T2Kオープンスパコンの概要

東京大学
石川　裕

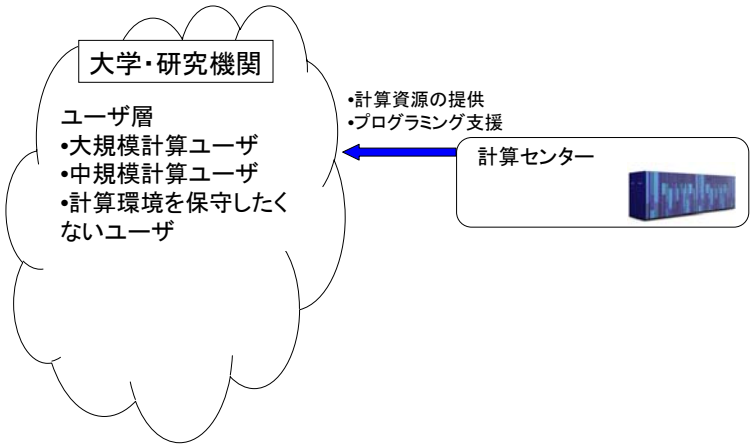
概要：筑波大学、東京大学、京都大学の3大学は、共同で、次期スパコン調達の仕様を検討してきた。本講演では、次期スパコン設計にあたっての基本思想および共通仕様部分を紹介する。さらに、東京大学情報基盤センターにおけるスパコン利用促進および今後の研究開発の取り組みを紹介する。

キーワード：オープンスパコン、センター運用、Linux

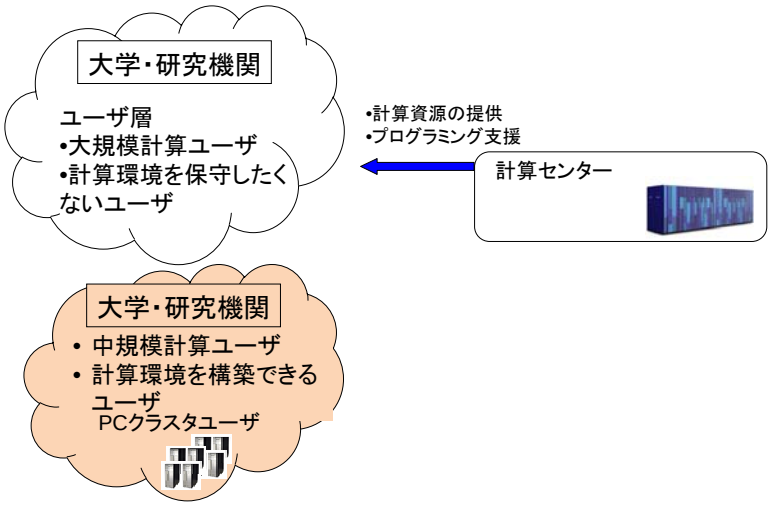
Outline

1. 背景
   1. センターユーザーの変遷
   2. PCクラスタの台頭と限界
   3. 計算ニーズ、利用形態の多様性
   4. 次世代スーパーコンピュータ開発
2. 筑波大・東大・京大の共同研究
3. オープンスパコンの理念
4. オープンスパコンの基本要件・基本仕様
5. ネットワーク性能の検証
6. 構成例
7. センターの今後
8. まとめ

背景：計算センターとユーザ



背景：計算センターとユーザ



背景：PCクラスタユーザの現状

- 満足しているユーザ
  - 小中規模科学技術計算ユーザ
  - Embarrassingly Parallel Applicationユーザ
    - モンテカルロ手法など、並列度はあるが通信をほとんどしない計算
- 不満を感じているユーザ
  - 性能が出ない
    - ネットワークの問題
  - 保守・維持が大変
    - よく壊れる
      - 熱設計がされていない
    - システムソフトウェアの保守
    - 電気代がかかる

大規模化への道筋

背景：計算ニーズの多様性

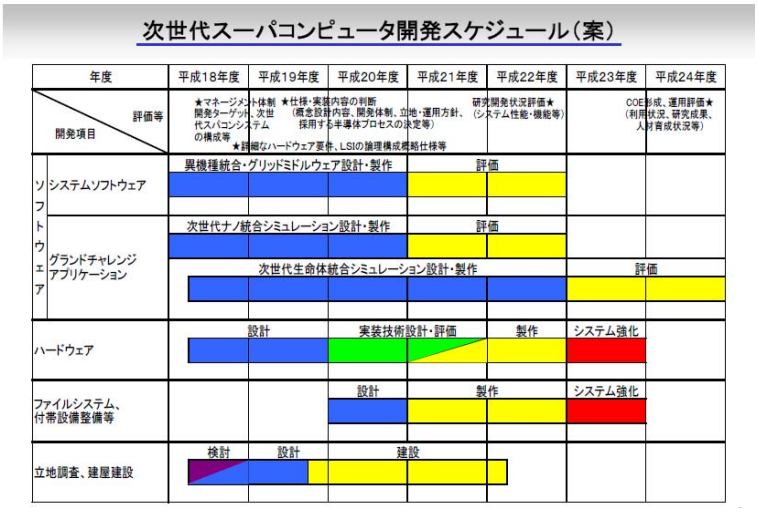
- 従来スパコン応用分野
  - 超規模計算科学、超規模計算工学
  - 100テラフロップス超、数十 Tbytes超主記憶
- 新興応用分野
  - 大規模ゲノム情報処理
  - 超規模アーカイブ検索
  - コンピュータグラフィックス
- PCクラスタユーザ
  - 小中規模科学技術計算
  - Embarrassingly Parallel Applications

新しい計算センターはこれらニーズを吸収するマシンが必要

背景：利用形態の多様性

- バッチ・インタラクティブ
  - 従来の利用形態
- WEBユーザインターフェイス
  - ポータル
- グリッド
  - データ共有、ワークフロー

背景：次世代スーパコンピュータ開発

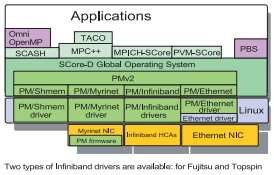


出展：http://www.mext.go.jp/b_menu/houdou/18/06/06061214/002.pdf

背景：大学の知

- 筑波大学
  - PACS-CS
    - 筑波大学計算科学研究センターの宇川彰教授、佐藤三久教授、朴泰祐教授らにより設計開発している並列コンピュータ。2006年6月に公表されたTOP500リストにおいて34位にランクされた。日本国内メーカ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を有する。

- 東京大学
  - SCore
    - 経済産業省（当時、通商産業省）リアルワールドコンピューティングプロジェクトで開発されたシステムソフトウェア。東京大学の石川裕教授（当時、並列分散システムソフトウェア研究室室長）がSCoreの開発を主導。現在、PCクラスタコンソーシアム（会長：石川裕）により開発・配布が継続されている。SCoreは、PACS-CSや理化学研究所のSuper Combined Clusterなど、計算センター運用されているスーパーコンピュータで使用 されている。



Two types of Infiniband drivers are available: for Fujitsu and Topspin

- 京都大学
  - MegaProto
    - 独立行政法人 科学技術振興機構の戦略的創造研究推進事業における「超低電力化技術によるディペンダブルメガスケールコンピューティング」課題において研究開発された省電力クラスタ。京都大学 中島浩教授主導の元に研究開発された

筑波大・東大・京大の共同研究

- 目的
  - 2007年～2008年に出荷されるCPU技術を想定し、センタ運用できるスパコン仕様を提示
    - ハードウェア仕様
      - ノード構成
      - ネットワーク
    - ソフトウェア仕様
      - オペレーティングシステム
      - コンパイラ＆プログラミング環境
      - ライブラリ
      - 商用アプリケーション
- 手法
  - 現状のコモディティCPU技術の検証
  - 必要とされる通信性能とその実現性の検証
  - 現状のコンパイラ技術、プログラミング環境、オペレーティングシステムの検証
  - 新しいタイプのユーザに対する運用方法の検討
  - 共通ベンチマーク策定

オープンスパコンの理念

- 基本アーキテクチャのオープン性
  - コモディティ高性能プロセッサを基本
    - コンピュータ市場を牽引しているコモディティ高性能プロセッサを使用することにより、最新技術を使用することにより、高性能かつ低消費電力を実現したシステムを導入することが可能
- システムソフトウェアのオープン性
  - オープンソースに基づく先端ソフトウェア技術を基本
    - 多くのユーザが使用するこれら資産をシームレスに利用できる環境を提供することにより、より多くのユーザが大規模並列処理環境へ移行することが促進できます
- ユーザのニーズに対するオープン性
  - 従来の計算センターユーザでないニーズに対して応える
    - 大規模ゲノム情報処理、大規模データマイニング

計算センターから研究室までをカバーするアーキテクチャ

オープンスパコンの基本要件

- 大規模科学技術計算ユーザに対するサービス
  - 大規模ノード数を有するプラットフォーム
- 京速コンピュータへのブリッジ
  - ペタ規模を利用するアプリケーション開発ユーザのサポート
  - ペタスケールプログラミング言語＆ライブラリ開発ユーザのサポート
- 新しいユーザに対するサービス
  - 従来の科学技術計算ユーザに加えて新しいユーザのサポート
    - データマイニング/データ検索
    - Computer Graphics
    - Embarrassingly Parallel Applications
- グリッド技術提供
  - データ共有
  - キャンパスからの利用
- 最新プログラミング環境提供
  - GNU BSD/Linux上で提供されているコモン環境の提供
- 各基盤センターの現ユーザプログラミング環境の継続＆サポート
- 電力消費量・発熱量

オープンスパコン仕様

規定するもの

- ハードウェア
  - 基本構成
  - ネットワーク性能＆ネットワークトポロジ
  - 管理系ネットワーク＆機能
- 基本ソフトウェア
  - オペレーティングシステム
  - MPI通信ライブラリ性能
  - 数値計算ライブラリの一部
  - プログラミング環境の一部
  - 商用アプリケーションの一部
- 基本ベンチマーク

規定しないもの

- 運用の継続性を必要とする可能性があるもの
  - バッチ処理システム
  - ファイルシステム
  - コンパイラ
- サイズ的要素
  - ノード＆メモリサイズ
  - ディスクサイズ

ノード性能とネットワーク性能

- コモディティ高速ネットワーク
  - Infiniband DDR : 16Gbps = 2.0GB/s
    - シグナルレベルでは20Gbpsだが、8B10Bエンコーディングされているので実効性能は16Gbps
  - Myrinet 10G : 10Gbps = 1.25GB/s
  - 10G Ethernet : 10Gbps = 1.25GB/s
- PCI Expressによる制限
  - 2.5Gbps/lane, 8B10Bエンコーディング
  - 128/256/512/… /4096 byte Payload, 12-16 Byte header, 4 Byte ECRC 8 Byte Data Link and Physical Layer overheads
  - すなわち、PCI Express x8では、Infiniband DDRの性能を生かせない

- ノード理論性能150Gflops時

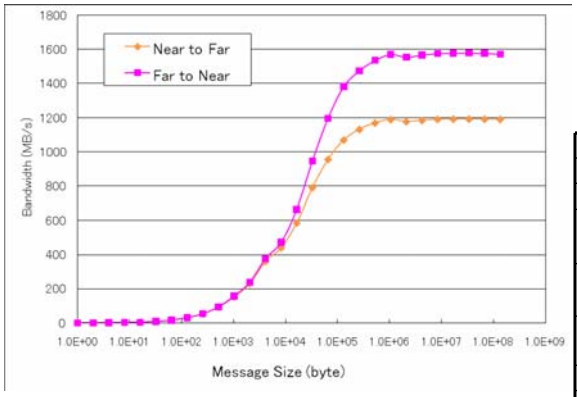
本数	Gbyte/sec	byte/flops
1	2	0.013333
2	4	0.026667
4	8	0.053333
8	16	0.106667

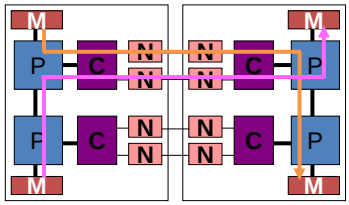
本数	Gbyte/sec	byte/flops
1	1.25	0.008333
2	2.5	0.016667
4	5	0.033333
8	10	0.066667

	Node (Gflops)	Network (Gbyte/sec)	Byte/Flop
SCore-III	1.866	0.25	0.134
Riken	12.24	1.064	0.087
TSUBAME	76.8	2	0.026
T2K	150	5	0.033

ネットワーク性能の検証

- NUMAとネットワーク
  - メモリ位置とNIC(Network Interface Card)位置による性能の違い



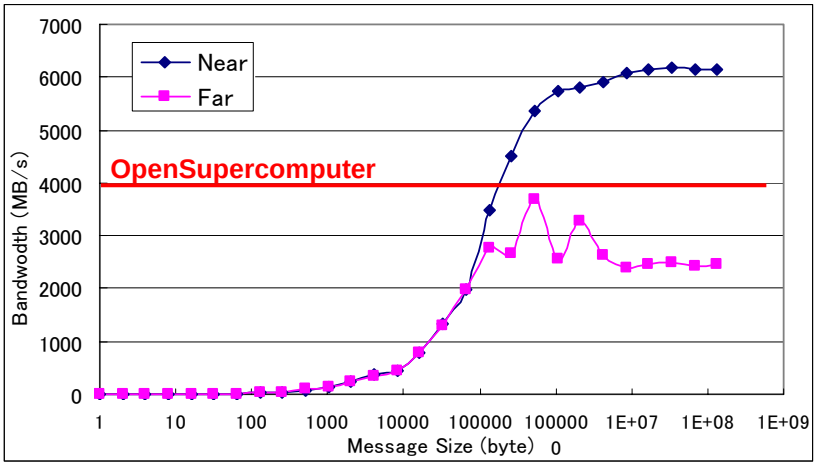


注：性能はBIOS設定に依存

CPU	Opteron 2212 (Dual Core, 2.0GHz) x 2
メモリ	DDR2 667 (Dual Channel) 2GB
チップセット	nVIDIA nForce Pro 3600 + 3050
マザーボード	Tyan Thunder n6650W (S2915) BIOS version: 1.01D beta
バス	PCI-Express x8 2本 (from NFP3600) PCI-Express x8 2本 (from NFP3050)
Infiniband	Mellanox InfiniHost III Lx (x4 DDR) Memfree(NICにメモリを持たない)
OS	Fedora Core 6 (Linux 2.6.18)
MPI	YAMPI

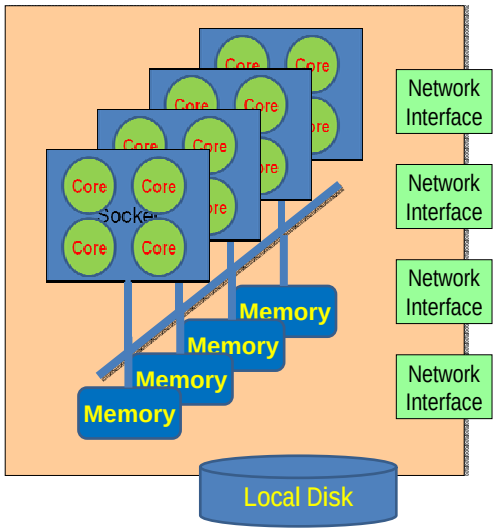
ネットワーク性能の検証

- NUMAとネットワーク　x　4

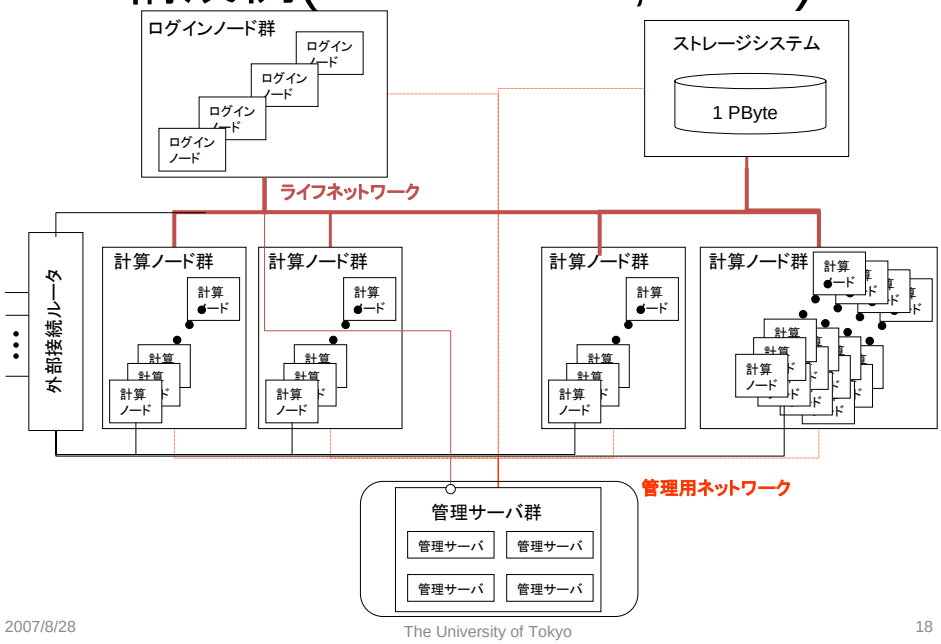


基本構成 (As of March, 2007)

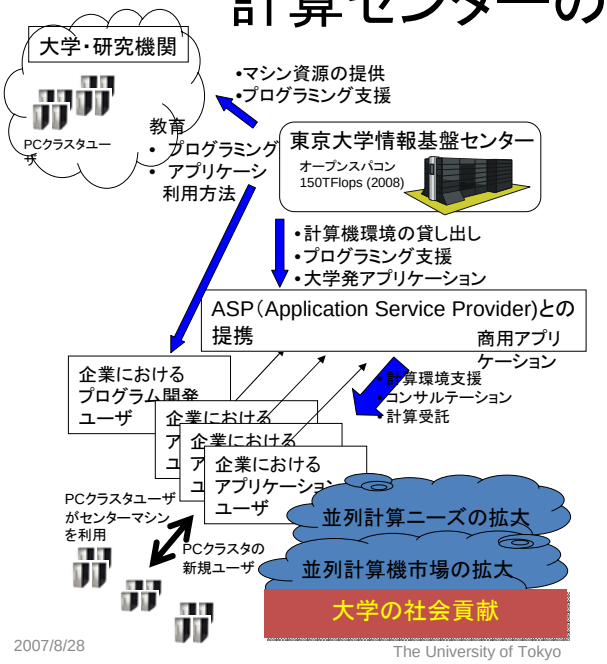
- ノードアーキテクチャ
  - 64ビット IA32アーキテクチャ
  - 16コア以上
  - 32 Gbyteメモリ、40Gbyte/sec以上の物理メモリ転送容量
  - 128 Gbyte以上のローカルディスクあるいはネットワーク経由のディスク
- ノード間ネットワーク
  - 物理性能
    - 5 Gbyte/sec
  - MPI性能
    - 4Gbyte/sec



構成例(As of March, 2007)

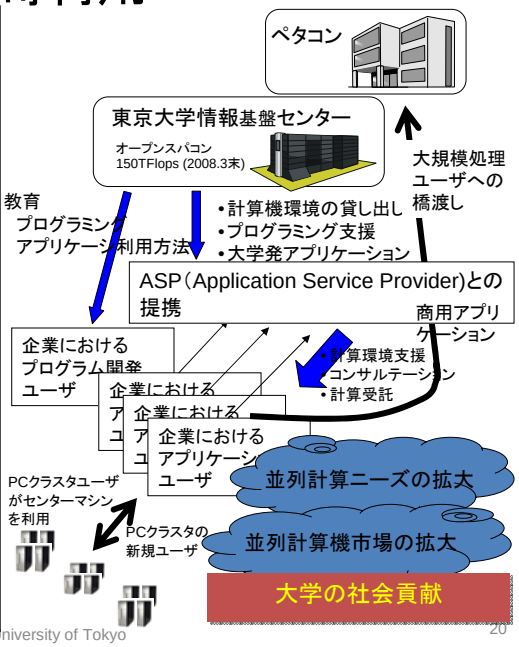


計算センターの展望

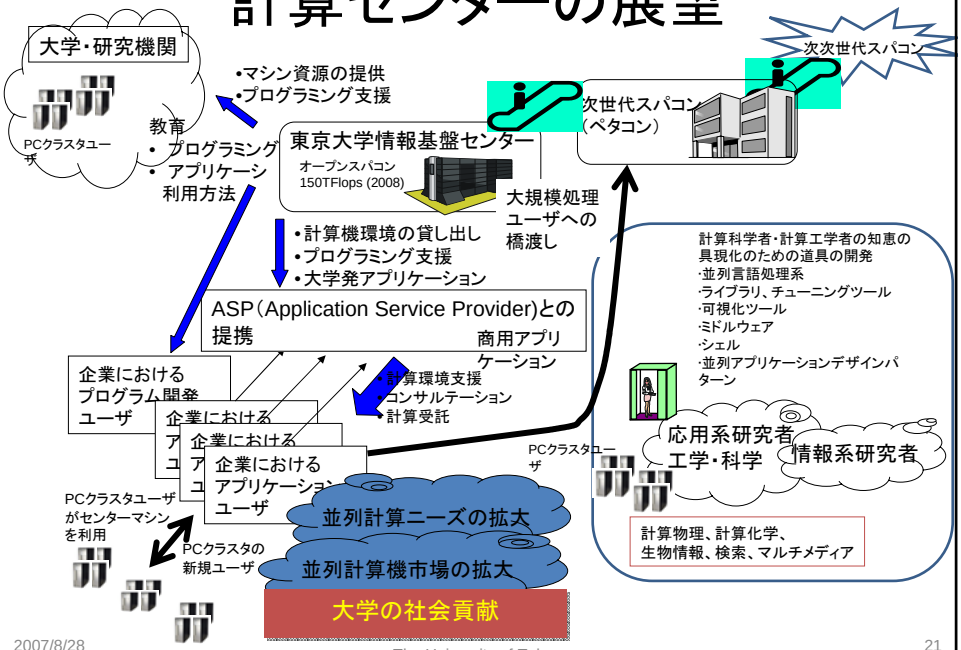


民間利用

- 民間利用の意義
  - 大学が持つ計算資源および大学発の先端的アプリケーションを産業界に開放し、従来の科学技術計算のみならずゲノム情報処理や超規模データ検索などの新しい応用分野を開拓していく。
  - センターの計算資源（オープンスパコン）を軸にしてPCクラスタユーザの取り込み、PCクラスタからペタコンピュータへの橋渡し、あるいはPCクラスタへの下方展開を担う。
  - 並列計算機ユーザが拡大することにより、計算需要が増大し、スパコン市場も増大
    - PCクラスタ、オープンスパコン、ペタコンピュータの普及
- 体制
  - センターによる1次サービス供給
    - 安定かつ高性能計算システム環境を提供
    - ユーザが開発するソフトウェアに対して、並列プログラミング＆最適化支援
    - アプリケーションサービスプロバイダに対する支援
  - 企業による2次サービス供給
    - アプリケーションサービスプロバイダ
      - 大学が開発したアプリケーションや商用アプリケーションを使用してもらうためには、支援体制が必要であり、ASPビジネスモデルをベースにする



計算センターの展望



まとめ

- T2Kオープンスパコン
  - PCクラスタユーザから大規模計算ユーザを支援
  - ペタコンへの橋渡し
- 今後のセンターの役割
  - 計算資源提供
  - プログラミング支援
  - ユーザ教育
  - 新規ユーザ発掘
  - HPCの牽引