超並列クラスタPACS-CSの概要

2006年6月27日

筑波大学　計算科学研究センター

Center for Computational Sciences, Univ. of Tsukuba

計算科学研究センターにおけるスーパーコンピュータ

- CP-PACS (1996年11月世界最高速)
  ⇒ 2005年9月に運用停止

- この数年間はPCクラスタをベースとしたシステムが対価格性能比的に有望

- 一般的な高性能PCクラスタの問題点を見直し、本センター独自の発想により超並列クラスタシステムを設計・開発

一般的な高性能PCクラスタの問題点

- 対象アプリケーション、問題の解法、問題規模等を限定できないため、極めて汎用的・広範な利用を想定せざるを得ない
  - 野心的なノード／ネットワークアーキテクチャを採用しない
- ピーク性能は高いが、実効性能はそれほど追求しない
  - CPU性能に対し、メモリ性能やネットワーク性能が相対的に低い
- ノード数が多くても実運用上は分割運転が多い
  - フルシステム規模で運用した場合の性能を軽視（ネットワーク性能等）

我々の求める超並列クラスタのコンセプト

- **CPU性能:メモリ性能:ネットワーク性能**
  の性能バランスを極力保つことを強く意識したPCクラスタ
- 超並列アプリケーションの日常的運用
- 現在のコモディティ(汎用・大量生産)技術を最大限に利用

コモディティ技術による超並列計算機(MPP)を作る

- チップ等の開発はしない
- コモディティCPU、コモディティネットワーク、コモディティソフトウェア
- ボード・レベルでの開発は行う
- 必要なドライバソフト等の開発は行う

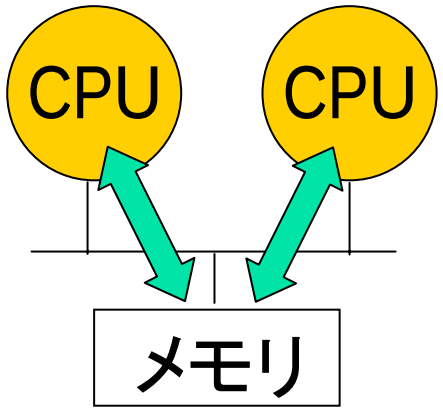
PACS-CS：計算科学向け超並列クラスタ

- Parallel Array Computer System
  for Computational Sciences

- 計算科学における大規模シミュレーションの効率的実行を目指した新しいコンセプトの超並列PCクラスタ
- 実効性能を追求するため、メモリと通信のバンド幅を重視
- 対価格性能比の良い汎用ネットワーク媒体を利用し、我々の超並列計算モデルに適合する相互結合網を構築

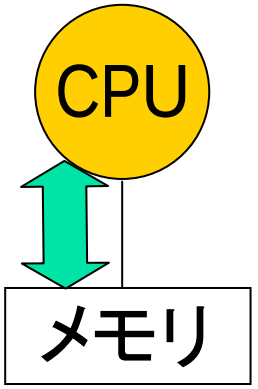
メモリに対する高バンド幅設計

SMP構成（一般のクラスタ）



- ノード上のメモリを複数CPUで共有
- 頻繁なメモリアクセスを行うプログラムで性能低下
- 複数CPUでネットワークアクセスを共有するので性能低下

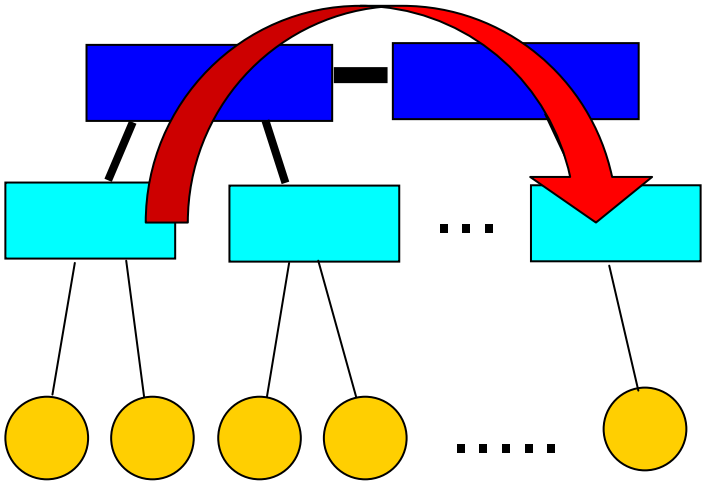
単一CPU構成（PACS-CS）



- ノード上のメモリを単一CPUで独占
- 頻繁なメモリアクセスを行うプログラムで性能向上
- 単一CPUでネットワークアクセスを占有

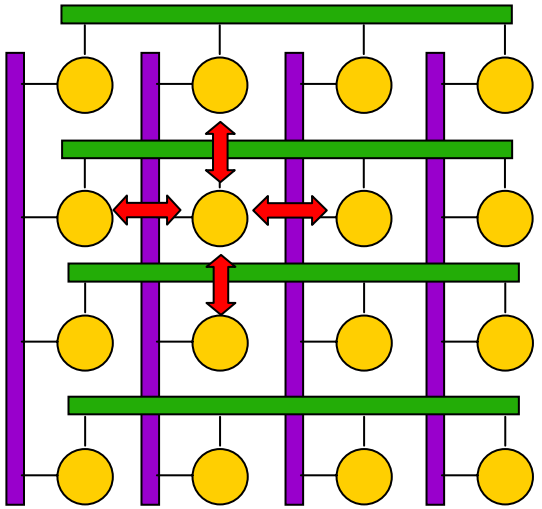
ネットワークに対する高バンド幅設計

ツリー型ネットワーク
（一般の高性能クラスタ）



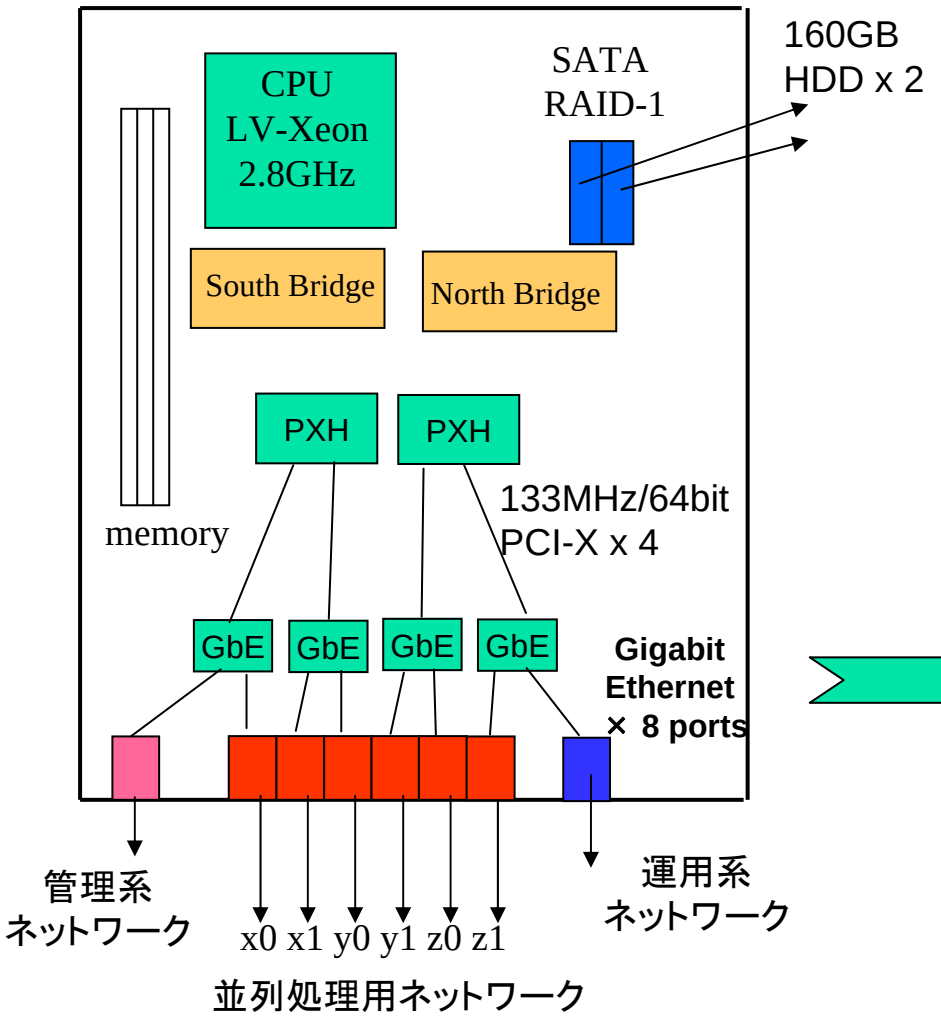
- 木構造のルート付近で混雑発生
- 多次元格子上の近接通信であっても大域的通信が発生
- 数千ノード規模での実装でコストが大幅に増加

ハイパクロスバネットワーク
（PACS-CS）

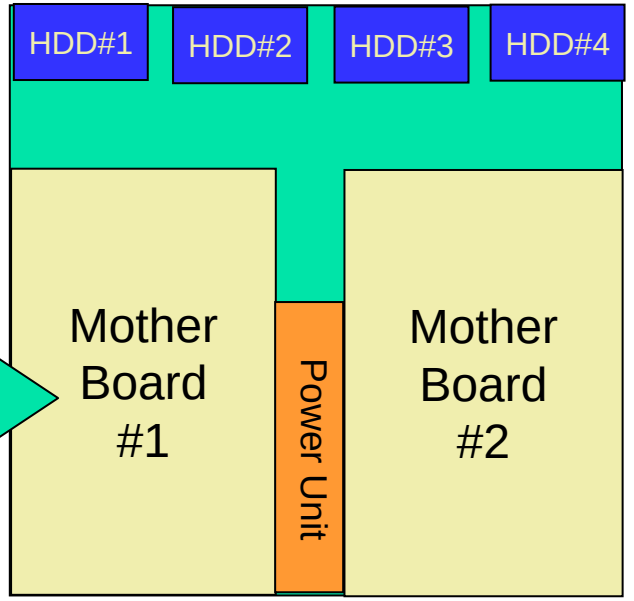


- 近接通信・集団通信ではボトルネックが発生しない
- 数千ノード規模に容易に拡張可能
- 対価格性能比の高いネットワークを構築可能

ボードの構成：1ノード当り1 CPU



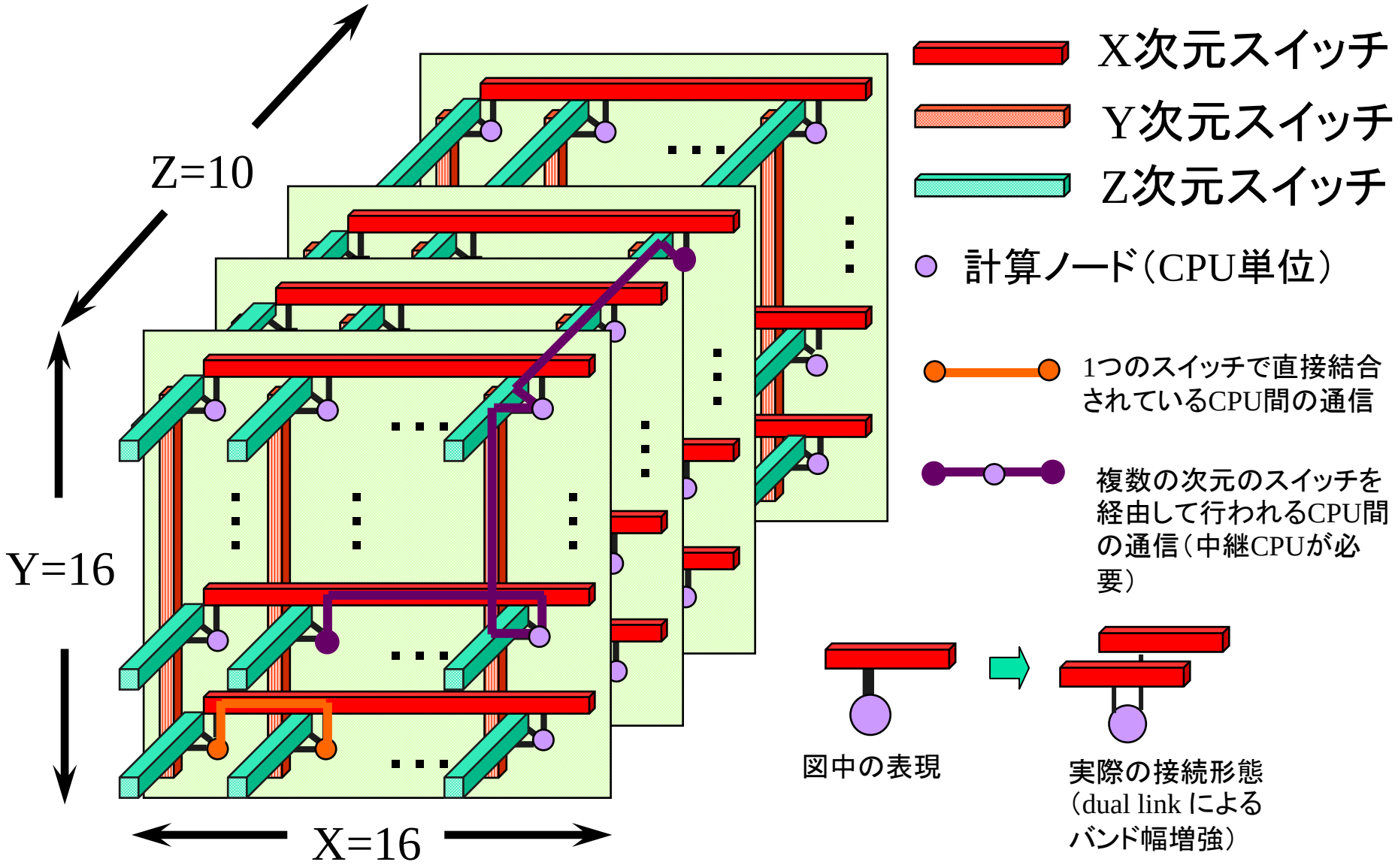
従来の2CPU／ノード構成の
ハイエンドPCクラスタと同じ
実装密度を保つ新規ボードの開発



1ユニット内に2台のマザーボード

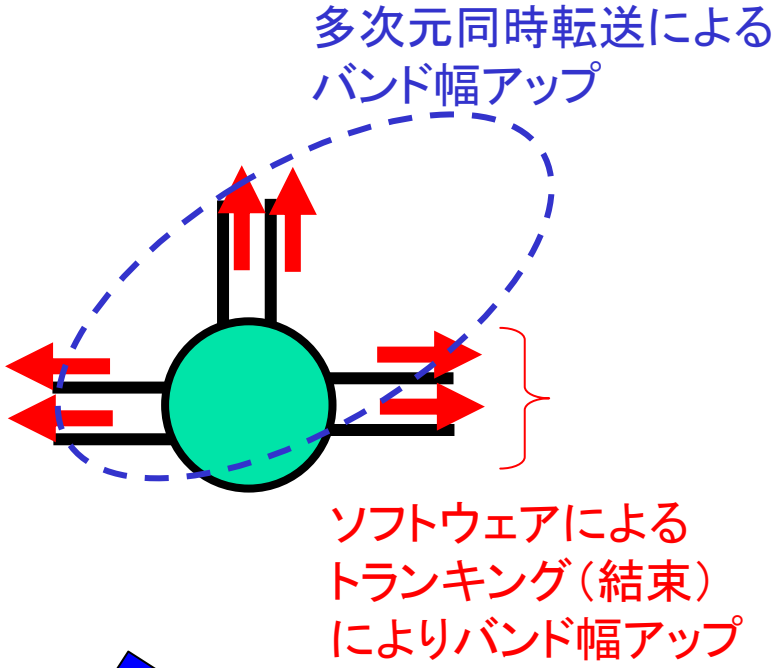
ノード(CPU)間の論理的結合

(3次元ハイパクロスバ網) 2560 node 構成



Gigabit Ethernetトランクによるハイパクロスバ網

- ターゲットアプリケーションの絞込みによりネットワーク構成の無駄を省く
  ⇒ 近接通信と集合通信を高性能で実現
- 対価格性能比に極めて優れている Gigabit Ethernet を積極的に利用
- ＋ソフトウェア技術により高価な高性能ネットワークと同等の性能を実現



汎用Gigabit Ethernetと比較的安価なスイッチのみで
高価な高性能クラスタ向けネットワークに匹敵する性能

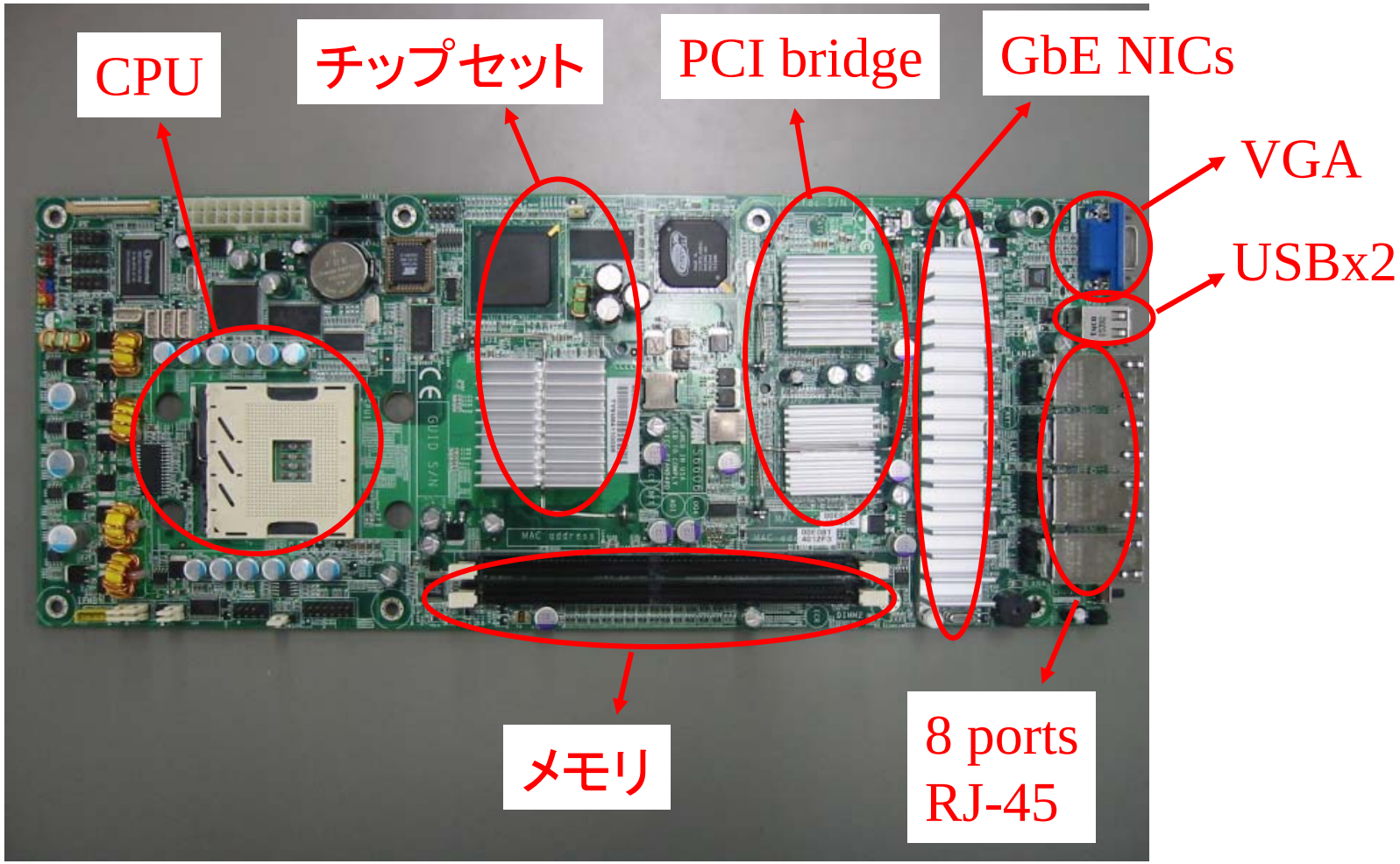
PACS-CSシステム諸元

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	2GB/ノード　(5.12 TB/システム)
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	320GB/ノード⇒RAID-1 (160 GB/ノード)
総システムサイズ	59ラック
総消費電力（推定）	545 kW

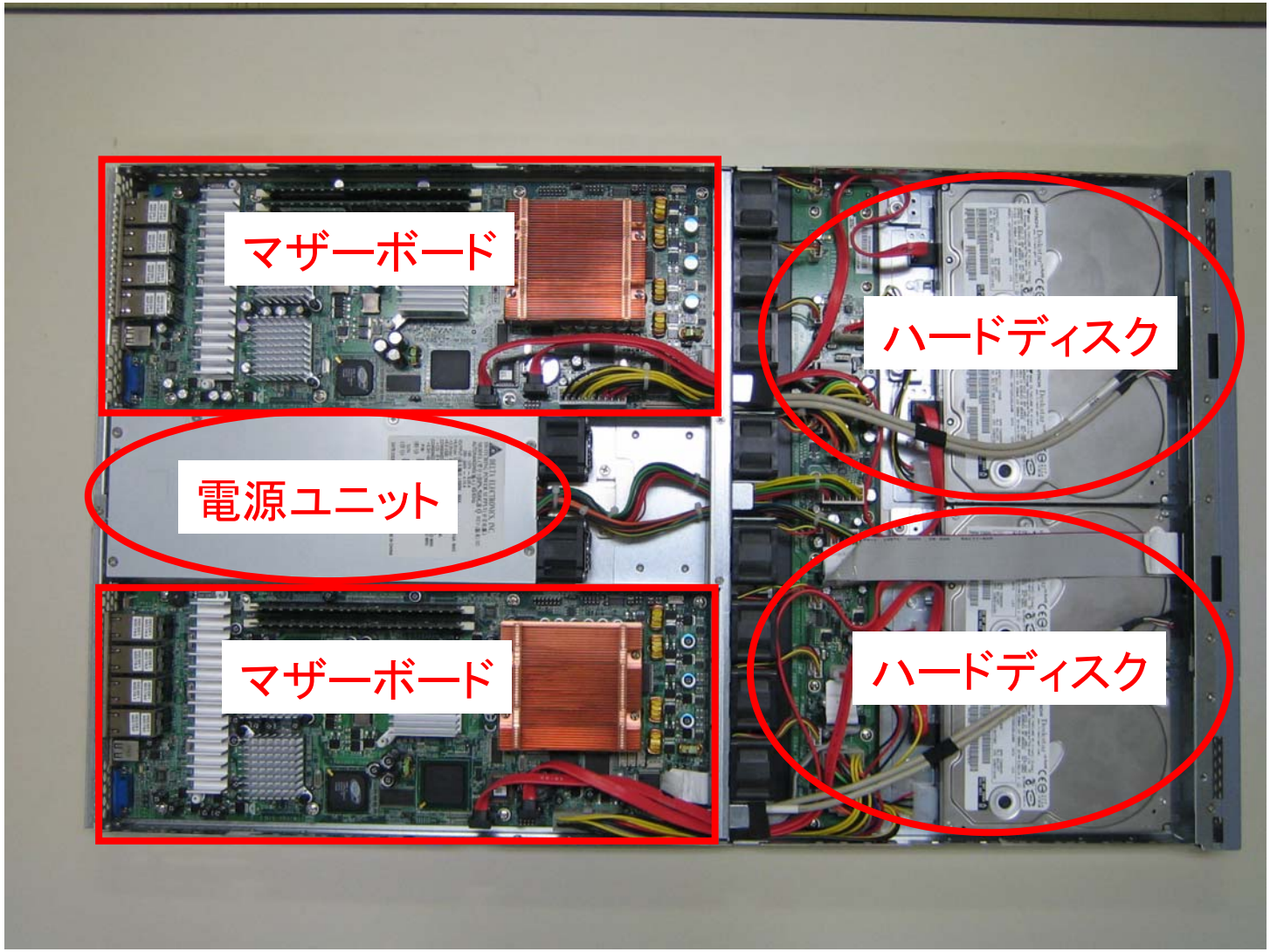
一般的なHPC向けクラスタとPACS-CSの比較

	一般的なPCクラスタ	PACS-CS
ノード構成	SMP（共有メモリ）構成 ピーク性能重視	単一CPU構成 実効性能重視
ネットワーク	高価なクラスタ向けネットワーク（SAN）を利用	安価なGigabit Ethernetを束ねて多次元化して利用
実装密度	SMP構成の汎用PCサーバを利用して高密度化	専用ボードを開発し、一般のHPCクラスタと同等の実装密度を実現
ターゲット応用・モデリング	あらゆる演算・通信特性を持つアプリケーションを対象とする（特定の手法に特化しない）	大規模シミュレーションに実空間アプローチを適用し、汎用性を保ちつつ大規模化を実現
総合性能	Linpackベンチマークでは高性能だが一般応用では高性能は困難	Linpackベンチマークを含め、高いメモリ、ネットワークバンド幅を要求する問題に対応

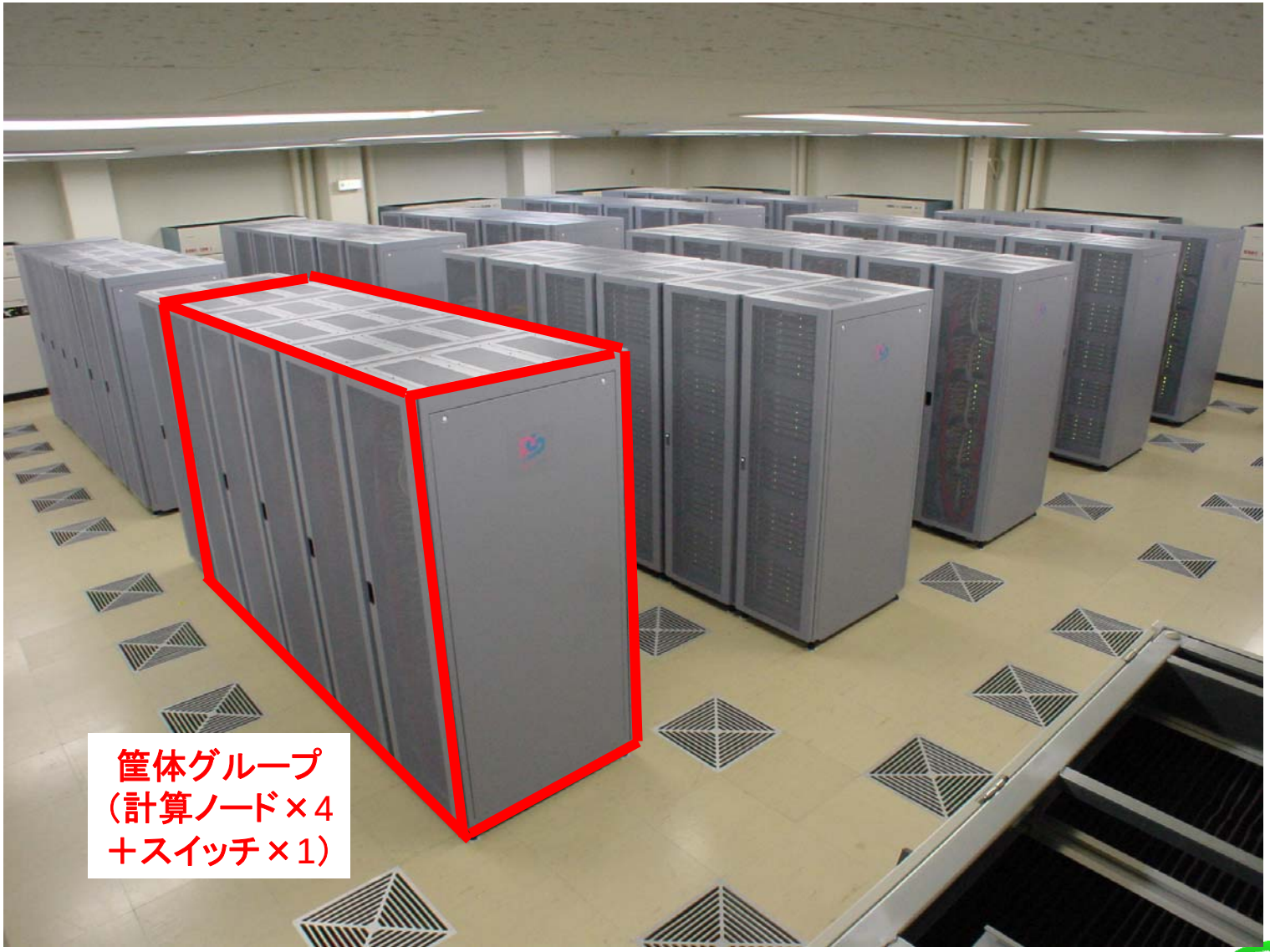
開発したマザーボード



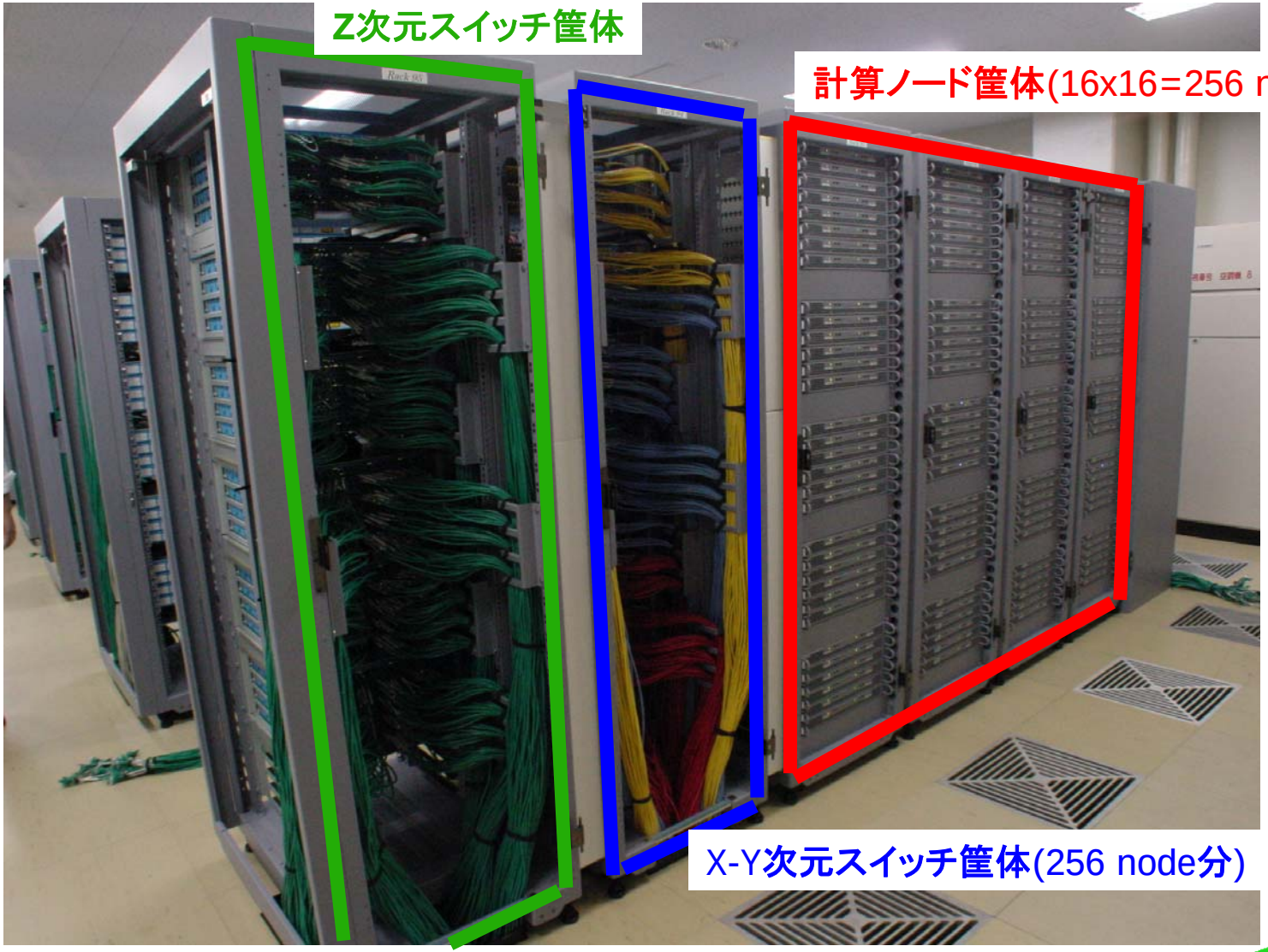
ユニット(19inch x 1U)



PACS-CS (2560 node)



１筐体グループの構成



ソフトウェア

- Linux + SCore
  - PM/Ethernet-HXBドライバ
    （GbEトランク3次元ハイパクロスバ用の専用ドライバ）
  - パーティショニング、モニタリング
- MPIによる並列プログラミング
  - MPICH（標準）とYAMPII（東京大学開発）を切り替えて使用
- 言語：Fortran90, C, C++
- 数値計算ライブラリ：MKL
  FFT-E（筑波大開発）

国産システムとして第二位

システム名称	メーカー名	設置機関	設置年月	Linpack性能 (TFLOPS)
TSUBAME	SUN	東工大	2006/4	38.18
地球シミュレータ	NEC	海洋科学技術センター	2002/3	35.86
Blue Protein (BlueGene/L)	IBM	産総研・生命情報科学研究センター	2005/3	18.20
BlueGene/L	IBM	高エネルギー加速器研究機構	2006/3	18.20 x 2台
Altix3700	SGI	日本原子力研究開発機構	2005/3	11.81
PACS-CS	日立＆富士通	筑波大・計算科学研究センター	2006/6	10.35

国内メーカーによるスーパーコンピュータとして
地球シミュレータに次ぐ第二位の性能

SCore利用クラスタとして世界最高速

- SCore
  - 経済産業省RWCPにて開発
  - 現在、PCクラスタコンソーシアムで開発・管理
  - 日本発の大規模クラスタ用ミドルウェアとして世界的に有名
- 主なSCoreクラスタ
  - 筑波大 PACS-CS (2560 node, 14.3 TFLOPS)
  - 理研 RIKEN Super Combined Cluster (1024 node, 12.53TFLOPS)
  - 産総研 AIST Super Cluster (1024 node, 8.8TFLOPS)
- PACS-CSはSCore利用クラスタとして世界最大規模・最高速

まとめ

- 大規模計算科学シミュレーション用超並列クラスタ
- 「CPU性能：メモリ性能：ネットワーク性能」のバランスを重視
- アプリケーションの特性を生かしたコストパフォーマンスの高いネットワーク（GbE trunked 3D-HXB）
- 従来の dual CPU SMP ノードと同等の演算ノード実装密度
- 2560 CPU, 14.3 Tflops システムが2006年7月に稼動開始
- SCoreクラスタとして世界最大・最高性能
- Linpack性能だけに満足せず、実アプリケーションにおける性能にコストをかける