理研スーパー・コンバインド・クラスタ（RSCC）の運用報告とリプレース計画について

独立行政法人理化学研究所　情報基盤センター
重谷　隆之

はじめに

2004 年 3 月に導入した理研スーパー・コンバインド・クラスタ（以下、RSCC）は今年 2 月末で 4 年が経過する。本報告ではこの 4 年間の利用状況を報告する。また、2009 年 2 月末で RSCC のリースは終了予定であり、次期システムへのリプレースのための仕様作成を現在行っている。本報告ではその概要を紹介し、我々が次期システムで目指している内容についても紹介する。

RSCC システム

RSCC（図 1）は 1024 台(2048CPU)の大規模 Linux クラスタ、256GB のメモリを搭載したベクトル計算機（SX-7)、そして理研で開発した分子動力学専用ボード（MDGRAPE）を搭載した Linux クラスタの 3 種類の計算機からなる複合システムである。Linux クラスタは 512 ノード（1024CPU）、128 ノード（256CPU）×4 という 5 つのサブ･システムに分割されている。512 ノード(1024CPU)では 32CPU 以上の高並列ジョブを優先的に実行し、128 ノード(256CPU)では 32CPU 以下の並列ジョブを優先的に実行している。
2007 年 4 月には、それまで運用していた MDGRAPE-2 から、その後継である MDGRAPE-3（約 60TFLOPS）に更新し、運用を継続している。

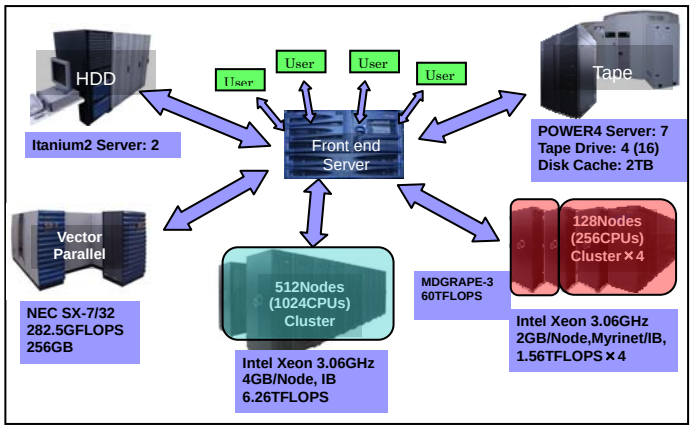

図 1　RSCC システム構成図

統計情報

運用開始から 4 年間で利用者数（登録者数）（図 2）は順調に増加している。研究分野別の利用者数の割合（図 3）を見ると、ライフサイエンスと物理学が７０％以上を占めている。これは、RSCC 以前に運用していたベクトル計算機（VPP700E）で利用が少なかった分野でも利用可能にするという、RSCC 導入時のコンセプトが実現した結果である。

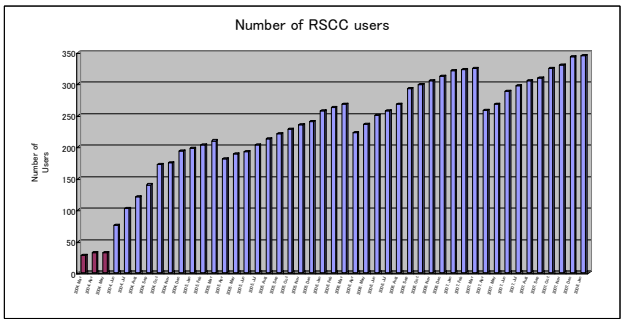

図 2

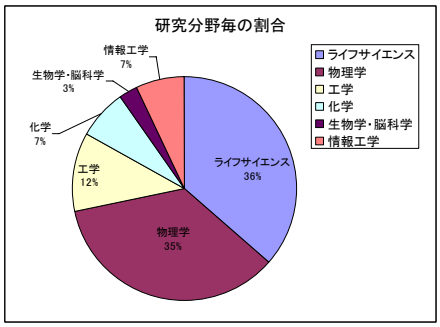

図 3

各サブ･システムの利用率のグラフ(図 4)を見ても分かる通り、全体として利用率が上昇している。特に Linux クラスタでは、2006 年 11 月を境に傾向が変化している。これは、理研で開発したメタ・ジョブ・スケジューラを運用に適用した結果である。
５つのサブ・システムに分割して運用している RSCC の Linux クラスタでは、バッチ型ジョブのスケジュー

リングを制御するジョブ・スケジューラは、それぞれのサブ・システム毎に動作している。ユーザーからのジョブリクエストを一度ジョブ・スケジューラが受け付けると、別のサブ・システム上のジョブ･スケジューラに移動する機能は無い。従って、利用状況によっては利用されるジョブ・クラスに偏りが発生し、CPU リソースの利用状況非効率になっていた。そこで、複数のクラスタを統一的に上位で管理し、資源管理ベースのフェアシェアスケジューリングを含む柔軟なスケジュール機能を備えたスケジューラ（メタ・ジョブ・スケジューラ）を開発し、効率的な CPU リソースの利用を実現した。このメタ・ジョブ・スケジューラによって、ジョブの投入から実行開始までのジョブ実行待ち時間をそれほど増加させることなく、各サブ・システムの利用率は上昇した。

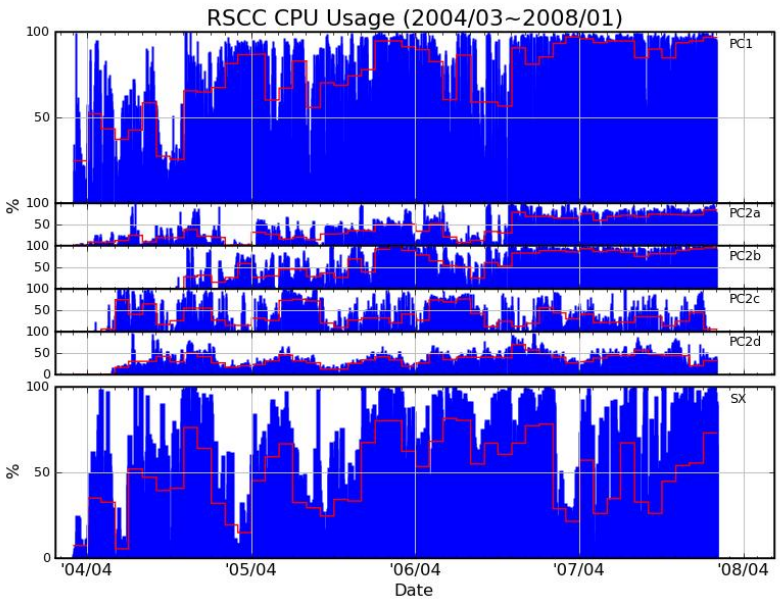

図 4

次期スパコン・システムについて

理化学研究所では、2004 年 3 月にそれまでのベクトル型計算機（Fujitsu VPP700E/160）から、汎用スカラ CPU を用いた大規模 PC クラスタを中心とした複合型システムである RSCC へと大きな変更を行った。それまでのシステムでの利用率が常に 90%前後で、ユーザジョブの処理能力はほぼ限界であったという状況の改善のためばかりでなく、ライフサイエンス分野および高エネルギー物理学における研究の推進を決定した本研究所の中期計画を受けて、需要の増加が見込まれる新規分野としてライフサイエンス分野（特にバイオ・インフォマティックス）や高エネルギー実験データ解析での利用が可能な計算機システムの導入を検討した結果である。

今後も本研究所ではライフサイエンス分野における研究推進を推し進め、それに伴い、本センターのスーパーコンピュータシステムに対するライフサイエンス分野からの需要(データ処理・データ解析)は増加すると予想されている。また、高エネルギー実験からのデータは今後も増大し、それを処理する計算機能力への要求も増加する見込みである。したがって、次期システムは RSCC の設計思想やシステム機能などを継承し、性能（演算性能、主記憶容量、特に磁気ディスク容量とアーカイブ容量）と利便性、運用の効率化などの向上を目指す予定である。さらに、本システムは、現システムの置き換えとなるため、現システムで利用されているプログラム、データ、商用アプリケーションなどの資産の高い移植性も目指している。

具体的には、次期システムは概念的に３つの計算機サブ・システムから構成する予定である。

- １つのプロセスで大規模なメモリが使用できる大容量メモリ計算機部
- 多数の計算用 PC により構成される超並列 PC クラスタ部
- 大容量メモリ計算機部と超並列 PC クラスタ部の中間程度のメモリ容量を利用可能で、SIMD 型専用加速ハードウェアや分子動力学専用計算機（MDGRAPE-3）などを接続可能な多目的 PC クラスタ部

次期システムでは、現システムの２つの計算機部に加えて多目的 PC クラスタ部（含 MDGRAPE-3）を導入することで、システムの適応範囲の拡大を目指している。さらに、計算機部以外にフロントエンドサーバ、磁気ディスク装置、アーカイブ装置を接続し、システム外部からはセキュリティを保ったまま、１システムのイメージで利用可能とする予定である。

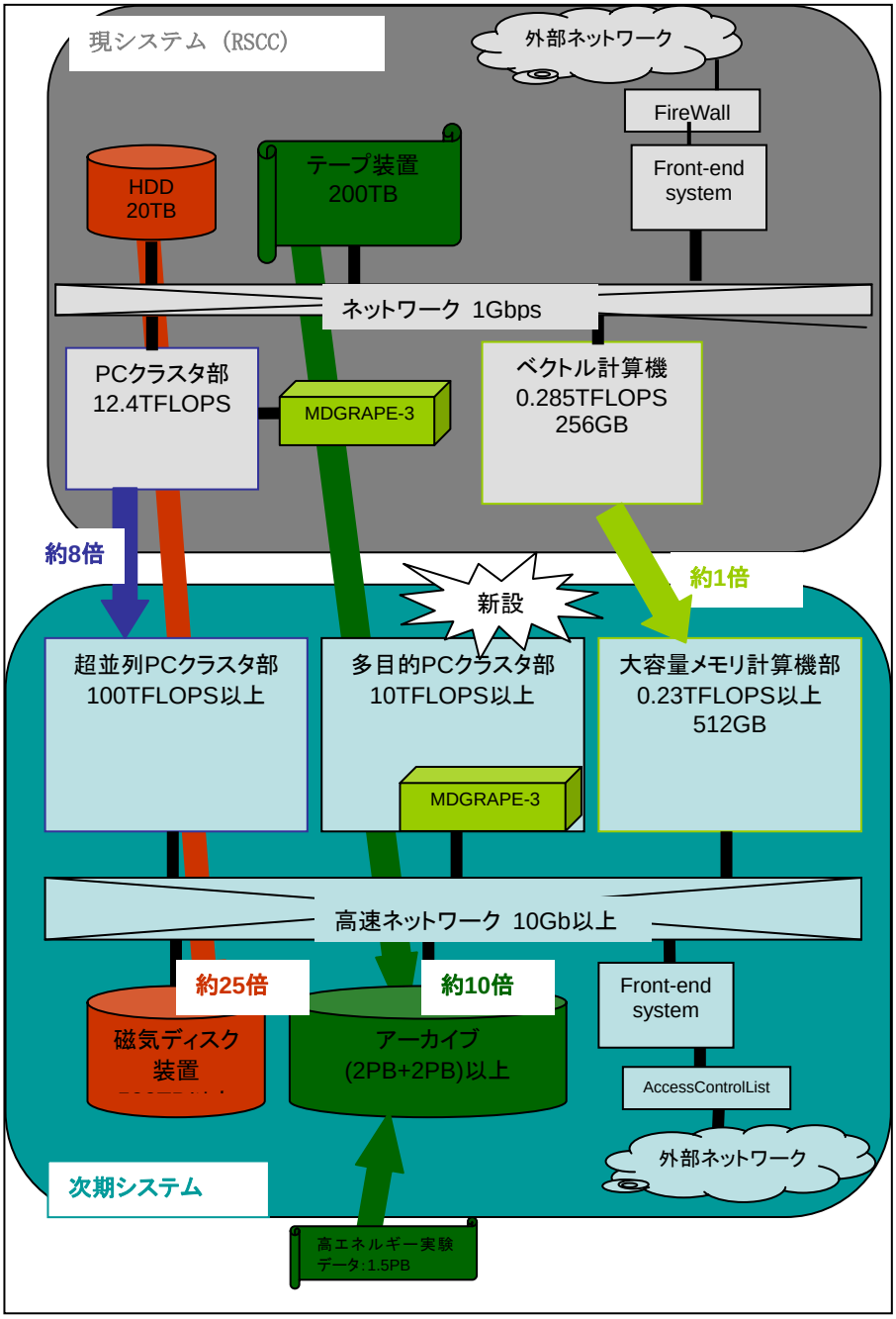

図５　RSCC システムと次期システム（予定）との比較

おわりに

RSCC システムは、それ以前に運用していた単一の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に変更し、いくつもの新しい機能と試みを取り入れた新しいシステムである。情報基盤センターでは、導入からの 4 年間で大規模 Linux クラスタの運用における課題を明らかにし、それらの課題を克服するためのシステム開発を行い、より効率的なシステム運用を行ってきた。

1 年後に予定しているリプレースでは、これまで RSCC で培った技術や運用のノウハウを継承し、ユーザの利便性と演算・データ処理性能の高いシステム構築を行っていくつもりである。