RSCCの運用報告

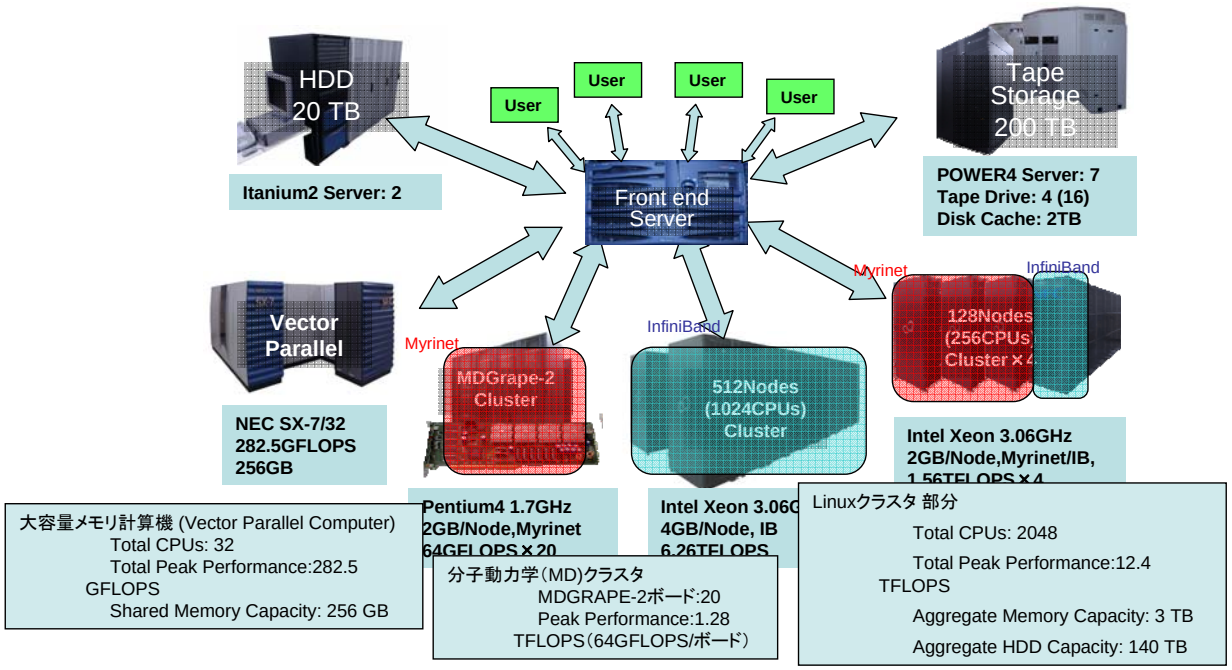
独立行政法人 理化学研究所
情報基盤センター
重谷　隆之

目次

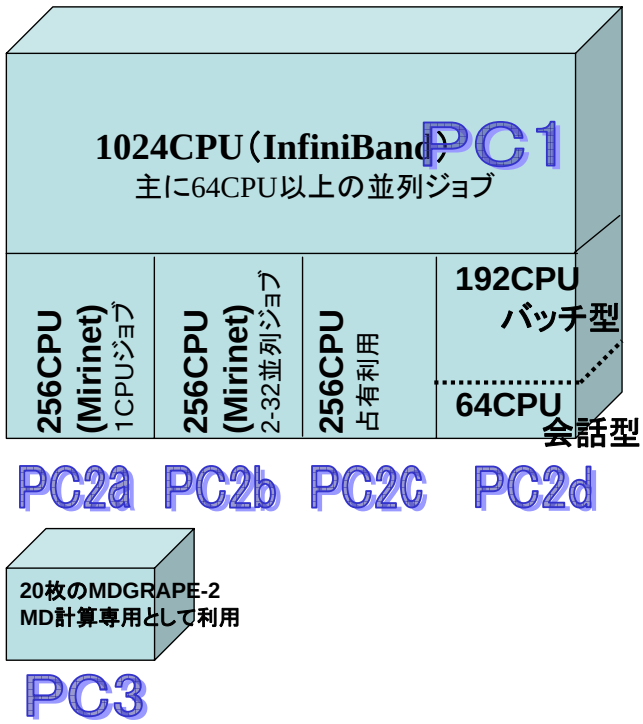
- RIKEN Super Combined Cluster
  - 日本で初めてのPCクラスタ@計算機センター
- よく受ける質問に対して
  - 本当に使えてますか？／使われていますか？
  - ちゃんと動いていますか？／PCクラスタで大丈夫なんですか？
  - 困ったことは無いですか？
- MDGRAPE-3の導入

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



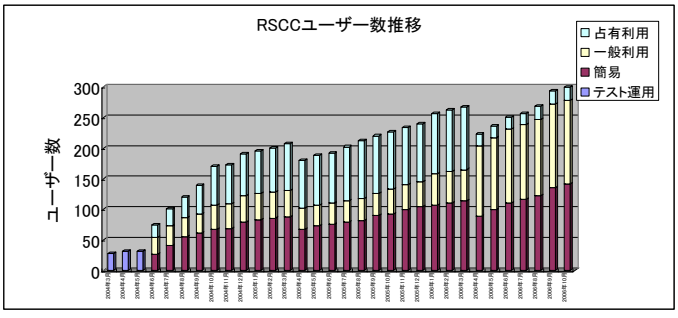
Linuxクラスタの構成



- 2048CPUは5つで構成
- 高並列のジョブ実行
  - 通常128CPU並列，最大1024CPU並列
  - 週末利用など
- 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）

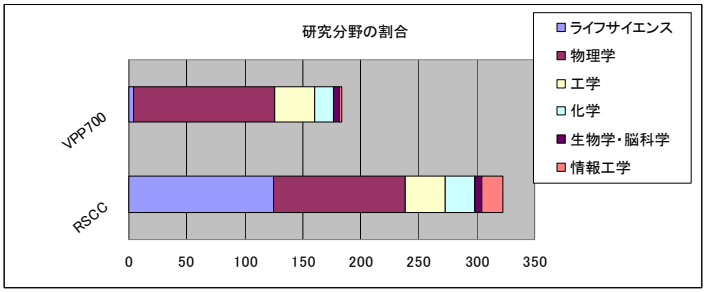
本当に使えてますか？／使われていますか？

ユーザー数の推移



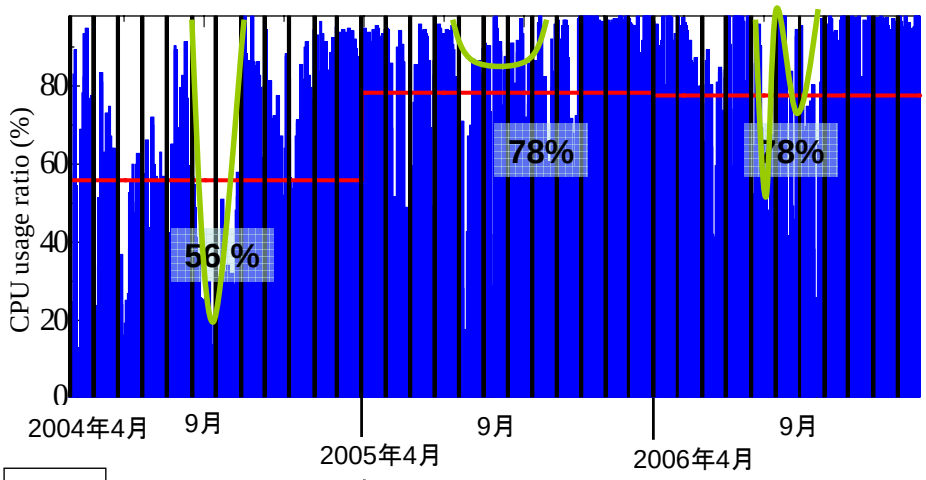
- 初めの3ヶ月間はテスト運用
- 課題審査委員会を設置、利用課題の審査を開始
- 現在までのユーザー数は約300（VPPの最後は180）

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

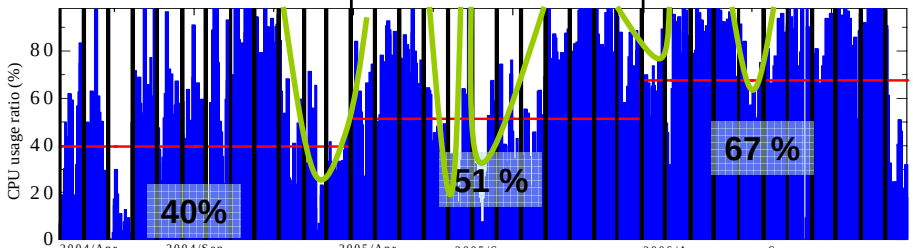


CPU利用率

Linuxクラスタ



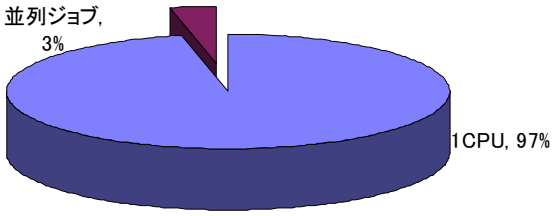
SX-7



- 年間平均利用率：クラスタは1年目より上昇。2，3年目はほぼ同じ。SX-7は年々上昇している。
- 年に4回の定期保守、2回の計画停電ではシステム停止
- 夏季休暇、学会時期など季節による閑・繁の差
- 利用率100%のときもある

利用CPU数別CPU時間とジョブ件数

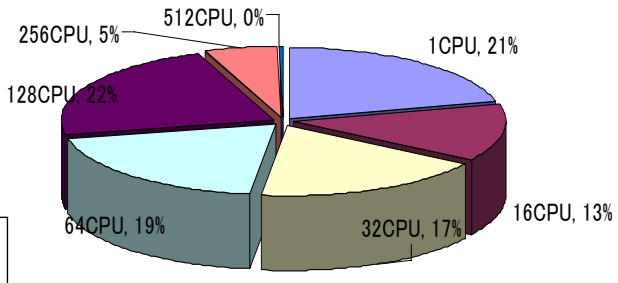
全実行ジョブ数での1CPUジョブ



- 大半（97%）が1CPUジョブ

- MPIによる並列以上のジョブ（128CPU以上）が8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2007年2月末までの統計情報
- 傾向は3年間変わらず

2006年度RSCCにおける主な利用課題（一部）

高エネルギー原子核物理学
格子QCDにおける非摂動効果の研究
QCD閉じ込め機構の解明
レプトン異常磁気能率の精密理論計算
不安定原子核生成時の熱負荷計算
原子核理論(反K中間子と原子核の深い束縛状態の理論的研究)
ウラン元素生成過程の理論的研究
ハイブリッド大規模系成長シミュレーションの開発
重力崩壊型超新星爆発の2次元数値シミュレーション
平均場的計算による原子核構造におけるテンソル相関の役割の研究
第一原理電子状態計算による地球惑星物質科学
ナノスケール分子磁性体の電子スピン共鳴
スーパーコンピューターを用いた乱流予混合燃焼の研究
生体情報を用いた血流シミュレーション
遺伝的アルゴリズムによる生物の持つパターン形成
VCAD データに基づく流体シミュレーション技術の開発
不規則電子系臨界点における共形不変性の数値的実証
荷電させた二酸化チタン表面上での水分子の反応性
不規則電子系臨界点における共形不変性の数値的実証

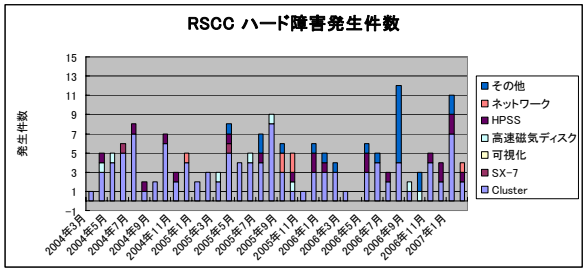
機能性アート錯体の創製と機能
理論計算による合理的な試薬設計と有用な有機反応の開発
機能性アート錯体創製による芳香環化学の新反応・新現象・新機構
新規機能性分子の開発と反応への応用
亜鉛アート錯体を用いたハロゲン−メタル交換反応の解析
有機金属錯体を用いた新規反応の開発
量子化学計算を用いる複核金属酵素の機能解明
自己組織化単分子膜の形成機構の全容解明
ゲノム創薬研究とRSCCシステム
薬剤設計におけるab initio MO法によるパラメータ決定
DNA、RNA塩基分子の光イオン化過程の理論的研究
タンパク質の予測ドメイン特定システムの開発
FRET効果を持つタンパク質の分子設計
分子動力学計算の並列化と応用
水分子チャネルタンパク質の計算機シミュレーション
遺伝統計学に基づく疾患遺伝子探索
in silicoスクリーニング精度向上のための研究
PC クラスタを用いた大量ゲノムマッピング

：
：

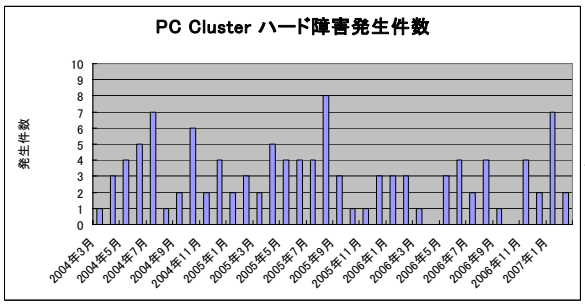
ちゃんと動いています？
PCクラスタで大丈夫なんですか？

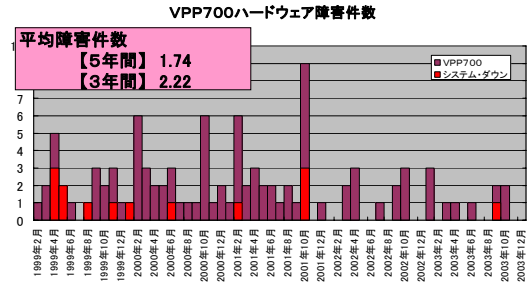
Linuxクラスタのハードウェア故障

- 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



- Linuxクラスタだけのハード障害は、月平均3台
- VPP（160PE)の平均障害件数：【5年間で1.74】、【3年間で2.22 】。システム停止もあった。

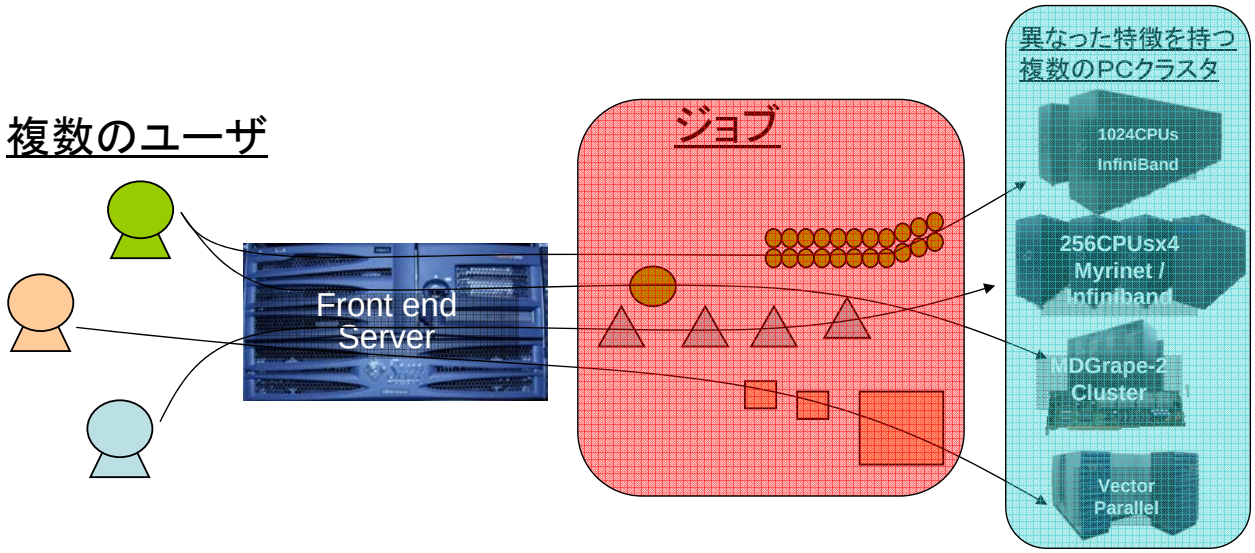




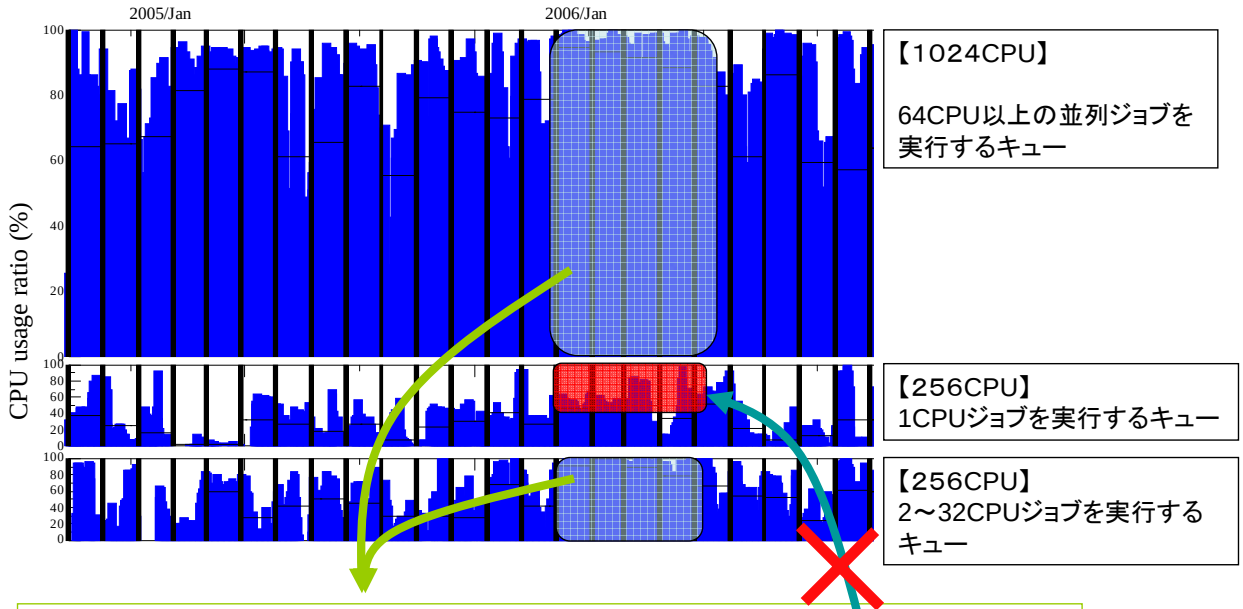
困ったことは無いですか？

ジョブ・キューイング・システムの問題

- Linuxクラスタ（2048CPU）は6つのサブ・クラスタに分割
- サブ・システムごとにジョブ・キューイング・システムが動作（NQS）。SX-7にも別のキューイング・システム（NQSⅡ）
- **場合によっては、特定のサブ・システムにジョブが集中！**
  - 空きリソース（他のクラスタ）があってもジョブが実行できない



実際の例

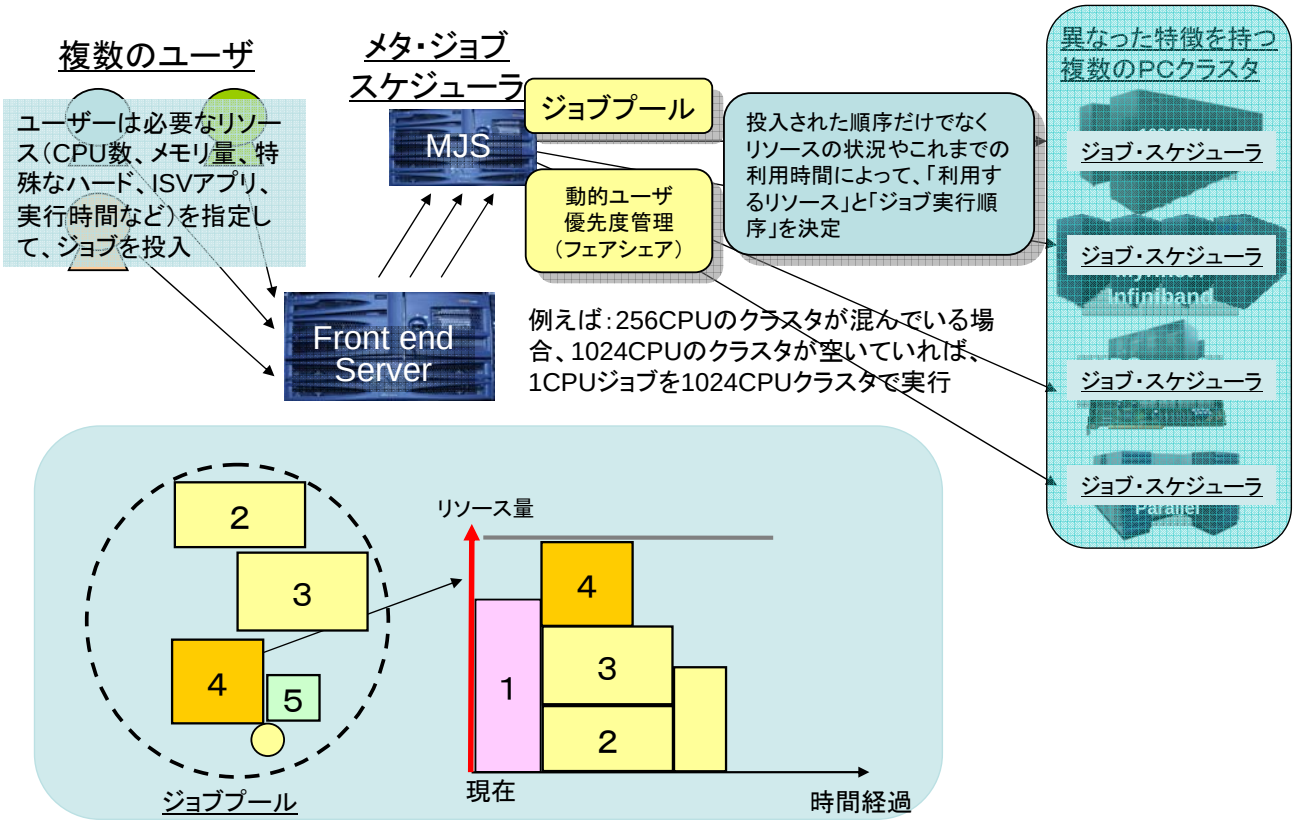


利用率はほぼ満杯。待ちジョブがあっても実行できない！
一度キューイングされてしまったジョブは、他のシステムが空いていても、移動できない！

メタ・ジョブ・スケジューラの開発

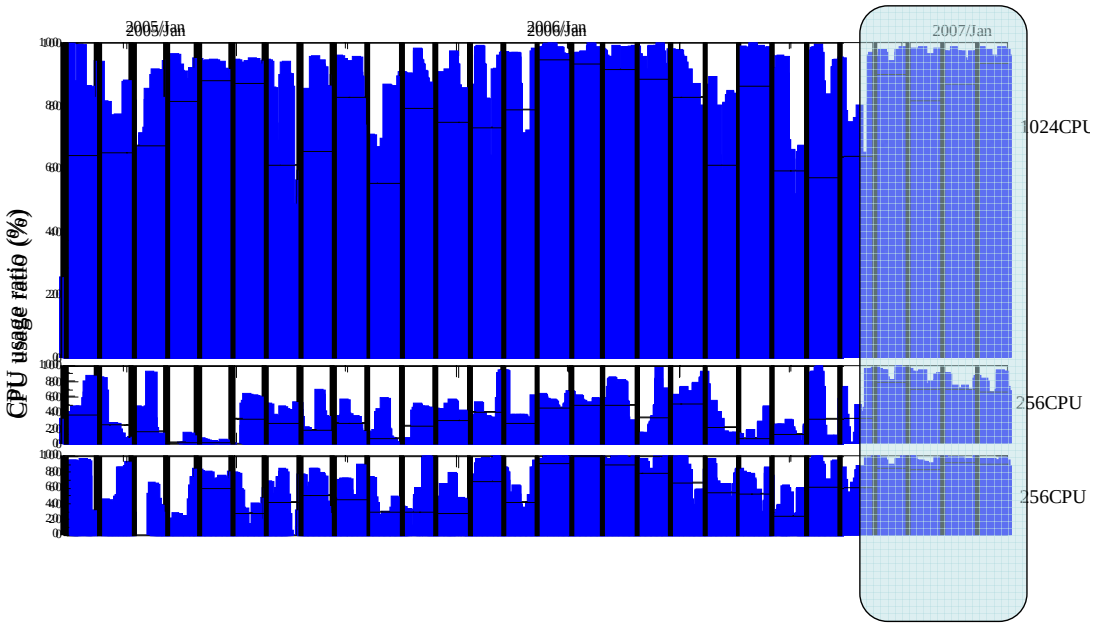
- 目的
  - ジョブ・キューイング・システムの問題を解決
  - 大域的ジョブ制御、リソースの効率的な利用
- 実装・機能
  - システムの既存スケジューラを生かしたまま上位で実行制御。色々なスケジューラへの対応の可能性を残す。
  - キューの概念を排除。リソース（CPU数、メモリ量、アプリケーション、ハードウェアなど）を基準とする。
  - フェアシェア機能、バックフィル機能も
  - SX-7とLinuxクラスタの連成計算にも対応

メタ・ジョブ・スケジューラ詳細



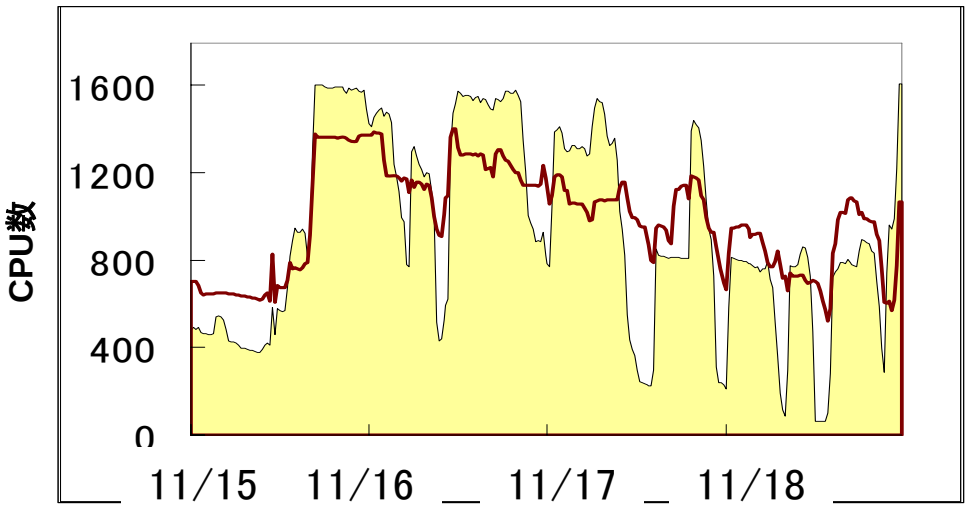
分割したLinuxクラスタごとの利用率

- 2006年11月より実運用
- 運用開始直後から、複数クラスタで満遍なくジョブが実行



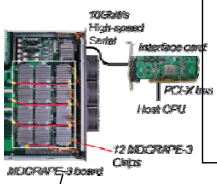
メタ・ジョブ・スケジューラの効果

- 導入前のログを元に導入後のメリットをシミュレーション
- CPU利用率は上昇！
- スループットも短くなった。リクエスト数は同じなので、CPUに空き。
  ⇒投入可能本数を増やすことが可能に



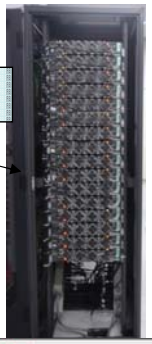
MDGRAPE-3の導入

- 2007年1月末にMDGRAPE-3（32ボード）をRSCCに増設
- これまでのMDGRAPE-2ボード×20【1.28TFLOPS】の50倍！
- 1ラックで理論性能が**64TFLOPS！**
- **ユーザー1人で全てを使うことも可能**
- プログラミングに不慣れな研究者でも使えるように、アプリケーション【AMBER】を用意して、Webインターフェースを開発。



MDGRAPE-3
- 理研で開発した分子動力学専用機
- Interfaceカードを計算ノードにさして接続
- 1つのボードに12個のMDGRAPE-3 LSIチップ【2TFLOPS】

1ラックに32ボード



最後に

- RSCCは5年間のリース
- リプレースは2年後（2009年2月末）
- 現在、理研内部で次期システムに向けた検討を開始

- もちろん、2年間RSCCの運用・開発も継続