京速コンピュータ「京」の現状について

平成23年9月12日

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開発グループ
横川三津夫

日米のスーパーコンピュータ開発（平成17年当時）

- 米国は，軍事利用を中心に産業，科学技術・学術研究での利用のため，複数の大規模プロジェクトを並行して推進．
- 日本は，地球シミュレータ計画（平成9～14年）の後，大規模スーパーコンピュータ開発の計画がなく，米国に大きく引き離されてしまった．

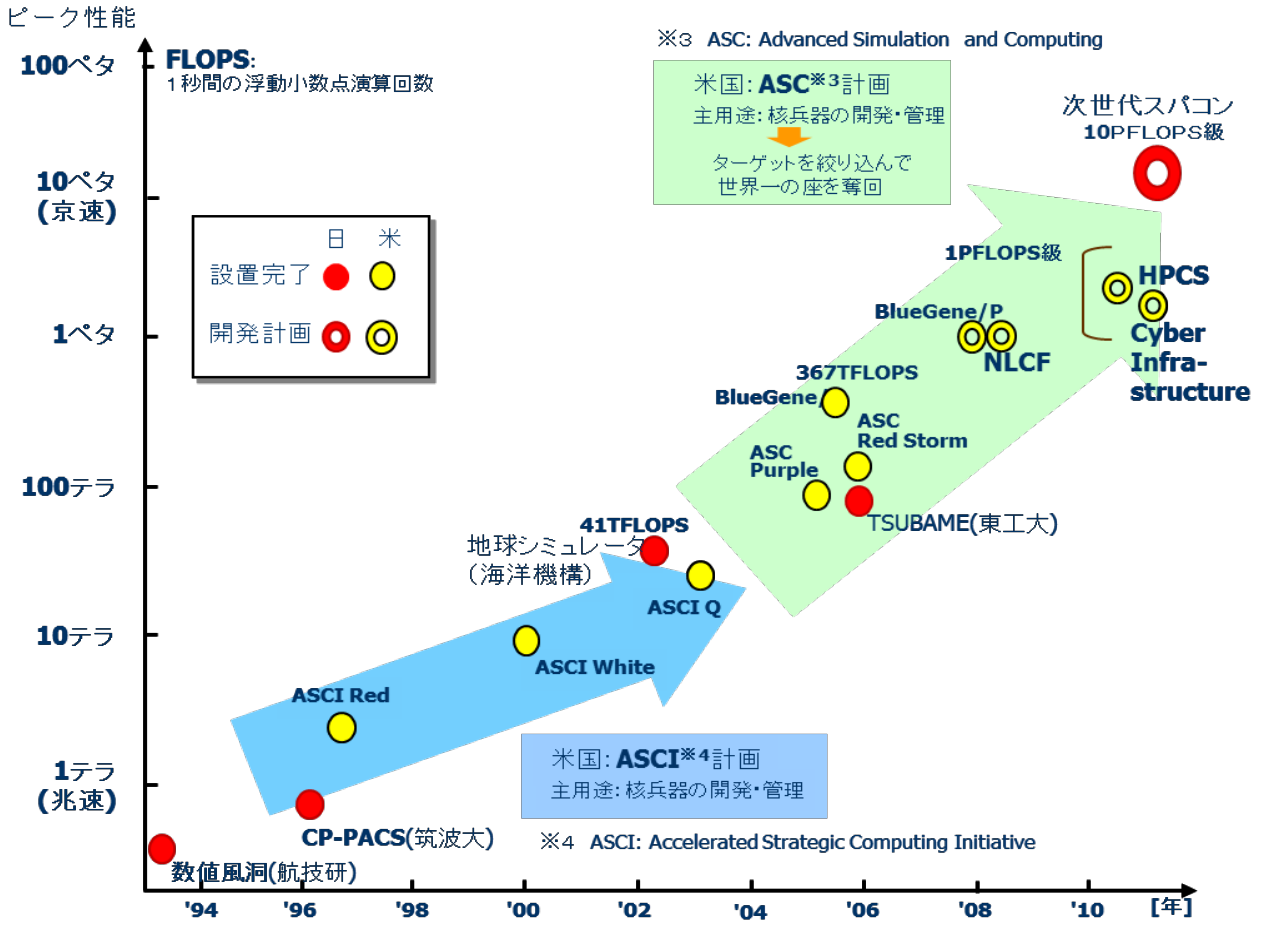
エネルギー省（DOE）
－ASC計画（旧ASCI計画）－
ターゲットを絞って世界最速を目指す(BlueGene)
－NLCF※1計画－
ライフサイエンスや核融合分野といった幅広い分野での利用を目指す

国防省（DoD）
－HPCS※2計画－
既存技術の延長線上にない新世代スパコンの開発を目指す

米国科学財団（NSF）の活動
－Cyber Infrastructure計画－
2011年に1ペタFLOPS超を目指す

※1 NLCF：National Leadership Computing Facility
※2 HPCS：High Productivity Computing System

注) IBMはBlue Gene/Q（ピーク性能10ペタFLOPS？）を2010～2012年頃を目指し開発するとしている．



革新的ハイパフォーマンス・コンピューティング・インフラ（高機能演算研究基盤）の構築（平成22年度から）

【概要】

次世代スーパーコンピュータプロジェクトを進化・発展させ，開発側視点から利用者側視点に転換し，多様なユーザーニーズに応える革新的な計算環境を実現する.

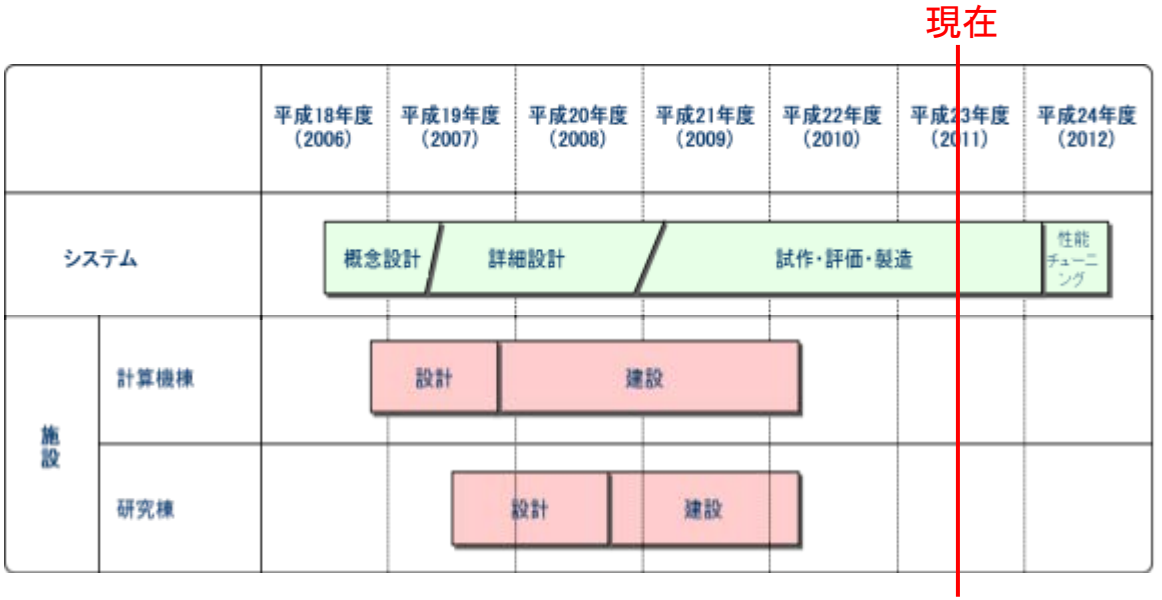
- ナンバーワンの世界最先端・最高性能を目指した次世代スーパーコンピュータの開発・整備
- 次世代スパコンと国内のスパコンをネットワークで結び協調的に利用するオンリーワンの「革新的ハイパフォーマンス・コンピューティング・インフラ（HPCI）」の構築

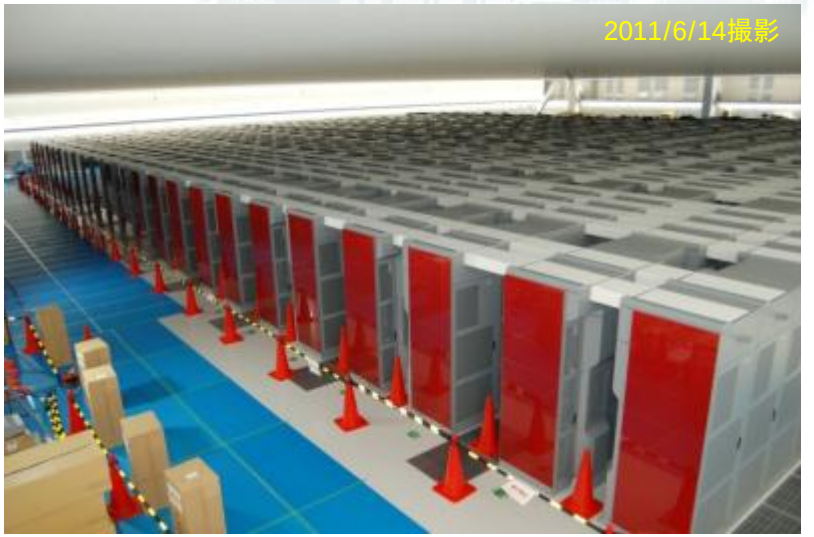
1. 次世代スーパーコンピュータの開発・整備
   我が国のハイパフォーマンスコンピューティングの中核となる次世代スパコンを平成24年の完成を目指し開発・整備する.
   （平成22年度末一部稼動，平成24年6月までに10ペタFLOPS級を達成）

2. 革新的ハイパフォーマンス・コンピューティング（HPC）に必要な研究開発

文部科学省資料より

システム整備状況(1/2)





システム設置状況

- 平成22年9月29日に計算機本体(筐体)の搬入開始.
- 平成22年11月の国際会議SC10(米国・ニューオリンズ)において, 4筐体(全体の約0.5%)のLINPACK性能を, TOP500及びGreen500に登録.
  - TOP500　　第170位 (48.03TFLOPS, 効率約92%)
  - Green500　　第4位 (828.67MFLOPS/W)
- 平成23年3月末 試験利用が可能な計算機環境の整備完了(一部稼働).

システム整備状況(2/2)

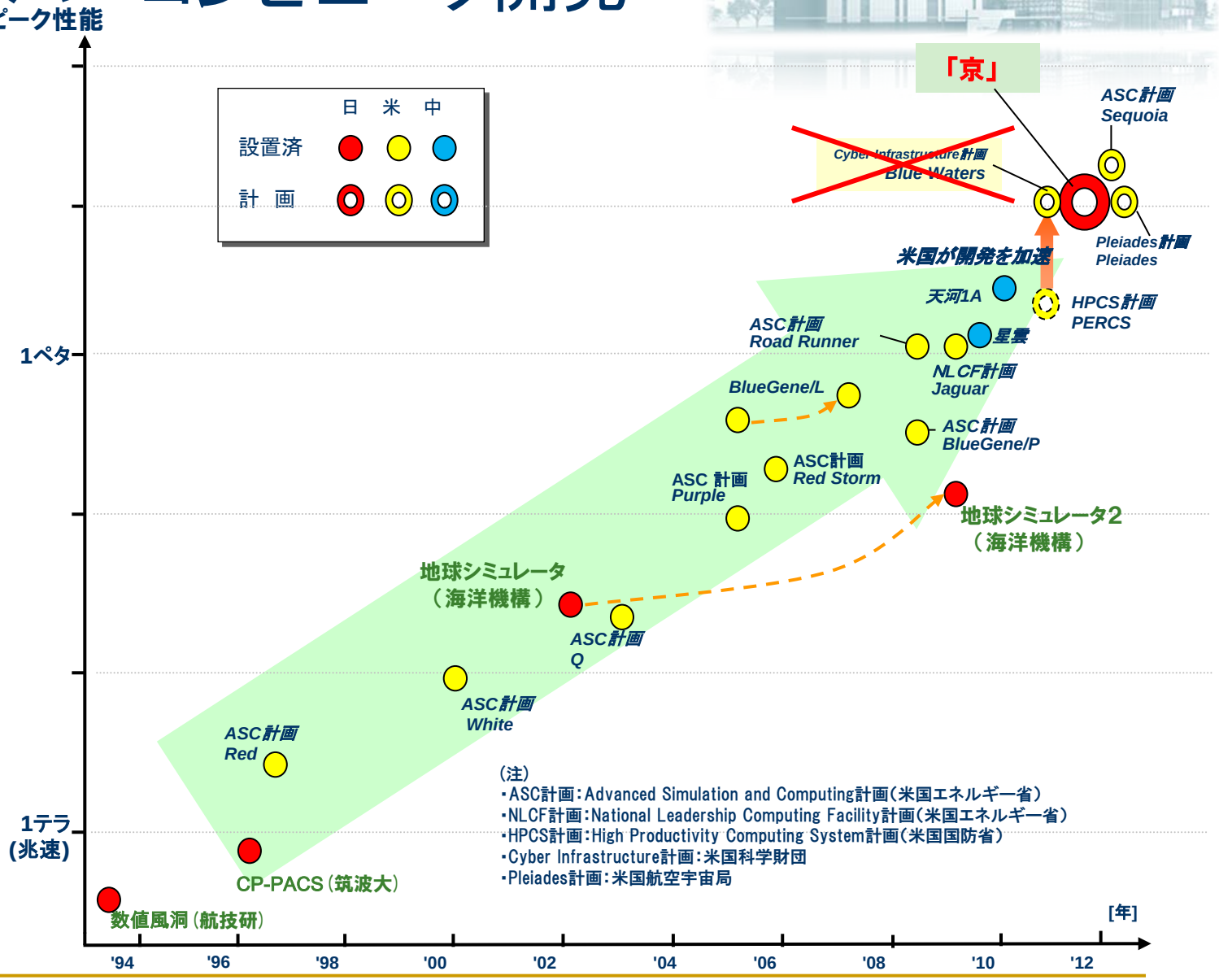
- 平成23年4月より，整備中の計算機本体の一部(16筐体)を，アプリケーション・ユーザ(グランドチャレンジ及び戦略分野の一部のユーザ)に提供し，試験利用を開始.
- 平成23年5月 システム動作，及び性能確認のひとつとして，LINPACK性能を計測し，TOP500サイトへ登録.
- 平成23年6月20日 ISC'11(国際スーパーコンピューティング会議2011，独・ハンブルク)にて，第37回TOP500リストで第一位を獲得.

  【LINPACK性能】
  - システム数: 672筐体(全体の約8割)
  - ピーク性能: 8.774ペタフロップス
  - 性能値: 8.162ペタフロップス(実行効率 93.0%)
  - 問題サイズ: 10,725,120次元
  - 実行時間: 100771秒 (約28時間)
  - 電力性能比: 825MFLOPS/W

- 8月末までに搬入，設置完了

世界のスーパーコンピュータ開発

- 米国は，軍事利用を中心に産業，科学技術・学術研究での利用のため，複数の大規模プロジェクトを並行して推進.
- 中国がスーパーコンピュータの開発で力をつけてきている.
  平成2010年11月に国防科学技術大学（NUDT）の天河1A（Tianhe-1A）が，TOP500で世界第1位になった.
- 我が国のスパコン性能は，「京」が7年ぶりに第1位（2011年6月）を奪還.



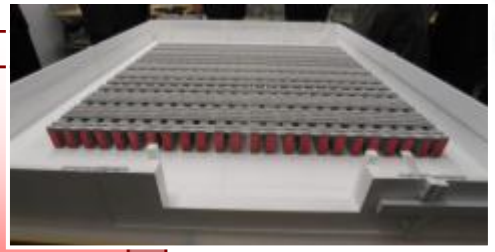
「京」の設置風景

- 平成22年9月29日に設置開始（8ラック）

「京」システムの全体概要

- ターゲット性能
  - LINPACK 10PFLOPS
- 多くのアプリケーションの高速実行が可能なシステム
  - ハイパフォーマンスコンピューティング向けCPUを採用
    - マルチコアCPU（8コア），演算アクセラレータ付き（SIMD機構），FPレジスタ数増強
  - ペタフロップス級のアプリケーション実効性能を目指す
- 高性能・低消費電力CPU（SPARC64 VIIIfx）を採用
  - 45nm CMOS プロセス
  - 2.2GFLOPS/W，ワット当たり世界トップレベル
- 高信頼性システム
  - 「故障しにくい」，「1か所故障しても全てが止まらない」．「故障個所はすぐ直せる」
  - ネットワークの高信頼性化： 自動代替経路，自動再構成機能
  - サーバ二重化，ファイル経路二重化など
- 利用者にとって使いやすいシステム
  - 世界標準のソフトウェア環境を提供

システム構成概要



京速コンピュータ「京」
計算ノード群

6次元メッシュ/トーラス結合

IOノード群

ローカルファイルシステム
(11PB～)

管理用・制御用ネットワーク

データ系ネットワーク

Pre/Post
処理サーバ

「京」
フロントエンド

外部接続系ネットワーク

外部
へ

利用者

管理用
サーバ群

制御用
サーバ群

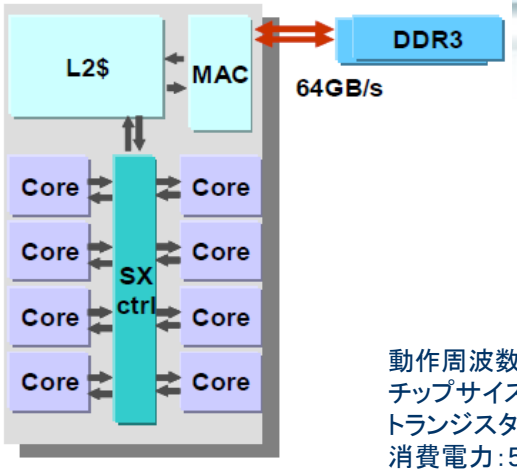
ジョブ管理
ユーザ管理

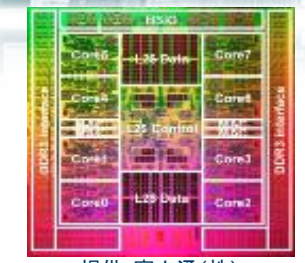
構成・制御

グローバルファイルシステム
(30PB～)

プロセッサ構成

- 8コア構成，各コア256本の浮動小数点レジスタを備えたスーパースカラ方式
  - SIMD拡張（積和演算器2個 x 2セット）
  - コア当り16GFLOPS，CPU当り128GFLOPS
- ハードウェアバリア機構
- プリフェッチ機構
- コア共有の2次キャッシュ（6MB，12way）
  - セクタキャッシュ機能





提供：富士通（株）

動作周波数： 2GHz
チップサイズ： 22.7mm x 22.6mm
トランジスタ数：760 M トランジスタ
消費電力：58W（水冷，30℃時）

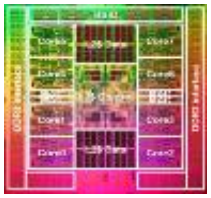
	仕　様
CPU性能	128GFLOPS（16GFLOPSx8コア）
コア数	8個
浮動小数点演算器構成（コア当り）	積和演算器：2×2個（SIMD） （逆数近似命令： SIMD動作） 除算器：2個 比較器：2個
	浮動小数点レジスタ（64ビット）：256本 グローバルレジスタ（64ビット）：188本
キャッシュ構成	1次命令キャッシュ：32KB（2way） 1次データキャッシュ：32KB（2way） 2次キャッシュ：6MB（12way），コア間共有
メモリバンド幅	64GB/s(0.5B/F)

より詳細な情報は，「SPARC64™ VIIIfx Extensions」を参照のこと
http://img.jp.fujitsu.com/downloads/jp/jhpc/sparc64viiifx-extensions.pdf

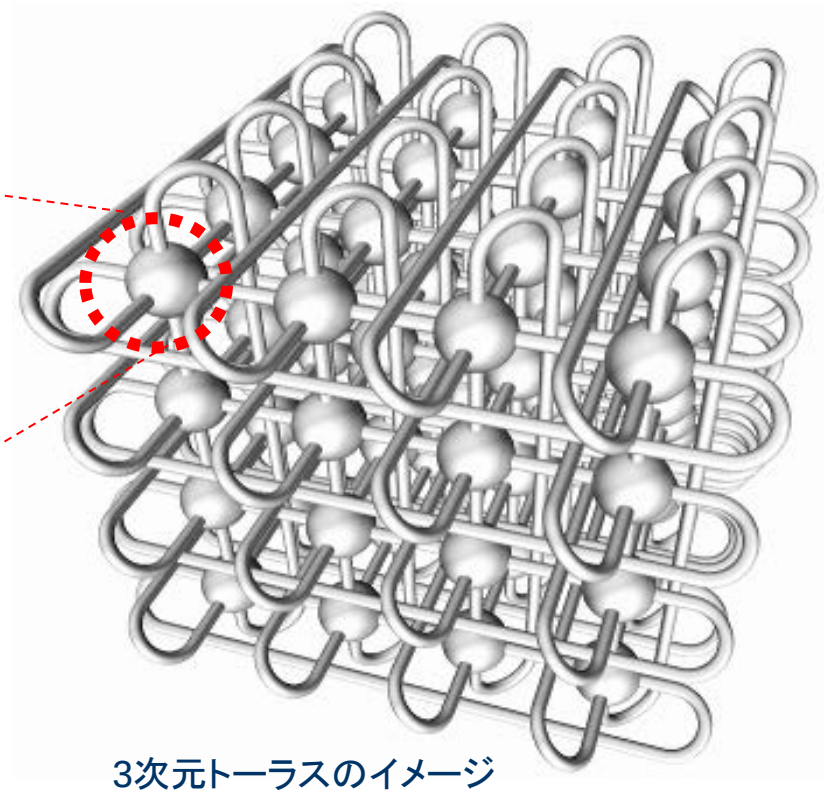
計算ノードとインターコネクトの構成

- 計算ノード
  - CPU: 1個
  - メモリ量: 16GB
  - ICC（Interconnect Controller）: 1個

- 3次元メッシュ/トーラスネットワーク: Tofu
  - ユーザ・ビューは3次元トーラス
  - 帯域: 3次元の正負各方向にそれぞれ5GB/s x 2（双方向）【理論ピーク】
  - ケーブル: 約200,000本，約1000km



SPARC64™ VIIIfx



3次元トーラスのイメージ

提供：富士通（株）

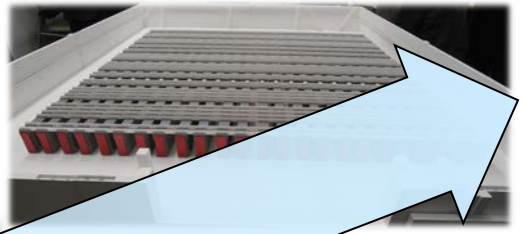
システムの基本性能（実測）

- CPUの性能
  - DGEMM (Matrix Multiply)
    - 123.6 GFlops（効率 96.6%）
  - コア間のハードウェアバリア同期性能
    - 49 ns
- メモリアクセス性能
  - STREAM Triad： 46.6 GB/s (peak：64 GB/s)
- インターコネクト（Tofu）性能
  - 2ノード間の帯域
    - 4.76 GB/s（peak：5 GB/s）
  - レンテンシ
    - 最大で112 ns/hop

「京」の構成

システム全体

計算ラック 800以上



計算速度：1京回/秒
＝10ペタフロップス
メモリ容量：1PB以上

計算ラック群

計算ラック×8

計算速度：98.4兆回/秒
メモリ容量：12TB

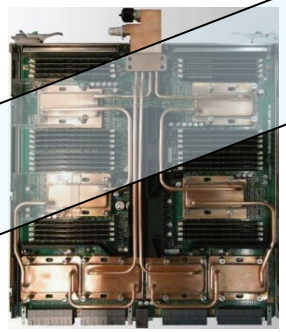
計算ラック

システムボード×24
IOシステムボード×6

計算速度：12.3兆回/秒
メモリ容量：1.5TB

システムボード

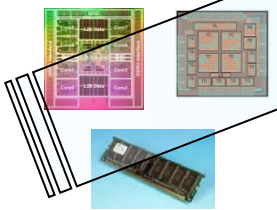
ノード×4



計算速度：5120億回/秒
メモリ容量：64GB

ノード

CPU×1
ICC×1
メモリ



計算速度：1280億回/秒
メモリ容量：16GB

システムソフトウェア環境

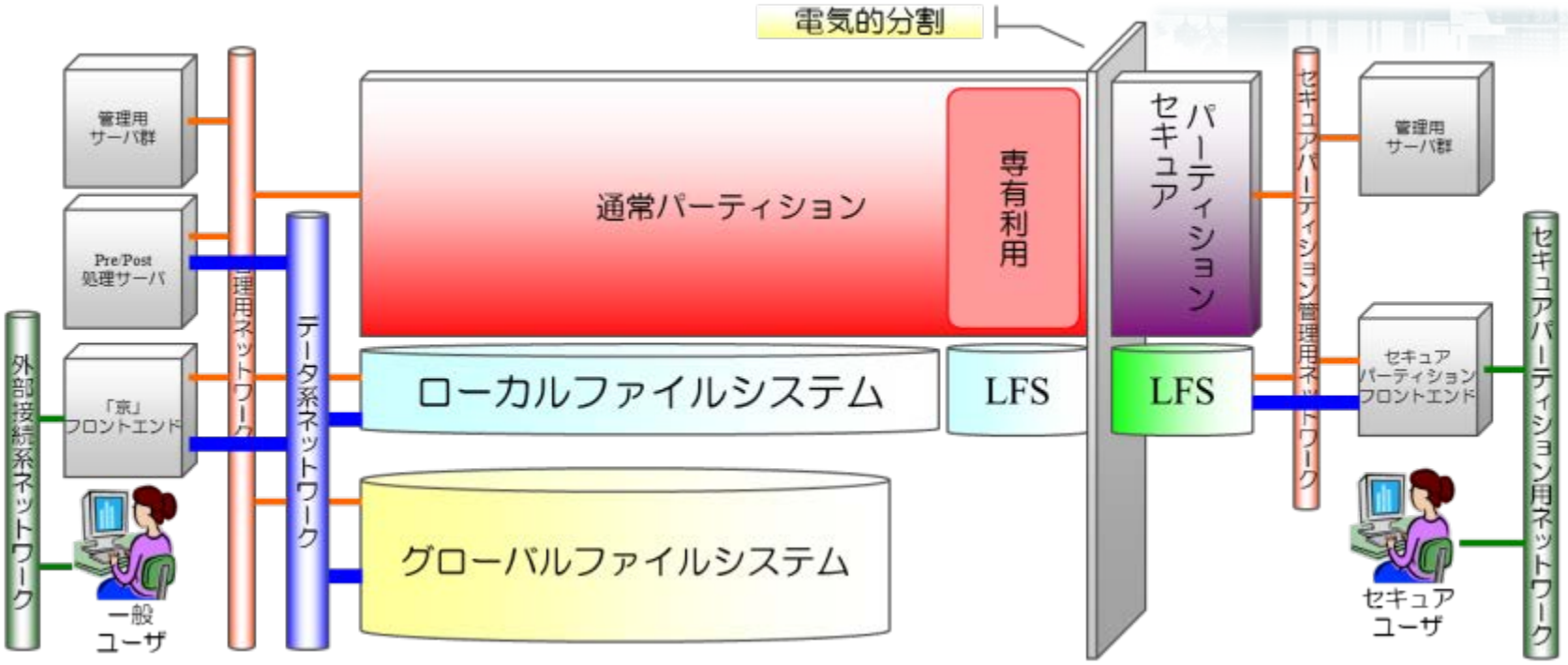
- 計算ノードOS
  - Linuxベース
  - 計算ノードではOSによる演算処理への外乱を極力排除
    - 超並列環境での低ノイズ環境（アプリ実行時間のブレが少ない環境）

- ファイルシステム
  - Lustreファイルシステムをベースとした拡張ファイルシステム（FEFS）を採用
  - グローバルファイルシステム（GFS）とローカルファイルシステム（LFS）の2階層のファイルシステム
    - ステージングによる運用
      - ジョブ実行前にGFSからLFSへ実行・入力ファイルを転送（ステージイン）
      - ジョブ終了後にLFSからGFSへ出力ファイル転送（ステージアウト）
    - LFSはラック単位にローカリティによる配置最適化やジョブ毎に帯域を（ある程度）保証する機能
      - ステージング帯域のQoS機能

プログラム開発環境(1/2)

- 提供言語
  - Fortran, XPFortran, C/C++
    - C/C++:ISO/IEC9899:1999(C99)準拠, ISO/IEC14882:2003(C++2003),(GCCのC/C++拡張を一部対応)
    - Fortran:ISO/IEC1539-1:1997(Fortran95)準拠およびFortran2003の一部機能に対応
    - OpenMP 3.0に準拠
    - CPUの機能拡張を踏まえた最適化コンパイラを提供
      (SIMD, 256 FPレジスタ, セクタキャッシュ, コア間バリアなど)
- 4倍精度演算のサポート
  - Fortran, C/C++
    - ソフトウェア処理としてbinary128データ形式をコンパイラで提供
    - ハードウェア処理としてdouble-doubleデータ形式の演算ライブラリを提供
- MPIライブラリ
  - MPI2.1規格(OpenMPI 1.4.1をベース)に対応
  - Tofuネットワーク(直接網)に最適化されたライブラリ
    - 低レイテンシかつ広帯域

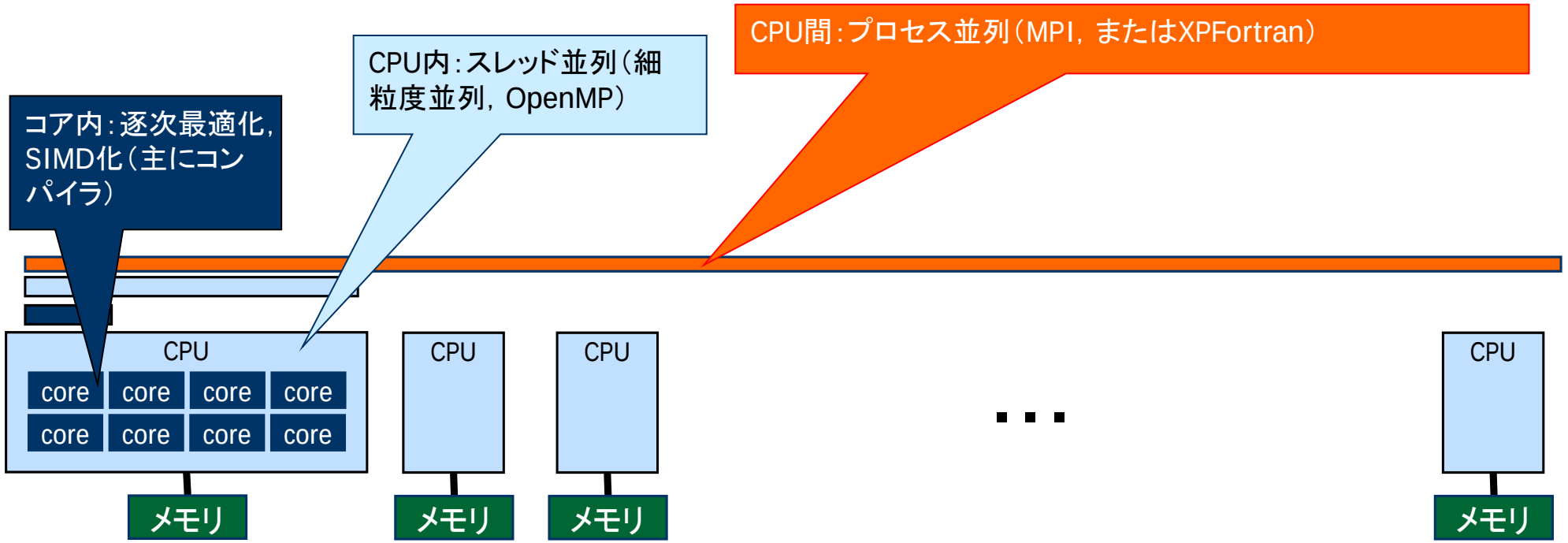
システム・パーティションについて



- 専有利用
  - 特定期間，通常パーティション内の特定の計算ノード群を論理的に分離
    - ローカルファイルシステムも専有
    - グローバルファイルシステム，フロントエンド等は他ユーザと共有
- セキュアパーティション
  - 電気的に計算ノード群を分離．管理系のサーバ群も専用に用意
    - グローバルファイルシステムへのアクセスはできない
    - ローカルファイルシステムをホーム領域として使用

プログラミングモデル

- スレッド並列＋プロセス並列のハイブリッド型を推奨
  - コア内：コンパイラによる逐次最適化，SIMD化
  - CPU内：スレッド並列（自動並列化，OpenMP）
  - CPU間：プロセス並列（MPI，XPFortran）
- フラット型も可能



プログラム開発環境(2/2)

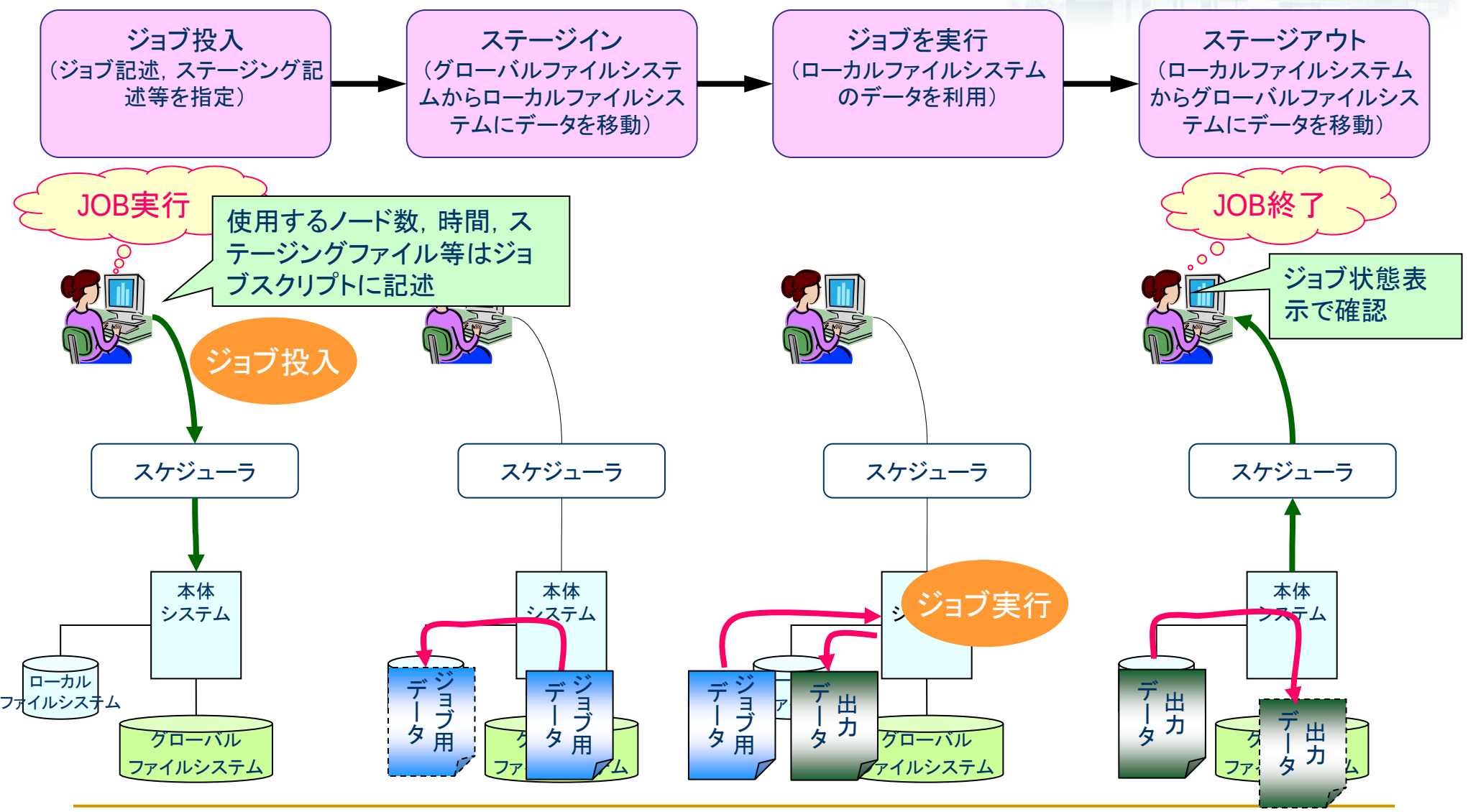
- 数値計算ライブラリ，科学技術計算ライブラリ
  - BLAS，LAPACK，FFTW(1.2，1.3)，およびSSL-II(富士通製科学技術計算ライブラリ)
- IDE環境
  - ユーザ端末で実行するリッチクライアントを提供(予定)
- デバッグ
  - 会話型デバッガ(ソースコードデバッガ)
  - プログラム埋め込みデバッグ(関数埋め込み)
  - 実行時デバッグ(実行時のデバッグ出力)
- パフォーマンス・チューニング
  - プロファイラ
    - ルーチン/プログラム区間の割合/経過時間/ハードウェアモニタ情報/通信時間/プロセス間・Thread間バランスなどのプロファイル情報を採取
  - トレーサ
    - 並列アプリケーションの時系列情報の取得手段を提供
      - OTF形式をサポート
  - パフォーマンスカウンタI/FとしてPAPIも提供

ジョブ実行環境

- ジョブ実行環境
  - ジョブタイプ
    - 会話ジョブ，バッチジョブ，ステップジョブ，バルクジョブなど
  - 様々なMPIジョブの実行が可能
    - ハイブリッドMPI，フラットMPIなどのジョブが実行可能
    - MPIジョブの形態としてSPMD，MPMDなどが実行可能
  - リソース指定
    - ノード数，仮想トポロジ，ステージングのファイル容量，経過時間，ノード割付ポリシーなどが指定可能

- MPIジョブの計算ノードへの割付
  - 全てのジョブに論理的1次元~3次元トーラス空間を提供
    - ジョブスクリプトに1~3次元のトーラス形状を指定
      - rankマッピングはいくつかの方法が可能
        - Rank番号をXYZで自動的にマッピングする方法
        - hostfileによるユーザ指定の方法
    - IO占有を指定
      - IO性能を確保したいジョブのための指定

バッチジョブ実行時の処理の流れ



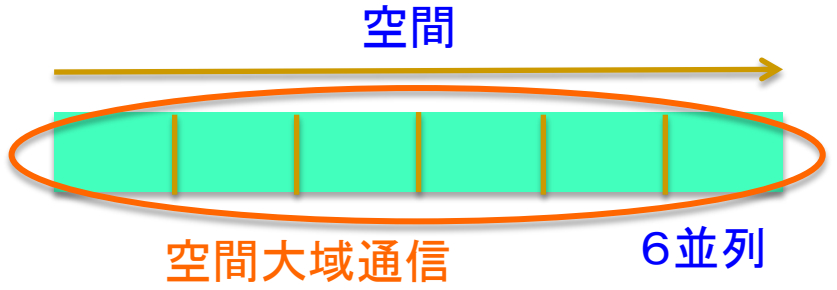
アプリケーションによる性能実証に向けて

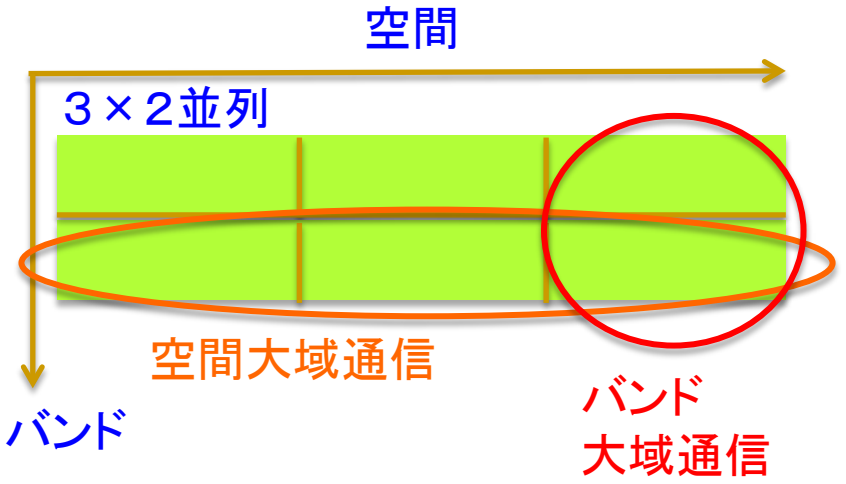
- 京の性能を確認するため，6本のアプリケーションについて，性能最適化（単体性能向上，超並列化），及びその評価を実施中．

プログラム	分野	アプリケーション
NICAM	地球科学	全球雲解像大気大循環モデル
Seism3D	地球科学	地震波伝播・強震動シミュレーション
PHASE	ナノ	平面波展開第一原理分子動力学解析
FrontFlow/Blue	工学	Large Eddy Simulation (LES)に基づく非定常流体解析
RSDFT	ナノ	実空間第一原理分子動力学計算
LatticeQCD	物理	格子QCDシミュレーションによる素粒子・原子核研究

RSDFTコードのチューニング

- 空間のみの並列化から，空間+エネルギーバンド並列(2軸並列)へ
  - 10万並列レベルが可能
- 2軸並列により，空間・バンドそれぞれに対するコレクティブ通信が局所化された.
  - 空間並列のみの場合は全プロセッサ間の大域通信が必要
- トーラスネットワークへのタスク・マッピングに工夫
- バンド並列のロードインバランスを改善





RSDFTの測定結果（暫定）

- 「京」の96筐体，9216CPU（ピーク性能 1.18Pflops）で計測
- 格子数 576 × 576 × 40，原子数 25236，バンド数 53280
- 空間分割数 32 × 48 × 2，バンド分割数 3

区間	全体（秒）	演算（秒）	通信時間（秒）			性能（Tflops）
			隣接/空間	大域/空間	大域/バンド	
DIAG	322.49	272.49	4.04	41.72	4.24	353.71
DTCG	80.41	20.87	36.38	23.15	0.01	8.38
Gram-Schmidt	110.38	74.87	-	18.70	16.81	684.74
Total	513.28	368.23	40.42	83.57	21.06	370.85

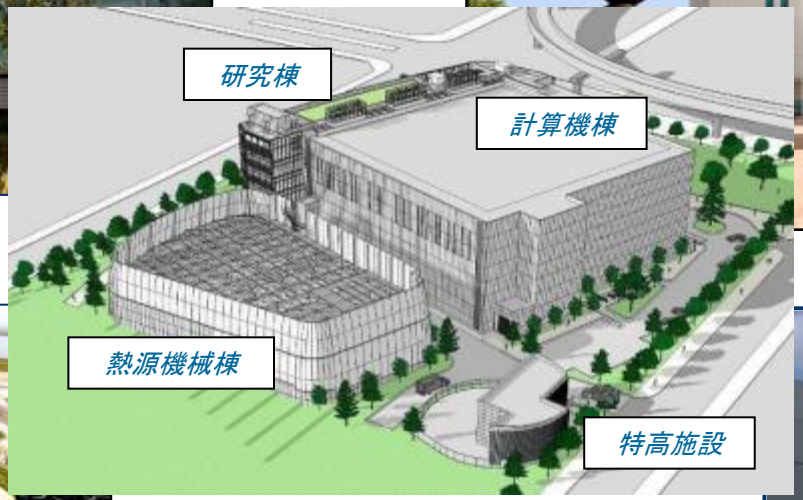
- 実効性能 371Tflops（ピーク比 31.4%）という高い性能を確認

試験利用について

- 平成23年4月1日から試験利用を開始.
  - 平成23年7月11日から試験利用期間Ⅱ
  - 平成23年10月24日（予定）から試験利用期間Ⅲ

- 利用できる計算資源を順次拡大
  - 計算ノード数（利用期間Ⅱ）
    1,536ノード（ピーク性能 約200TFLOPS）→4,608ノード（約590TFLOPS）
  - ログインノードを1台から2台に増加
  - 利用可能なディスク容量を増加（0.5TB/4TB → 6TB/32TB）

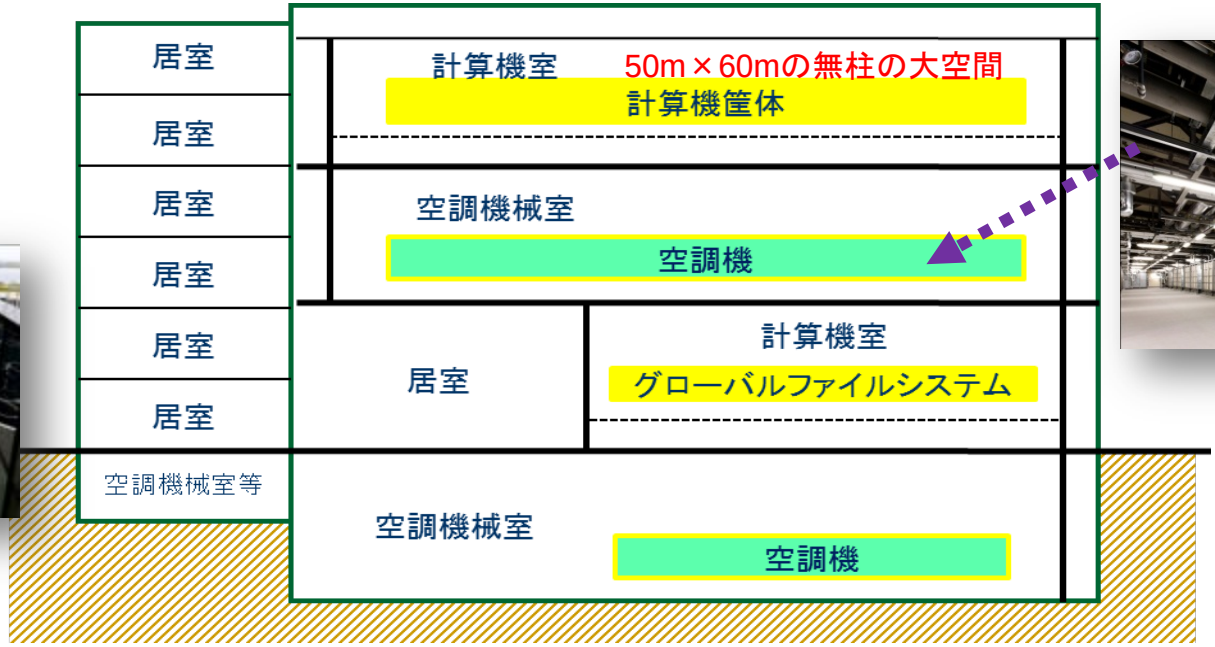
施設概要



施設概要

研究棟
- 地上6階，地下1階（鉄骨造り）
- 建築面積 ~1,800m^2，延床面積 ~9,000m^2

計算機棟
- 地上3階，地下1階（鉄骨造り）
- 建築面積 ~4,300m^2，延床面積 ~10,500m^2



熱源機械棟（面積 1900m^2）

吸収式冷凍機 x 4

ターボ型冷凍機 x 3

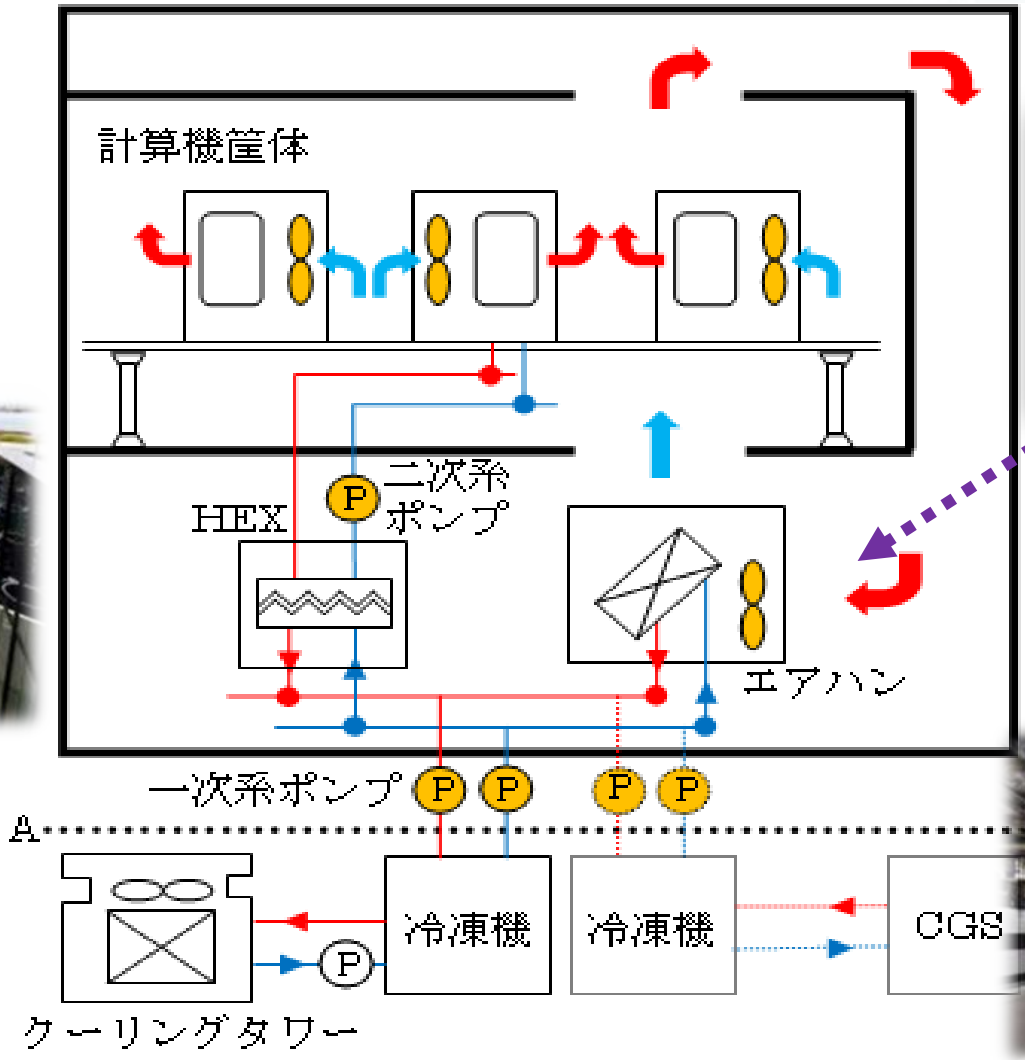
CGS (5MW) x 2

特別高圧変電施設（面積 200m^2）

30MW
77,000V（受電）
→ 6,600V

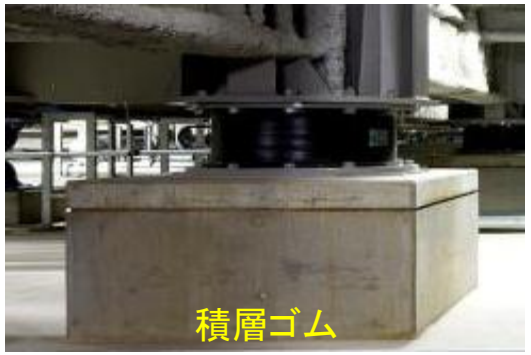
冷却システム

クーリングタワー



施設の特徴(1)

- 基礎
  - 地盤支持力の強化，液状化対策等を目的として，埋立部分（地下20m）の地盤改良を実施地盤改良部を支持層とする2mの直接基礎
- 耐震性
  - 研究棟及び計算機棟は免震構造．積層ゴムによる免震装置を49か所設置．振動を素早く抑える鉛ダンパーや鋼性U型ダンパーを設置
  - 耐震グレードはSグレード（震度6強レベルの大地震が起きても主要な機能を確保）


積層ゴム


鉛ダンパー


U型鋼性ダンパー

- 耐久性（海浜地区の塩害対策）
  - 重要な構造体には防錆対策などを実施
  - 計算機棟外壁をアルミパネル，研究棟外壁をガラス基調として建設

施設の特徴(2)

計算機室構造

- 設置レイアウトの自由度の確保，均一配置，スパコンの相互接続時の通信ケーブル長の短縮化に対応するため，計算機室は無柱.
  - 約200,000本のケーブル
- 計算機設置フロアの床耐荷重は平均約1 トン/m^2

まとめ

- 「京」の設置，調整作業は順調
  - 8月末時点で，「京」のすべての筐体の設置が完了
  - システムソフトウェアの開発，調整はこれから本格化
- 試験利用として，グラチャレ機関，戦略機関の一部のユーザに計算資源を提供中
  - アプリ本数 約20本，アカウント数 約200
  - 状況をみながら試験利用の計算資源を拡大.
- 今後の予定
  - 平成24年3月末　ストレージ等の周辺機器の設置完了
  - 平成24年6月末　計算機システムの開発完了
  - 平成24年11月　　共用開始