



理研・新スパコン・システム (RICC) の紹介

重谷隆之

理化学研究所 情報基盤センター



独立行政法人 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、ライフサイエンス、脳科学...
- 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)





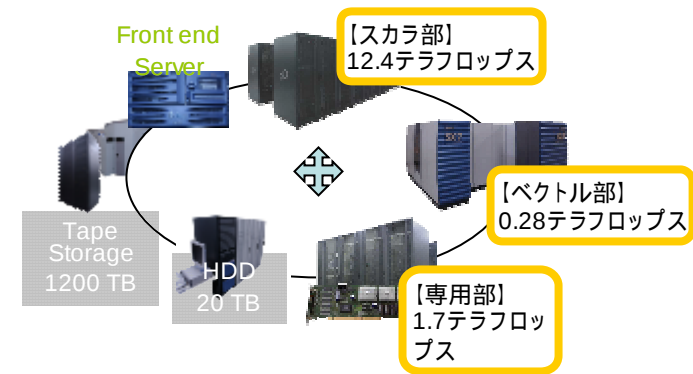
情報基盤センター

- 全理研を対象とする研究支援部門
 - 理研の研究をHPC技術で支援・サポート、推進
- 情報基盤センター内のチーム
 - ネットワーク・チーム
 - 所外、拠点間、所内ネットワーク・インフラの整備、運用
 - メール・Webサーバ運用、ネットワーク・セキュリティ
 - HPCチーム
 - **共同利用計算機(スパコン)**の設計、運用：スパコン資源を効率的に使うようにマネジメント
 - ユーザー・サポートチーム
 - 図書・研究成果も



RIKEN Super Combined Cluster (RSCC)

- 2004年3月テスト稼働, 2004年6月本格稼働(TOP500 7位)
- PCクラスタ・ベクトル・専用計算機をユーザ利用に対してシームレスに結合
- 2005年度「日本産業技術大賞」文部科学大臣賞受賞
- 理研の研究者のための計算資源(課金はないが, 利用には審査が必要)
- PCクラスタを主計算資源に採用
 - 計算機センターの共同利用計算機でPCクラスタを日本で初めて採用
 - ベクトル並列(VPP700E/160)からスカラ並列への転換
 - OS等ソフトウェアのプロプライエタリからオープンスタンダードへの転換
 - InfiniBandを採用した初めての大規模クラスタ
 - 増大する計算(研究)規模に対応
 - ただし, 利用するCPU数の増加による並列化は必須
 - ライフサイエンス・ユーザ(スパコンの新規ユーザ)の利用拡大
 - 分断されたクラスタを一括管理・高機能スケジューリングするジョブ・スケジューラの開発



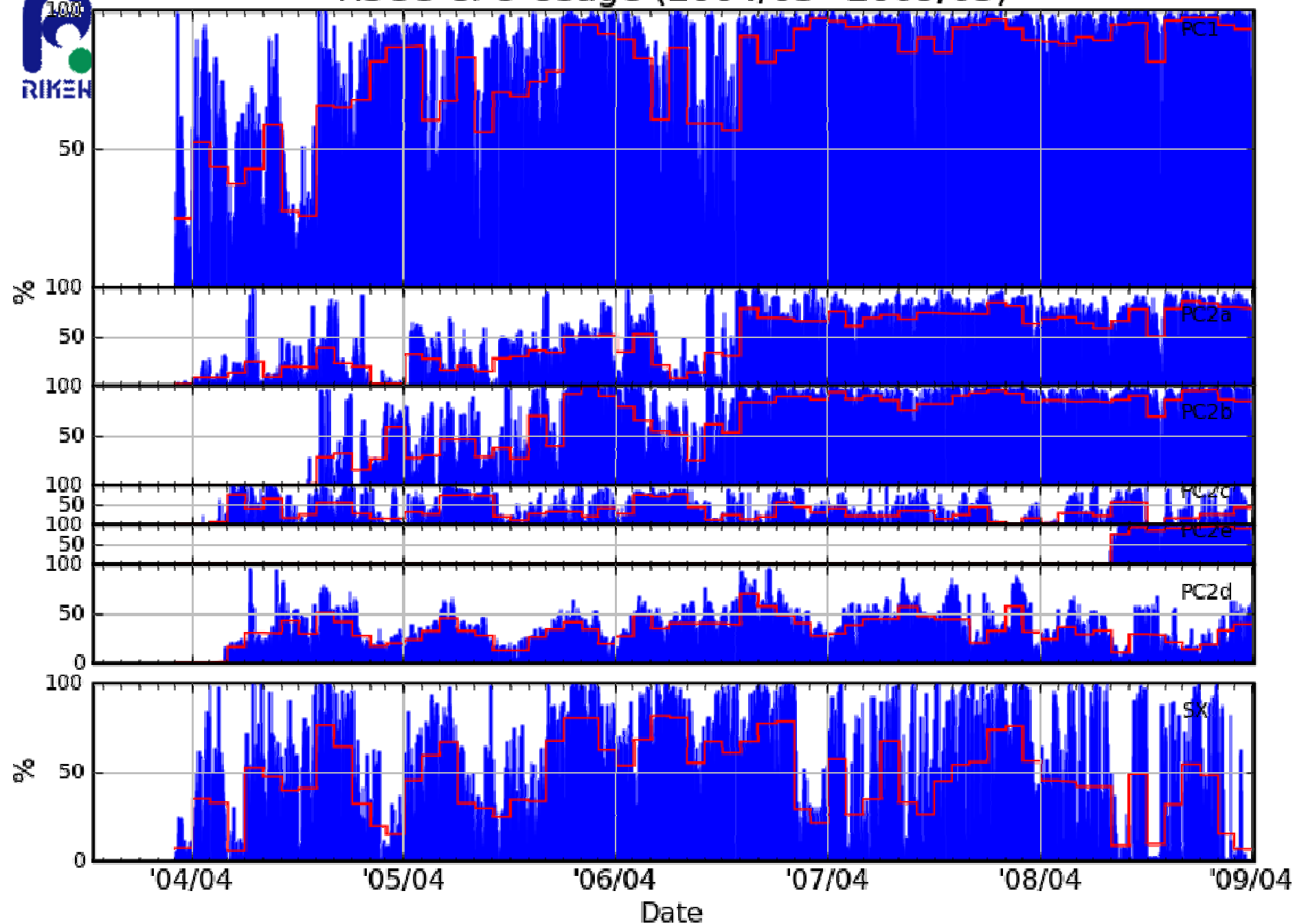


RSCCの特徴

- オープン・スタンダード
 - オープンソースソフトウェアの導入
 - Linux採用による様々なサイエンティフィックなオープン・ソース・アプリが利用可能に
 - 特にライフ系・ナノ系のユーザの利用が多い
 - スクリプト言語系の利用者も増加
 - スパコンセンター利用に不慣れなユーザへのケア
 - Webポータルを利用したスパコンの利用を推進
 - 並列化・チューニングに向けた講習会やチューニングサービス
 - 一般参加可能な講習会を実施
 - 並列化やチューニングをセンターで請け負って実施
- PCクラスタが計算機センターの運用に耐えられることを実証
 - 導入当初, 本当にPCクラスタがスパコンセンターの計算資源として本当に耐えられるのか, という声・不安があった
 - PCクラスタの運用中の全系停止はないが、クラスタなので, 個々のPCの障害は絶えない
 - ソフトウェアの運用時障害はまだ枯れていないが、ハードウェアの障害数は予想以下



RSCC CPU Usage (2004/03~2009/03)





RSCCの良かった点(ユーザの声)

- 演算性能が格段に上がった(使えるCPU数が増えた)
- チューニング(ベクトル化)をしなくても、ある程度の性能
- Webポータルによる利用で使ってみる気になった
- IA系CPUとLinuxでありOpen Sourceのツールやアプリ等が使えた
- スクリプト言語系が使えて良い
- ステージング機能を採用しているので、ノードが故障しても元データが消えなかった



RSCCで良くなかった点(ユーザの声)

- 使えないISVアプリがある
 - Kernel VersionとlibcのVersionが違っている
 - OSのアップデートはしないのか？
- エラー表示の意味が分からない
 - 変なコードを表示するだけで、意味が分からない
- MPIだけでなくPVMが使えるようにして欲しい
 - ISVアプリで使っているものがある
 - システム上の問題で難しい
- ディスク領域が少ない、メモリが小さい
- ステージング機能は使いにくい



RSCCで良くなかった点(管理者として)

- ログ項目不足
 - ジョブの利用メモリ、HDD、性能(浮動小数点カウンターなど)……
- ネットワーク(インターコネクト)構成
 - 計算用ネットワーク(IB, Myrinet)をクラスタ毎に分断
 - コスト対効果では有益だったが、同時に全システム利用は難しい
 - クラスタ毎に負荷の偏りが発生
- ストレージ容量が少なかった
 - メモリ:HDD(ホーム領域):テープアーカイブ 3:20:200
- ステージングはコスト・システムの的には正解だったが、ユーザビリティは賛否両論
- 制御ノードでの実行ジョブの管理
 - 1人で同時に数千本のジョブ投入をするケースも
 - 制御ノードの負荷が高くなる
- ジョブ・マネージャの制御パラメータ不足や柔軟性の欠落
 - 「少ないCPU数を大勢で使う」から「大量のCPUを上手く使う」へのパラダイム変換がうまくいってなかった
 - 商用アプリにはライセンス数制限も
 - ジョブ・スケジューラの開発で回避
 - スケジューリングポリシーや様々なユーザジョブ特性を念頭に開発
 - 様々なリソース要求によるジョブの優先順位の決定, フェアシェア機能, バックフィル機能
 - リソース: 特殊なハード(MDGRAPE-3など)の有無, ISVソフトのライセンス数, 並列度の大小, 時間の長短, メモリ量, ジョブ間依存関係
 - 稼働率の劇的な向上, ユーザに物理的な資源状態を意識させない, ジョブ実行待ちのユーザ間の平準化
 - 90%を超える稼働率を達成

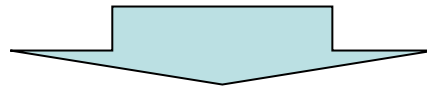


次期システムの検討



システム構築で最初に考えたこと

- TOP500のトレンドからも中長期的に見て、並列数(プロセッサ数)の増加は避けられない
 - ネットワーク帯域(FatTree)をそれなりに維持したシステムを構築するにはお金が掛かる
 - そもそもシステム全体でFBB (Full-Bisection Bandwidth)が必要か
- システムとしてTOP500上位を狙うのは難しい
 - LINPACK性能のみを追うのが理研のシステムとして正しいのか？
- もしかすると(GPGPUのような)アクセラレータが主流に？
- もちろん利用者の要求要件も聴かないと



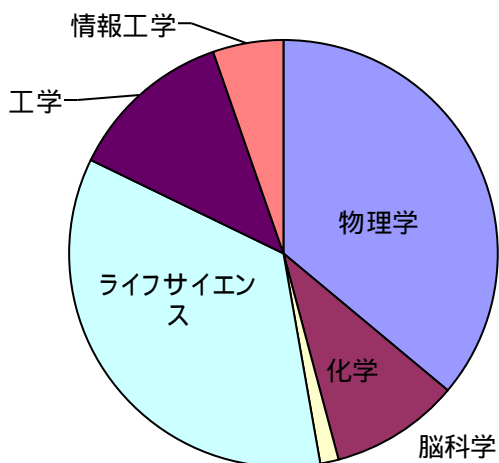
- トrend、利用者の要求、コストなどを総合的に検討

利用者(研究者)からの要望

- 研究分野

- ゲノム情報データの解析
- 古典分子動力学シミュレーション
- 大規模ゲノムデータセットに対するパターン検索
- 第一原理分子動力学計算(VASP)
- 天体物理学
- 重イオン加速器での衝突実験で発生するデータ解析処理
- 量子電気力学
- 第一原理量子化学計算(Gaussian)
- 計算力学シミュレーション
- 可視化

研究分野毎の割合



- CPU (Core)

- 多ければ多いほど良い

- メモリ

- 1GB/Core以上欲しい
- 3GB/Core以上欲しい
- 200GBを1プロセスで扱いたい

- インターコネクト

- InfiniBandでFat-Treeトポロジで十分な帯域が必要
- あまりインターコネクトにお金をかけるべきではない

- ローカルHDD

- 高速なI/O性能が必要

- オンラインディスクストレージ

- 多ければ多いほど良い
- Home/Data領域として計算ノードにマウントが必要

- テープストレージ

- 3PB以上必要
- 500MB/s以上のI/O性能が必要

- 拡張機能

- GPU/アクセラレータが使いたい
- MDGRAPE-3も継続してほしい

- アプリ・ライブラリ

- 4倍精度計算が高速に行えるライブラリを有する
- Gaussian/ANSYS/Amber等が動くこと

- 外部ネットワーク

- FW経由とスイッチACLによる帯域確保



次期システム要件の整理

- 利用者(研究者)からの要望
 - 計算性能、メモリディスク容量の増強
 - インターコネクト(計算用ネットワーク)性能への要件は両極端
- RSCCからの改良・拡張
 - RSCCの問題点・拡張要望点を反映
 - トータルな演算性能よりも研究分野での成果が出せる構成
 - 利用者ユーザビリティの向上
- 情報基盤センターとして
 - 理研の研究者がシステムトレンドから外れないように
 - 次世代スパコンに向けたプログラム開発
 - 新しいユーザ領域の開拓
 - 実験データ処理とスパコンとの連携拡大を模索
 - XFELや次世代シーケンサーやDNAマイクロアレイのデータ処理など
 - アクセラレータの利用形態とその応用利用
 - スパコンセンターとしての運用にチャレンジ

次期システムのコンセプト

基本コンセプトはRSCCを継承しつつ、
新たな要望やこれからの傾向をキャッチアップ

各研究室では用意出来ない研究開発のための計算資源

実験のデータ処理や
実験系研究者の
サポート

次世代スーパー
コンピュータに
向けた開発環境

新しい計算技術
に挑戦

データ処理との連携強化
ストレージ性能強化

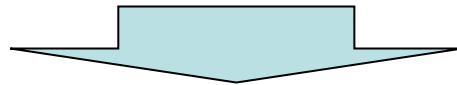
大規模並列に対応
計算能力強化

GPUアクセラレータの
利用と応用検討



計算用ネットワーク設計の考え方

- アプリケーションとして性能を出す3つのパターンを想定
 - 今現状の並列アプリケーションのプロダクション実行
 - 大規模並列アプリケーション開発
 - 本質的にネットワーク性能が不要なアプリ
- それぞれ並列度と通信パターンを考慮すると
 - 一般的な並列ジョブ: 並列度64～128ぐらい、通信量もある程度
 - 大規模並列アプリ開発: 並列度が最低1024ぐらい、通信量は少なく
 - 本質的に通信を減らさなければ、高並列までスケールしない
 - ネットワーク性能が不要な場合: 並列度はいくらでも
 - 全系システムでジョブ実行がいつでも出来るように



- 一般的な並列ジョブをリーフ・スイッチ配下に閉じ込めて、上位帯域を絞る方向
- システム全体でのFBBは諦める



新システムの概要



新システム構成 (RIKEN Integrated Cluster of Clusters: RICC)

【システム構成】

PCクラスタ + 大容量メモリ計算機 + アクセラレータ

演算性能: 8.5倍
メモリI/O性能: 2.5倍

【超並列PCクラスタ】 1024Nodes
96.0TFLOPS, 12TB(mem), 435TB(hdd)
12GB/Node, DDR IB × 1/Node

【多目的PCクラスタ】 100Nodes
9.3TFLOPS, 2.3TB(mem), 25TB(hdd)
24GB/Node, DDR IB × 1/Node
PCI-ex16レーン × 1

アクセラレータ × 100

【分子動力学専用計算機】
64TFLOPS
ホストノード: 32Nodes
32GB/Node, DDR IB × 1/Node

【大容量メモリ計算機】 1Node
0.24TFLOPS, 512GB(mem)
PCI-X, 10GbE

メモリ容量を2倍

実験データ



利用者

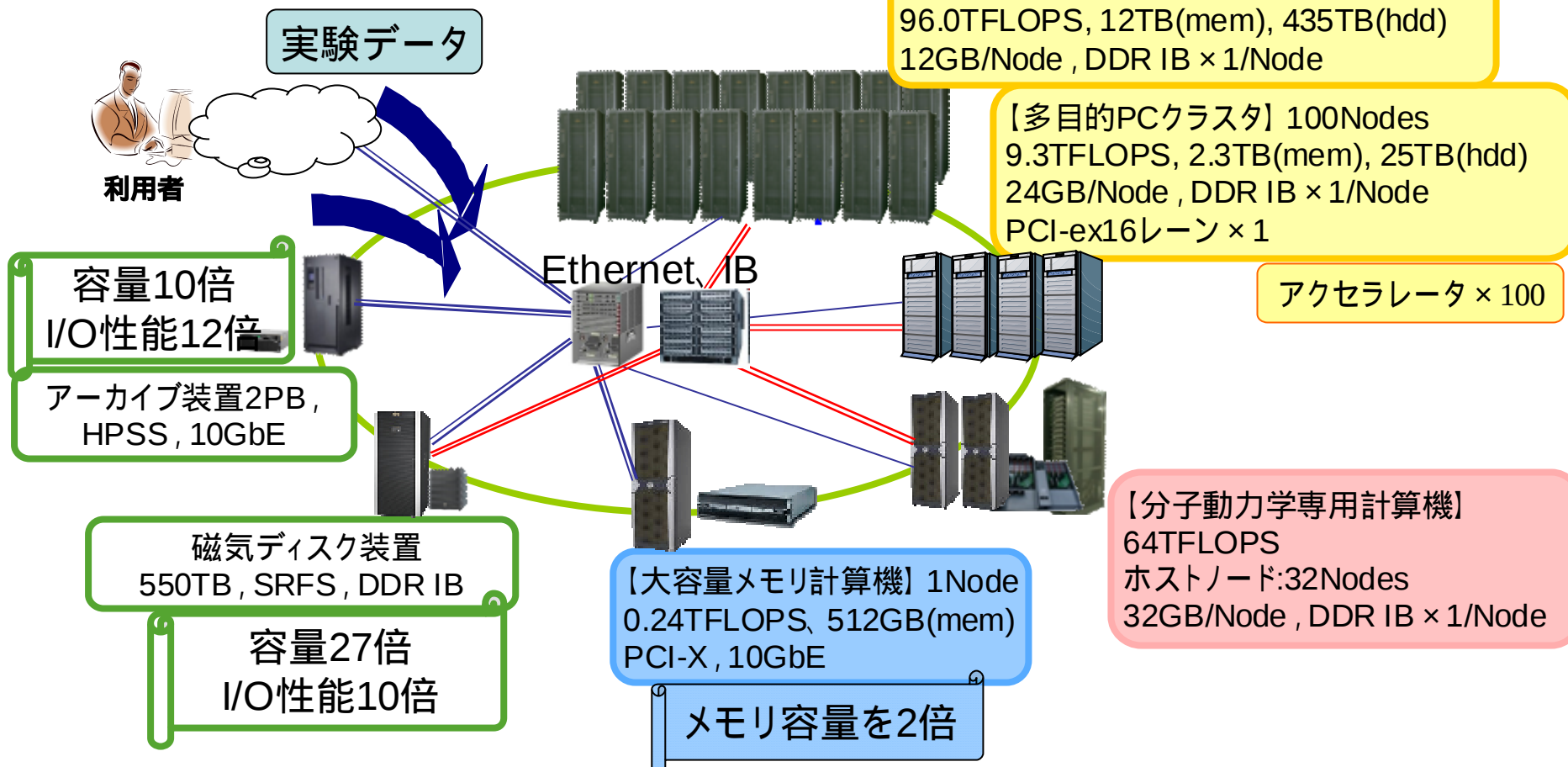
容量10倍
I/O性能12倍

アーカイブ装置2PB,
HPSS, 10GbE

磁気ディスク装置
550TB, SRFS, DDR IB

容量27倍
I/O性能10倍

Ethernet, IB

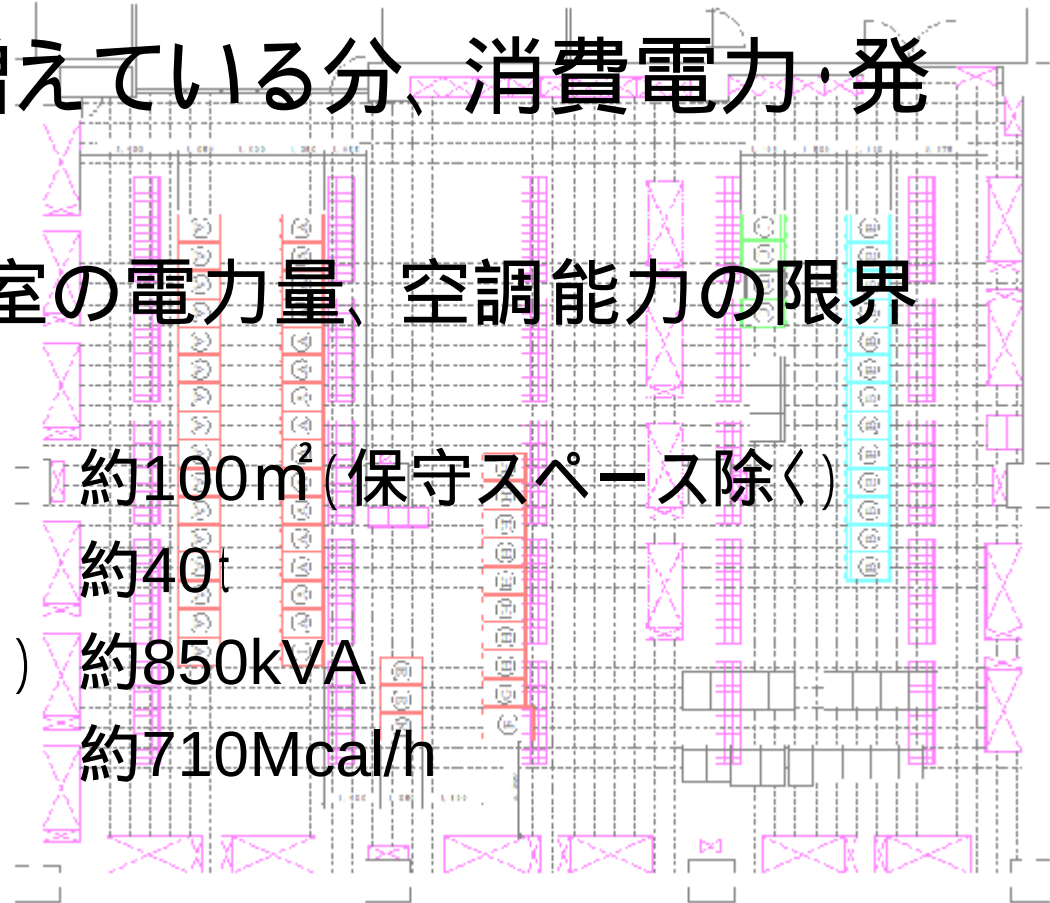


システム緒元

- サイズは500m²のマシン室で十分
- 性能に対する消費電力は小さくなった
- トータル性能が増えている分、消費電力・発熱量も増加

ーもう少しでマシン室の電力量、空調能力の限界

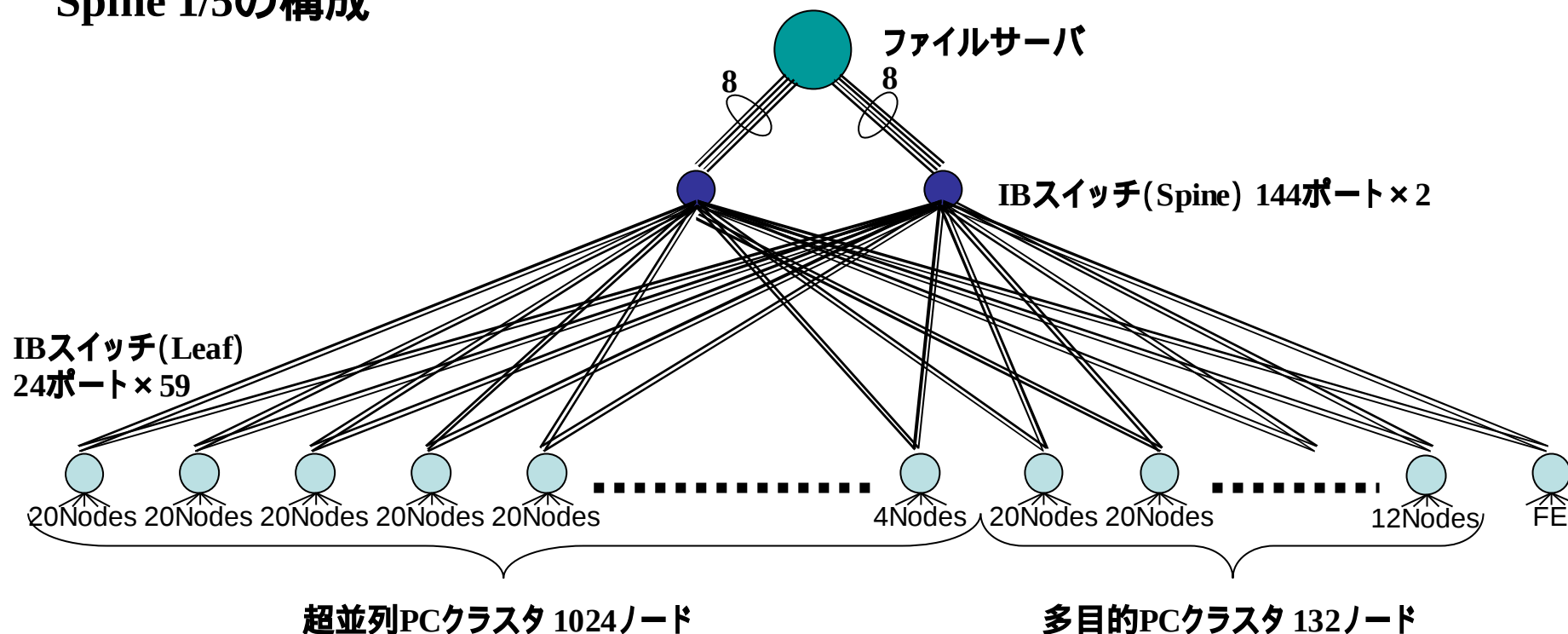
- 設置面積 約100m² (保守スペース除く)
- 重量 約40t
- 消費電力(ピーク) 約850kVA
- 発熱量 約710Mcal/h





インターコネクト (InfiniBand) 構成

FBB構成よりもLeaf 2/3,
Spine 1/5の構成





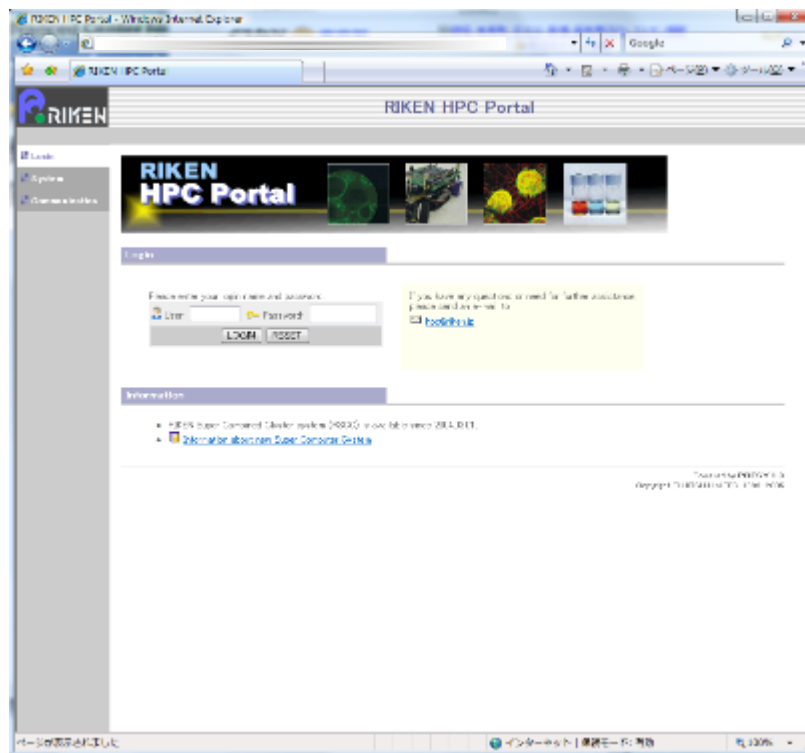
RSCCでの問題点への対応

- ジョブ・スケジューリング・ソフトウェアの機能強化
 - RSCCで開発したもの + α の機能
 - マルチコア・システムでの効率的なジョブ管理・リソース管理システム
 - ネットワーク・トポロジと運用方針を踏まえたノード・アロケーション管理
 - 単一CPU利用ジョブの同時数万本投入にも対応
 - サーバ障害時のフェイルオーバーに対応
- ユーザへの助言のためのロギング強化
 - ユーザジョブのリソース情報の把握
- ISVアプリケーションの可搬性の向上
 - 最新バージョンのRedHat ELを採用
 - PVMも利用可能
- ジョブ実行中のファイルI/O
 - ステージングとダイレクトアクセスの両立
 - ローカルHDDのコストパフォーマンスとストレージシステムの性能を両立させる構成



Webポータルの継承と新規サービス

- スパコンをWeb画面から利用
- ポータルのサブセットを携帯電話からも利用可能に



システム状態の表示
ジョブ状態の表示
結果確認
ジョブ操作
など



WebシステムとWebサービス

- Webサービスでの利用を促進する基盤を提供
- リモートからのファイルやジョブのハンドリングをWebサービスとして提供
 - Webシステムから一歩前進
 - 固定的なサービスではなく、ユーザ要望を広く取り入れることが可能
 - 自分のPCでワークフロースクリプト(Perlなど)を作成して、ファイル転送、ジョブ実行、結果取得等が可能
 - ユーザが自分でPortalを作成するためのツール
 - 利用の柔軟性, サービスの柔軟性の向上
 - 利用シーン想定
 - PCやサーバ上の処理とスパコン上の処理の連携利用
 - グループで独自のWebポータルを作成して、重いデータ処理をスパコン側に処理を依頼する



まとめ

- 新システム (RICC) のコンセプト
 - 超並列型アプリケーション (次世代スパコン) に向けた開発
 - スーパーコンピュータと実験データ処理の融合のサポート
 - 新しい計算パラダイム (アクセラレータ) への挑戦
- 新システムのハード・ソフト
 - 異なる特性の計算資源の結合 (ハードウェア・ソフトウェア)
 - ジョブスケジューラの機能改良による使い勝手, 利用効率の向上
 - 新しい使い方・サービスを展開
 - アクセラレータの導入、携帯端末への対応、Webサービスなど