FUJITSU
shaping tomorrow with you

スーパーコンピュータ「京」の開発

2011年11月4日

富士通株式会社
次世代テクニカルコンピューティング開発本部
システム開発統括部長　新庄

※「京」は2010年7月に理化学研究所様が発表した「次世代スーパーコンピュータ」の愛称です

FUJITSU

1.「京」の概要

- システムの概要
- ソフトウェアの概要
- システムの信頼性

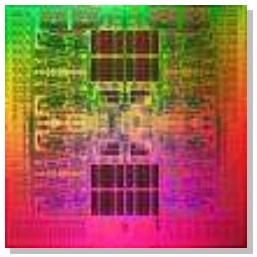
FUJITSU

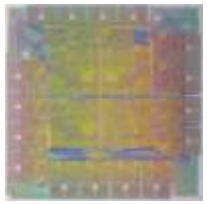
1.「京」の概要

- システムの概要
- ソフトウェアの概要
- システムの信頼性

「京」(注) システム概要

FUJITSU





プロセッサ: SPARC64™ VIIIfx

- 富士通の最先端半導体テクノロジ(45nm)
- 8プロセッサコア，キャッシュメモリ及びメモリコントローラを1チップに集積
- 高性能・高信頼と低消費電力を両立

インターコネクトコントローラ:ICC

- 直接網6次元メッシュトーラス(Tofu)を実装

システムボード:高効率冷却

- 4計算ノードを実装
- プロセッサ、ICCほか主要部品を水冷
- LSI温度を抑制し，消費電力を低減，部品寿命向上

ラック:高密度実装

- 1ラックに約100ノードを搭載
  - 24枚のシステムボード
  - IO用システムボード
  - システム用磁気ディスク装置
  - 電源 など
- 従来比10倍以上のラックあたり性能を実現

(10PFlops: 800ラック以上)

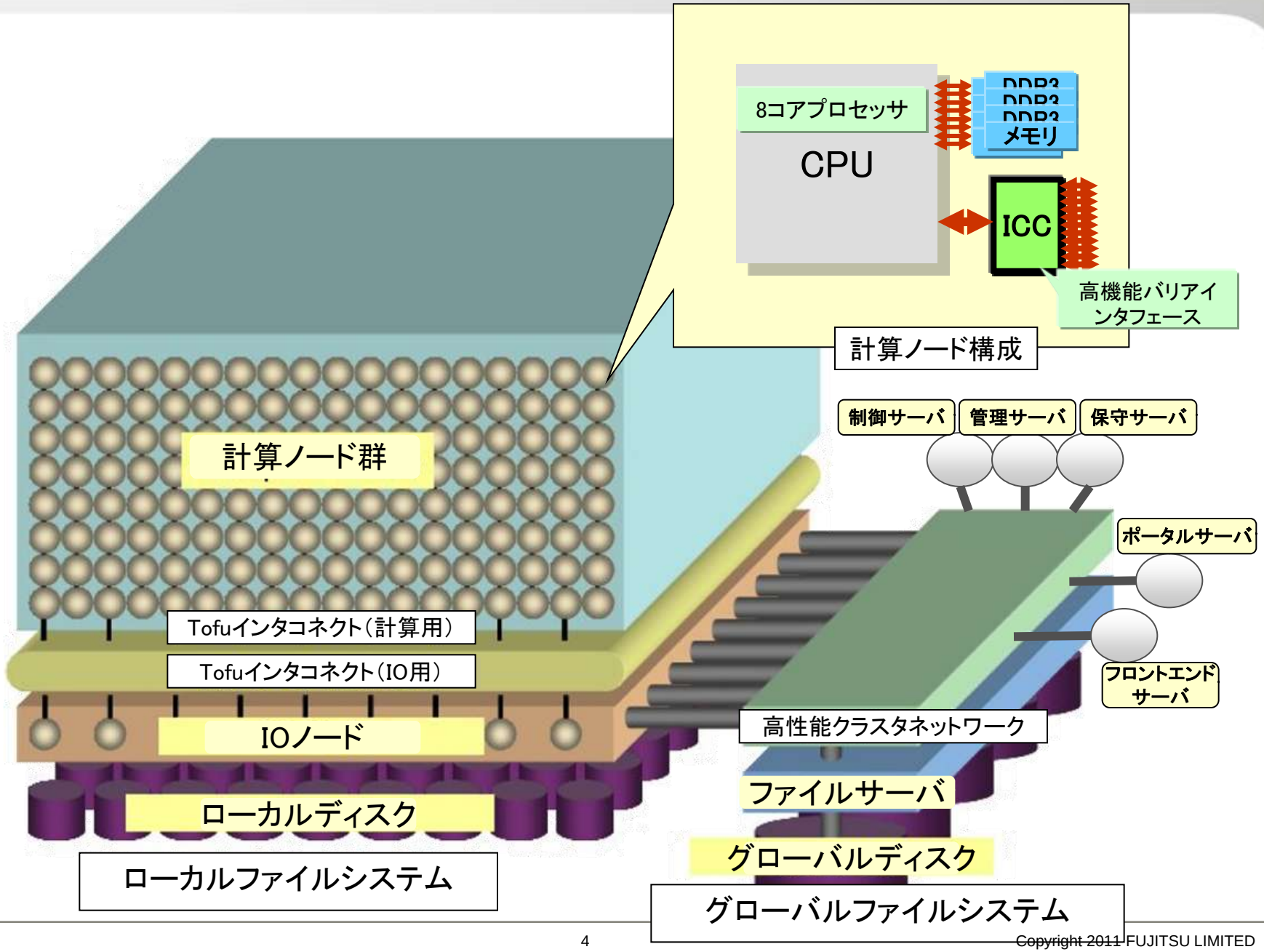
注: 2010年7月に理化学研究所様が発表した「次世代スーパーコンピュータ」の愛称です
「京」は理化学研究所様と共同開発中です

システム

- 世界最高性能への挑戦 (初代地球シミュレータの250倍以上)
- 超大規模システム(8万プロセッサ以上)を安定稼動

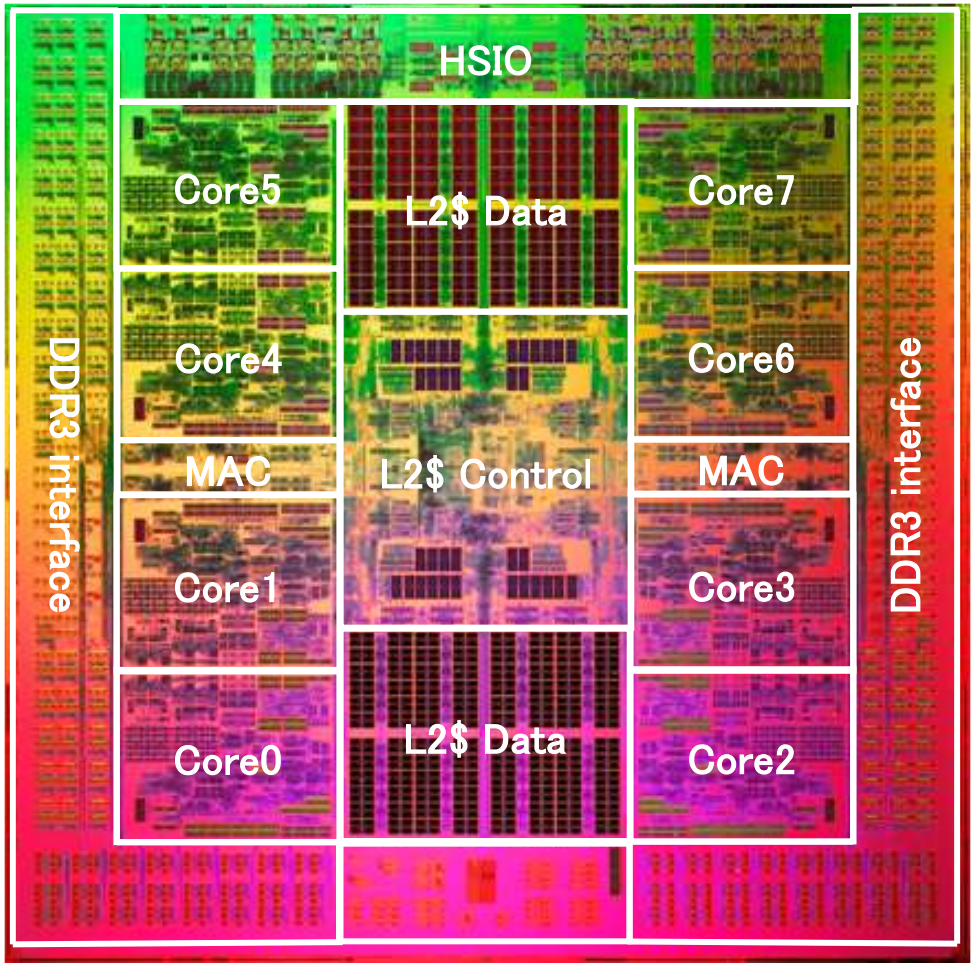
システムイメージ

システム構成



SPARC64™ VIIIfx Chip 概要

設計目標：高性能、省電力かつ高信頼性



- **基本仕様**
  - 8 コア、6 MB 共有L2キャッシュ
  - メモリコントローラ内蔵
  - クロック 2 GHz
  - HPC向け命令拡張(HPC-ACE)
- **FMLの 45nm CMOS**
  - 22.7mm x 22.6mm
  - 760M トランジスタ、1271信号
  - 信号ピン数 1271
- **ピーク性能**
  - 演算性能　128GFlops
  - メモリスループット　64GB/s
- **消費電力**
  - 58W (TYP, 30℃)
  - 水冷
    - リーク電流削減、信頼性向上

HPC-ACE

FUJITSU

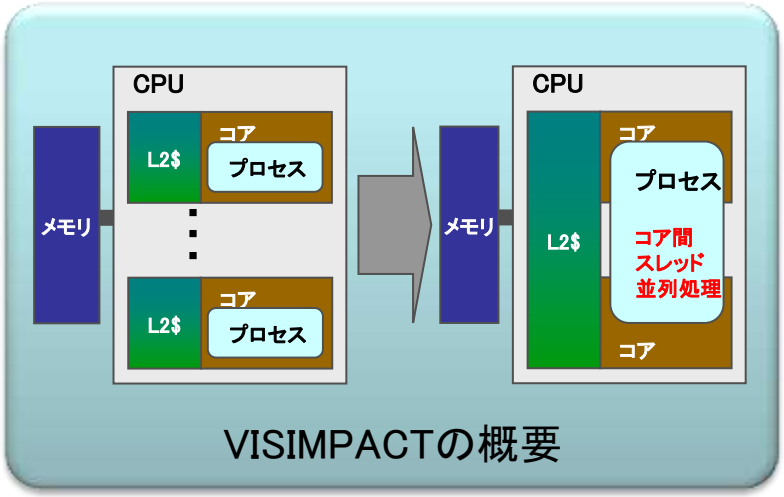
(High Performance Computing - Arithmetic Computational Extensions)

- 富士通独自のHPC向け命令セット拡張
  - 準拠仕様
    - **SPARC-V9 仕様**
    - **JPS (Joint Programmer's Specification): SPARC-V9拡張仕様**
  - 主な拡張内容
    - **ハードバリア**
    - **レジスタ数の拡張**
    - **SIMD (single instruction multiple data) 命令**
      - コア当たり2SIMD x 2pipe
      - 128レジスタ
      - クロス演算
      - マスク演算
    - **セクターキャッシュ**
    - **科学技術計算を加速する命令**
      - 除算/平方根の逆数近似命令
      - 三角関数補助命令
      - 複素数計算の効率化 (クロスSIMD演算命令の活用)

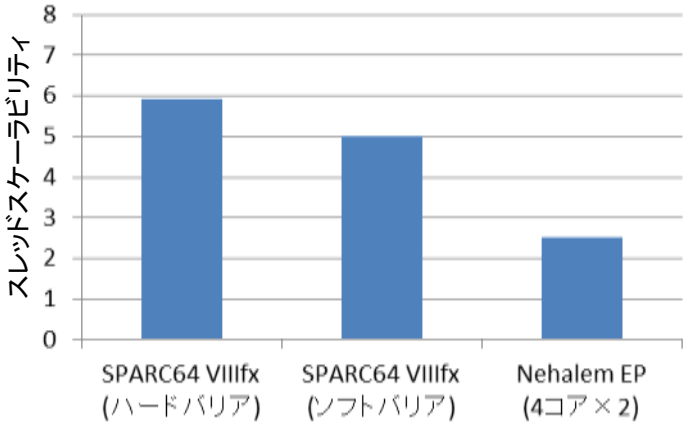
VISIMPACT (Virtual Single Processor by Integrated Multi-core Parallel Architecture)

FUJITSU

- VISIMPACTは、プロセス数を減らすために、マルチコアCPUを一つの高速なプロセッサとして扱う仕組み
  - CPU技術とコンパイラ技術を統合して、高効率なスレッド並列を実現
    - 低オーバヘッドを実現して最内ループ並列化も可能とするCPU技術
      - ソフトバリアより10倍高速なコア間ハードバリア
      - コア間のデータのfalse sharingを防止するコア間共有L2キャッシュ
    - 複雑な多重ループを最適に並列化するコンパイラ技術
      - 自動ベクトル化技術を発展させた高度な自動スレッド並列化コンパイラ



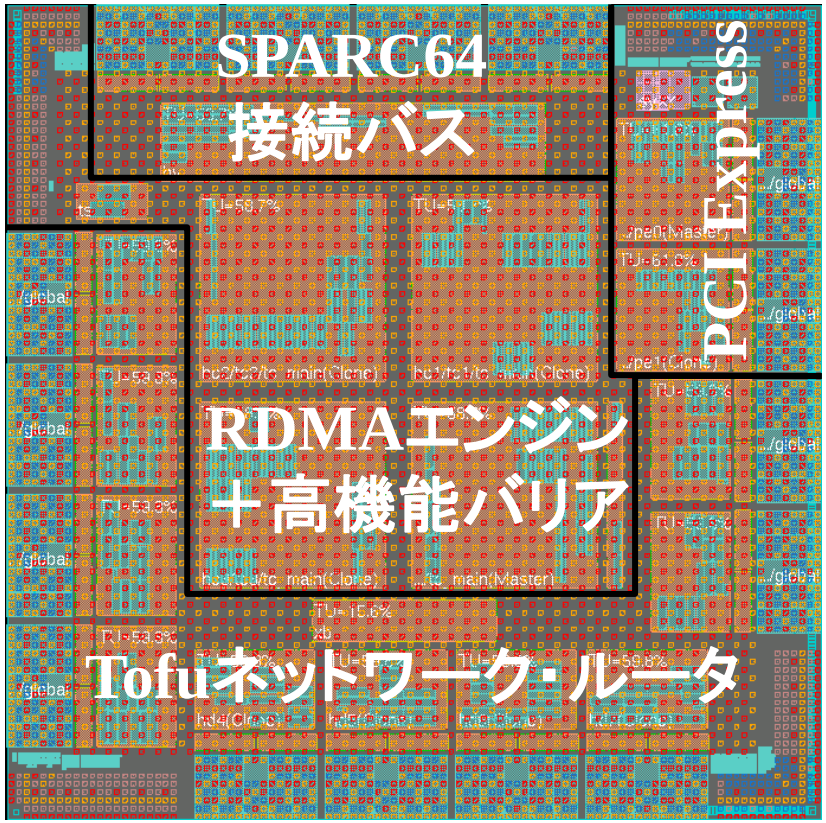
VISIMPACTの概要



実アプリカーネルにおける最内ループ並列時の
ハードバリアと共有L2キャッシュの効果の例

InterConnect Controller（ICC）Chip 概要

設計目標：高帯域、低遅延、
省電力、高信頼性

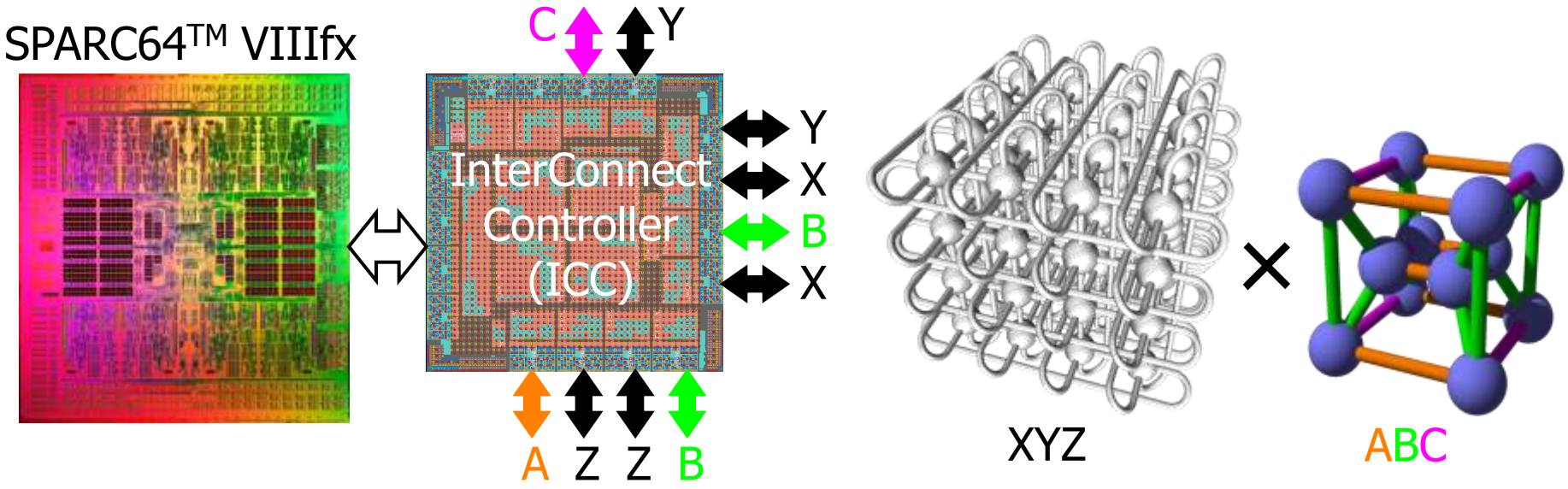


- **基本仕様**
  - RDMAエンジン×4＋高機能バリア
  - ネットワーク次数 10
  - PCI Expressルート機能内蔵
  - クロック 312.5 MHz
- **65nm ASIC**
  - 18.2mm x 18.1mm
  - ゲート数 48M
  - SRAM 12Mbit
  - 差動入出力信号 128レーン
- **高帯域**
  - リンク帯域 5GB/s×双方向
  - スイッチング容量 100GB/s
- **低遅延**
  - Virtual Cut-Through転送 ~100ns

Tofuインターコネクト 概要

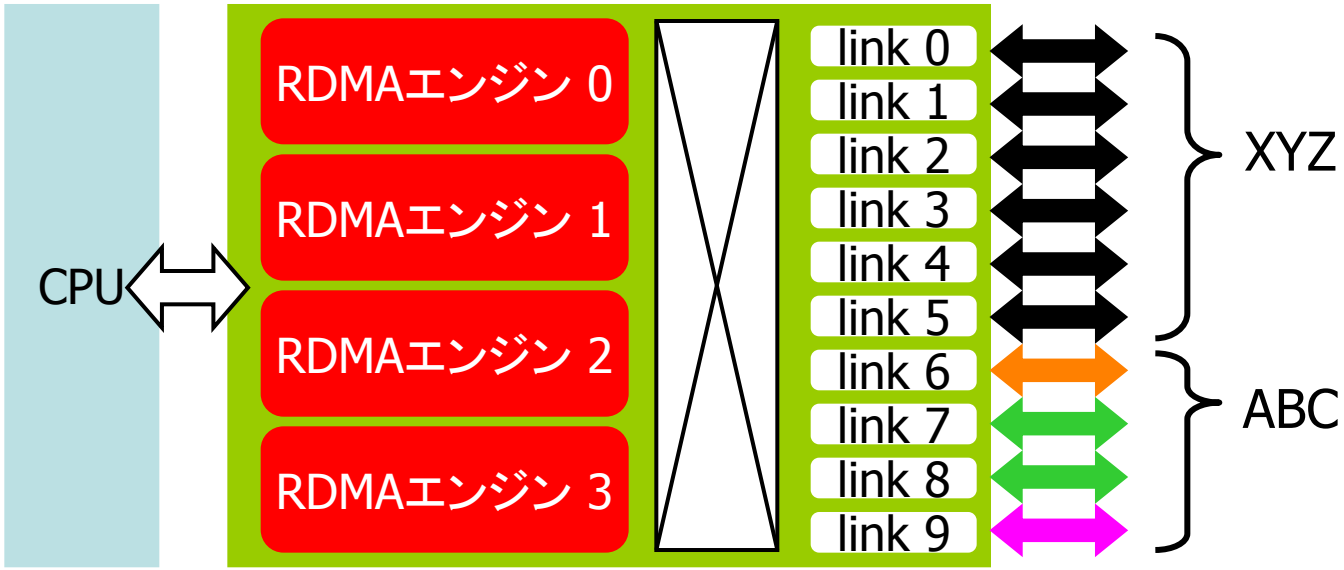
- SPARC64™ VIIIfx専用のノード間インターコネクト
- Tofu: “**To**rus **fu**sion”

ネットワーク・トポロジ	6次元メッシュ／トーラス
座標軸	X, Y, Z, A, B, C
最大ネットワーク・サイズ	32, 32, 32, 2, 3, 2
「京」システム構成	トーラス軸: X, Z, B ／ メッシュ軸: Y, A, C 計算ノード: Z = 1～16 ／ IOノード: Z = 0



インターコネクト性能、同時通信数

- ポート数10（XYZ軸6ポート＋ABC軸4ポート）
- 4つのRDMAエンジンを搭載、同時に4送信4受信が可能

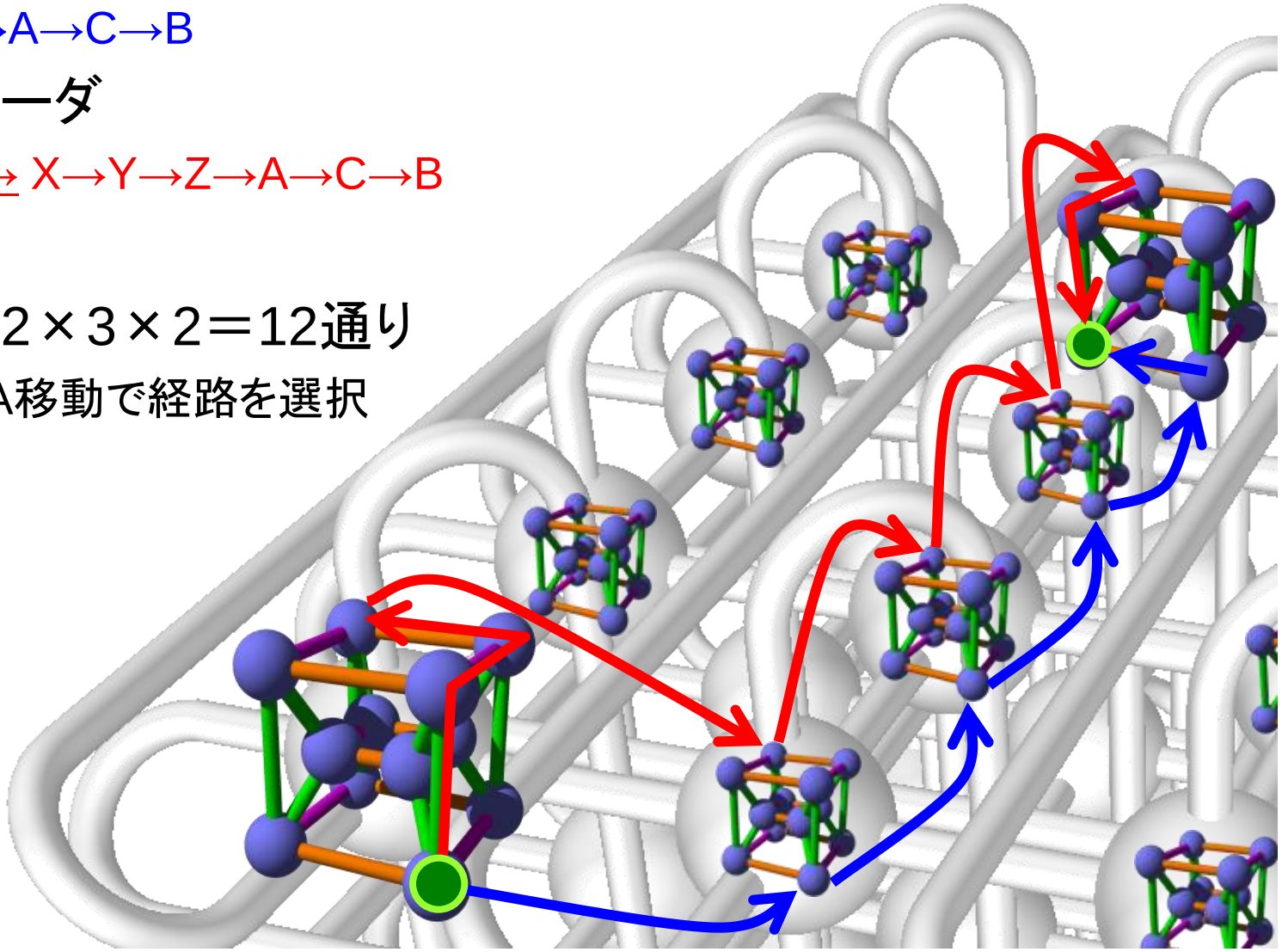


ノードあたり 理論性能	TSUBAME 2.0 InfiniBand QDR	Cray XE6 Hopper Gemini 1.2	「京」 Tofu Interconnect	IBM Blue Gene/Q 5D-Torus
演算性能	2391 GFlops	153.6 GFlops	128 GFlops	204.8 GFlops
リンク帯域（片方向）	4 GB/s	5.8 GB/s	5 GB/s	2 GB/s
同時通信数	2	1	4	10
同時通信帯域（片方向）	8 GB/s	8.3 GB/s	20 GB/s	20 GB/s

ルーティング・アルゴリズム

FUJITSU

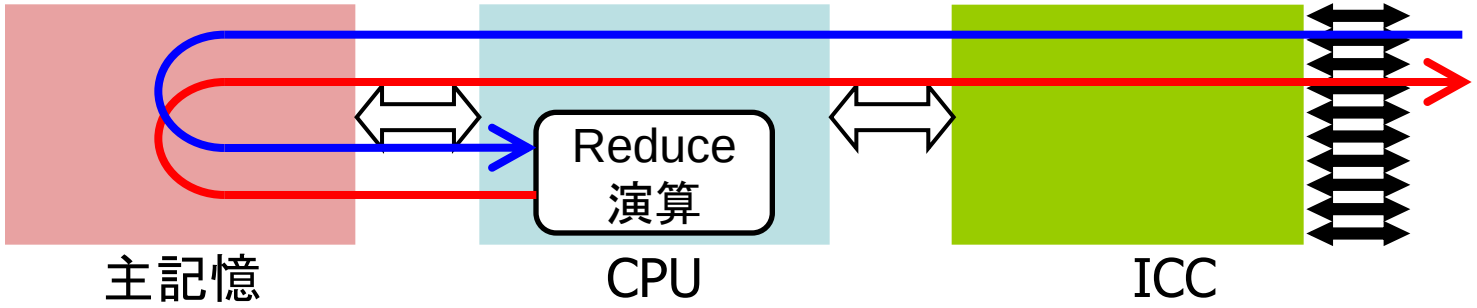
- デフォルトの次元オーダ
  - X→Y→Z→A→C→B
- 拡張次元オーダ
  - **B→C→A→** X→Y→Z→A→C→B
- XYZ経路は2×3×2＝12通り
  - 最初のBCA移動で経路を選択



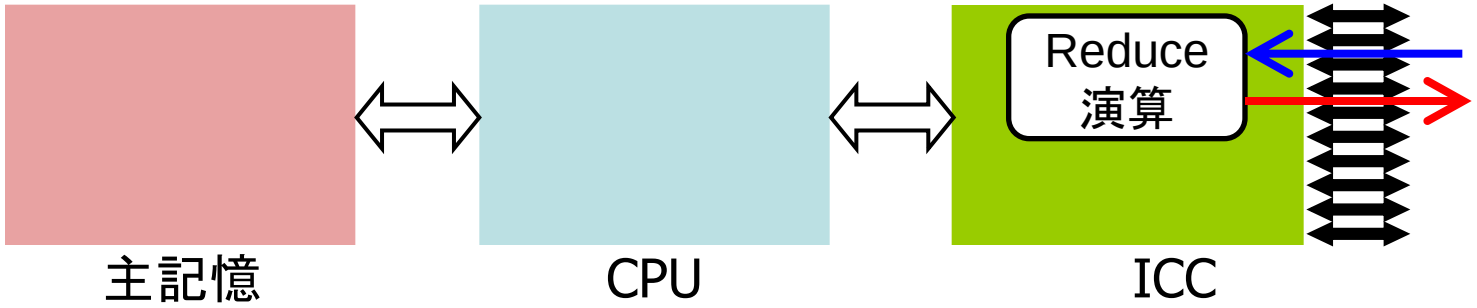
高機能バリア

- バリア同期とAllreduce集団通信に対応
  - 64ビット整数: AND, OR, XOR MAX, SUM
  - 独自160ビット浮動小数点: SUM
- Nステップ（バタフライ通信）または2Nステップ（ツリー通信）で2^Nノードを同期
- 高機能バリアは低遅延かつOSジッタ影響を受けない

ソフトウェアによる通信処理（1ステップ）



高機能バリアによる通信処理（1ステップ）



1.「京」の概要

- システムの概要
- ソフトウェアの概要
- システムの信頼性

ソフトウェア体系

FUJITSU

ユーザ/ISVアプリ

ポータル/可視化ツール

OS/運用管理

ジョブ運用管理
- ジョブ投入・実行・状態管理
- 資源割当・配分制御
- 統計・課金情報

システム運用管理
- システム導入、ソフト保守
- システム起動・停止、障害監視
- システム構成制御、保守資料採取

ファイルシステム

高性能ファイルシステム
- Lustre ベースのクラスタファイルシステム (FEFS)

言語システム

コンパイラ
- Fortran
- C/C++
- XPFortran

並列言語
- 自動並列
- OpenMP
- MPI

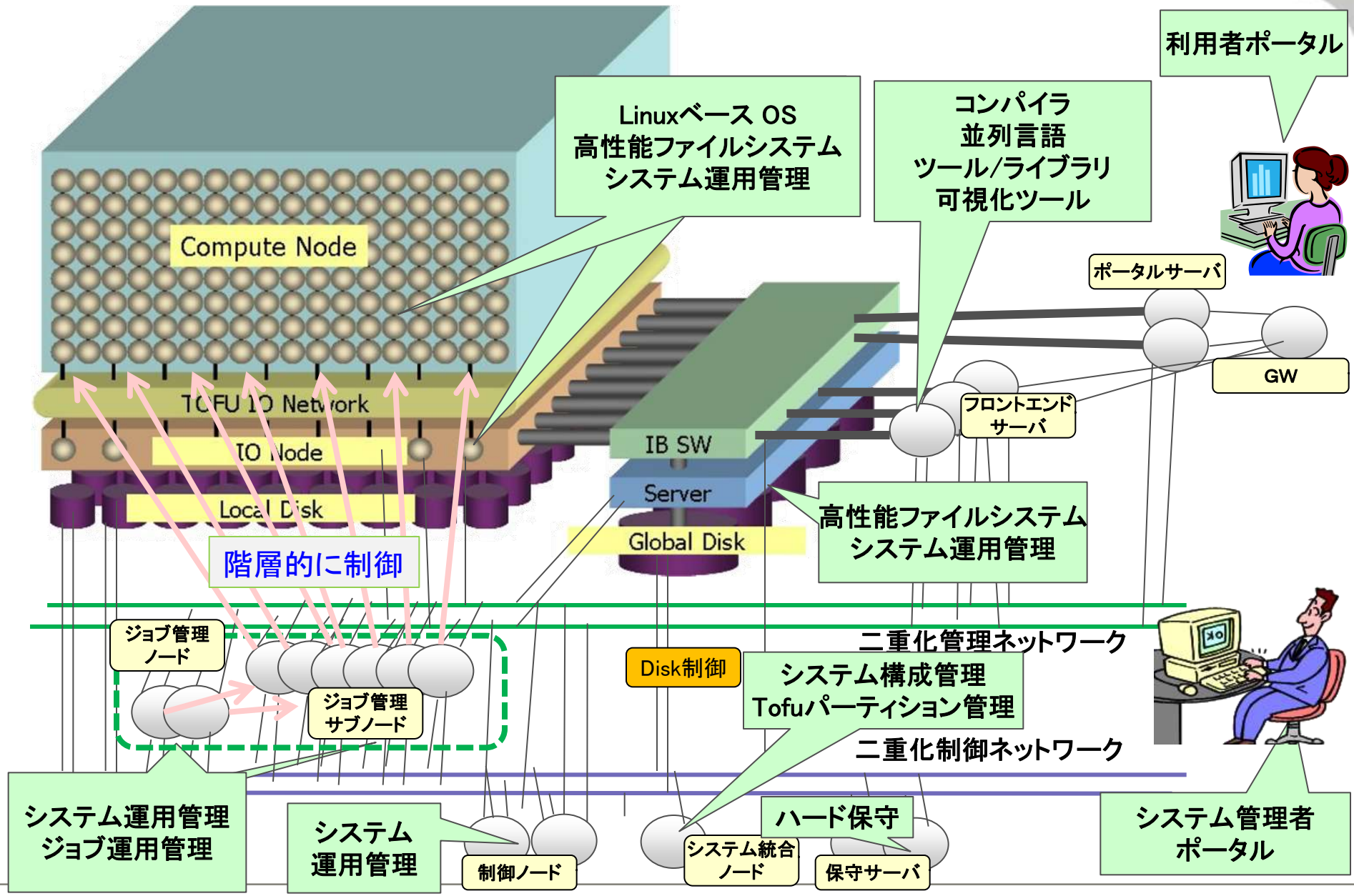
ツール/ライブラリ
- プログラミングツール
- 数学ライブラリ (SSL II/BLAS etc.)

Linux ベース OS

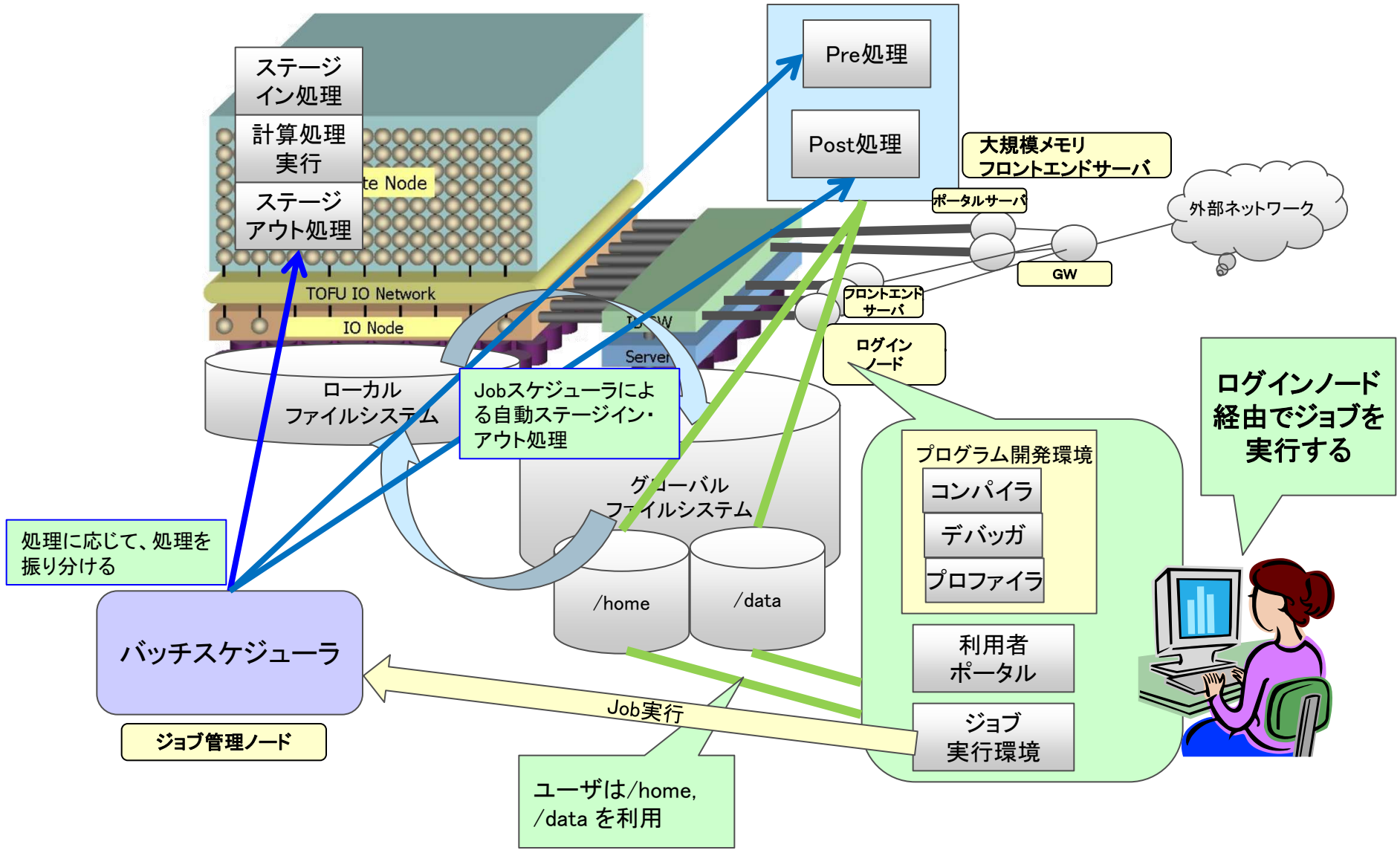
OS拡張
- 拡張ハードウェア, 高速インターコネクトサポート
- 信頼性・保守性向上
- スケーラビリティ向上(同期スケジューラ)

「京」ハードウェア

全体システム構成図と各ソフトウェアの配置



ユーザのシステム利用イメージ



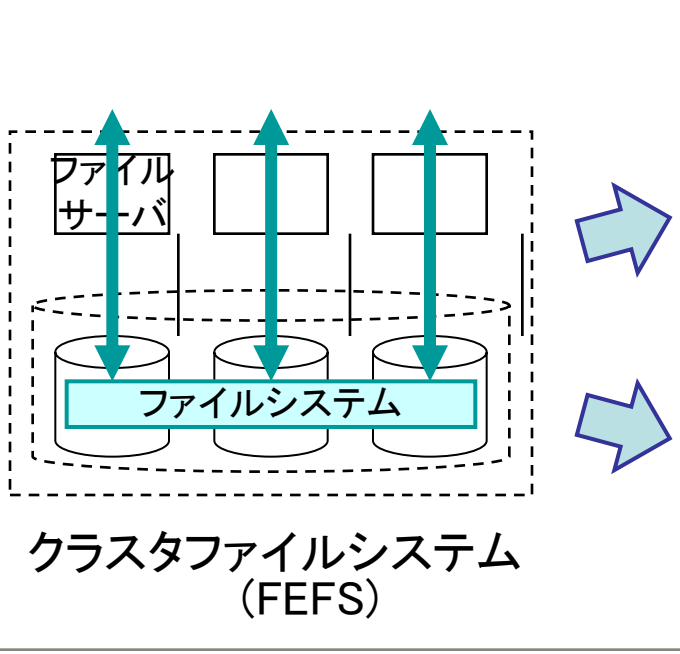
システムソフトウェアの概要

FUJITSU

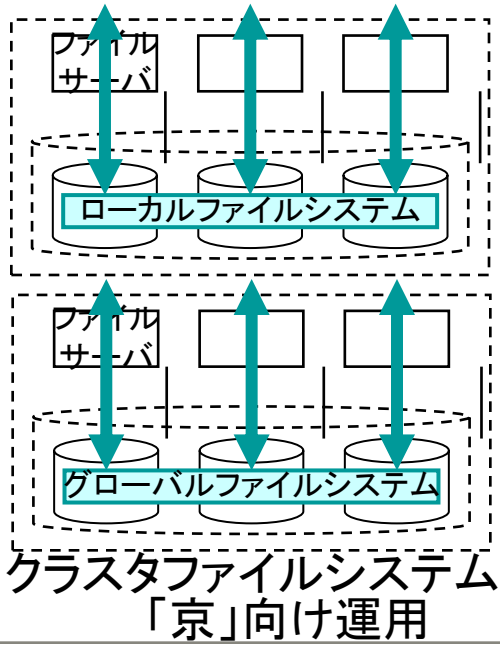
- 目標：「京」とPCクラスタ向けに統一した実行環境の提供
- OSとしてLinuxを採用し、各コンポーネントにOSSを最大限に活用： Lustreファイルシステム、Open MPIなど
  - アプリケーション移植性、オープンソースソフトウェア(OSS)の移植性を最優先に考慮
  - ただし、ノウハウが必要な運用系ソフトは独自開発
- Linuxを活用する際の課題
  - 通常のLinuxシステムは数多くの管理プロセスが存在するため、OSジッタが問題となり並列プロセス間で大きな実行バラツキを引き起こす： OSジッタ対策が必須

Lustreベースファイルシステム(FEFS)の概要

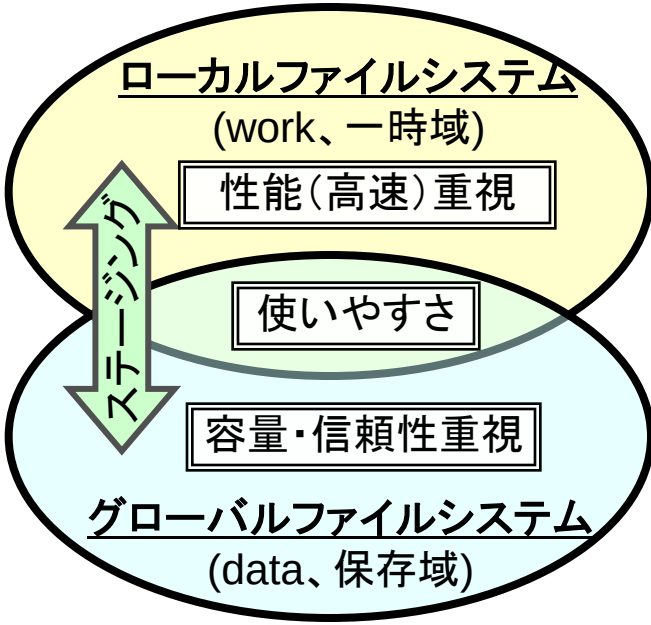
- 大規模対応のため従来の単一サーバ型でなくクラスタ型である「Lustreファイルシステム」(GPLv2)をベースに開発
  - グローバルとローカルのファイルシステムからなる運用にも対応
- 世界トップクラスに相応しい最大規模、最速IO性能が目標
  - **目標　2011年：　100PB, 1TB/s**
- Lustreコミュニティ(Open SFS)に参画し、Lustre標準化を推進
  - Open SFS:　Lustreの標準化と開発を担う非営利組織



クラスタファイルシステム
(FEFS)



クラスタファイルシステム
「京」向け運用



言語処理系の概要

- HPC向けの主要な言語と並列手法をサポート
- HPC-ACE向けの高度な命令レベル最適化、VISIMPACTを実現するループレベル最適化をサポート
- 超高並列向けデバッグ・チューニングツールをサポート
- SSL IIに加えてデファクトな数学ライブラリをサポート

*1: eXtended Parallel Fortran（分散並列Fortran言語）
*2: Rank Map Automatic Tuning Tool（ランクマッピング最適化）

FUJITSU

1.「京」の概要

- システムの概要
- ソフトウェアの概要
- システムの信頼性

システムの信頼性の向上

FUJITSU

- 実績のある高信頼化技術の適用
  - CPU命令リトライ
  - Tofuインターコネクトのリンクレベルリトライ
  - 運用ソフト（ParallelNavi）による障害ノードの自動切り離し
  - 活性保守
- 単一点故障でダウンしないシステムを目指し二重（多重）化を徹底
  - Tofuインターコネクト： 冗長経路を12経路取れるようにし、障害ノードを迂回
  - IOノード、IOパス： 2重化により、ファイルIOを確実に処理
  - 管理ノード、制御ノード、ネットワーク（管理ノード、制御ノード間）
  - サービスプロセッサ(SP)： 筐体内に2重化、障害時には交代して動作を継続
- 水冷の効果
  - LSIの動作温度を低減し、CPU/ICCの障害率を下げる

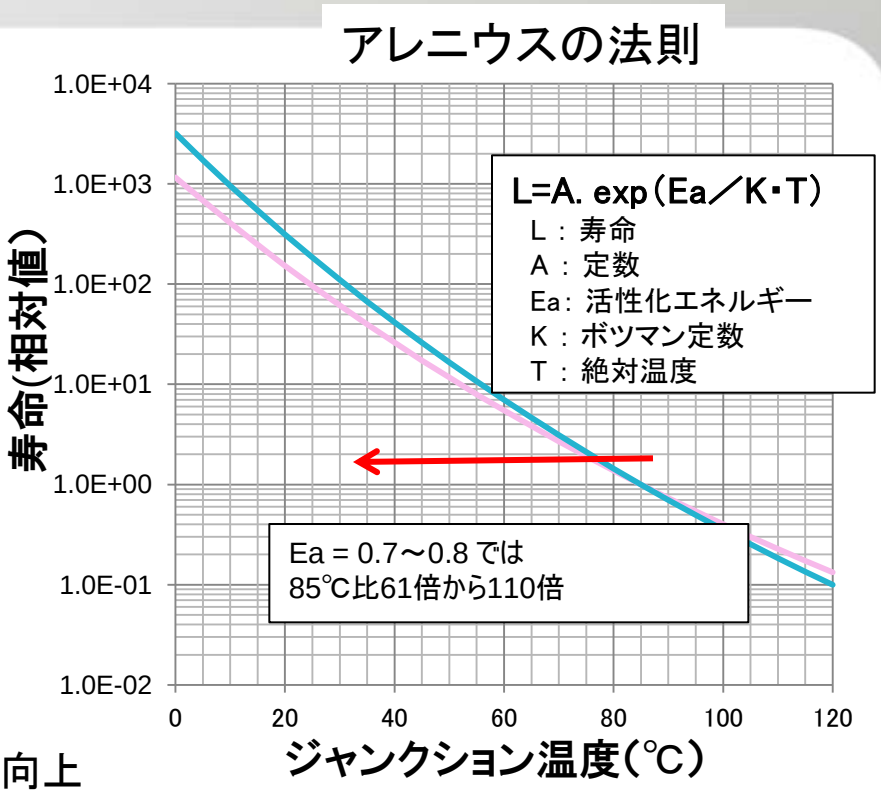
水冷の効果

■ 大規模システムの課題

■ 年間故障率（AFR）が数%でも10万ノード構成で、数時間に1回の故障

1%（100ノードで年間1回の障害）でも約9時間に1回発生

⇒実用的な連続稼働時間を確保するためには、故障率の低減が必須



■ 液冷方式の効果

■ 半導体のジャンクション温度を下げると部品寿命が向上

アレニウスの法則：温度を10度下げれば寿命は約2倍向上

■ ジャンクション温度を85℃から30℃程度に下げれば、

部品寿命は約60から100倍

⇒1万ノードを超える大規模構成における稼働時間の確保に貢献

FUJITSU

2. 余談

開発プロジェクトの歴史

- 2005年春：文科省、要素技術開発プロジェクト開始
- 2005年夏：文科省、次世代スパコン開発プロジェクト了承
- 2006年春：開発主体を理研として次世代スパコン開発プロジェクトを開始
- 2006年秋：概念設計、富士通とNEC日立連合が参加
- 2007年初：富士通案(スカラ)とNEC日立連合案(ベクトル)を併用の方針
- 2007年3月：施設立地点を神戸市ポートアイランドに決定
- 2007年9月：スカラ＋ベクトルの複合計算機構成と決定
- 2008年4月：建屋着工

開発プロジェクトの歴史

- 2008年　　：粛々と開発を実施
- 2009年1月：CPU初版をテープアウト
- 2009年5月：CPU初版PON、川崎工場で試験開始
  NEC日立連合がプロジェクト離脱
- 2009年9月：筐体PON、沼津工場で試験開始
- 2009年11月：事業仕分け、「予算計上見送りに近い縮減」
- 2010年9月：出荷開始
- 2011年3月：震災影響により出荷中断
- 2011年6月：TOP500で一位獲得
- 2012年6月：システム完成予定
- 2012年11月：共用運用開始予定

TOP500とは

FUJITSU

世界で最も速いスパコン上位500システムランキング

- 1993年に発足
- LINPACKベンチマークの結果に基づいてランキング
- 年2回（6月、11月）公表

LINPACKベンチマーク

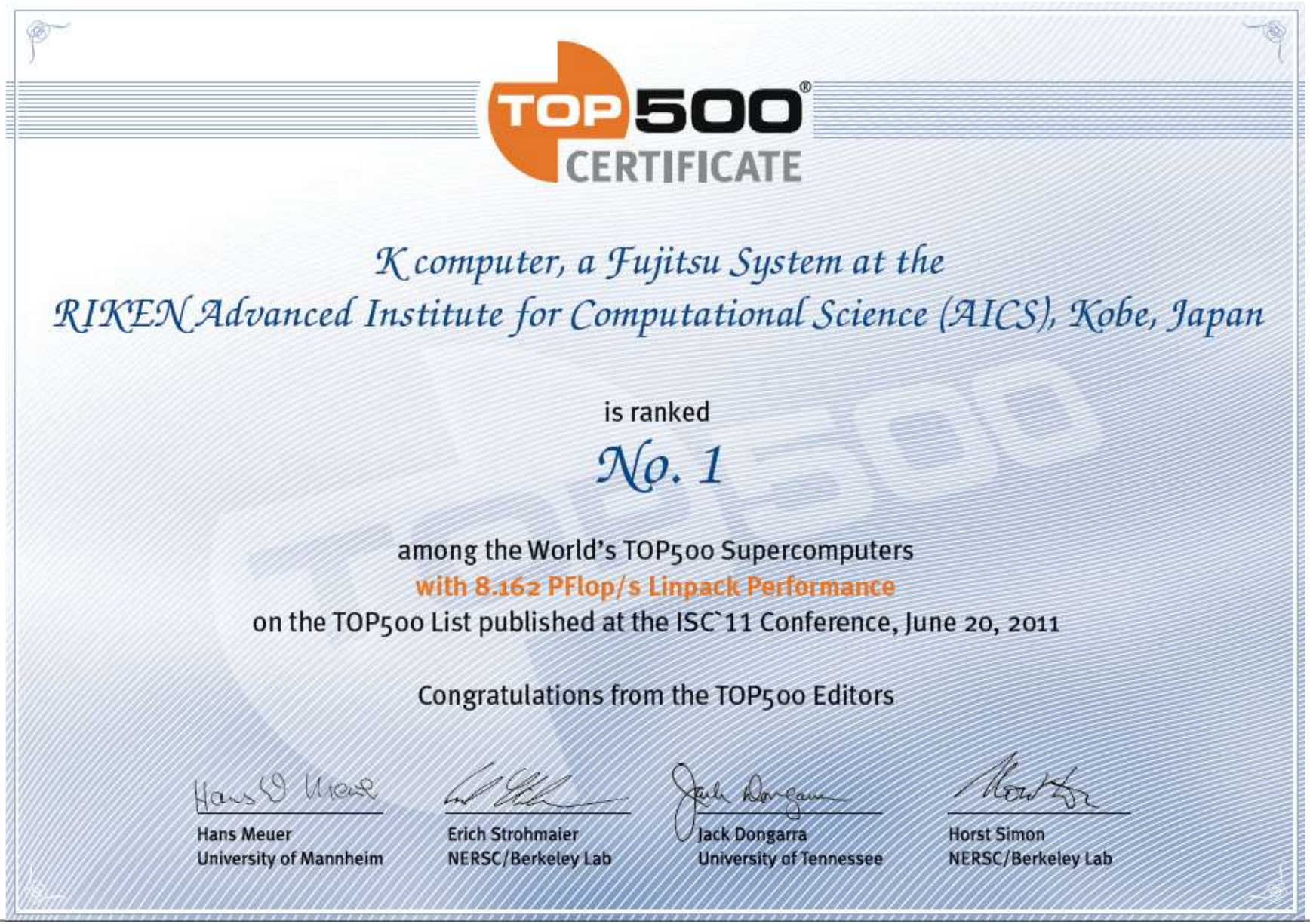
理学・工学で一般的な連立一次方程式をLU分解法で解く速度を測定し、システムの浮動小数点演算性能を評価

最近の動向

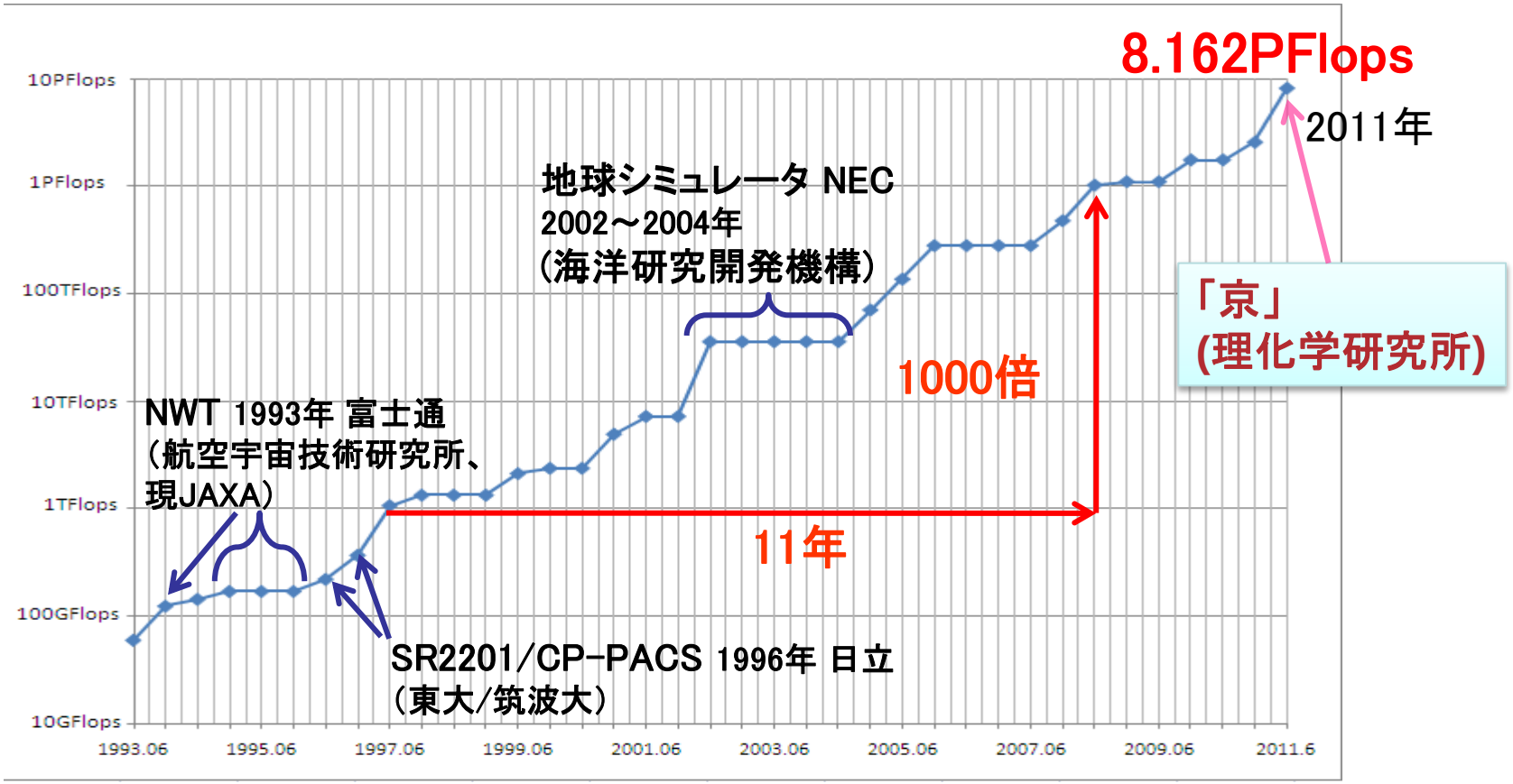
- Intel, AMDなどのx86系プロセッサを利用したシステムが大半を占める
- 上位にはGPUを用いたシステムが多数
- 近年中国などアジアのシステムが増加傾向

GPU・・・3Dグラフィックスの表示に必要な計算処理を行う半導体チップ

Certificate 2011/6/20



TOP500歴代実行性能1位



2011年6月、「京」が世界第一位を獲得
地球シミュレータ以来7年ぶりの国産スパコン快挙

*NWT：Numerical Wind Tunnel(数値風洞システム)

TOP 500　BEST10

順位	サイト名（国名）	システム名	開発担当	ﾌﾟﾛｾｯｻｱｰｷﾃｸﾁｬ	実行性能 (PFlops)
1位	RIKEN AICS（日本）	K computer	Fujitsu	Sparc	8.162
2位	NSCT（中国　天津）	Tianhe-1A	NUDT (国防科学技術大学)	Intel EM64T	2.566
3位	ORNL（米国）	Jaguar	Cray	AMD x86_64	1.759
4位	NSCS（中国　深圳）	Nebulae	Dawning	Intel EM64T	1.271
5位	Tokyo Tech（日本）	TSUBAME-2	NEC/HP	Intel EM64T	1.192
6位	LANL/SNL（米国）	Cielo	Cray	AMD x86_64	1.110
7位	NASA Ames（米国）	Pleiades	SGI	Intel EM64T	1.088
8位	LBNL/NERSC（米国）	Hopper	Cray	AMD x86_64	1.054
9位	CEA（フランス）	Tera-100	Bull	Intel EM64T	1.050
10位	LANL（米国）	Roadrunner	IBM	Power(cell)	1.042

圧倒的実行性能
8.162PFlops
2～6位の合計値
（7.898PFlops）を上回る

性能測定

FUJITSU

- Linpackの測定は一発勝負ではありません
  - 3000ノード超の並列化は未経験の世界
  - 理屈の上では動く筈だが実証しない限り不確実
  - 段階的に規模を拡大して実施
    - 2010年10月　:408ノード、48TFLOPS
    - 2011年1月　:9744ノード、1.1PFLOPS
    - 2011年3月　:27648ノード、3.2PFLOPS
    - 2011年4月　:48960ノード、5.7PFLOPS
    - 2011年5月　:58752ノード、6.8PFLOPS
    - Top500登録値:68544ノード、8.1PFLOPS

記者会見

- 日本時間6月20日17時頃にハンブルグでTOP500発表
- 同時に理研・富士通でプレスリリース
- 日本では6月20日19時から記者会見を実施
- 狭い部屋に大勢プレスが来た上ジャケット着用で大変暑かった
- たった2時間の間に蓮舫大臣コメントを取って来たのは驚いた

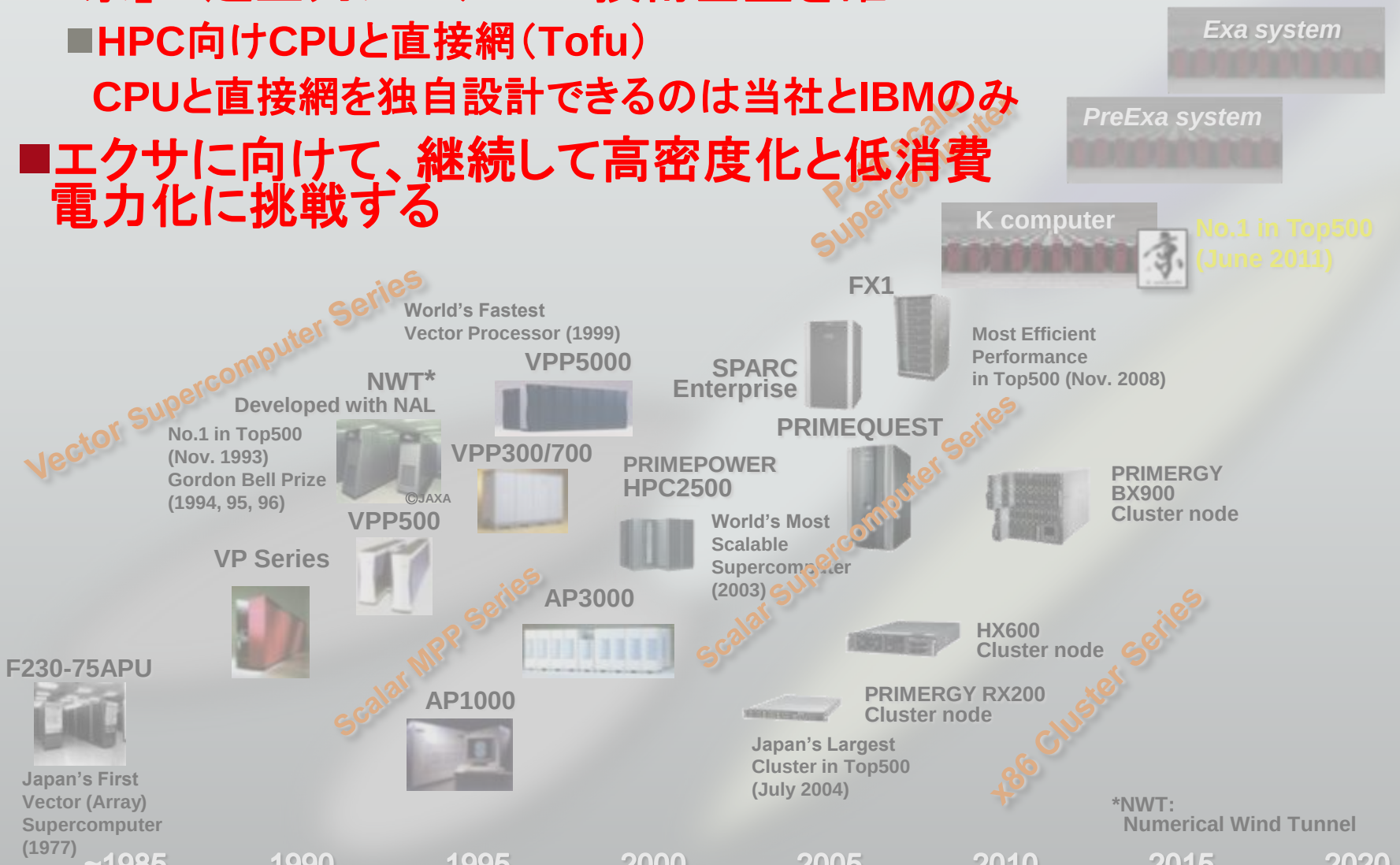
FUJITSU

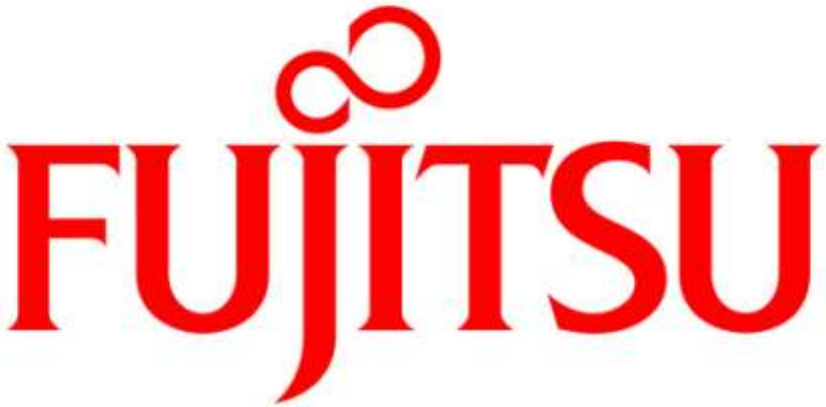
4. まとめ

まとめ

FUJITSU

- 「京」で超並列システムの技術基盤を確立
  - HPC向けCPUと直接網（Tofu）
    CPUと直接網を独自設計できるのは当社とIBMのみ
- エクサに向けて、継続して高密度化と低消費電力化に挑戦する





shaping tomorrow with you