

エクサスケールコンピューティング 実現に向けて

石川 裕

東京大学情報基盤センター

東京大学情報理工学系研究科

理化学研究所計算科学研究機構

本発表の概要

- 国内におけるポストペタスケールコンピューティングの状況
- ポストT2Kに関する私案
 - Manycore-based PC Clusterの概要
 - System Software Stackの概要
 - プロトタイプシステムの状況

国内ポストペタスケールコンピューティングに向けての取り組み

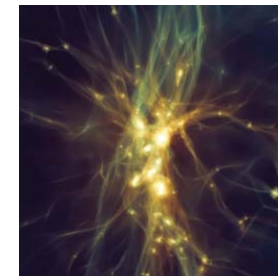
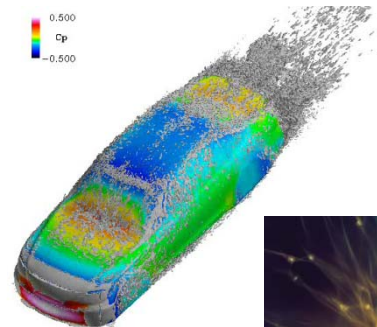
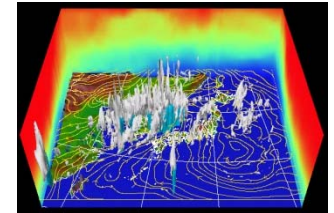
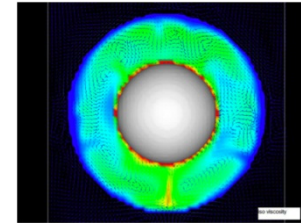
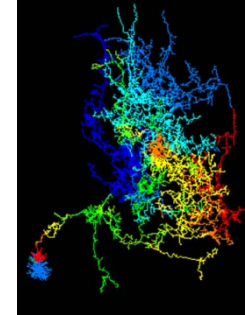
- HPCI計画の推進にあたり国として必要な事項等を検討するため、研究振興局長の諮問会議「HPCI計画推進委員会」のもとに「今後のHPC技術の研究開発のあり方を検討するWG」が設置
 - 今後の開発を担う若手を中心に、幅広い産学官の関係者を集めた作業部会による検討
 - 全体取りまとめ: 石川裕(東大)
 - アプリケーション作業部会
 - 取りまとめ: 富田(理研AICS)、牧野(東工大)
 - コンピュータアーキテクチャ・コンパイラ・システムソフトウェア作業部会
 - 分野取りまとめ: 丸山(東工大)、近藤(電通大)、石井(NEC/東大)、野村(東工大)、滝澤(東北大)、須田(東大)、片桐(東大)
 - 2月10日、HPCI計画推進委員会に作業部会報告を行う

作業部会での議論の経緯

- アプリケーション作業部会
 - 2020年までに社会的・科学的課題を如何に解決できるかの視点に立ったサイエンスロードマップをまとめるとともに計算機システムに対する要求事項を検討
 - 生命科学、物質科学、地球科学、ものづくり、宇宙・素粒子・原子核
- コンピュータアーキテクチャ・コンパイラ・システムソフトウェア作業部会
 - サイエンスロードマップを達成するために、サイエンス達成年の2年前である2018年までに消費電力20～30MW、設置面積2000平米に設置できる計算機システムに必要となる技術開発を検討
 - アーキテクチャ、コンパイラ、OS、ミドルウェア、プログラミングモデル・言語、数値計算ライブラリ
- 合同作業部会：3回開催（参加延べ人数：368人）
 - <http://www.open-suercomputer.org/workshop>
- Co-designの重要性

合同作業部会報告概要:サイエンス分野概要

- 生命科学
 - 生命体を構築する要素である生体分子、細胞、臓器、脳神経に対し、様々なスケールにて計算機シミュレーションを行うとともに、膨大な生命データを解析し、予測可能な生命科学と医療および創薬に貢献できる理論モデルを構築
- 物質科学
 - 物性科学、分子科学、材料科学の3つに学問分野において、さまざまな物質や材料をミクロな視点に立って研究し、基礎研究と応用研究とをつなぐと同時に、現代社会の産業基盤の形成に貢献
- 地球科学
 - 地震・津波分野では、力学系としての固体地球の包括的理解にもとづき、M9級以上の大規模低頻度現象を含む被害予測システムを構築。気象・気候分野では、局所的集中豪雨の予測を目指す気象予報の実用化を目指すとともに、地球温暖化などの地球環境変化の予測の再解析システムを構築
- ものづくり(連続系)
 - 「数値シミュレーションによる設計」の実現。大規模数値シミュレーションを用いた高精度解析による試験の代替、最適設計、設計の準自動化による開発の効率化、製品の革新的高性能化
- 宇宙・原子核・素粒子
 - 宇宙の探求、複数の階層にわたる現象の解明、極限状況下での物質形態の解明



合同作業部会報告概要:サイエンス課題達成に必要なシステム

- 現時点の計算機構成を4つ検討
 - 汎用(従来型)、容量・帯域重視、メモリ容量削減、演算重視
- 現時点で想定される今後の世界的な技術開発が進んだ場合に、各構成で2018年に実現すると推測されるシステム性能諸元を予想
 - 消費電力20~30MW、設置面積2000平米
- サイエンスロードマップに基づいて2018年ごろのアプリケーションに必要な性能要求と技術トレンドに基づいたシステム性能とのギャップを示す
 - サイエンスロードマップについては、「今後のHPC技術開発に関する報告書」参照

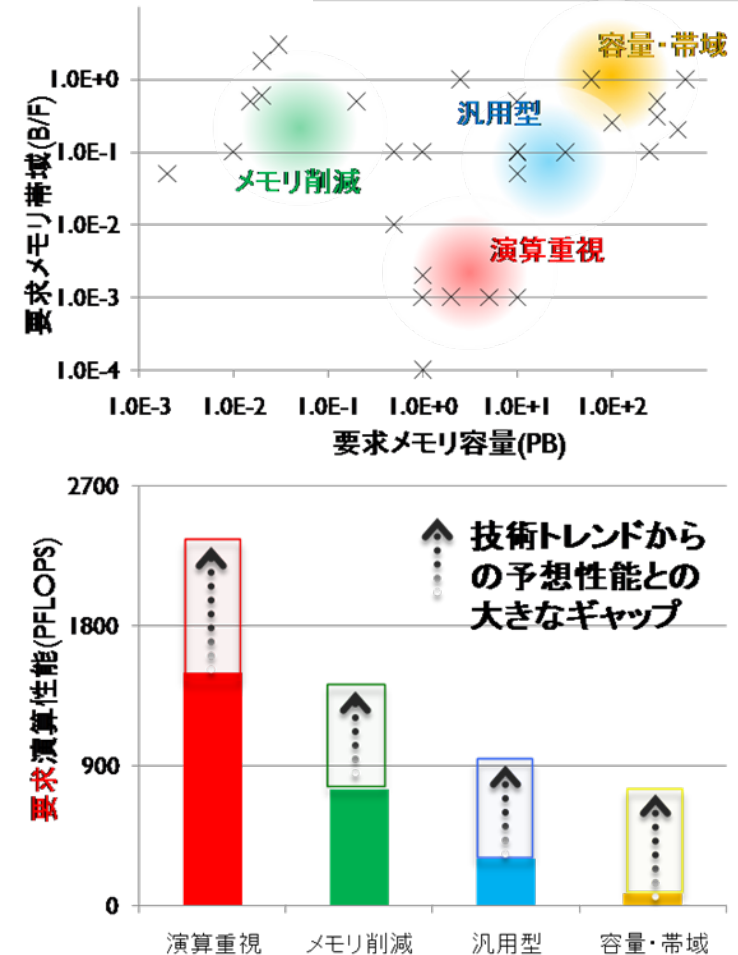


図1. 要求性能と予測性能の相関
上: 各アプリの要求するB/F値とメモリ容量
下: 各分類に対する要求性能値と予想性能

出展: 今後のHPCI技術開発に関する報告書

合同作業部会報告概要：研究開発計画案

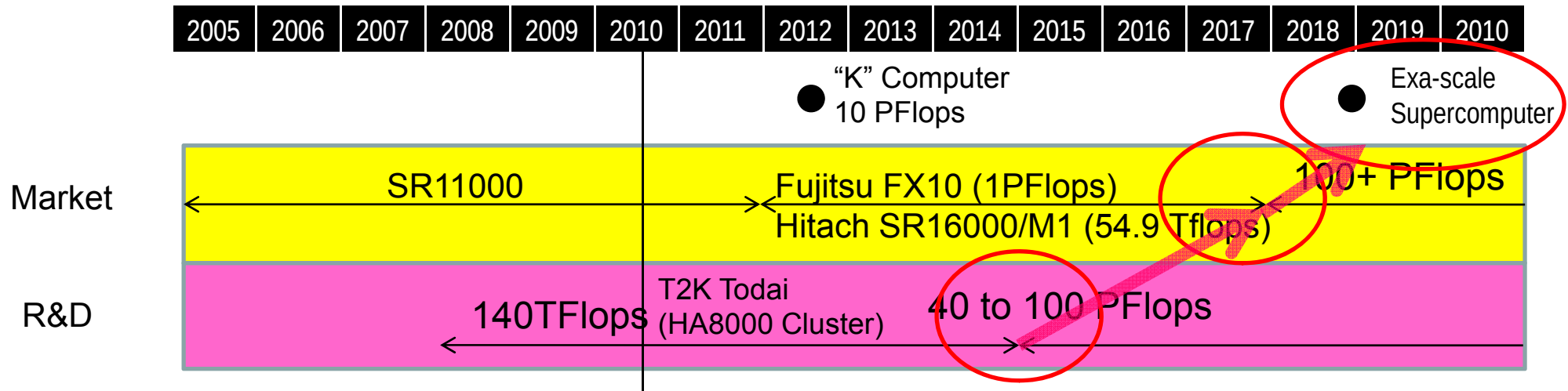
・ 報告書における研究開発案

	2012	2013	2014	2015	2016	2017	2018	2019	2020
JST CREST: ポストペタスケール高性能計 算に資するシステムソフトウェ ア技術の創出	1期		2期		3期				
ライブラリ	要素技術方式検討	メモリ・通信 I/F, FW	非均質 精度・低電力・耐故障						
プログラミングモデル・ 言語、フレームワーク	要素技術方式検討	要素技術開発	実証実験・改良						
	統合環境方式検討	試作・評価	要素技術統合	実証実験・改良	機能整備				
OS、ミドルウェア	要素技術方式検討	要素技術開発・評価・改良							
	API国際標準化への検討・働きかけ		技術統合・移植・最適化						
アーキテクチャ・コンパ イラ	技術調査	要素技術開発	開発						
	方式検討	評価	方式改良						

合同作業部会報告概要：課題（一部）

- アーキテクチャ
 - 電力効率に優れたスループットコア導入（ヘテロ化）
 - 計算モデルに適した記憶階層と高機能メモリ
 - 三次元積層DRAM、NVRAM等の活用
 - 超大規模並列化への対処
 - 通信遅延削減、トポロジ検討、集合通信支援
 - 低電力化と電力あたり性能の向上
 - 新デバイス活用、スループットコア利用技術
 - システムレベル電力制御技術の確立
 - 耐故障へのアーキテクチャ支援
- システムソフトウェア
 - $O(100K) \sim O(1M)$ 計算ノード管理
 - コモディティソフトウェアに対する互換性確保
 - 演算コア向け軽量OS（ヘテロ構成時）
 - 電力・メモリ階層ウェアOS・ランタイム
 - 低遅延非同期通信ライブラリ
 - 超規模並列ファイルシステム
 - タスクスケジューリング・データ配置
 - 耐故障機構
 - ユーザAPIの国際標準化推進
- プログラミングモデル・言語・フレームワーク
 - ヘテロジニアスアーキテクチャの透過性
 - 大規模並列性の透過性
 - 複雑化するメモリアーキテクチャの透過性
 - 電力制御の透過性
 - 耐障害機能の透過性
 - 生産性：デバッグおよび性能チューニング支援
 - 各要素技術を統一的に利用するための環境の構築
- 数値計算ライブラリ
 - ライブラリインターフェース：移行コストを軽減、アーキ依存性を吸収
 - 非均質プロセッサ対応：高演算効率のアクセラレータを最大限利用
 - 通信量およびメモリアクセス回数の削減：データ移動オーバーヘッドの最小化
 - 演算精度：倍精度では不足の事態も
 - 低電力化：電力を制御可能なライブラリ
 - 耐故障性：ライブラリでFTを補助
 - フレームワーク：主要領域のアプリ開発共通基盤により、開発コストを低減

Post T2K東大



本郷キャンパス

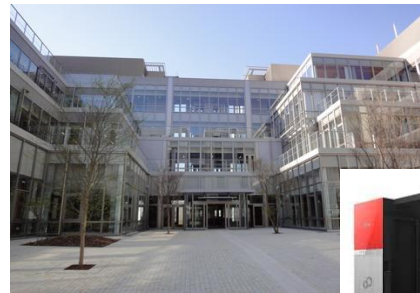


SR16000/M1



HA8000

柏キャンパス

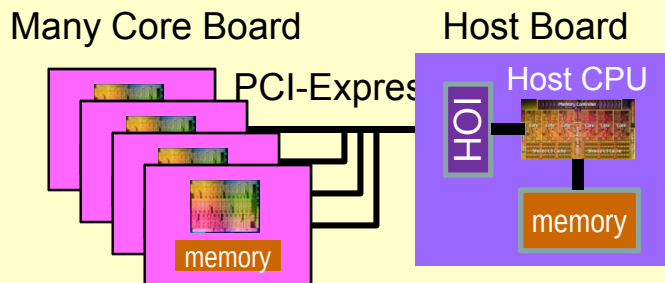


FX10

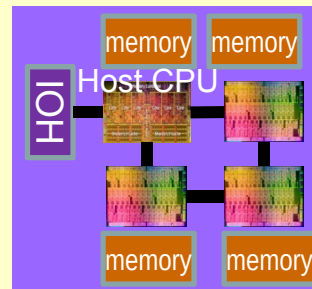
- PRIMEHPC FX10
 - 4800 ノード (16コア/ノード)
 - 1.13 PFlops
 - 150 テラバイト主記憶
- Hitachi SR16000/M1
 - 56ノード (32コア/ノード)
 - 54.9 TFlops
 - 11200 ギガバイト主記憶

Variations of Many-core based machines

Many-core board connected to PCI-Express
e.g., Intel Knights Ferry, Knights Corner



Many-core chip connected to system bus
Not existing so far

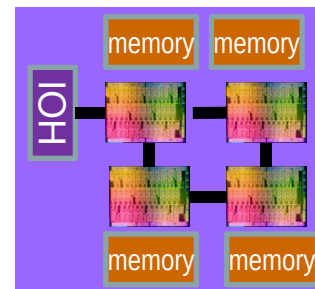


Many-core inside CPU die
c.f., Intel Sandy Bridge with GPU



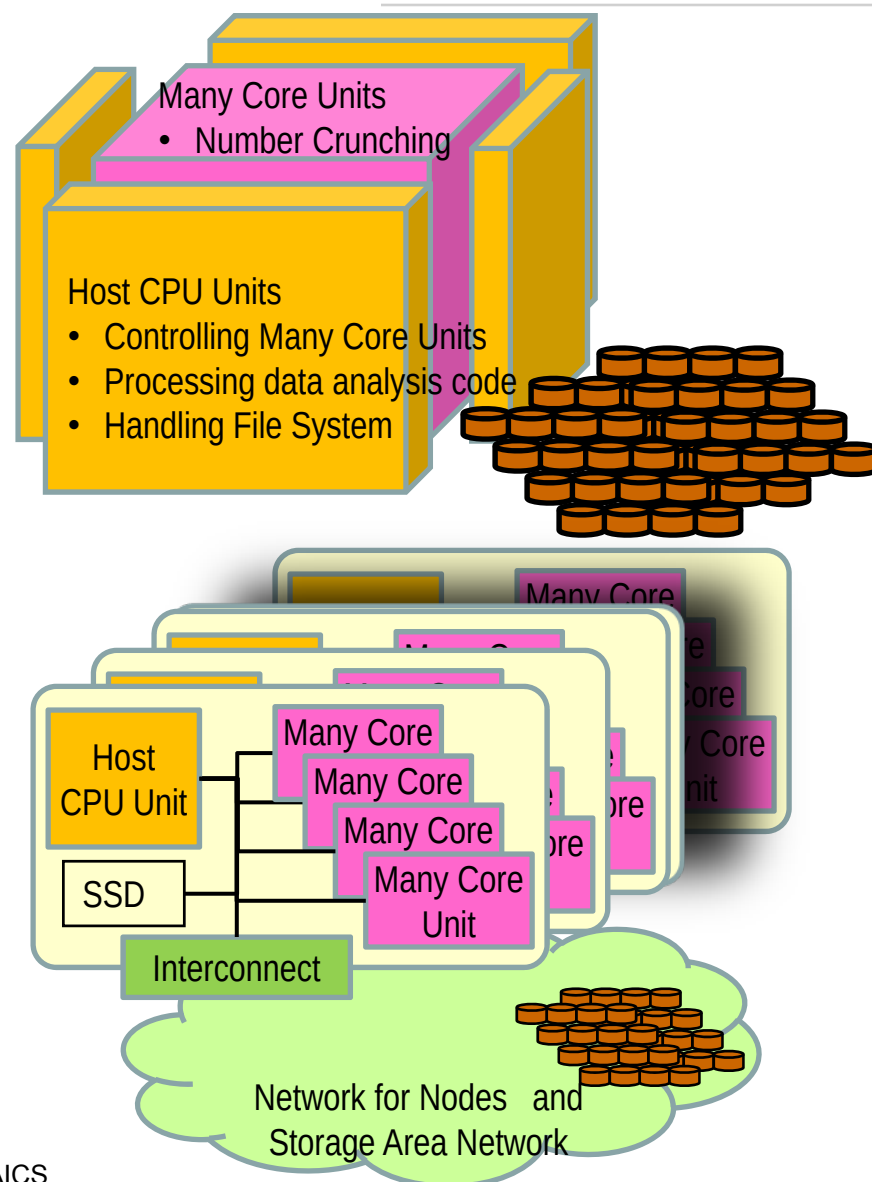
http://pc.watch.impress.co.jp/docs/column/kaigai/20100412_360173.html

Many-core only
Not existing so far



ポストT2Kシステムイメージ: 要求

- 大規模データ解析ならびに数値計算アプリケーションの要求が満たされる必要がある
 - I/O性能
 - 浮動小数点演算性能
 - 並列性能



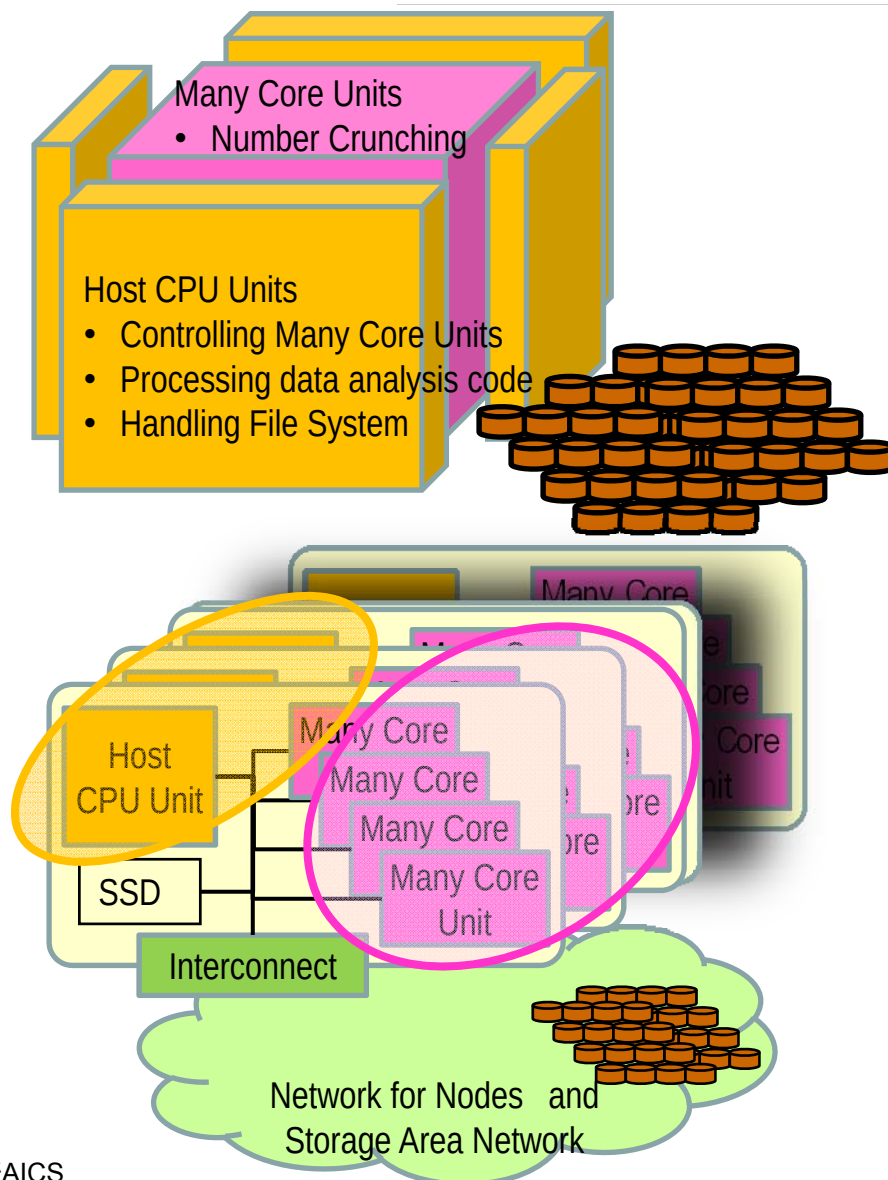
ポストT2Kシステムイメージ: ジョブ実行イメージ

- 大規模データ解析ならびに数値計算アプリケーションの要求が満たされる必要がある
 - I/O性能
 - 浮動小数点演算性能
 - 並列性能

パーティション内2タイプジョブの同時実行

ManyCore群: 演算重視アプリケーション実行
ファイルI/O時等Host CPU使用

Host CPU群: I/O重視アプリケーション実行



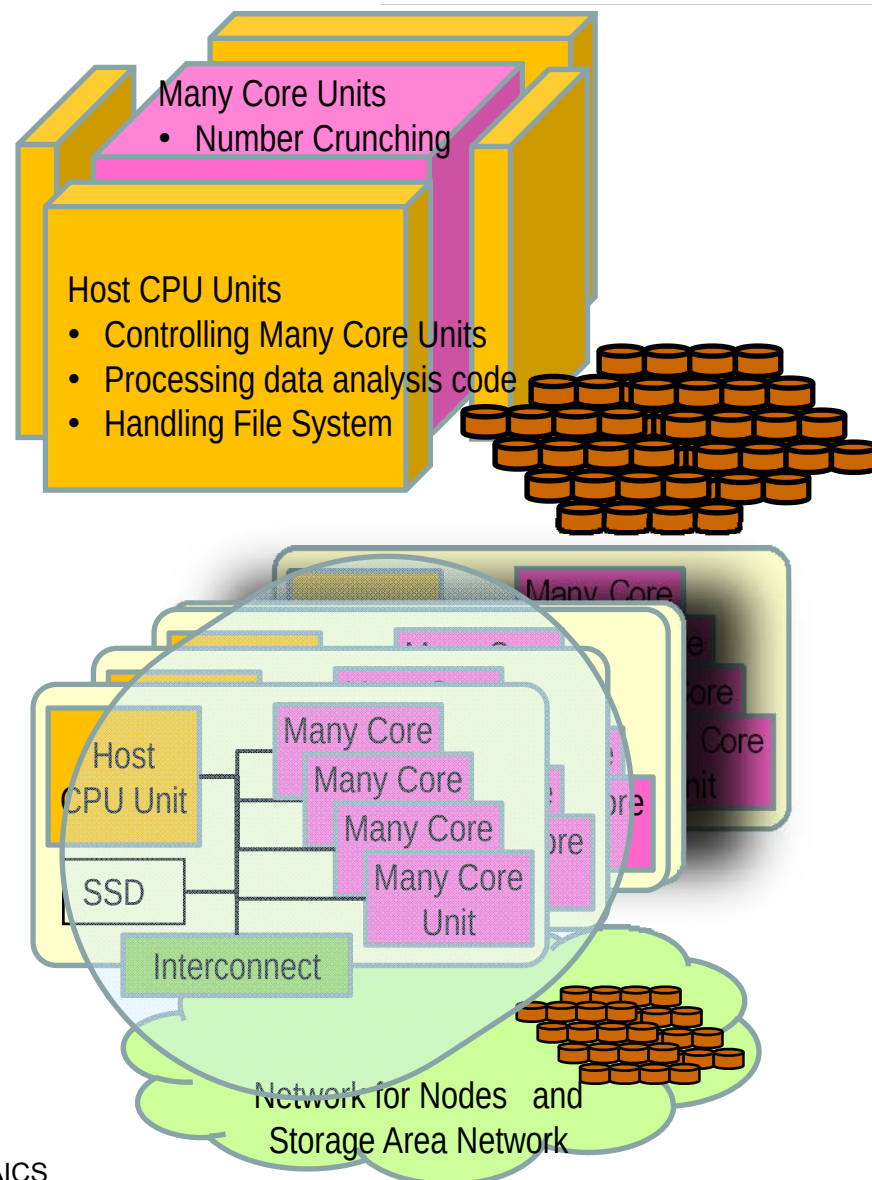
ポストT2Kシステムイメージ: ジョブ実行イメージ

- 大規模データ解析ならびに数値計算アプリケーションの要求が満たされる必要がある
 - I/O性能
 - 浮動小数点演算性能
 - 並列性能

パーティション内1ジョブ実行

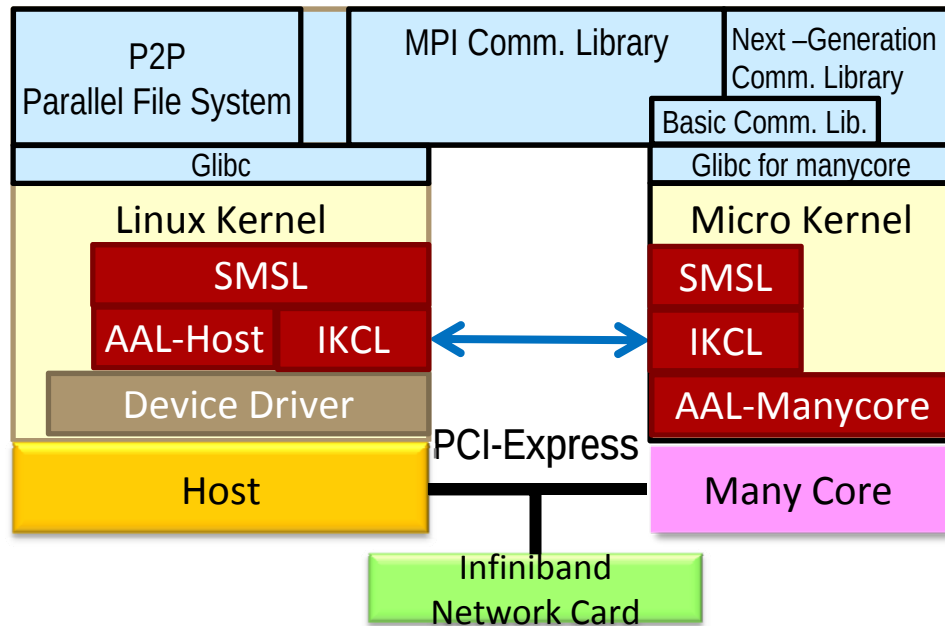
ManyCore群: 演算 & 通信

Host CPU群: メモリスワップ & 通信 & I/O



ポストT2Kソフトウェアアーキテクチャ

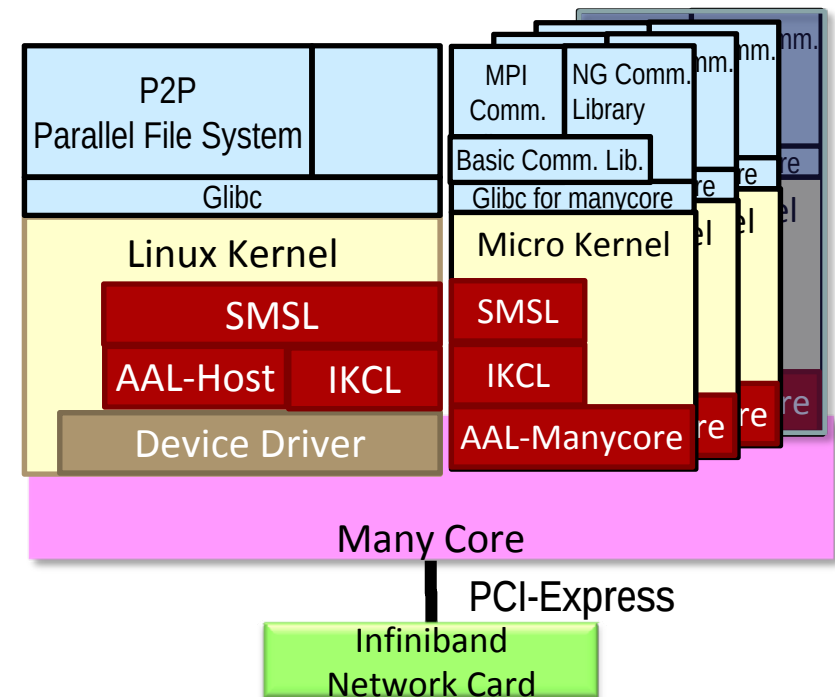
In case of Non-Bootable Manycore



- AAL (Accelerator Abstraction Layer)
 - Provides low-level accelerator interface
 - Enhances portability of the micro kernel
- IKCL (Inter-Kernel Communication Layer)
 - Provides generic-purpose communication and data transfer mechanisms
- SMSL (System Service Layer)
 - Provides basic system services on top of the communication layer

石川裕@東大／理研AICS

In case of Bootable Manycore

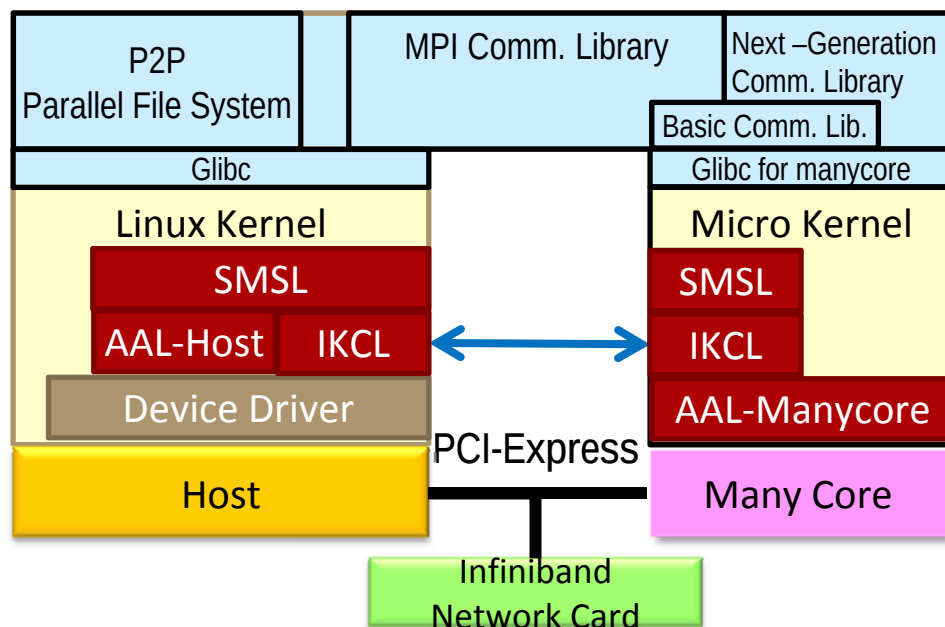


Design Criteria

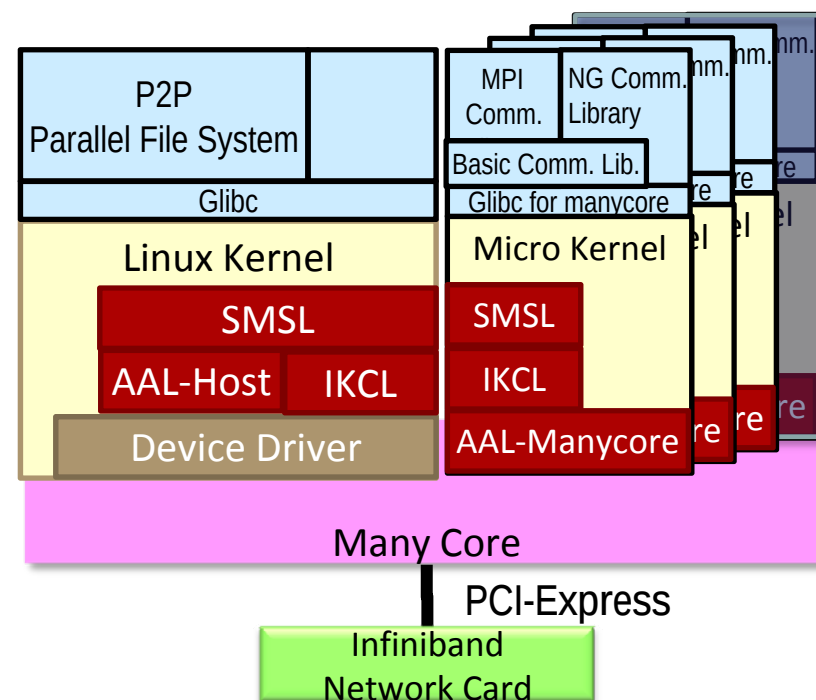
- Cache-aware system software stack
- Scalability
- Minimum overhead of communication facility
- Portability

ポストT2Kソフトウェアアーキテクチャ

In case of Non-Bootable Manycore



In case of Bootable Manycore



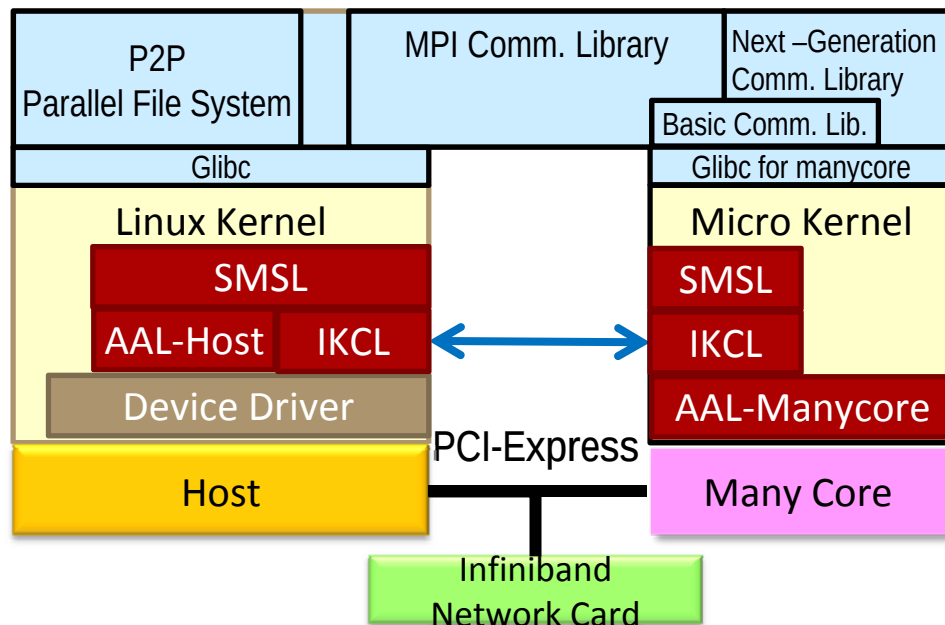
- AAL (Accelerator Abstraction Layer)
 - メニコアアーキテクチャはキャッシュが小さくメモリバンド幅も小さいため、ユーザアプリケーションとシステムプログラムのキャッシュフットプリントを最小にすべき
- IKCL (Inter-Kernel Communication Library)
 -
- SMSL (System Service Layer)
 - Provides basic system services on top of the communication layer

Design Criteria

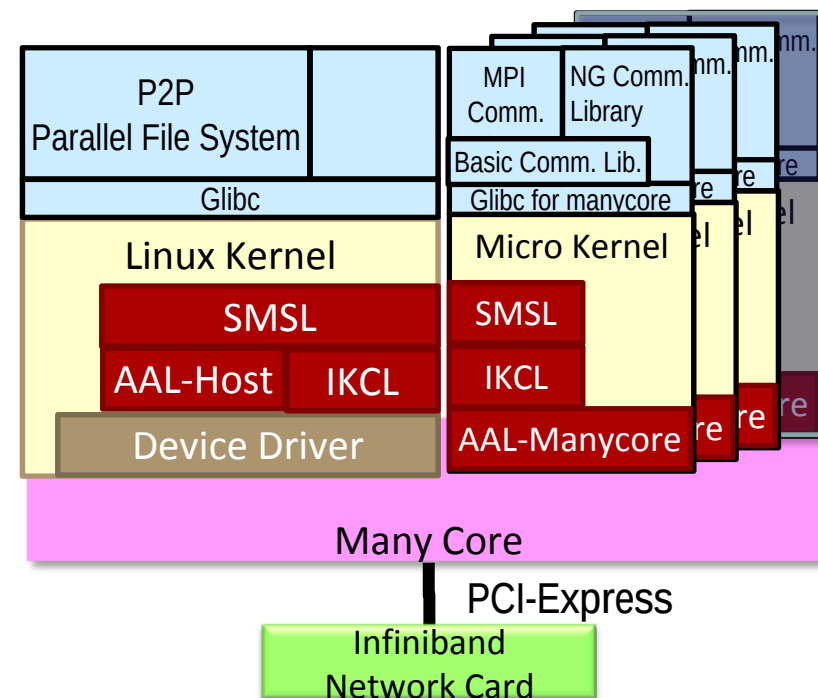
- Cache-aware system software stack
- Scalability
- Minimum overhead of communication facility
- Portability

ポストT2Kソフトウェアアーキテクチャ

In case of Non-Bootable Manycore



In case of Bootable Manycore



- AAL (Accelerator Abstraction Layer)
 - Provides low-level accelerator interface

エクサスケールマシンでは、OSは、 $O(100K)$ ~ $O(1M)$ ノードを管理しなければならない。これら資源管理に要するメモリおよび制御コストを低減する必要がある

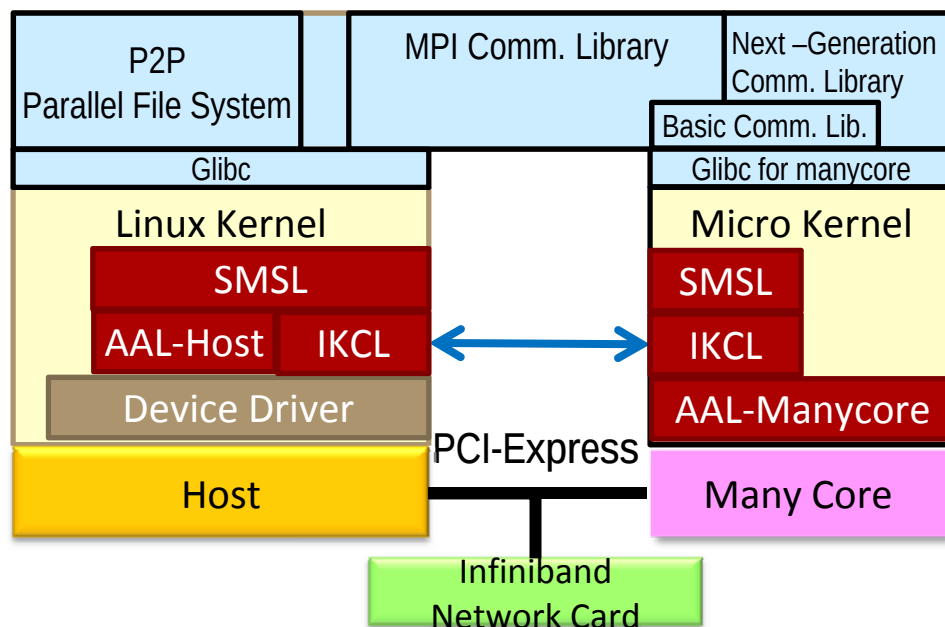
- SMLS (System Service Layer)
 - Provides basic system services on top of the communication layer

Design Criteria

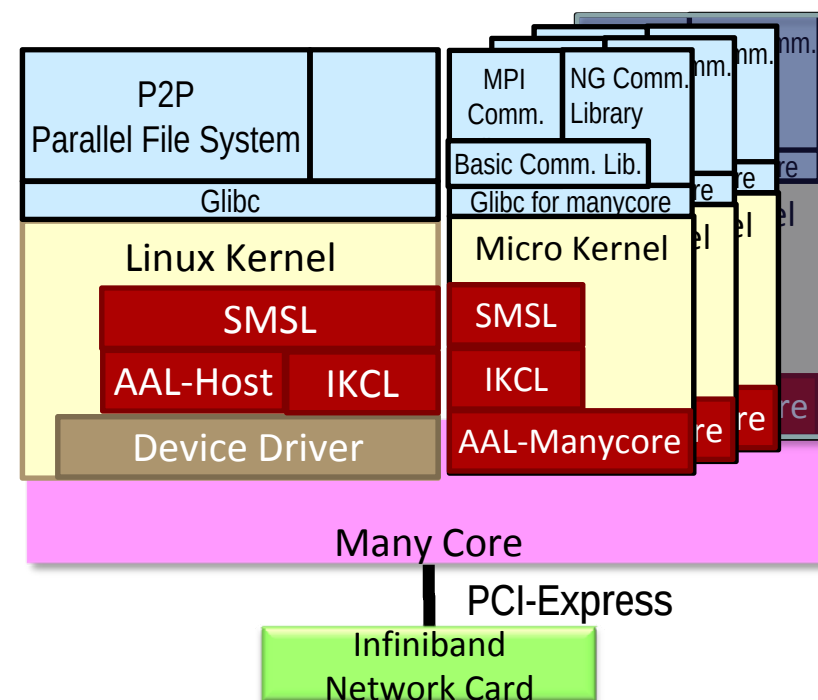
- Cache-aware system software stack
- Scalability
- Minimum overhead of communication facility
- Portability

ポストT2Kソフトウェアアーキテクチャ

In case of Non-Bootable Manycore



In case of Bootable Manycore



- AAL (Accelerator Abstraction Layer)
 - Provides low-level accelerator interface
 - Enhances portability of the micro kernel

- IKCL (Inter-Kernel Communication Layer)
 - 低コストメニコア間通信が必要。本件は、この発表で現在の取り組みを紹介。

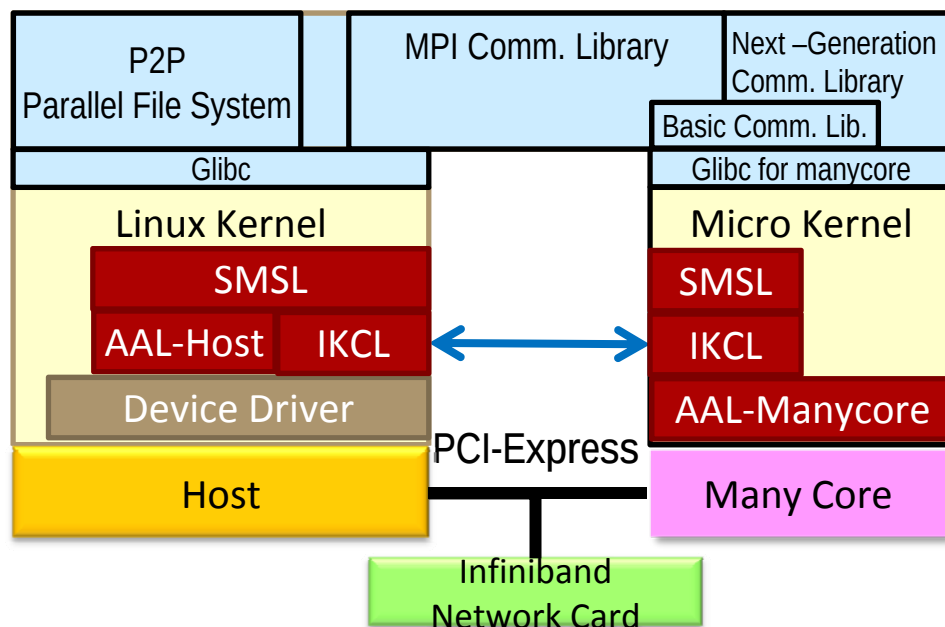
- SMLS (System Service Layer)
 - Provides basic system services on top of the communication layer

Design Criteria

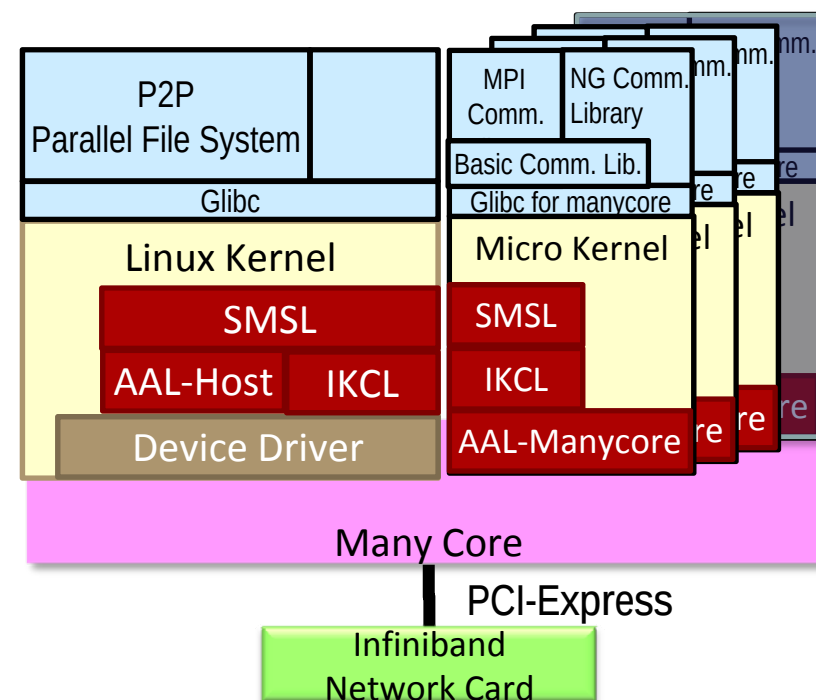
- Cache-aware system software stack
- Scalability
- Minimum overhead of communication facility
- Portability

ポストT2Kソフトウェアアーキテクチャ

In case of Non-Bootable Manycore



In case of Bootable Manycore



- AAL (Accelerator Abstraction Layer)
 - Provides low-level accelerator interface
 - Enhances portability of the micro kernel
- IKCL (Inter-Kernel Communication Layer)
 - Provides communication and data
- SMLS
 - Provides basic system services on top of the communication layer

現在のPCクラスタ上のアプリケーションが最低限の修正で動作することが重要

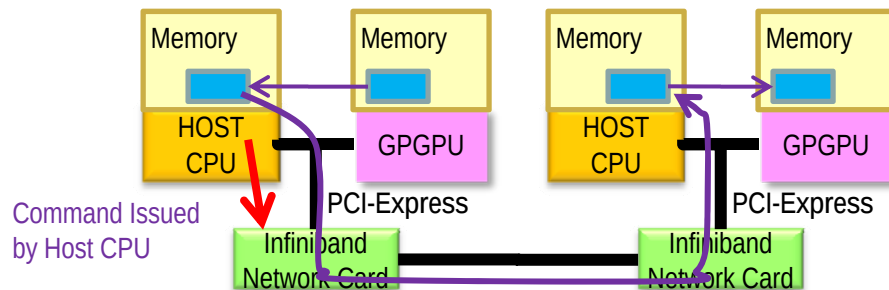
Design Criteria

- Cache-aware system software stack
- Scalability
- Minimum overhead of communication facility
- Portability

DCFA: Direct Communication Facility for Accelerator

Acceleratorにおける通信に関する問題

- AcceleratorはPCI-Expressの1デバイスであるため、Acceleratorが通信デバイスを直接Configureできない
 - Acceleratorは単独で他のデバイスのアドレスを知ることはできない
- Acceleratorが通信デバイスのアドレスを知ったとしても、GPUの場合通信デバイスを制御するコマンドが発行できない
- Mellanox社GPU DirectはホストCPU介在



Many Core型Acceleratorの場合

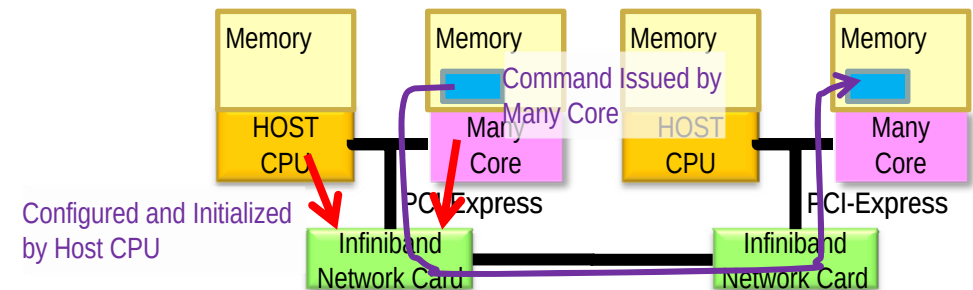
- Many Coreが通信デバイスのアドレスを知ることができれば、通信デバイスを制御するコマンドを発行できる
- ただし、通信デバイスから割り込みを受信することはできない



提案手法

DCFA

(Direct Communication Facility for Accelerator)



Min Si and Yutaka Ishikawa, "Design of Direct Communication Facility for Manycore-based Accelerators," to appear at CASS2012 in conjunction with IPDPS2012.

おわりに

- 「今後のHPCI技術開発に関する報告書」の概要説明
 - 2012年度から2年間Feasibility Study
 - エクサスケールに向けたFeasibility Studyでは、社会ニーズに合致した実現性のあるマシン設計が望まれる
 - 継続的マシン開発・設置・運用が重要
- ポストT2K開発構想
 - コモディティ部品によるPCクラスタとの親和性
 - 理研計算科学研究機構および企業との連携

