PACS-CSと
次世代スーパーコンピュータ
～計算機システム屋からの提言～

朴　泰祐

筑波大学　システム情報工学研究科／計算科学研究センター

PACS-CSプロジェクト

- CP-PACS (1996〜2005)の後継システム構築とそれを用いた大規模計算科学問題の解決
- 2560 node (=CPU)からなる超並列クラスタ
  - メモリとネットワークの実効バンド幅の追求
  - 実空間アプローチに適したシステムアーキテクチャ
- 2005〜2007年度の3ヵ年計画
  - 特別教育研究経費
    「計算科学による新たな知の発見・統合・創出」
  - システム稼動開始予定：2006年7月
  - 当面の大型アプリケーション：RS-DFT & QCD

バンド幅を強く意識した超並列クラスタPACS-CS

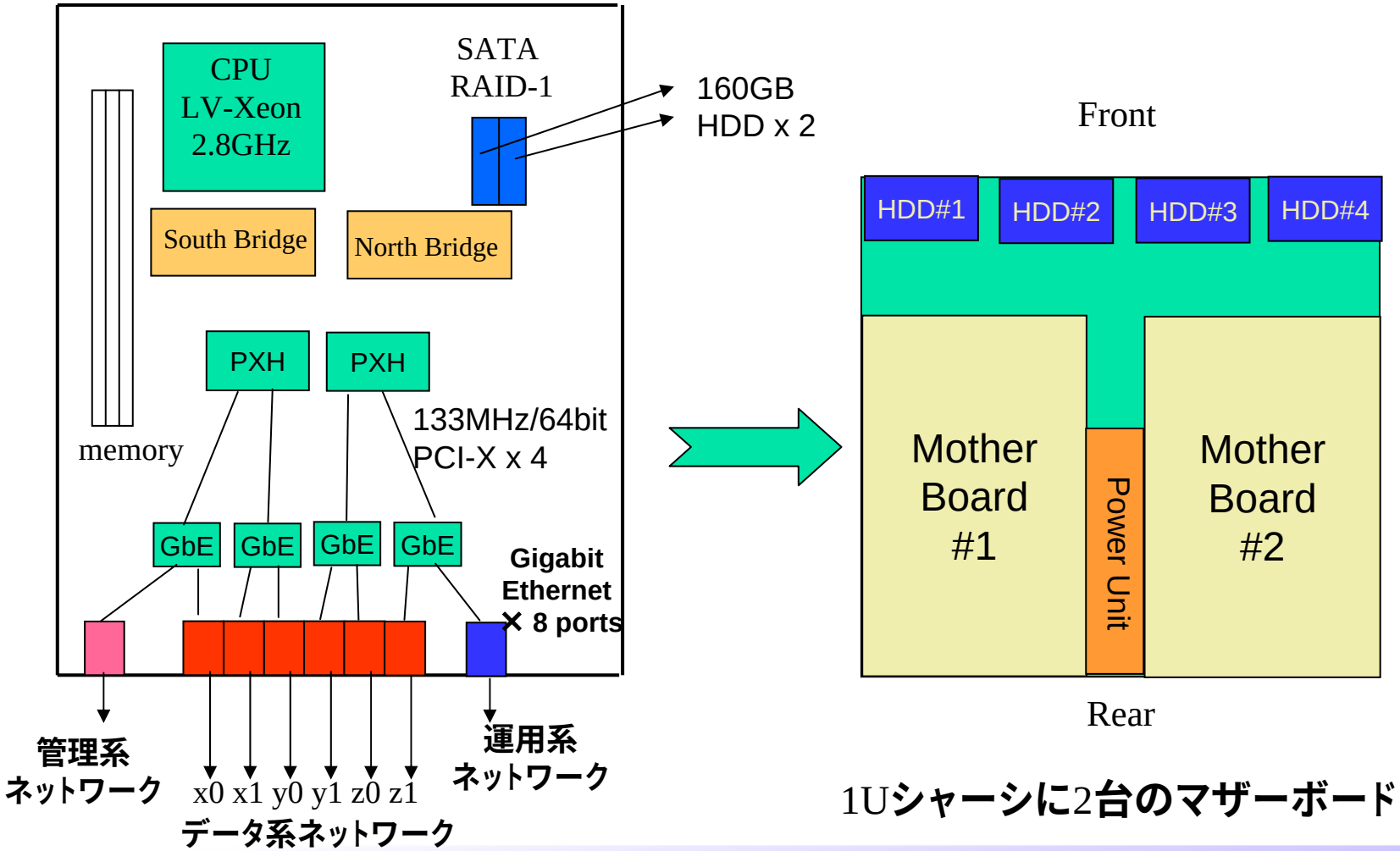
- PACS-CS =

  Parallel Array Computer System for Computational Sciences
- **メモリバンド幅に対する意識**
  - Single CPU / node = CPU**当たりの有効バンド幅の向上**
  - **従来と同様の高密度実装を実現 (専用ボード開発)**
- **ネットワークバンド幅に対する意識**
  - **「アプリケーションとアルゴリズムを意識した」高バンド幅ネットワーク**
  - **高い対価格性能比を持つ**Gigabit Ethernet**をトランク結合＋多次元化**
  - **3次元ハイパクロスバ網** (3D-HXB) **をコモディティネットワーク技術＋ソフトウェアで実現** (PM/Ethernet-HXB)

PACS-CSシステム諸元

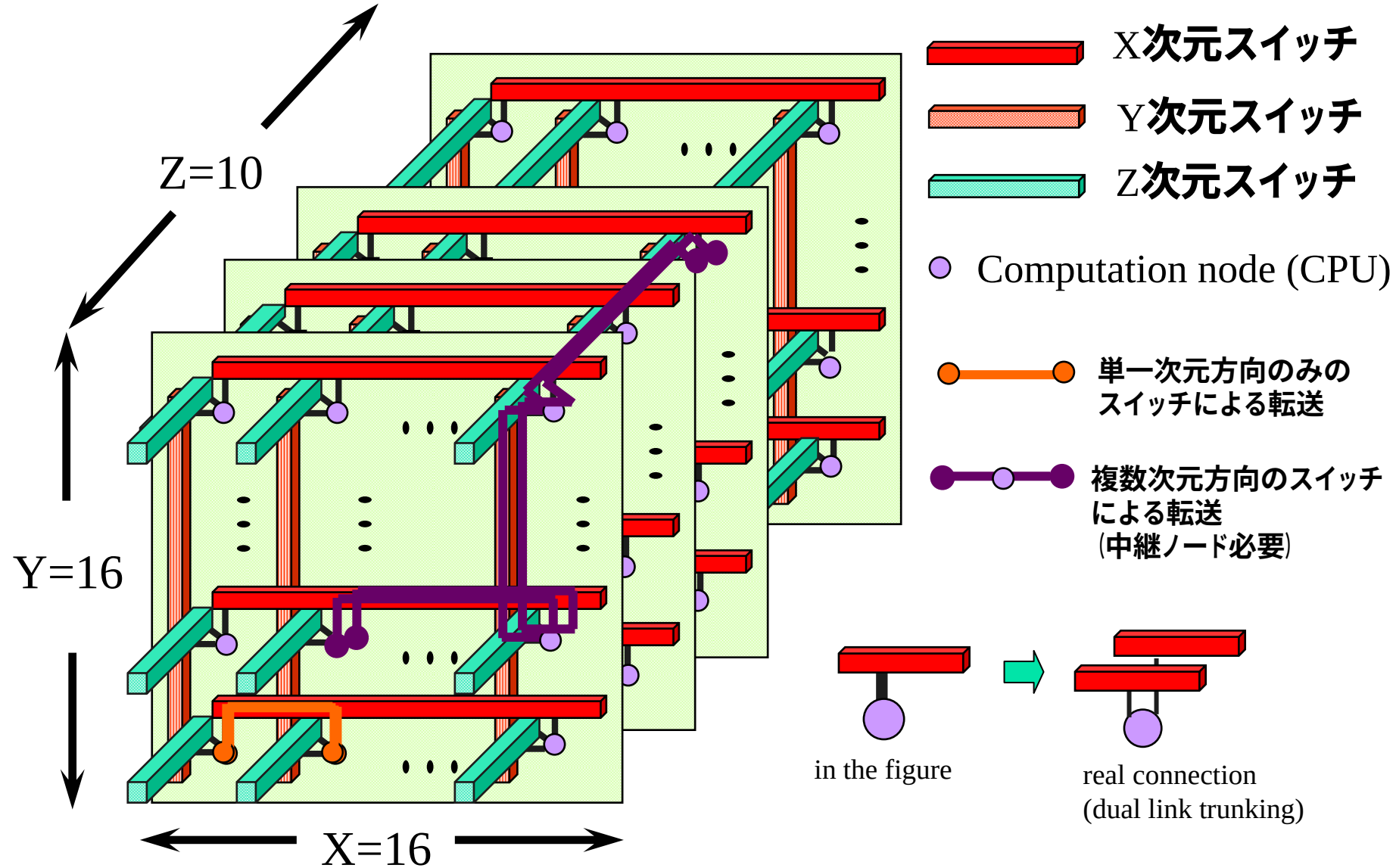
ノード台数	2560 (16 x 16 x 10)
理論ピーク性能	14.3 Tflops
ノード構成	**単一**CPU / **ノード**
CPU	Intel LV Xeon EM64T, 2.8GHz, 1MB L2 cache
メモリ容量・バンド幅	2GB/CPU　6.4GB/sec/CPU
並列処理ネットワーク	**3次元ハイパクロスバ網**
リンクバンド幅	**単方向** 250MB/s/**次元** **単方向** 750MB/s　(**3次元同時転送時**)
ローカルHDD	160 GB/**ノード** (RAID-1)
総システムサイズ	59**ラック**
総消費電力 (推定)	545 kW

専用マザーボード開発

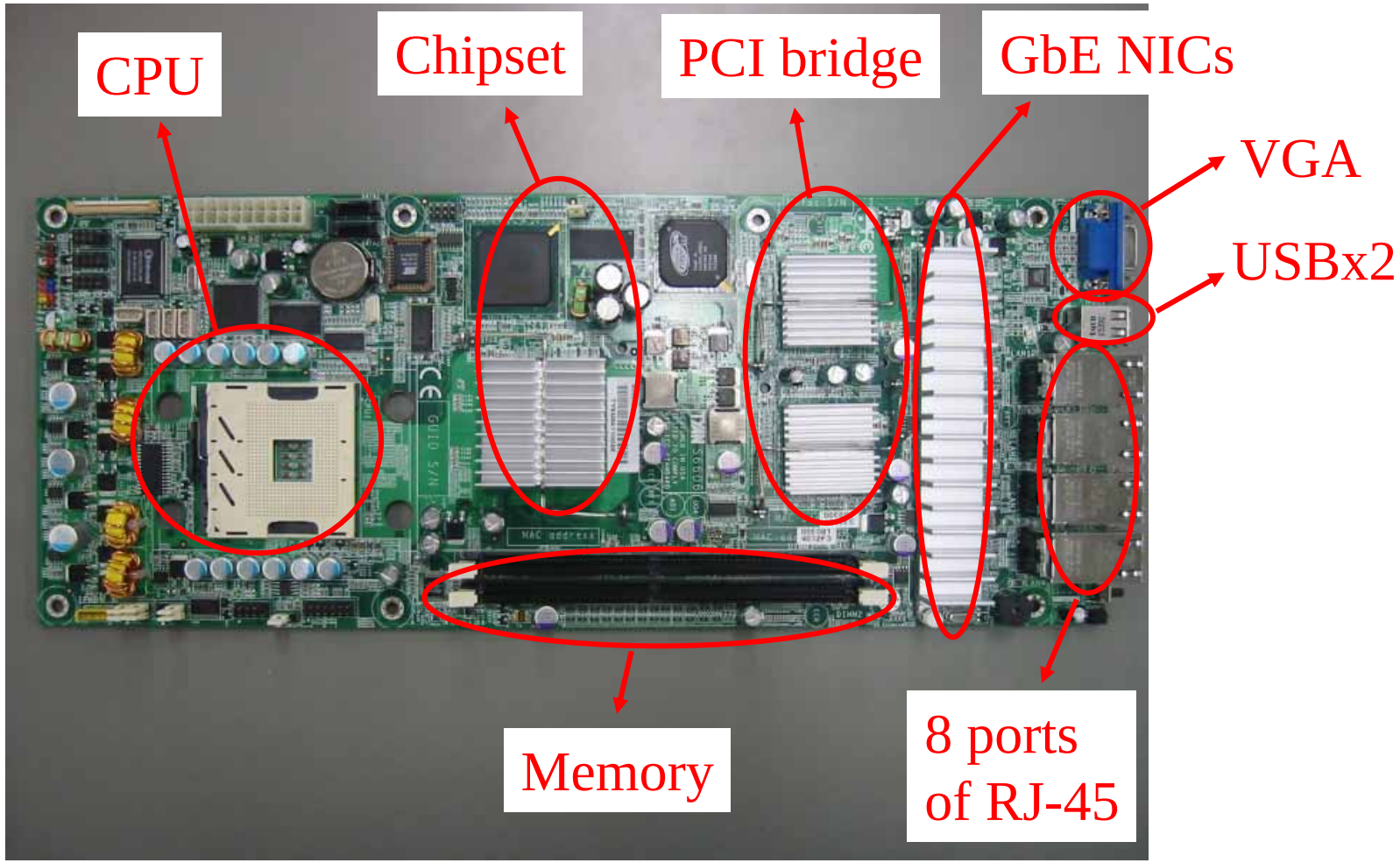


1Uシャーシに2台のマザーボード

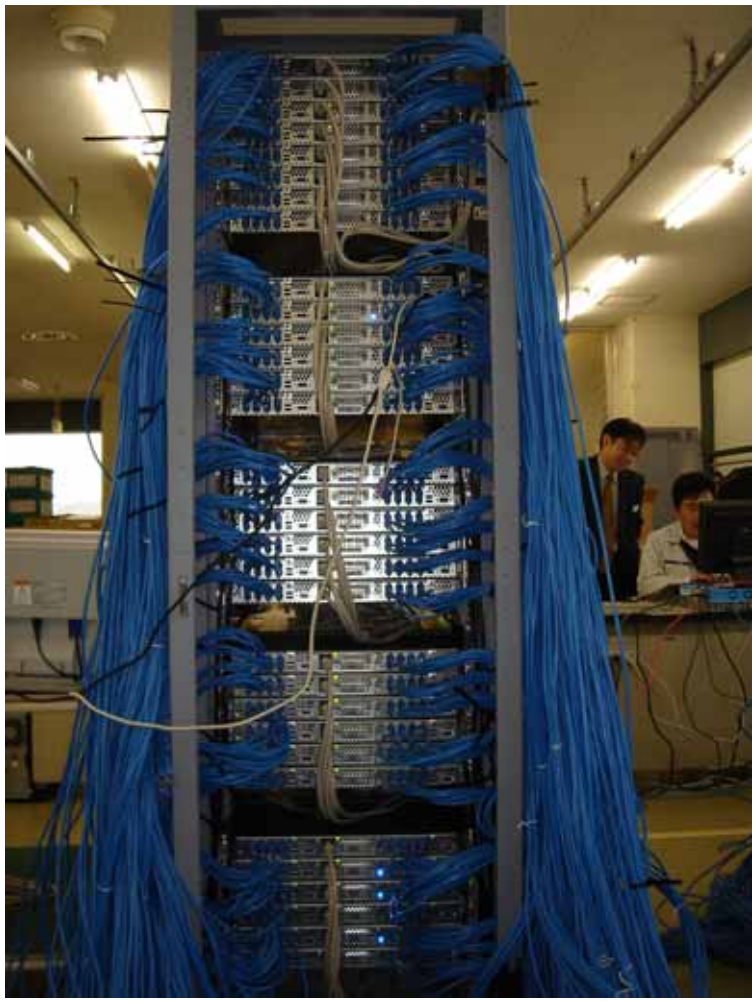
3次元HXBネットワーク (16x16x10=2560 node)



マザーボード実物写真


CPU
Chipset
PCI bridge
GbE NICs
VGA
USBx2
Memory
8 ports
of RJ-45

ノード筐体 (左) とスイッチ筐体 (右)

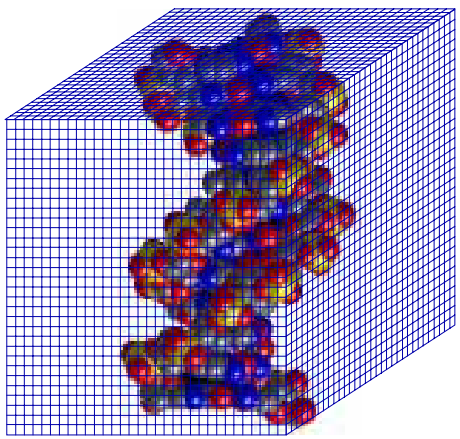


体制･スケジュール

- 産学連携プロジェクト
  - 全体構想･推進：筑波大学計算科学研究センター
  - ハードウェア：日立製作所
  - ソフトウェア：富士通
- 現状と予定
  - マザーボード、シャーシ完成
  - 512ノード単位で建造中
  - 2006年6月中完成
  - 2006年7月運用開始

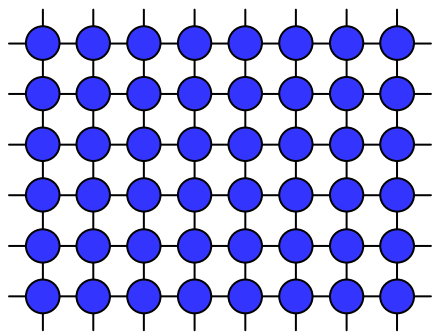
実空間モデリング・アプローチ

- 問題空間を物理的実空間のままで解く
- 近接相互作用を中心に超並列処理
  - 膨大な演算量＝超並列処理
  - N次元近接通信＝HXBにより構築可能 (物理的　or 論理的)
- 様々な物理問題・工学問題に対して応用可能
- PACS-CSシステム上で大規模問題を解くことにより方法論の確立・大規模化への対応を目指す



実空間離散化
から直接マップ

物理空間を直接
シミュレーション

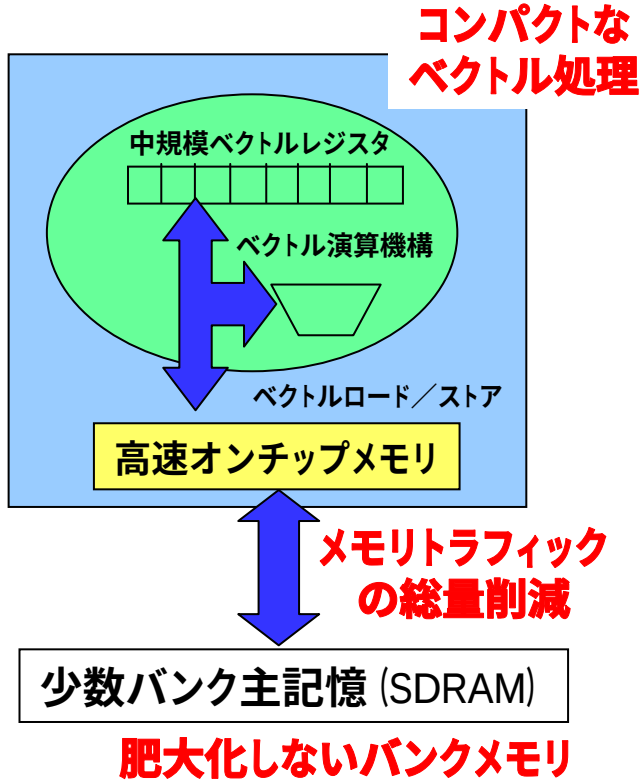


PFLOPS超級を目指す低電力・超並列処理方式

- PFLOPSをいかに実現するか
  - 現在のPCクラスタで作ってみると…
    - 演算性能電力比の問題
      3GHz Xeon x 2 / 300W = **1PFLOPS / 25MW**
    - 設置スペースの問題
      3GHz Xeon x 2 / 1U ⇒ ラック当たり30U分詰めたとして
      360 GFLOPS/1.2m^2 = **1PFLOPS / 3000m^2**
  - 性能当たりの**電力密度・実装密度をいかに向上させるか**が最大のポイント
- **実効性能を確保**
  - CPU性能：メモリバンド幅：ネットワークバンド幅のバランスを保つ
  - メモリバンド幅ボトルネックを brute force でなく、解決する
- **超低消費電力・超高密度システム実現の鍵**
  - プロセッサ及びネットワークにおける超低消費電力化技術
  - 超高密度実装を実現するアーキテクチャ
  - 実アプリケーションにおける実行効率を落とさないアプローチ

コンパクトなベクトル処理＋低消費電力

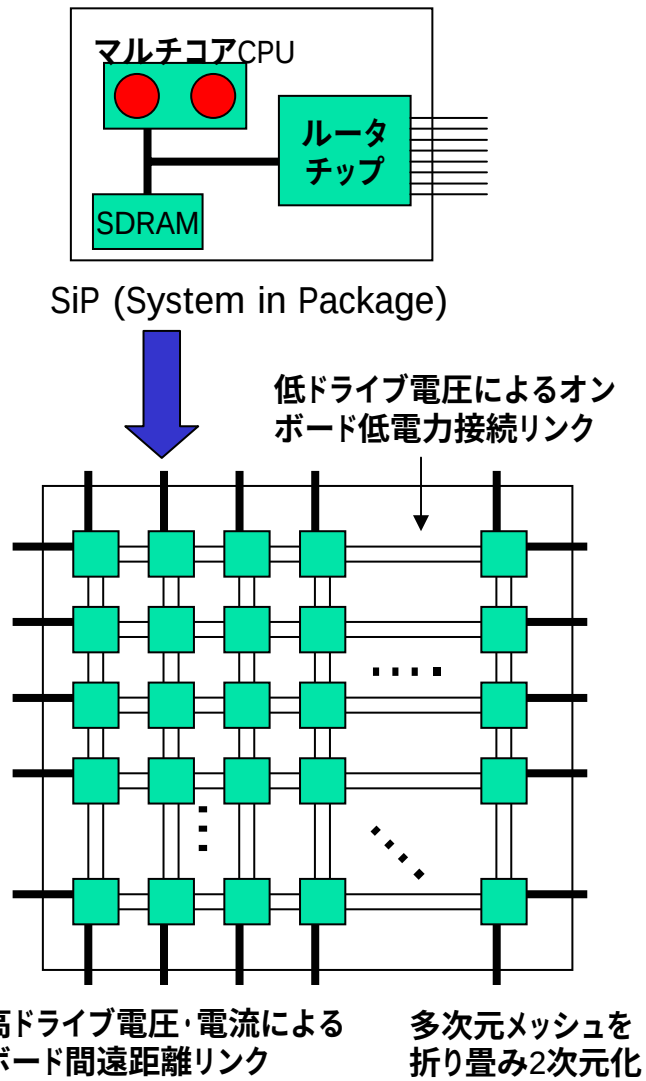
- 軽量･高性能プロセッサの開発
  - 組み込みプロセッサ／システム実装技術のHPC向け応用
  - 将来に向けての持続的開発
- 低電力･高性能プロセッサ向けアーキテクチャ
  - コンパクトで足回りの良いショート･ベクトル処理
  - メモリトラフィックの絶対量の削減
    オンチップメモリの有効利用 (SCIMAプロセッサ)
- プロセッサアーキテクチャ
  - 4core/chip, 4FMA/core, 1.5GHz
    ＝ 48GFLOPS/chip
  - 各種低消費電力技術 (DVS, SOI, 閾値制御, SoC, SiP) の導入



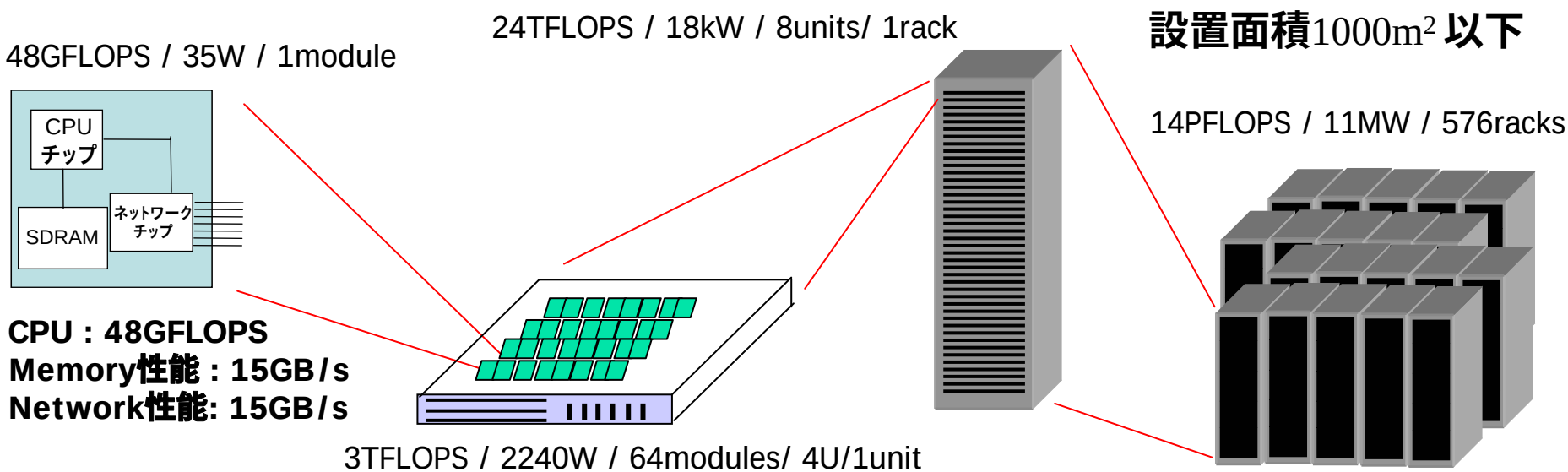
高性能･拡張性を持つ低電力相互結合網

- **基本方針**
  - プロセッサ性能･電力に見合った結合網
  - 多次元展開により100万ノード規模まで拡張化
  - 対象とするアルゴリズムを実空間離散化に絞り、その上でバンド幅に余裕のある超並列通信網を構築

- **技術的課題：低消費電力･高密度ルータチップの開発**
  - 10～20本の高速シリアルリンクをトランク結合してバンド幅と次元数を柔軟に組み合わせる
  - SiP化により相互結合網リンク接続のみをボード上で結線
  - **通信所要電力を通信距離に応じて最適化**
    - **ユニット内 (短距離)** はリンク当たりの転送レートを上げられるので少数リンクで高性能化可能
    - **ユニット間 (長距離)** はリンク当たりの電力を抑え、転送レートを低くする代わりにリンク数を増やす
    - メッシュ型であれば両者を組み合わせて同一バンド幅を維持できる



超並列方式での１０PFLOPS超級システムのイメージ



- 基本方針
  - 超並列方式の採用
  - 低電力高性能プロセッサの開発
  - 低電力高性能相互結合網の開発
  - 高密度実装技術による大規模集積
- 電力、設置面積、信頼性が課題

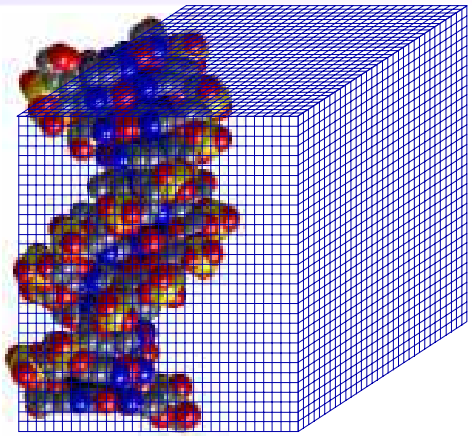
- 目標仕様
  - ピーク性能：14PFLOPS
  - 消費電力：11MW以下
  - ノードの演算/電力性能　：48GFLOPS/35W
  - 64ノード/ ユニット、8ユニット / rack、576 rack/system
  - 総ノード数：64x8x576 = 294,912
  - 総主記憶 2304TB (8GB/node)
  - 相互結合網　：３次元メッシュトーラス(64x64x72)
  - ネットワーク性能：　10Gbps/link × 12link

提言：協調関係とパラダイムシフト

- PFLOPS超級のための「思い切り」
  - CPU性能・メモリバンド幅・ネットワークバンド幅のバランスを取ることにより実効性能を確保
  - 現実的な電力・スペースを維持（大規模センターだけでなく計算センターレベルで導入可能なもの）
  - プロセッサ及びネットワークのアーキテクチャにある種の「妥協」は必要
- グランドチャレンジが可能なアーキテクチャを
  - プロセッサ／ネットワークを最大限に利用するモデリング／アルゴリズム／プログラミング
  - PFLOPS超級の性能を引き出すためにはアルゴリズム／ソフトウェア／ハードウェアの統一的な連携が必要

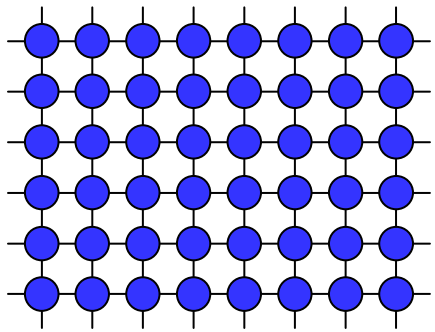
    ⇒ アプリケーション側の努力にも大いに期待したい
  - 計算科学と計算機工学の研究者の綿密な連携が必要
- 超並列＝高性能＝実アプリケーションをつなぐパラダイムシフトを！

超並列方式へのパラダイムシフト



物理空間を直接
シミュレーション

実空間離散化
から直接マップ



問題空間の高密度離散化 → 物理空間を解く直接解法へ

↓

非常に高い演算性能が要求されるが近接通信のみでよい

← 空間局所性を生かしたプロセッサ＋ネットワークであれば超並列は可能

↑

問題空間の高密度離散化

超並列方式と実空間物理シミュレーションは相補的な関係にあり、実空間アプローチへのパラダイムシフトが、それ自身が必要とする十分な並列演算性能を生み出す