京速コンピュータ「京」の最新状況

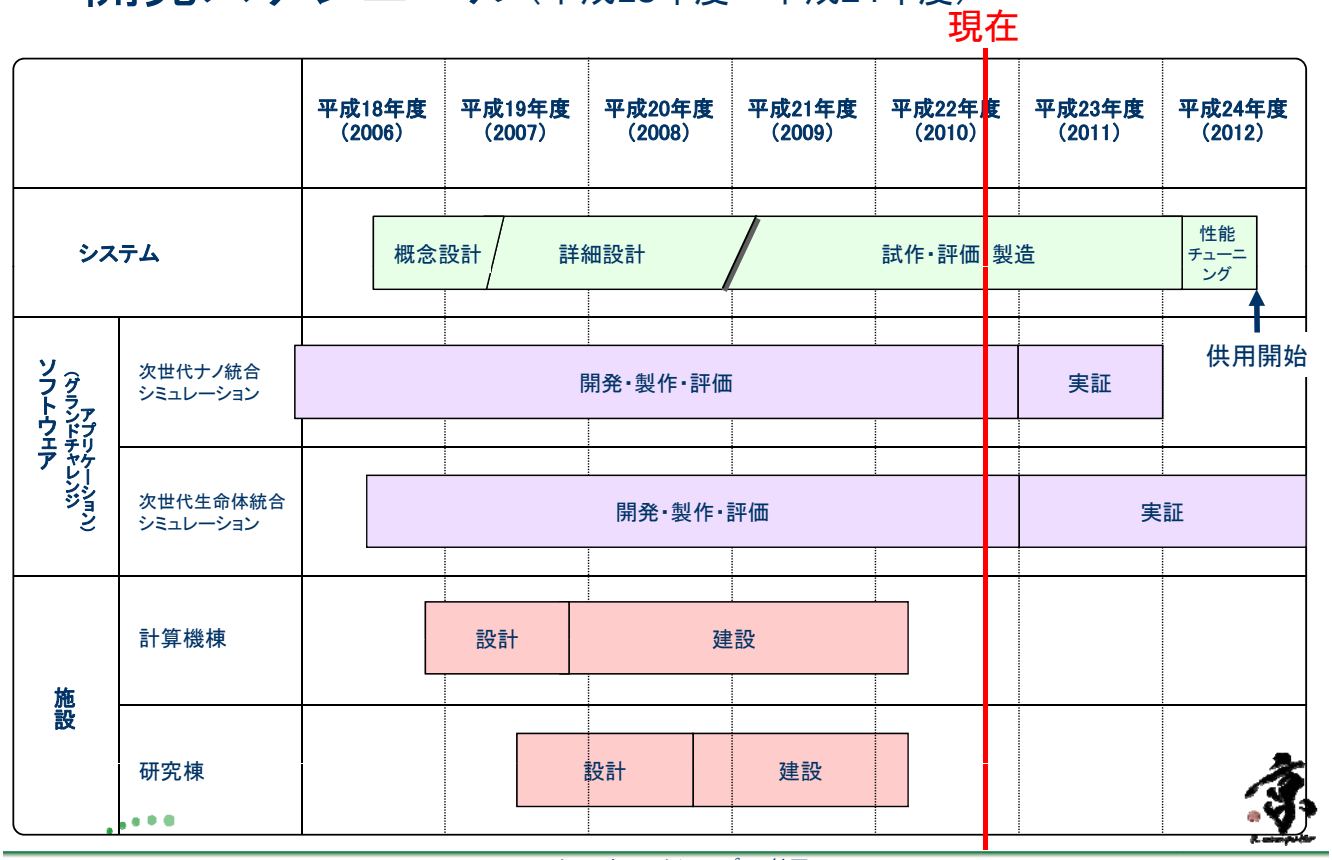
独立行政法人 理化学研究所
次世代スーパーコンピュータ開発実施本部
計算科学研究機構 運用技術部

黒川 原佳 <motoyosi@riken.jp>

Outline

- プロジェクト・スケジュール
- 京速コンピュータ「京」の概要
  - システム概要
    - ハードウェア・システム構成
    - 利用環境
    - ソフトウェア（システム・ツールなど）
- 計算科学研究機構 施設の概要

開発スケジュール（平成18年度－平成24年度）



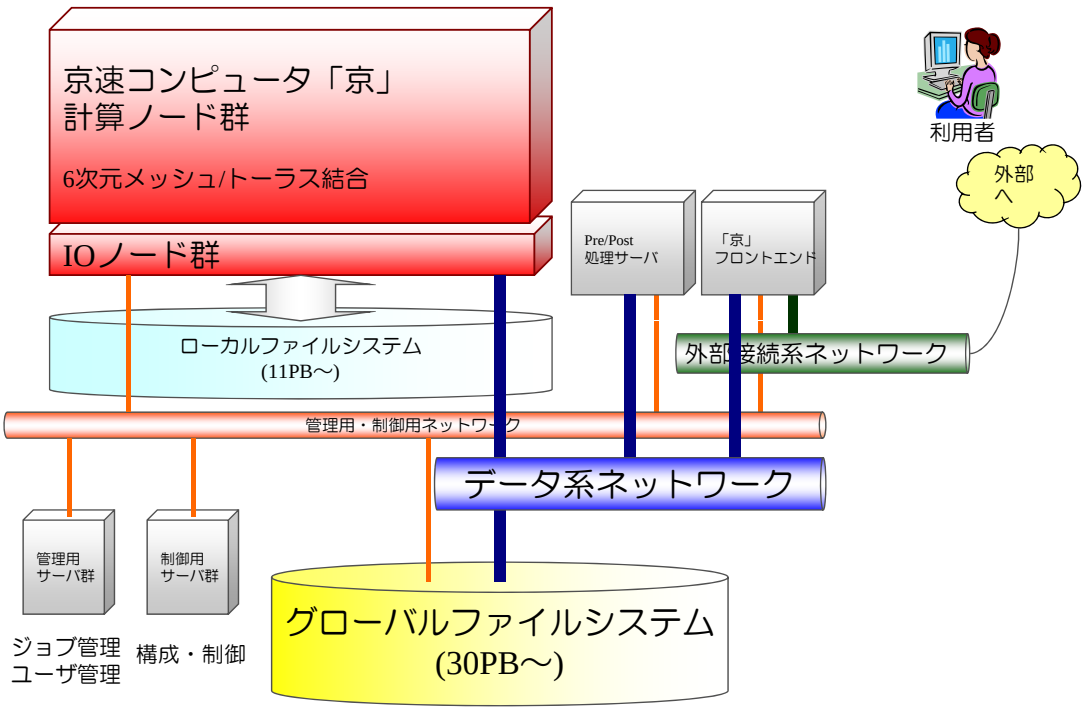
京速コンピュータ「京」の概要

※本発表内容は2011年2月1日時点のものであり，
実際に供用が開始される時点で変更されている
可能性（特にソフトウェア関連）があります．

システム全体概要

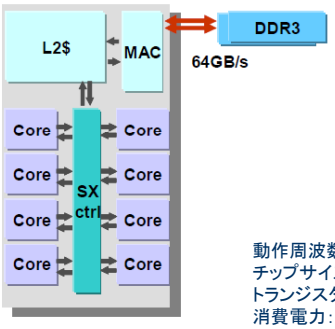
- ターゲット性能
  - LINPACK 10PFLOPS
- 多くのアプリケーションに高い実効性能を提供
  - ハイパフォーマンスコンピューティング向けCPUを採用
    - マルチコアCPU（8コア），演算アクセラレータ付き（SIMD機構），FPレジスタ数増強
  - ペタフロップス級のアプリケーション実効性能を目指す
- 高性能・低消費電力CPU
  - 2.2GFLOPS/W，ワット当たり世界トップレベル
    - 45nm CMOS プロセス
    - 小規模システムとして829MFLOPS/W（GREEN500 4位、TOP500 170位）
- 高信頼性システム
  - 「壊れにくい」,「壊れても全てが止まらない」.「壊れた部分はすぐ直す」
- 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
    - デファクトスタンダードの開発環境を提供

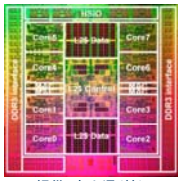
システム構成概要



プロセッサ構成

- スーパースカラ・マルチコアプロセッサ
  - 128GFLOPS（16GFLOPS/コア）
    - 8コア
  - 各コア256の浮動小数点レジスタ
  - SIMD拡張（積和演算器2個 x 2）
  - 共有2次キャッシュ（6MB，12way）
- データ供給能力
  - レジスタ-L1キャッシュ間：4B/FLOP
  - L1キャッシュ-L2キャッシュ間：2B/FLOP
  - L2キャッシュ-主記憶間：0.5B/FLOP
- 拡張機能
  - ハードウェアバリア機構
    - コア間の高速バリア同期が可能
  - セクタキャッシュ機能
    - キャッシュをプログラマブルに使うための機能





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）

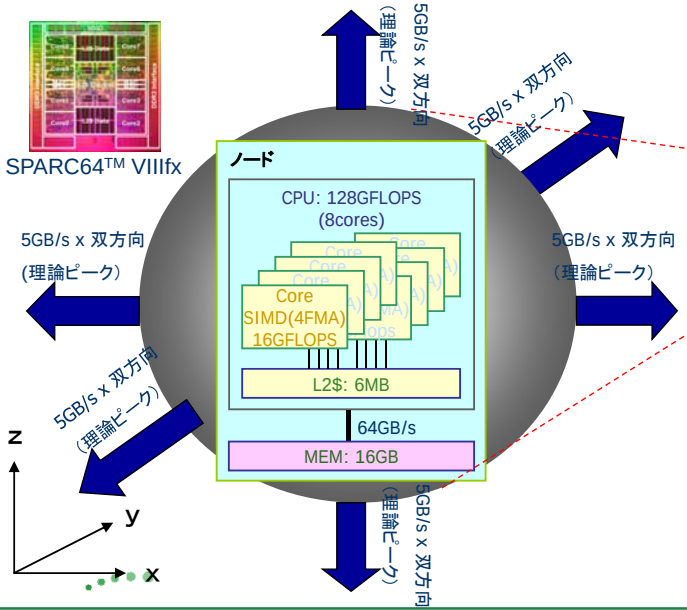
	仕　様
CPU性能	128GFLOPS（16GFLOPSx8コア）
コア数	8個
浮動小数点演算器構成（コア当り）	積和演算器：2×2個（SIMD） （逆数近似命令：SIMD動作） 除算器：2個 比較器：2個
	浮動小数点レジスタ(64ビット)：256 グローバルレジスタ(64ビット)：188
キャッシュ構成	1次命令キャッシュ：32KB(2way) 1次データキャッシュ：32KB(2way) 2次キャッシュ：6MB(12way)コア間共有
メモリバンド幅	64GB/s(0.5B/F)

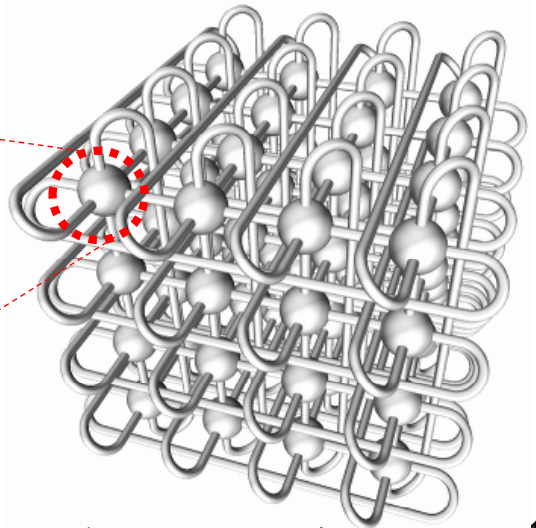
より詳細な情報は，「SPARC64™ VIIIfx Extensions」を参照のこと
http://img.jp.fujitsu.com/downloads/jp/jhpc/sparc64viiifx-extensions.pdf

計算ノード群の構成

- 計算ノード数（CPU数）：8万以上
  - コア数：64万以上
- ピーク演算性能：10PFLOPS以上
- メモリ総容量：1PB以上（ノード当り16GB）
- ネットワーク：6次元メッシュ／トーラス(Tofu)
  - リンク数は最大10
  - ユーザービューは3次元トーラス
  - 帯域：リンクにそれぞれ5GB/s x 2（双方向）【理論ピーク】
  - ケーブル：約200,000本，約1,000km

SPARC64™ VIIIfx

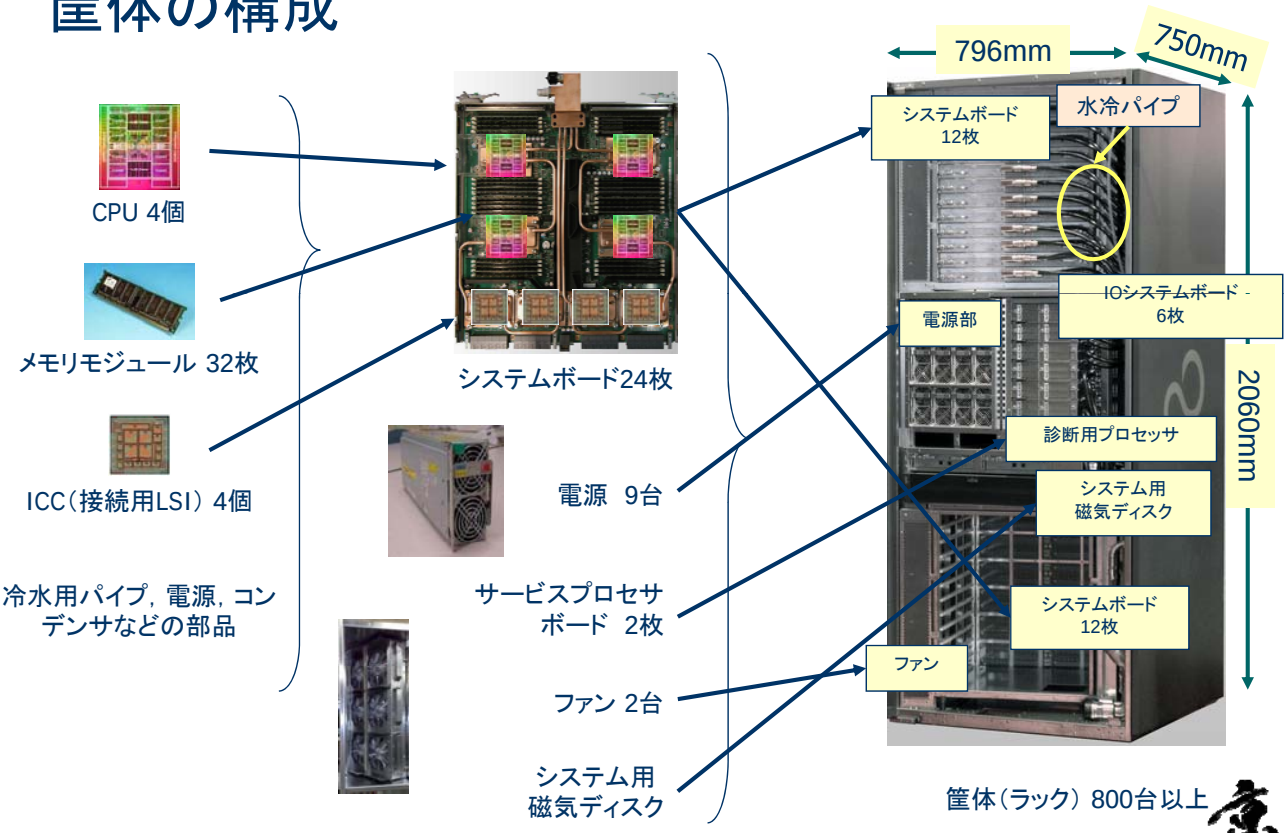




3次元トーラスのイメージ

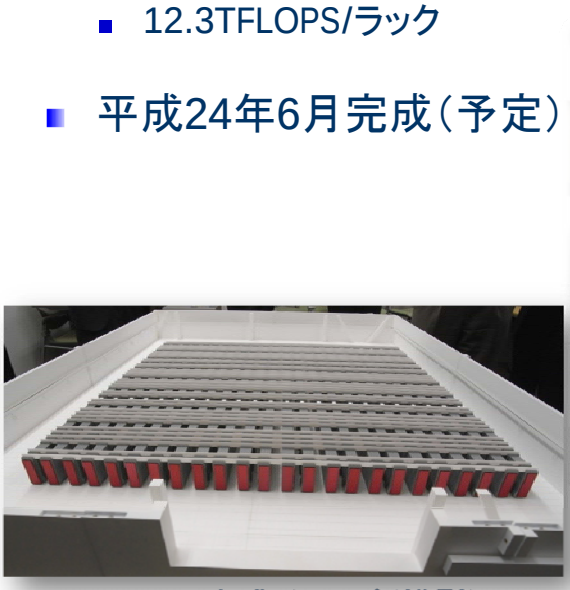
提供：富士通（株）

筐体の構成



平成22年10月1日設置開始

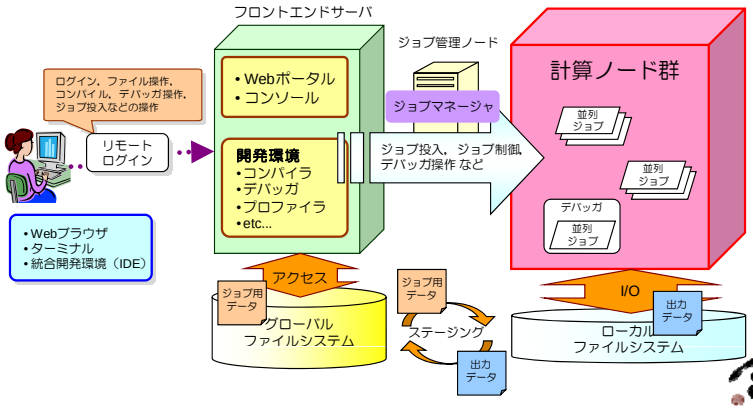
- ピーク性能 98.3TFLOPS（8台）
  - 12.3TFLOPS/ラック
- 平成24年6月完成（予定）



システム完成イメージ（模型）
800台以上の筐体が並ぶ

システム利用環境概要

- フロントエンドサーバ
  - Linux（RHEL）上で開発環境を提供
  - ターミナルとWebブラウザ（およびWebサービス）のUIを提供
- ステージングによる2階層のファイルシステム
  - 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）
    - 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）
- バッチジョブを主体としたジョブ実行環境
  - デバッグ用に会話型環境を用意（予定）



ジョブ実行シナリオ

- グローバルファイルシステム上にファイルを用意
  - 外部から転送の場合もあり
  - 必要に応じてPre/Postサーバなどでデータを処理
- フロントエンド（IDE，ポータル，CUI）からジョブスクリプトを投入
- フロントエンドからジョブを操作
  - 状態監視，キャンセルなど
- 実行結果を取得
  - 必要に応じてPre/Postサーバなどでデータを処理
- 必要に応じて外部へ転送

システムソフトウェア環境

- 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）を目指す
    - パフォーマンスカウンタI/FとしてPAPIも提供
    - IOノード
      - PCI-ex を有し、Tofu, GbE, FC, IBの各インターフェイスを搭載
- ファイルシステム
  - Lustreをベースとした拡張ファイルシステムを採用
    - GFS・LFS供にLustreベースのファイルシステム
      - 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

ジョブ実行環境

- ジョブ実行環境
  - 様々なジョブの実行が可能
    - ハイブリッド・フラットMPI、バルクなどが実行可能
      - フラットMPIは実行環境として推奨しません
    - MPIジョブの形態としてSPMD、MPMD、などが実行可能
      - 実行に際して様々な制約はあります
  - 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
  - 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
- ノード割付
  - 全てのジョブにVirtual 1~3D Torus空間を提供
    - ジョブスクリプトに1~3次元形状を指定
      - Rankマッピングは様々な方法が可能
        - Rank番号をXYZで自動マッピングやhostfile的なユーザ指定など
    - IO占有を指定
      - IO性能を確保したいジョブのために指定可能
        - Z軸に2つ以上のジョブを混ぜない指定が可能

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
    - 低レイテンシかつ広帯域
    - Bcast, AllReduce, AlltoallやAllgatherなどをTofuに適したアルゴリズムで処理

プログラム開発環境(2/2)

- 数値・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
- 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(ユーザ関数埋め込み)
  - 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
    - サンプリング情報取得の手段を提供
      - 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
  - トレーサー
    - 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

計算科学研究機構 施設の概要

計算科学研究機構

東京から西に450km

兵庫県神戸市中央区港島南町7丁目（ポートアイランド第2期内）
ポートアイランド南駅より徒歩約3分，JR新神戸駅から25分

平成22年5月31日完成

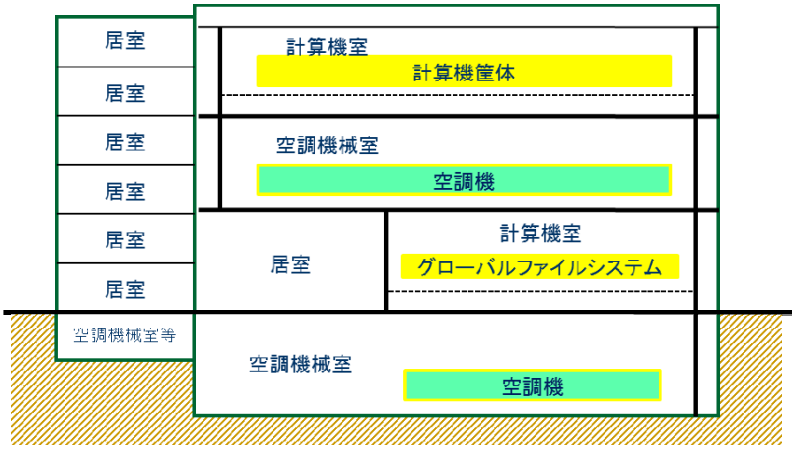
施設の概要

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



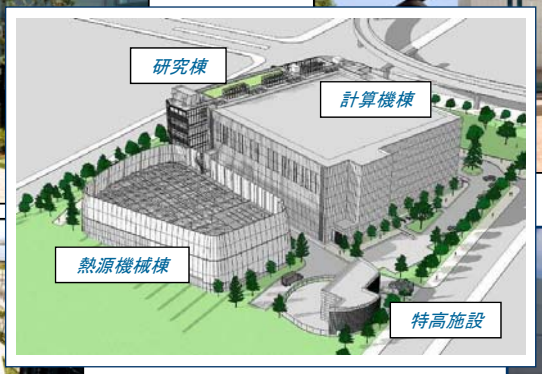
熱源機械棟（面積 1900m^2）

- 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）

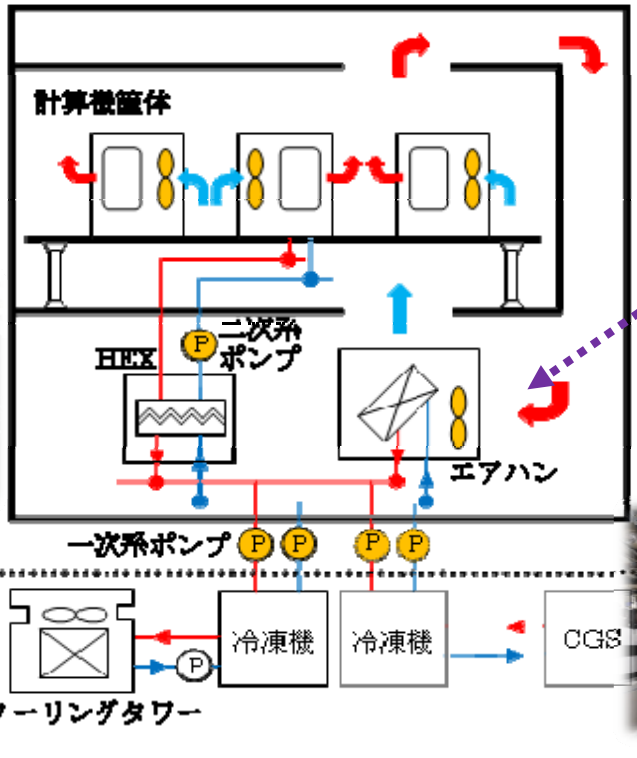
- 30MW 77,000V（受電） → 6,600V

建物の配置



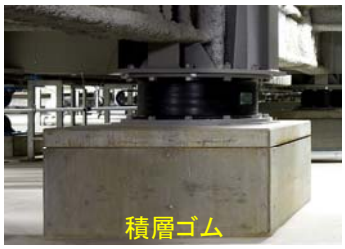
冷却システム

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．
    - 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
    - 「京」筐体当り最大1,500kg

フリーアクセス架台

計算機室（3F）

研究スペース