次期システム（RICC）の導入に向けて
“RSCC リプレースについて”

黒川 原佳
理化学研究所 情報基盤センター

RSCCおさらい(1/3)

- **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，大資源利用は審査あり）
- **PCクラスタを主計算資源に採用**
  - 日本初の計算機センターのサービスマシンでPCクラスタを採用
  - 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
  - ライフサイエンスユーザ(スパコンの新規ユーザ)の利用拡大を狙う
  - 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スケジューラの開発

RSCCおさらい(2/3)

- **オープン・スタンダードの得点と利用環境の構築**
  - 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が増加
  - ライフサイエンスや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
  - 従来と変わらぬ利用環境を構築
    - ただし，利用するCPU数の増加による並列化が必須に
  - 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

RSCCおさらい(3/3)

- **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クラスタがスーパーコンピューティングセンターの計算資源として本当に耐えられるのか．という不安はあった
  - 現状，PCクラスタの運用中の全系停止はない
  - ソフトウェアの運用時障害はまだ枯れていない
  - クラスタなので，個々のPCの障害は絶えない．

RSCCは良かった？（ユーザからの反応）

- 単純に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，そこそこの性能で動作する
- コンパイル時間が格段に短くなった
- 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いやすい
- スクリプト言語系が使えていい
- データ処理・データベース処理も普通に行える

どこがイマイチだった？（ユーザ編）

- **ISVアプリが使えないことがある**
  - 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**
  - 変なコードを掃くだけで，何のことだかさっぱりわからない
  - まだまだこなれてない一例
- **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**
  - . . . . .
- **ステージングはやっぱり使いにくい**
  - . . . . .

どこがイマイチだった？（システム屋目線）

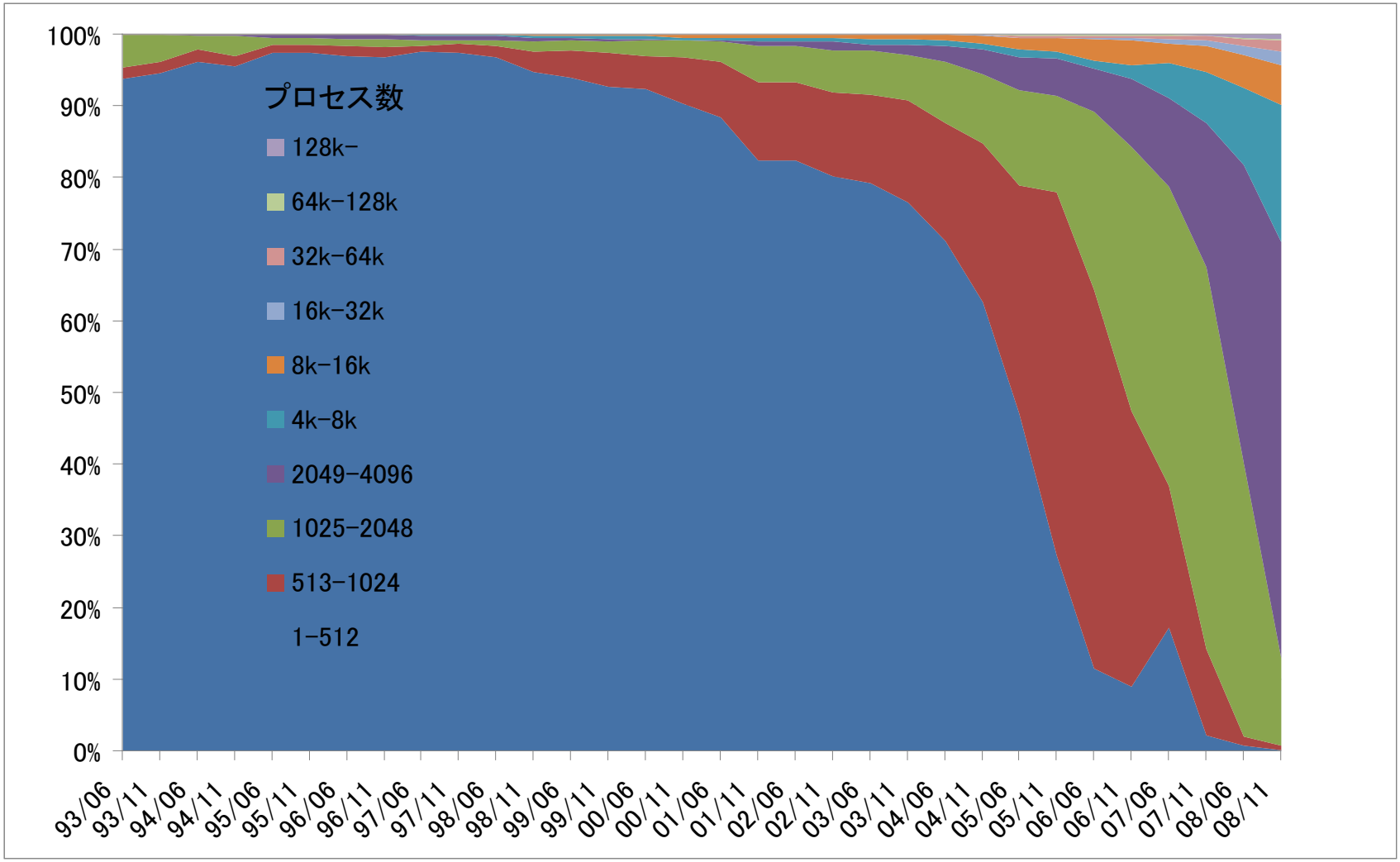
- **制御ノードでの実行ジョブの管理**
  - バルクジョブの対応とか
  - 制御ノードの負荷が高くなる
- **ジョブマネージャの制御パラメータ不足や柔軟性の欠落**
  - CPUが貴重でそれを上手く使うのと，大量のCPUを上手く使うのは違うというパラダイム変換がうまくいってなかった
- **ログ項目がイマイチ，ログ出力ない．．．**
  - 特に性能に関わるログは皆無だった気がする
- **ネットワーク構成も良い点・悪い点があったなぁ**
  - 計算用ネットワーク(IB,Myrinet)をクラスタ毎に分断するのはコスト対効果では有益だったが，全系を利用するのはやはりバリアがあった
- **ストレージがかなり少なめだった**
- **データI/O帯域も細かったなぁ**
  - ステージングはコスト・システム的には正解だったが，ユーザビリティはほどほどだった

次期システムについて

準備段階

システムトレンド

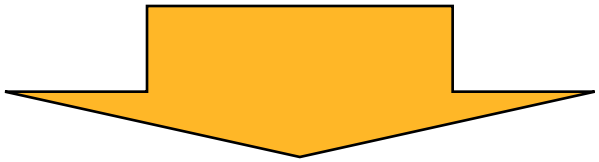
- 中長期的にシステムを見据えて並列数の増加は避けられない



引用:TOP500

システム構築について(システム屋目線)

- **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)が必要なのか
- **システムとしてTOP500上位(10位以内)を狙うのは難しい**
  - ここ2，3年で年間でTOP20ぐらいは性能向上が加速している
  - 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
- **もしかするとアクセラレータが主流の世界に？**



- **利用者要件とシステムトレンドおよびコストのバランスが非常に難しい**

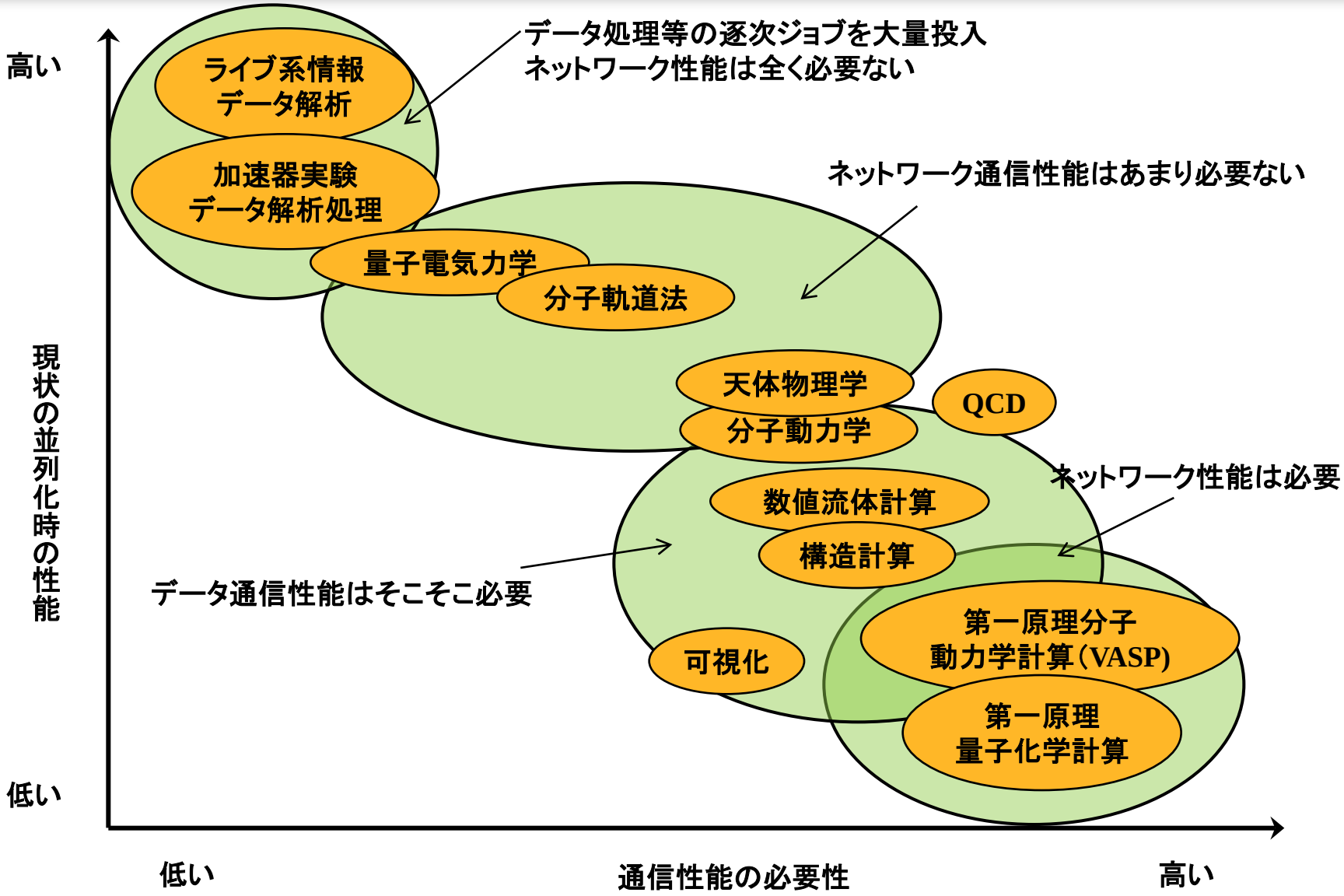
利用者からの要件

- **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

- **ユーザ要件要約**

- CPU（Core）
  - 多ければ多いほどよい
- **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**
  - 存在が必要で高速なI/O性能が必要
- **オンラインディスクストレージ**
  - Home/Data領域として計算ノードにマウントが必要
  - 広帯域で複数ノードからのランダムアクセスに強い必要がある
- **テープストレージ**
  - 3PB以上が必要
  - 500MB/s以上のI/O性能が必要
- **拡張機能**
  - GPU/アクセラレータが搭載出来ること
  - MDGRAPE-3が利用可能であること
- **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

利用者の研究分野

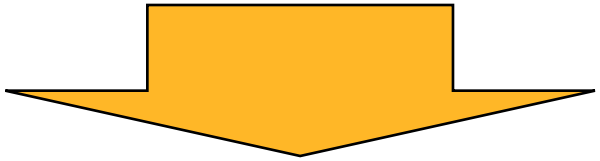

高い
低い
現状の並列化時の性能
低い
通信性能の必要性
高い
データ処理等の逐次ジョブを大量投入
ネットワーク性能は全く必要ない
ライブ系情報
データ解析
加速器実験
データ解析処理
ネットワーク通信性能はあまり必要ない
量子電気力学
分子軌道法
天体物理学
QCD
分子動力学
ネットワーク性能は必要
数値流体計算
構造計算
データ通信性能はそこそこ必要
第一原理分子
動力学計算(VASP)
可視化
第一原理
量子化学計算

次期システムへの要件の整理

- **ユーザ要件の整理**
  - 真っ向から衝突しているのは計算用ネットワーク
- **RSCCからの改良・拡張**
  - RSCCの問題点・拡張要望点から反映
  - 演算性能よりも研究分野での成果が出せる構成
  - 更に言えば，理研の研究者のユーザビリティの向上
- **情報基盤センターとして**
  - 上記の項目から実現可能な構成を検討
  - 理研の研究者がシステムトレンドから外れないように
    - スパコン黎明期やオンリーワンの計測器ならいざ知らず，シミュレーション用計算ハードウェアのガラパゴス島にしてはまずい
  - 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ーパーコンピュータの連携の拡大を模索
      - 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にも広げられる可能性を高める
  - 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，どのような利用形態を考えていくのか
  - オープンソース・ツールorアプリケーションorライブラリの導入

ネットワークの構成の考え方

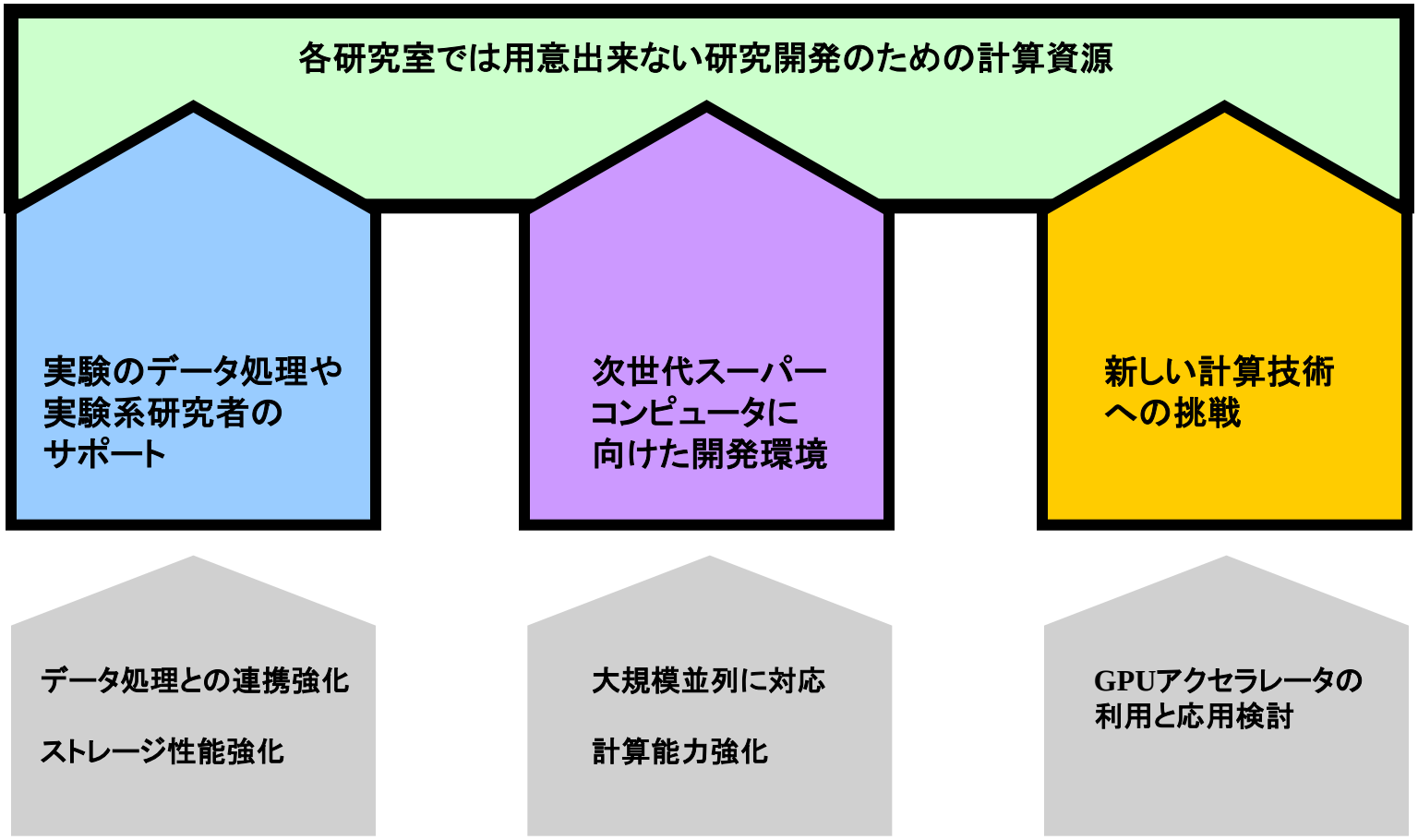
- **アプリケーションとして性能を出す3つのパターンを想定**
  - 今現状の並列アプリケーションのプロダクション実行
  - 次世代スパコンに向けた並列アプリケーション開発
  - 本質的にネットワーク性能が不要なアプリ
- **それぞれ並列度と通信パターンを考慮すると**
  - 現状のプロダクション：並列度128ぐらいで割と通信を行う
  - 次世代に向けた開発：並列度が最低1024ぐらいで通信は少なく
    - 本質的に通信を減らさなければ，高並列までスケールしない
  - ネットワーク性能が不要な場合：並列度はいくらでも
  - 全系システムでジョブ実行がいつでも出来るように



- **プロダクションをリーフに閉じ込めて，上位帯域を絞る方向**
  - ジョブスケジューラのノードアロケーションと3つのユーザモードの切換で対応

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**



システム詳細

ハードウェア編

次期システム構成（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

実験データ

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

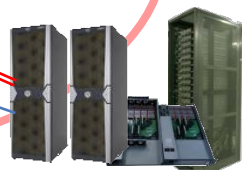
【多目的PCクラスタ】100Nodes
9.3TFLOPS, 2.3TB(mem), 25TB(hdd)
24GB/Node，DDR IB×1/Node
PCI-ex16レーン×1

アクセラレータ×100

容量20倍
I/O性能12倍

Ethernet、IB

アーカイブ装置4PB，
HPSS，10GbE



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

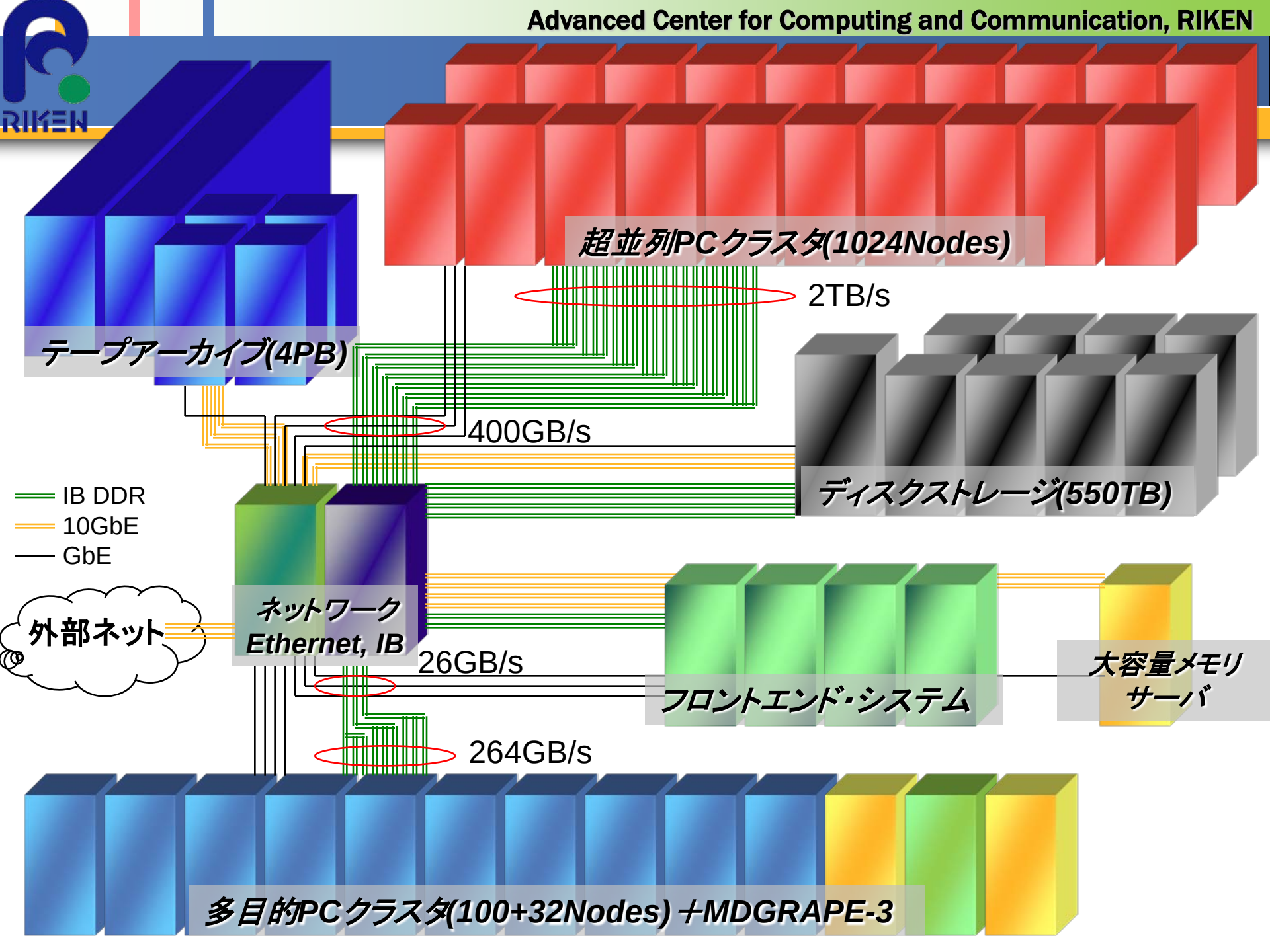
磁気ディスク装置
550TB，SRFS，DDR IB

容量27倍
I/O性能10倍

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

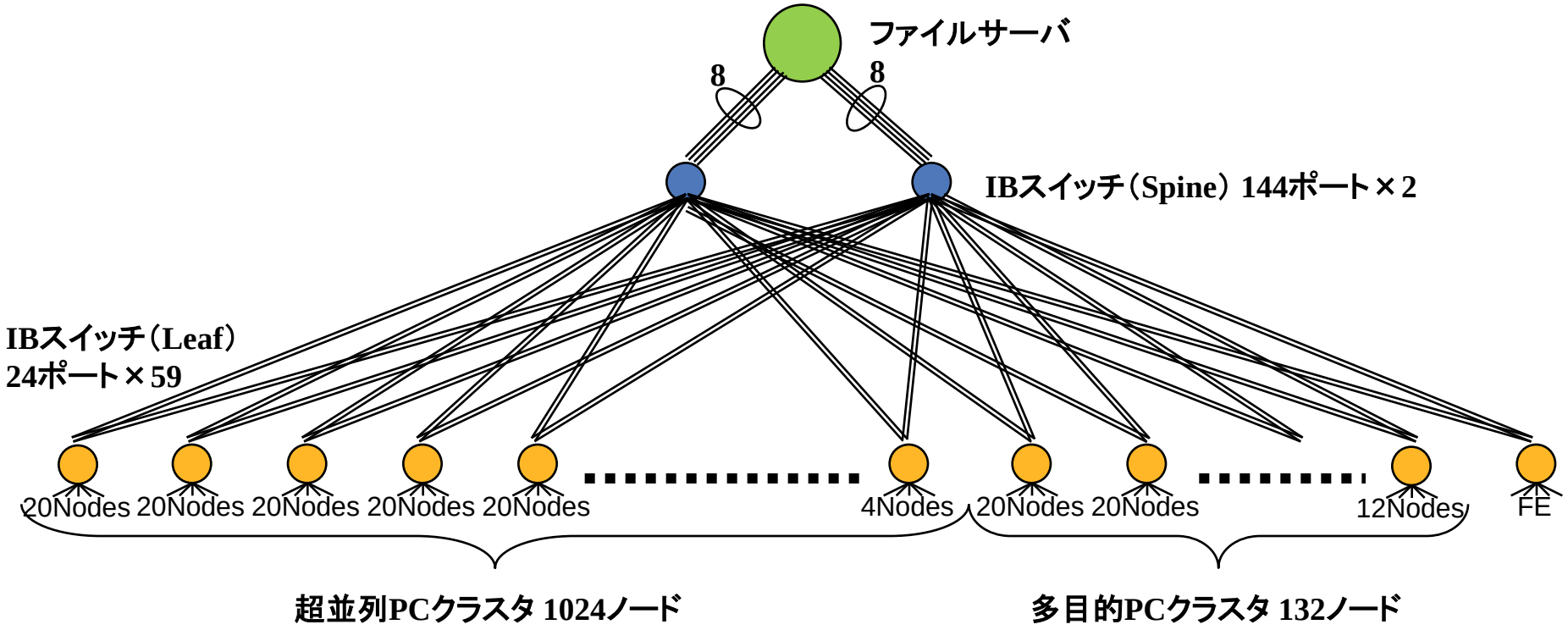
メモリ容量を2倍

総演算性能：106+100+64TFLOPS


超並列PCクラスタ(1024Nodes)
2TB/s
テープアーカイブ(4PB)
400GB/s
ディスクストレージ(550TB)
IB DDR
10GbE
GbE
外部ネット
ネットワーク
Ethernet, IB
26GB/s
フロントエンド・システム
大容量メモリ
サーバ
264GB/s
多目的PCクラスタ(100+32Nodes)＋MDGRAPE-3

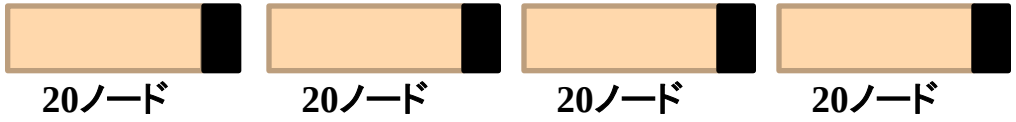
InfiniBandネットワーク構成

FBB構成よりもLeaf 2/3,
Spine 1/5の構成

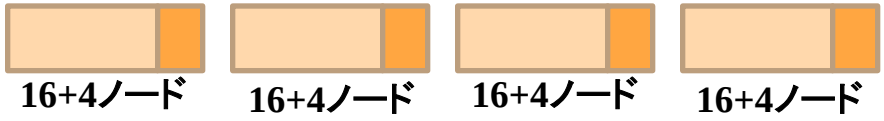


ノードアロケーション指針

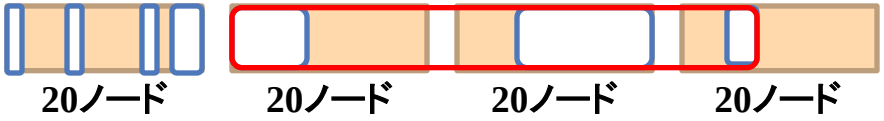
物理配置



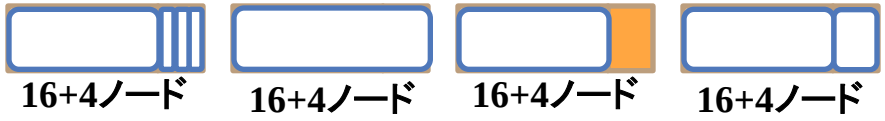
ノードアロケーションの
考え方



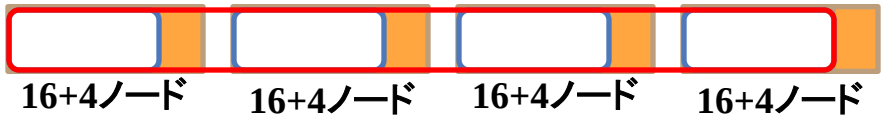
ネットワーク性能が不要



小規模プロダクション

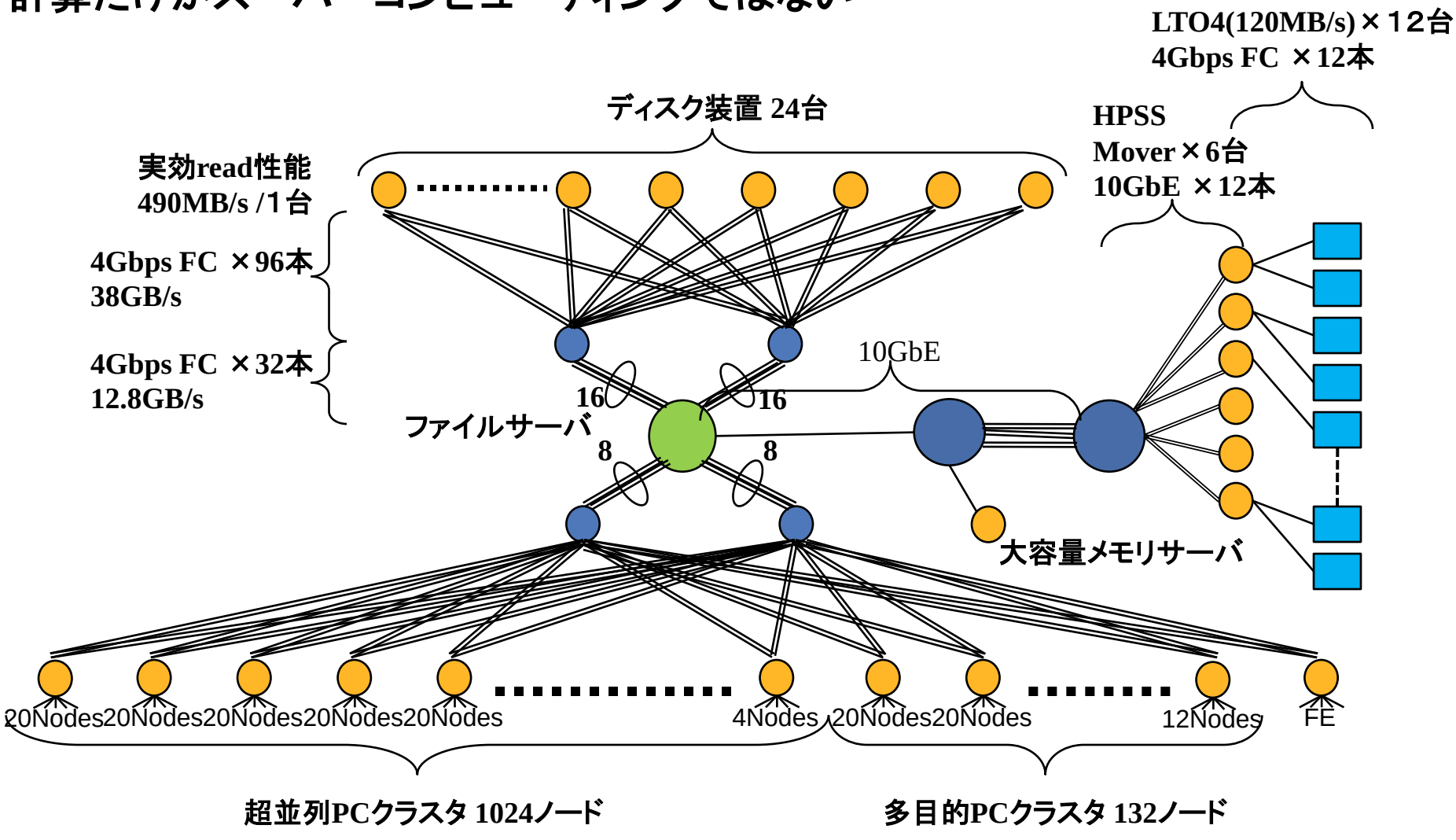


大規模並列ジョブ



データ処理系ネットワーク構成

計算だけがスーパーコンピューティングではない



アクセラレータ

- **グラフィックボードの導入**
- **稼働開始時(2009/8)に搭載し，サービスインから利用も可能**
- **当面は限定された利用者に開放**
  - アプリケーション分野：分子動力学，宇宙物理，可視化など
  - センターとして，サポートと利用方法や応用分野の検討
- **計算機センターでサービスするプログラマブル・アクセラレータは．．．**
  - 使える人が使える → それでいいのか
  - パラメータサーチの道具 → それでいいのか

  - 多くの人が使えるように，GPU readyのプログラム・ライブラリの導入，プログラミング講習
  - スケールアウトの手段として，パラメータサーチも重要．ただ，アクセラレータを使った，並列計算の検討も進めないと

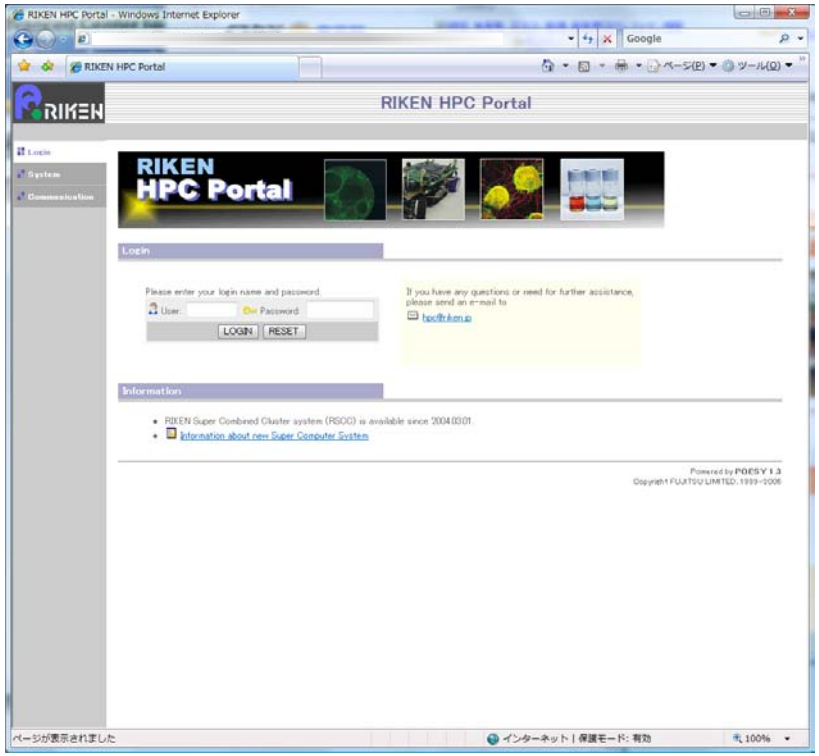
通信がプアというのは，PCクラスタ黎明期を彷彿とさせるなぁ．

システム詳細

ソフトウェア編

Webポータル(HPC・Bioポータル)

- スーパーコンピュータの操作をWeb画面から
- HPCポータルのサブセットを携帯電話から利用可能に



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
などなど

WebシステムとWebサービス

- **Webサービス(REST)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
    - ユーザさんが自分で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
  - スーパーコンピュータのSaaS (Supercomputing as a Service)化を実現

ユーザジョブの運用管理

- **スケジューリング・ソフトウェアを開発**
- **RSCCの様々なノード構成とサブクラスタ間のロードバランス等の解消を目指す**
  - 昨年度実績では90%を超える稼働率を達成
  - 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
    - スケジューリングポリシー：様々な優先順位，フェアシェア，バックフィル
    - 特性：MDGRAPE-3の有無，ISVソフトのライセンス，並列度の大小，時間の長短，メモリ量，ジョブ間依存関係
    - 効果：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
- **RICCシステムでも同様＋αのジョブ運用を目指す**
  - ジョブスケジューラはスパコン・システムの縁の下の力持ち
    - 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・トポロジと利用方針を踏まえたノード・アロケーション管理
    - 数万の単一CPU利用ジョブのスケジューリングに対応
    - サーバ障害時のフェイルオーバーに対応

様々な苦情・問題点への対処

- **ユーザジョブへの助言のためのロギング強化**
  - ユーザジョブの性能情報を把握する
  - 超並列アプリの開発にはプロファイル・トレースは必須
- **ISVアプリケーションの可搬性の向上**
  - RHELを採用，PVMも動きます
- **メモリ量，ディスク量を増強したが．．．．**
- **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

まとめ

- **次期システム（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
- **次期システムのハードウェア・ソフトウェアの詳細**
  - 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を利用することによる使い勝手，利用効率を向上
  - 新しい使い方・サービスを展開・サポート