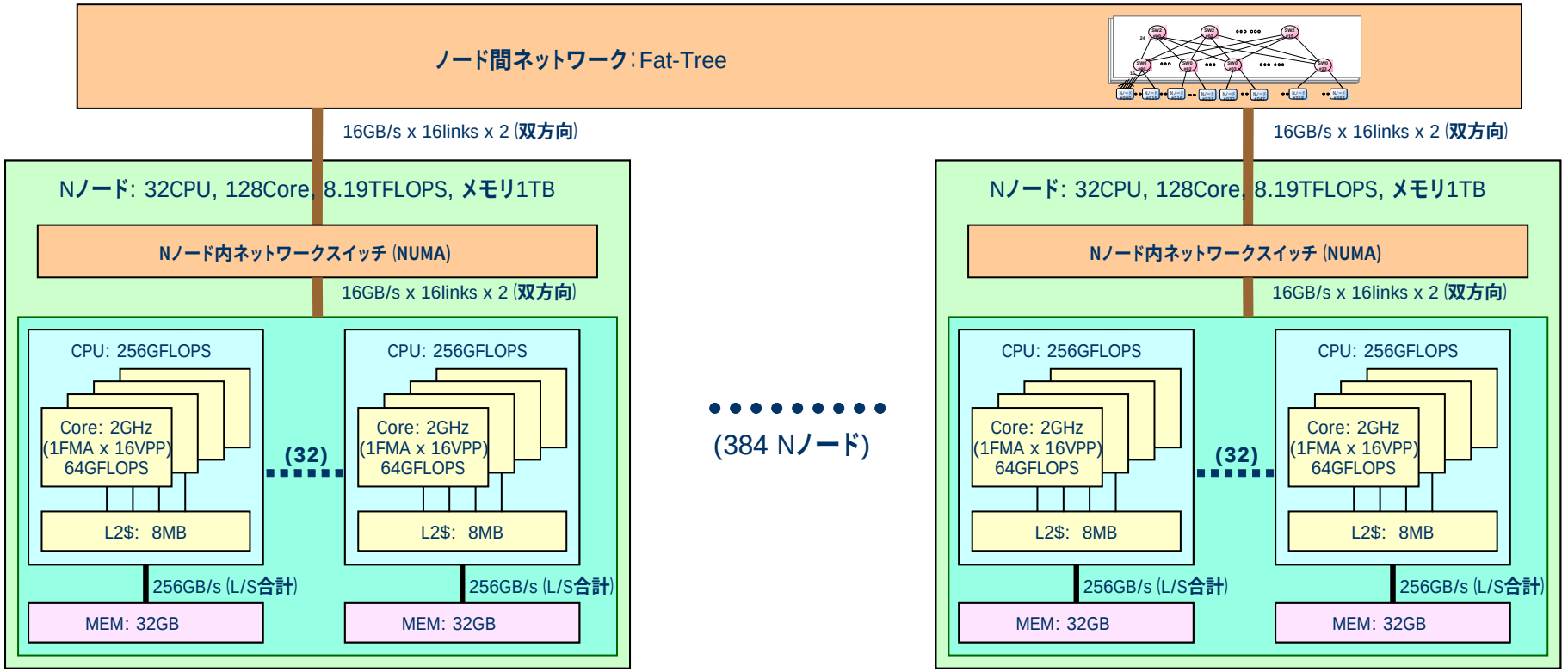
ベクトル部 (ユニットB) の構成

ベクトル部の構成図

- **計算ノード数**：12,288（384 Nノード）
  - CPU**数**：　12,288
  - **コア数**：　49,152
- **ピーク演算性能**：　3.14PFLOPS
- **メモリ総容量**：0.375（**計算ノード当り**32GB）
- 2**段**Fat-tree**ネットワーク**：　(24 x 16) x 16**プレーン**
- **消費電力**：　**約**7MW（**周辺機器を含む**）
- **設置面積**：　**約**1070**m²**（**周辺機器を含む，柱部分含む**）

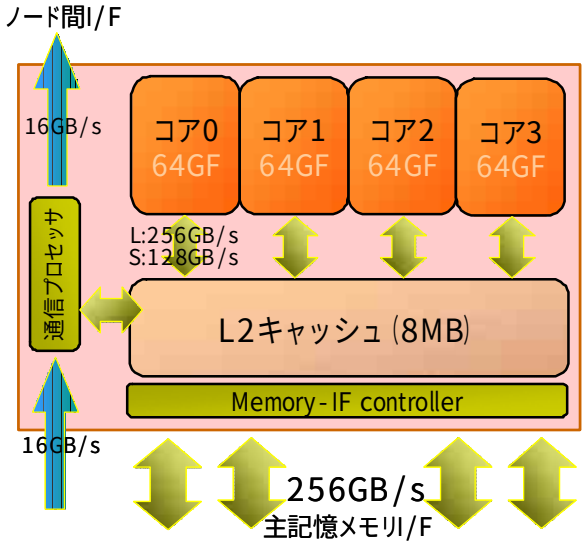


ベクトル部の特徴

- **プロセッサ**
  - 45nmプロセスによる1CPUチップ当り256GFLOPSの高性能演算器を実装
  - 1CPU当り4コア構成，動作周波数2GHzで駆動
  - コア当り1FMAx16セットの演算器と64本（256要素/本）のベクトルレジスタ
  - ソフトウェア制御可能なRDB（Reusable Data Buffering）機能を持つ8MBのL2キャッシュを4コアで共有
  - Nノード内の32CPUが論理的にメモリ空間を共有し，一つのOSで動作
  - 消費電力: 200W/CPU（Linpack実行時）
- **ネットワーク**
  - バイセクションバンド幅98TB/s（双方向），2段のFat treeで384 Nノードを接続
  - 光インターコネクトの採用
  - 非同期転送，同報機能，高速バリア同期機能付きのデータ転送機能
  - 入出力ポートの構成制御によるパーティショニング

プロセッサ構成

- 4演算コア，8MB共有L2キャッシュ，通信プロセッサを1チップ化
  - 動作周波数2GHz
  - コア当たり64本（256要素/本）の大容量ベクトルレジスタ
  - コア当り64GFLOPS，CPU当り256GFLOPS
- キャッシュ
  - 8MBのコア間共有L2キャッシュ
  - 各コアとのバンド幅は，ロード256GB/s，ストア128GB/s
  - 主メモリ間バンド幅は，ロード･ストア合わせて256GB/s
  - データ供給能力は，L2キャッシュから各コアのベクトルレジスタまで4B/FLOP，主メモリからL2キャッシュまでロード･ストア合わせて1B/FLOP
  - ソフトウェア制御可能な選択的データ･キャッシング（RDB（Reusable Data Buffering））機能

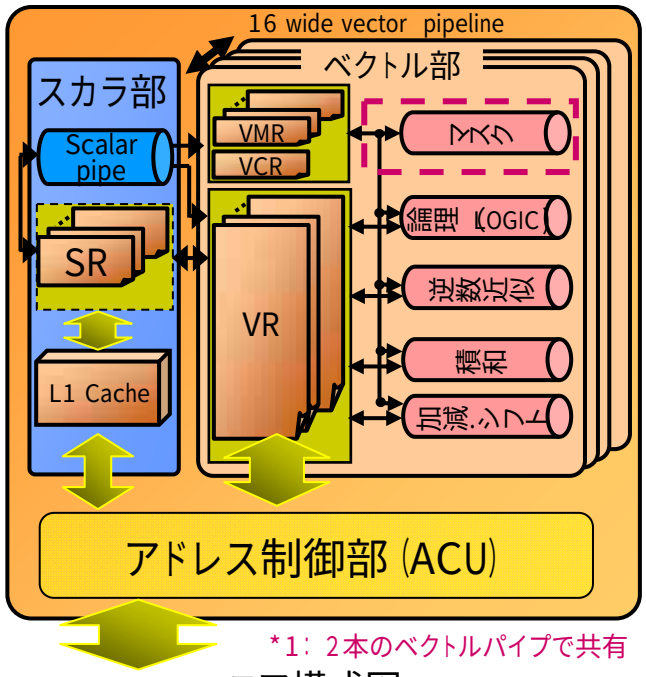


CPUチップ構成

	仕様
ピーク演算性能	256GF（64GFx4コア）（ベクトル･プロセッサ部のみ）
L2キャッシュ	8MB（8way-セットアソシアティブ），64B/ライン，4コア共有のUnifiedキャッシュ，選択的データキャッシング機能(RDB機能)あり
メモリバンド幅	256GB/s（ロード･ストア合わせて）
ノード間I/Fバンド幅	16GB/s（片方向のみ）

コア構成

- スカラ部
  - 128本の汎用レジスタ
  - 投機実行
  - 4wayスーパースカラ
- ベクトル部
  - 16セットの多重ベクトルパイプライン
  - 積和演算器x16個/コア
  - 論理演算器x16，逆数近似演算器x16，加減/ｼﾌﾄx16、マスク演算器x8，ロード/ストア・パイプラインx8/コア

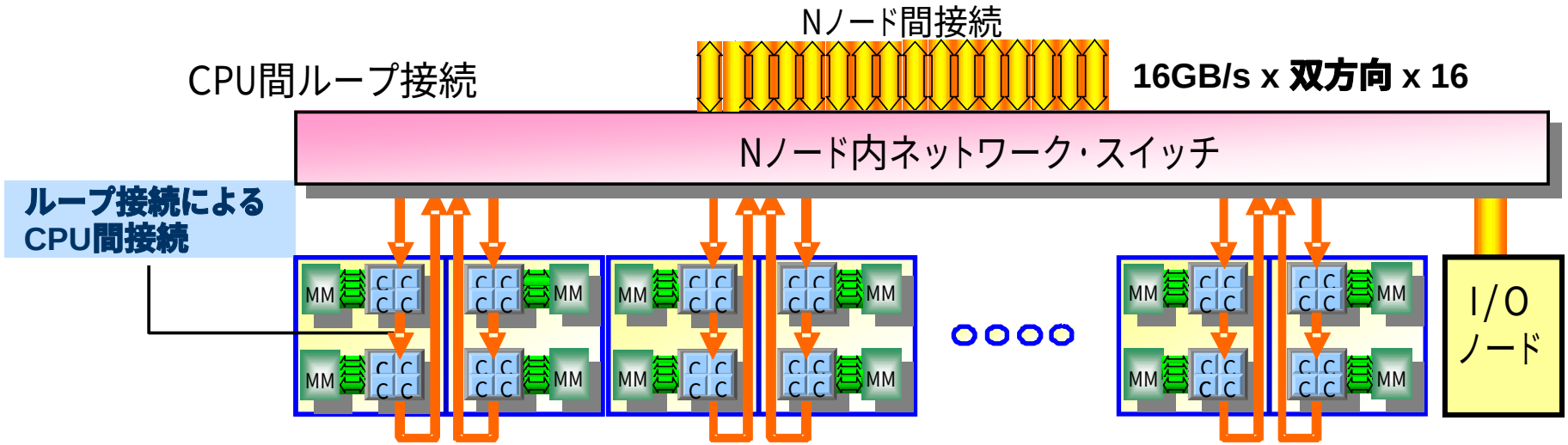


コア構成図

		仕様
ピーク演算性能		ベクトル：64GF，スカラ：4GF
動作周波数		2GHz (ベクトル／スカラ)
ベクトル部 (16VPP構成)		
	演算器構成	積和x16，論理x16，逆数近似x16，加減/ｼｽﾄx16、マスクx8，ロード/ストアx8
	レジスタ	VR：64本 (8Bx256要素/本) VCR:1本(256ビット/本) VMR：16本 (256ビット/本)
スカラ部		
	演算器構成	乗算x1，加算x1，除算x1，整数x2
	レジスタ	SR:128本 (8B/本)
	L1キャッシュ	命令：64KB (2wayｾｯﾄｱｿｼｱﾃｨﾌﾞ) データ：64KB (2wayｾｯﾄｱｿｼｱﾃｨﾌﾞ)

Nノード構成

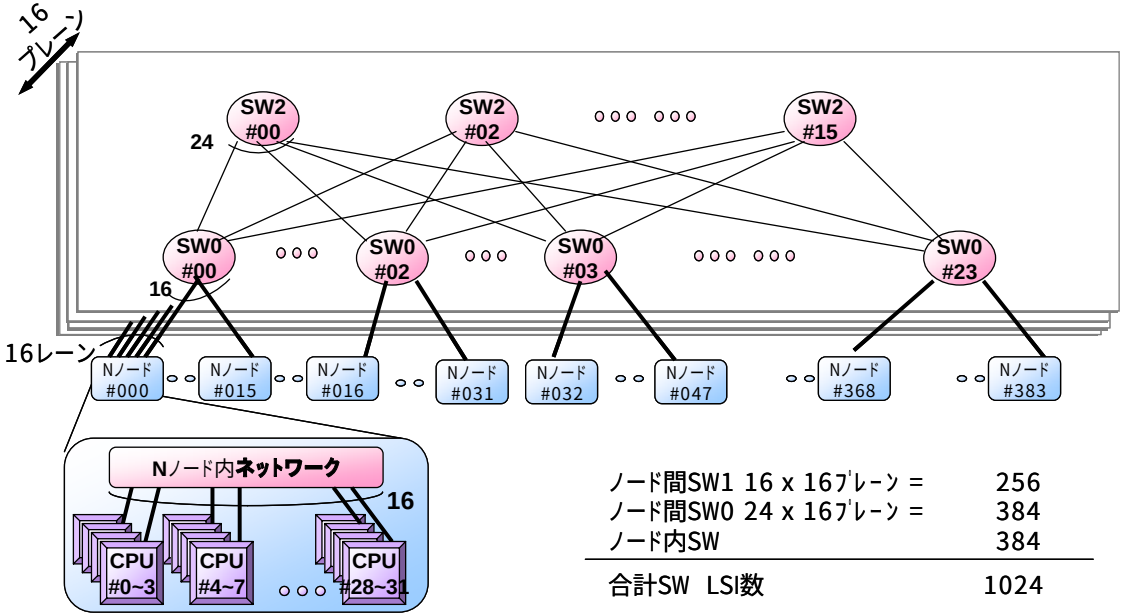
- 1個のCPUを1カードに搭載
- 32個のカード (32CPU) と1つのI/Oノードをネットワーク・スイッチで接続しNノードを構成
  - 2CPUが, Nノード用ネットワーク・スイッチ (33x33) にループ接続. 通信バンド幅は, 16GB/s x 2 (1接続当り)
  - 1つのI/Oノード (x86ベース)
    - Nノード内入出力処理, Nノード内計算ノード管理など



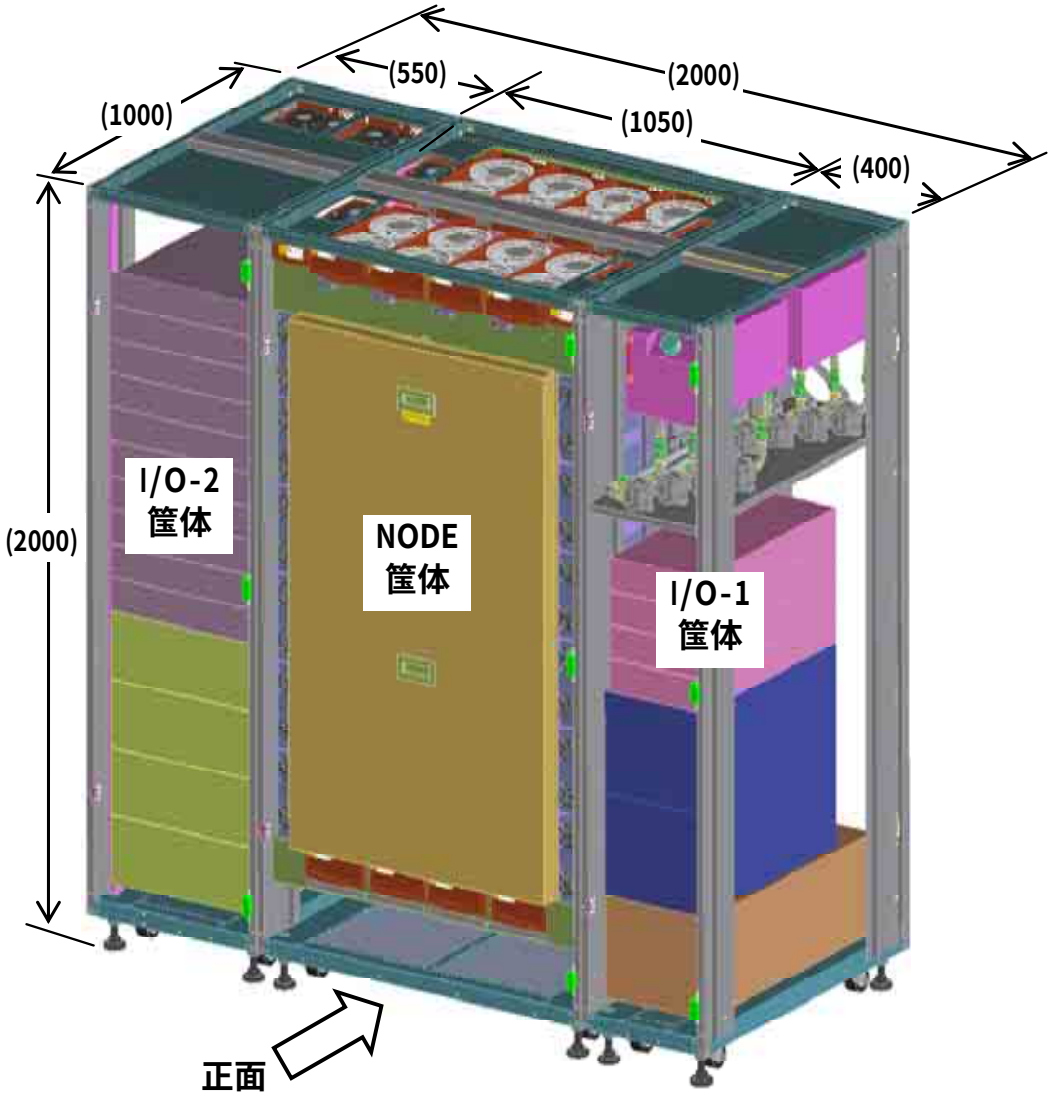
Nノード間ネットワーク

- 2段のFat-treeネットワーク構成
  - ポート当り16GB/s (双方向) の32×32スイッチ
  - スイッチ間の接続に，20Gbpsの光インターコネクト技術を採用
  - 各Nノードから出る16レーンの接続をレーン毎にプレーン構成とし，システム全体で16プレーン構成
  - (24 x 16) x 16プレーンで384 Nノードを接続．バイセクション・バンド幅 98TB/s (双方向) となる

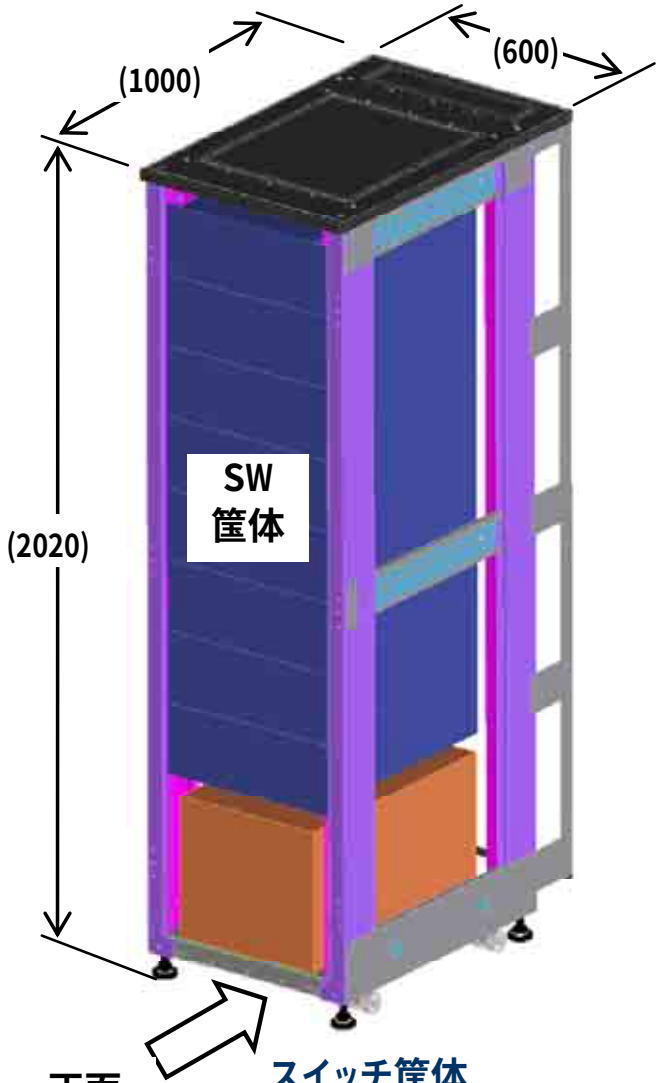
- 特徴
  - 光インターコネクトの採用
  - 非同期転送
  - 同報機能
  - 高速バリア同期機能付きのデータ転送機能
  - 入出力ポートの構成制御によるパーティショニング



【ベクトル部】実装設計 － 構造



計算ノード筐体 (2Nノード)

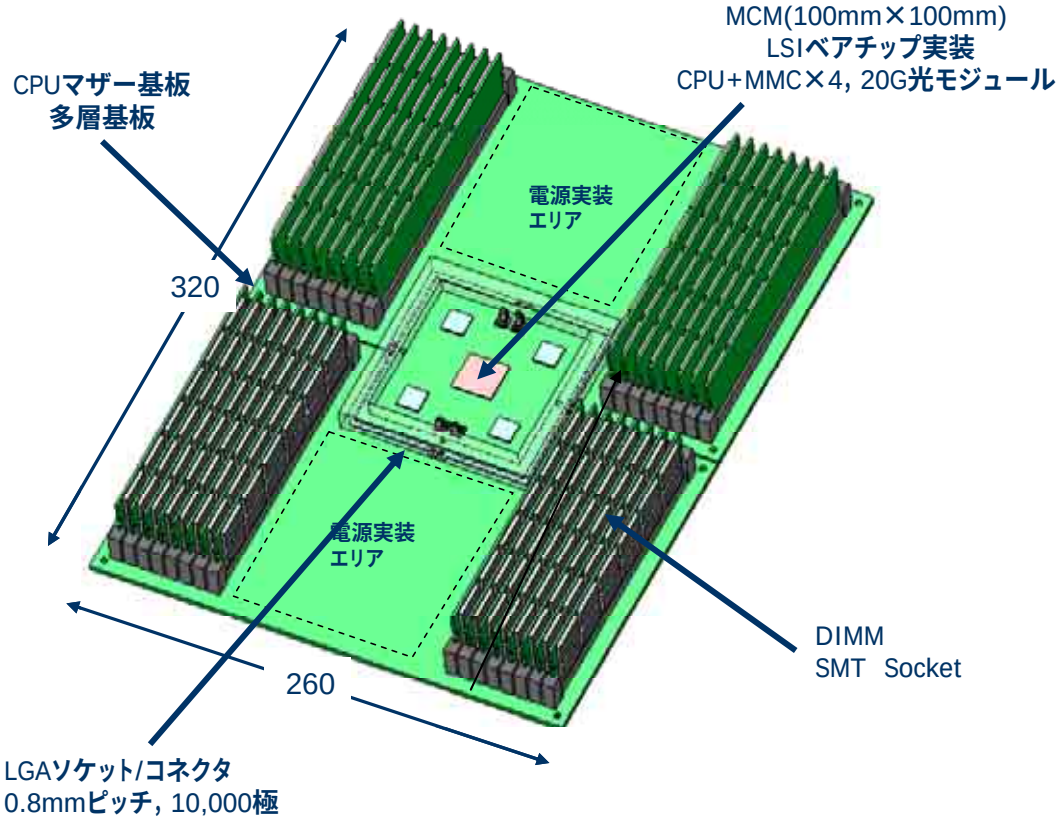


スイッチ筐体
(ノード間スイッチモジュール8台)

【ベクトル部】実装設計 ー CPU，NW実装

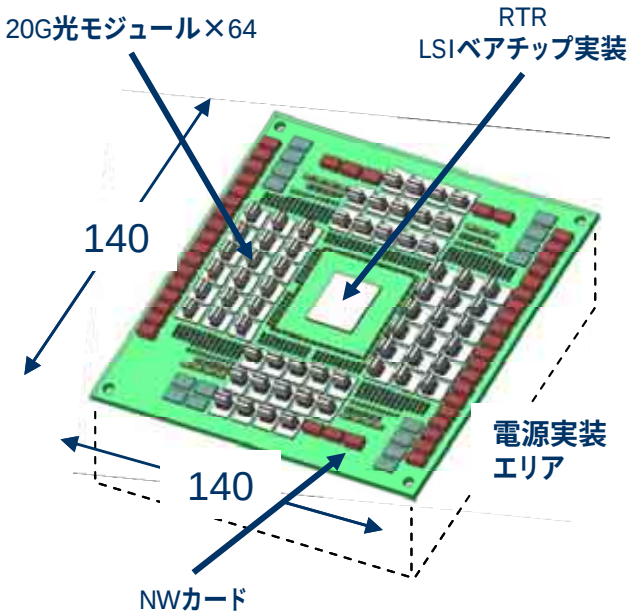
CPU実装

高速信号伝送が求められるCPUとMMCのLSIをMCM実装．MCMとDIMM（メモリ）間は，約10000極の多極LGAソケットにて接続する構造



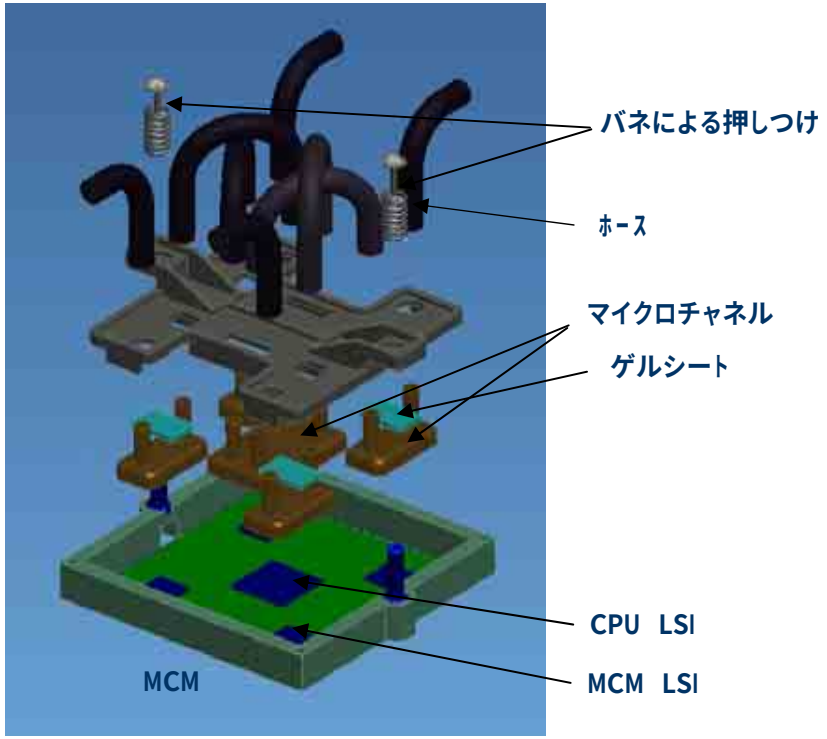
NW実装

RTRのLSIを基板中央にベアチップ実装．RTRを取り囲むように，20Gbps光モジュールを実装する構造．

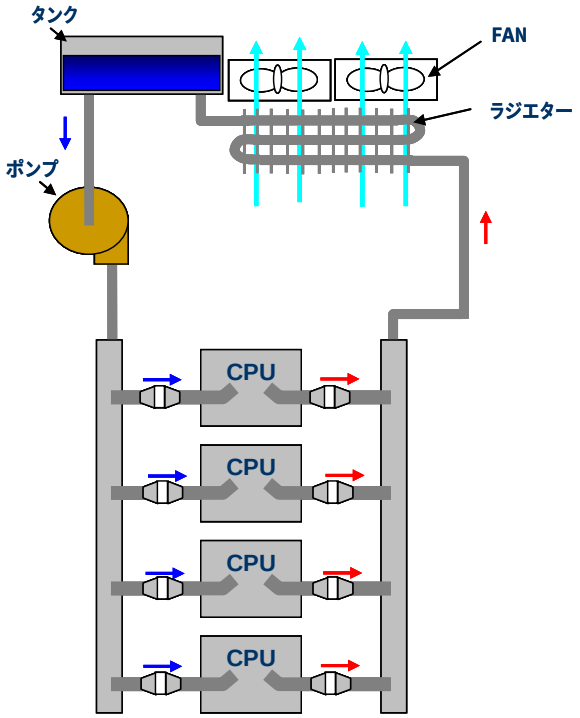


【ベクトル部】実装設計 ー CPU冷却

- MCM実装されたCPUとMMCのLSIを水冷にて一括冷却する構造を検討中.
- 冷却水は，筐体内のラジエタへ行き放熱する方式.
- 筐体風量は，118.5㎥/min以下の見込み
- CPU：200W@（ジャンクション温度 77℃），MMC：25W @（ジャンクション温度 65℃）



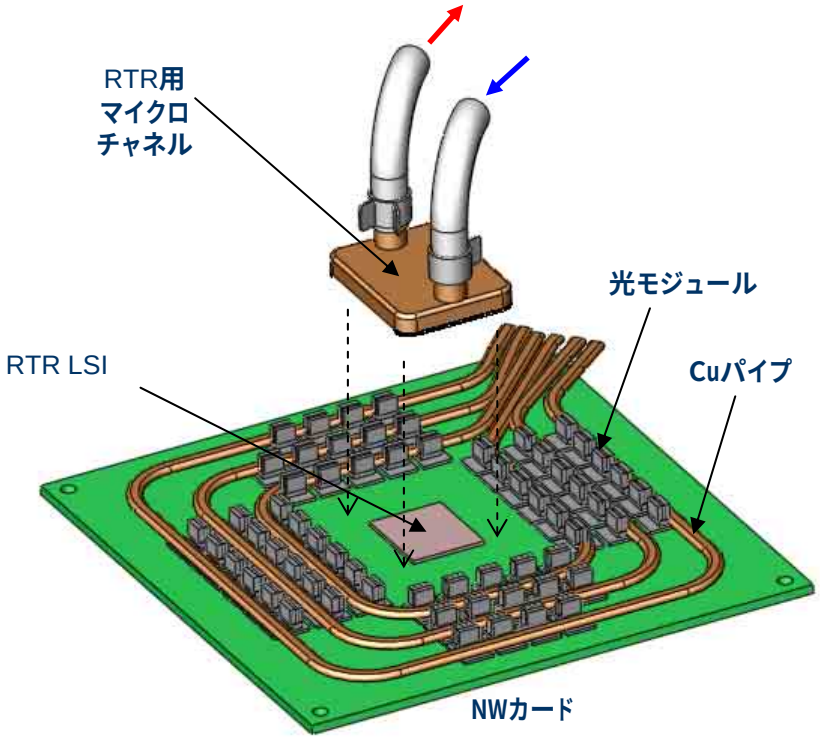
CPU部冷却構造図



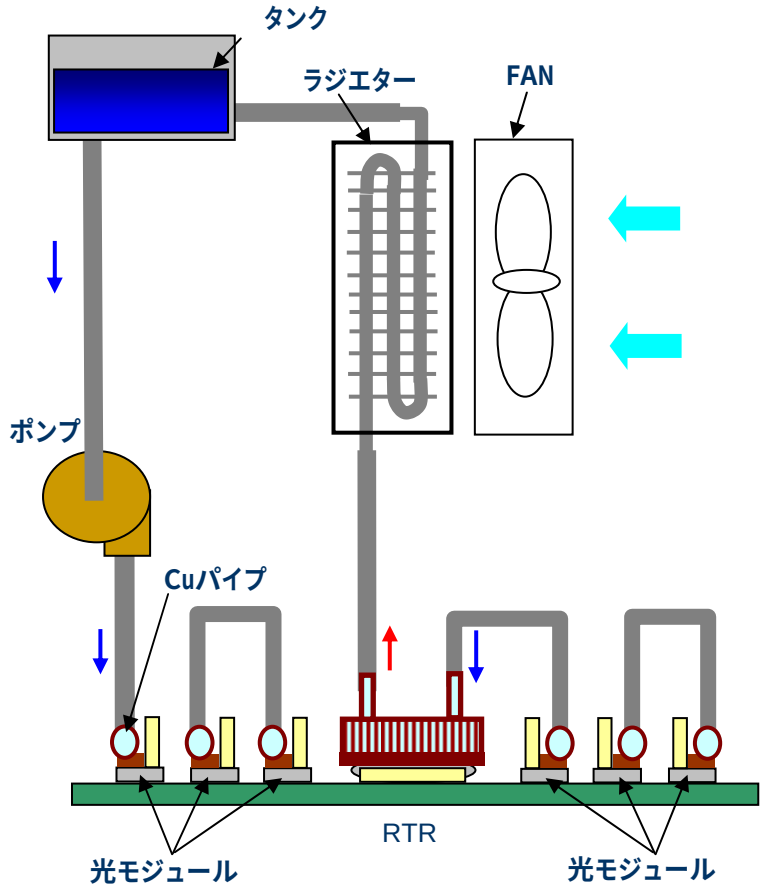
冷却システム構成図

【ベクトル部】実装設計 ーネットワークスイッチ冷却

- ベアチップ実装されたRTRのLSIを水冷にて冷却．RTR周囲の光モジュールも水冷にて冷却



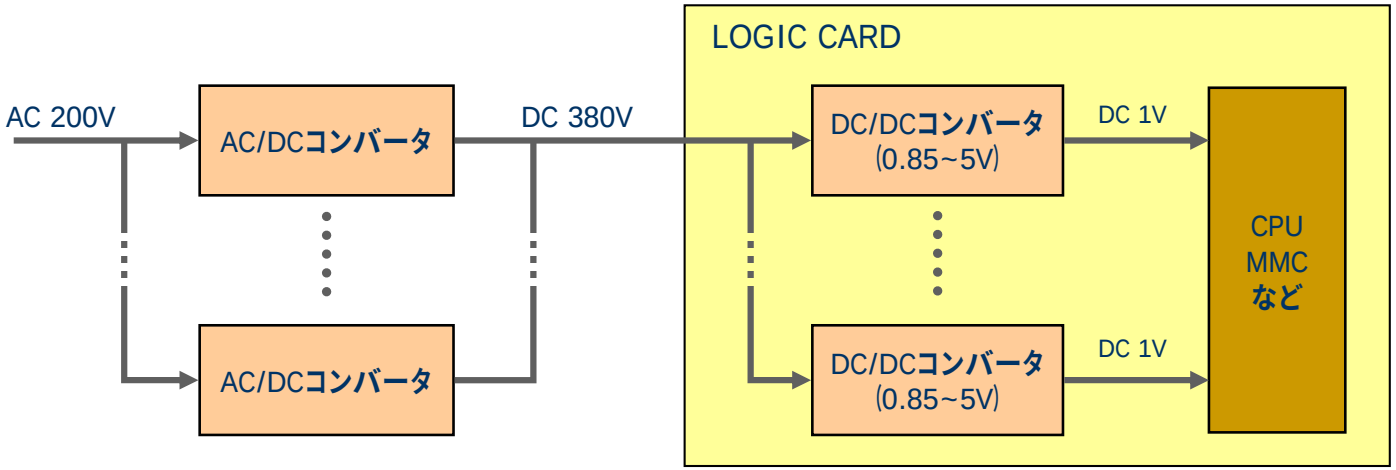
NW部冷却構造図



NW冷却システム構成図

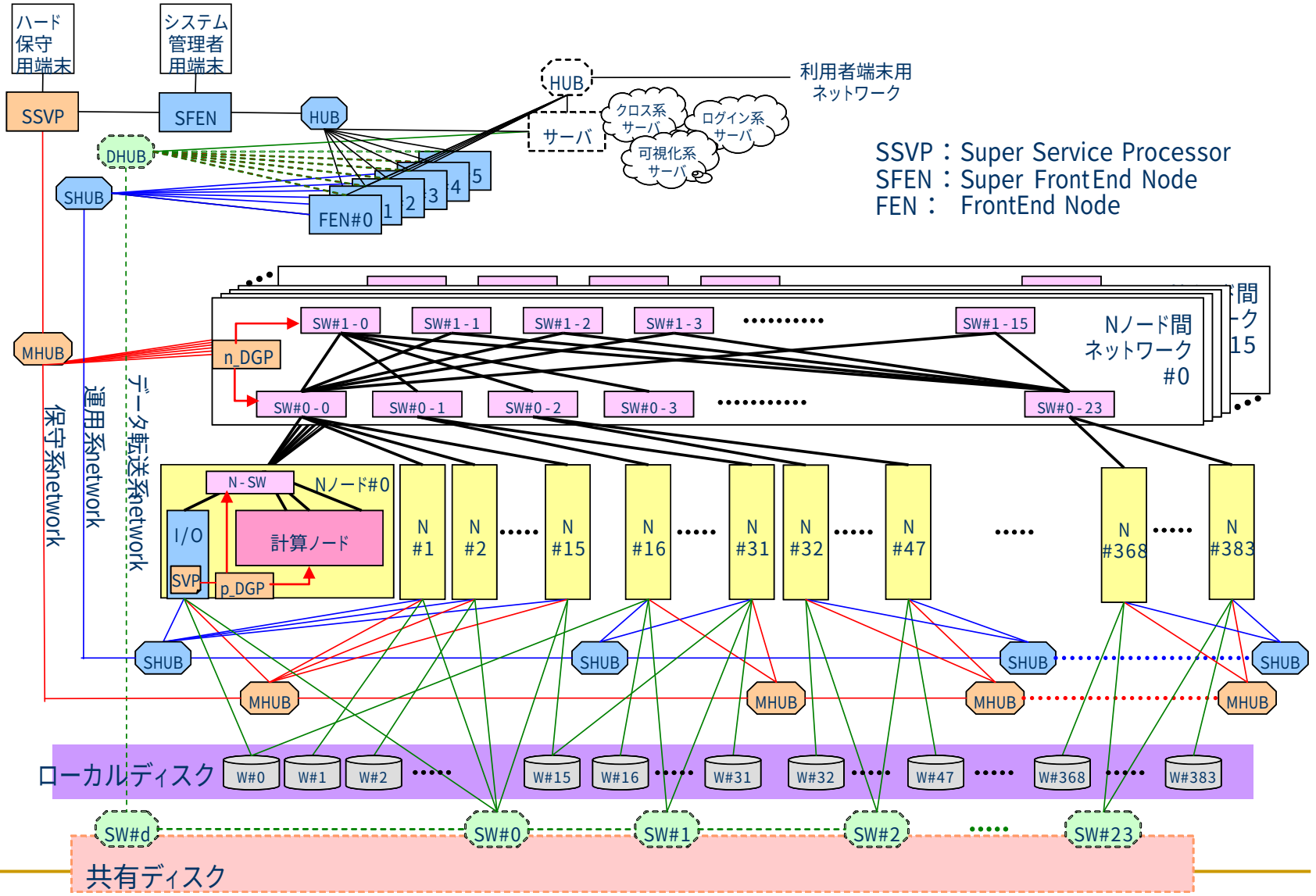
【ベクトル部】実装設計 ー電源

- 電源の変換回数を減らし電源高効率化を図るため，380V給電方式を採用．
- DC-DCコンバータは，電圧変動，電圧ドロップを抑える目的で給電路を短縮するためLOGIC CARD内に実装．



仕様	AC/DCコンバータ	DC/DCコンバータ
入力	AC 200V	DC 380V
出力	DC 380V / 2kW	DC 0.85~5V / ~200W
目標効率	97%	82% (出力 1V / 100W)
高効率施策	ゼロボルトスイッチング回路 新材料半導体部品	同期整流回路 最新半導体部品

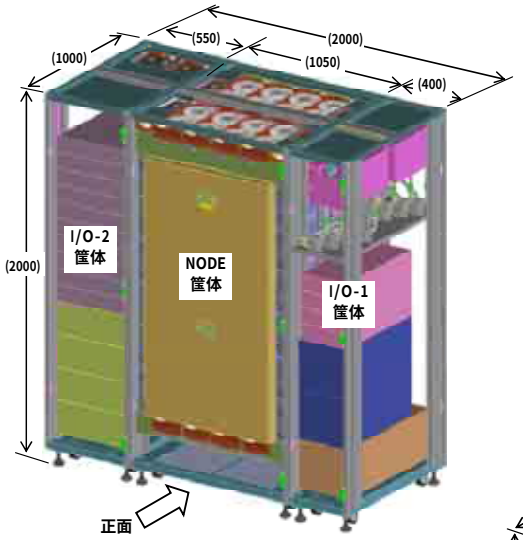
ベクトル部のシステム構成図



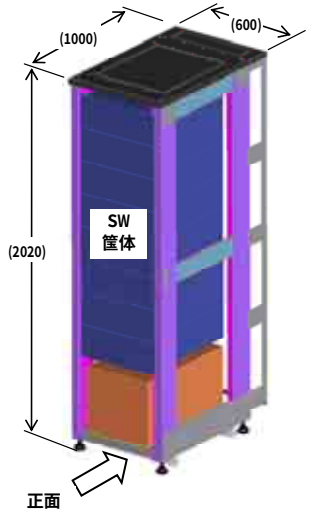
筐体実装とフロアプラン

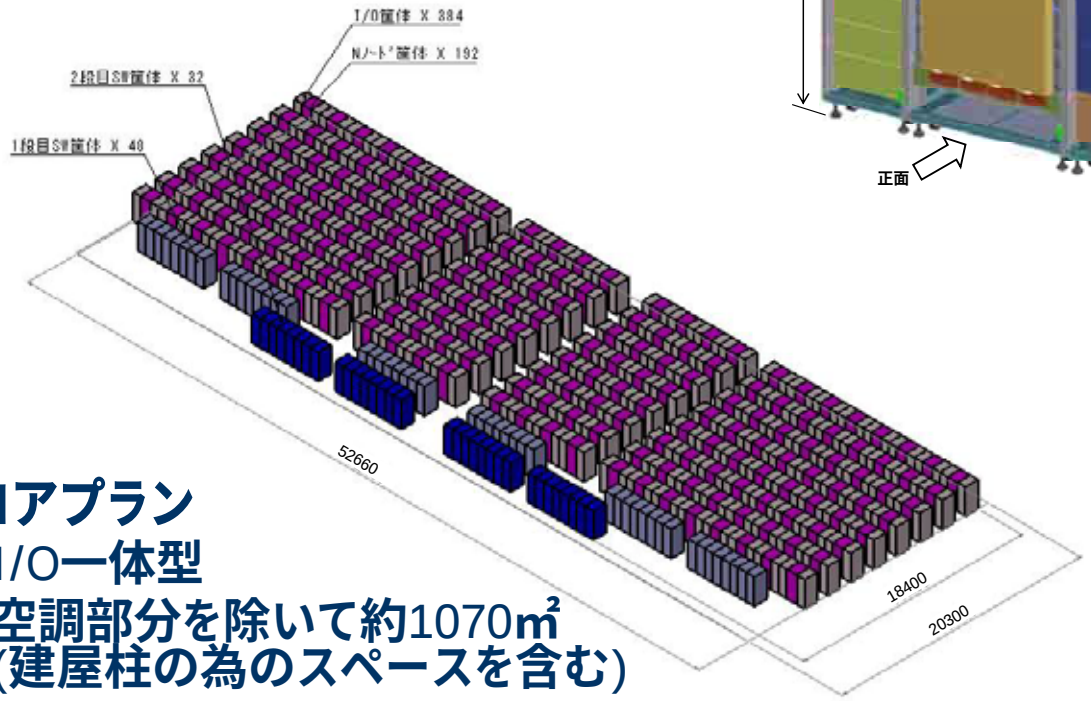
- 計算ノード筐体
  - ノード筐体1つとI/O筐体2つを1組として，2000mm×2000mm×1000mmに格納
  - ノード筐体には，2つのNノード
- スイッチ筐体
  - スイッチ･モジュール 8個を収容.

計算ノード (2 Nノード、I/Oノード)



SW筐体 (8SW)





- フロアプラン
  - I/O一体型
  - 空調部分を除いて約1070㎡ (建屋柱の為のスペースを含む)

システムソフトウェア

- **ソフトウェア構成**
  - OS: **専用OS（計算ノード）**, Linux (**フロントエンドノード, I/Oノード**) ,

- **計算ノードOSの機能**
  - **カーネル**
    - **Nノード間ネットワーク制御機能**
    - **マルチノード／マルチプロセッサ／マルチコア制御機能**
    - **ギャングスケジューリング機能**
    - **グローバルメモリ機能 (MPI通信のOSバイパス)**
    - **チェックポイントリスタート機能**
    - **I/Oノードに対するクライアント機能**
  - POSIX**に準ずるライブラリおよびコマンド**

言語・コンパイラ・開発支援ソフトウェア

- 言語・コンパイラ
  - Fortran, C/C++, HPF
  - 4倍長精度演算 (IEEE754R及びdouble-double形式)
  - RDB制御をサポート
- プログラミングモデル
  - コア内：自動ベクトル化
  - CPU内：スレッド並列 (自動並列化・OpenMP) またはプロセス並列 (MPI/HPF)
  - CPU間：プロセス並列 (MPI/HPF)
- 開発支援ソフトウェア
  - デバッガ
  - 性能解析ツール
- パラメータスィープ支援機能
  - バルクジョブ型実行モデル：　スクリプトにより実現
- 数値計算/科学技術計算ライブラリ
  - BLAS, LAPACK, ASL, fftwなど

運用系ソフトウェア

- 電源管理
  ベクトル部全体およびNノード個別の電源状態確認と起動および停止制御
- 運転スケジュール管理
  ベクトル部の起動および停止スケジュール設定
- パーティション管理
  ベクトル部全体のパーティション構成の確認および変更管理
- システム稼働状況管理
  ベクトル部全体の稼働状況の確認
- ジョブスケジューリング管理
  ベクトル部バッチシステムのジョブの予約状況の確認および管理
- ジョブ実行状況管理
  ベクトル部において投入されたリクエストのステータスおよびジョブの実行状況の確認および管理
- ログ管理
  ベクトル部において収集したログの内容確認およびログファイルの管理
- 課金／利用統計管理
  ベクトル部の課金情報および利用統計情報の確認および管理
- 障害情報管理
  ベクトル部内の障害情報の通知および障害履歴の管理

RAS機能

- CPU
  - ハードウェア診断回路
    - ECCチェック：大規模RAM (L2キャッシュ)，チップ間I/F
    - パリティチェック：その他RAM，各データバス
    - 一部ユニット二重化チェック
    - MOD-Nチェック，Out-of-Nチェック回路
    - 制御回路のシーケンスチェック，タイミングチェック，タイムアウトチェック
    - BIST (Built-In Self Test) 回路
  - 診断プログラム
    - 自動診断プログラムによるパトロールチェック機能
  - モニタ回路
    - 温度，ノイズモニタ回路による異常状態検出/モニタリング機能
- メモリ
  - LSIに関しては上記と同等
  - ECCによる1ブロック (8 b) エラー訂正，2ブロックエラー検出
  - チップ故障救済機能
- Nノード間ネットワーク
  - エラー検出/訂正
    - パリティチェック，コードチェック，シーケンスチェック，データ長チェック
    - 診断プロセッサによるOSストール監視，CPUなどの温度異常検出
  - リトライ／縮退運転
    - Nノード自動再立ち上げ，I/Oリトライ
    - Nノード縮退，Nノード間スイッチのプレーン縮退
- ストレージシステム
  - ディスクアレイRAID5
  - パス／I/Oノード冗長化
- 運用ソフトウェア
  - 計算ノード，ファイルシステム，フロントエンドの的確な連携とシステム全体の信頼性の確保

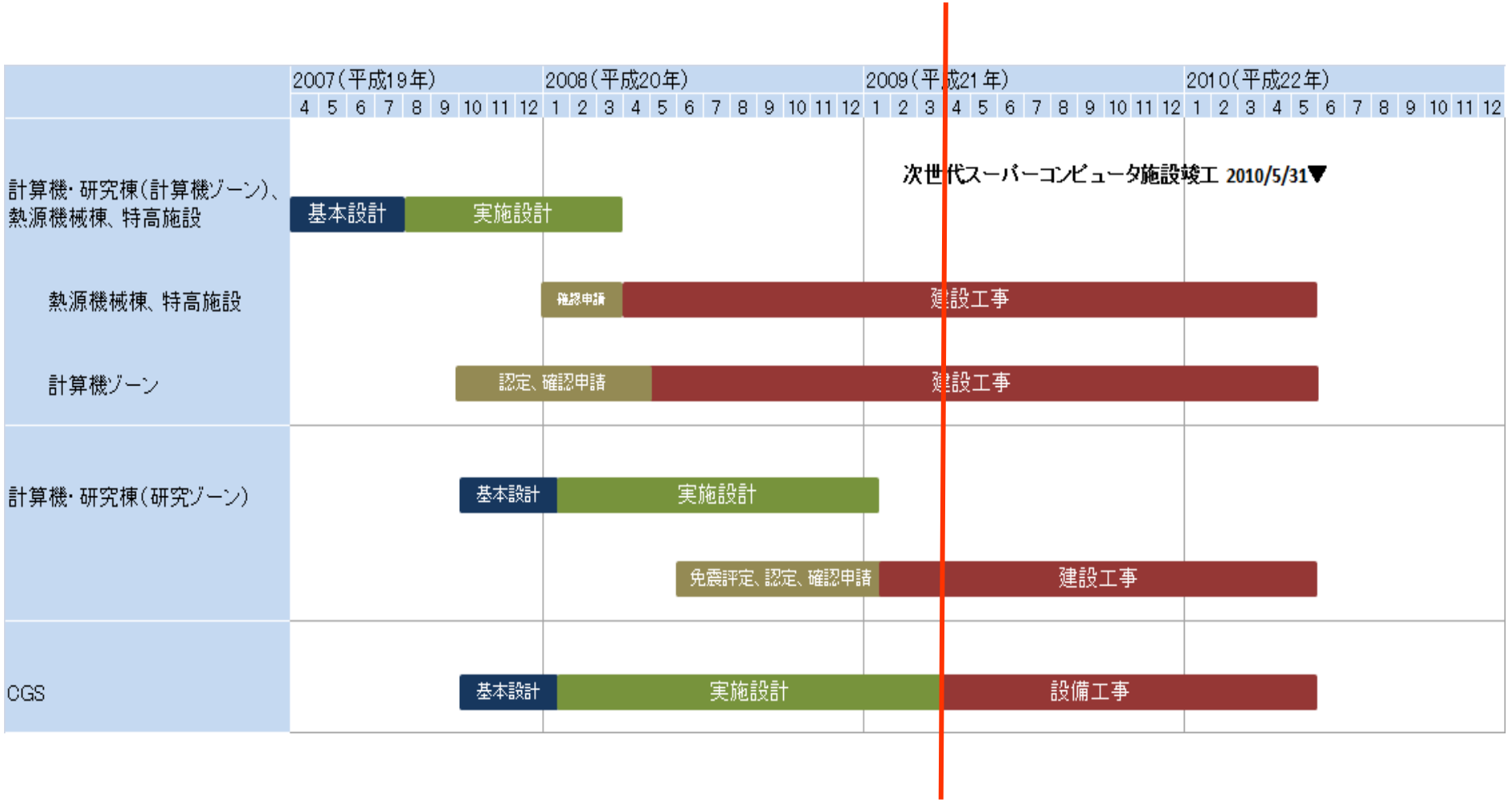
施設/設備

次世代スーパーコンピュータ施設の設置場所

450km (280miles)
west from Tokyo

次世代スーパーコンピュータ施設のイメージ

施設整備スケジュール


2007(平成19年)
2008(平成20年)
2009(平成21年)
2010(平成22年)
4 5 6 7 8 9 10 11 12 1 2 3 4 5 6 7 8 9 10 11 12 1 2 3 4 5 6 7 8 9 10 11 12 1 2 3 4 5 6 7 8 9 10 11 12
次世代スーパーコンピュータ施設竣工 2010/5/31▼
計算機・研究棟(計算機ゾーン)、熱源機械棟、特高施設
基本設計
実施設計
熱源機械棟、特高施設
建設工事
計算機ゾーン
認定、確認申請
建設工事
計算機・研究棟(研究ゾーン)
基本設計
実施設計
免震評定、認定、確認申請
建設工事
CGS
基本設計
実施設計
設備工事

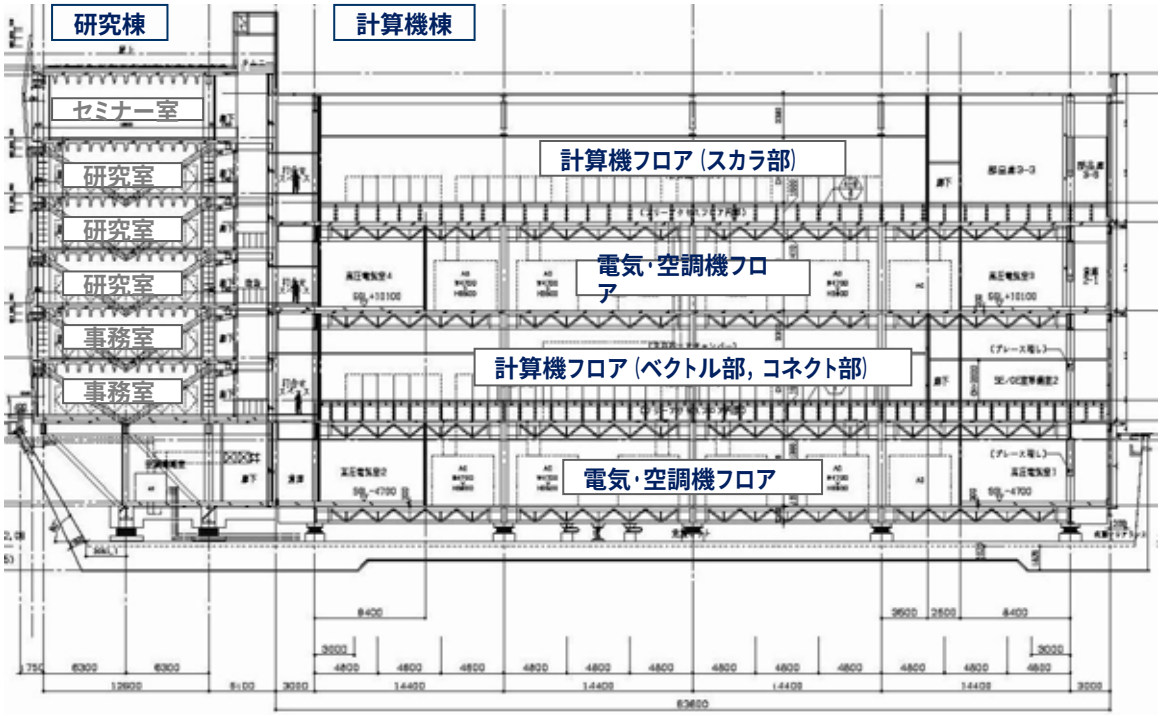

三宮
ポートアイランド南駅
次世代スーパーコンピュータ施設
建設地
2008年10月撮影


研究棟
計算機棟
熱源機械棟
特高施設

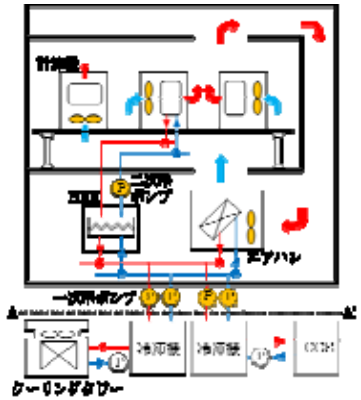
計算機棟設備概要

- 免震構造
- 4層構成
  - 3階　計算機設置階（スカラ部）
  - 2階　電気・冷却設備設置階
  - 1階　計算機設置階（ベクトル部，コネクト部）
  - 地下1階　電気・冷却設備設置階

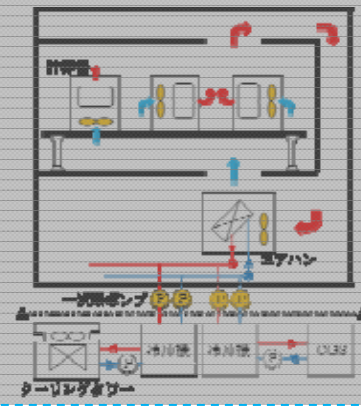
計算機施設東西断面図



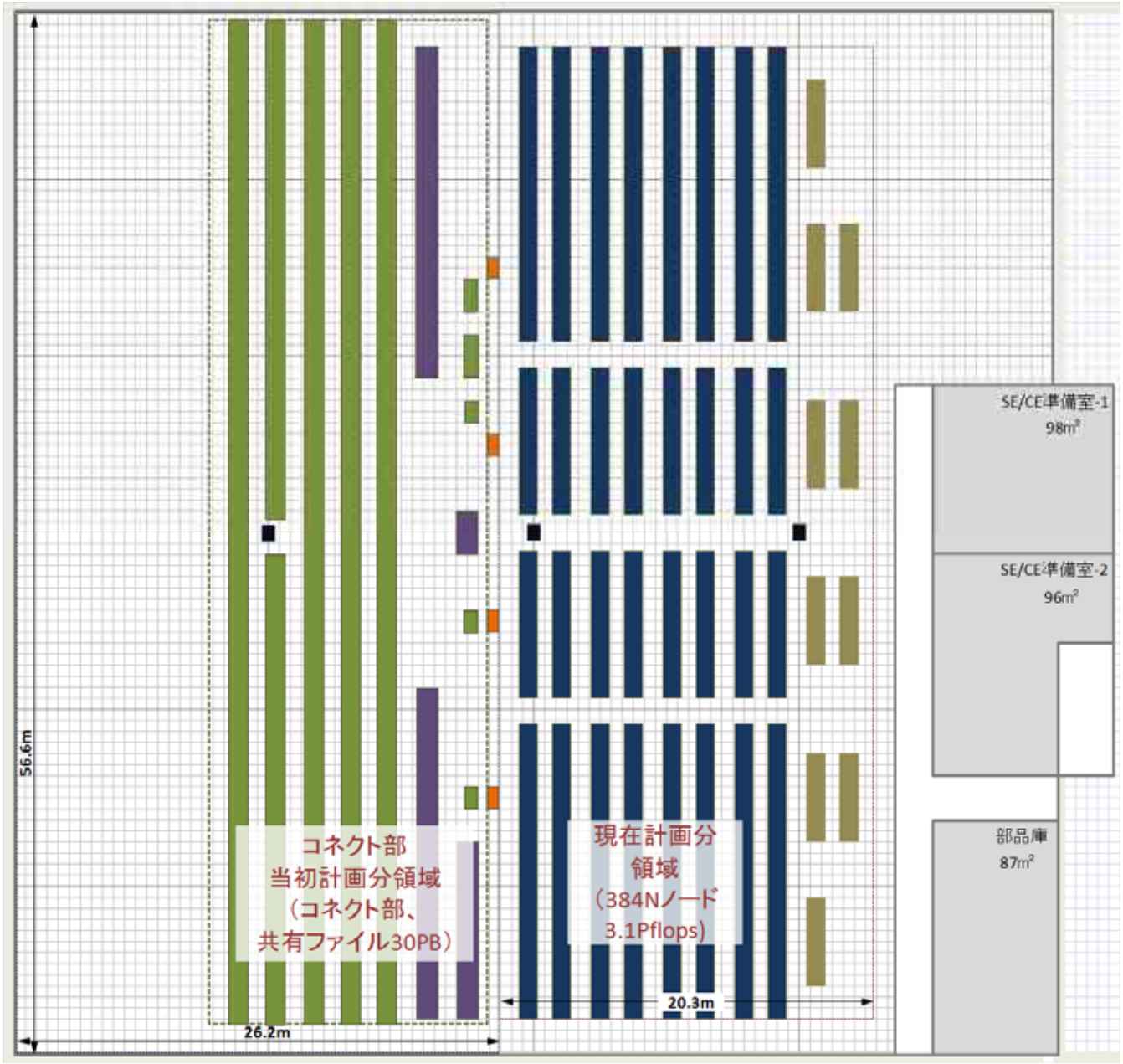
- 冷却設備概要
  - スカラ部
    - 空冷水冷併用方式



  - ベクトル部，コネクト部
    - 空冷方式



1階部分レイアウト


56.6m
26.2m
コネクト部
当初計画分領域
（コネクト部、
共有ファイル30PB）
現在計画分
領域
（384Nノード
3.1Pflops）
20.3m
SE/CE準備室-1
98m²
SE/CE準備室-2
96m²
部品庫
87m²

3階部分レイアウト


39.2m
37.6m
倉庫
倉庫
倉庫
倉庫
倉庫
倉庫
倉庫
現在計画分：1080ラック
計算機ラック24列×36列⇒11.28Pflops
ディスクラック216ラック
x
y

HPC Challenge Awardの推定性能と重点化アプリケーション

HPC Challenge Awardの推定性能

HPCCA　4項目	概念設計時	概念設計時（平成18年11月）の最大値	詳細設計 その3（予測値）		BlueWaters（HPCS Goals）	現在の最大値（平成20年11月）
			スカラ部	ベクトル部		
Global HPL (PFLOPS)	9.9	0.259 [BlueGene/L]	10+	< 3.0	10+	0.9 [Cray XT5]
Global Random Access (GUPS)	2300※1)	35 [BlueGene/L]	202	77	50000	103 [BlueGene/P]
Global FFT (TFLOPS)	140-180	2.3 [BlueGene/L]	180	226	N/A	5.1 [BlueGene/P]
EP Stream (Triad) per system※2)	6100-6900	160 [BlueGene/L]	3800	< 2250	5700	330 [Cray XT5]

※1) 概念設計評価時の推定については正確な見積りは不可能であると説明済み.
※2) 概念設計時には複合システムの値として提示. HPCC Awardのレギュレーションに合致すれば 両演算部合わせて計測する.

重点化アプリケーションについて

- 開発者等の協力を前提として，以下に挙げるアプリケーション群から，5〜6本のアプリケーションに対して，高性能化支援を重点的に実施．
  - グランドチャレンジアプリケーション (GCA) ※
  - ターゲットアプリケーション (TA)
  - 「革新的シミュレーションソフトウェアの研究開発」プロジェクト
  - JST/CREST「マルチスケール・マルチフィジックス現象の統合シミュレーション」
  - コミュニティなどから，ターゲットアプリケーションと同格に位置付けるべきとの意見が公式に提出されたもの (例えば核融合)

- 支援の重点化にあたっては，性能評価の結果や期待される科学技術・学術上の成果を考慮すると共に，次世代スーパーコンピュータの汎用性を踏まえ，分野 (TAの分野設定による) バランスや，計算手法のバランスに配慮する．

- アプリケーション検討部会，ナノ統合拠点，生命体統合拠点，革新プロジェクト，CRESTプロジェクト等の代表者と意見交換の上，重点化アプリケーションを決定 (今後，可能な限り追加予定) ．

重点化アプリケーション (6本)

プログラム名	分野	アプリケーション概要	期待される成果	手法
NICAM	地球科学	全球雲解像大気大循環モデル	大気大循環のエンジンとなる熱帯積雲対流活動を精緻に表現することでシミュレーションを飛躍的に進化させ，現時点では再現が難しい大気現象の解明が可能となる.	FDM (大気)
Seism3D	地球科学	地震波伝播･強震動シミュレーション	既存の計算機では不可能な短い周期の地震波動の解析･予測が可能となり，木造建築およびコンクリート構造物の耐震評価などに応用できる.	FDM (波動)
PHASE	ナノ	平面波展開第一原理分子動力学解析	第一原理計算により，ポスト35nm世代ナノデバイス，非シリコン系デバイスの探索を行う.	平面波 DFT
FrontFlow/Blue	工学	Large Eddy Simulation (LES)に基づく非定常流体解析	LES解析により，エンジニアリング上重要な乱流境界層の挙動予測を含めた高精度な流れの予測が実現できる	FEM (流体)
RSDFT	ナノ	実空間第一原理分子動力学計算	大規模第一原理計算により，10nm以下の基本ナノ素子(量子細線，分子，電極，ゲート，基盤など)の特性解析およびデバイス開発を行う.	実空間 DFT
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	モンテカルロ法およびCG法により，物質と宇宙の起源を解明する.	QCD

重点化アプリケーションの整備状況

- 重点化アプリケーションに対し，次世代スーパーコンピュータの両演算部の特長
  - 数万オーダの高並列性
  - キャッシュアーキテクチャ

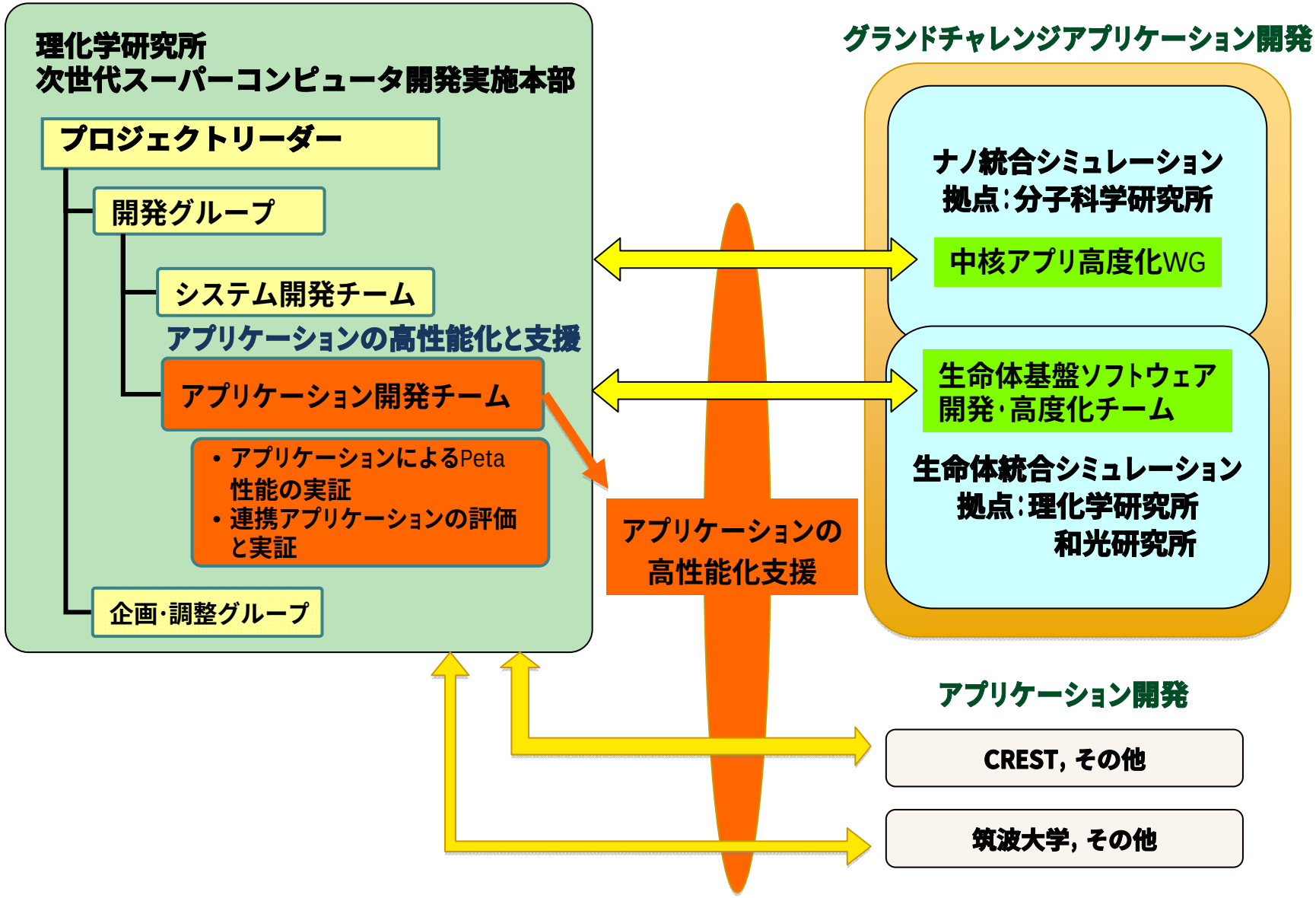
  に配慮した高性能化に着手．

- 以下の手順により，整備を進めている．
  1. 現状のマシン上で高並列性を確認する．
  2. 現状のキャッシュアーキテクチャで性能が出る事を確認する．
  3. 現状のキャッシュアーキテクチャ上で，高並列かつ単体性能も良好なアプリケーションにチューニングする．
  4. 上記段階まで達したアプリケーションについて特性を見極めた上でスカラ部/ベクトル部に特化した高度な性能最適化を実施する．

- 高並列性およびキャッシュアーキテクチャ上での単体性能に関する重点化アプリケーションの現時点の評価結果は次ページのとおり．

コード名	高並列性	キャッシュアーキテクチャ上での単体性能
NICAM	640プロセスまでの実測でスケーラビリティは良好(理研実施). 隣接通信がほとんどであり, 実行時間に占める通信時間の割合は1割以下である. さらなる高並列が可能と判断している.	カーネル部分によるベンチマークコードを用いた評価では, 数万プロセッサで14%程度の性能が得られている (スカラ部での予測評価) .
Seism3D	2048プロセスまでの実測でスケーラビリティは良好(理研実施). 隣接通信が主体である. 集団通信も一部あるが, 参加するプロセスは一部であり, 高並列時ほど影響は小さくなる. 65000プロセスまで十分にスケールすると見込んでいる (スカラ部での予測評価) .	現在, 単体性能チューニング作業中であるが, 高性能が得られる感触を得ている.
PHASE	地球シミュレータで3072並列, クラスタ等にて数百〜2000並列程度の計算実績あり. 演算量O(N3), 通信O(N2)であり, 大規模系の計算では演算が支配的. 現在のバンド並列に加え, 波数空間方向も並列化することで更なる超並列に対応可能であり現在作業中. 現状のバンド並列のみでも6万原子系であれば数万並列までスケールする評価を得ている (革新プロジェクト) .	Gram Schmidt法による直交化計算等の主要部分は行列-行列積の形で書き換えが可能であり, BLAS Level3のDGEMMを使うことで, 90%〜95%の実行効率が期待できる.
FrontFlow/Blue	1024プロセスまでの実測でスケーラビリティは良好(理研実施). 単純問題かつESの通信性能を前提とすれば, 数万プロセッサまでのスケーラビリティがある (革新プロジェクト) . 複雑問題, 現実的な通信性能での評価を実施中.	革新プロジェクトにおいてキャッシュアーキテクチャでのチューニングを実施済み. 2-3倍の性能向上を達成し, ピーク比15%程度の性能が得られている. 現在更なるチューニングを実施中.
RSDFT	・空間方向とバンド方向の2軸で並列処理が可能であり, 数万規模の並列処理が可能である. 2軸の並列処理により, 集団通信の対象プロセス数を少なくでき, 通信コストの大幅削減が可能. ・詳細設計において, カーネル部分によるベンチマークコードを用いて, 2軸並列に対する万オーダーの並列化効率を評価し, 良好な結果が得られている(スカラ部/ベクトル部). PACS-CS上で空間方向のみ並列化した. ・実アプリケーションに対して, 1,000並列以上で30%程度の実効性能の実績あり.	演算コストの中心は行列-行列積となり全体の約90%を占める. 行列-行列積はBLAS Level 3のZGEMMを使っており, 90-95%の実効効率が期待できることから, 単体性能が十分得られると判断している.
LatticeQCD	T2K-TSUKUBAにて1000プロセス規模の実績あり. 隣接通信主体のため高並列性は高い. 昨年度の評価ではスカラ部で7万プロセス程度, ベクトル部で1万プロセス程度まで良好にスケールするとの結果が出ている. さらに, アルゴリズムの改善によりノード間通信を劇的に削減できることがわかっており, 今年度の評価では当該アルゴリズムを用いたコードで推定を行っている.	現状のアルゴリズムではデータの再利用性が限定的で, 3B/FLOPS程度のメモリバンド幅が必要であることが分かっているが, 現在再利用性を高めるアルゴリズムの開発を行っており, これを適用できれば, キャッシュの利用効率を格段に向上できると見込んでいる.

アプリケーション開発支援体制



今後のスケジュール

システム導入･設置スケジュール

- 平成23年3月末： スカラ部 5PF (LINPACK測定可能構成)，ベクトル部 1PF
- 平成24年3月末： スカラ部10PF，ベクトル部3PF
- 平成24年9月末： 統合システム (スカラ部＋ベクトル部) 完成 (予定)
- 平成24年10月： 正式運用