PCクラスタによる大規模計算物理学の可能性

朴　泰祐

筑波大学　計算科学研究センター

(システム情報工学研究科コンピュータサイエンス専攻)

taisuke@cs.tsukuba.ac.jp

http://www.hpcs.is.tsukuba.ac.jp/~taisuke/

概略

- CP-PACSの時代を振り返って
- HPC (特に大規模計算物理学) のプラットフォームに必要なもの
- 現在のセンターのクラスタ
- PCクラスタによる post CP-PACS の可能性
- 新クラスタ構想：**CP-PACS II**
- まとめ

CP-PACSの時代を振り返って

- 1990年代前半～中盤：
  大規模ベクトル、超並列計算機等のハイエンドコンピュータの絶頂期
- 文部省：新プログラム「専用計算機による『場の物理』の研究」
- ベクトル機に代わるHPC向け超並列計算機へのチャレンジ
- 数々の並列処理関連プロジェクト（RWCP，重点領域）

良い時代でした！

CP-PACS成功の背景

- 天の時
  - ちょうど(超)並列処理研究と実用化の機運が高まりつつあるタイミング
- 人の和
  - 計算物理学者と計算機工学者の協調作業
  - シーズとニーズのベストマッチング
  - 研究者とメーカーの要望のマッチング(産学連携)
- 地の利(?)
  - 筑波大学の並列計算機／計算物理学の伝統
  - つくばという土地(HPC研究のメッカ?)

かつての世界最高速計算機も…

1996年11月のTOP500
第一位
ピーク性能 614 GFLOPS
Linpack性能
368 GFLOPS
(地球シミュレータの前
に日本が一位を取った
最後の計算機)

2003年11月のTOP500
ついに drop off !!

現在のHPCの状況

- ベクトル計算機の棲家が狭まりつつある
  - 性能向上に対する規模･価格･電力の増大
  - 商業的な成り立ちの難しさ
- なかなか広がらない裾野
- クラスタの台頭
  - コモディティ部品の圧倒的な低価格性
  - メーカーのサーバー向け商品の充実
  - ネットワーク性能の向上

我々に必要なHPCプラットフォーム

- 演算性能
  - もちろん必要だが、
  - 性能自体を見れば現在のマイクロプロセッサでも十分
    単体プロセッサでピーク 6Gflops、1000台で6 Tflops
- ネットワーク性能
  - お金をかければ一昔前のスパコンを凌駕する性能
    Infiniband (x4): 1 Gbyte/s
    MyrinetXP (dual): 500 Mbyte/s
- 結局はメモリバンド幅
  - ベクトル計算機のお金＝メモリ（バンド幅）
  - CP-PACSは擬似ベクトル処理（ソフトウェアによる）だったが、
    メモリは16 bank持っていた

現在の計算科学研究センターのクラスタ

Orion cluster
Compaq AlphaServer DS20L
Alpha EV68 833MHz dual
30 nodes, 100 GFLOPS
Fast Ethernet
Linux + SCore
「CPUリソースばら撒き型」利用

Perseus cluster
HP ProLiant DL360G3
Xeon 2.8GHz dual
37 nodes, 414 GFLOPS
Myrinet2000
Linux + SCore
プロダクトラン(HMCS)
＋クラスタ実験

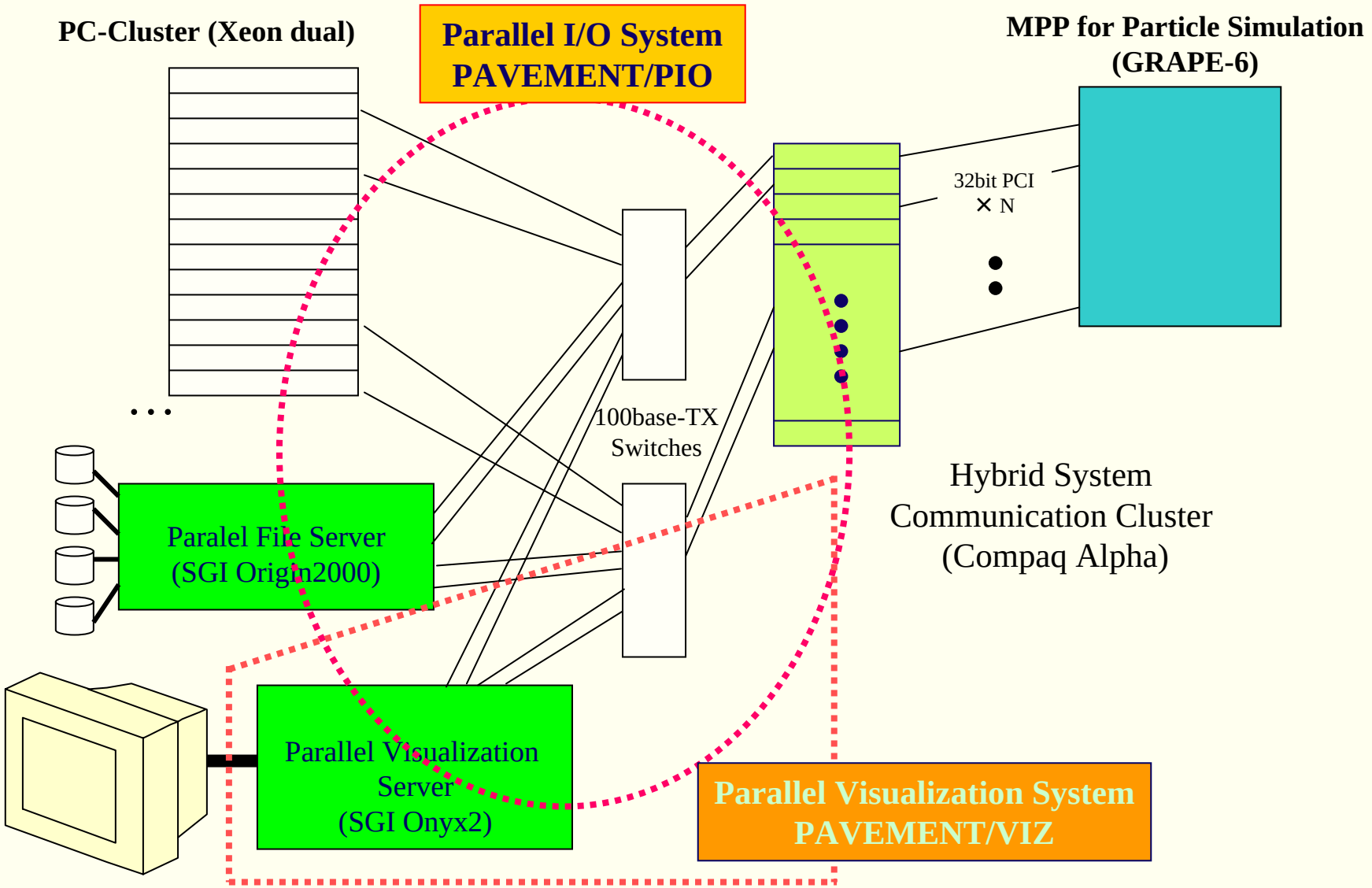
Flare cluster
DELL PowerEdge 1750
Xeon 3.06GHz dual
12 nodes, 72 GFLOPS
Gigabit Ethernet
Linux
「CPUリソースばら撒き」利用

Corona cluster
HP ProLiant DL380G3
Xeon 3.06GHz dual
8 nodes, 48 GFLOPS
Gigabit Ethernet x 6
Linux+SCore
ネットワークtrunk実験

現在のセンタークラスタの位置付け (1)

- 中規模計算のプロダクトラン用エンジン
  (Perseus custer – Xeon dual, Myrinet2000, 37 nodes)
  - 主に計算宇宙物理学
  - SCore+PBS+CMU による標準クラスタ運用
  - MPI on PM/Myrinet, no SCore-D
  - 重力専用並列計算機GRAPE-6との協調動作
    (HMCS: Heterogeneous Multi-Computer System)
  - 平均的なジョブサイズ
    6万～13万粒子の銀河形成シミュレーション
    200GFLOPS単位程度
  - Myrinet2000 full connection 程度でほぼ満足

Block Diagram of HMCS



現在のセンタークラスタの位置付け (2)

- 単なるCPU集合体として
  (Alpha EV68 dual,　Xeon dual)
  - QCD, 宇宙物理において単純なpost processingが多数発生する
  - 計画的計算というよりアドホックな計算
  - 従来並列処理用として用いていたものを転用
  - 基本的には逐次処理のばら撒き
    (WSの代わりに使う)

センターのリソースの今後の展望

- 中･長期的
  - CP-PACSの代替となる大型計算機の導入を目指す
  - センターの特色を生かした効率的なシステムを
    ⇒単に「できあいのスパコンを買う」のではなく
- 短期的
  - 現在のCPU性能とネットワーク性能、全体的な価格も含めたバランスを考えると、この数年間ではクラスタが有利
  - ベクトル化率99%の大型ベクトルプロセッサと、実効処理効率2～3割程度の多数の汎用スカラープロセッサの最終的な総合実効効率の比較

クラスタによる10～20TFLOPSクラスのシステム

センターのHPCリソースの考え方

- 計算科学研究センターの特徴
  - 対象とする分野の重要課題の認識
  - ある程度アプリケーション（あるいは手法）を絞り込める
    - full QCD　（小規模倍精度複素数行列計算＋近接通信）
    - ナノ物性　（FFT + CG法）
    - 宇宙物理　（HMCSの枠組み）
    - その他の今後の拡張分野
- MPPからの移行
  - 格差の大きい演算／通信性能比に対する最適な構成・技術をちゃんと考える

一般的なHPC向けクラスタ

- プロセッサはIntel互換 (Xeon, Opteron, Itanium2)
- Dual CPU 構成が一般的
  - 全体ピーク性能を保ちつつネットワークインタフェースの数と設定スペースを減らす
  - Network boundなアプリケーションは不得意
- ネットワークは SAN (System Area Network)
  - MyrinetXP: 現在dual connection に対応
  - Infiniband: 次世代の期待、x4 まで利用可能
  - ネットワーク性能が必要ない分野ではGbEthernetが中心

我々が目指すクラスタ

- ネットワークバンド幅対策：
  - 「ネットワークにはお金をかけるのは当たり前」ではつまらない
    ⇒我々独自の工夫で乗り切れないか
  - クラスタにおけるネットワークコストの増大
    - 「クラスタは大きくするとコスト効率が悪くなる」
  - 高性能ネットワーク (ex. InfiniBand, Quadrix)
    vs コモディティネットワーク・トランク
    (ex. GbE × n)
  - 対象問題の特性を生かしたクラスタを！

Mult_Bench: QCD計算ベンチマーク

(資料提供:石川@広島大)

- PCクラスタを想定した、QCD計算のカーネル部の計算・メモリアクセス・並列通信をまとめたベンチマーク
- Lattice QCDの処理における性能を、各次元方向 (x, y, z, t) に関して分析・出力可能
- 単体プロセッサ性能・並列処理性能を任意のノード構成で測定可能
- PC及びそのクラスタにおける性能測定から、大規模クラスタを実現した場合の性能を類推できる

CPU性能に関するMult_Benchの緒元

- 格子点1点当たりの演算数、データロード、データストア

方向	演算数 [flop]	Load [byte]	Store [byte]	比率 [byte/flop]
t	288	672	192	**3.00**
x	336	864	192	**3.14**
y	336	864	192	**3.14**
z	336	864	192	**3.14**
clover	600	864	192	**1.76**

全体で　5088[byte]/1896[flop] = 2.68 [byte/flop]

並列化通信におけるMult_Benchの緒元

- 並列化に関しては、4次元格子 (Nx*Ny*Nz*Nt) のうち3次元空間方向を3次元のプロセッサ格子 (nx*ny*nz) に分割する
- 各次元方向データの通信量

次元	通信量 [byte]
x	12*2*(Nt/2+1)*(Ny/ny)*(Nz/nz)*16
y	12*2*(Nt/2+1)*(Nx/nx)*(Ny/ny)*16
z	12*2*(Nt/2+1)*(Nx/nx)*(Ny/ny)*16

1ループの演算における総通信量と総演算量の比
(簡単のため Nx=Ny=Nz=Ns, nx=ny=nz=ns **とする)**
0.608 * ((Nt+2)/Nt) / (Ns/ns) [byte/flop]

実システムにおける性能評価

- 評価システム
  - P4 (Pentium4, 2.4GHz):
    HyperThreading、PC3200 memory、single CPU
  - Xeon (Pentium Xeon, 2.8 GHz):
    PC2100 memory, 2 CPU SMP
  - EV7 (Alpha EV7, 1.15 GHz):
    PC3200 memory, 16 CPU HyperTransport connected

（＊Alpha EV7 **は日本HPの協力による参考データ**）

単体CPU性能

- 単体CPUにおける性能 [Mflops]

(＊P4, Xeonに関しては、SSE2により浮動小数点処理を最適化)

格子サイズ	P4		Xeon		EV7	
	No copy	Copy	No copy	Copy	No copy	Copy
2*2*2*64	1251	957	811	598	1190	949
4*4*4*64	1020	878	633	536	1144	1034
8*8*8*64	1045	958	686	625	1140	1082
16*16*16*64	N/A	N/A	604	573	1122	1101

No Copy: 単体CPUだけで全演算を行う場合を想定
Copy: 並列化により、通信部分の「のりしろ」データをバッファにコピーする処理を含む

メモリバンド幅が重要！！

並列化性能 (1)

- EV7の16 CPUにおける並列性能
  (問題サイズ16*16*16*64の場合)

CPU数	1	2	4	8	16
並列化効率	**1**	**0.94**	**0.86**	**0.72**	**0.32**

- 8 CPUまで性能が5割を上回る
  - 高メモリバンド幅
  - HyperTransportの通信性能

並列化性能 (2)

- 32*32*32*64の大規模計算を想定
- 1 trajectory当たりの総演算数=20284 [Tflop]

ノード数	演算量／ノード [Tflop]	通信量／ノード [Tbyte]	演算･通信比 [Tflop/Tbyte]
1	20284.0	-----	-----
8	2535.50	77.1	32.9
64	316.938	19.3	16.4
512	39.6	4.82	8.22
4096	4.95	1.20	4.11
32768	0.619	0.301	2.06

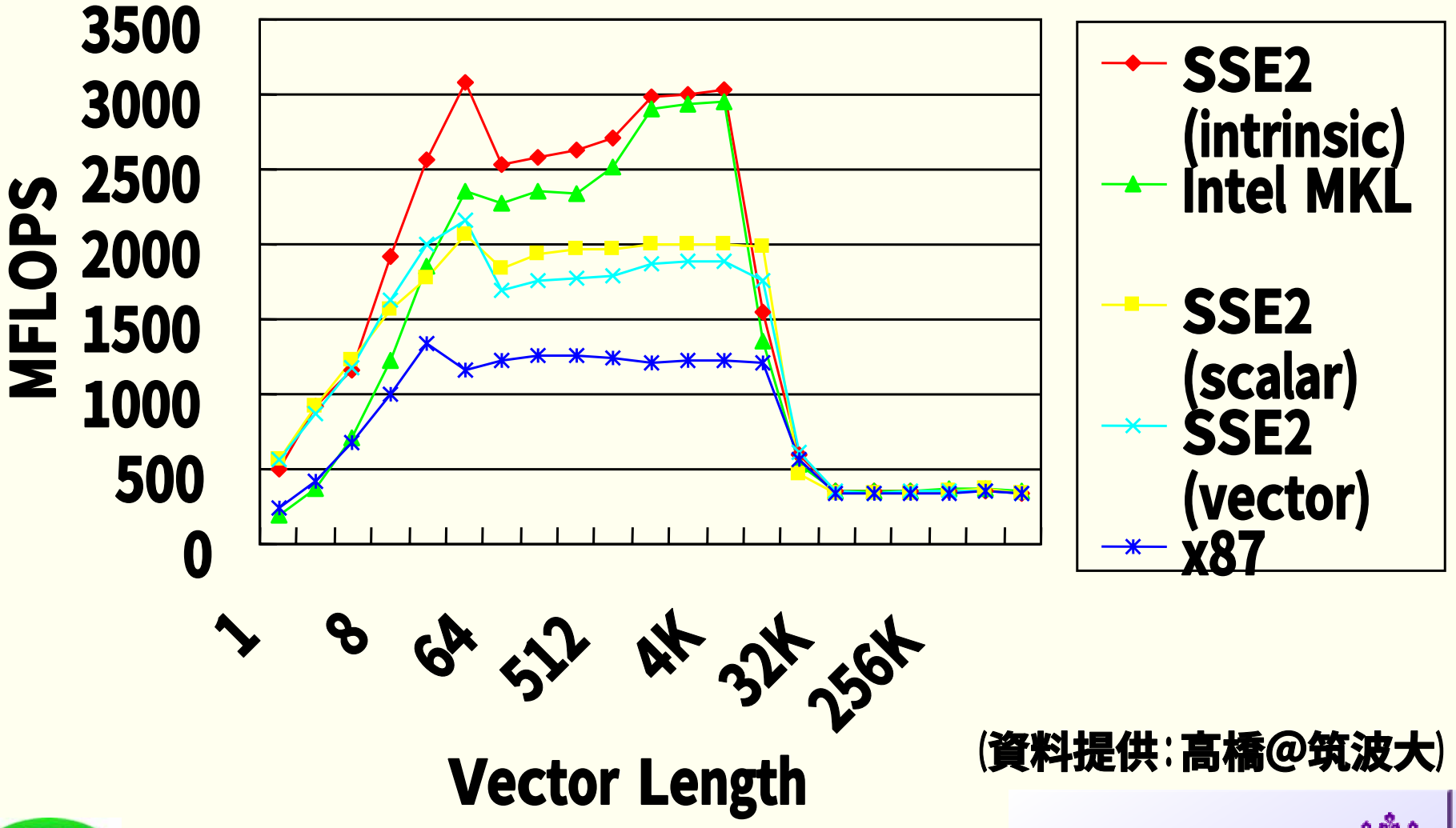
並列化性能 (3)

- 実効プロセッサ性能が1Gflops、実効通信性能が0.6Gbyte/sだとした場合の、32*32*32*64の問題サイズの5000 trajectory当たりの計算時間

CPU数	時間 [日]	通信時間の割合 [%]
512 (ns=8)	2768	17
4096 (ns=16)	405	29

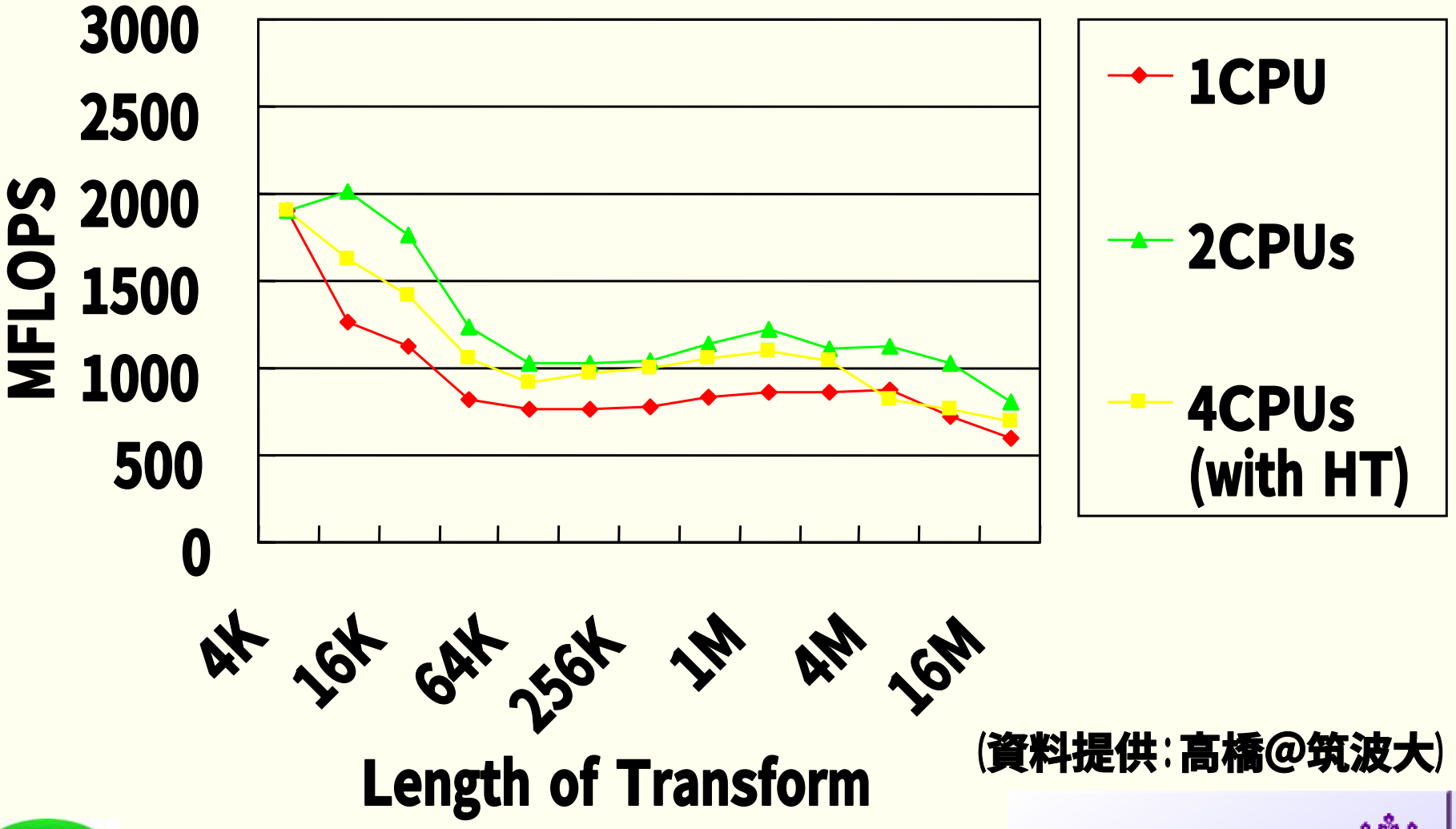
通信性能は非常に重要
(レイテンシよりもバンド幅)

SSE: ZAXPYの性能 (Xeon 3.06GHz, 1CPU)



(資料提供：高橋@筑波大)

FFTE 3.2 (SSE2)の性能 (Xeon 3.06GHz)



(資料提供:高橋@筑波大)

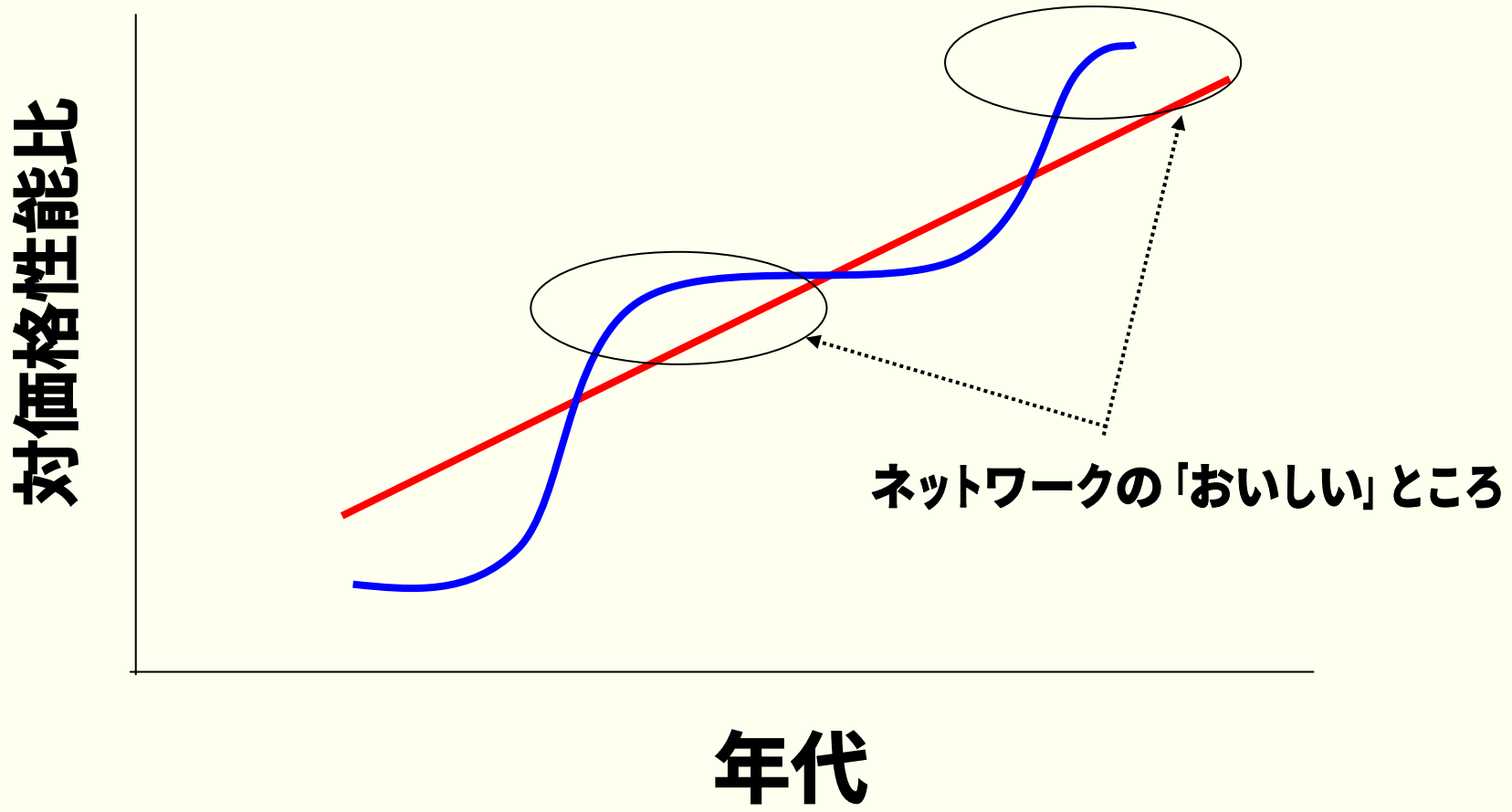
CPU性能の見通し

- メモリバンド幅の拡充
  - PC3200 (3.2 Gbyte/s) x 2等
- 大容量音チップキャッシュ
  - 3MB L2等
- short vector 機能の充実
  - SSE2
  - SSE3
- 設置スペース、熱･消費電力等を考えると現在のコモディティでかなりいける

ネットワークに関しては…

- 現在の高性能クラスタ向けネットワークはコモディティとは言えない
  (私の) コモディティの定義:
  「我々は開発費を払わなくて良い」
- CPU及び周辺の対価格性能比の延びに対し、現在の対応するネットワークの対価格性能比は不利
  (数年後はわからない)
- CPUの対価格性能比の延びとネットワークのそれはうまく釣り合っていない
- 大規模化すればするほど、特にスイッチのコストが大きくなる

CPUとネットワークの対価各性能比の変化



ネットワークのまとめ

- CPU性能はもちろん，ネットワークバンド幅が必要
- センターの性質から，アプリケーションをある程度絞り込める
- 「どことでもランダムに速く通信できる」ネットワークは実はいらない

  ⇒ネットワークコストを大きく抑える工夫ができるのでは？

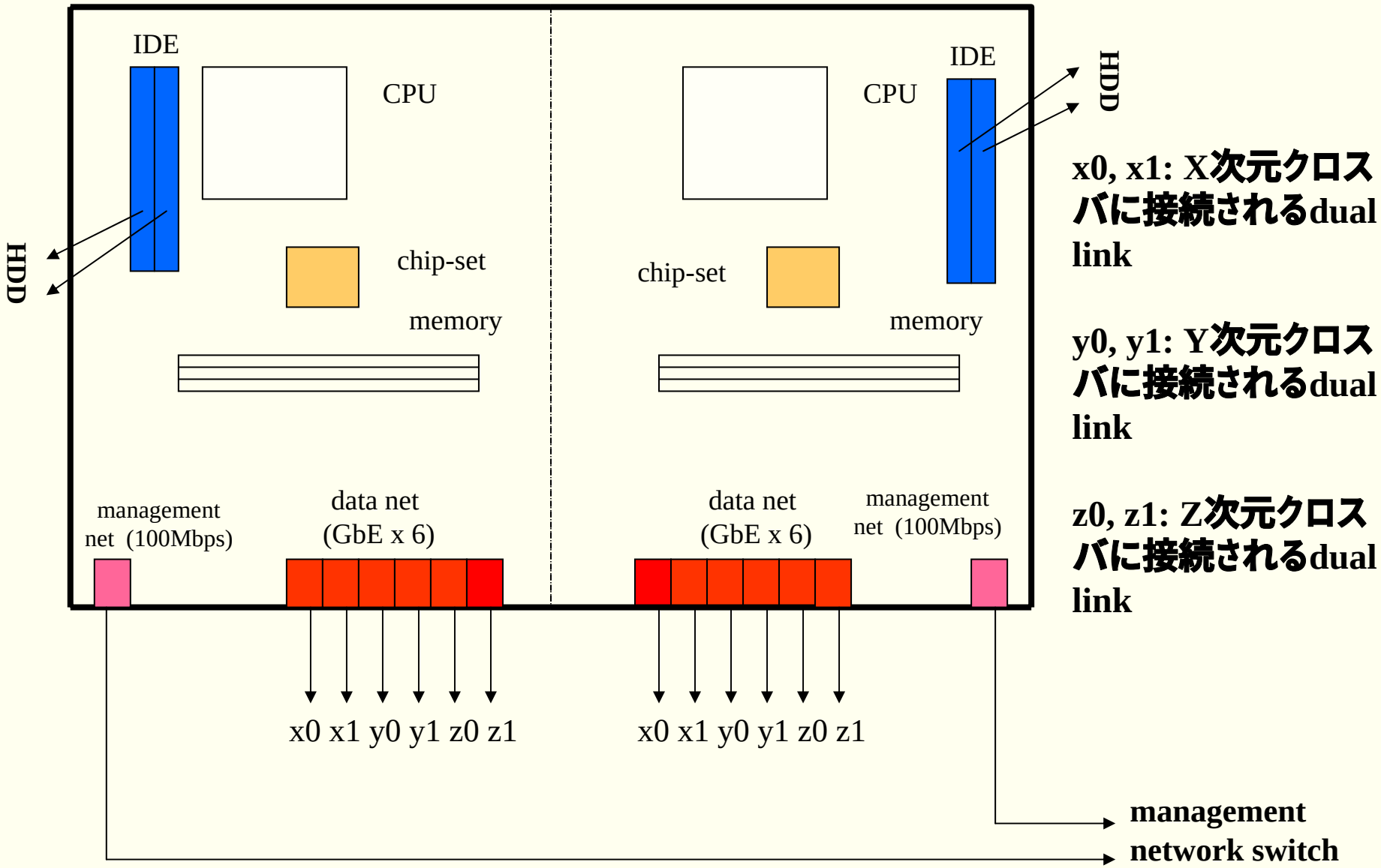
  ⇒例えば近接通信だけができればよい？

  ⇒高価なNIC+Switch 必ずしも必要ではない

目標とするクラスタ：CP-PACS II (仮称)

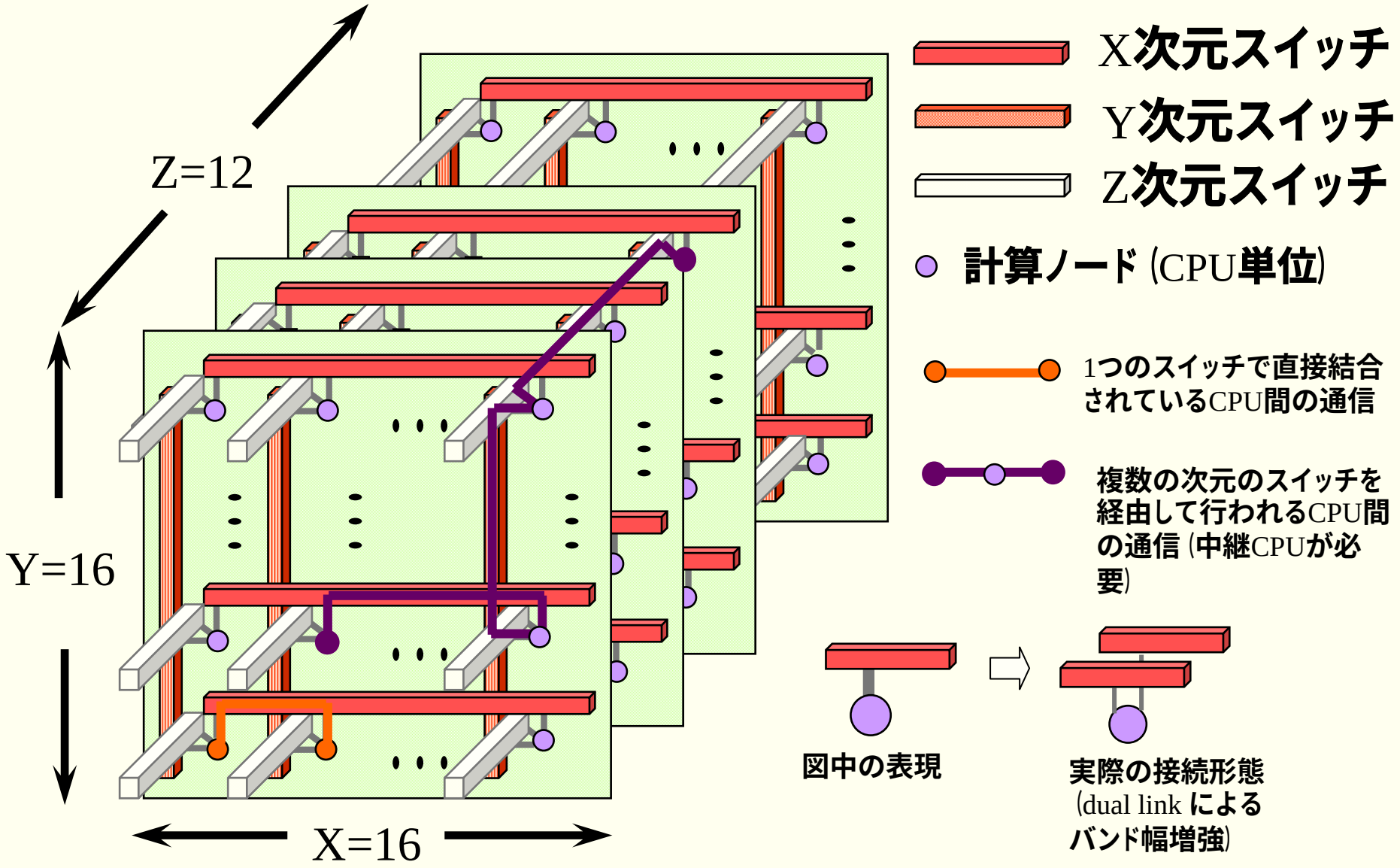
- ピーク性能：24.6 Tflops (4GHz CPUを想定)
- プロセッサ数：3072 CPU (1536 boards)
- ボード構成：独立した 2 CPU を1ボードに実装
  (各 CPU が独立にネットワークインタフェースを装備)
- 総メモリ容量：6.1 TB
- 総ディスク容量：1.05 PB (RAID0 mirror)
- ネットワーク：GbEthernet trunk (dual link × 3方向)
  ホストインタフェース：PCI-X dual
  Gigabit Ethernetトランクによるバンド幅増強
- ネットワーク構成：3-D Hyper Crossbar
  小規模コモディティスイッチ＋ソフトウェアルーティング
- 総筐体数：88 (Node=48, Switch=40)

ボード構成概念図


IDE
CPU
chip-set
memory
HDD
IDE
CPU
chip-set
memory
HDD
management
net (100Mbps)
data net
(GbE x 6)
data net
(GbE x 6)
management
net (100Mbps)
x0 x1 y0 y1 z0 z1
x0 x1 y0 y1 z0 z1
x0, x1: X次元クロスバに接続されるdual link
y0, y1: Y次元クロスバに接続されるdual link
z0, z1: Z次元クロスバに接続されるdual link
management
network switch

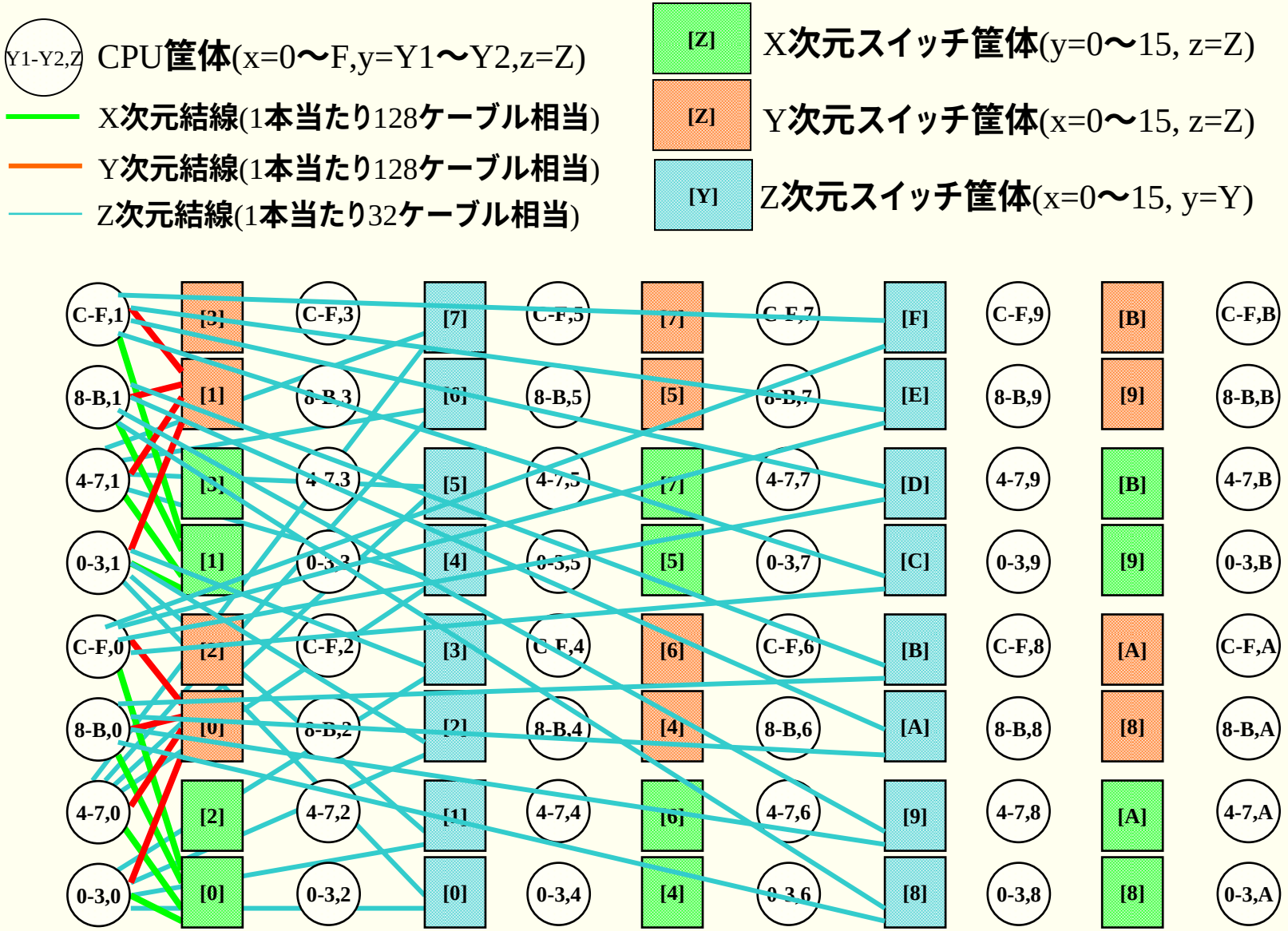
ノード(CPU)間の論理的結合 *(3次元ハイパクロスバ網)*

3072 CPU (1536 node) 構成



筐体の配置及び物理的結線

CPU([0-F],[0-F],[0-1])の範囲（全体の1/6=512 CPU）分の結線のみ図示



CP-PACS II アーキテクチャのまとめ

- ボードは作成するが部品は全てコモディティ
  ⇒短期間での開発と低コスト化
- CPUの実効性能はある程度のメモリバンド幅とSSE等のショートベクトル処理で稼ぐ
- 最低限のメモリバンド幅を稼ぐために dual CPU SMP にはしない
- コモディティベースの3次元ハイパークロスバーネットワーク
  ⇒近接通信は全てダイレクト通信
  　それ以遠は少しバンド幅が落ちるが通信可能
  ⇒実質的なバンド幅を犠牲にせずコストを大幅ダウン

まとめ

- 計算科学研究センターの主要計算リソースの更新を目指す
- 今後数年間はコモディティベース（＋ちょっとした工夫）のPCクラスタが有利
- 我々のアプリケーションの特性を生かし、無闇に全てのノード間のバンド幅を追求しない
  ⇒対価格性能比を追求
- このシステムを礎に、さらなる高性能・大規模システムを目指す