AICS

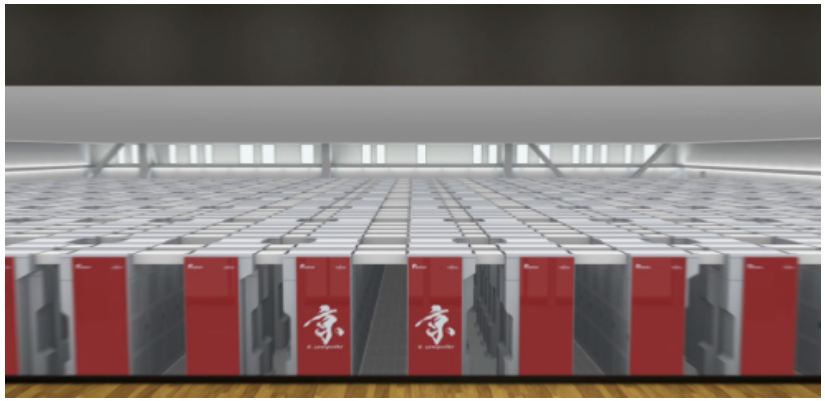
京速コンピュータ「京」

黒川 原佳 < motoyosi@riken.jp >

独立行政法人理化学研究所
次世代スーパーコンピュータ開発実施本部
計算科学研究機構

講演のアウトライン

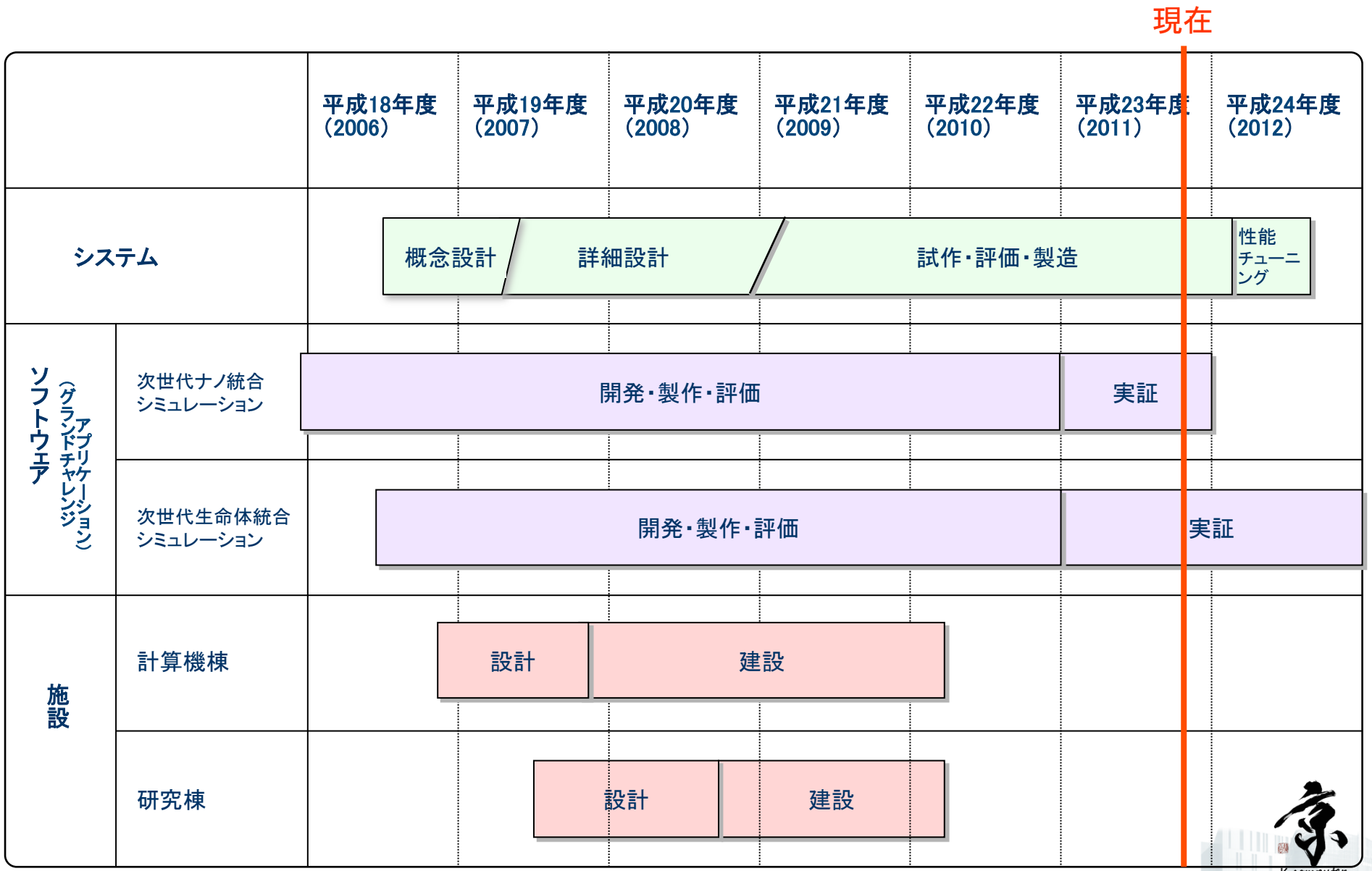
- 開発プロジェクトについて
- 「京」の概要
- 「京」を構成する要素技術
- まとめ



開発目標

- LINPACK性能で10PFLOPSを達成する
- 実アプリケーション性能で数PFLOPSを達成する
- 高性能かつ低エネルギー消費のシステムであること
- 高信頼性・高可用性のシステムであること
- ユーザが使いやすいように、運用出来るシステムであること

開発スケジュール


現在
平成18年度（2006）
平成19年度（2007）
平成20年度（2008）
平成21年度（2009）
平成22年度（2010）
平成23年度（2011）
平成24年度（2012）
システム
概念設計
詳細設計
試作・評価・製造
性能チューニング
ソフトウェア（アプリケーション・グランドチャレンジ）
次世代ナノ統合シミュレーション
開発・製作・評価
実証
次世代生命体統合シミュレーション
開発・製作・評価
実証
施設
計算機棟
設計
建設
研究棟
設計
建設

次世代スパコンの愛称とロゴマークについて

http://www.nsc.riken.jp/aisho/kekkahoukoku.htmlより

■選定結果

愛称： 「京」（けい）

英語表記 “K computer”

※今後、理研は次世代スパコンを京速コンピュータ「京」と呼びます。

■選考理由

- 「京」は、開発目標性能の10ペタを表す万進法（注）の単位で、ほかの科学技術分野の名称・愛称に多く用いられるものと重複がなく、「単位」で表現するという新規性を評価。
- 漢字一文字とシンプルで分かりやすく、ロゴデザインもしやすい。
- 外国人も発音しやすい。
- 日本人は漢字に意味を込め、語呂や響きを大切にする。日本的な漢字と数学（単位）が重なっている点がよい。
- 歴史と未来が同居する言葉。「京」はもともと大きな門を表し、「計算科学の新たな門」にもつながる。
- 次世代スパコンは、すでに計算科学のコミュニティでは、京速（けいそく）のコンピュータとして「京（けい）」にはなじみがある。

注）万進法
4桁（万倍）ごとに単位を区切る方式。

京
K computer

TOP500

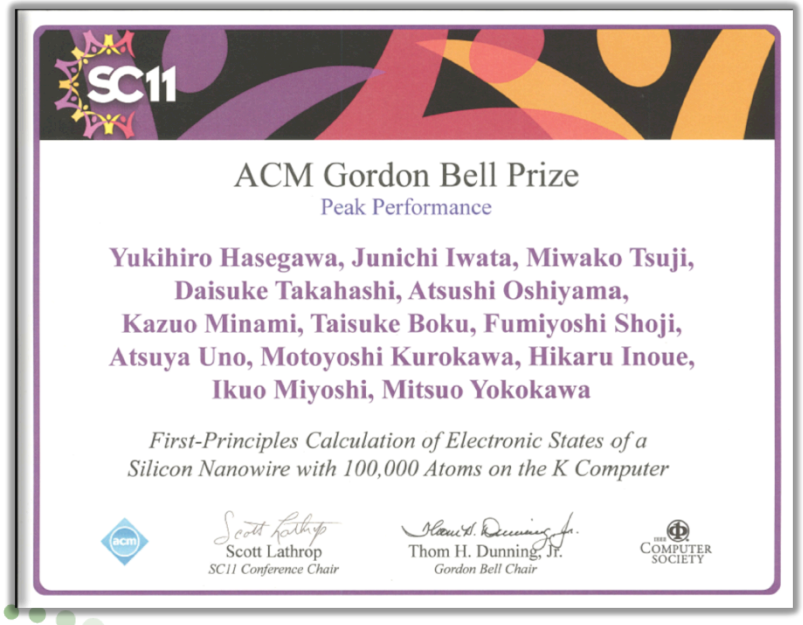
- 「京」はLINPACK 10PFLOPSを達成

	K computer（2011年11月）	K computer（2011年6月）
R_{max}	10.51PFLOPS	8.162PFLOPS
R_{peak}	11.28PFLOPS	8.774PFLOPS
efficiency	93.2%	93.0%
N_{max}	11,870,208	10,725,120
Execution time	29h28m	27h59m

- TOP500 listでは2度のNo.1を獲得
  - 第38回（2011年11月）
  - 第37回（2011年6月）
- 「京」の設置は2011年8月末で終了
  - 現在はシステムソフトウェアの開発チューニングに注力

ゴードン・ベル賞（最高性能賞）

- "First-principles calculations of electron states of a silicon nanowire with 100,000 atoms on the K computer
  - 107,292個のシリコン原子からなるナノワイヤ（直径20ナノメートル、長さ6ナノメートル）に対する、量子論に基づく電子状態計算（世界初）を行い、実効性能 3.08ペタフロップス（実行効率 約43.6%）を達成
    - 電界効果トランジスタの電子チャネルの状態計算
    - 約10,000個から40,000個のシリコン原子からなる、円、楕円、ダンベルなどさまざまな断面形状を持つ長さ約10ナノメートルのナノワイヤの電子状態を解明し、電子輸送の特性がナノスケールの形状に大きく依存することを初めて明らかにしました。


SC11

ACM Gordon Bell Prize

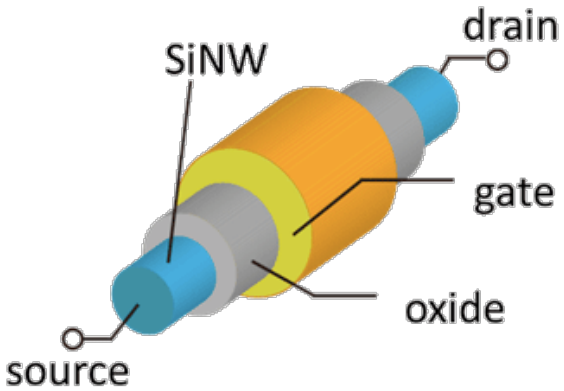
Peak Performance

Yukihiro Hasegawa, Junichi Iwata, Miwako Tsuji,
Daisuke Takahashi, Atsushi Oshiyama,
Kazuo Minami, Taisuke Boku, Fumiyoshi Shoji,
Atsuya Uno, Motoyoshi Kurokawa, Hikaru Inoue,
Ikuo Miyoshi, Mitsuo Yokokawa

First-Principles Calculation of Electronic States of a Silicon Nanowire with 100,000 Atoms on the K Computer

Scott Lathrop
SC11 Conference Chair

Thom H. Dunning, Jr.
Gordon Bell Chair

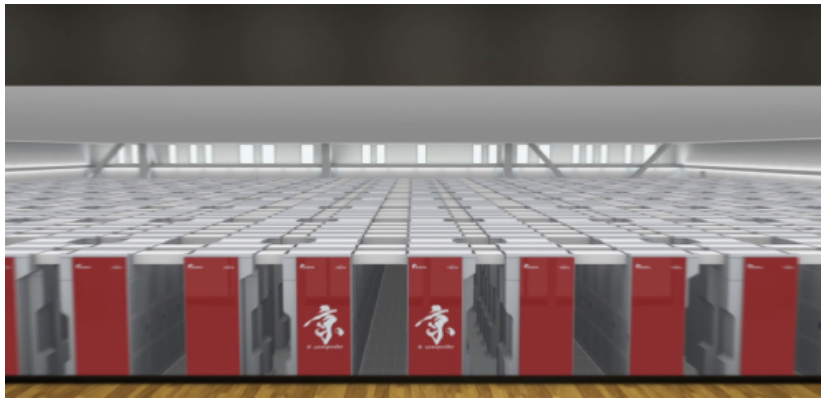


HPC Challenge Award (四冠)

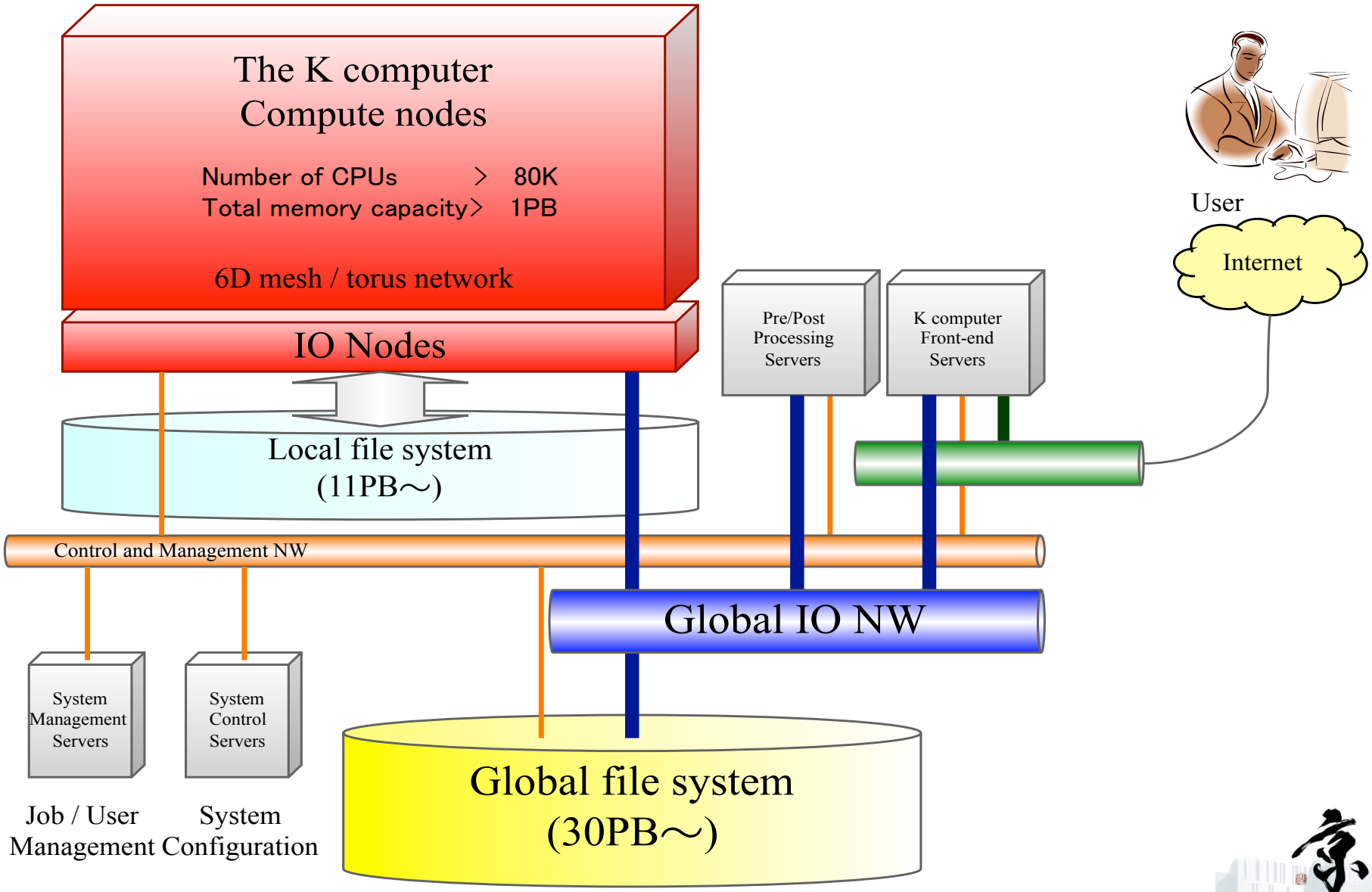
- HPCチャレンジベンチマークは、科学技術計算で多用される計算パターンから抽出した28項目の処理性能によって、スパコンの総合的な性能を評価するベンチマークプログラム
  - このうち特に重要なベンチマークが以下の4つ:
    - Global HPL (大規模な連立1次方程式の求解における演算性能)
    - Global RandomAccess (並列プロセス間でのランダムメモリアクセスの性能)
    - EP STREAM (Triad) per system (多重負荷時のメモリアクセス性能)
    - Global FFT (高速フーリエ変換の性能).
  - The HPC Challenge Class 1 Awardsは上の4つのベンチマークの1位が表彰対象
  - 四冠はBlueGene/L以来

講演のアウトライン

- 開発プロジェクトについて
- 「京」の概要
- 「京」を構成する要素技術
- まとめ



「京」概要

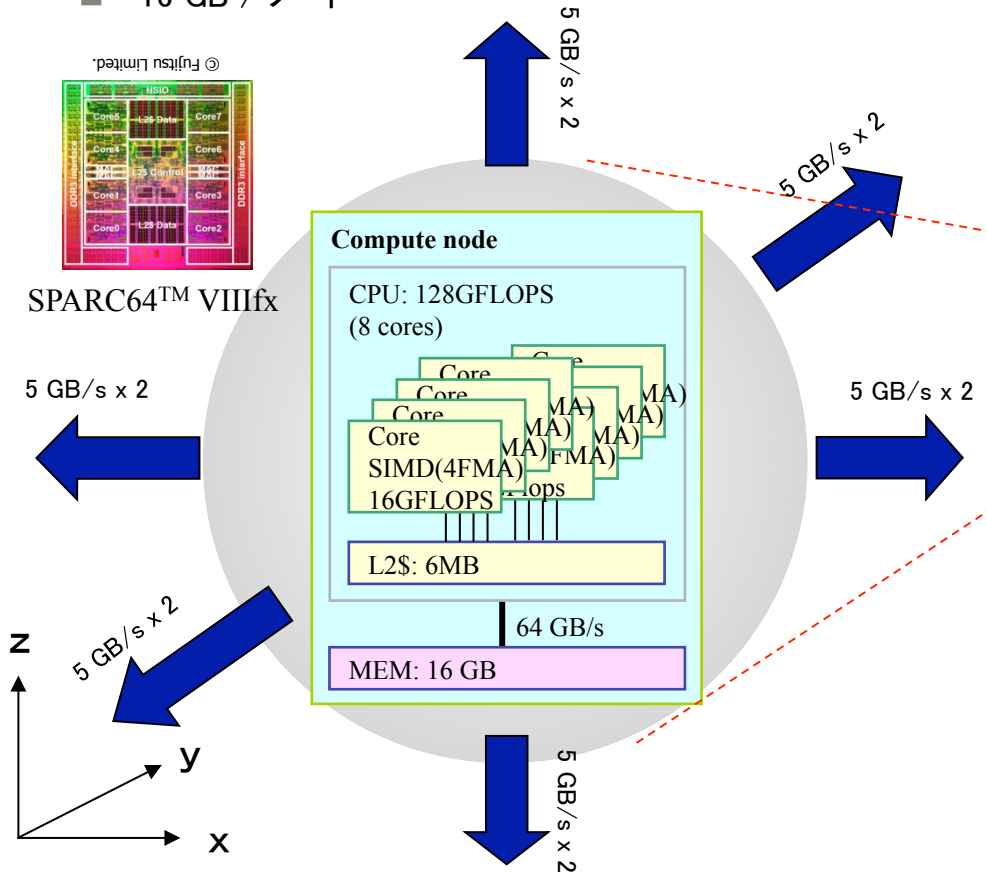

The K computer
Compute nodes
Number of CPUs > 80K
Total memory capacity> 1PB
6D mesh / torus network
IO Nodes
Local file system
(11PB～)
Control and Management NW
Global IO NW
Global file system
(30PB～)
Pre/Post
Processing
Servers
K computer
Front-end
Servers
User
Internet
System
Management
Servers
System
Control
Servers
Job / User
Management
System
Configuration

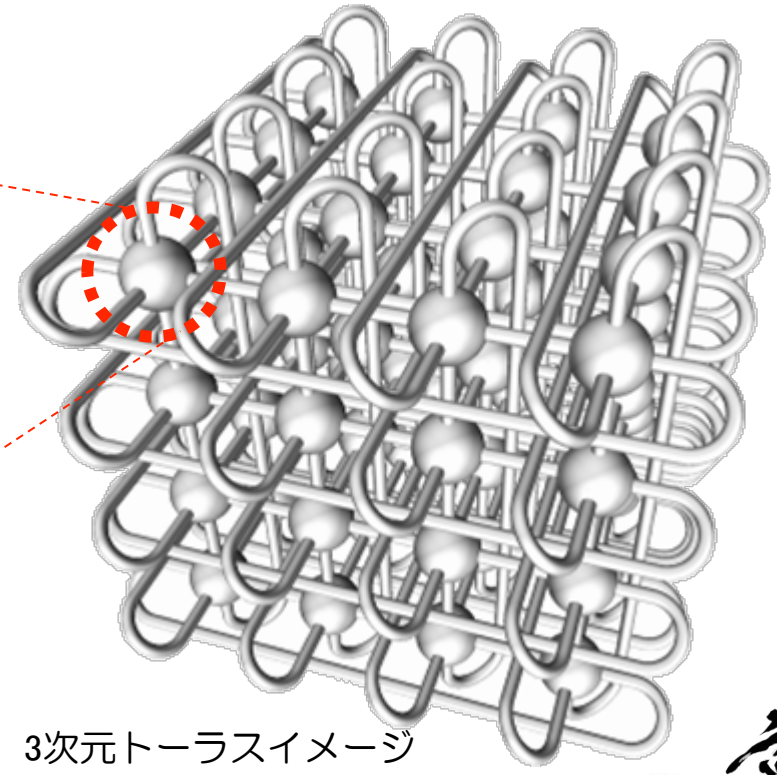
計算ノードとインターコネクトネットワーク構成

- 計算ノード (CPU) : 82, 944
  - IOノード:5, 184
  - コア数 : 663,552 (705,024)
- 理論ピーク性能 : 10.6 (11.3) PFLOPS
- 総メモリ量 : 1.27(1.34)PiB
  - 16 GB / ノード

- ネットワーク : Tofu インターコネクト
  (6D メッシュ / トーラス ネットワーク)
  - 10 リンク (5 GB/s x 2) / リンク
  - 論理的な 3D, 2D or 1D トーラスネットワークをユーザジョブに提供
  - ケーブル:約20万本、総延長約1000km

SPARC64™ VIIIfx



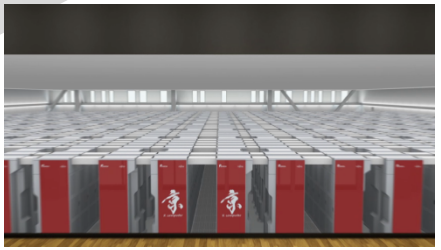


3次元トーラスイメージ

© Fujitsu Limited.

「京」のシステムスケール

40m × 40m
Full System
Compute Rack × 864



10.6(11.3)PFLOPS
1.27(1.34)PiB

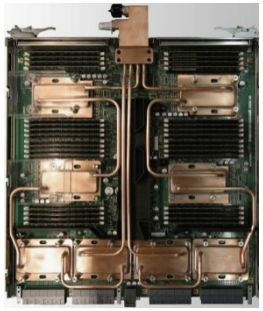
400cm × 80cm
2 Cabinet
Compute Rack × 4
Disk Racks × 1

49.2(52.4)TFLOPS
6.00(6.38)TiB

80cm × 80cm
Compute Rack
SB×24
IOSB×6

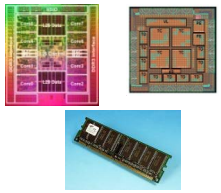
12.3(13.1)TFLOPS
1.50(1.59)TiB

50cm × 50cm
System Board(SB)
Node×4

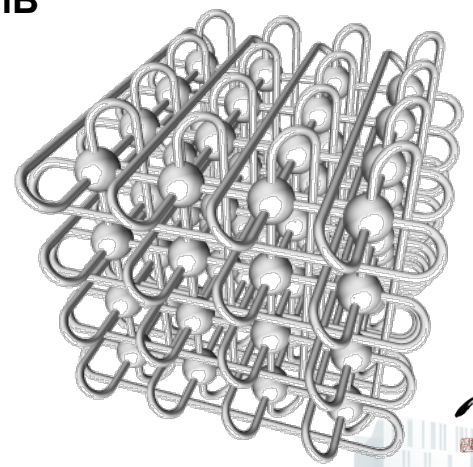


512GFLOPS
64GiB

Node
CPU×1
ICC×1
memory



128GFLOPS
16GiB



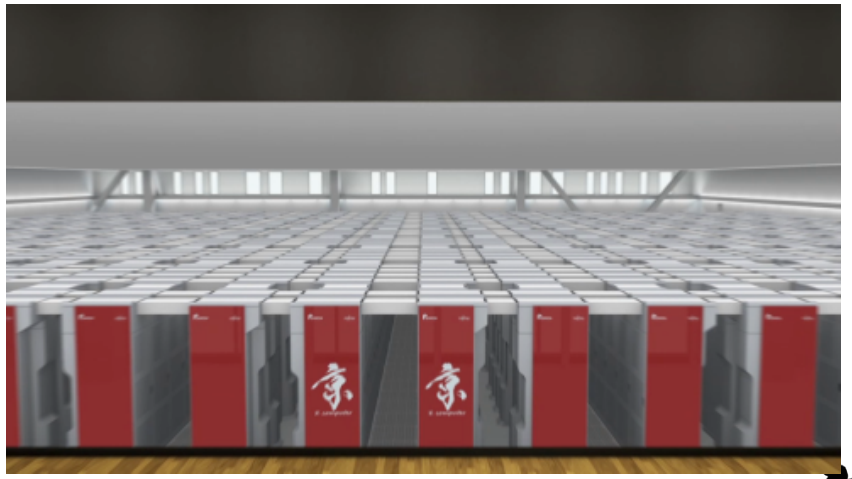
() 内はIOノードを含めた数値

「京」のソフトウェア環境

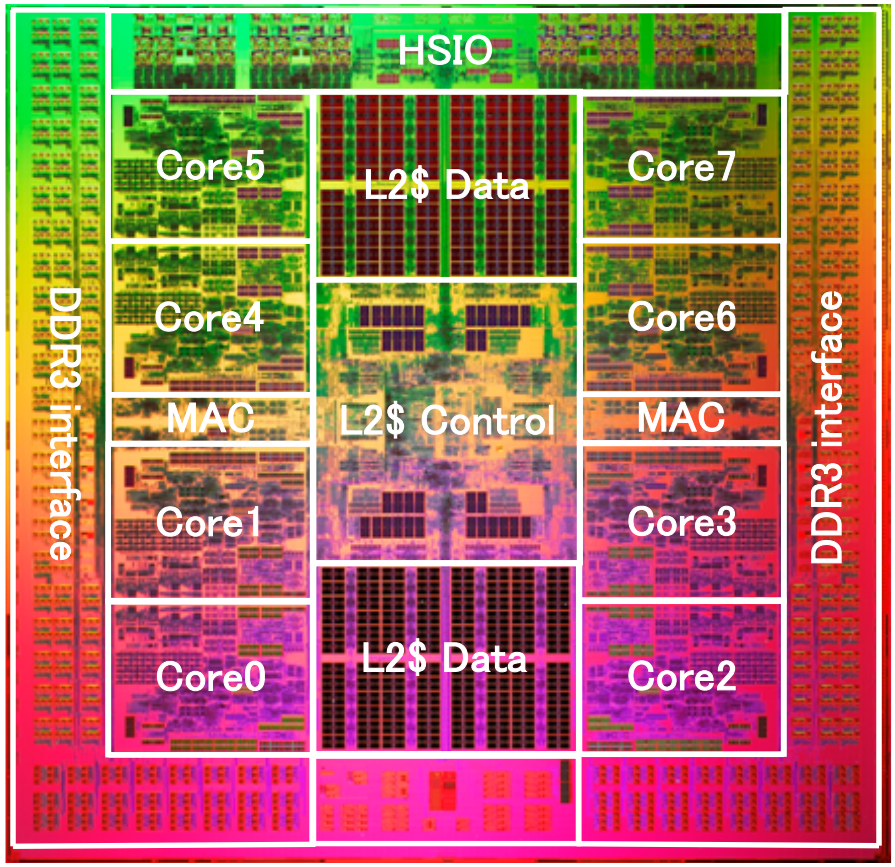
- システムソフトウェア
  - OS : Linuxベース(2.6.38.8) OS （計算ノードおよびIOノード）
  - ファイルシステム : Lustre(1.8.5)ベースのFEFS (Fujitsu Exabyte File System)
  - システムマネジメント、JOBマネジメントなどなど
- プログラミング環境
  - プログラム言語・コンパイラ
    - Fortran 95とFortra2003のサブセットに準拠したFortran、XPFortran（富士通のPGAS言語）
    - C99 and C++ 2003 に準拠したC/C++
      - いくつかのGNU C/C++拡張仕様も取り込み
    - OpenMP 3.0 をサポート
  - 数値演算ライブラリ
    - BLAS, LAPACK, FFTW, SSL II(富士通製科学技術計算ライブラリ), double-double datatype (4倍精度)演算ライブラリ
  - MPIライブラリ
    - OpenMPI 1.4.1 (MPI 2.1準拠)
    - 低遅延で高スループットを実現
      - OpenMPIのソフトウェアスタックをバイパス
    - Tofuネットワークに最適化した集団通信を提供
      - Broadcast, Gather, Reduce

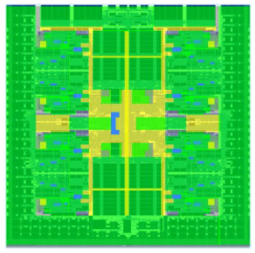
「京」を構成する要素技術

- CPU
  - 低消費電力かつ高性能
  - 高い信頼性と可用性
- インターコネクトネットワーク
  - 高い拡張性と広帯域、低遅延
  - 高い信頼性と可用性
- システムソフトウェア＆MPIライブラリ
  - 低ノイズ環境提供
  - 柔軟なジョブマネジメント
  - 高性能MPIライブラリ
- 開発環境/可視化ライブラリ
  - デバッガ/プロファイラ
  - 可視化ライブラリ
- ファイルシステム
  - 高い拡張性と高スループット
  - センター利用と柔軟性
- 施設
  - 電源施設
  - 冷却施設



CPU（SPARC64™ VIIIfx）



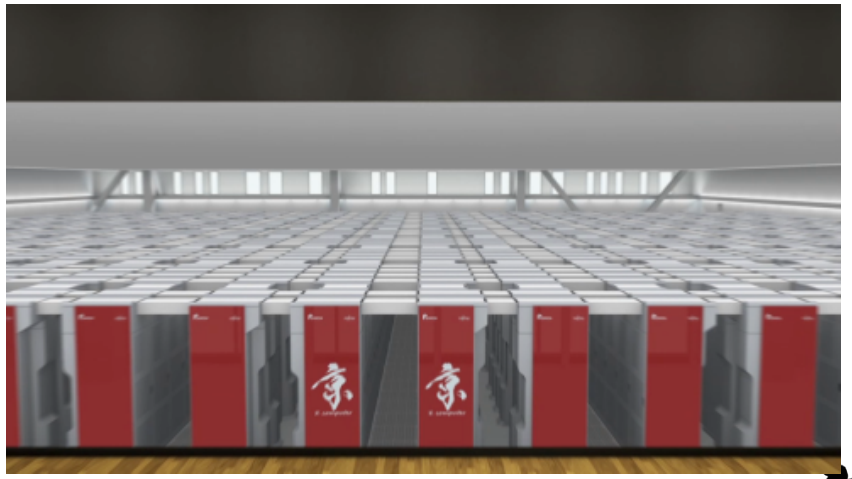


check and correct by hardware
check by hardware

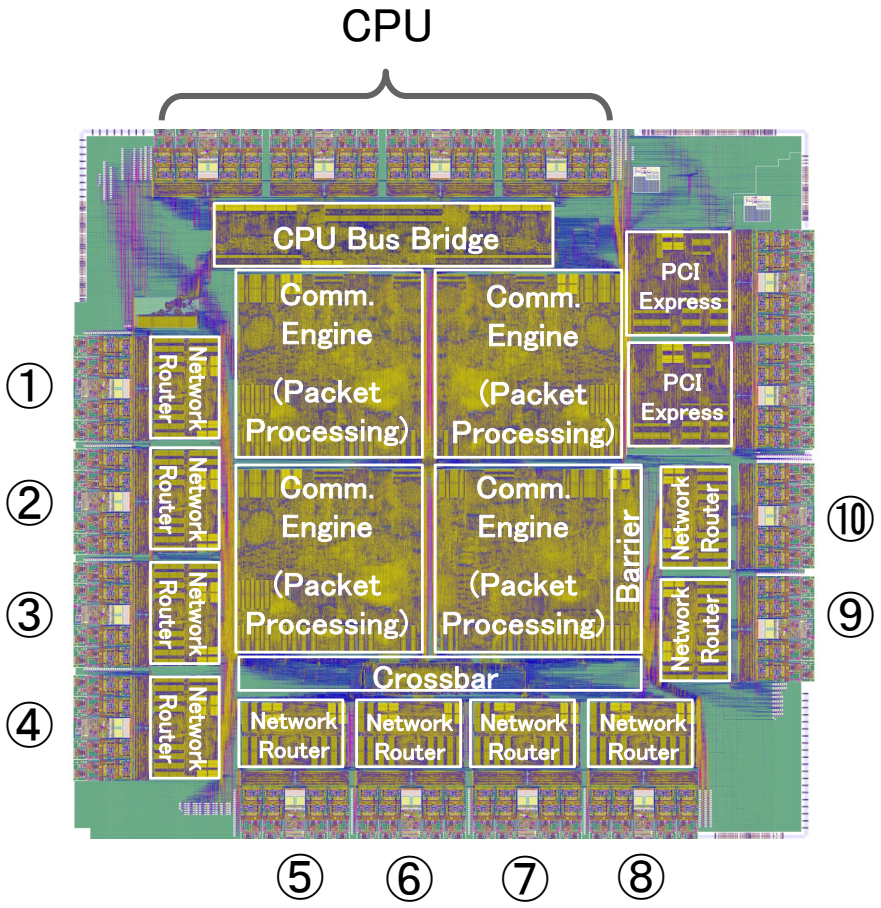
- 基本仕様
  - 8 コア, 6 MB共有 L2 キャッシュ
    - FMA×4（2 SIMD）
    - セクターキャッシュ（ソフトウェアによる置き換え制御）
    - 32KB (I)/ 32KB (D) L1 キャッシュ / コア
    - 256 (64bit) DPレジスタ. 188 (64bit) GPレジスタ
  - 8ch DDR3 DIMM MAC
  - 周波数 2 GHz
  - HPC拡張のISA (HPC-ACE)
  - 消費電力 58W（Tj=30℃）
- 45nm CMOS
  - 22.7mm x 22.6mm
  - 760M トランジスタ
  - 1271 信号ピン
- 理論ピーク性能
  - PF 演算性能 128GFLOPS
  - メモリ帯域 64GB/s
- 信頼性と可用性
  - 水冷
    - リーク電流削減と信頼性の向上
      - Tj=85℃に比べて60 ～ 100 倍高寿命（アレニウスの法則）
  - RAS
    - エラーチェックとエラー訂正
    - ハードウェアによる自動命令リトライ
      - 間欠エラー対策

「京」を構成する要素技術

- CPU
  - 低消費電力かつ高性能
  - 高い信頼性と可用性
- インターコネクトネットワーク
  - 高い拡張性と広帯域、低遅延
  - 高い信頼性と可用性
- システムソフトウェア＆MPIライブラリ
  - 低ノイズ環境提供
  - 柔軟なジョブマネジメント
  - 高性能MPIライブラリ
- 開発環境/可視化ライブラリ
  - デバッガ/プロファイラ
  - 可視化ライブラリ
- ファイルシステム
  - 高い拡張性と高スループット
  - センター利用と柔軟性
- 施設
  - 電源施設
  - 冷却施設



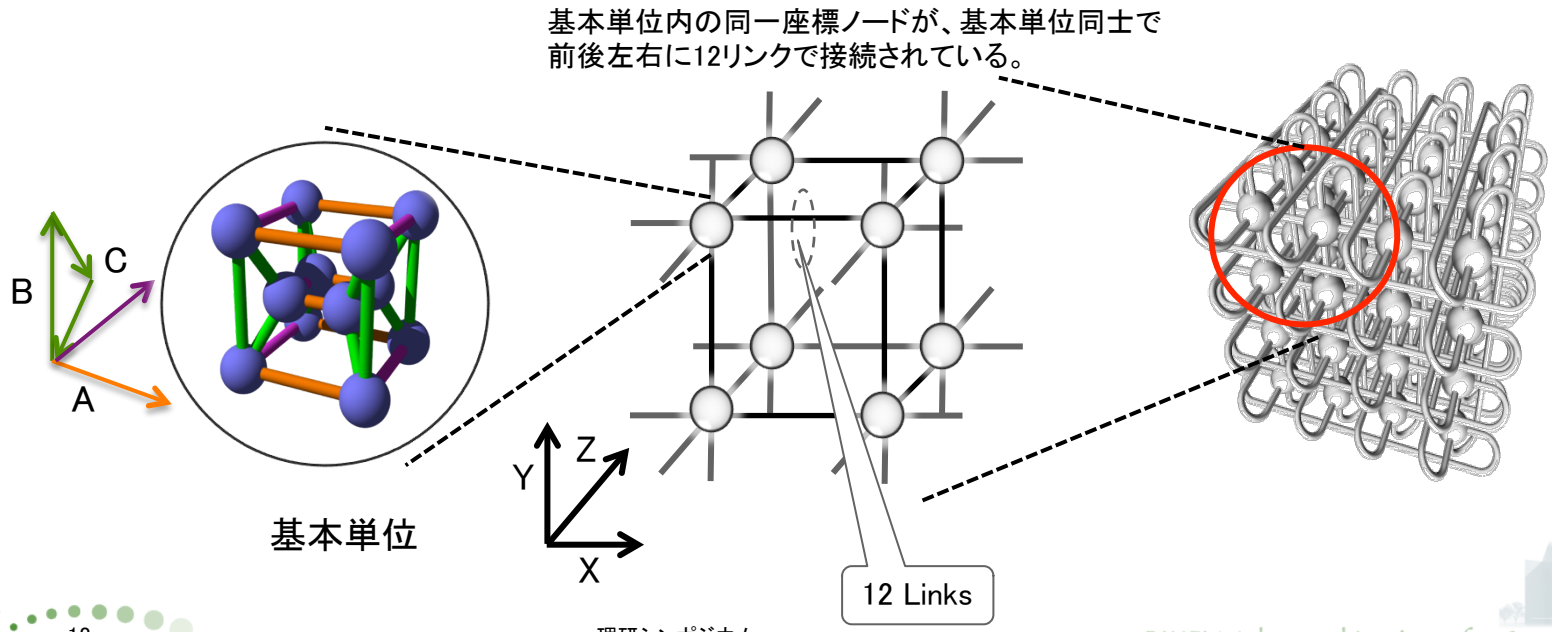
InterConnect Controller（ICC）



- **基本仕様**
  - RDMA通信エンジン×4
  - ハードウェアBarrier + Reduction
  - ネットワーク次元 :10
  - PCI Express搭載
  - 動作周波数 312.5 MHz
- **65nm ASIC**
  - 18.2mm x 18.1mm
  - ゲート数 48M
  - SRAM 12Mbit
  - HSIO数 128 lanes
- **ピークバンド幅**
  - リンクバンド幅 5GB/s x 2
  - スイッチング容量
    - 外部100GB/s（内部・外部含めて140GB/s）
- **低遅延**
  - Virtual Cut-Through転送 ~100ns

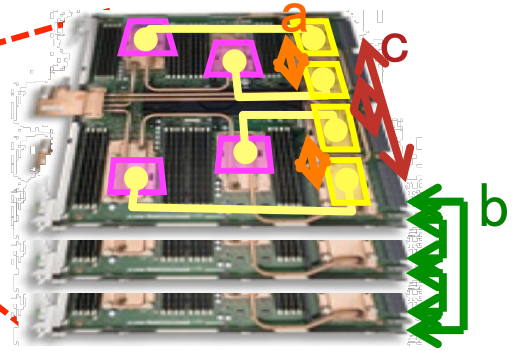
Tofuインターコネクト

- 高機能、広帯域と高い耐故障性を実現
- Torus Fusion : 3D トーラス + 3D メッシュ
  - ネットワークトポロジ : 6次元 メッシュ / トーラス ネットワーク
  - 軸 : X, Y, Z, a, b, c
    - X,Z,b : トーラス (Z=0: IOノード), Y,a,c : メッシュ
  - 軸サイズ : 24, 18, 17, 2 ,3, 2
  - 高可用性 : XYZabc軸による拡張次元ルーティングにより故障ノードを回避



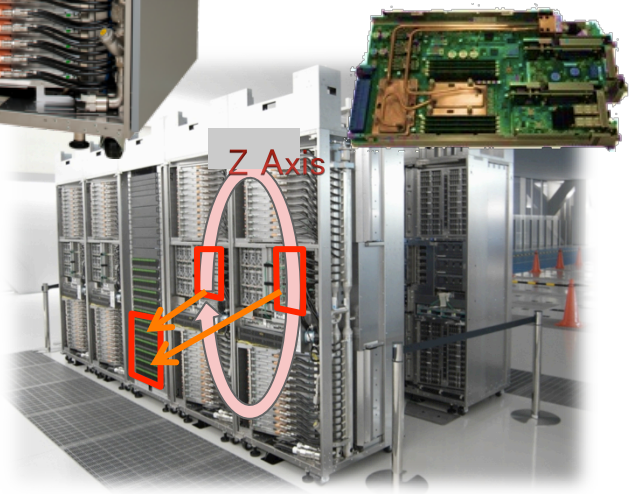
「京」の物理配置

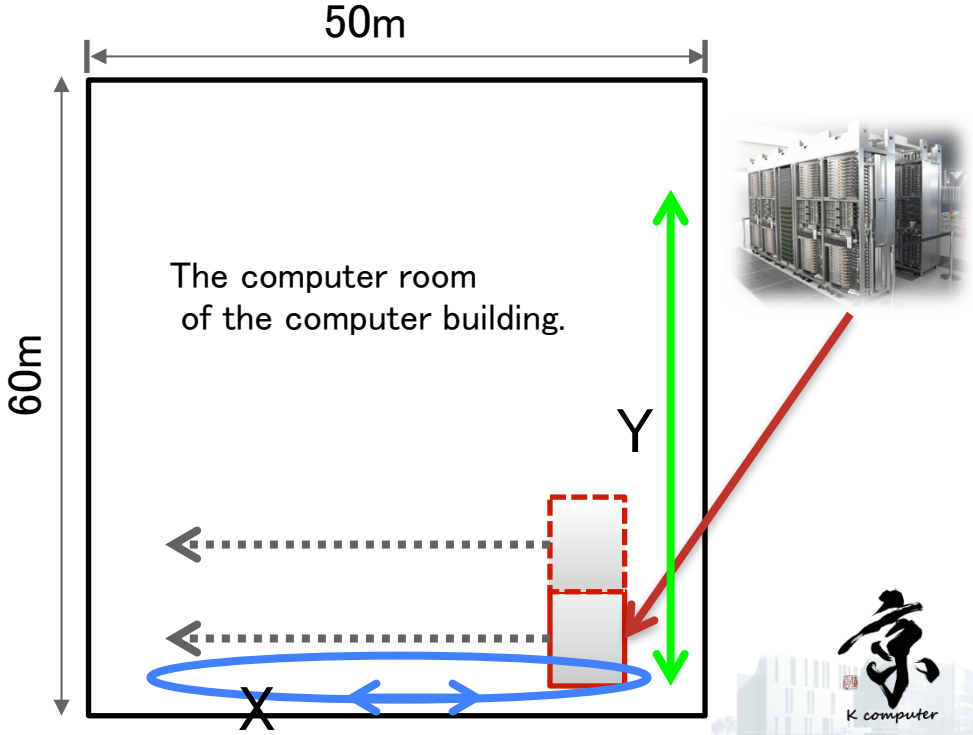
12 SB
4 Basic Unit



柱が無い空間
設置およびケーブリングの
自由度が非常に高い

12 SB
4 Basic Unit

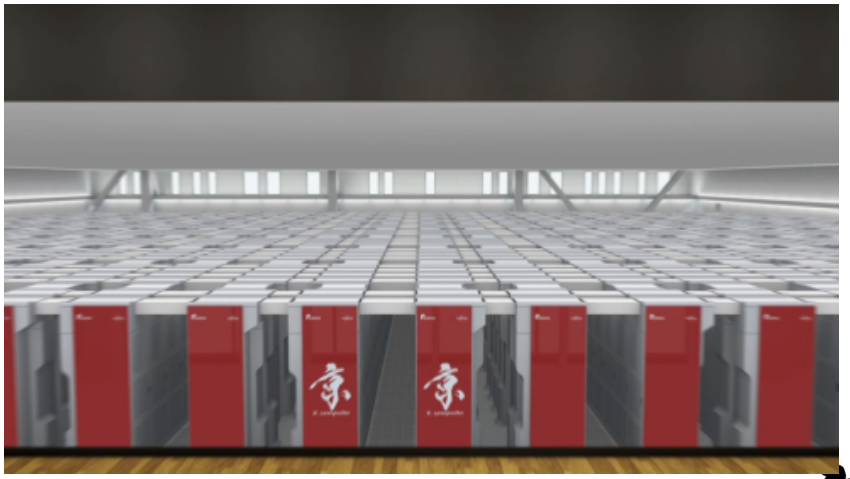




「京」の構築の様子

「京」を構成する要素技術

- CPU
  - 低消費電力かつ高性能
  - 高い信頼性と可用性
- インターコネクトネットワーク
  - 高い拡張性と広帯域、低遅延
  - 高い信頼性と可用性
- システムソフトウェア＆MPIライブラリ
  - 低ノイズ環境提供
  - 柔軟なジョブマネジメント
  - 高性能MPIライブラリ
- 開発環境/可視化ライブラリ
  - デバッガ/プロファイラ
  - 可視化ライブラリ
- ファイルシステム
  - 高い拡張性と高スループット
  - センター利用と柔軟性
- 施設
  - 電源施設
  - 冷却施設

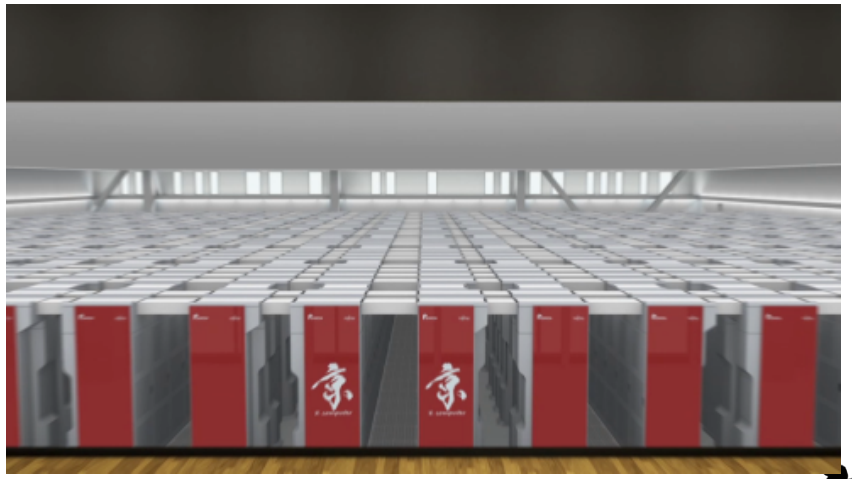


システムソフトウェア＆MPIライブラリ

- OS
  - SPARC64 VIIIfx拡張への対応など
    - SIMD、256FP Reg.など
    - ラージページ
  - アプリに対して低ノイズ環境を提供
    - OSジッタ排除
    - デーモンの割り込み最小化
    - 割り込み同期スケジューラ
- ジョブマネージャー
  - ACL（Access Control List)による柔軟な制御機能
  - 様々な優先順位付け(フェアシェアなど）によるスケジューリング
  - トポロジマッピングの最適化
  - ステージング機能との連携
    - 非同期ステージング機能
- MPIライブラリ
  - OpenMPI 1.4.1 (MPI 2.1準拠)ベース
  - ICC機能への対応
    - 高性能ハードウェアバリア
    - ハードウェアリダクション
    - 4DMAエンジンの活用
  - 3Dトーラス網（Tofuを含む）に適した集団通信アルゴリズムの適用
    - 3Dトーラスをターゲットに開発
    - バイセクションバンド幅が優位なものは6次元を適用

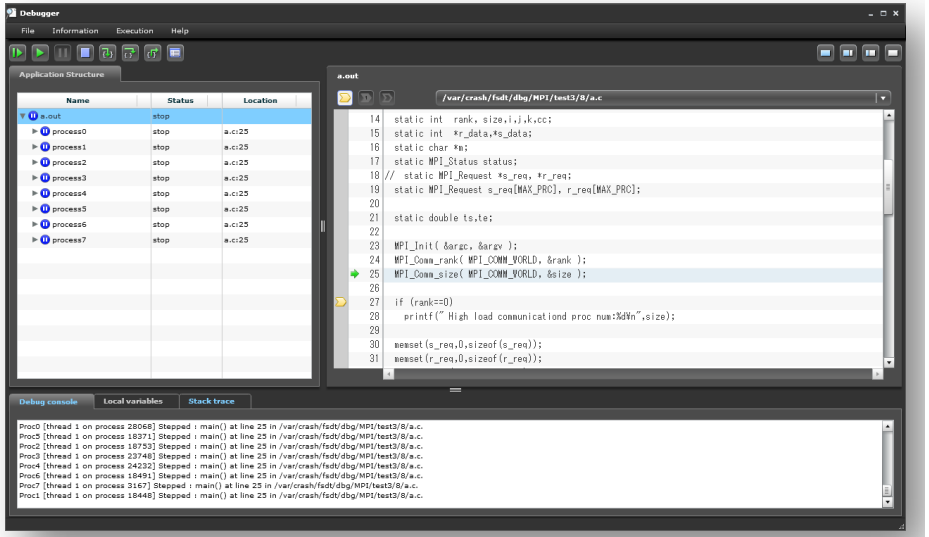
「京」を構成する要素技術

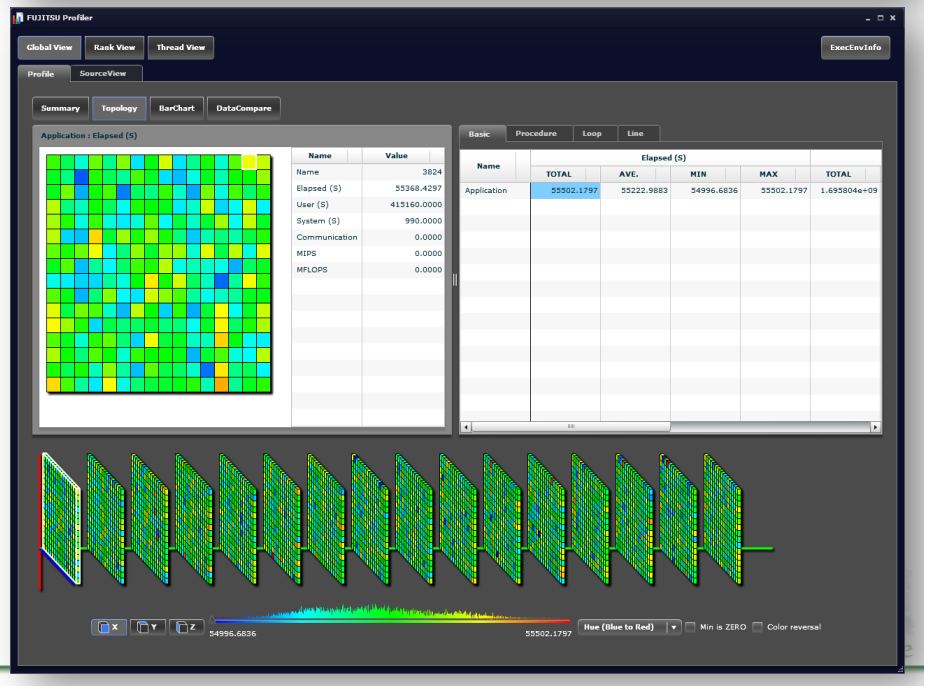
- CPU
  - 低消費電力かつ高性能
  - 高い信頼性と可用性
- インターコネクトネットワーク
  - 高い拡張性と広帯域、低遅延
  - 高い信頼性と可用性
- システムソフトウェア＆MPIライブラリ
  - 低ノイズ環境提供
  - 柔軟なジョブマネジメント
  - 高性能MPIライブラリ
- 開発環境/可視化ライブラリ
  - デバッガ/プロファイラ
  - 可視化ライブラリ
- ファイルシステム
  - 高い拡張性と高スループット
  - センター利用と柔軟性
- 施設
  - 電源施設
  - 冷却施設



開発環境

- IDE
  - Adobe AIRを用いた統合開発環境
  - 以下のデバッガやプロファイラを統合
- デバッガ
  - 会話型デバッガ
    - DWARF-2対応
  - 実行型デバッガ
    - プログラムコード中に割込コードを埋め込み
- プロファイラ
  - 大規模並列時にプロファイル情報表示表現をより見やすく
  - 様々なプロファイル情報を取得可能
    - サンプリング処理
    - イベントカウンタ
    - トレーサ

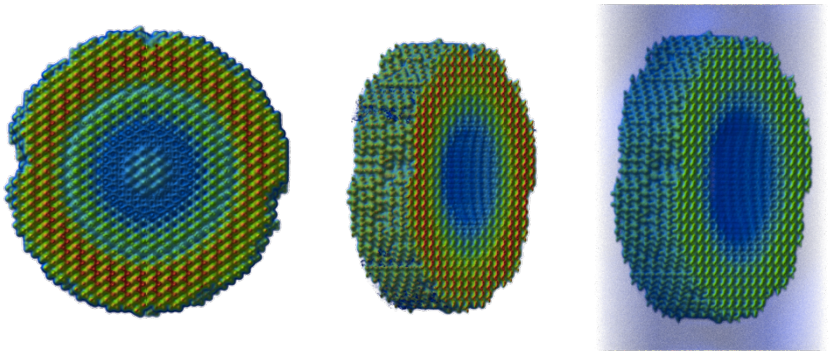


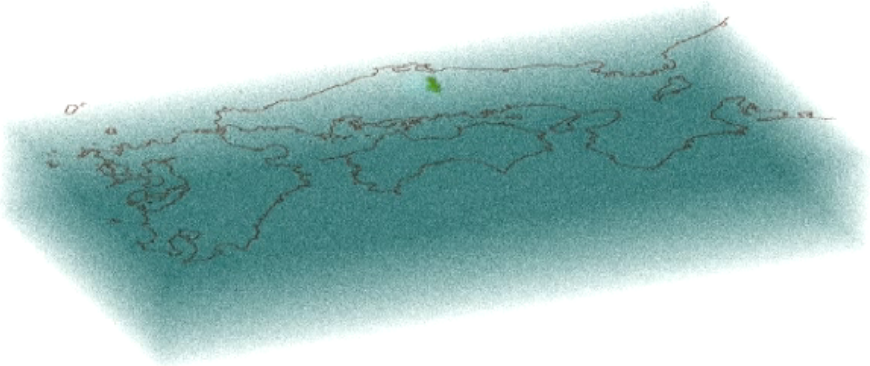


可視化ライブラリ

- 目的
  - 大規模データをそのまま可視化する。
  - シミュレーション中の状況を絵にする。
- 機能・特徴
  - 普通のジョブとして動作
  - ボリュームレンダリング（Particle Based Volume Rendering）
  - 構造格子、非構造格子、スライス画像
- 利用方法
  - レンダラーライブラリとAPI
  - パラメータファイル生成ツール
  - データ分割ツール

- 利用シーン
  - データを可視化するジョブにする
  - シミュレーションコード中に埋め込みコード中のデータを可視化する

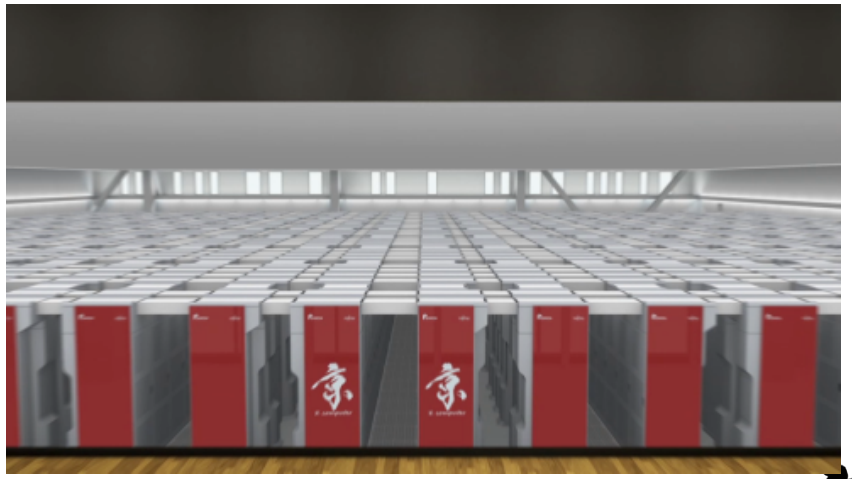




東京大学／地震研究所 古村孝志・ 岩井一央・前田拓人

「京」を構成する要素技術

- CPU
  - 低消費電力かつ高性能
  - 高い信頼性と可用性
- インターコネクトネットワーク
  - 高い拡張性と広帯域、低遅延
  - 高い信頼性と可用性
- システムソフトウェア＆MPIライブラリ
  - 低ノイズ環境提供
  - 柔軟なジョブマネジメント
  - 高性能MPIライブラリ
- 開発環境/可視化ライブラリ
  - デバッガ/プロファイラ
  - 可視化ライブラリ
- ファイルシステム
  - 高い拡張性と高スループット
  - センター利用と柔軟性
- 施設
  - 電源施設
  - 冷却施設



「京」のファイルシステム

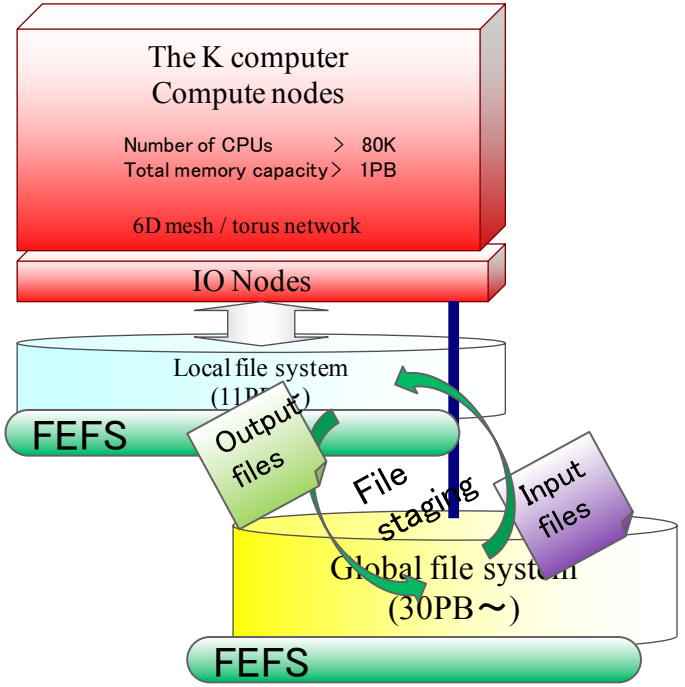
- 高性能と信頼性の両立
  - 1TB/sクラスの並列IO性能
  - 単一障害点がない構成で24H365Dの運用が可能
- 「京」の演算性能を妨げない
  - ジョブ間やユーザ間、システム間のIO処理を分ける
  - ファイルIOによるOSジッタの排除
- センター運用が可能な信頼性と柔軟性
  - QoSの導入
  - ACLやQuotaやアカウンティングが可能
  - システムのモニタリングが可能

Features		Current Lustre	Our 2012 Goal
System limits	Max file system size	64PB	100PB
	Max file size	320TB	1PB
	Max #files	4G	32G
	Max OST size	16TB	100TB
	Max stripe count	160	10k
	Max ACL entries	32	8191
Node Scalability	Max #OSSs	1020	10k
	Max #OSTs	8150	10k
	Max #Clients	128k	1M

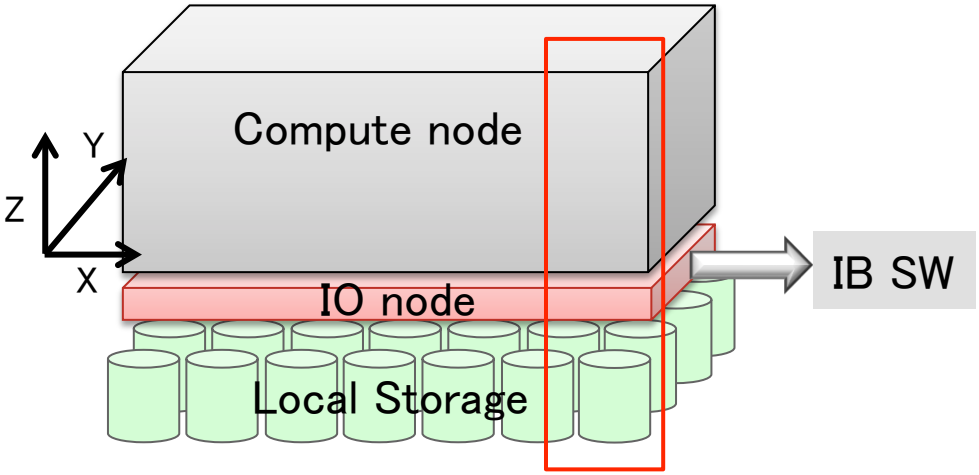
- FEFS (Fujitsu Exabyte File System)
  - Lustreファイルシステムをベースに様々な仕様を拡張
- グローバル＆ローカルの2段階ストレージ
  - グローバルファイルシステム：ユーザホーム領域
  - ローカルファイルシステム：ユーザジョブのスクラッチ領域
- ファイルシステムの開発過程と評価における経験
  - Lustreファイルシステムの品質は微妙。。。
  - 特にOSSやOSTが多い場合には、多くのバグやパフォーマンス問題、想定外がかなり内在している

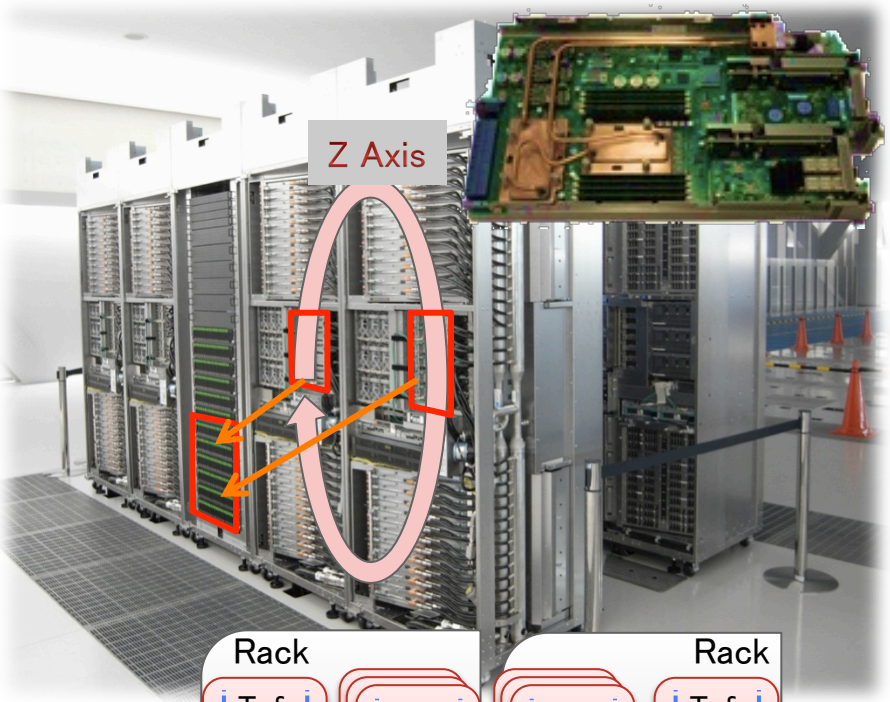
ファイルステージング概要

- ステージング機能
  - ステージ・イン（非同期）
    - ジョブ中で用いるグローバル・ファイルシステム上のファイルをローカル・ファイルシステム上にコピーする。
  - ステージ・アウト（非同期）
    - ジョブ実行によって生成あるいは更新されたファイルをジョブ終了後にグローバル・ファイルシステムに移動する。
- ステージングに関する指定はジョブスクリプト中にユーザが記述
- ステージング処理はジョブマネジメントシステムが実行
  - 非同期ステージング処理により、ジョブの隙間は極力空かないようにスケジューリング
  - ジョブ間のIO処理を極力分離するように調整
    - ステージングされるローカルファイルシステムのボリュームはジョブがアサインされる領域に最も近いLIOノード（RAIDがつながっているIOノード）上に配置
    - 利用するGIOノード（外向きのIOノード）も極力近い物を利用
  - ユーザによりジョブにとって最適なストライピングの設定が可能

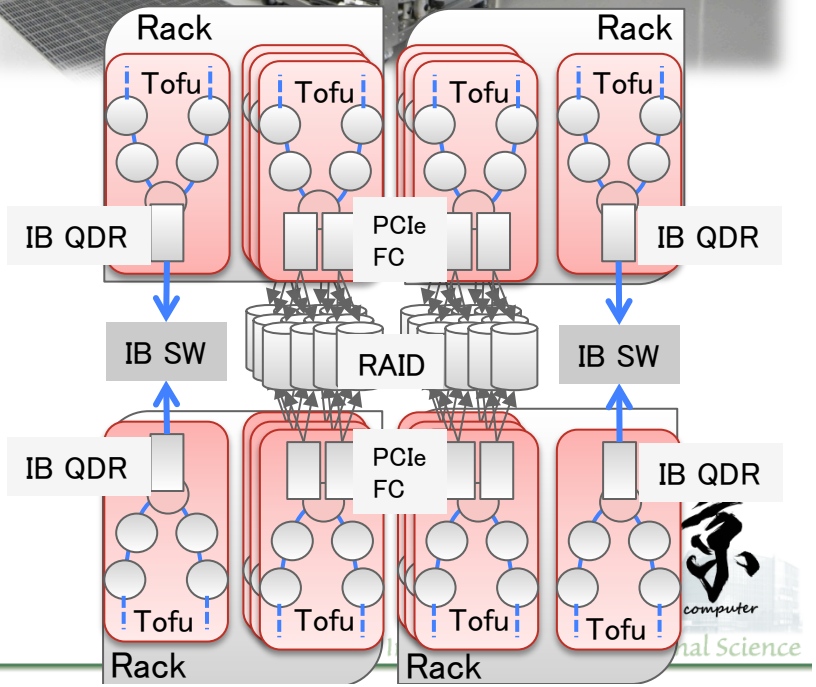


ローカル・ストレージの構造



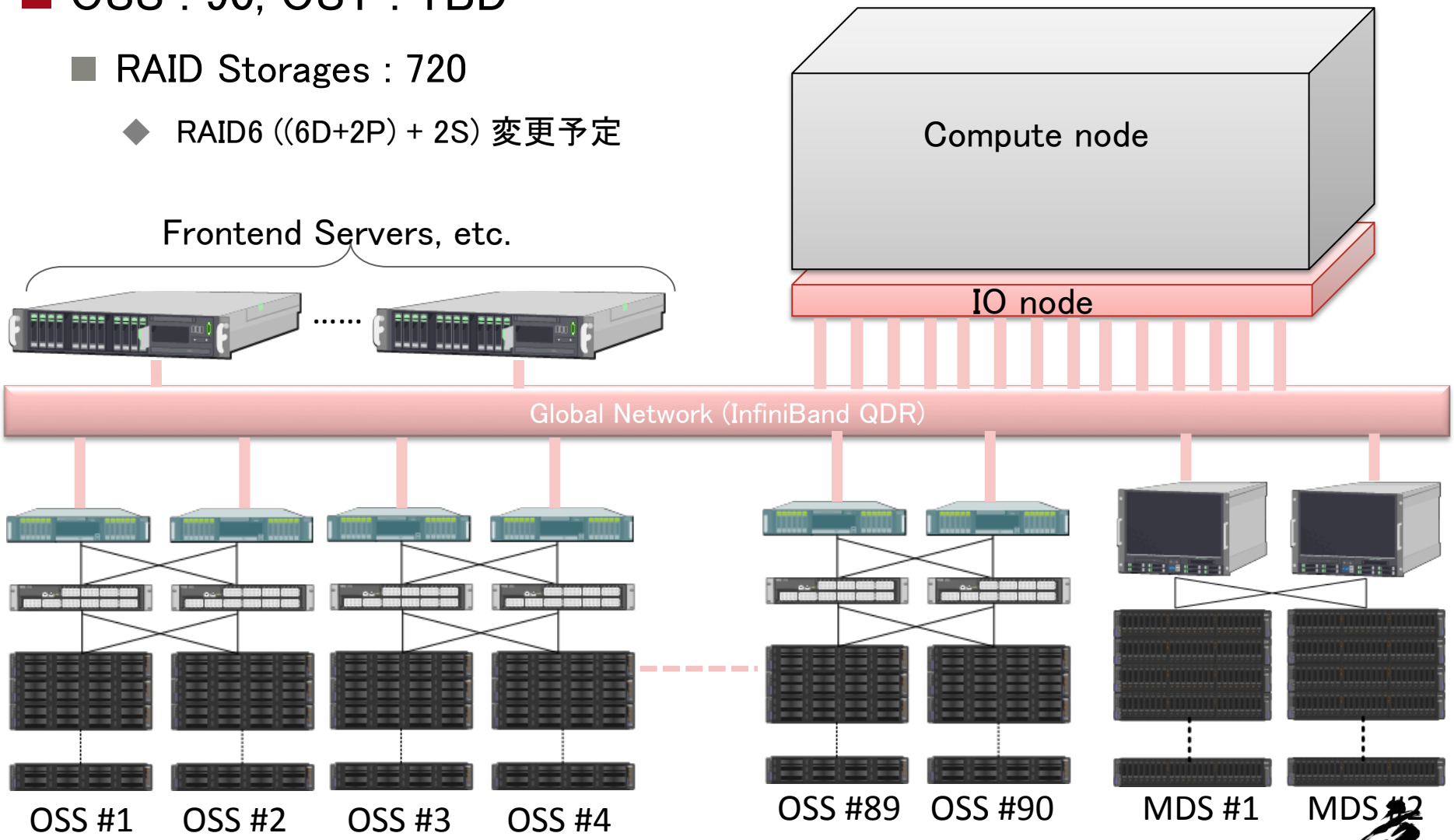


- MDS: 2, OSS: 2592, OST: TBD
  - RAID5 (4D+1P)x4+2S 変更予定
- IO系の利用パスの限定化
  - ファイルI/Oは基本的には同一Z座標のみで通信を行い、Z=0に対して通信
    - 計算ノードは同一XY軸のIOノードに通信する
    - ループバックデバイスをローカルFS上に作成して、計算ノード個別のファイルシステムを作成して、利用可能



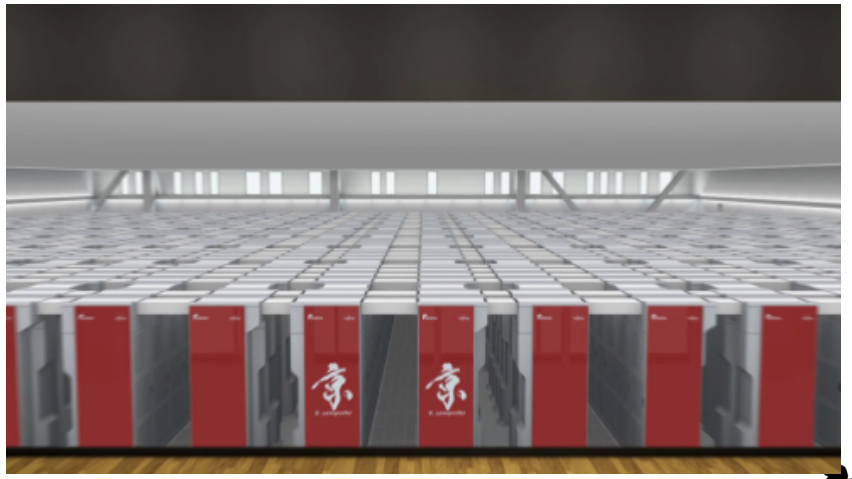
Global Storage I/O Architecture

- OSS : 90, OST : TBD
  - RAID Storages : 720
    - RAID6 ((6D+2P) + 2S) 変更予定



「京」を構成する要素技術

- CPU
  - 低消費電力かつ高性能
  - 高い信頼性と可用性
- インターコネクトネットワーク
  - 高い拡張性と広帯域、低遅延
  - 高い信頼性と可用性
- システムソフトウェア＆MPIライブラリ
  - 低ノイズ環境提供
  - 柔軟なジョブマネジメント
  - 高性能MPIライブラリ
- 開発環境/可視化ライブラリ
  - デバッガ/プロファイラ
  - 可視化ライブラリ
- ファイルシステム
  - 高い拡張性と高スループット
  - センター利用と柔軟性
- 施設
  - 電源施設
  - 冷却施設

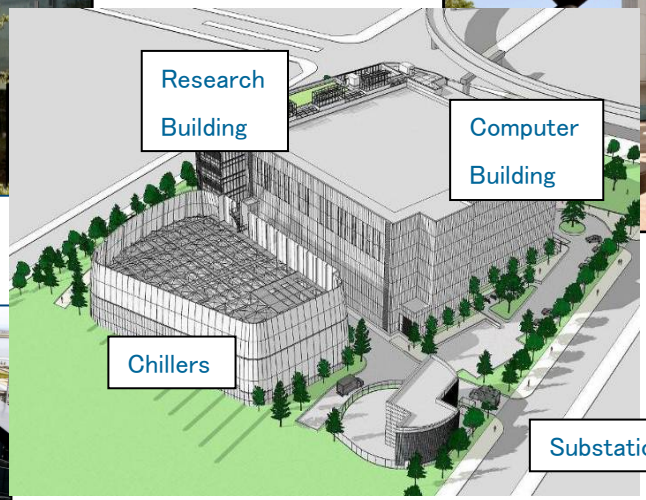


「京」の設置場所

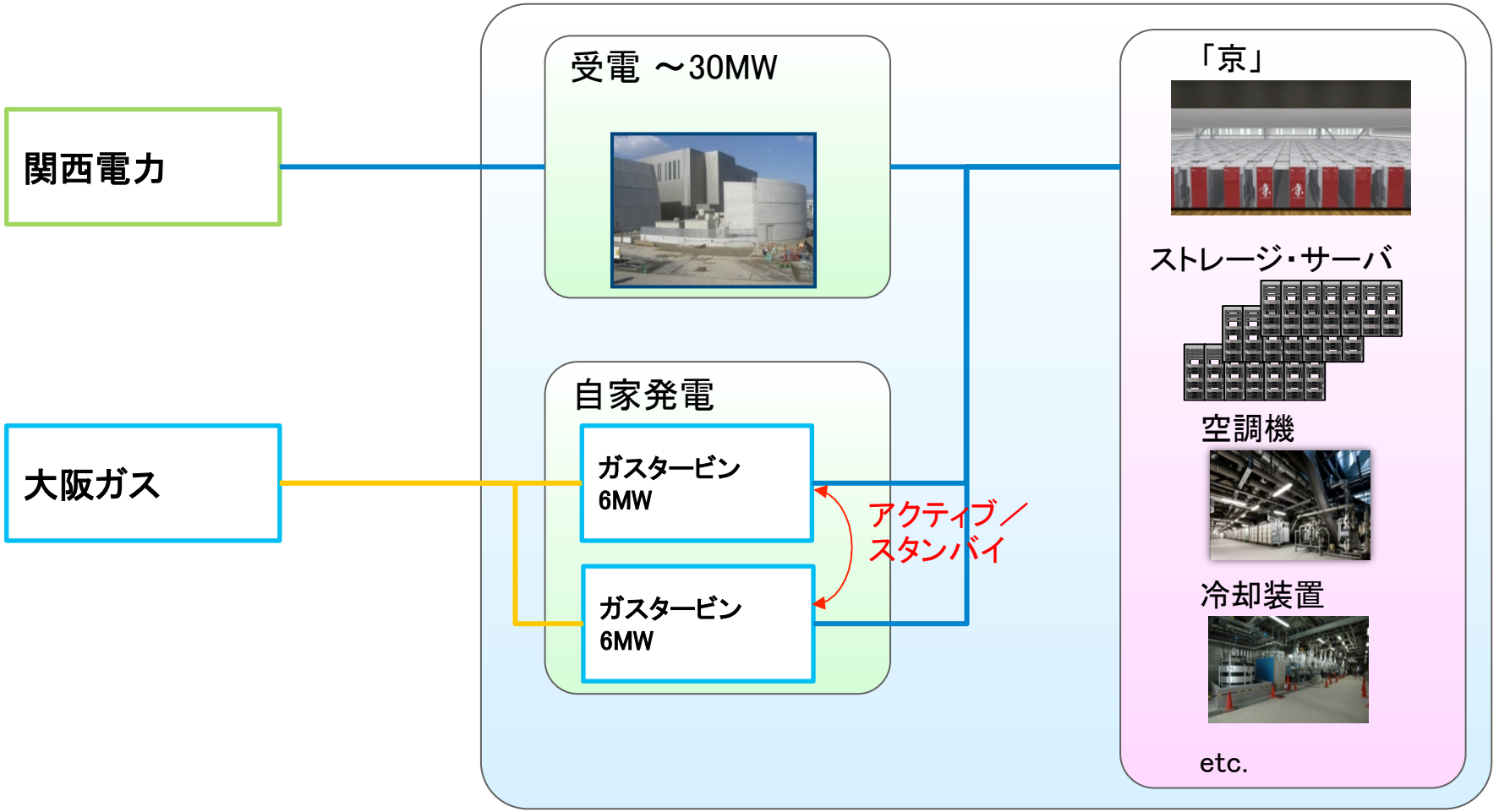
計算科学研究機構は2010年7月に設立されました。
AICS (Advanced Institute for Computational Science)

京コンピュータ前駅
から徒歩4分

建物の概要

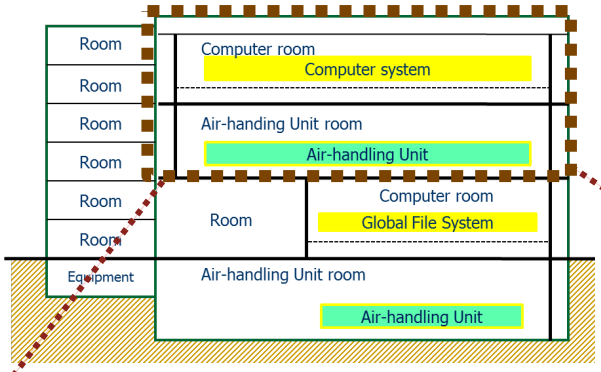

Research
Building
Computer
Building
Chillers
Substation

電源施設

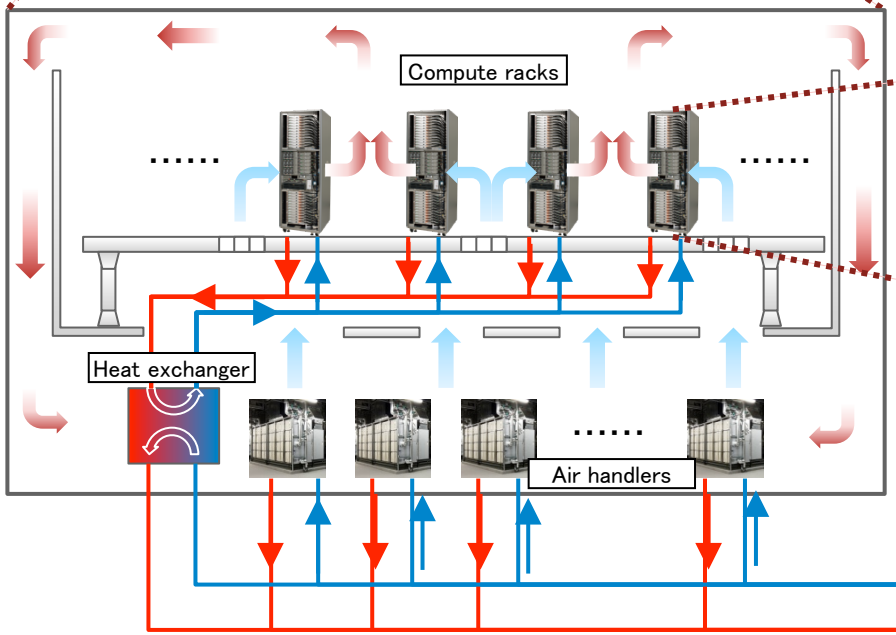

関西電力
大阪ガス
受電 ～30MW
自家発電
ガスタービン
6MW
ガスタービン
6MW
アクティブ／
スタンバイ
「京」
ストレージ・サーバ
空調機
冷却装置
etc.

冷却システム

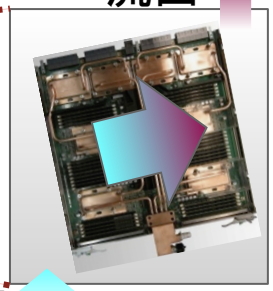
■ 「京」は空冷と水冷のハイブリッド



	冷却部分	温度
水	CPU と ICC	入力 :～15℃ 出力 :～17℃
空気	上記以外	室温 ～20℃



流出



流入

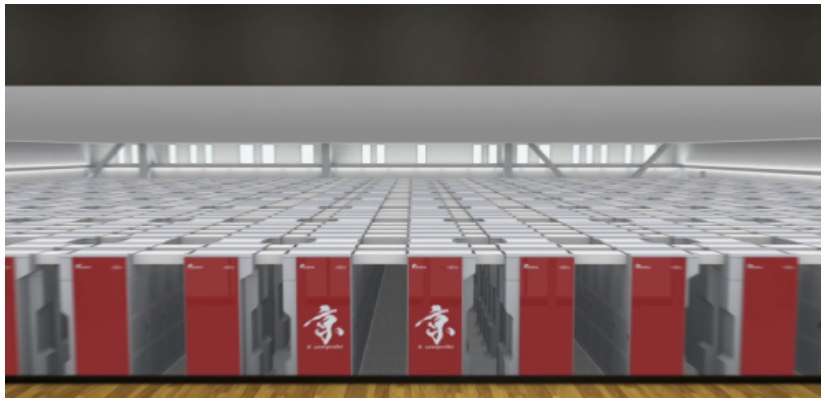
ラック上部からの視点

システムボードは空冷のために斜めに搭載

吸収式冷凍機やターボ冷凍機

講演のアウトライン

- 開発プロジェクトについて
- 「京」の概要
- 「京」を構成する要素技術
- まとめ



現在

- プロジェクト目標における性能目標は全て達成

- 京速コンピュータ「京」の開発プロジェクトはスケジュール通り
  - システムソフトウェア、運用ソフトウェアやライブラリの開発中
    - その後、実運用を行うための調整・ドキュメント作成を行う予定
  - 2012年6月にシステム完成
  - 2012年秋に共用開始

- 試験利用環境を提供
  - グランドチャレンジおよび戦略機関などを対象にアプリケーションのチューニングおよび最適化作業を行っている

Thank you for your attention !

A photo in the early evening