PostT2K構想

石川　裕
最先端共同HPC基盤施設　副施設長
東京大学情報基盤センター

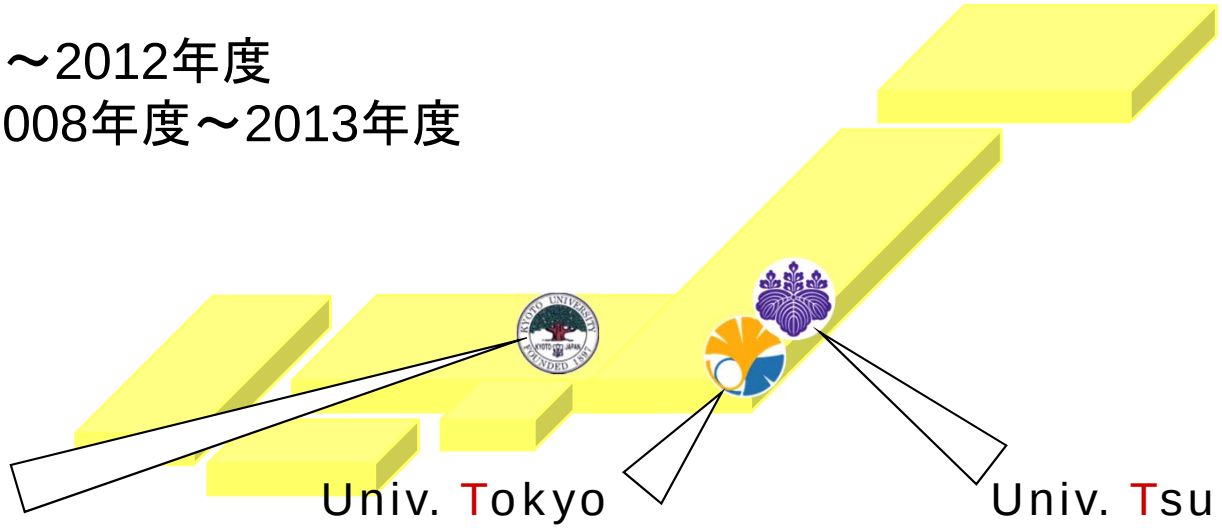
最先端共同HPC基盤施設発足記念シンポジウム

2013/7/24

14:40～15:10

T2Kオープンスパコン

京大：2008年度～2012年度
筑波大、東大：2008年度～2013年度



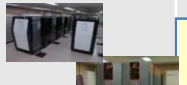
Kyoto Univ.
416 nodes (61.2TF) / 13TB
Linpack Result:
Rpeak = 61.2TF (416 nodes)
Rmax = 50.5TF

Univ. Tokyo
952 nodes (140.1TF) / 31TB
Linpack Result:
Rpeak = 113.1TF (512+256 nodes)
Rmax = 83.0TF

Univ. Tsukuba
648 nodes (95.4TF) / 20TB
Linpack Result:
Rpeak = 92.0TF (625 nodes)
Rmax = 76.5TF

筑波大＆東大＆国内7大学情報基盤センター

Post京？

Fiscal Year	2012	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Hokkaido	 Hitachi SR16000/M1 (172 TF, 22TB) Cloud System Hitachi BS2000 (44TF, 14TB)											
Tohoku	 NEC SX-9 + Exp5800 (31TF)											
Tsukuba	 HA-PACS (800 TF)				(Manycore system) (700+ TF)							
Tokyo	 T2K Todai (140 TF)						PostT2K (20+10? PF)			100+ PF		
	Fujitsu FX10 (1PFlops, 150TiB, 408 TB/s), Hitachi SR16000/M1 (54.9 TF, 10.9 TiB, 5.376 TB/s)								50+ PF			
Tokyo Tech.	 Tsubame 2.0 (2.4PF, 97TB, 744 TB/s)			Tsubame 2.5 (5.7 PF, 110+ TB, 1160 TB/s)								
Nagoya	 Fujitsu M9000(3.8TF, 1TB/s) HX600(25.6TF, 6.6TB/s) FX1(30.7TF, 30 TB/s)			Fujitsu FX10 (90.8TF, 31.8 TB/s), CX400(470.6TF, 55 TB/s) Upgrade (3.6PF)								
Kyoto	 Cray XE6 (300TF, 92.6TB/s), GreenBlade 8000 (243TF, 61.5 TB/s)			Cray XC30 (400TF)								
Osaka	 SX-8 + SX-9 (21.7 TF, 3.3 TB, 50.4 TB/s)											
Kyushu	 Hitachi SR1600(25TF)			Hitachi HA8000tc/HT210(500TF, 215 TiB, 98.82TB/s), Xeon Phi (212TF, 26.25 TiB, 67.2 TB/s), SR16000(8.2TF, 6 TiB, 4.4 TB/s)								
	Fujitsu FX10（270TF, 65.28 TB/s), CX400(510TF, 152.5 TiB, 151.14 TB/s), GPGPU(256TF, 30 TiB, 53.92 TB/s)											

2013/6/16版

設置場所

2013年7月22日　資料招請説明会
2015年4月以降　設置・運用開始

- 柏キャンパス 第2総合研究棟2階 スーパーコンピュータ室2

アウトライン

- T2Kでなにをやったか、設計思想
- PostT2Kではなにをやるか
  - 設計思想
  - 検討状況
  - 研究開発状況

オープンスパコン3原則

- 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

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
  - 実装密度、性能/電力
- 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
- 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

ポストT2K仕様と開発

- ハードウェア
  - CPU、ネットワークなどのハードウェア開発は行わない
  - 仕様を決める
    - ノード性能、メモリ容量、メモリバンド幅、ネットワーク性能、ストレージ容量、ストレージ性能
- ソフトウェア
  - 仕様を決める
    - アプリケーションプログラムインターフェイス：Linux
    - 通信ライブラリ：MPI-3
    - 数値計算ライブラリ群
    - 階層化ファイルシステム
    - バッチジョブシステム
- 仕様に基づいた開発
  - オペレーティングシステム
  - 低レベル通信ライブラリ＆MPI通信ライブラリ
  - 並列プログラミング言語
  - 数値計算ライブラリの一部

なぜ開発しなければならないか？

- メニーコア上での軽量OSが必要
  - キャッシュ容量
  - 共有資源管理
  - コア間通信
- エクサに向けて使えるプログラミング言語・ライブラリの必要性

オープンスパコン3原則 + TCO削減

- 基本アーキテクチャのオープン性
- システムソフトウェアのオープン性
- ユーザのニーズに対するオープン性

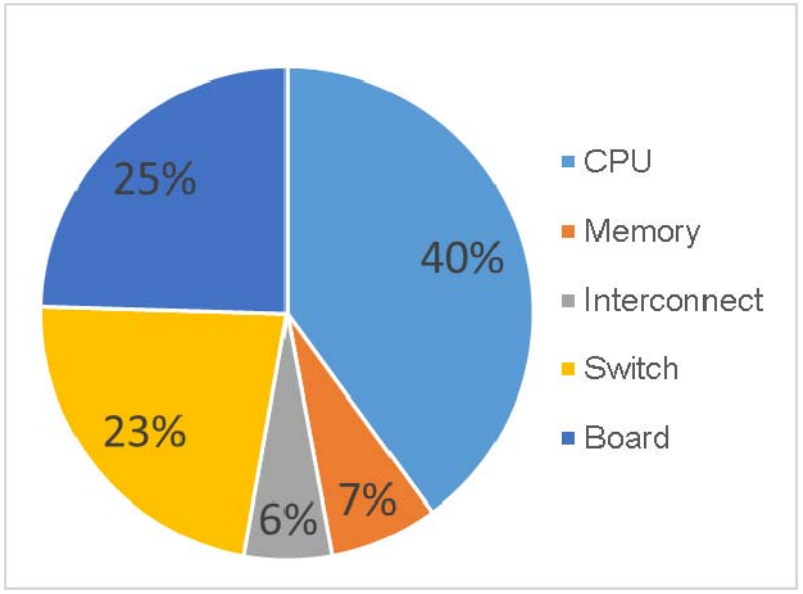
Total Cost of Ownership

- 電力削減
  - 計算機システム＋冷却装置の最大電力容量は4 MW
  - 電気代はおよそ1億円～/MW/年

ネットワークスイッチとボードレベルの電力で
全体の半分位の電力を必要としている

ネットワーク構成、
電力ロス、冷却
の工夫が必須

計算機システム電力量割合



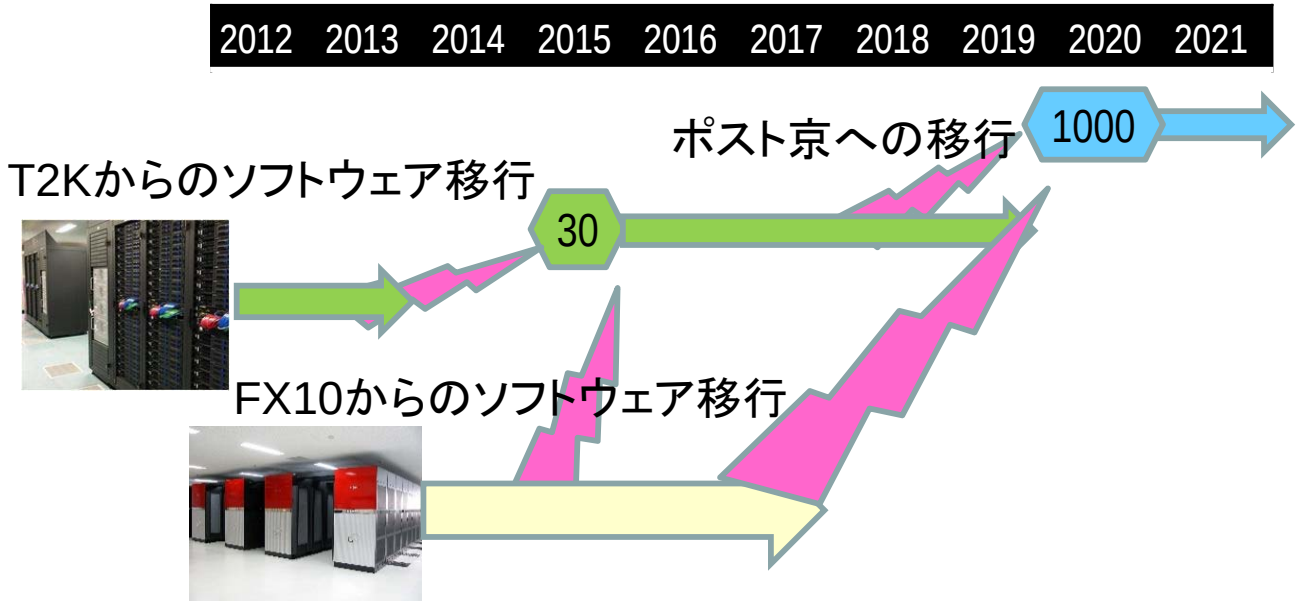
Board：FAN, AC/DC, DC/DC変換

オープンスパコン3原則 + TCO削減

- 基本アーキテクチャのオープン性
- システムソフトウェアのオープン性
- ユーザのニーズに対するオープン性

Total Cost of Ownership

- 電力削減
- ソフトウェア移行性



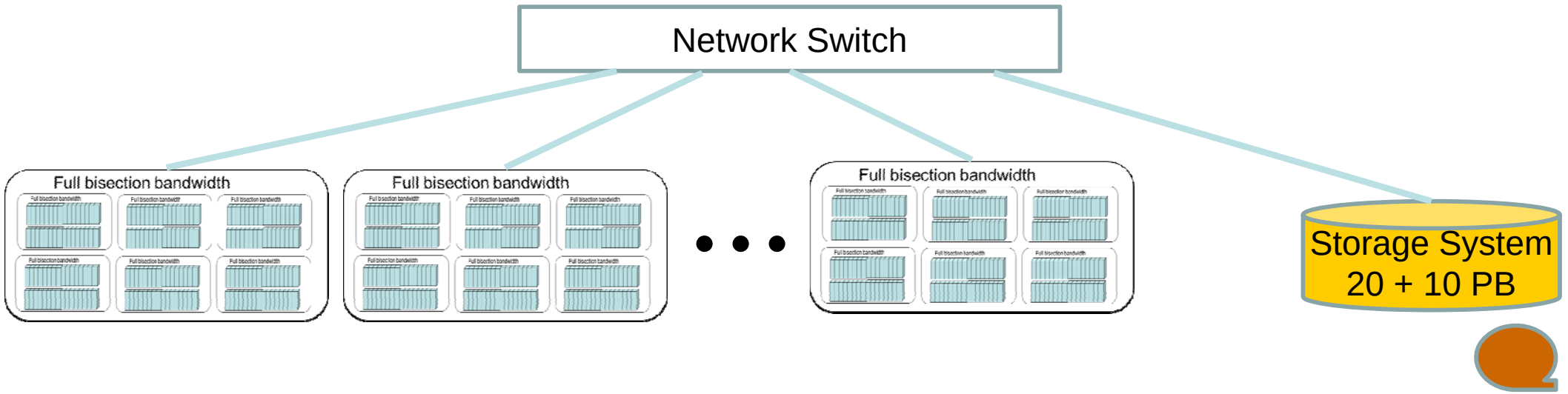
ノード単体性能、メモリバンド幅、ネットワークバンド幅

ノード単体性能、メモリバンド幅

- 総メモリ容量Oakleaf-FXとほぼ同じだが、総メモリバンド幅10倍以上
- 総メモリ容量Oakleaf-FXの４倍だが、総メモリバンド幅２倍

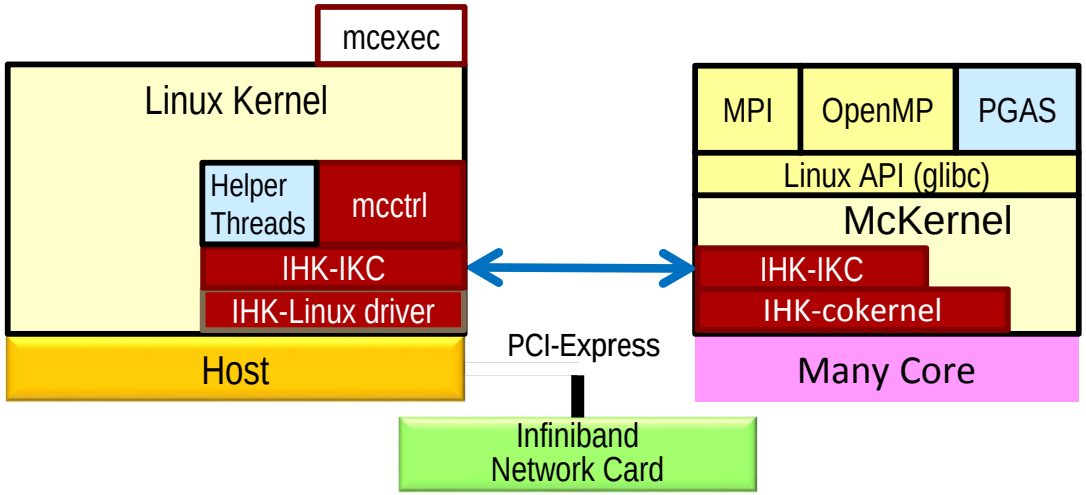
ネットワーク性能

- ネットワークバンド幅が小さくなる
  - 通信ソフト最適化
  - 通信と計算のオーバラップ
- ネットワークB/FをT2Kと同じになるCPU理論演算性能値
  - 357Gflops相当

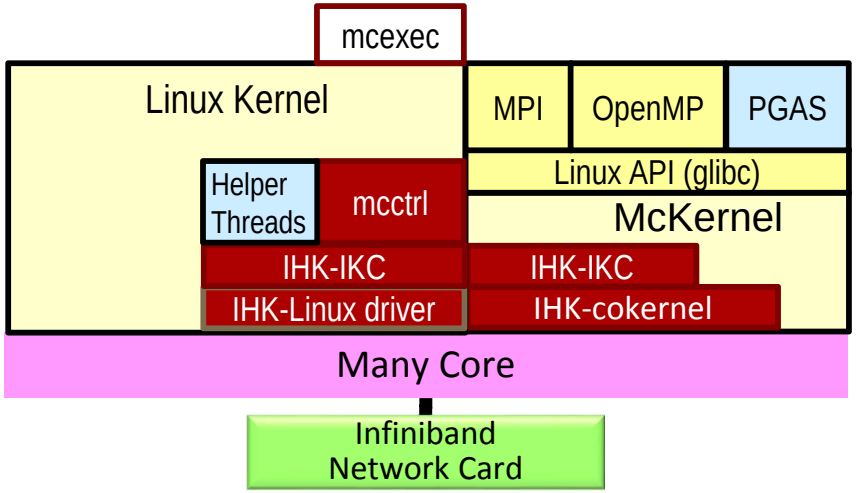


System Software Stack

In case of Non-Bootable Many Core



In case of Bootable Many Core



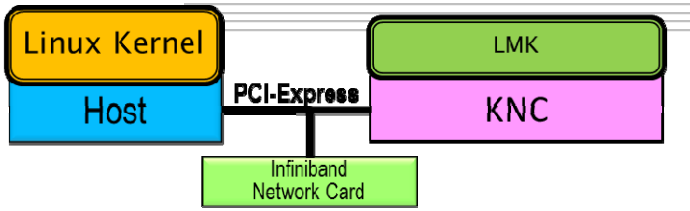
- IHK (Interface for Heterogeneous Kernel)
  - Provides interface between Linux kernel and micro kernels
  - Provides generic-purpose communication and data transfer mechanisms
- mcKernel
  - Micro lightweight kernel

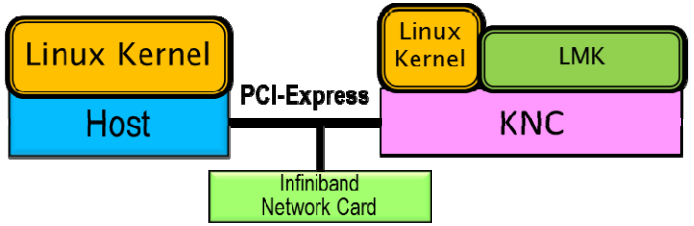
```
ihk/
|--- linux/
|    |--- driver/
|    |    |--- attached/
|    |    |    |---mic/
|    |    |--- builtin/
|    |--- core/
|    |--- include/
|    |--- user/
|
|--- ikc/
|--- include/
|--- cokernel/
|     |--- attached/
|     |    |--- mic/
|     |--- builtin/
|          |--- mic/
|          |--- sparc/
|          |--- x86/
|--- doxygen/
|--- test/
```

```
mckernel/
|--- arch/
|    |--- x86/
|    |    |--- elfboot/
|    |    |--- kboot/
|    |    |--- kernel/
|    |--- sparc/
|--- kernel/
|    |---- config/
|    |---- include/
|    |---- script/
|--- include/
|--- executer/
|    |--- user/
|    |--- kernel/
|    |--- include/
|--- lib/
|--- doxygen/
|--- test/
```

Prototype System on Xeon Phi

- Features implemented and being tested
  - glibc and pthread
    - Thread and memory management
    - File I/O, delegated to Linux in host
    - Memory map and dynamic link library
  - Process launcher in host
  - Direct Communication with Infiniband
  - MPI library (not fully) running on Xeon Phi
  - OpenMP environment with Intel compiler
- Features being developed and planned
  - Hierarchical Memory Management
  - PVAS, supporting the PGAS model
  - Direct SSD
  - Single OS kernel image for partitioned multiple light-weight kernels





Online Supercell Simulation

- Supercell is a rotating thunderstorm, which causes significant damages by heavy rainfall and strong winds, and sometimes involves tornados.
- The simulation code, SCALE, is developed at RIKEN AICS, Japan
- The simulation is an ideal experiment based on the standardized test case prepared in the Weather Research and Forecasting (WRF) Model.
  - mesh size: 200 m x 1500 m x 1500 m (z, x, y)
  - domain size: 18 km x 150 km x 150 km (z, x, y)

Target: 20 second real-time visualization for 2 hour phenomena in the real-world using 4 node Xeon-phi cluster with Our OS/Runtime

おわりに

- 省電力化の検討
- アプリケーション性能予測
- ハードウェア・ソフトウェア仕様策定
- システムソフトウェア開発