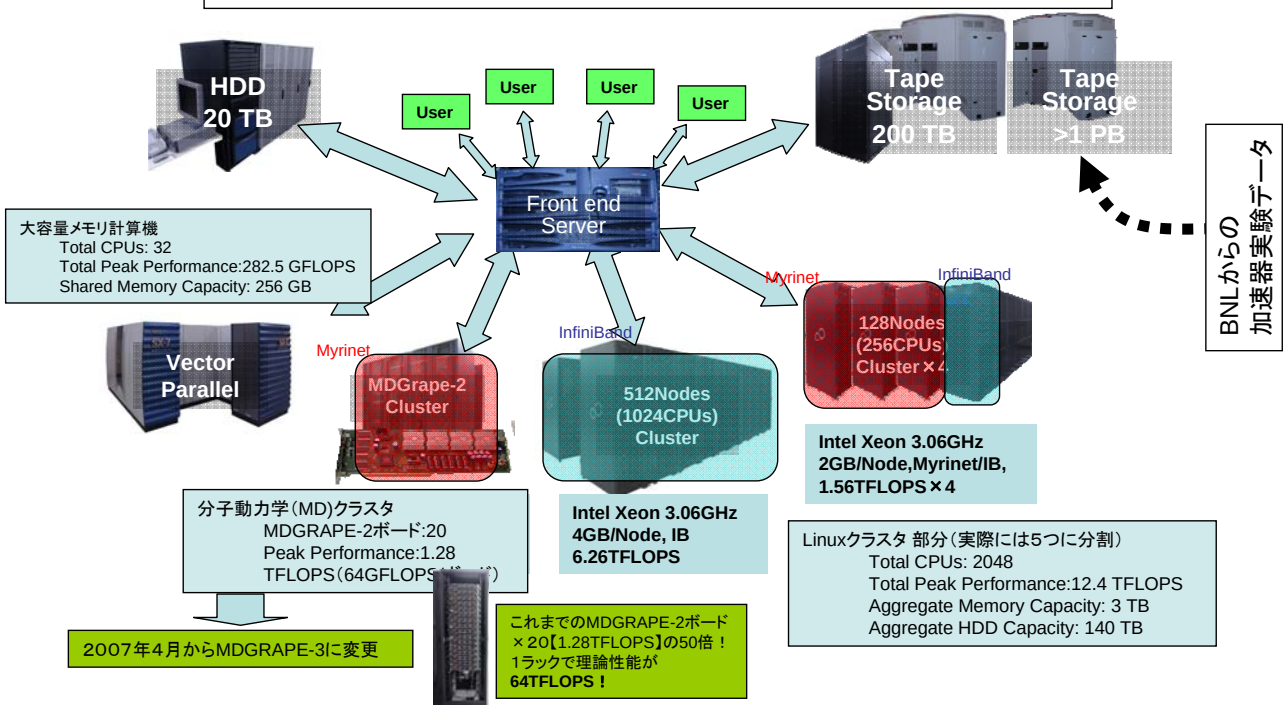
RSCCの運用報告とリプレース計画について

理研・情報基盤センター
重谷　隆之

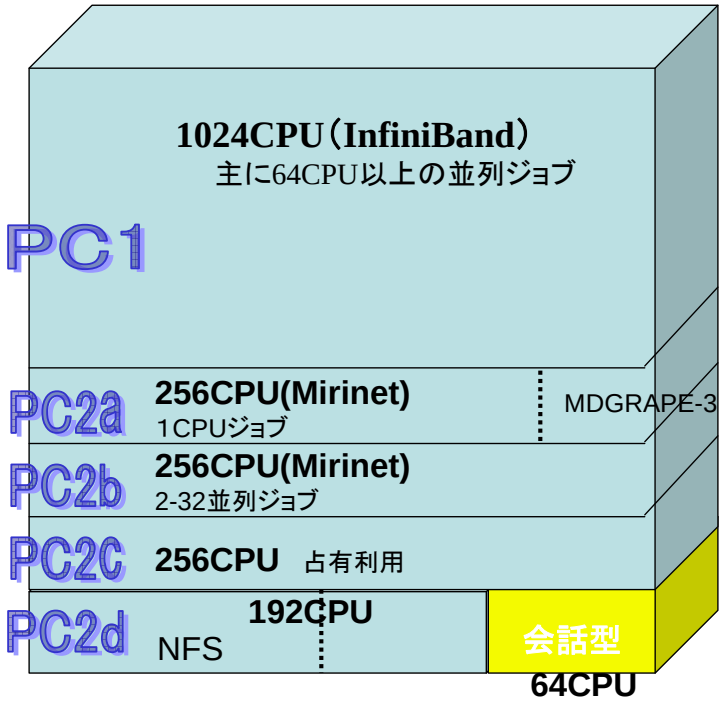
導入から5年目を迎えるRSCC

RIKEN Super Combined Cluster

2004年3月に導入：それまで運用していた単一システム（VPP700E/160PE）から複合型システムへ



5つのLinuxクラスタ

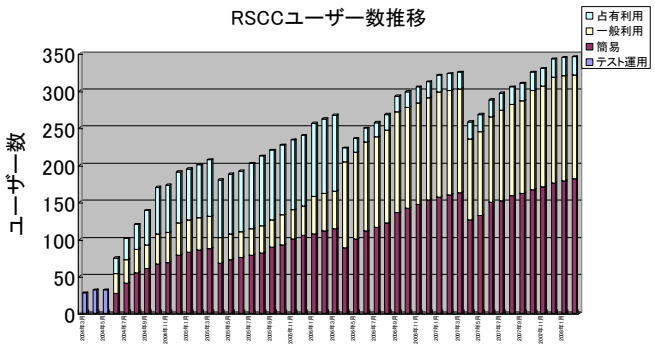


- 2048CPUは5つで構成
- 256CPUのクラスタは占有利用も可能
- 長時間のジョブ実行
  - 24時間から最長1ヶ月間（Gaussian）
- 基本的には

PC1	PC2a	PC2b	PC2c	PC2d
64以上	1CPU MD	32~2	占有利用	ISVアプリ 会話型

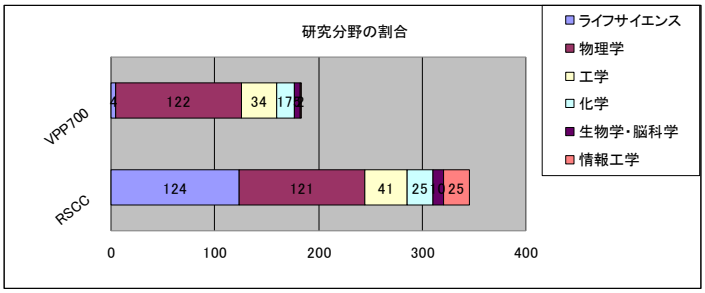
- メタジョブスケジューラの負荷分散機能により、最適なクラスタで実行
- 会話型部分のノード数は固定
- バッチ部分は、ジョブ実行の前後に必要なファイル・データをステージング：ホームはマウントしていない！
- アプリによっては、ホームをNFSマウントする必要もある

ユーザー数の推移



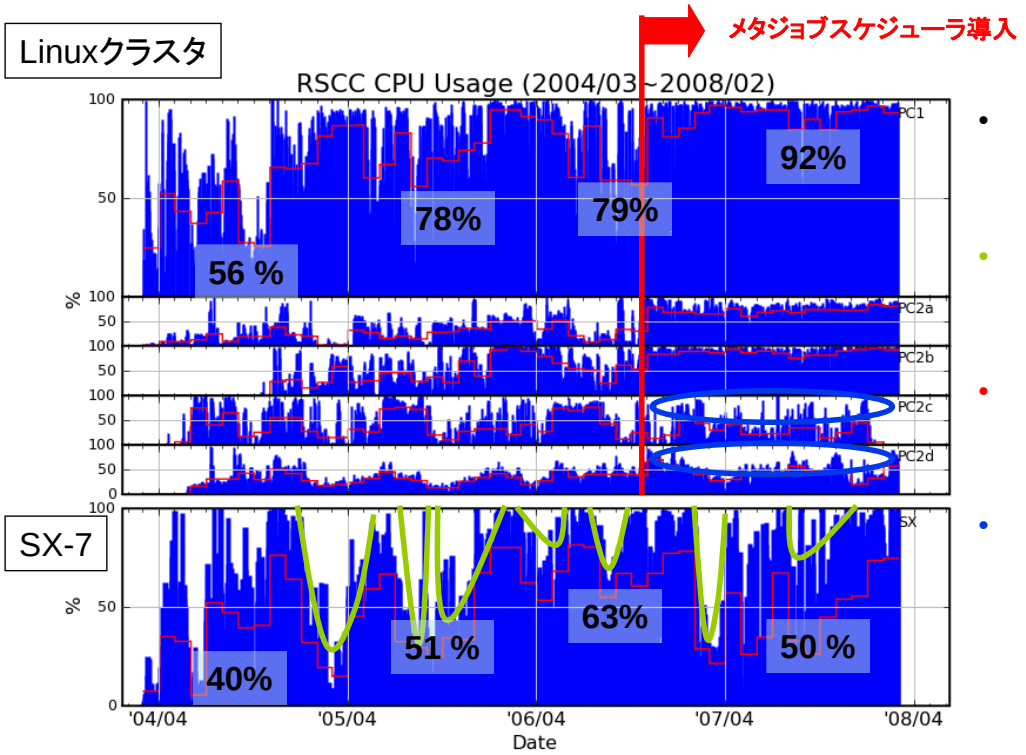
- 初めの3ヶ月間はテスト運用
- 課題審査委員会を設置、利用課題の審査を開始
- 1%未満は簡易利用
- 現在までのユーザー数は約350

- ライフサイエンス分野のユーザー増大
- VPP700（180名）では、物理学、工学が6割



CPU利用率

Linuxクラスタ

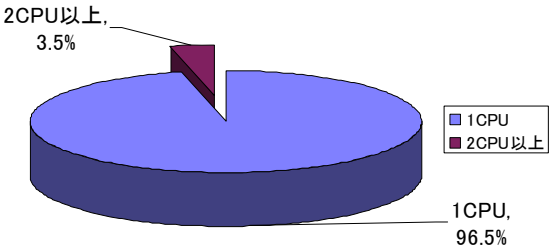


SX-7

- 年に4回の定期保守、2回の計画停電以外は24時間運転
- 夏季休暇、学会時期など季節による閑・繁の差（特にSX）
- メタジョブスケジューラ導入後はクラスタの負荷が分散
- 占有利用（PC2c）とISV実行と会話型ジョブ実行用クラスタ（PC2d）は、まだ余裕があるようにも見える。

利用CPU数別CPU時間とジョブ件数

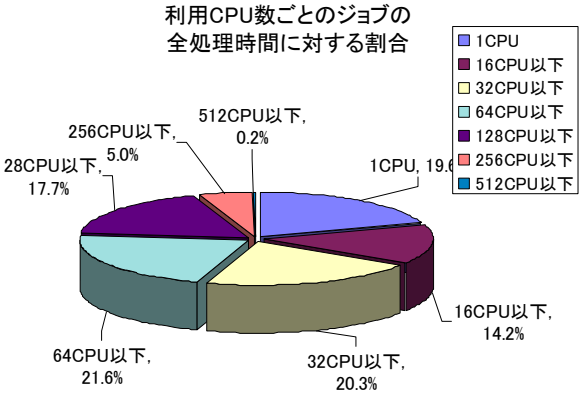
全実行ジョブ数での1CPUジョブ



- **大半（96.5%）が1CPUジョブ**

- **MPIによる32CPU並列以上のジョブが約7割**
- **1CPUジョブは実行時間が短い**

利用CPU数ごとのジョブの全処理時間に対する割合

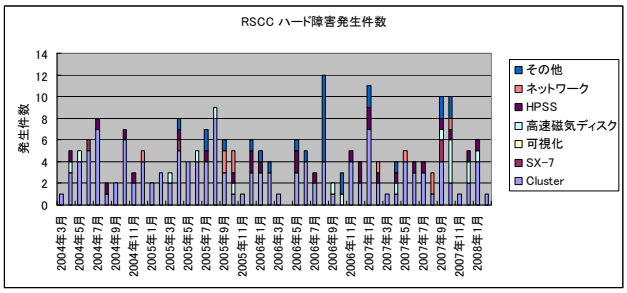


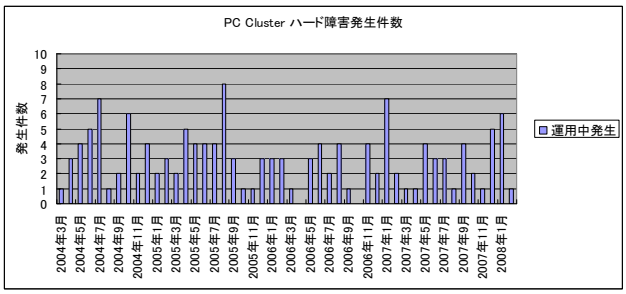
- 2004年3月から2008年2月末までの統計情報
- 傾向は4年間変わらず

Linuxクラスタのハードウェア故障

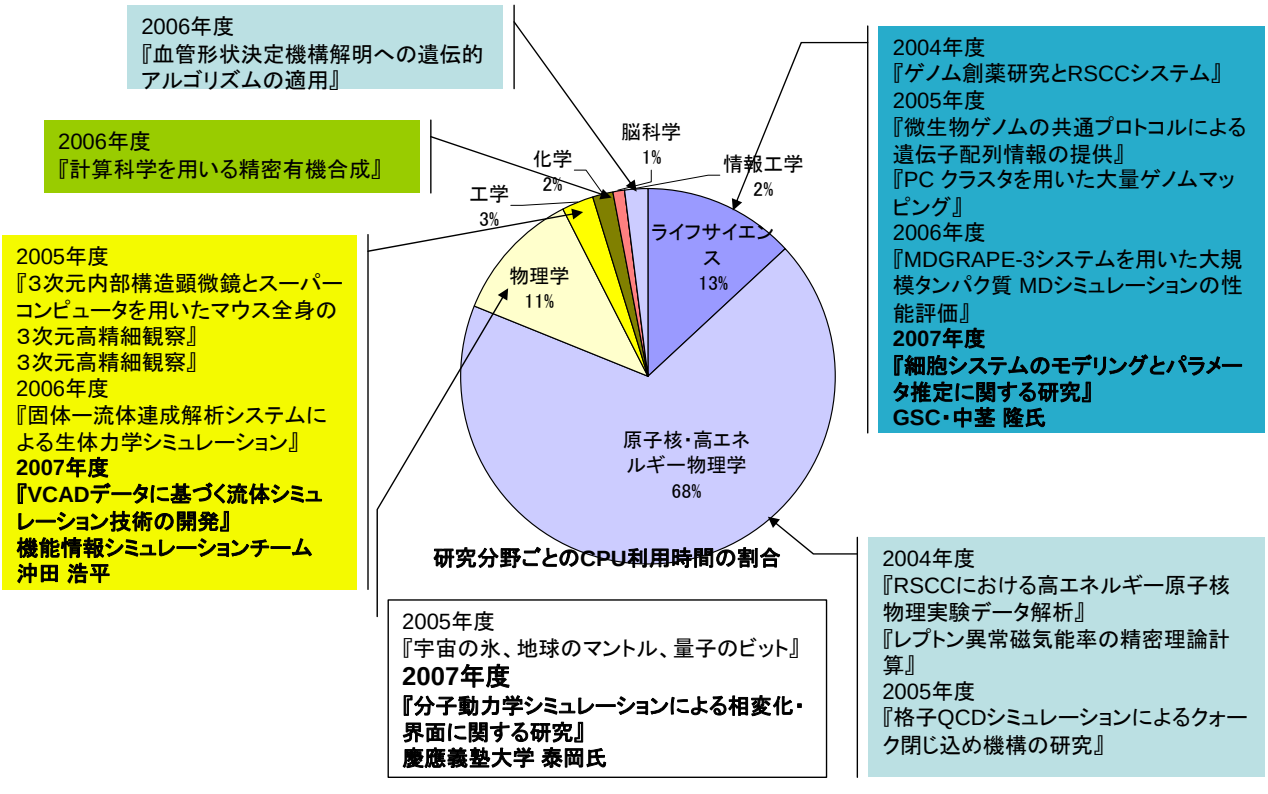
- ハードウェア（ IPMI ）、ネットワーク、OS、ミドルウェア、ソフトウェア（NQS, Pitsaw*）による障害の自動検知
  - Pitsaw：理研で開発したログ収集＆解析ソフト：各計算ノードでログを収集

- 全システム停止なし
- ハード障害は月平均5件、Linuxクラスタだけで月平均3件
- 導入直後はメモリ、マザーボード、SCSIカードなどが多く、最近はHDD、ファン等の故障が目立っている





Linuxクラスタを利用している研究分野



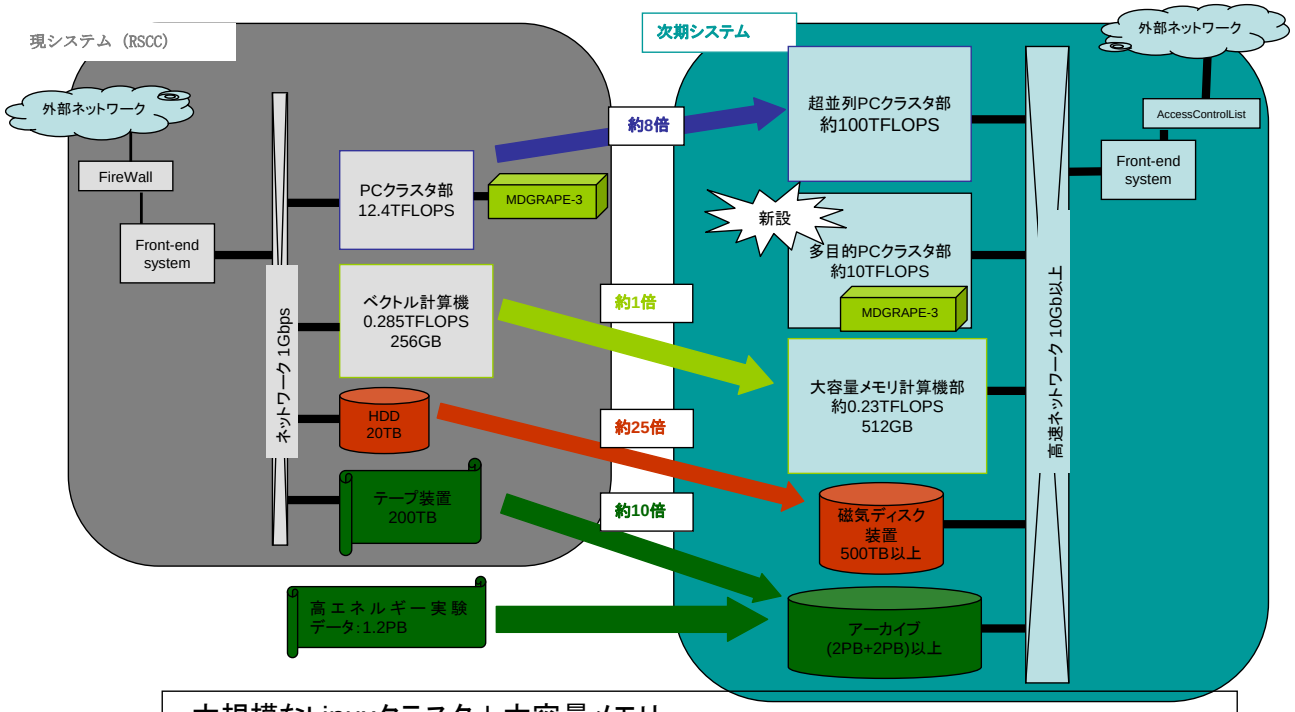
運用で判明した主な課題

- 効率的なリソースの利用
  RSCCは複数のサブ・システムに分割
  →各サブ・システムで負荷の偏りが発生
  →各サブ・システムで動作するジョブ・キューイング・システムが問題
  ⇒ **メタ・ジョブ・スケジューラ**の開発で対応
- 高エネルギー実験データ解析のための占有利用
  →利用するツールはプロジェクトで開発
  →利用するOSが異なる＝占有利用部分だけOSが違う
  →異なるOSでのツールの動作確認は困難（毎日、ツールが更新される）
  →実験がないときには利用率が低い
  ⇒ **2つ以上のOSの切り替え（しかも性能を落とさずに）は可能か？**
- 共有領域が必要
  →基本はステージングでローカルHDDを利用
  →ISV、フリーソフトによっては、NFS領域が必要（中間ファイルなど）
  →NFSしているノードと、NFSしていないノードを自動的には切り替えできない
  →ジョブの量によってCPUリソースの不足・余りが出てしまう！
  ⇒**共有領域とローカルHDDの使い分ける機能が必要**
- 会話型ジョブも必要
  →同じく自動的には切り替えできないので、リソースの不足・余りが出る
  ⇒**インタラクティブ・バッチジョブで解決できる？**

次期システムに向けて

- スーパー・コンピュータ作業部会を発足
  - 各研究分野からメンバーを選出
    - 10名+情報基盤センター
- 2007年1月から検討を開始
  - 次期スーパー・コンピュータ・システム導入に向けた要望調査
  - 全理研計算機リソース調査結果の報告
  - 次期システムに対する要望と基本的な方針
    - **基本的には大規模なLinuxクラスタ**
    - **1プロセスで大容量のメモリが使える環境は必要**
    - ストレージ（HDDやテープ）に関しては、容量を増加
    - バックアップ領域として，ハードディスク(HDD)と同等の使い勝手
    - 可視化用グラフィックスボード（GPU）やSIMD型加速器の導入希望
    - MDGRAPE-3を継続利用
    - RSCCで問題になっている部分の改良・強化
  - 昨年9月に資料招請、12月に仕様書原案を作成
  - 現在、意見招請の最中！

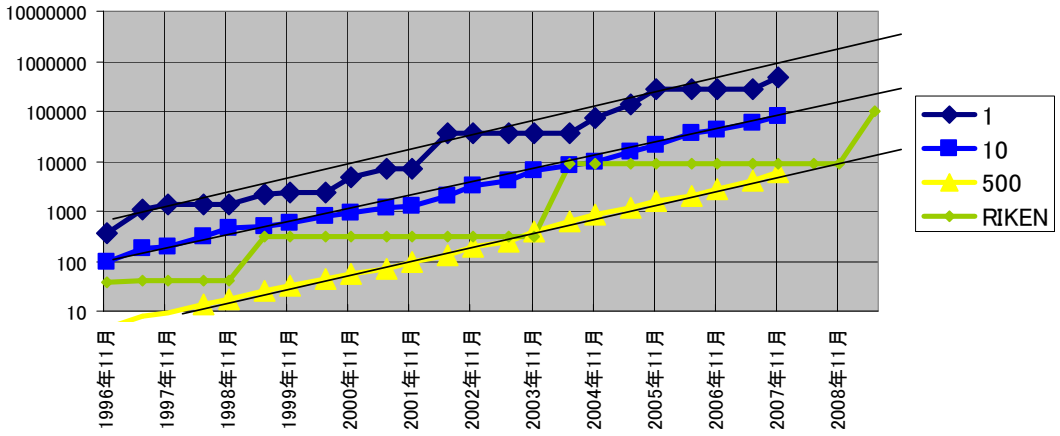
次期システム概略



・大規模なLinuxクラスタ＋大容量メモリ
・クラスタは均一なネットワーク構成
・ISVアプリ、MDGRAPE-3（、SIMD型加速器）などのために多目的クラスタを導入
・HDD・アーカイブ（テープ）を増量

計算機センターにおける主なPCクラスタ

TOP500リストにみる性能予測



- これまで、理研のスパコンシステムの導入直後の順位は
  - 7位（VPP500）、26位（VPP700E)、7位（RSCC)、2007/11の時点でRSCCは、212位
- 次期システムは？
  - 2007/11の8位：IBM Thomas J. Watson Research Center（Blue Gene）
    [91.29TFLOPS (Peak:114.7TFLOPS)]
  - 2007/11の9位：NERSC/LBNL（Cray XT4, 2.6 GHz）
    [85.368TFLOPS (Peak:100.5TFLOPS)]
  - 30位から40位くらい・・・

まとめ

- RSCCは導入から４年が経過して、順調に稼動・利用されている
- 大規模Linuxクラスタの運用における課題
  - 克服した課題もあれば、まだのものもある
- ５年間のリースで残り、約１年
  - 最後の１年間も、より効率的なシステム運用を目指す
- 現在、次期システムに向けた最終仕様を作成中
  - １００TFLOPS級の大規模Linuxクラスタを中心にした、複合型システム
  - RSCCで培った技術や運用のノウハウの継承
  - ユーザの利便性と演算・データ処理性能の高いシステム