世界最高性能を目指すシステム開発
ー 次世代スパコンのシステム構成と施設の概要 ー

平成22年3月2日

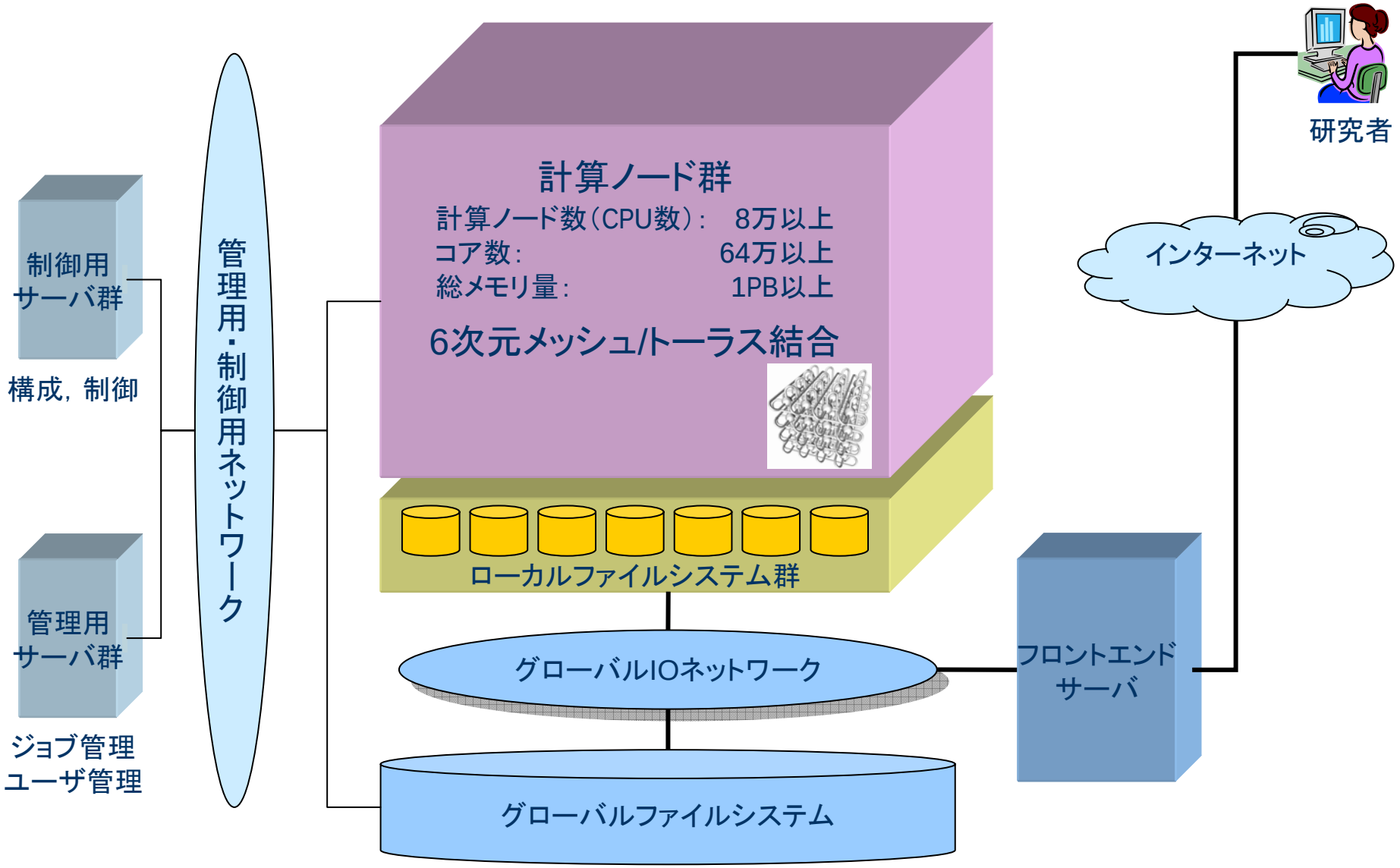
理化学研究所
次世代スーパーコンピュータ開発実施本部
横川 三津夫

高性能かつ大規模システムの課題と対応

- 演算性能の向上
  - CPUのマルチコア化，SIMD（ベクトル化）機構
- 主記憶へのアクセス頻度の削減
  - CPU性能とメモリアクセス性能のギャップ（メモリウォール）
  - レジスタ数増，ソフトウェア制御可能なキャッシュ（セクタキャッシュ）の導入
- 消費電力の削減
  - CPUの適切な動作周波数の選択
  - 直接ネットワークの採用
- 実運用に耐えられる安定動作可能なシステムの提供
  - ECC機構などエラー修正に考慮したシステム設計
  - 単一障害を回避する冗長性あるネットワーク構成
  - 使いやすい利用環境の提供

次世代スーパーコンピュータの概要

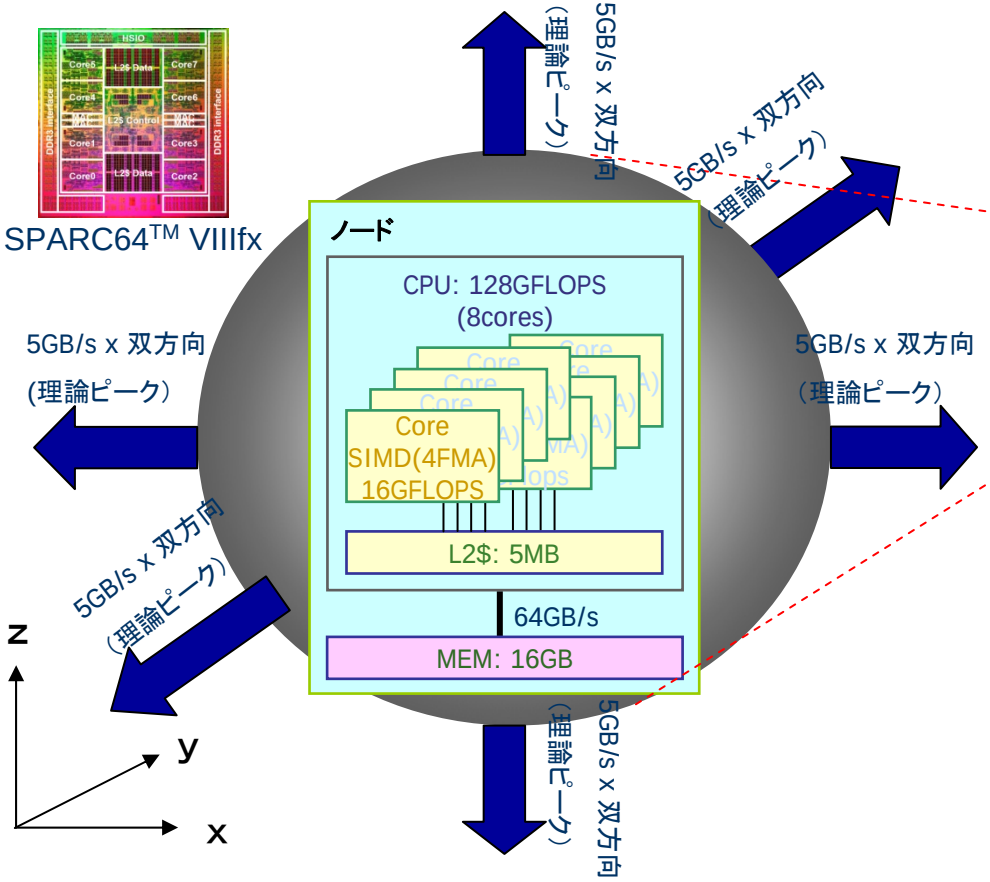
システム構成概要

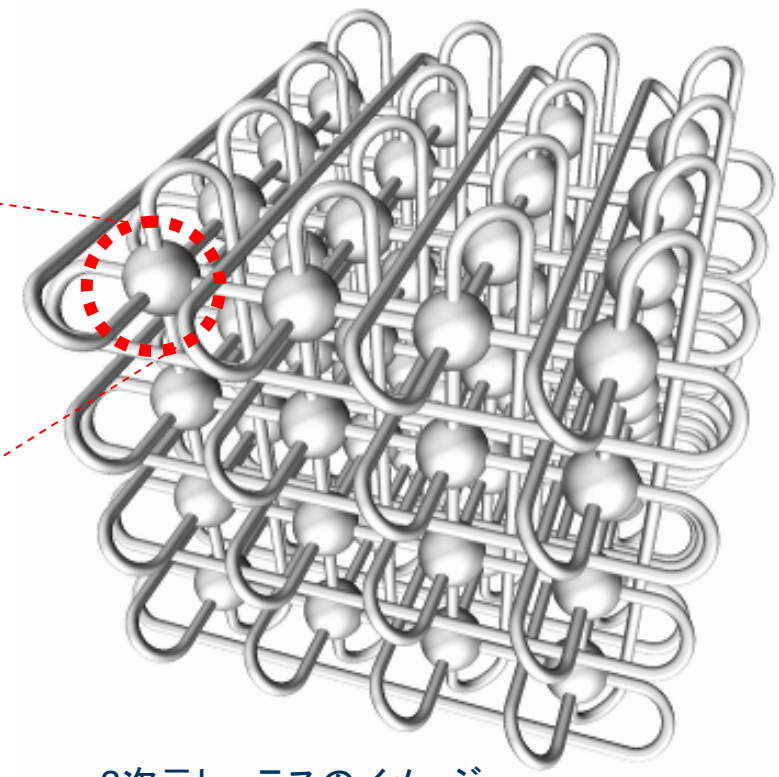

計算ノード群
計算ノード数(CPU数): 8万以上
コア数: 64万以上
総メモリ量: 1PB以上
6次元メッシュ/トーラス結合
ローカルファイルシステム群
グローバルIOネットワーク
グローバルファイルシステム
管理用・制御用ネットワーク
制御用サーバ群
構成，制御
管理用サーバ群
ジョブ管理
ユーザ管理
フロントエンドサーバ
インターネット
研究者

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:ユーザービューは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1200km



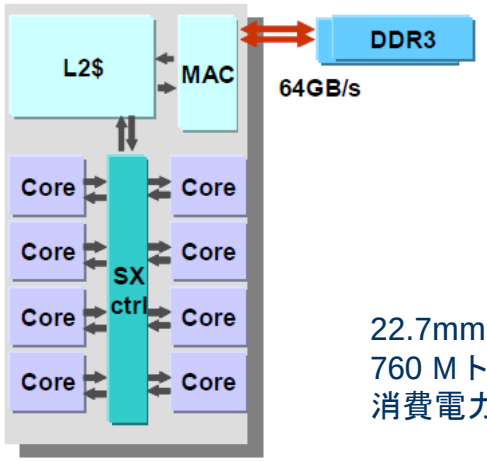


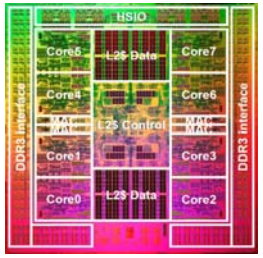
3次元トーラスのイメージ

提供:富士通(株)

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
- コア共有の2次キャッシュ（5MB，10way）
  - ハードウェアバリア機構
  - プリフェッチ機構
  - セクタキャッシュ機能（次ページ）
- 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





提供：富士通（株）

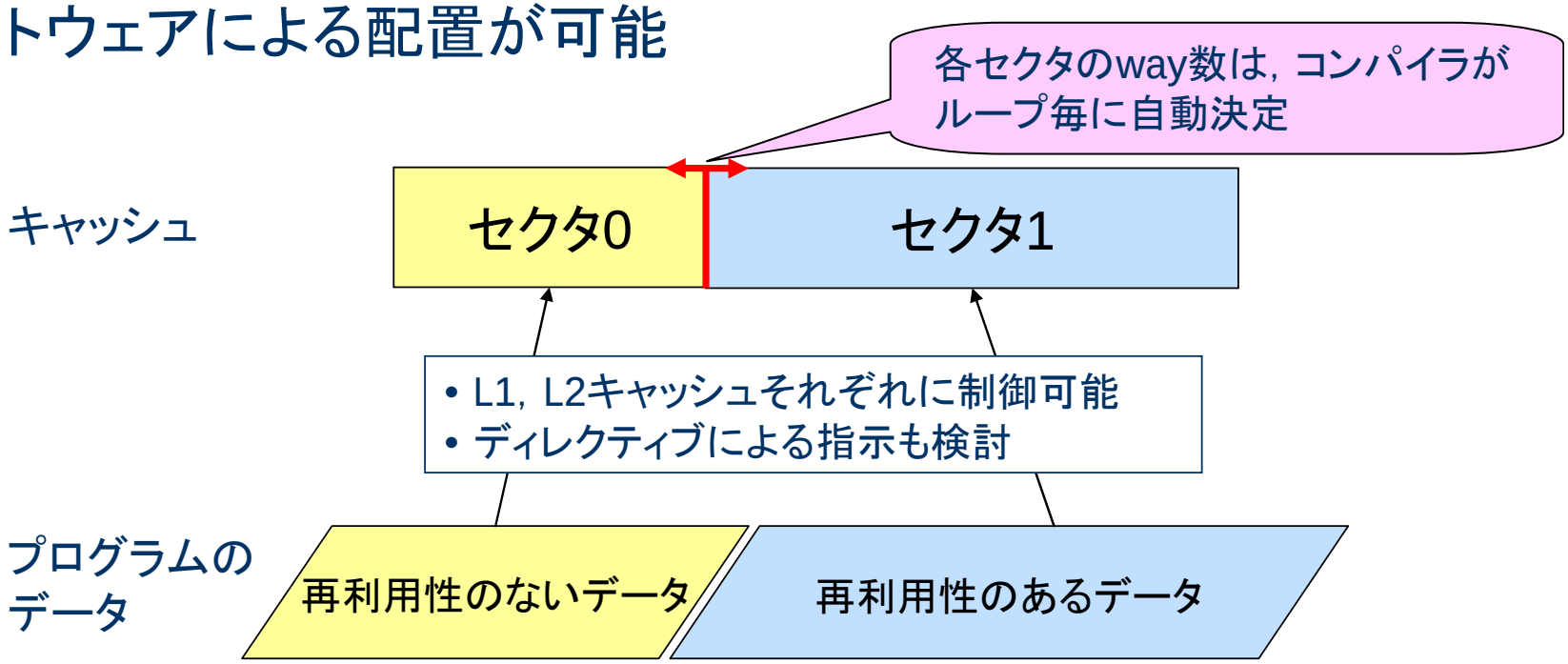
22.7mm x 22.6mm
760 M トランジスタ
消費電力：58W（水冷，30℃時）

	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 ビジュアル演算器：1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：5MB(10way)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

セクタキャッシュとは？

- 再利用性のあるデータを選択的にキャッシュに残す仕組み
- ソフトウェアによる配置が可能

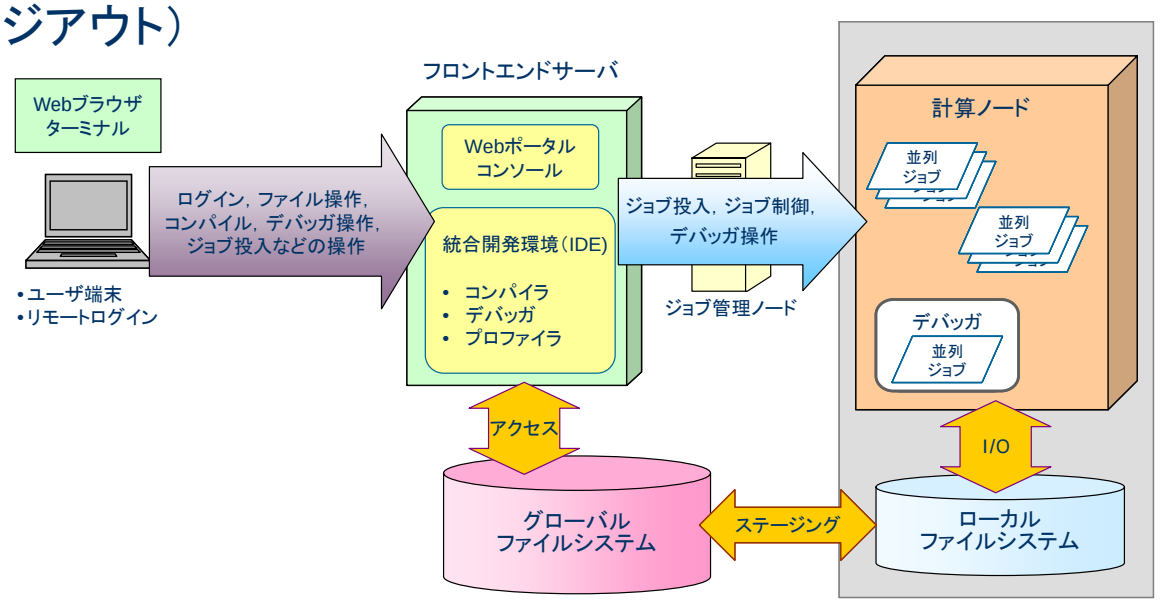


- 再利用性のあるデータを含むプログラム例

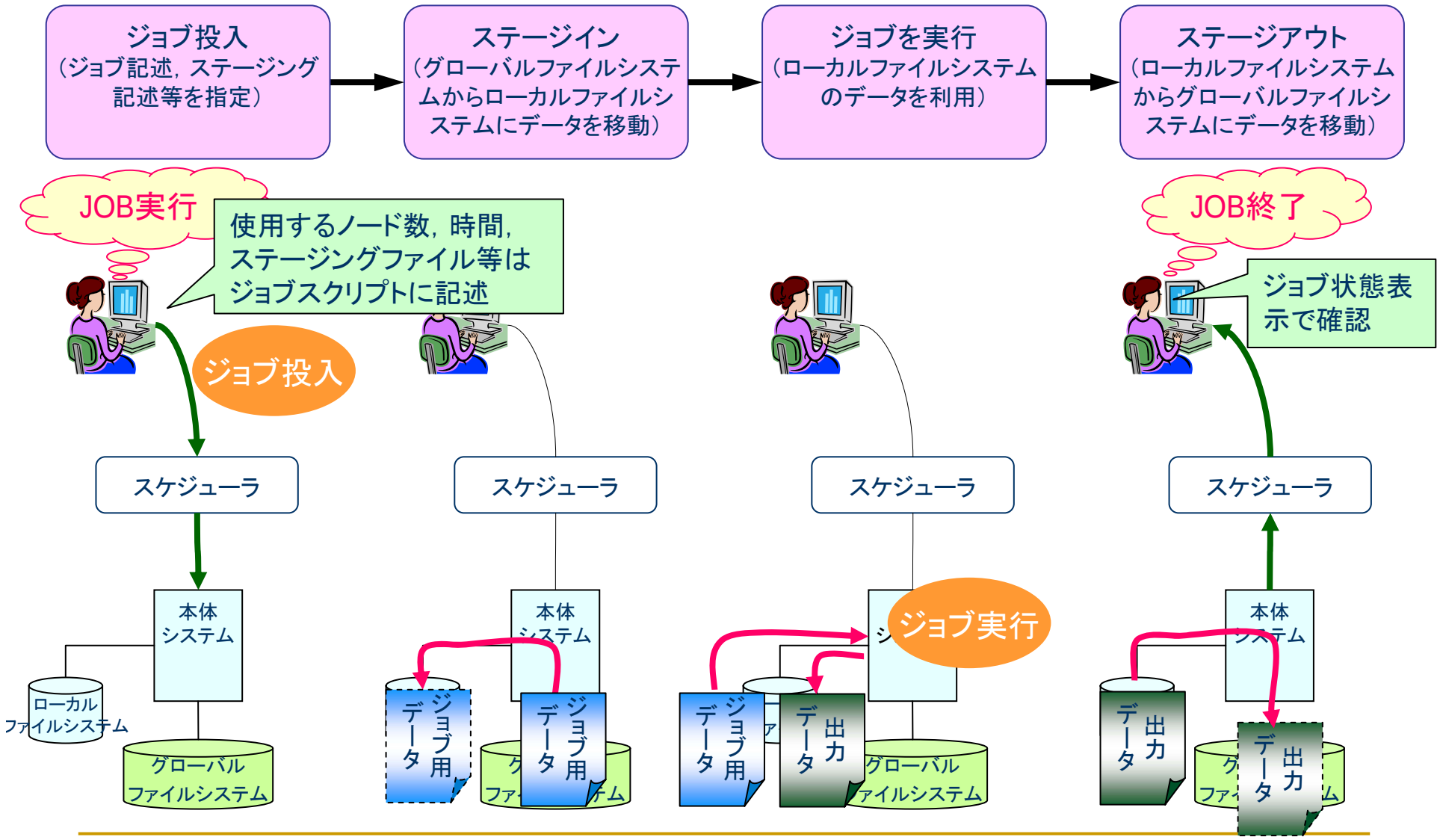
```
do j=1,n
 do i=1,n
   a(i) = a(i) + b(i,j)
 enddo
enddo
```

システム利用環境

- OS：Linuxをベースとしたオペレーティングシステム
  - POSIX規格に準ずるコマンド群を提供
- 大規模分散ファイルシステム（2階層のファイルシステム）
  - ファイルステージング機能
    - ジョブ実行前にグローバルファイルシステムからローカルファイルシステムへファイルを転送（ステージイン）
    - ジョブの出力ファイルをローカルファイルシステムからグローバルファイルシステムへ転送（ステージアウト）
  - ファイル共有機能
- 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）



バッチジョブ実行時の処理の流れ



プログラム言語，コンパイラ

- Fortran 2003，XPFortran，C，C++
- GNU C/C++拡張仕様
- 4倍長精度演算をサポート：IEEE754R及びdouble-double形式

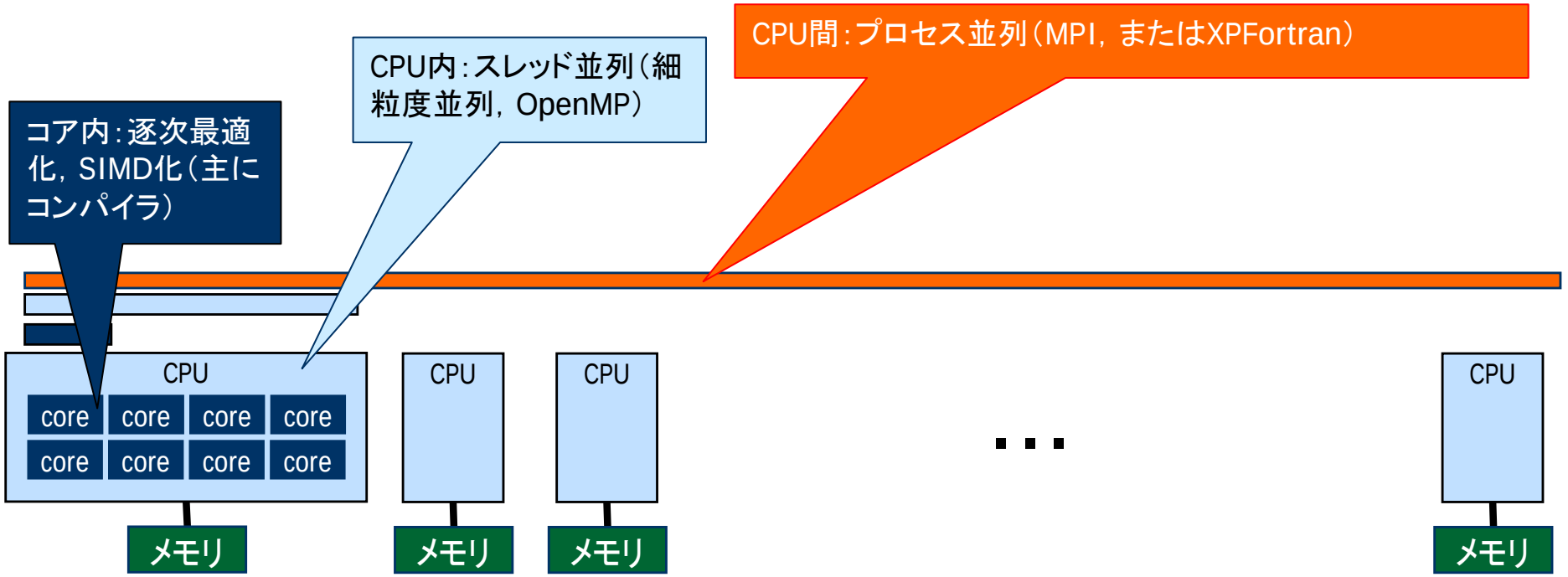
- SPARC64™ VIIIfxの機能を有効活用するコンパイラ機能
  - SIMD機構の活用
    - 自動ベクトル化を応用したSIMD命令の自動生成
    - IF文を含むループのSIMD化（マスク付きSIMD化）
  - 大容量レジスタ（倍精度浮動小数点 256本）の有効活用
- セクタキャッシュの利用
  - セクタキャッシュを考慮したプリフェッチ命令の自動生成
  - セクタキャッシュをユーザが意識して利用するためのディレクティブ
- 自動並列化
  - マルチスレッド化，パイプライン並列化機能

ライブラリ及びプログラム開発支援環境

- MPIライブラリ（MPI-2.1に対応）
  - 低レイテンシ・高スループットの実現
  - トポロジ構成を意識した集団通信関数を提供
    - Bcast /Allgather /Alltoall /Allreduce
  - インターコネクトのハードウェアバリア機構を用いたハードバリア/リダクション演算への活用
- 数値計算/科学技術計算ライブラリ
  - システムにチューニングされたBLAS，LAPACK，SSL II（富士通製科学技術計算用ライブラリ），FFTWを提供
- 開発支援ソフトウェア
  - デバッガ：DWARF2対応
  - 性能解析ツール：デバッグツール，プロファイラ，MPIトレーサ等の連携

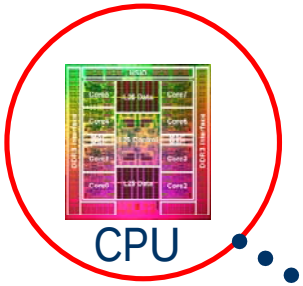
プログラミングモデル

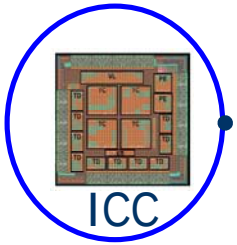
- スレッド並列＋プロセス並列のハイブリッド型を推奨
  - コア内： コンパイラによる逐次最適化，SIMD化
  - CPU内： スレッド並列（自動並列化，OpenMP）
  - CPU間： プロセス並列（MPI，XPFortran）
- フラット型も可能



システム開発の状況

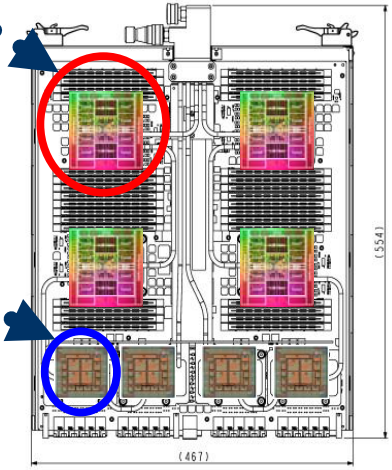
- LSI開発（45nm半導体プロセス）
- 試作機が完成．ハードウェア及びソフトウェアの試験を実施中．

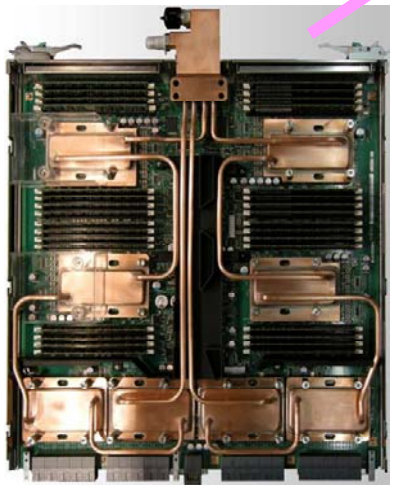


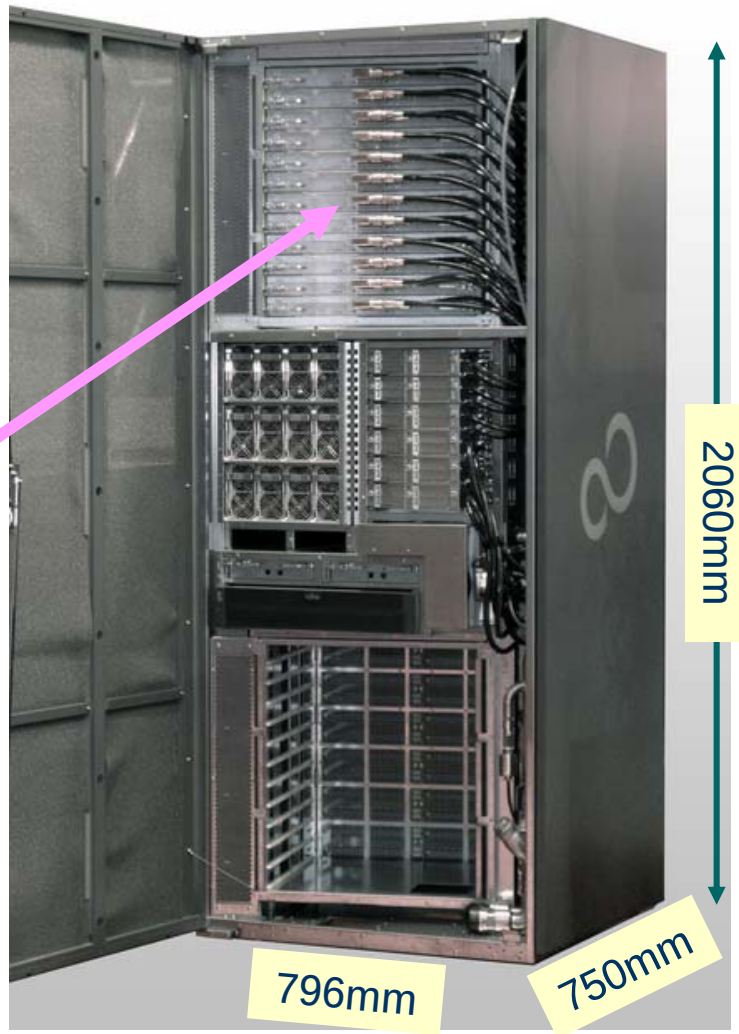


トーラス・ネットワーク用LSI

システムボード







提供：富士通（株）

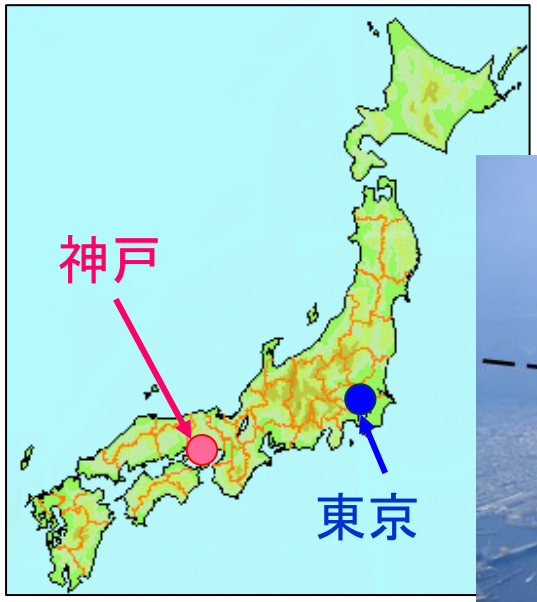
プロセッサ及びシステムの比較

ベンダ	チップ名	プロセスルール(nm)	理論性能(GFLOPS)	キャッシュ容量(MB)	消費電力(W)	ワット当たりの性能
Fujitsu	SPARC64VIIIfx	45	128.00	5	58	2.21
IBM	Power7	45	256.00	32	200	1.28
Intel	Xeon W5590	45	53.28	8	130	0.41
AMD	Opteron 8439SE	45	67.20	9	105	0.64

ベンダ	システム名	Linpack性能(PFLOPS)	CPU数	ネットワーク構成	備考
Fujitsu	次世代スパコン	10(目標値)	8万以上	3次元トーラス	2012年完成予定
IBM	BlueWaters	6-8?	2万5千以上	ツリー?	2011年完成予定
IBM	Sequoia (BlueGene/Q)	20(理論性能)	10万以上	3次元トーラス	2011年完成予定
Cray	XT5(Jaguar)	1.76	3万5千以上	3次元トーラス	2009年11月度世界一

次世代スーパーコンピュータ施設について

次世代スーパーコンピュータ施設



450km (280miles)
west from Tokyo

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1分，JR新神戸駅から25分

建屋イメージ

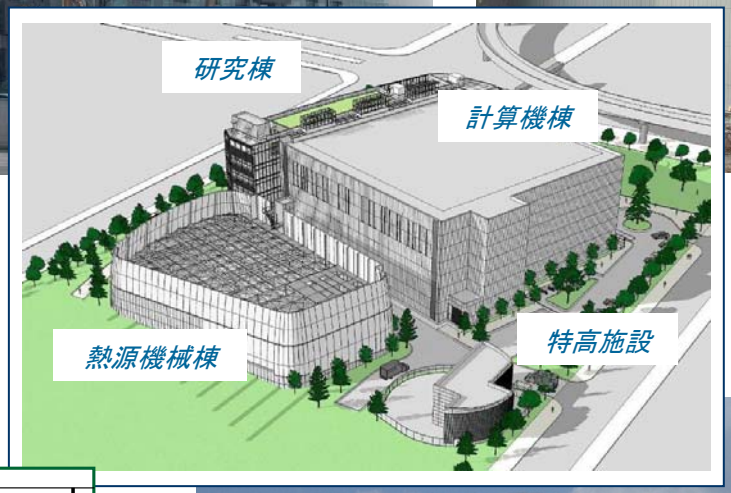

正面

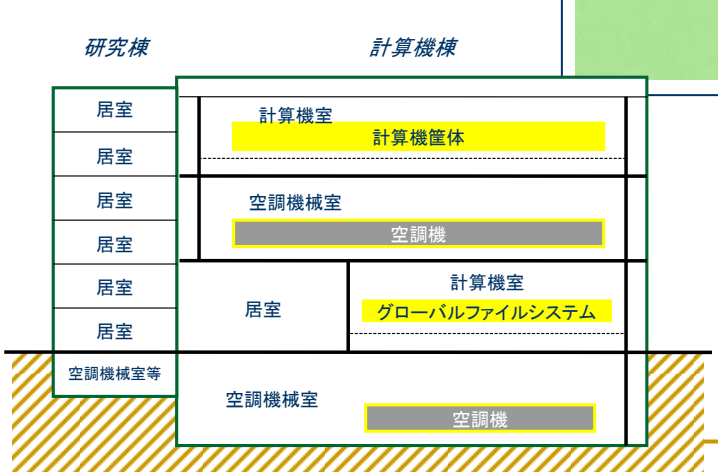

北西から研究棟側を見る

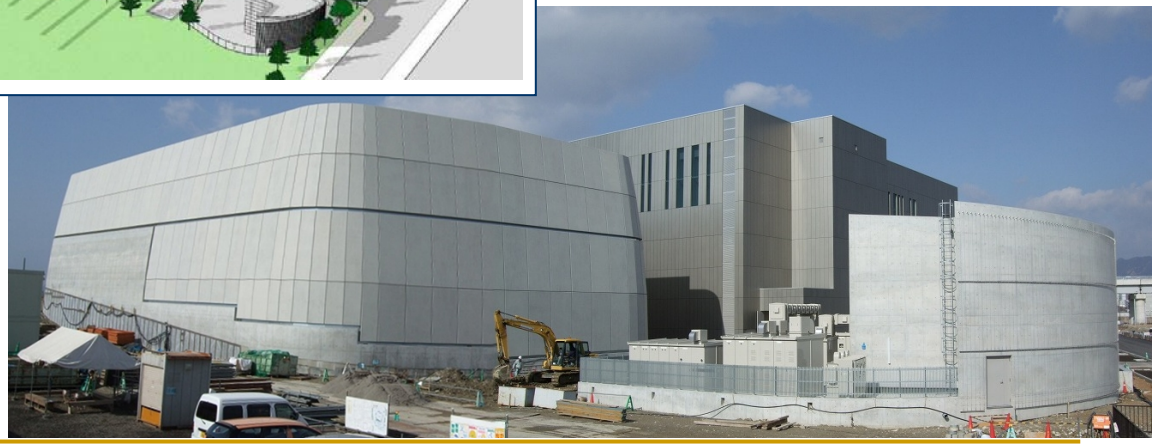
【計算機棟】
- 延床面積　約10,500㎡
- 建築面積　約 4,300㎡
- 構　　造　鉄骨造り地上3階地下1階

【研究棟】
- 延床面積　約9, 000㎡
- 建築面積　約1, 800㎡
- 構　　造　鉄骨造り地上6階地下1階

施設の建設風景（平成22年2月2日）



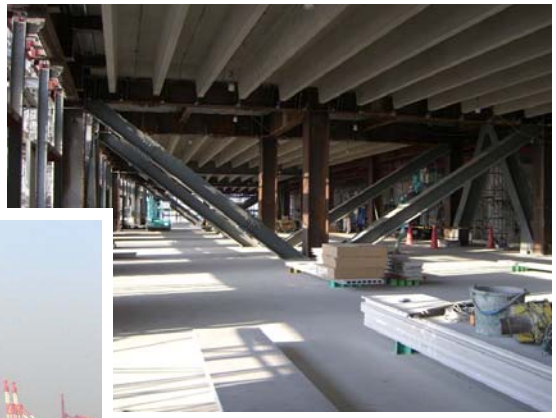




施設内部

次世代スパコン設置フロア（計算機棟3階）

フリーアクセス架台

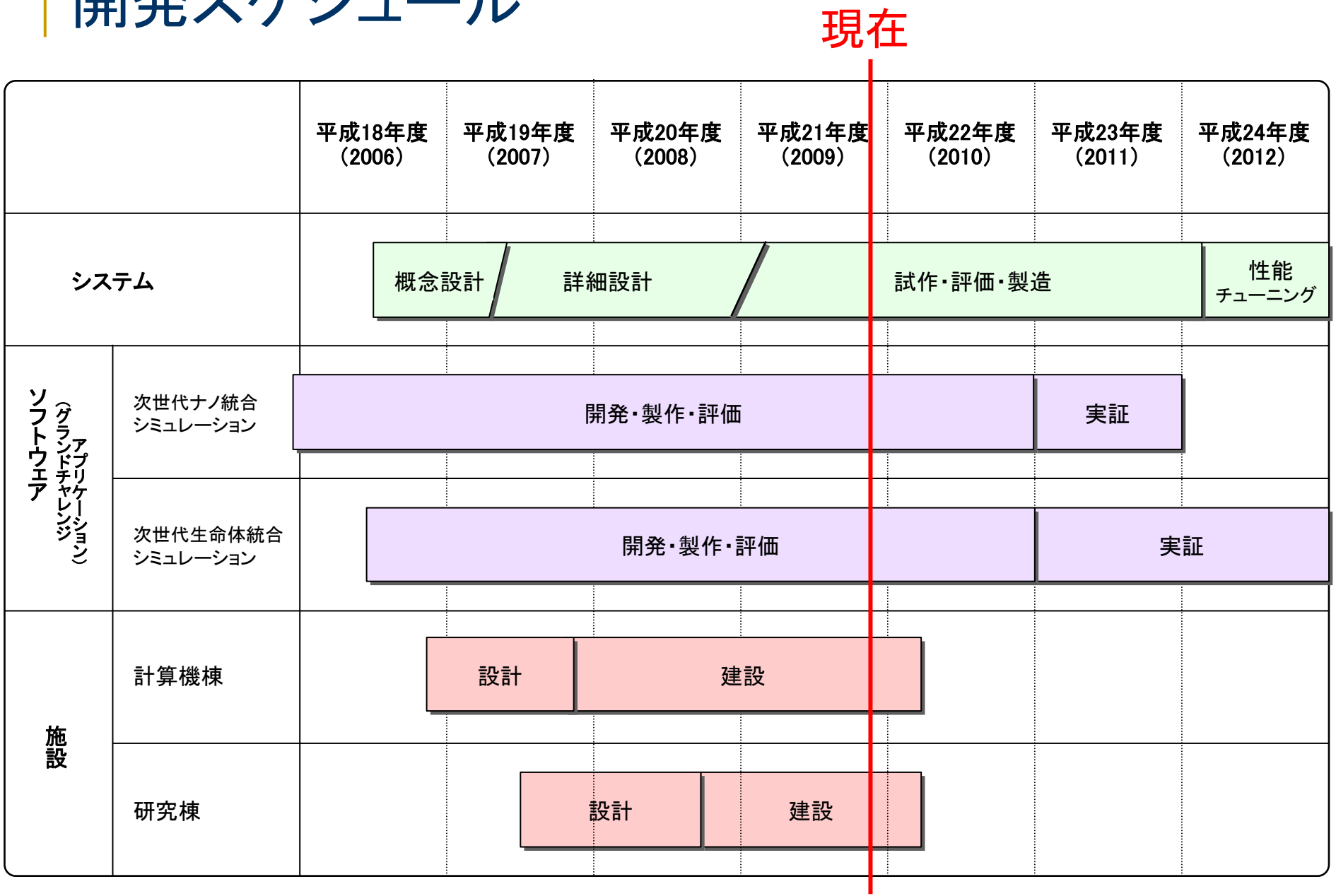


研究棟

冷凍機（熱源機械棟）

太陽光発電パネル（屋上）

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
ソフトウエア（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