今回は、MPI と XP Fortran により並列化されたものを用いた実効性能を示した。特に MPI バージョンでは、計算サイズを変化させ、インターコネクトとして Infiniband（IB）と Myrinet を用いたときの性能を 256CPU まで示した。ネットワーク帯域の理論値は IB が片方向 8Gbps、Myriniet が片方向 2Gbps と IB の方が有利であるが、レイテンシーは Myrinet の方が僅かに小さい。計算サイズが小さい場合、各計算用 PC が担当する計算量に比べてデータ通信量の比率が大きいので、並列度（利用 CPU 数）を大きくしていくと、レイテンシーの小さい Myrinet の方が高い実測値を出すことがわかる。そのため、大規模な計算量で通信量が多い場合、IB を用いたが方が良い。実測値をみると、256CPU 用いた場合に約 100GFLOPS と非常に高い性能を示している。

次に、VPP Fortran バージョンの姫野ベンチマークを用いて、XP Fortran の性能を測定した。RSCC の計算用 PC は 1 ノードあたり 2CPU 搭載している。並列計算は 2CPU/1node を使用して計算を実行させているが、XP Fortran の場合 2CPU/1node でジョブを実行するより、１CPU/1node で実行する方が高性能なことが分かる。これは、VPP システムの DTU（ハードウェア）が行っていたデータ通信処理を、XP Fortran ではソフトで処理するために、データ通信用のスレッドが CPU 資源を必要とするからである。データ通信用スレッドのために、ベンチマークを 1CPU/1node で実行した結果は、VPP700E/160 の結果より高い性能を示している。

TOP500 リスト（http://www.top500.org/）を決定するために用いられている LINPACK ベンチマークの測定では、5 つに分割した Linux クラスタ全てを用いて、1024 台（2048CPU）での測定を行った。各 Linux クラスタ内は IB もしくは Myrinet を利用し、各クラスタ間

は、16node 単位で Gigabit Ethernet を利用するように工夫して測定を行った。その結果、8.728 TFLOPS という非常に高い結果を出し、6 月に発表された TOP500 では 7 位にランクされた。今回の LINPACK 測定では、異なるネットワークを備えた複数の Linux クラスタを同時に利用した初めての試みということで、高い評価を受けた。

７． 利用状況と障害件数

3 月に運用を開始した RSCC の利用状況を、各サブ･システムごとに紹介する。導入直後ということもあり、平均して 50%前後の CPU 利用率は比較的良いスタートである。3 月から５月までの３ヶ月間はテスト運用を行った。6 月からの本運用では、「スパコン課題審査委員会」による審議の結果、許可された申請者だけが利用出来ることになっている。本運用開始にあたってシステム設定の変更などを行ったために、6 月当初は利用率が一時的に低迷したが、7 月に入ってからは徐々に上昇してきている。今後情報基盤センターでは、ユーザー教育・プログラム相談などのサポートを強化するだけでなく、MPI、スカラーチューニング、可視化などの講習会の実施や、潜在的な利用者を発掘するための「プログラム高速化・並列化支援」、「プログラム高度化支援」などのサービスを強化していく予定である。

次に、Linux クラスタを計算機センターで運用する上で重要となる、ハードウェア故障について報告する。RSCC では、システムの状態把握を目的としたログの収集と、障害検知を目的とする自動監視を行っている。ログの収集には、理研で開発したログ収集&解析ソフト「Pitsaw」と、Linux クラスタに実装されているハードウェア「IPMI（Intelligent Platform Management Interface)」の２つにより行っている。また、障害検知のための監視は、ハードウェア、ネットワーク、OS、ミドルウェア、ソフトウェアなど様々なレベルで行い、障害検知を行っている。

実際に起こったハードウェアの故障台数は、3 月からの約 5 ヶ月間で 33 台であるが、全て個々の計算用ノードでの障害であったため、全システム停止ということは無い。また、33 台のうち、19 台の故障は LINPACK 測定中に起こったものである。事後の調査によりそのうちの 16 台のハードウェアには潜在的な問題があったことが判明し、まだ故障していないハードウェアも全て予防交換した。また、実際に障害は起こらなかったが、IB のケーブルに潜在的な問題があることも判明し、512 ノード（1024CPU）のサブ･クラスタの IB ケーブル 1024 本のうち半分を予防交換した。

８． おわりに

これまで運用してきたベクトル型並列計算機から大規模 Linux クラスタを中心とした複合システムへのリプレースは、我々にとって大きな挑戦であった。日本で初めて計算機センターのマシンとして Linux クラスタの採用であればなおさらである。そのため、RSCC の設計においては、不足していた機能を補い利便性を高めるために、いくつもの新しい機能と試みを取り入れた。導入後のベンチマークによる実効性能は非常に高く、ハードウェア障害も予想以上に少なく安定稼動している。しかし、バッチ・ジョブ・スケジューラの修正、ジョブ凍結機能の改善などという課題が残っている。これらの課題を克服し、RSCC としてのより効率的なシステム運用を可能とし、利用者の研究活動を支援するシステムを目指していきたい。

以上

付録：プレゼンテーション資料
http://www.ssken.gr.jp/lib/nl/2004/sci/1/ppt1.pdf