Advanced Center for Computing and Communication

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

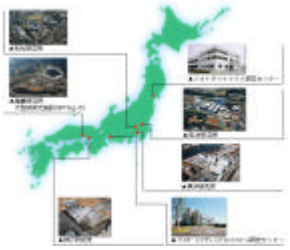
Advanced Center for Computing and Communication

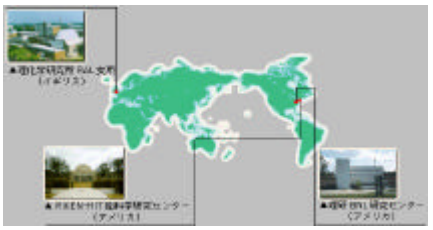
目次

- 独立行政法人理化学研究所と情報基盤センター
- 理研スーパー・コンバインド・クラスタ（RSCC）
- リプレースの背景
- RSCCシステムの構成と設計
- BMTによる性能評価
- 利用状況
- まとめ

独立行政法人理化学研究所と情報基盤センター

- 理化学研究所
  - 日本で唯一の自然科学の総合研究所（文部科学省所管）
  - 物理学、工学、化学、生物学、医科学・・・
  - 拠点：埼玉県和光市以外にも、国内6箇所、海外3箇所
  - 人員：約3000名（これ以外に外来研究者等が約3000名）
  - 2003年10月：特殊法人から独立行政法人へ
- 情報基盤センター
  - 情報環境室⇒情報基盤センター
    - 全理研が対象
  - 共同利用計算機
  - 所内ネットワーク





Advanced Center for Computing and Communication

理研のスーパー・コンピュータの歴史

- 大型汎用機　（メインフレーム）
  - 1994年まで
- スーパー・コンピュータ
  - 1994年2月～
    - ベクトル型並列計算機：VPP500/32PE
  - 1999年2月～
    - VPP700E/128PE（途中から160PE）
  - 2004年3月～
    - 理研スーパー・コンバインド・クラスタ導入

理研スーパー・コンバインド・クラスタ

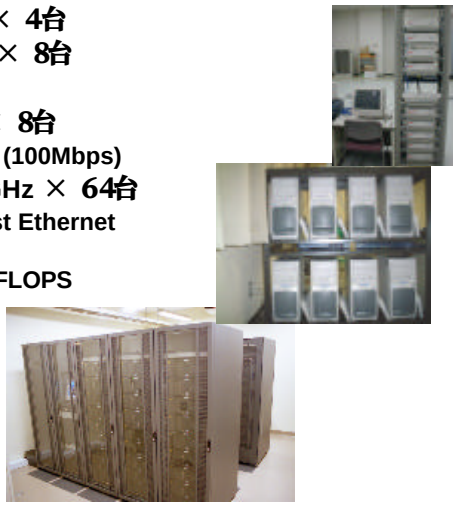
- 計算資源は2種類
  - PCを高速ネットワークで接続したLinuxクラスタ
    - 2048 CPU (Intel Xeon 3.06GHz、Dual)
    - 12.5 TFLOPS
  - 大容量メモリ計算機（SX-7）
    - 32CPU/256GB
    - 282.5 GFLOPS
- 計算機センターのスパコンにLinuxクラスタの採用は日本で初めて

Advanced Center for Computing and Communication

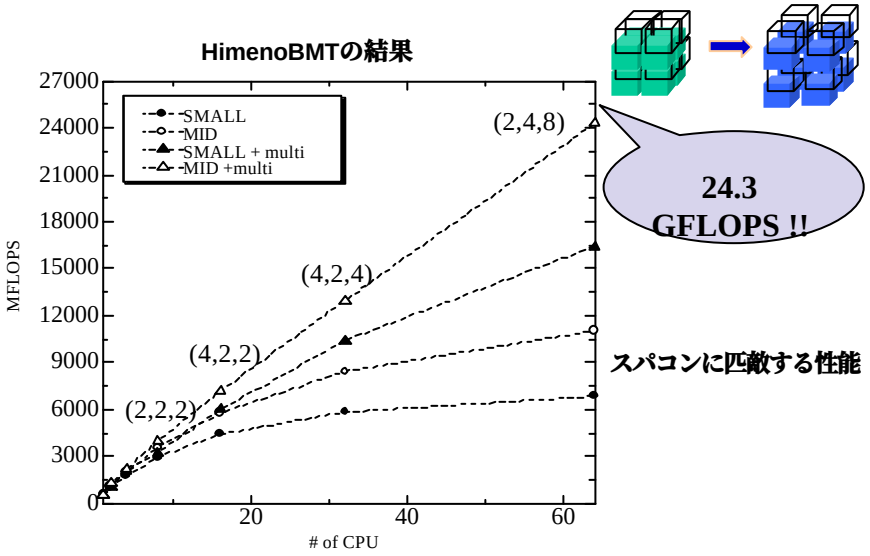
リプレースの背景

リプレースの背景（1）

- **情報環境室におけるPCクラスタのテスト**
  - **1999年 :Pentium II 450MHz × 4台**
  - **2000年 :Pentium III 450MHz × 8台**
    - Network : 100Base
  - **2001年 :Pentium 4 1.5GHz × 8台**
    - **Network : Fast Ethernet x 1 (100Mbps)**
  - **2001年10月 :Pentium 4 1.7GHz × 64台**
    - **Network : Myrinet 2000+Fast Ethernet (100Mbps)**
    - **Peak Performance :217.6 GFLOPS**

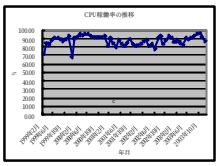


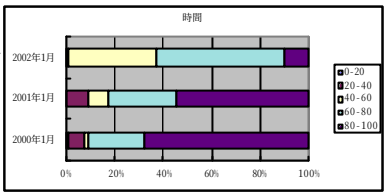
Pentium4×64台での結果（多次元分割の結果）



リプレースの背景（2）

- **Fujitsu VPP700E/160**
  - ベクトル型並列計算機（1999年導入）
    - 計算性能、メモリ容量などPCなみ
    - プログラムがベクトル化されていないと、性能はPC以下
- **稼働率は90%前後で、ジョブは常に満杯**
- **主なユーザーは、計算科学分野**
  - 計算化学、計算流体力学、分子動力学
- **新規のユーザーにとってベクトル化は困難**
  - スカラー向きソフトの利用が増加
  - フリーソフトの利用希望の増加
- **並列化できないプログラムの増加**
  - 1CPUで大きなメモリを使用





Advanced Center for Computing and Communication

システム導入の基本的な考え方

- コスト性能比の良い計算機
- できるだけ多くのフリーソフトが利用できる環境
- 1CPUで大容量のメモリが利用できる環境
- 新規分野の開拓
  - 高エネルギー物理実験データ処理、ライフサイエンス
- 旧来のユーザーも利用できる環境

異なるサブシステムから構成される複合型システム
- Linux PCを高速ネットワークで接続したLinuxクラスタ
- 大容量メモリ計算機（SX-7）

RSCCシステムの導入

Advanced Center for Computing and Communication

RSCCシステムの構成と設計

RSCCシステムの基本設計（1）

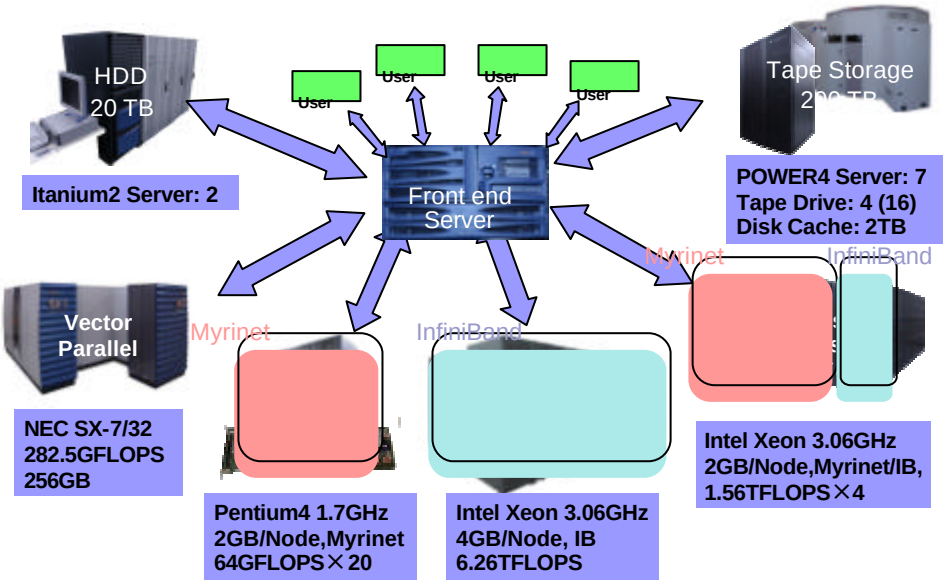
- 用途ごとに異なるサブシステムで構成
- Linux クラスタは6つに分割
  - 高並列のジョブ実行(通常128CPU並列，最大1024CPU並列)
  - 長時間のジョブ実行（24時間から最長1週間）
  - MDGRAPE-2（20枚）を搭載したLinuxクラスタも用意
  - 占有利用も可能（現在は、高エネ物理実験データ解析、ゲノム DB用）
- SX-7
  - 巨大なメモリを必要とするプログラム（Linuxでは最大2GB）
  - ベクトルチューニングされたプログラム
  - VPPからLinuxクラスタへの中継的役割

ゲノム DB用

高エネ物理実験データ解析

PC	512n 1 PU）	InfiniBand (8Gbps)
PC2a	8node 256CPU）	Myrinet XP (2Gbps)
PC2b	128node 256CPU）	Myrinet XP (2Gbps)
PC2c	128node 256CPU）	Myrinet XP (2Gbps)
PC2d	128node 256CPU）	InfiniBand (8Gbps)
PC3 (MDGRAPE-2)	64node 64CPU）	Myrinet XP (1Gbps)

RSCCシステムの構成

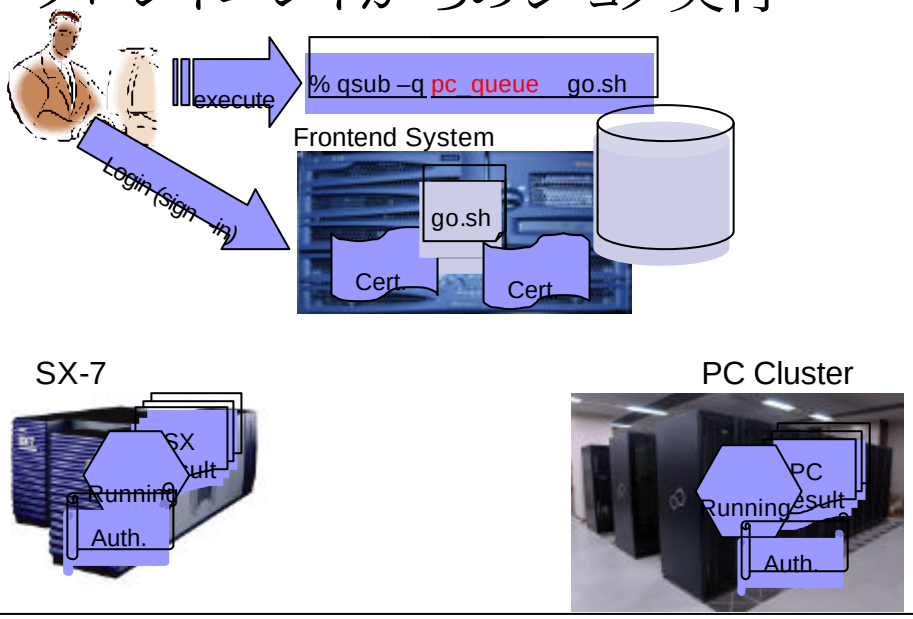


RSCCシステムの基本設計 (2)

- 複合システムを意識させないための工夫
  - フロントエンドからのジョブ実行
    - コンパイルやバッチジョブのサブミットはフロントエンドで完結
  - 利用者には、どのサブ・システムを使うかは明示させない
    - 現在は、キュー名（使用CPU数）を指定
  - データはプログラム実行時にローカルHDDへコピー
    - 1024nodeからのNFSは非現実的
    - I/Oはネットワーク経由よりローカルHDDの方が高速

Advanced Center for Computing and Communication

フロントエンドからのジョブ実行



バッチジョブ・スケジューラ

- バッチジョブスケジューラは各サブ・システムごと
- フロントエンド計算機からジョブをサブミット
- ユーザーは、キュー名だけを指定

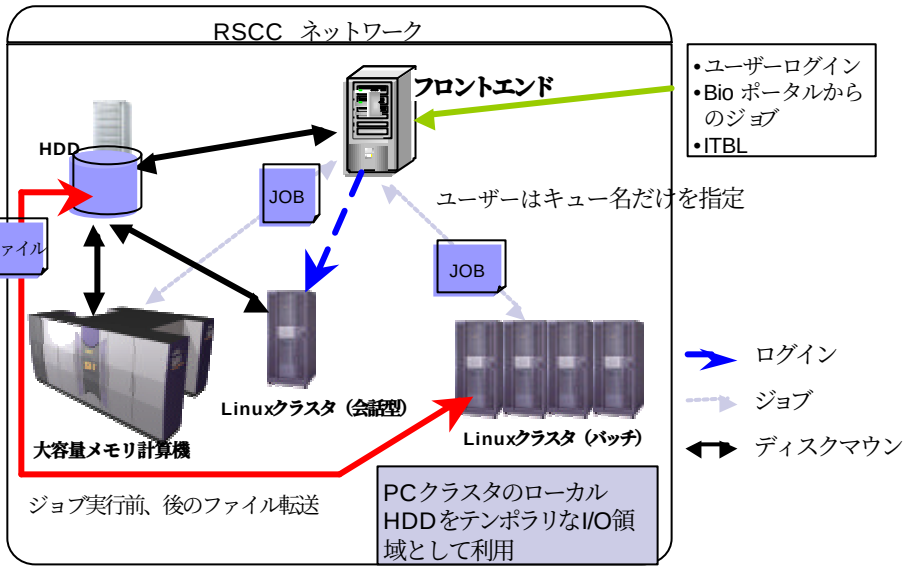
PC1	主に32CPU以上の並列ジョブ	Fujitsu NQS
PC2a	1CPUから16CPUまで	Fujitsu NQS
PC2b		Fujitsu NQS
PC2c	大量の逐次（1CPU）ジョブ	NQS II
PC2d	計算化学アプリ	Fujitsu NQS
	会話型ジョブ	
PC3	MDGRAPE-2	Fujitsu NQS
SX-7		NQS II

Advanced Center for Computing and Communication

ジョブスケジューラの問題

- 予想以上に多い逐次（1CPU利用の）ジョブ
  - RSCCのLinuxクラスタは、2CPU/node
  - Fujistu NQS＋SCoreでは、1nodeに2本の逐次ジョブを同時に実行できない
    - 半分の計算資源が無駄になる
    - InfiniBandドライバはSCoreシステム用
    - チェックポイント・リスタート機能もSCoreと密に連携
  - PC2c：ジョブスケジューラをNQS IIに変更（ただしSCore＝チェックポイントリスタートなし）
  - PC2b：サブシステムを占有して、ジョブスケジューラの改良、テスト
- 256CPUまでは、もうすぐ対応予定
- 1人で数千本のジョブ投入の要求
  - システム全体で、数万本のジョブ
- フェアシェア（一定期間のジョブ実行実績でジョブ・ユーザーの優先度を変更する）機能も必要

プログラム実行時にファイル転送



ファイル転送

- ノード数が多いので、NFSの信頼性に“？”
- ネットワーク経由より、ローカルHDDへのファイルIOの方が高速
- 計算ノードが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
- サブシステムを追加した場合でも対応可能
  - OS、システム構成に依存しない
  - ITBLでのシステム連携も視野

RSCCシステムの基本設計（3）

- 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
    - 実行中のジョブのチェックポイントを取りジョブ凍結、リスタート
    - SCoreの機能を利用
  - チェックポイント・リスタート機能の必要性
    - 通常は128CPUまでの並列ジョブ⇒週末に512CPUを使う場合
    - 停電時などのジョブ凍結

Advanced Center for Computing and Communication

チェックポイント・リスタート

- SCore チェックポイントの対象とならないジョブ
  - SCore以外のジョブ
    - たとえば商用アプリ（Gaussianなど）
    - Fortran、C以外のジョブ（perlなど）
  - ソケット通信を利用するジョブ
  - OpenMP、スレッドなどを利用するジョブ
- 実際の運用では、ジョブ凍結成功は**1割程度**！
  - 原因は調査中
- それでも、異なるノード上でのジョブ再開は不可能

その他の特徴

- HPCポータル：Bioポータルの構築
  - 簡便にスパコンを利用
  - 常に最新の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：バイオ関連DBミラーサービスと連携
- 実時間可視化環境の提供
- LinuxクラスタでVPP Fortranをそのまま実行
  - Linux用**XP Fortran**の導入
- 高エネルギー物理実験データ解析
  - 高速テープライブラリ装置（HPSS）を導入
  - 巨大な解析結果をテープライブラリ装置に直接格納
- ITBLプロジェクト
  - 計算機センターアライアンス
  - GRID技術の導入による計算資源の仮想化
    - ITBLプロジェクトにおける計算リソースとして利用

Advanced Center for Computing and Communication

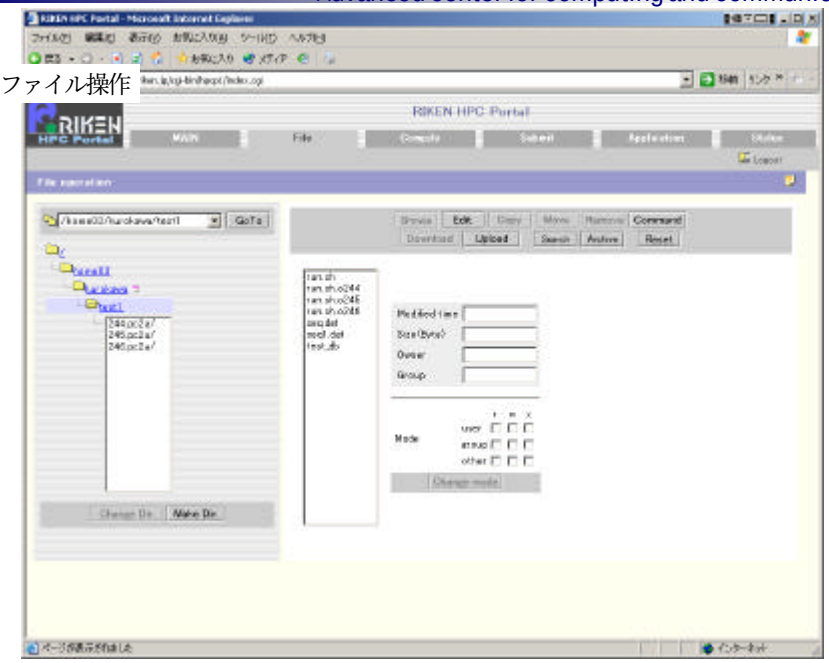
HPCポータル

- もっと簡便にスパコンを利用するために
  - 現状，キーボード入力と普段使わないコマンドを覚える必要がある
  - スパコン利用のハードルを下げる
- リモートでのジョブ実行が簡便に
  - Web技術：ブラウザからマウスでスパコン利用
  - Grid技術：計算資源の柔軟な利用技術
- ターミナル操作を全てWeb操作に置き換え
  - ファイル操作(編集・圧縮・コピー・削除・等々)
  - コンパイル・リンク・ジョブ投入
  - ジョブ・システムのステータス表示
  - 定型作業アプリの実行画面

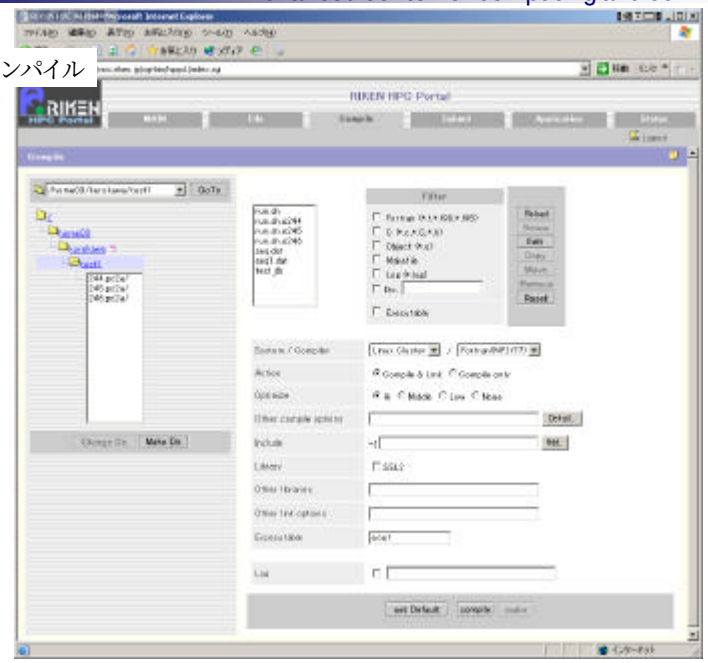
ログイン画面

Advanced Center for Computing and Communication

ファイル操作

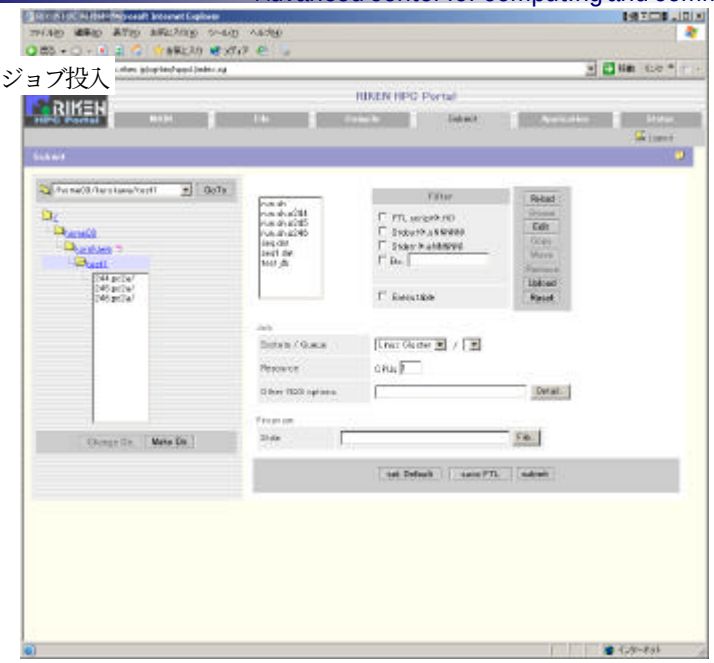


コンパイル



Advanced Center for Computing and Communication

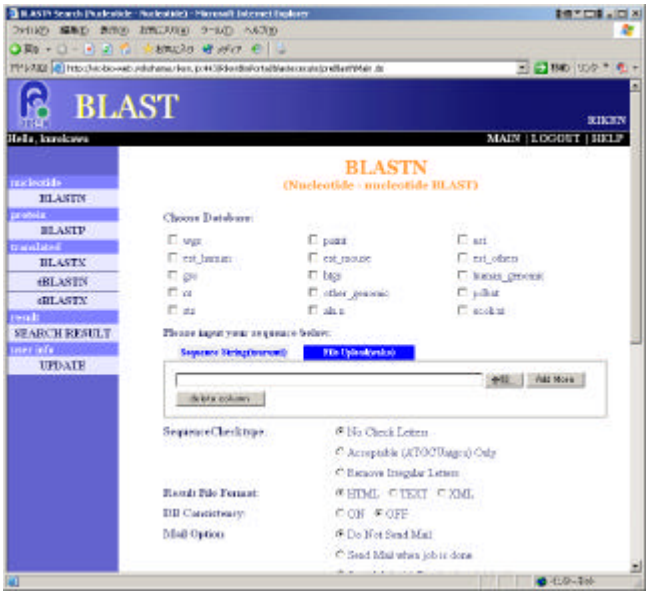
ジョブ投入



ジョブ状態表示

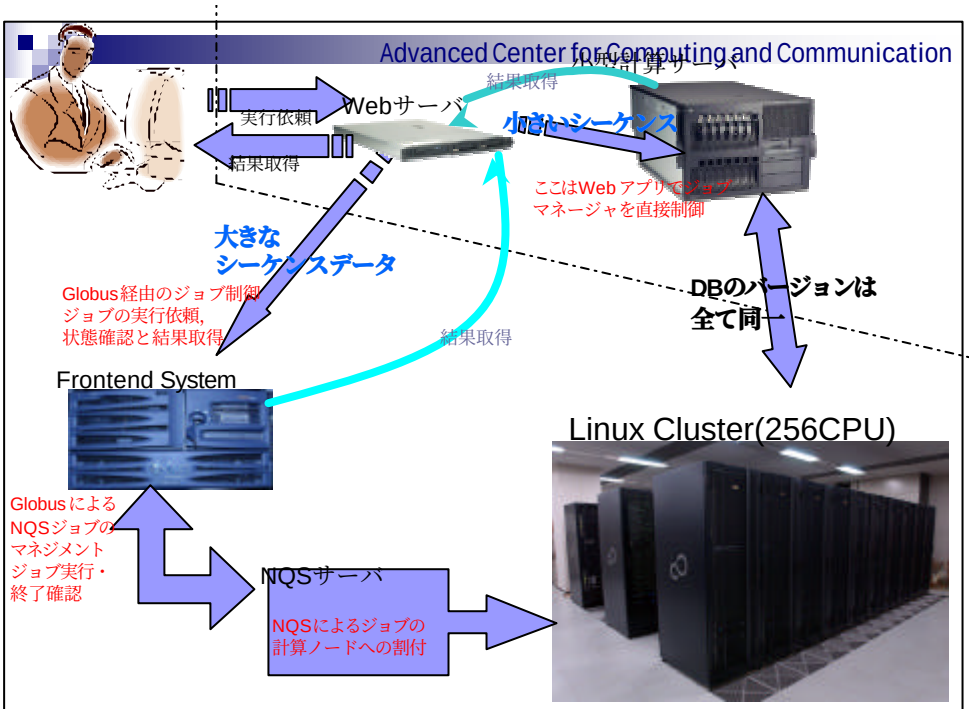
Bioポータル

- DBミラーサービスと共に情報基盤センターが行っている計算処理サービス
- 研究者が容易に利用できることを念頭に開発
- 取り敢えず、要望の一番多かったBLASTをターゲットにシステム開発
- 横浜研究所に設置
  - 大きな規模のジョブは、RSCCで処理
- DBもミラーサーバと同期（常に最新）
  - DBはミラーサーバの更新に合わせて同期



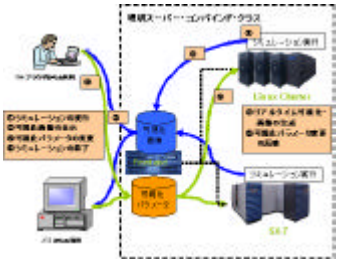
複数DBの指定
複数ファイルの投入

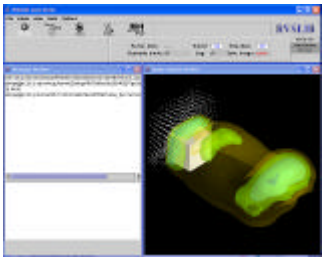
選択されたDBと
投入ファイル数の
組合わせで最もハイ
スループットになる
処理方法で実行



実時間可視化

- RVSLIBの導入
  - SX-7、Linuxクラスタどちらでも利用可能
- 計算実行中にデータをリアルタイム可視化、保存、操作
  - ポスト処理での可視化が不要
  - 並列化プログラムの分散したデータを１つにする必要が無い（サーバサイド並列可視化処理）
  - 保存するデータサイズを小さく出来る
- 可視化画像はWebブラウザ/可視化端末（AVS)で表示・変更可能
  - リモートからの可視化





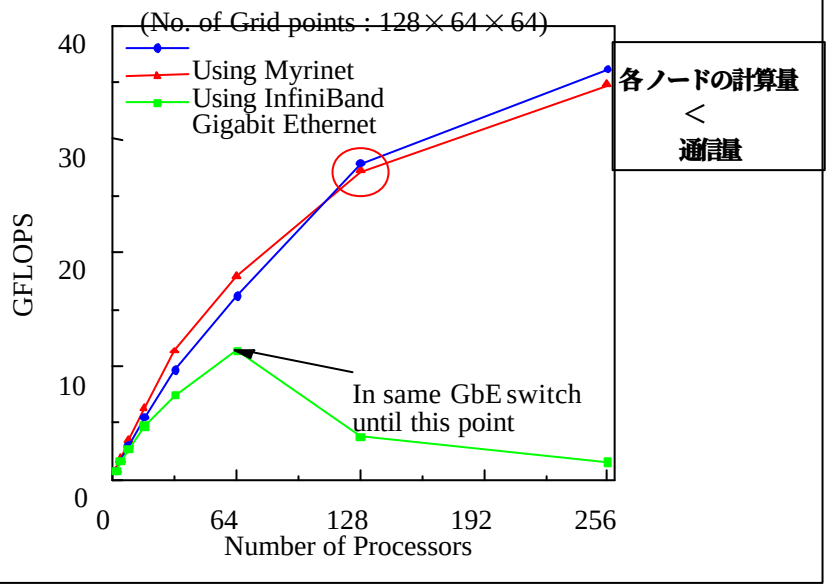
BMTによる性能評価

姫野ベンチマーク（HimenoBMT）

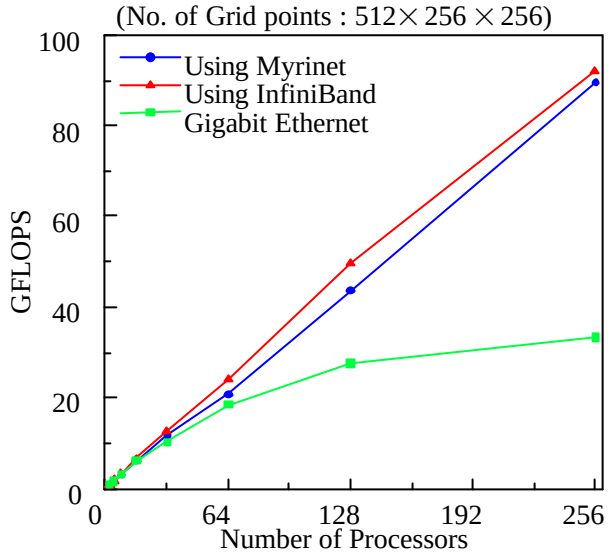
- 非圧縮性流体解析ソフトのカーネル
- 3D Poisson 方程式
- 絶対値（MFLOPS)を測定可能
- 非常に簡単なソースコード
- MPIによる並列化

  - http://accc.riken.jp/

MPI Results (size S)

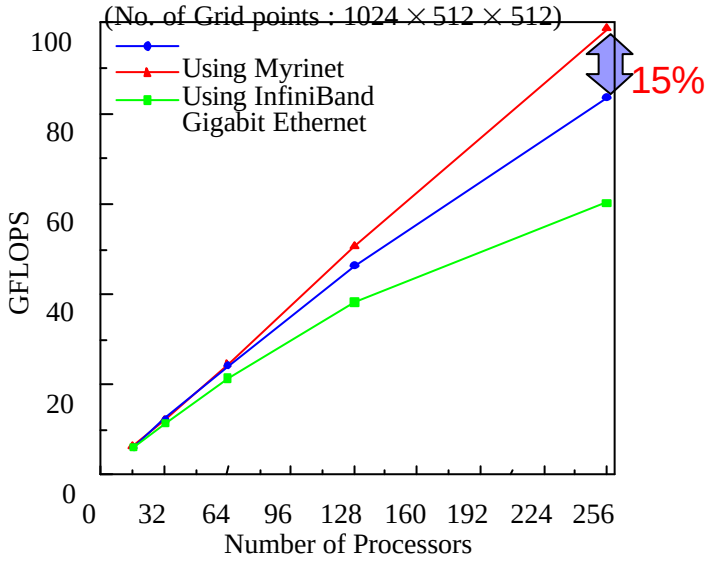


MPI Results (Size L)



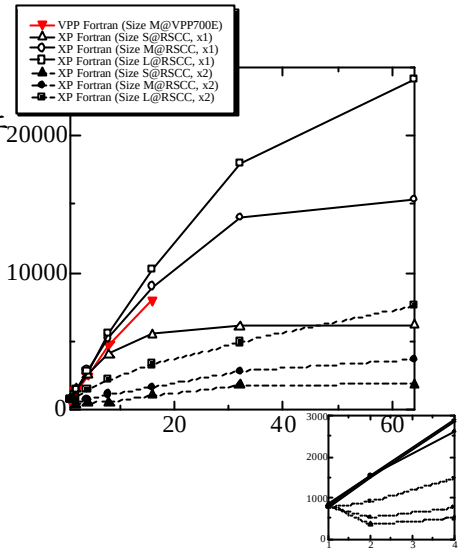
Advanced Center for Computing and Communication

MPI Results (Size XL)



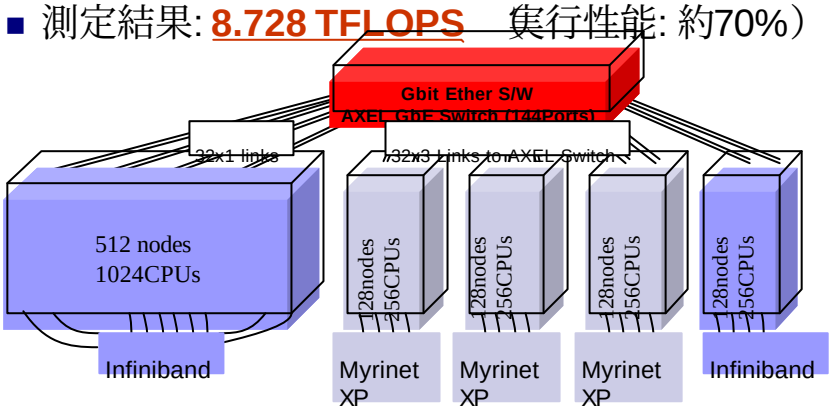
XP Fortran Results

- VPP FortranではDTU（ハードウェア）が行っていた処理を、XP Fortranではソフトウェアで処理
  - 1node内にデータ通信用のスレッドが必要
- 2CPU/1nodeでジョブを実行するより、1CPU/1nodeで実行する方が早い
- 1CPU/1nodeで実行すると、VPP700Eより高速



LINPACKベンチマークの測定

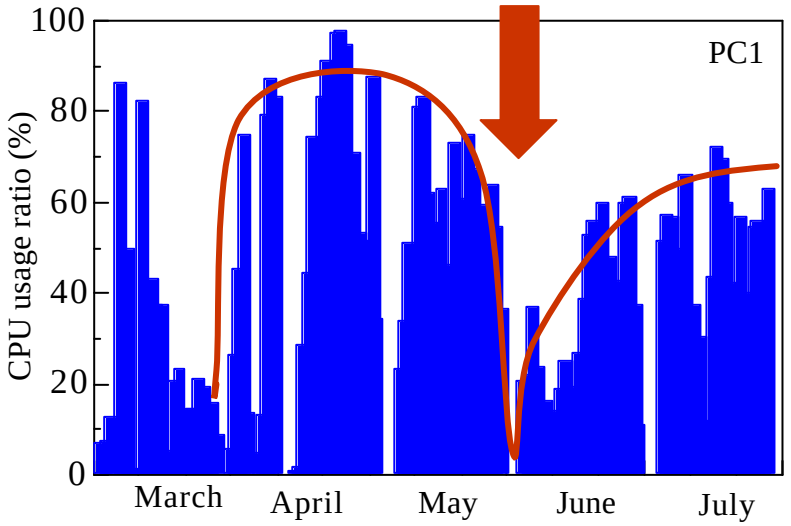
- 1024node(2048CPU)を使うために、Cluster of clusterでの実行
  - Linuxクラスタ内は高速インターコネクトを利用、各クラスタ間は、16node単位でGbit Etherを利用
- 測定結果: **8.728 TFLOPS**　実行性能: 約70%）



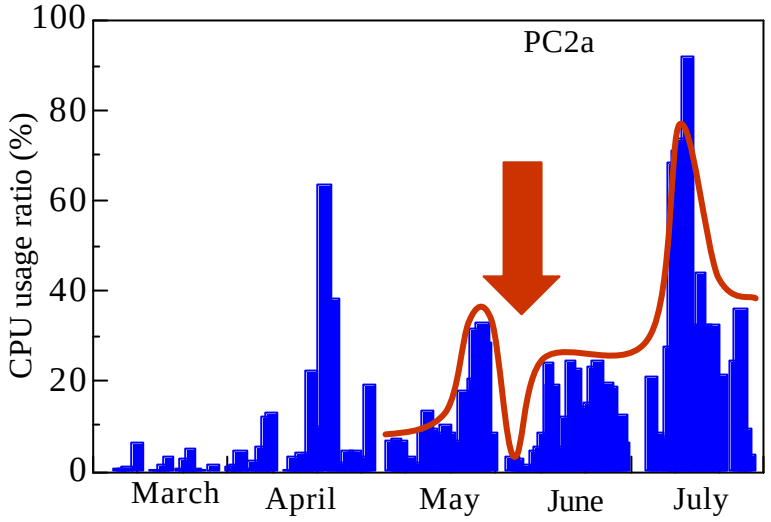
利用状況

Advanced Center for Computing and Communication

CPU利用率（PC1 :高並列ジョブ）

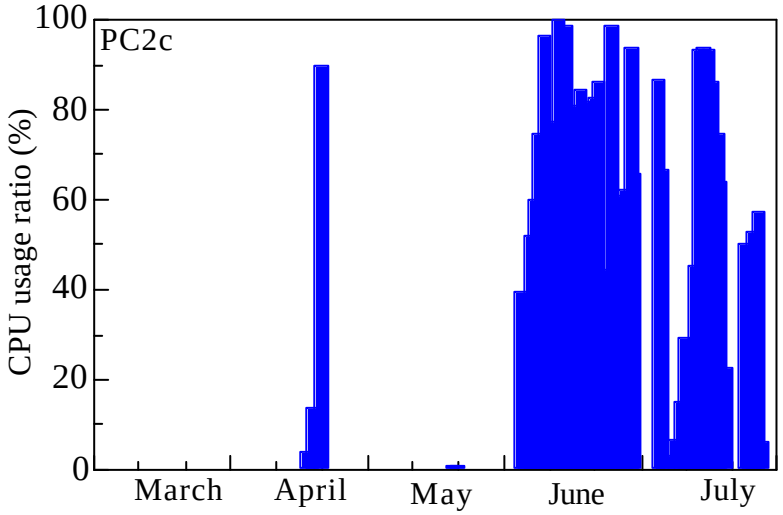


CPU利用率（PC2a：バイオ＋低並列ジョブ）

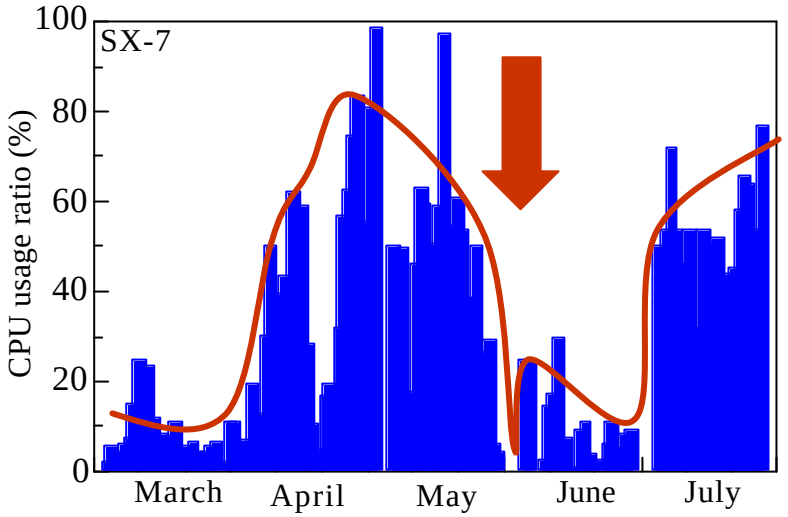


Advanced Center for Computing and Communication

CPU利用率（PC2c：高エネルギー実験データ解析）



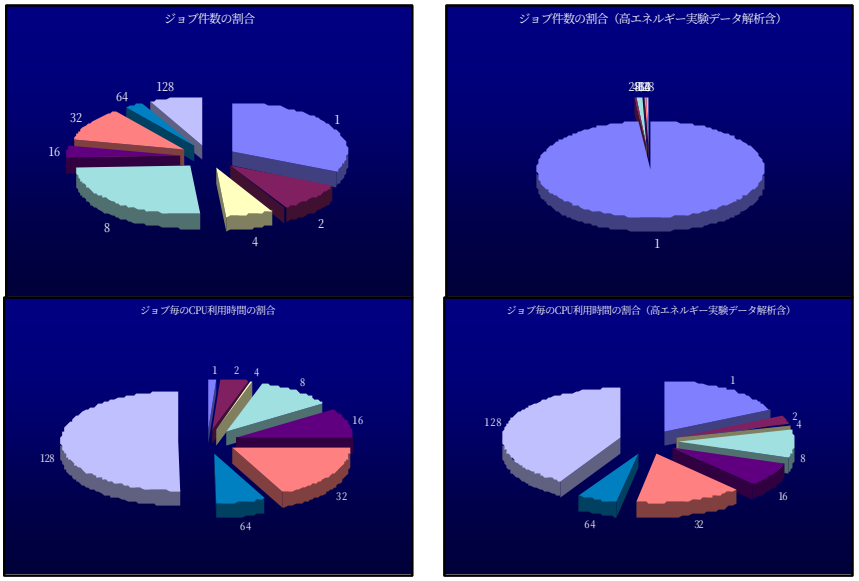
CPU利用率（SX-7）



テスト運用から本運用

- 3月から5月末までテスト運用
  - 自分のプログラムに必要なCPU時間を見積もるために利用
  - CPU時間に特に制限無し
- 6月からは本運用
  - ユーザーはCPU時間を申請して利用
  - 課題審査委員会で審査
  - 研究成果が義務
  - 許可されたCPU時間まで利用可能

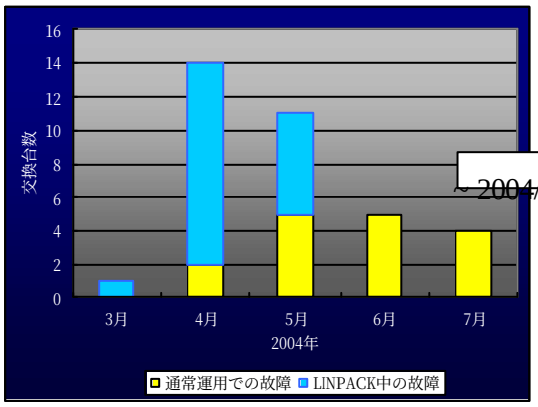
利用CPU数別ジョブ件数とCPU時間



ハードウェアの故障

- 全システム停止なし
- LINPACK測定中のハード障害：

3月1/1、4月12/14、5月 6/11（トータルで、19台/33台）



~~2004/Jul/24~~

予防交換
- Blade server nodes:128/128
- InfiniBand ケーブル:512/1024

理研での小規模クラスタの実績
8台のPCクラスタ、5年で約5回
5回×1024台/8台/5年/12ヶ月=10.7台/月

16台のクラスタ、2年で5回
5回×1024台/16台/2年/12ヶ月=13.3台/月

64台のクラスタ、2年で19回
19回×1024台/64台/2年/12ヶ月=12.67台/月

まとめ (1)

- 計算機センターマシンとして、日本で初めてLinuxクラスタを導入
- 複合システムをカバーする機能と利便性
  - 編集、コンパイル、実行をフロントエンドで完結
  - バッチジョブの実行はキューを指定
  - ポータルサイトの構築
  - ジョブの実行前後にファイル転送
    - Linux クラスタではローカルHDDでのI/O
- 旧システムからの移行
  - ベクトル化されたプログラムはSX-7で実行
  - VPP FortranはXP Fortranでそのまま実行可能
- BMTによる実効性能
  - LINPACK TOP 500で世界7位

Advanced Center for Computing and Communication

まとめ (2)

- ハードウェア障害
  - 月に4台程度は、予想以上に少ない
- CPU利用率
  - システム導入時にしては比較的高い？
  - 今後、ユーザー教育・サポートを強化して利用率を上げる予定
    - MPI、スカラーチューニング、可視化などの講習会の実施
    - チューニング、並列化のサポート
- バッチジョブ・スケジューラ
  - 逐次ジョブを1nodeで2本同時実行は今秋に対応予定
  - 大量のジョブ対応、フェアシェア機能も必要
- チェックポイント・リスタート機能
  - 全てのジョブを対象とするのは、現状では不可能
  - カーネルレベルでのチェックポイントが必要