NICT委託付共同研究「大規模分散コンピューティングのための
高機能 ネットワークプラットフォーム技術の研究開発」

SDN + HPC = ? : -大規模コンピューティングのための高機能ネットワークプラットフォーム技術にむけて-

大阪大学サイバーメディアセンター
准教授　伊達　進

講演の内容

I. 委託付共同研究
  - 背景
  - 全体概要
  - 目的と体制

II. HPC+SDN= ?
  - 背景
    - High-Performance Computingの世界
    - HPC視点でのネットワーク
  - SDNが拓くHPCの新展開：委託付共同研究で目指す大規模コンピューティングの世界
    - SDN MPI：通信特性にfitしたプロセス間通信技術
    - SDN JMS: ネットワーク資源を計算資源とともに制御可能資源として扱うリソース管理技術

III. まとめ

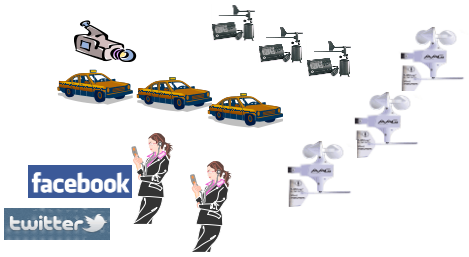
委託付共同研究の背景

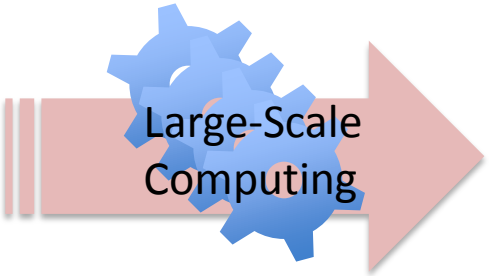
クラウドやスマートフォン・センサー・小型デバイス等の台頭に伴うネットワークサービスに対する要求の高度化

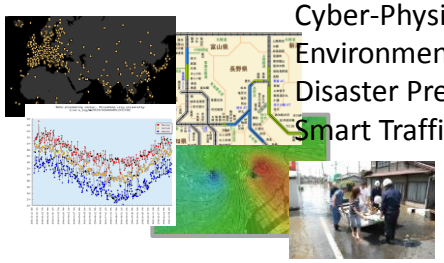
- サービスにおいて行うべき計算処理の複雑化・大規模化
  - 大容量画像マッチング
  - 高精細可視化処理
  - 大量に発生するセンサー値の意味的解析処理
- サービスが扱うべき対象データの多様化・大容量化
  - 分散して発生する大量のストリームデータ
  - 分散データベース上の大容量データ
  - 災害時等，ネットワーク不安定状況下におけるデータ

高度化する要求に対応可能な大規模分散コンピューティング環境の実現

高度ICT立国を目指す上で本技術を推進・確立していくことは必須

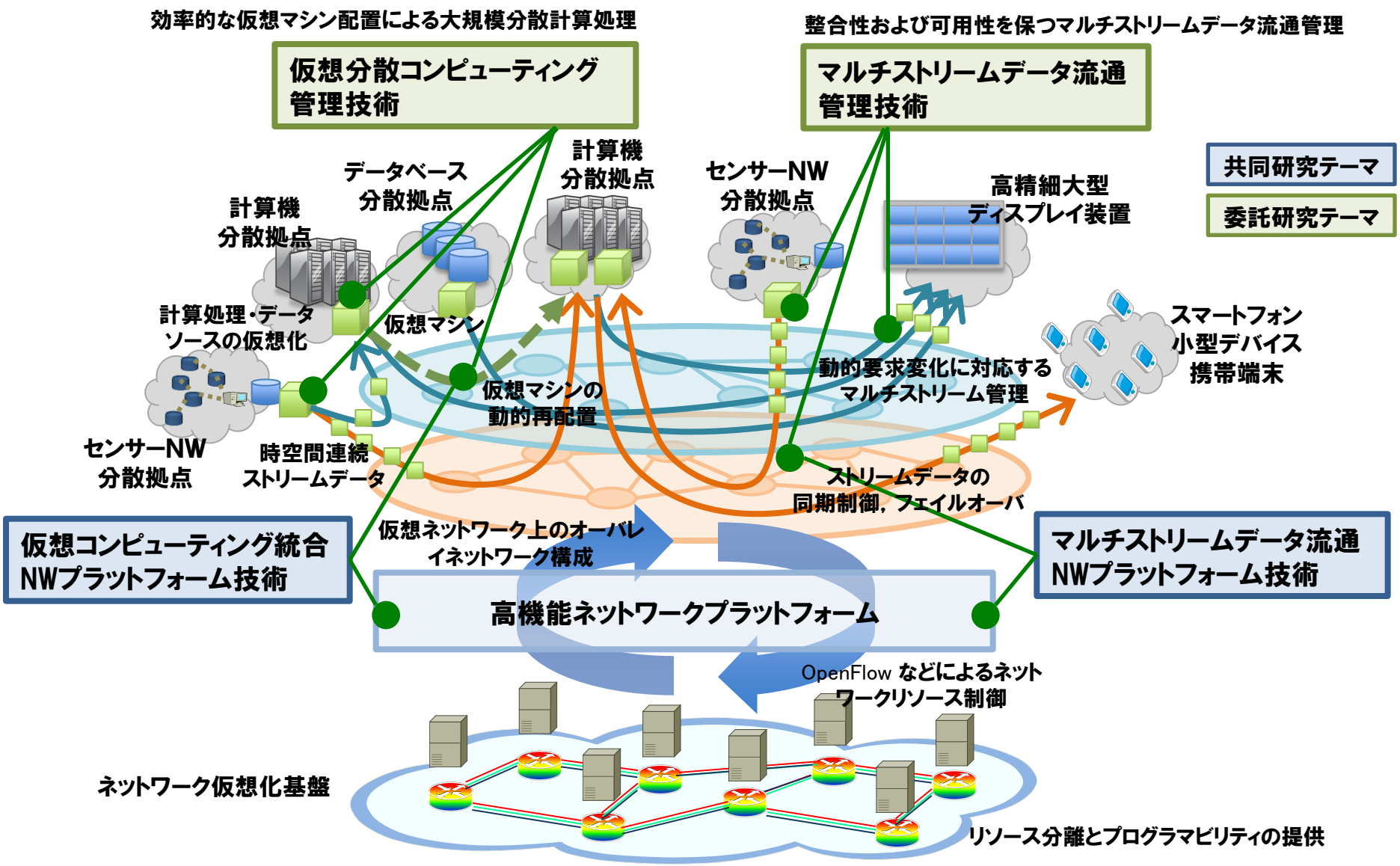






Cyber-Physical Coupling Services
Environmental Observations
Disaster Predictions
Smart Traffic Managements …

本委託付共同研究の全体概要



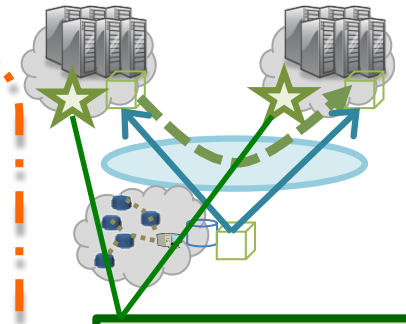
本委託研究内容

- 目的:
  - 要素技術としての仮想コンピューティング管理技術，マルチストリームデータ流通管理技術の確立
- テーマ:

本日のお話

A: 仮想分散コンピューティング管理技術

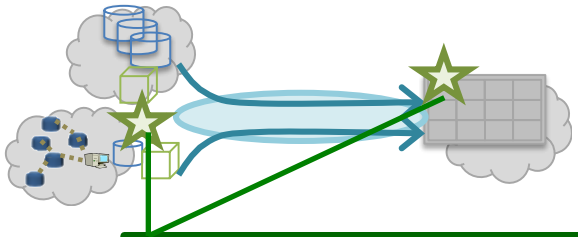
コンピューティングとネットワークのリソースを仮想化し，分散計算機能を動的に再配置することで大規模分散計算を効率的に実現



リソース仮想化によるユーザ分離と動的かつ柔軟な機能配置を行う

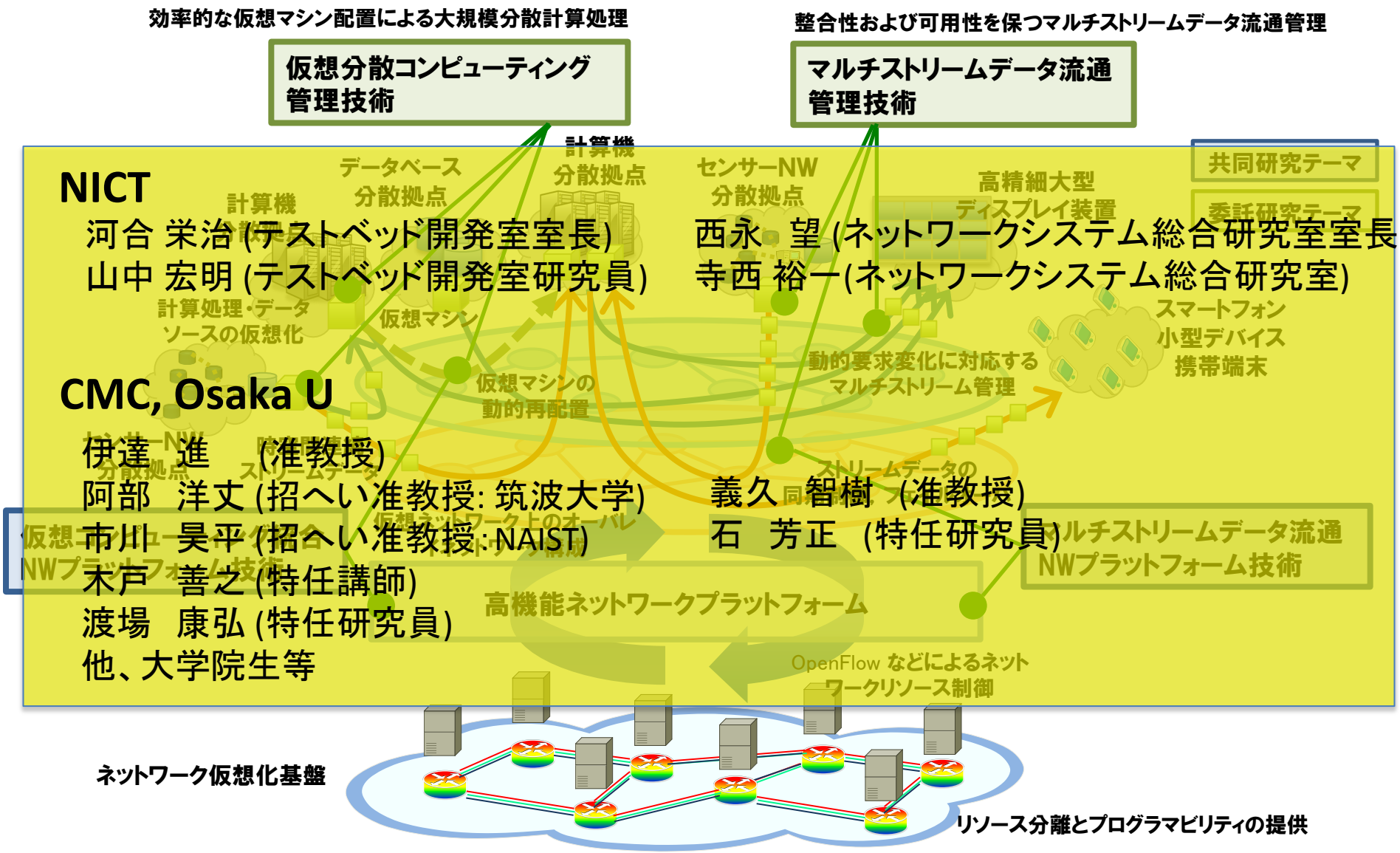
B:マルチストリームデータ流通管理技術

サービスレイヤの要求や障害等の状況を管理し，複数のストリームデータの動的な再構成や同期制御をネットワークリソース管理，フロー制御により実現



サービスとネットワークの相互連携によるリソース，フロー管理を行う

本委託付共同研究の体制



講演の内容

I. 委託付共同研究
  - 背景
  - 全体概要
  - 目的と体制

II. HPC+SDN= ?
  - 背景
    - High-Performance Computingの世界
    - HPC視点でのネットワーク
  - SDNが拓くHPCの新展開：委託付共同研究で目指す大規模コンピューティングの世界
    - SDN MPI：通信特性にfitしたプロセス間通信技術
    - SDN JMS: ネットワーク資源を計算資源とともに制御可能資源として扱うリソース管理技術

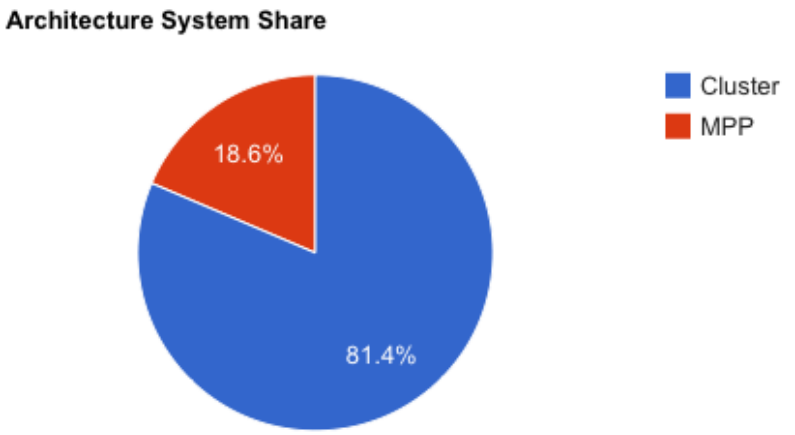
HPCの動向 (1)

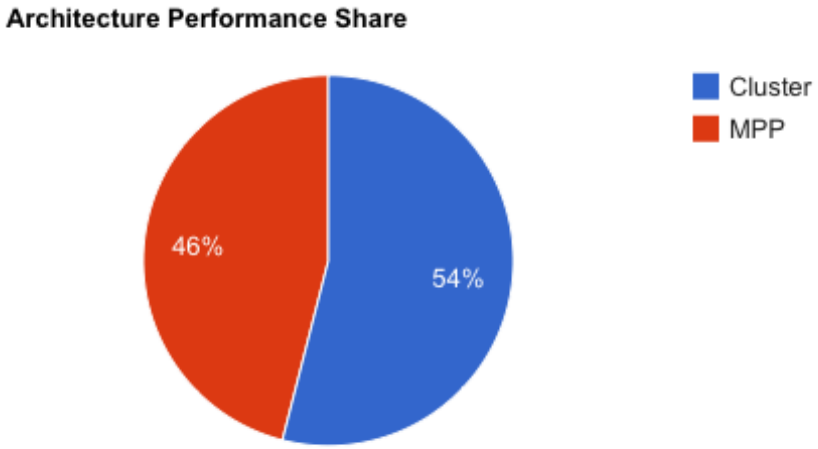
- TOP500 http://www.top500.org
  - LINPACKベンチマーク: 線形代数演算を使った性能評価
    - High-Performance Linpack Benchmark, http://www.netlib.org/benchmark/hpl/
  - 高速な計算機システムの上位500位までをランキング
  - 1993年以来、6 月と11月に開催される国際会議にあわせ年2回発表
    - June: International Supercomputer Conference (ヨーロッパで開催)
    - Nov. : IEEE Supercomputing Conference (米国で開催)

Rank	Site	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	DOE/SC/Oak Ridge National Laboratory United States	**Titan** - Cray XK7 , Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x Cray Inc.	560640	17590.0	27112.5	8209
2	DOE/NNSA/LLNL United States	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom IBM	1572864	16324.8	20132.7	7890
3	RIKEN Advanced Institute for Computational Science (AICS) Japan	K computer, SPARC64 VIIIfx 2.0GHz, Tofu interconnect Fujitsu	705024	10510.0	11280.4	12660
4	DOE/SC/Argonne National Laboratory United States	**Mira** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	786432	8162.4	10066.3	3945
5	Forschungszentrum Juelich (FZJ) Germany	**JUQUEEN** - BlueGene/Q, Power BQC 16C 1.600GHz, Custom Interconnect IBM	393216	4141.2	5033.2	1970
6	Leibniz Rechenzentrum Germany	**SuperMUC** - iDataPlex DX360M4, Xeon E5-2680 8C 2.70GHz, Infiniband FDR IBM	147456	2897.0	3185.1	3423
7	Texas Advanced Computing Center/Univ. of Texas United States	**Stampede** - PowerEdge C8220, Xeon E5-2680 8C 2.700GHz, Infiniband FDR, Intel Xeon Phi Dell	204900	2660.3	3959.0	
8	National Supercomputing Center in Tianjin China	**Tianhe-1A** - NUDT YH MPP, Xeon X5670 6C 2.93 GHz, NVIDIA 2050 NUDT	186368	2566.0	4701.0	4040
9	CINECA Italy	**Fermi** - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM	163840	1725.5	2097.2	822
10	IBM Development Engineering United States	**DARPA Trial Subset** - Power 775, POWER7 8C 3.836GHz, Custom Interconnect IBM	63360	1515.0	1944.4	3576

HPCの動向 (2)

- TOP500にみるコンピュータアーキテクチャ





Top500 2012年6月時点で、Top500リストに
含まれている計算機アーキテクチャの割合

MPP: Massive Parallel Processor

Architecture	Count	System Share (%)	Rmax (GFlops)	Rpeak (GFlops)	Cores
Cluster	407	81.4	66633050.71	99141890.55	7169077
MPP	93	18.6	56784736	72729253.98	6258868

ほとんどがクラスタシステムになりつつある

HPC視点でのネットワーク

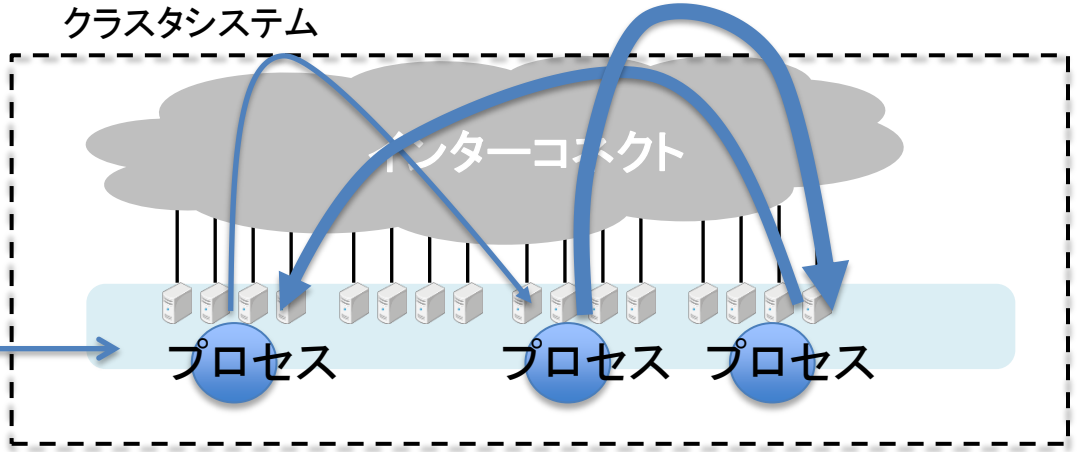
- ユーザからは細粒度な操作ができない”静的”な資源
  - ネットワークは“土管”でいい
  - ネットワークに余計な機能はもとめない。

ネットワークは十分余裕を持って設計、実装されている。
or
ネットワークは内部でうまく制御されていて、困らない。

どのような通信がどのように発生するかを考慮したネットワーク制御はしない。
(ネットワークが十分な資源量がある、ネットワーク内でうまく制御される)

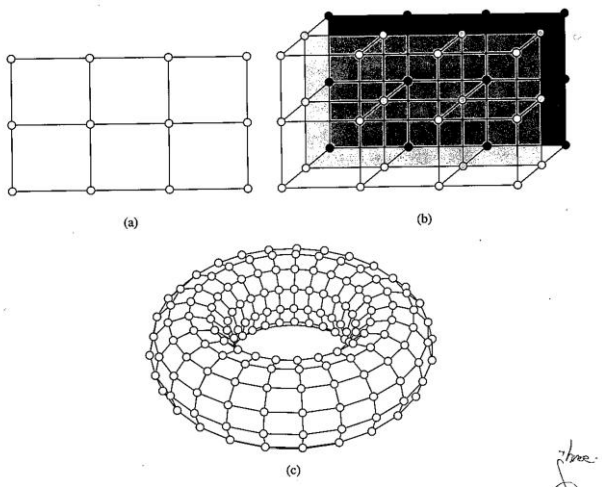
HPCユーザ

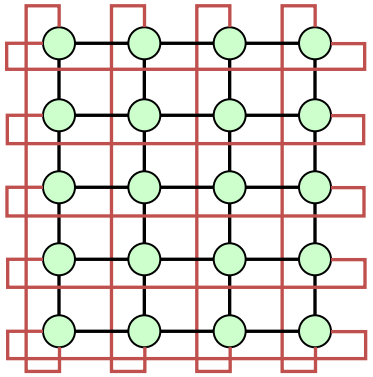
計算ジョブ投入



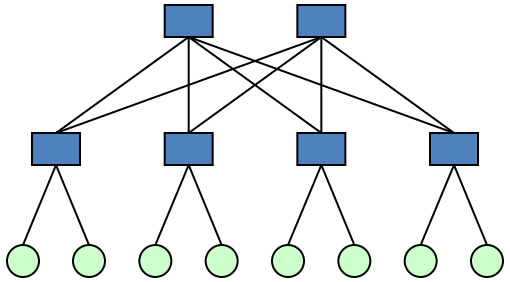
インターコネクト・トポロジのいろいろ

- 直接網
  - Full Connect
  - Mesh
  - Torus
  - Hypercube
  - Direct Tree
  - Recursive Diagonal Torus (RDT)
  - Tofu（6次元トーラス・メッシュ網）
- 間接網
  - Crossbar
  - Multi-stage Interconnection Network (MIN)
  - Multi-Dimensional Crossbar (MDX)
  - Tree
  - Fat Tree





2D Mesh/Torus



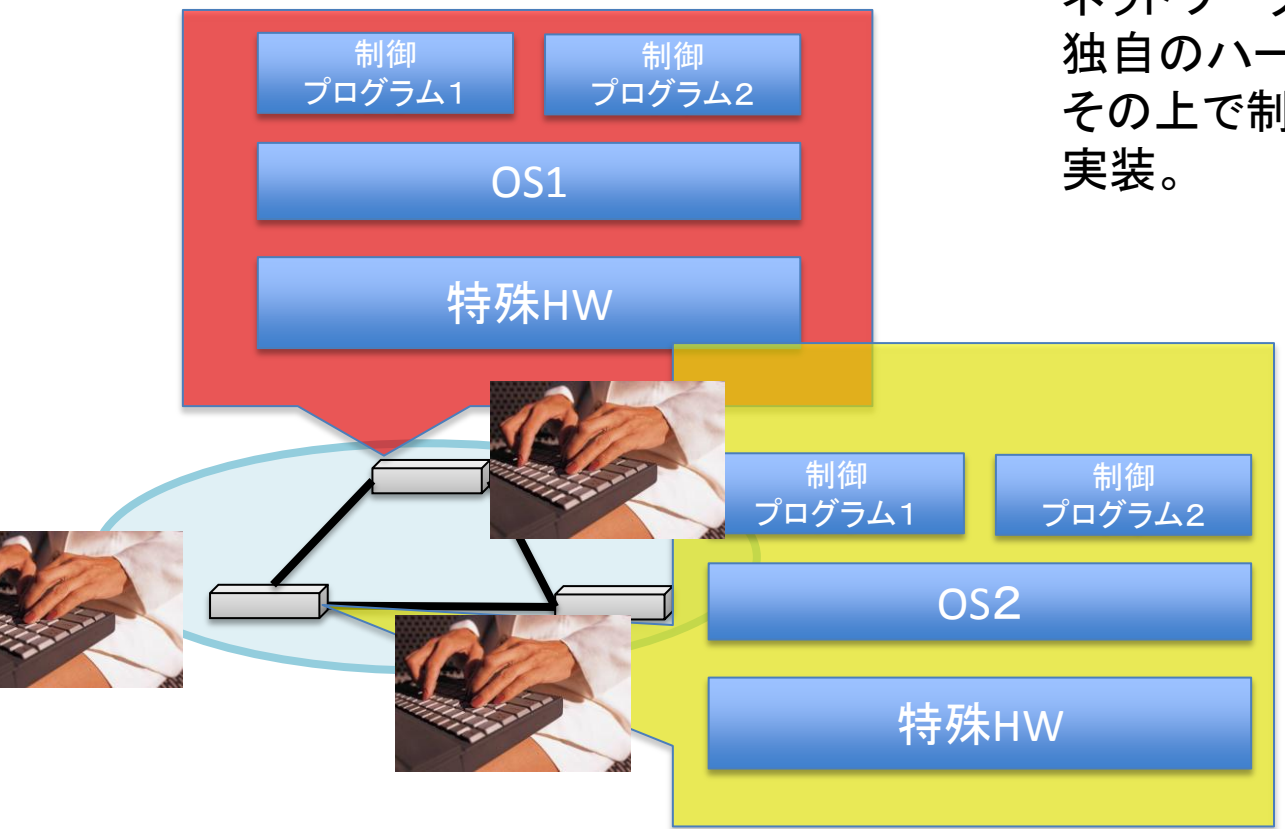
Fat Tree

ノード数が増加すればするほど
複雑化し、コスト高になる。
実装も難しくなっていく。

Software-Define Network at a glance

- 新しいネットワークアーキテクチャのコンセプト
  - 従来のネットワーク機能をコントロールプレーン(制御)とデータプレーン(パケット転送)に分離

従来のネットワーク



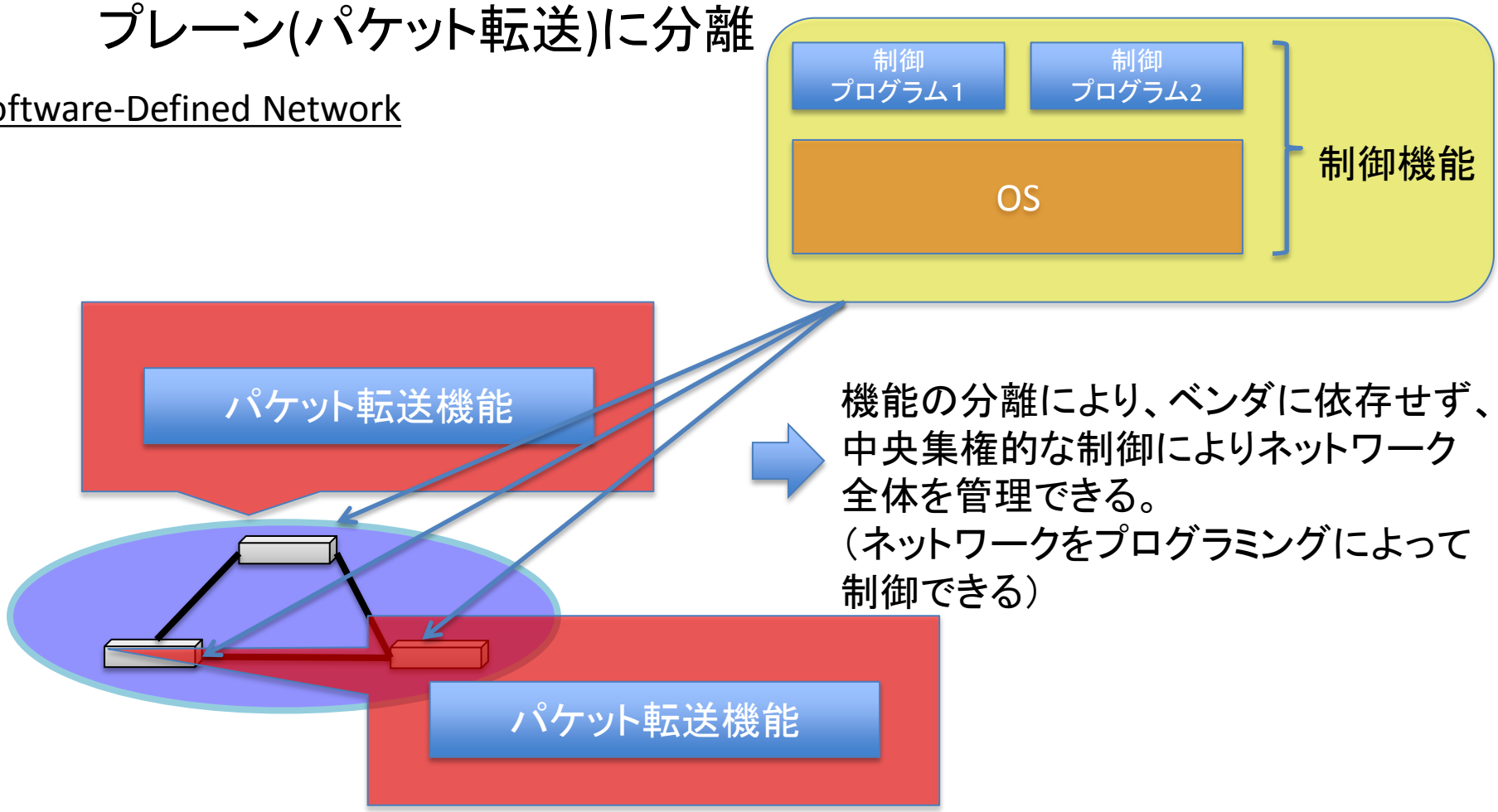
ネットワーク機器を提供するベンダ毎に独自のハードウェア、独自のOSを用意し、その上で制御プログラム(ルーティング)を実装。

- 機器ごとに異なるI/Fをつかって個別に設定、管理する必要がある。

Software-Define Network at a glance

- 新しいネットワークアーキテクチャのコンセプト
  - 従来のネットワーク機能をコントロールプレーン(制御)とデータプレーン(パケット転送)に分離

Software-Defined Network



機能の分離により、ベンダに依存せず、中央集権的な制御によりネットワーク全体を管理できる。
（ネットワークをプログラミングによって制御できる）

[目指す世界] SDN が拓く HPC の新展開

- SDNによってもたらされるネットワークプログラミング可能性を最大限に活用した、次世代型の分散並列計算・可視化のための要素技術に関する研究開発

静的なネットワーク資源を前提として計算性能を追求するHPCから、
ネットワークを動的制御できる資源として利用し、計算性能を追求する新たなHPCへのパラダイムシフト

計算要求

(1) インターコネクトを制御可能資源として、プロセッサ資源、ストレージ資源等とともに総合的に扱うRM(リソースマネージメント)技術

(4) 多拠点の計算資源をオンデマンドに集約・統合するクラウド構成技術

仮想クラスタ

(2) 通信特性にfitした実行時プロセス間通信制御技術

プロセス

プロセス

プロセス

プロセス

インターコネクトコントローラ

拠点内SDN

可視化要求

広域SDN

広域SDNコントローラ

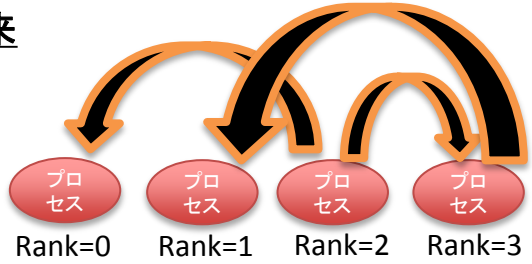
(3) ユーザの可視化操作に対して変動するネットワーク要求に応じたフロー制御技術

SDN MPI (Message Passing Interface)：実行時プロセス間通信制御

MPIによる分散並列処理実行時に発生する、複数プロセス間でデータ交換を行う集合通信の実行時間の短縮を目的とし、集合通信の通信パターンを考慮した効率的なデータ配送を実現する技術の実現を狙う。

MPI_Bcast(&x, 1, MPI_FLOAT, 2, MPI_COMM_WORLD)

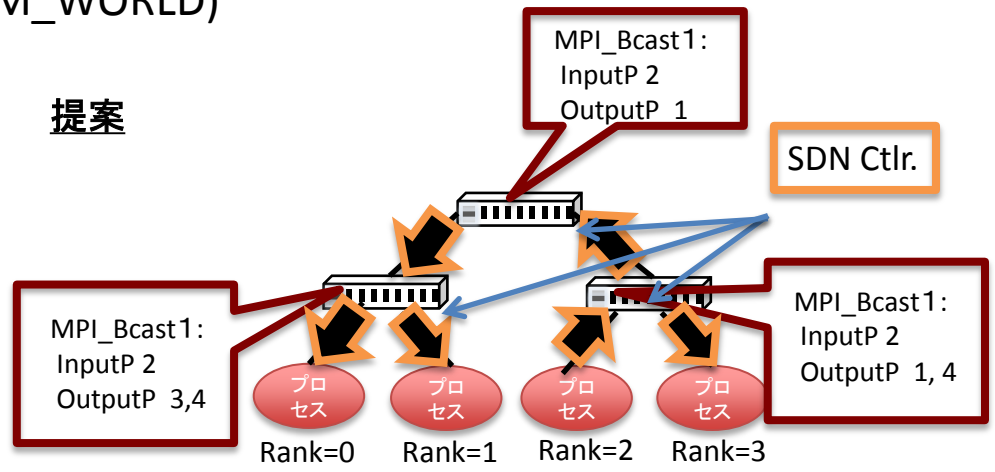
従来



MPIはAPIを規定するのみで、実装は委ねられている。

- 特定のアーキテクチャに依存しないMPI実装では、通信が行われるネットワークを想定せず、最適な通信トポロジ、アルゴリズムに基づき実装。
  - 実際のネットワーク構成と合致せず非効率な通信になりうる。
- ベンダ提供のMPIは、ベンダの提供するアーキテクチャに通信を最適化したMPIを提供
  - HPCシステムの大規模化により、コスト面、技術面ともに困難化

提案



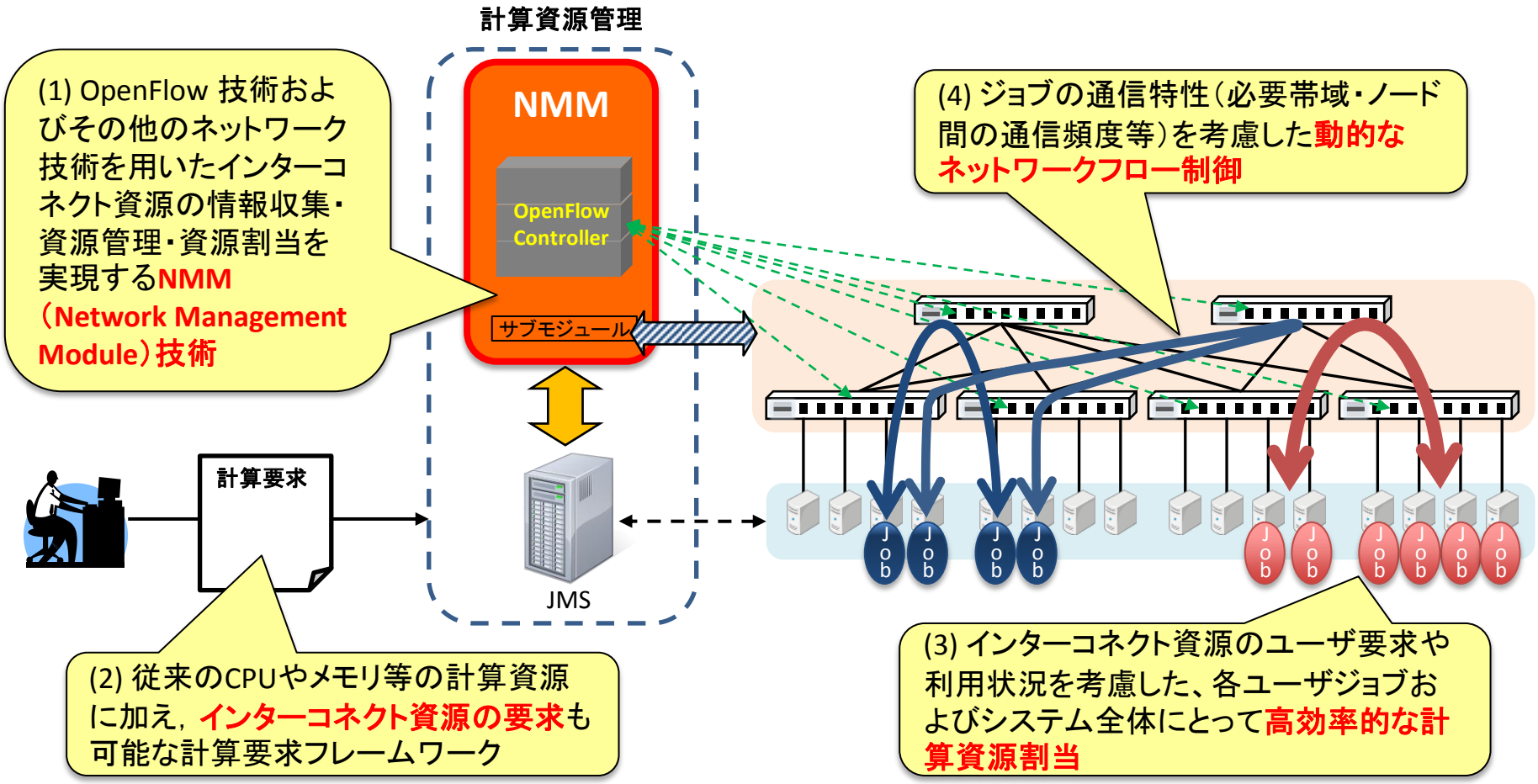
集合通信の通信パターンに、SDNのプログラマビリティを応用し、集合通信の実行時間を短縮（集合通信の効率化）する

今後の課題

- 初期設定時のオーバヘッド検証
- 実OpenFlowスイッチを用いた性能測定
- より複雑な集合通信(MPI_Reduce等)への応用

SDN JMS：リソースマネージメント技術

- インターコネクト資源の管理によるユーザジョブおよび計算資源全体の両面で高パフォーマンスかつ高効率な資源割当を実現するJMS（ジョブ管理システム）



まとめ

- NICT委託付共同研究「大規模分散コンピューティングのための高機能ネットワークプラットフォーム技術の研究開発」について概説した
  - SDNによってもたらされるネットワークプログラミング可能性を最大限に活用した、次世代型の分散並列計算・可視化のための要素技術に関する研究開発
    - SDN MPI：通信特性にfitしたプロセス間通信技術
    - SDN JMS: ネットワーク資源を計算資源とともに制御可能資源として扱うリソース管理技術