THE POSSIBILITIES ARE INFINITE FUJITSU

PM通信ライブラリの最新情報

富士通研究所
住元　真司

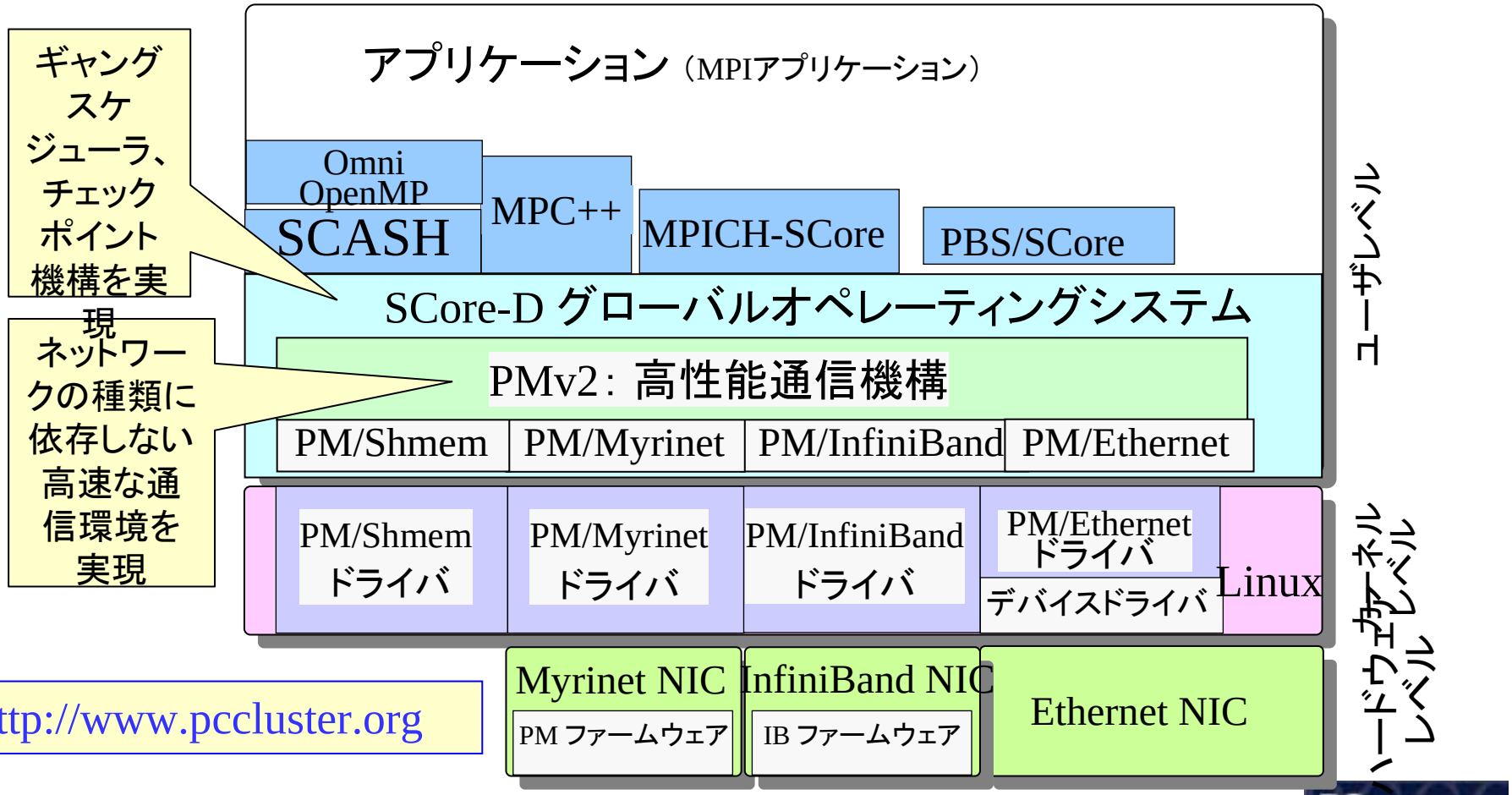
PC Cluster Consortium

発表の概要

- SCore＋PMの目指してきたもの
- PM開発の歴史
- １０年経ってみて
- 新しい通信機構PMXの開発状況
- 複数Gigabit Ethernetを用いた通信機構PM/Ethernet-HXBについて

SCoreクラスタシステムソフトウェア

- 1990年代にRWCプロジェクトで開発



http://www.pccluster.org

SCore＋PMが目指してきたもの

- シームレスクラスタコンピューティング
  - プロセッサ、ネットワークの違いを意識しない
- 高性能、高信頼
  - 高いアプリケーション性能
  - チェックポイントリスタート
- オールインパッケージ：プログラム開発から実行まで
  - プログラミング、デバッグ、チューニング用のツール
    - MPI、商用コンパイラ、デバッグツール
  - 実行ランタイム：　マルチユーザモード、チェックポイント

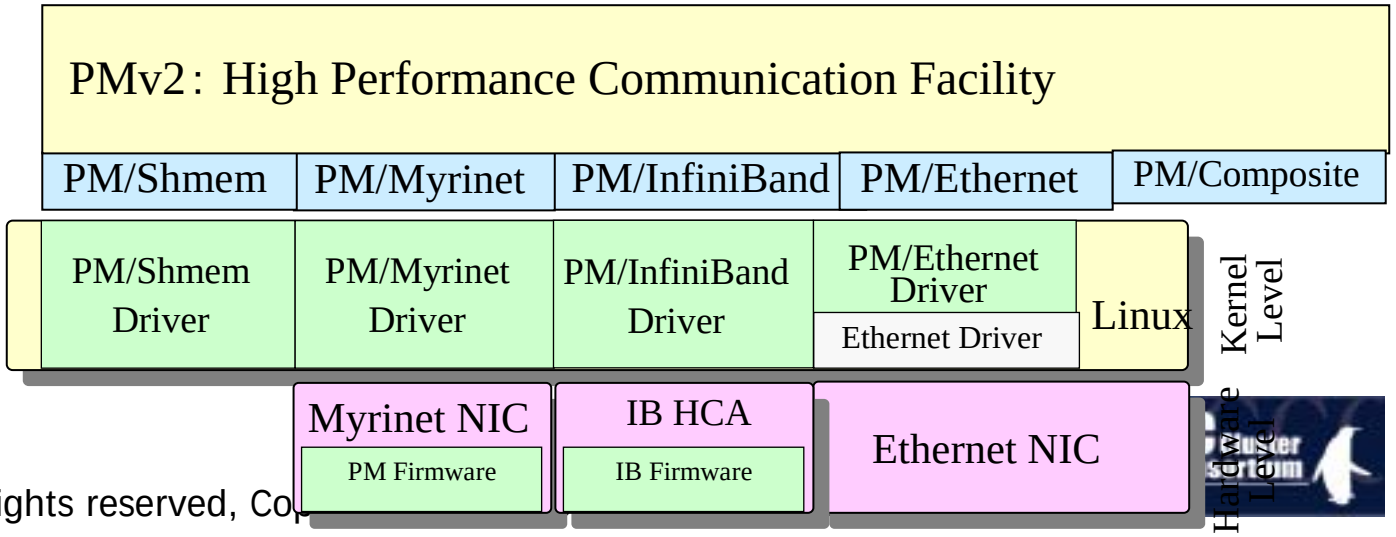
- PCクラスタのあるべき姿を追いかけて10年

PMv2の目指してきたもの

- 高い通信性能(1995-)
  - 高バンド幅、低遅延通信
- 先進機能のサポート
  - シームレスクラスタ通信（1999-）
    - マルチネットワークサポート
    - 単一ネットワーククラスタ、マルチネットワーククラスタ、クラスタof クラスタ
  - RDMA通信(Zero-Copy通信)（1997-）
  - チェックポイント・リスタートサポート(1998-)

PMv2通信機構の概要

- 既存OSの通信プロトコル処理オーバヘッドを取り除き高い通信性能（高バンド幅，低遅延）を実現するために開発された
- 開発当初のPMはMyrinet専用であったが、PMv2から複数種類のネットワークをサポートしている
  - PM/Myrinet, PM/Ethrernet, PM/Shmem, PM/SCI, PM/UDP(Agent),PM/InfiniBand (-FJ[Fujitsu], -TS[TopSpin])



PMv2の機能

- 高性能通信:
  - メッセージ通信とZero-Copy通信
- 複数種類のネットワークをサポート
  - シームレスな通信環境を実現
- チャネルとコンテキストによる多重化通信
  - チャネルは通信媒体、コンテキストは通信の状態
  - マルチユーザでのクラスタ利用を効率的に実現
- ノード構成をプログラムで変更可能
  - SMPクラスタ上の実行でSCore-Dが利用している
- いろいろな機能をもつPMを追加可能
  - 通信プロファイラ用など

PM Development History

- PMは1995年手塚氏（現在、産総研）により開発
  - Myrinet専用, Sun SS20s, SBUS, SunOS 4.1.x
- 1996年、Pentium(NetBSD)上に移植された
- 1997年、Linuxに移植、i386, Alpha(SCore 2.x)
- 1998年、GigaE PM開発
  - Gigabit Ethernet上のPM
- 1999年、異種ネットワーク対応(SCore 3.0)
  - Ethernet, Shmem, Myirnet, UDP
- 2001年、SCI, Myrinet 2000サポート
- 2003年、InfiniBand, Myrinet XP(2XP)サポート

PMv2 高性能通信のための技術

- 信頼性と順序性を確保したデータグラム通信: PMv2
  - TCP/IP等コネクションベースの通信は大規模クラスタには適さない
- ユーザレベル通信: PM/Myrinet, PM/InfiniBand
  - 低遅延ネットワークのためにシステムコールのオーバヘッド削減
- Zero-Copy 通信: PM/Myrinet, PM/InfiniBand
  - ホストCPUのコピーオーバヘッド、メモリバンド幅を節約し、高い通信バンド幅性能を実現
- Network Trunking : PM/Ethernet
  - 複数のEhternet NICを用いて高いバンド幅性能を実現
- カーネルレベルの RMA: PM/Ethernet-kRMA
  - 送受信の同期とコピーオーバヘッド削減で高い通信バンド幅を実現

異種クラスタ間をシームレスに結合
理研スーパコンバインドクラスタで採用

THE POSSIBILITIES ARE INFINITE

FUJITSU

PC Cluster A

512nodes/1024CPU(CPU:Xeon 3.06GHz)
Peak Performance: 6.2 TFLOPS,
Linpack Performance 4.5 TFLOPS
Interconnect: InfiniBand(8Gbps)

XP Fortran
MPI

Users

Firewall/VP
N router

Total: 1024nodes/2048CPU
Peak Performance : 12.4 TFLOPS
Linpack Performance: 8.7 TFLOPS

Sun Fire V480
etc..

Front-end Machines

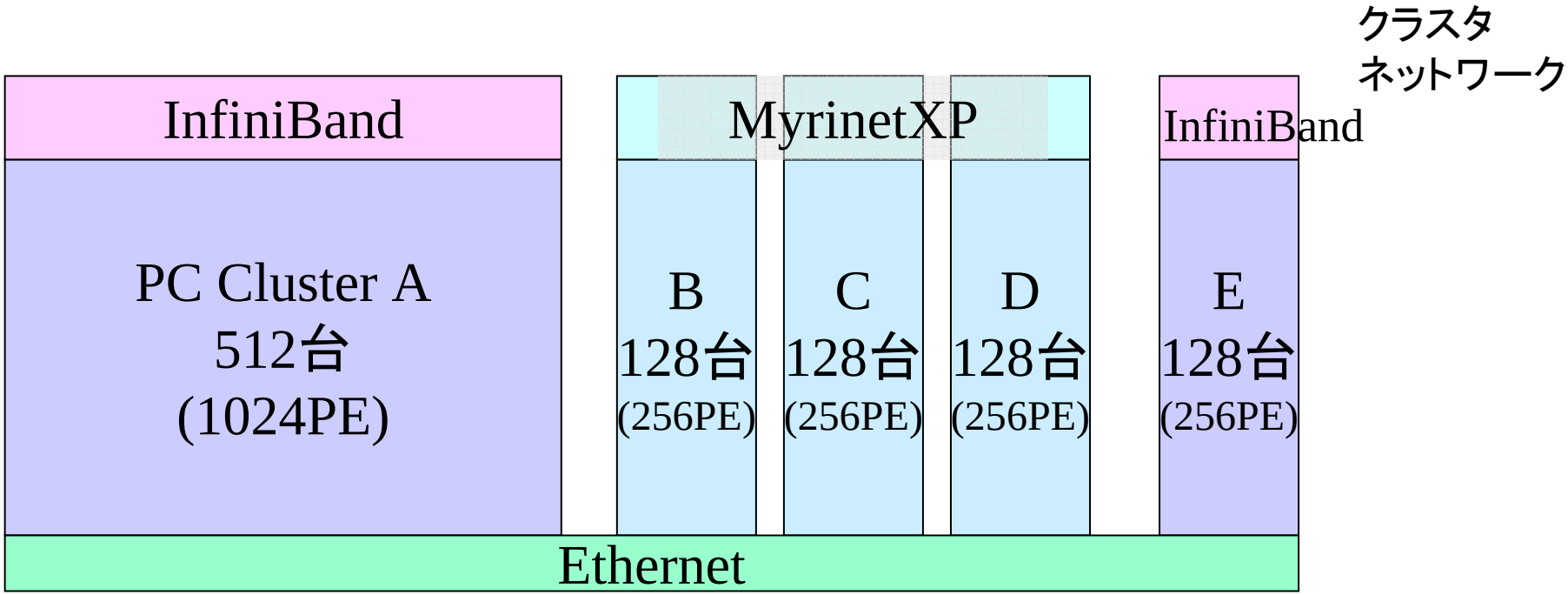
512nodes/1024CPU
[128nodes/256CPU]×4 sets
Peak Performance: 6.2 TFLOPS
Interconnect: MyrinetXP(2Gbps)
and TopSpin InfiniBand(8Gbps)

PC Cluster B/C/D/E

2004/6 : Top500 Rank 7

2005/4 : 第34回 日本産業技術大賞 文部科学大臣賞受賞

PMv2におけるクラスタofクラスタ



- 相手先によりPMライブラリが通信機構を自動選択
  - ノード内は共有メモリ: PM/Shmem
  - クラスタ内はクラスタネットワーク: PM/InfiniBand, PM/MyrinetXP
  - クラスタ間はEthernetネットワーク: PM/Ethernet

10年経ってみて、、

- 単なるMPIランタイムとしての利用が多い
  - 特に企業向けはアプリケーション開発は少ない
- ISVサポートが課題
  - 実行バイナリがstatic、バージョン毎にISVに対応依頼
- カーネルへの変更パッチが導入の障害に
  - RedHat ASなど
- サポートネットワークが多くなりライブラリ肥大
  - Myrinetだけでも5種類
- 新規ネットワークデバイス対応の問題
  - デバイス仕様を熟知しないと実装できない
  - サポートの問題

PMXの設計指針

- 高い通信性能にこだわる
  - これがPMXの存在意義
- より広く使ってもらえるように
  - Kernelへのパッチ不要化
  - ISVバイナリサポート強化
    - Intel MPIのDAPLデバイス提供(DAPL for PMv2)
  - 標準インターフェイスのサポート: OpenIB for PMv2
- シンプルかつ柔軟な構造
  - ネットワークデバイスを実行時にロード
  - ベンダー通信ライブラリ、標準ライブラリを活用
    - Myrinet, InfiniBand他

PMXの方針が決まるまで

- PMのAPI自体もやめてスクラッチから議論
  - 標準API(Ex. DAPL，MX etc.) ベースへの移行も検討して開発部会で議論を行なった
    - DAPLベース: 並列計算向きでなく、API処理が重い
- 結局、PM/Myrinet MX，DAPL for PMを実現することになった
  - ただし PM APIについては今後議論して改良する

PMX エンハンス

- 高い通信性能にこだわる
  - マルチリンクでの性能向上： PM/Ethernet-HXB etc..
  - 新規デバイスへの対応： PM/Myrinet-MX,
- より広く使ってもらうために
  - Linuxカーネルへのパッチをなくす
  - Intel MPIサポート（DAPL for PMv2）
- シンプルかつ柔軟な構造
  - ユーザライブラリのダイナミックロード化
  - ノード数制限の撤廃
  - ベンダー通信ライブラリ、標準ライブラリとの共存：
- その他
  - 2.6カーネル対応
  - RDMAの64bit対応

開発状況

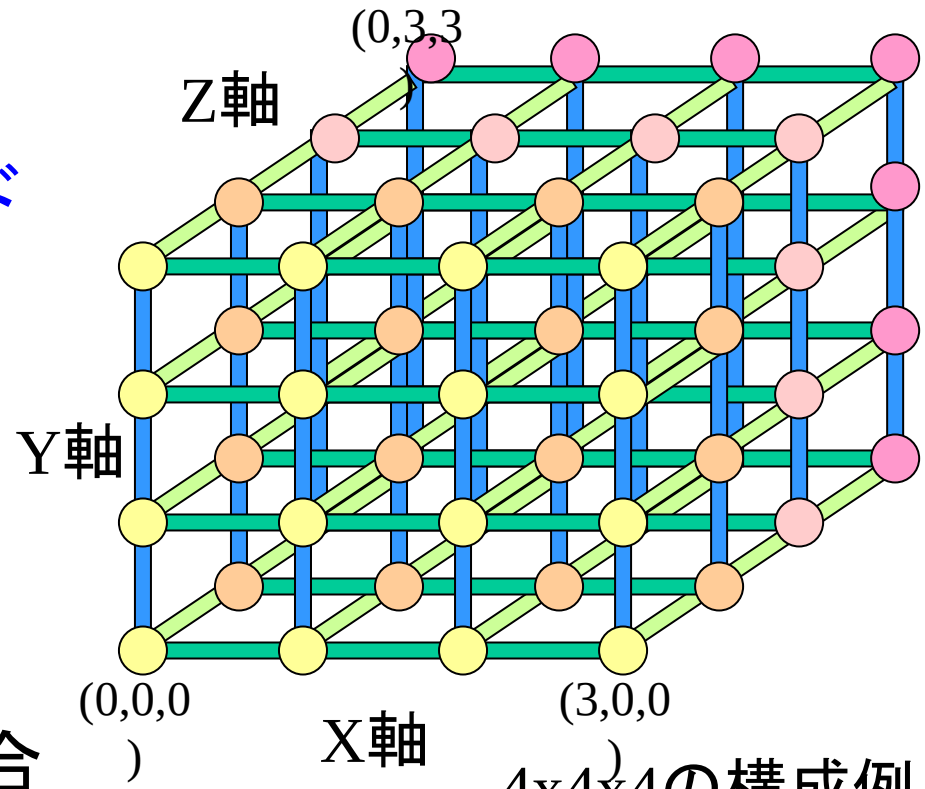
- 既に実装済み(SCore 5.8.3)
  - ユーザライブラリのダイナミックロード化
  - Linuxカーネルへのパッチをなくす
  - 2.6カーネル対応
- 実装中
  - ノード数制限の撤廃
  - Intel MPIサポート（DAPL for PMv2）
  - RDMAの64bit対応
  - PM/Ethernet-HXB
- 計画中
  - PM/Myrinet-MX: ターゲットとしてMyrinet10G
  - OpenIB for PM

PM/Ethernet-HXB

- PM/Ethernet(1999年開発)から6年、Commodityネットワークを使った究極の通信機構を作りたかった
  - 専用インターコネクトにどれだけ迫れるか？
- PACS-CSシステム向けに開発:
  - PACS-CS: 筑波大計算科学研究センターで開発中の14TFlopsのPCクラスタ
- 設計コンセプト
  - バンド幅スケーラブル: Gigabit Ethernet 10本クラスまで
  - 柔軟なトポロジー: Fat Tree, Nth-Crossbar結合
  - 安価な少数ポートスイッチで数千ノードを結合
- 筑波大からPCCCにContribution予定

PACS-CSの概要

- 計算ノード:
  - IA32互換 2.8GHz
    シングルプロセッサノード
  - クラスタ通信用:Gigabit Ethernet x 6系統
    750MB/s (1.5GB/s)
  - I/O用，管理用Ethernet
- ネットワーク結合方式
  - 3次元Hyper Crossbar結合
- ノード数
  - 2560ノード（16x16x10）



4x4x4の構成例

Z
Y
X
がそれぞれEthernet SW 2系統を示す

PM/Ethernet-HXBの特徴

- 市販Ethernetを用い通信バッファ間でのZero-Copy通信を実現 (これまでは1Copy)

- 極限まで通信プロトコルオーバヘッドを削減

詳細は、今後論文で発表予定

PM/Ethernet-HXBの性能

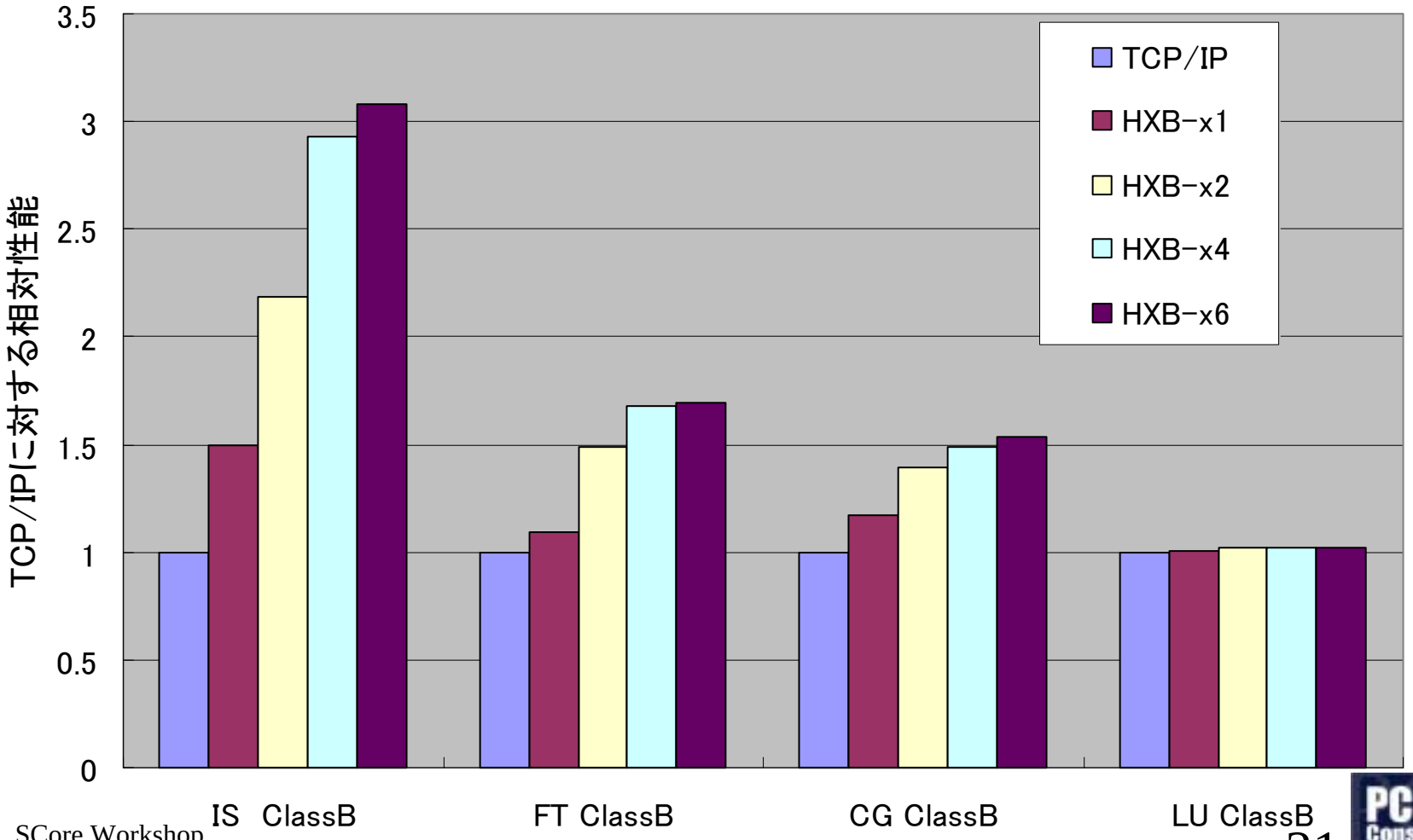
- Xeon 3.6GHz(EM64T)上の性能

	PM レベル 単方向	PMレベル 双方向	MPIレベル 単方向
Giga x 2	247MB/s	481MB/s	244 MB/s
Giga x 4	494MB/s	876MB/s	493 MB/s
Giga x 6	740MB/s	1190MB/s 3D:1400MB/s	737 MB/s
Giga x 8	918MB/s	1202MB/s	862 MB/s
Giga x 9	1035MB/s	1210MB/s	858 MB/s

3Dは3次元転送、その他は1次元転送、MPIは独自プログラム利用

NAS並列ベンチマーク性能：MPICH

- Xeon 3.6GHz(EM64T) 4 Nodeの結果 MPICH



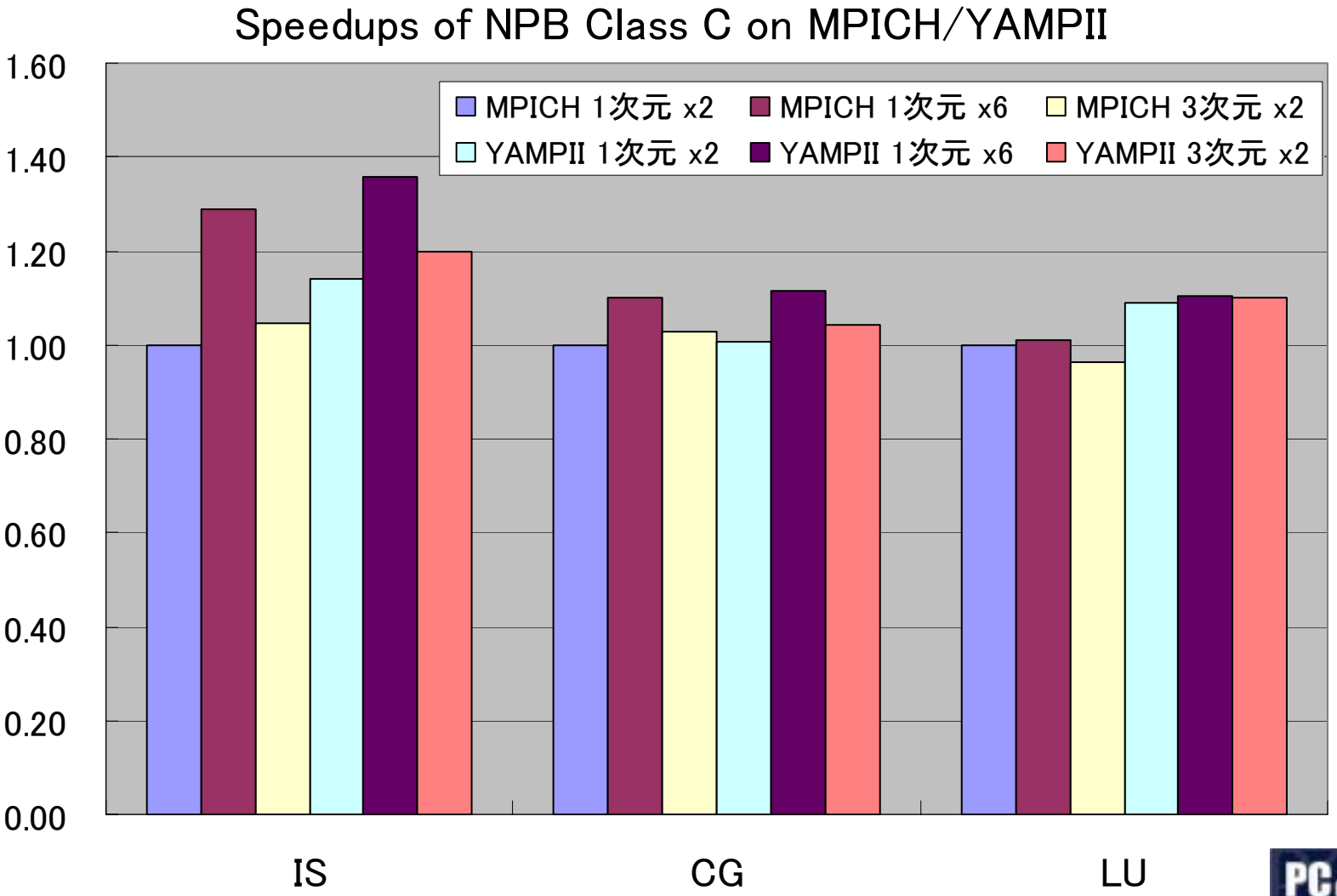
MPI通信性能比較：MPICH vs YAMPII

	MPICH 片方向	MPICH 双方向	YAMPII 片方向	YAMPII 双方向
Giga x 2	244 MB/s	478 MB/s	244 MB/s	490 MB/s
Giga x 4	493 MB/s	791 MB/s	494 MB/s	896 MB/s
Giga x 6	737 MB/s	811 MB/s	739 MB/s	937 MB/s
Giga x 8	862 MB/s	800 MB/s	915 MB/s	932 MB/s
Giga x 9	858 MB/s	790 MB/s	890 MB/s	921 MB/s

- PM/Ethernet-HXBはMPIレベルで800MB/s以上の性能を実現：専用インターコネクトに匹敵する性能
- YAMPII利用で900MB/sクラスの通信を達成

Powered by SCore

MPI性能比較：NAS並列ベンチマーク



PC Cluster Consortium

まとめ

- PM通信機構の１０年
- PMXの概要
  - 高い通信性能にこだわる
    - 複数ネットワークによる通信バンド幅の向上
  - より広く使ってもらうために
    - バイナリ互換、ISV対応
  - シンプルかつ柔軟な構造
- PM/Ethernet-HXBの概要と性能
- 要望、ご意見あればよろしくお願いします.