次世代ヘテロジニアスPCクラスタのための システムソフトウェア

石川　裕 (東京大学)
システムソフトウェア技術部会

システムソフトウェア技術部会

- 次世代PCクラスタ技術を想定した研究、開発体制を産学連携スキームで推進

学側の母体（H23.12.9現在）

理化学研究所計算科学研究機構

科学技術振興機構(JST) CREST
「メニイコア混在型並列計算機用基盤ソフトウェア」(代表：堀敦史)
近畿大学　辻田祐一、東京農工大学　並木美太郎

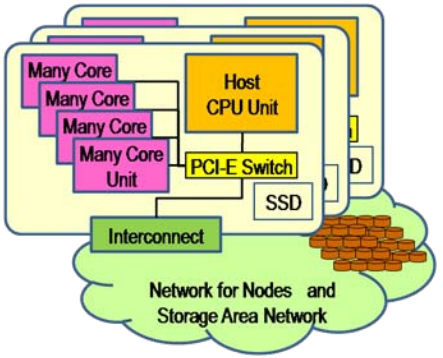
東京大学情報基盤センター

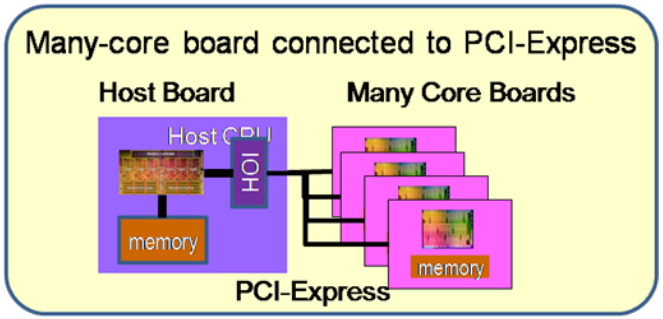
科学技術振興機構(JST) CREST
「自動チューニング機構を有するアプリケーション開発・実行環境 ppOpen-HPC」(代表：中島健吾)
東京大学　佐藤正樹、東京大学　奥田洋司、東京大学　古村孝志、
京都大学　岩下武史、海洋研究開発機構　阪口秀

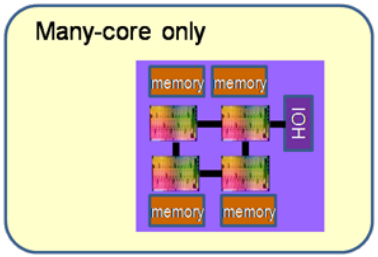
東京大学情報基盤センター＆情報理工学系研究科

石川裕研究室：ManyCore Architecture向けマイクロカーネル基本部＆通信機構

Manycore-based PC Cluster

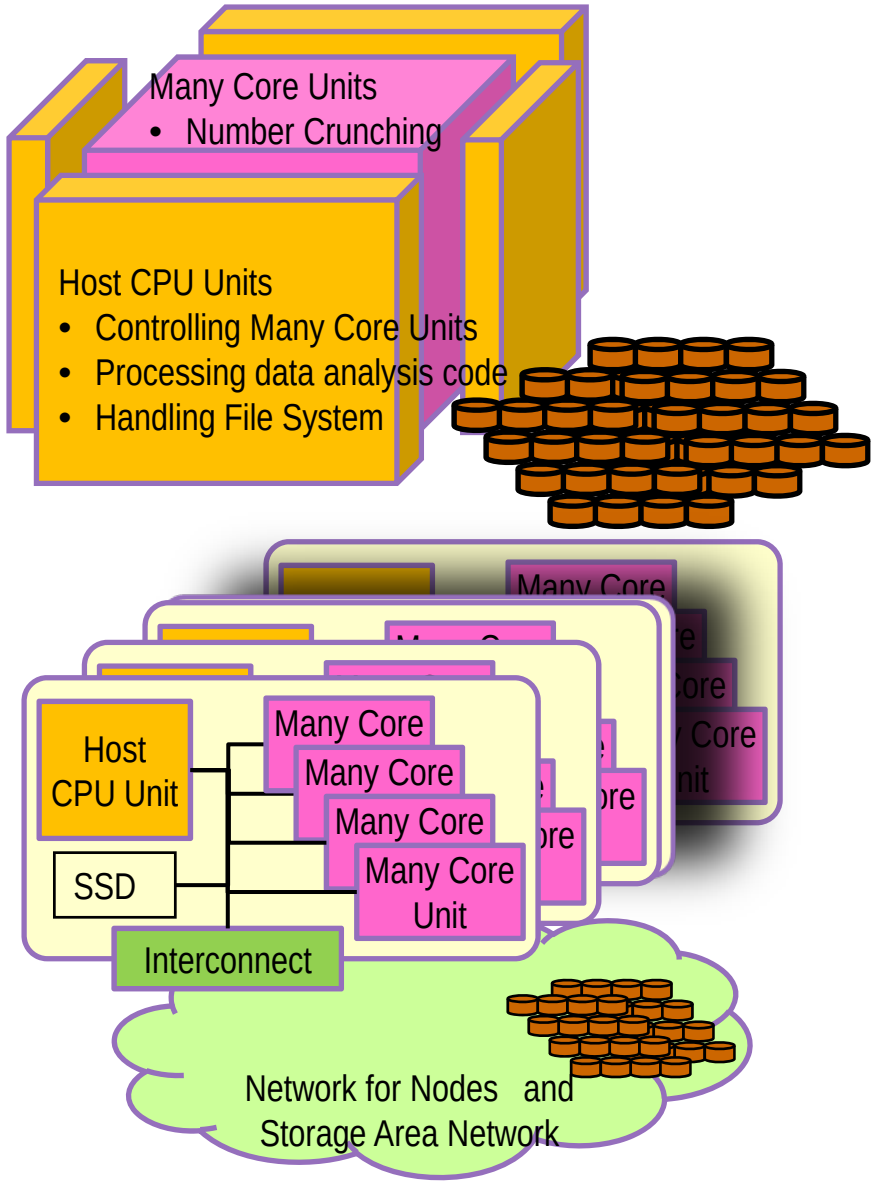






システムイメージ:要求

- 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



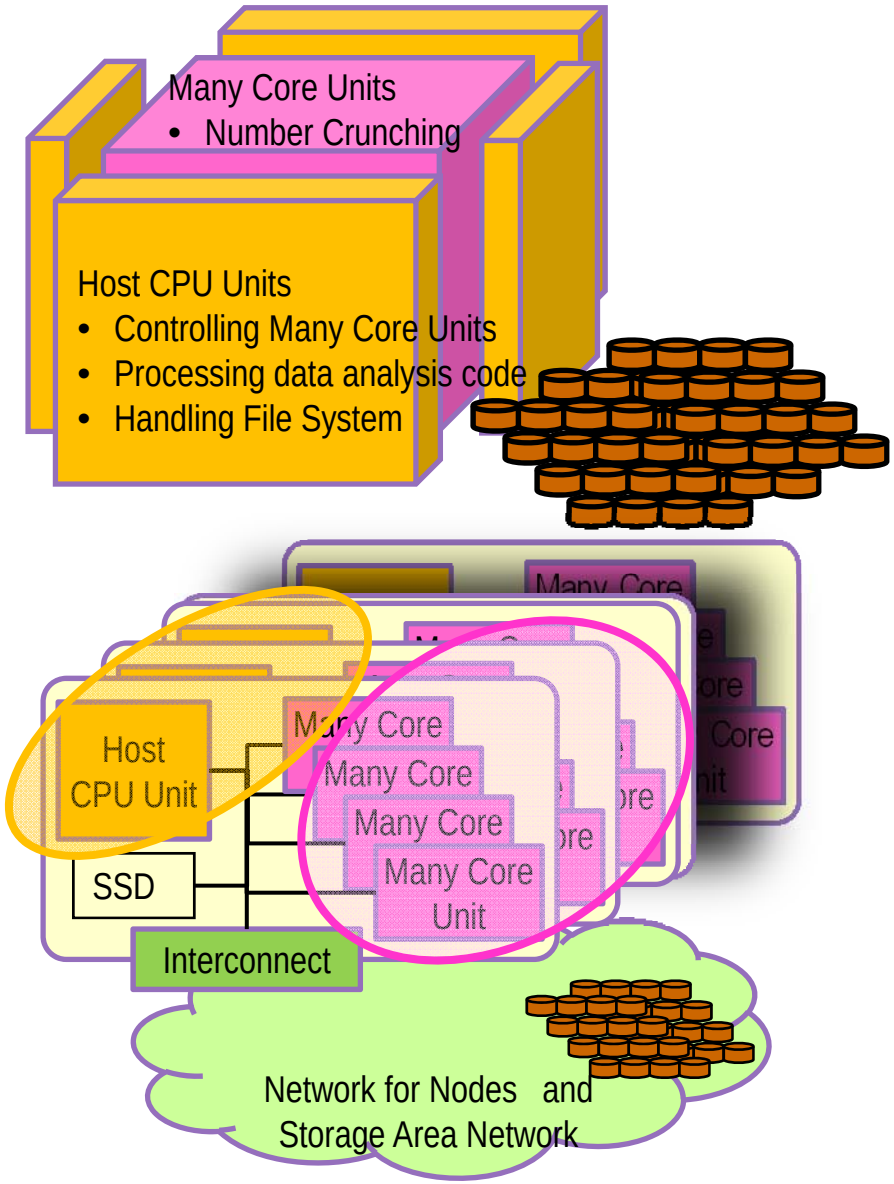
システムイメージ:ジョブ実行イメージ

- 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

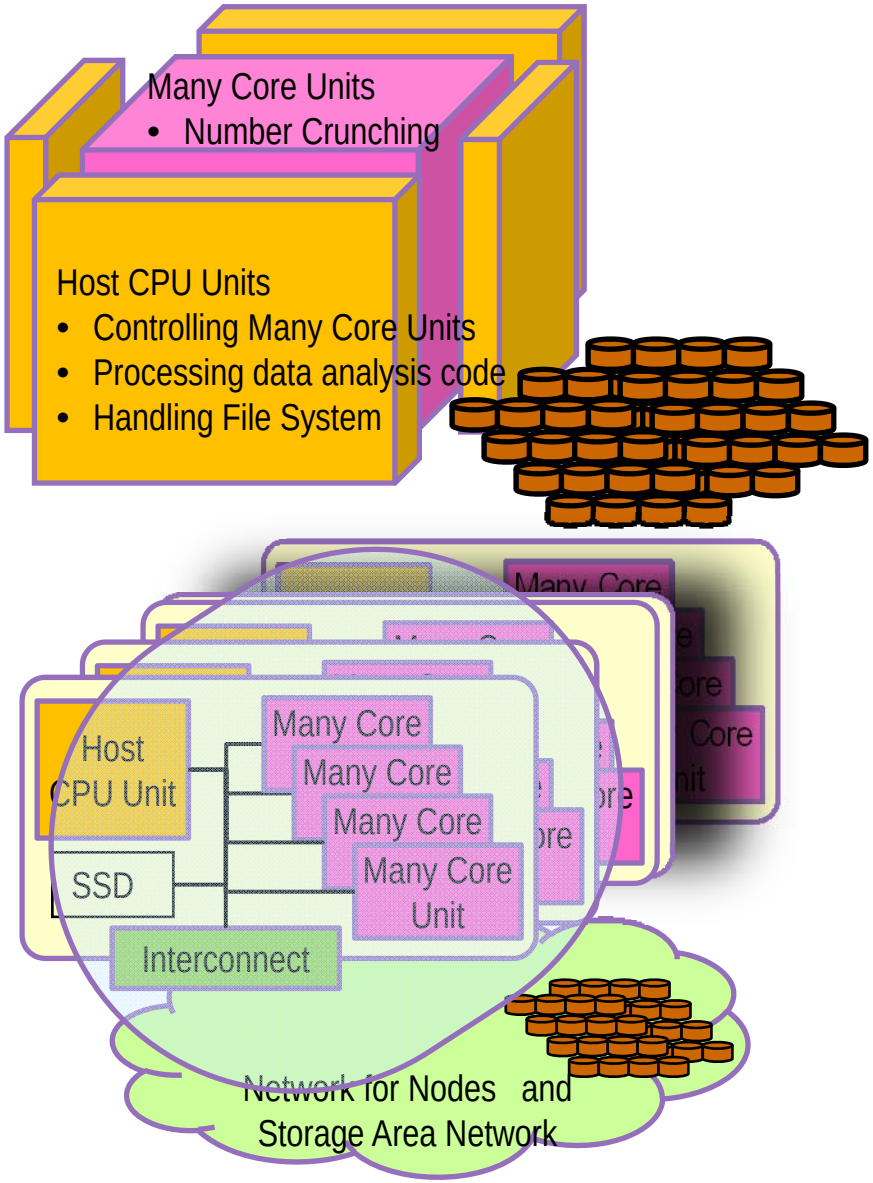
ManyCore群: 演算重視アプリケーション実行
ファイルI/O時等Host CPU使用

Host CPU群: I/O重視アプリケーション実行



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



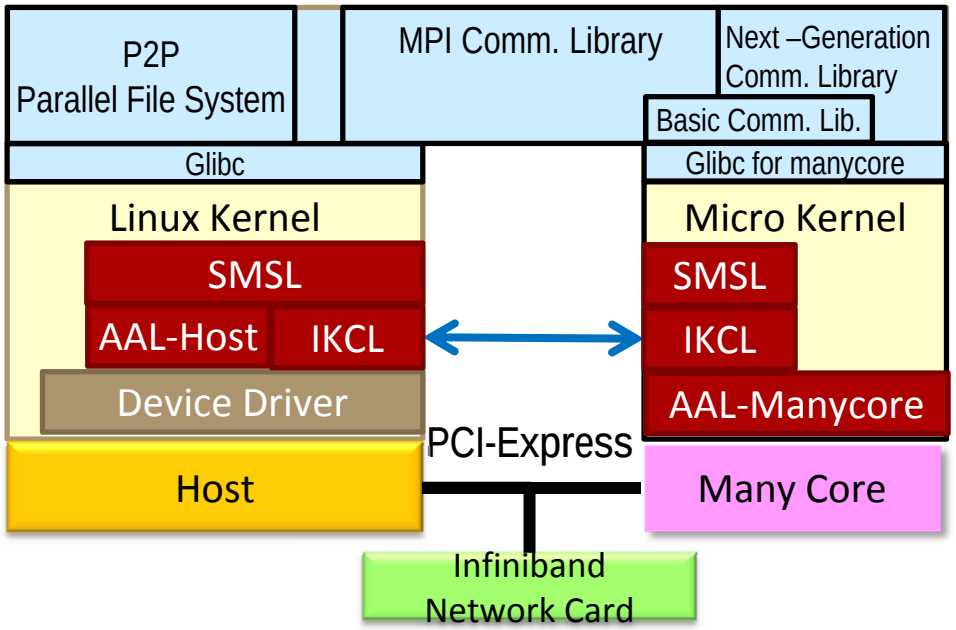
パーティション内1ジョブ実行

ManyCore群: 演算＆通信

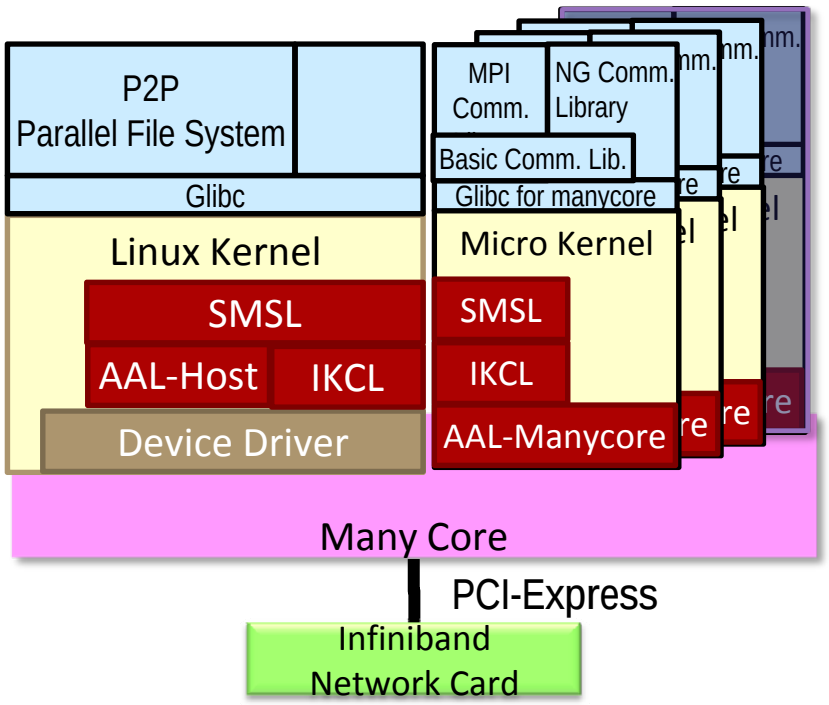
Host CPU群: メモリ共有＆通信＆I/O

ソフトウェアスタック

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 generic-purpose communication and data transfer mechanisms
- SMSL (System Service Layer)
  - Provides basic system services on top of the communication layer

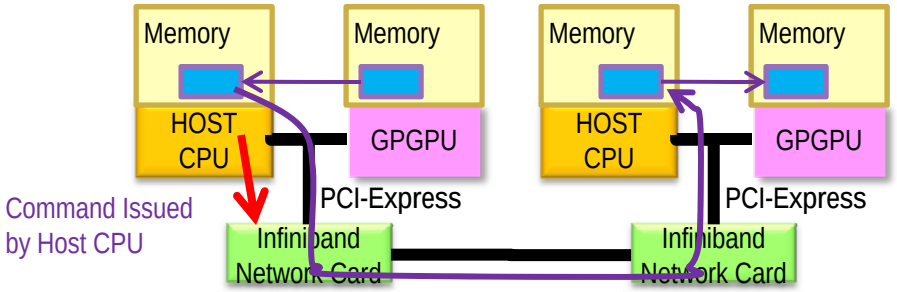
Design Criteria

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

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、Accelarator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

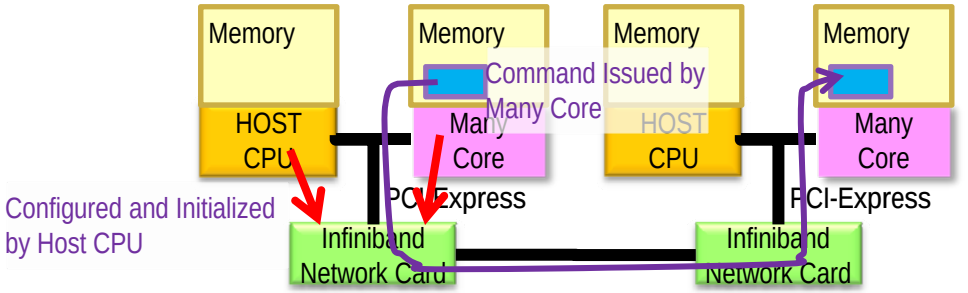


Many Core型Acceleratorの場合

- 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)



おわりに

- 次世代ヘテロジニアスPCクラスタ要件
  - 既存コモディティ部品によるPCクラスタとの親和性
    - 最小限のアプリケーションプログラム修正
  - PCサーバとアクセラレータ技術の将来動向をにらんだ設計
- 2012年度中にプロトタイプシステムリリース予定

関連発表

1. Taku Shimosawa, Hiroya Matsuba, Yutaka Ishikawa, "Logical Partitioning without Architectural Supports", 32nd IEEE Intl. Computer Software and Applications Conference (COMPSAC 2008), pp. 355-364, 2008
2. Taku Shimosawa, Yutaka Ishikawa, "Inter-kernel Communication between Multiple Kernels on Multicore Machines", IPSJ Transactions on Advanced Computing Systems Vol.2 No.4 (ACS 28), pp. 64-82, 2009
3. 下沢、石川、堀敦史、並木美太郎、辻田祐一、「メニーコア向けシステムソフトウェア開発のための実行環境の設計と実装」、情報処理学会、2011-OS-118(1), 1-7, 2011.
4. Min Si, Yutaka Ishikawa, “Design of Communication Facility on Heterogeneous Cluster ,” 情報処理学会、2012-HPC-133、2012.
5. Min Si and Yutaka Ishikawa, "Design of Direct Communication Facility for Manycore-based Accelerators, “ to appear at CASS2012 in conjunction with IPDPS2012, 2012