秘

回収資料

資料2-2

次世代スーパーコンピュータの概念設計について（続き）

平成19年3月27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

5. 共同研究によるアーキテクチャ検討，評価

共同研究（平成18年4月～8月）

- 平成18年4月6日「『次世代スーパーコンピュータ：概念構築に関する共同研究』参加機関の募集について」により，共同研究機関を公募．同時にプレス発表を実施．（4月21日締切）

- 16組織17提案．このうち，具体的なアーキテクチャ提案のあった6組織（東大，筑波大，国立天文台，富士通，日立，NEC）とアーキテクチャ評価に関する共同研究を実施．
  - 理研
    - 次世代スーパーコンピュータのアーキテクチャ検討
    - ベンチマーク・テスト・コードの開発（21本のターゲット・アプリケーションから）
  - 相手機関
    - 次世代スーパーコンピュータのアーキテクチャの提案
    - 提案アーキテクチャでのベンチマーク・テスト・コードの性能予測

- 6月末 評価結果を集計．

アーキテクチャ案の概要（汎用システム）【6月末時点】

- **基本要求仕様（性能評価の基準とするシステム構成）**
  - **理論ピーク性能**：10PFLOPS
  - **総メモリ容量**：2.5**ペタバイト**

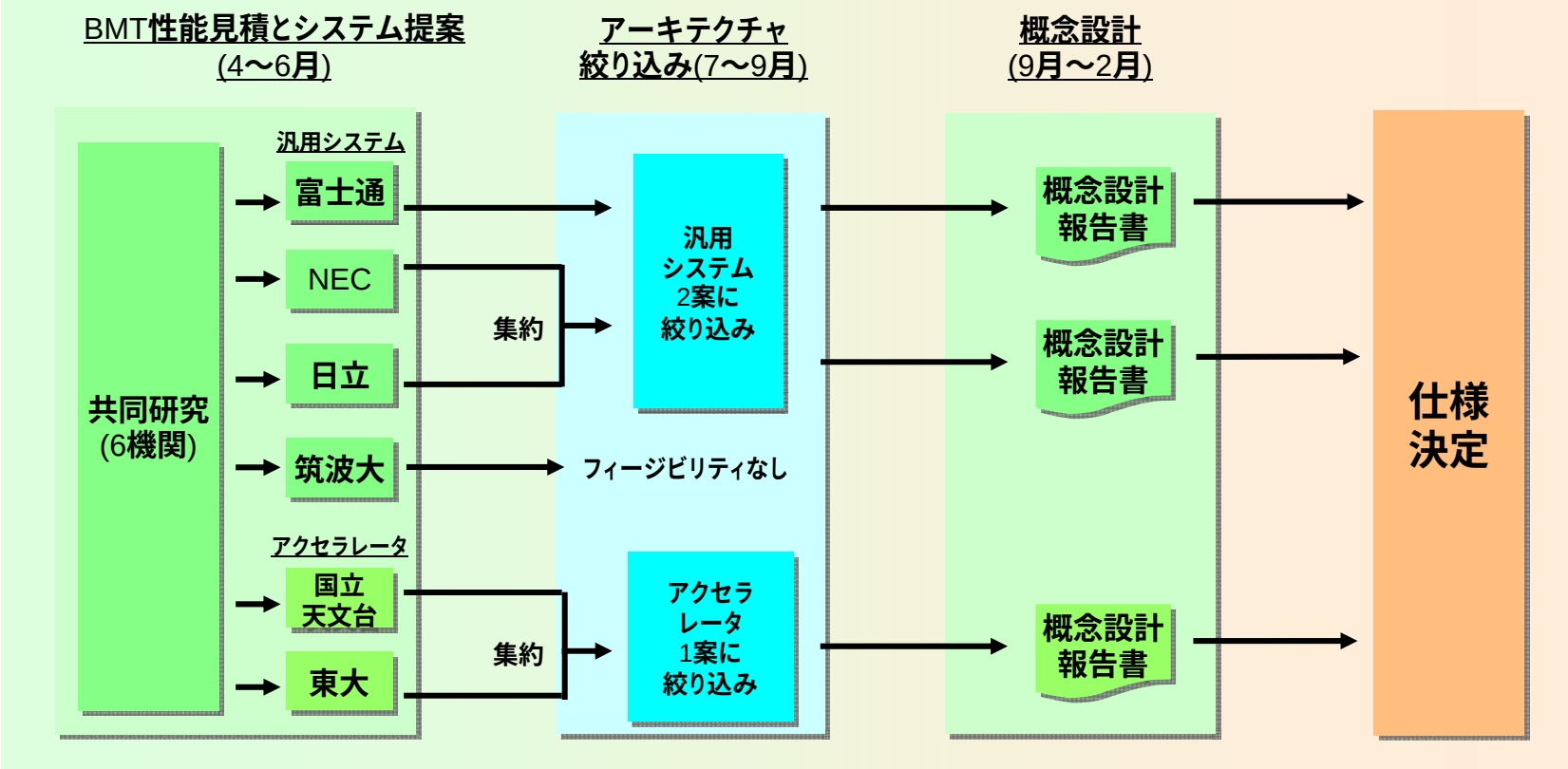
アーキテクチャ案	NEC	日立	富士通	筑波大学
コア数 （コア：1演算プロセッサ）	中並列 10万以下	高並列 10～50万	超並列 50～100万	
計算ノード数	～5万		10～15万	
高速演算機構	ベクトル		SIMD	
消費電力（本体のみ）	20-30 MW	10-20 MW	20-30 MW	10-20 MW
設置面積	3000 ㎡以上	1000-2000 ㎡	3000 ㎡以上	1000-2000 ㎡

アーキテクチャ案の概要（アクセラレータ）【6月末時点】

- **基本要求仕様**
  - **アクセラレータ部の理論ピーク性能**：10PFLOPS
  - **汎用サーバ（ホスト）のI/Oインターフェースに接続**
  - **ホストより指定された演算処理をアクセラレータで実行し，結果をホストに格納**

アーキテクチャ案		国立天文台	東京大学
アクセラレータ	アーキテクチャ	SIMD型 プロセッサアレイ	
	プロセッサチップ数	約 15,000	約 20,000
	ボード数	4,000	2,500
ホストサーバ数		2,000	2,500
アクセラレータ部 消費電力		-10 MW ※平成24年6月公開時の注意書き 消費電力については、10MW以下、10-20MW、20-30MWの範囲でまとめたもの。提案は、ホスト部を除いて、1.7MWであった。	-10 MW ※平成24年6月公開時の注意書き 消費電力については、10MW以下、10-20MW、20-30MWの範囲でまとめたもの。提案は、ホスト部を除いて、0.68MW（案1）、0.88MW（案2）であった。

システム候補の絞り込み



【各案の特徴】

	汎用システム				アクセラレータ	
	富士通	NEC	日立	筑波大	国立天文台	東大
特徴	性能優先のアプローチ		省電力優先のアプローチ		小規模･多数のSIMD演算器	

総合科学技術会議のフォローアップ結果

- 総合科学技術会議評価専門調査会では，平成17年度に「最先端・高性能汎用スーパーコンピュータの開発利用」プロジェクトの事前評価を実施．
- 平成18年度に，事前評価の指摘事項に対し，フォローアップを実施（平成18年8月～10月）．

- 指摘事項に関するフォローアップ結果
  - マネジメント体制の構築については，メーカー，大学，研究所の三者による協力体制が確立された．
  - ターゲットを明確にした開発の推進については，ターゲット・アプリケーションが選定された．
  - 京速計算機システムの構成の最適化については，システム構成の練り直しを実施した．
  - アーキテクチャ案の決定などについては作業の遅れが見られる．

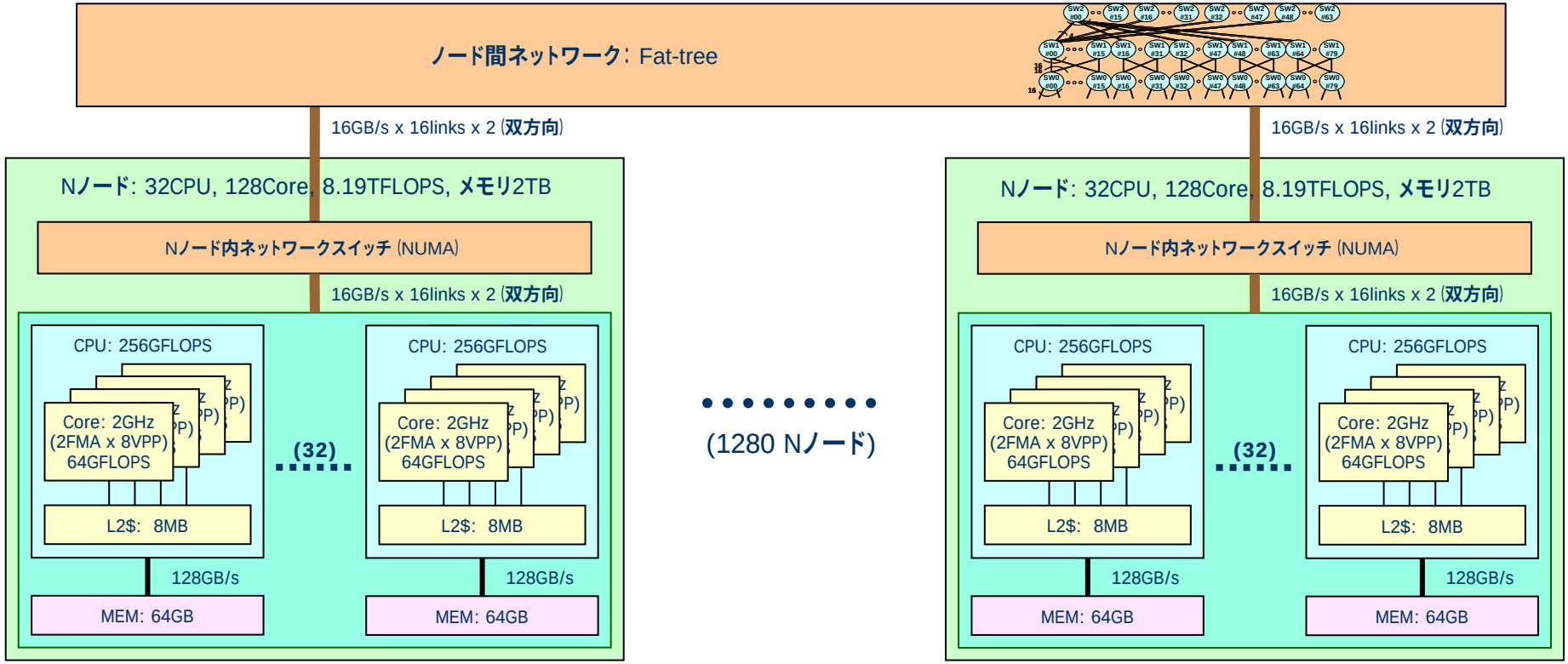
6. 概念設計について

概念設計の概要

- NEC＋日立チーム (NH) と富士通 (F) の2者が，次世代スーパーコンピュータ・システムの概念設計を実施．
  - 期間：平成18年9月19日 － 平成19年2月28日
  - 概念設計の主な要求仕様
    - ピーク性能10PFLOPS以上，メモリ容量2.5PB以上，消費電力30MW以下 (周辺機器，空調機器を含む)，設置面積3,200㎡以下 (周辺機器を含む)
    - ただし，最終仕様ではメモリ容量や磁気ディスク容量は変動の可能性あり．
- 平成18年12月1日，2者から中間報告を受領．内容は以下の通り．
  - システム構成
    - システム仕様及び構成図
    - システム諸元 (設置面積，消費電力等)
  - ソフトウェア・スタックと機能概要
  - ベンチマーク・テストによる性能予測結果
    - SimFold, GAMESS, Modylas, RSDFT, NICAM, LatticeQCD, LANS
    - HPL, NPB-FT
  - 中間報告結果を開発グループで評価．
- 最終報告書を受領 (平成19年2月28日)

NH案のシステム構成

- **計算ノード数**：1,280 (Nノード) , 40,960 (SMP)
  - CPU数： 40,960
  - コア数： 163,840
- **ピーク演算性能**： 10.48PFLOPS
- **メモリ総容量**: 2.5PB (Nノード当り2TB)
- **インターコネクトネットワーク**：Fat-tree
  - 3段のFat-treeを構成
  - ポート当り16GB/s双方向の32x32スイッチを採用
  - スイッチ間は20Gbpsの光接続
- **消費電力**： 17.5MW (Linpack時, 磁気ディスク除く)



提案システムの特徴 (NH案)

- **プロセッサ**
  - 45nmプロセスによる1CPUチップ当り256GFLOPSの高演算密度実装
  - 1CPU当り4コア構成，動作周波数2GHzで駆動
  - コア当り2FMAx8セットの演算器と128KBの大容量ベクトルレジスタ
  - 8MBを4コアで共有し，ソフトウェアでも制御可能としたRDB (Reusable Data Buffering) 機能付きL2キャッシュ
  - 1CPU内の4コアは (ハードによるキャッシュコヒーレンシ保証をした) SMP構成
  - 全システムを40,960CPUで構成し演算性能10.48PFLOPS，主記憶2.5PBを実現
  - システム運用のために，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 (MPIプロセスはCPUまたはコア単位)
  - 消費電力はCPUあたり140W (Linpack実行時)

- **ネットワーク**
  - バイセクションバンド幅328TB/s，3段のFat treeで1280 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

システム・ソフトウェアなど (NH案)

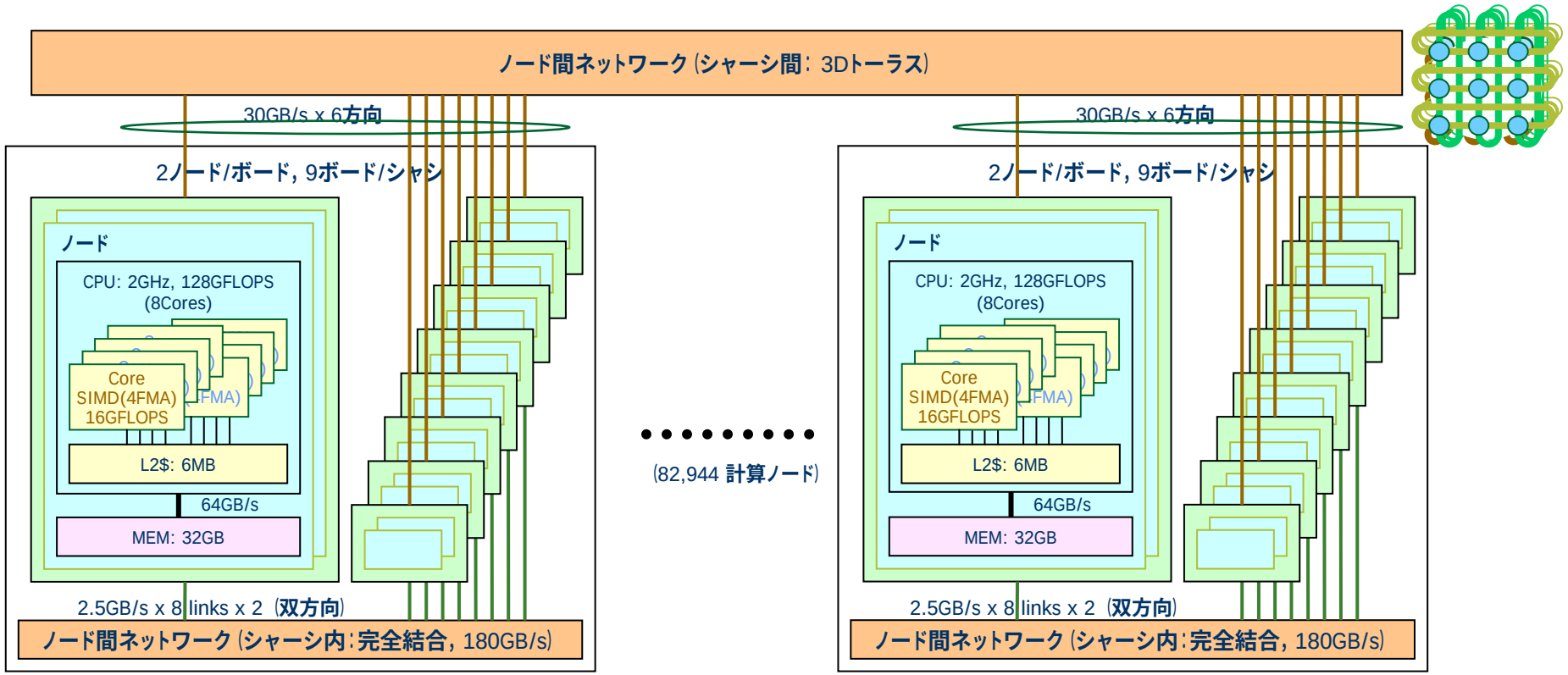
- **ソフトウェア**
  - OS: Linux (**フロントエンドノード, IOノード**) , **専用OS** (**計算ノード**)
  - **ミドルウェア**: **運用管理, ジョブ管理, ソフトウェア配布, 資源管理, グリッドミドルウェア** (**フロントエンド**)
  - **ライブラリ**: OpenMP, MPI, **科学技術計算ライブラリ**
  - **ツール**: **開発ツール, デバッグツール, チューニングツール**

- **プログラミングモデル**
  - **推奨モデル**: **分散メモリ並列, 共有メモリ並列**
  - **言語**: Fortran, C++, MPI, HPF

F案のシステム構成

- **計算ノード数**：82,944
  - CPU**数**：　82,944
  - **コア数**：663,552
- **ピーク演算性能**：10.61PFLOPS
- **メモリ総容量**2.53PB (**計算ノード当り**32GB)

- **インターコネクトネットワーク**
  - ToFu：**完全結合**+3Dトーラス
  - 18CPUを1セットとしたシャシ内を完全結合
  - シャシ間 (**総数**4608シャシ) を3Dトーラスで結合
    - リンク当り5.0GB/s×2，1シャシから30GB/s×6**方向**
- **消費電力**：15.5MW (Linpack**時，磁気ディスク除く**)



提案システムの特徴（F案）

- **プロセッサ**
  - 45nmプロセスによる1CPU（LSI）当り128GFLOPSの高密度実装
  - 1CPU当り8コア構成，動作周波数2GHzで駆動
  - コア当りFPレジスタ128本（SPARC-V9規格の4倍），SIMD拡張演算器（4FMA，4逆数近似等）によるHPC向け拡張
  - 6MBのL2キャッシュを8コアで共有，ハードバリア機構
  - パリティ/剰余チェック，命令リトライによる高信頼性
  - 全システムを82,944CPUで構成し，演算性能10.6PFLOPS，主記憶2.53PBを実現
  - 消費電力：Linpack時 58W/CPU（ジャンクション温度20℃時）

- **ネットワーク**：ToFu（Torus-Full connection）
  - 18CPUを1セットとしたシャシ内を完全結合，シャシ間を3Dトーラス結合した独自方式
  - 隣接通信を重視した設計思想
  - 次元毎に2シャシ単位で直方体分割することによるパーティション運用

システム・ソフトウェアなど (F案)

- **システムソフトウェア**
  - OS：POSIX**規格準拠の**UNIX**系オープン**OS
  - **ミドルウェア**：**運用管理，ジョブ管理，ソフトウェア配布，資源管理，グリッドミドルウェア (フロントエンドサーバ)**
  - **ライブラリ**：OpenMP，MPI，**科学技術計算ライブラリ**
  - **ツール**：**開発ツール，デバッグツール，チューニングツール**

- **プログラミングモデル**
  - 8**コア**SMP**の分散結合メモリ並列，または**8**コア**SMP**×完全結合×3D**トーラス (ToFu**トポロジに対するプロセス最適配置**)**
  - **言語**：Fortran，C++，XP Fortran，MPI，HPF，CAF

提案システム全体の比較

	NH案	F案
ピーク演算性能（PFLOPS）	10.48	10.61
総メモリ容量（PB）	2.50	2.53
総ディスク容量（PB）	140	140
設置面積：計算装置部/全体（m^2）	1,446 / 2,976	1,475 / 3,198
消費電力：計算装置部/全体（MW）	17.5 / 23 （Linpack時）	15.5 / 22.8 （Linpack時）
総計算ノード数（＝CPUチップ数）	40,960	82,944
総演算コア数	163,840	663,552
計算ノード間ネットワーク	Fat Tree	複合（完全結合＋ 3Dトーラス）

提案システムの演算部性能の比較

			NH案	F案
演算コア	動作周波数 (GHz)		2	
	演算性能 (GFLOPS)		64	16
	演算加速機構 (演算器数)		ベクトル型 (16: 2FMA x 8VPP)	SIMD型 (4FMA)
	レジスタファイル		ベクトルレジスタ 256要素×64本	スカラレジスタ 128本
CPUチップ (計算 ノード)	演算性能 (GFLOPS)		256	128
	演算コア数		4	8
	メモリバンド幅 (Byte/Flop)		0.5	
	L2 キャッシュ	容量 (MB)	8	6
		Byte/Flop	4	2
		特殊機構	選択的登録機構	ライン・ロック機構

提案システムに対する考察 (その1)

- **両者共通の設計思想：高性能・低電力システムを追求**
- **電力対性能を重視した並列アーキテクチャ**
  - **動作周波数 (2GHz) を押さえて電力低減**
  - **マルチコア：　半導体高集積技術の活用**
  - **Thinノード：　Fatノードに比べ電力対性能比で優位**
  - **超並列：　NH案 40,960ノード，　F案 82,944ノード**
- **演算加速機構(演算器数増強)とレジスタファイル**
  - **コアあたりの演算器数増強による効率よい高速演算**
  - **多数演算器に見合ったレジスタファイル装備**
- **HPC指向のオンチップ・メモリ・アーキテクチャ**
  - **キャッシュ，ローカルメモリ混在アーキテクチャ**

提案システムに対する考察（その2）

- **電力対性能比，及び面積対性能比はほぼ同等**

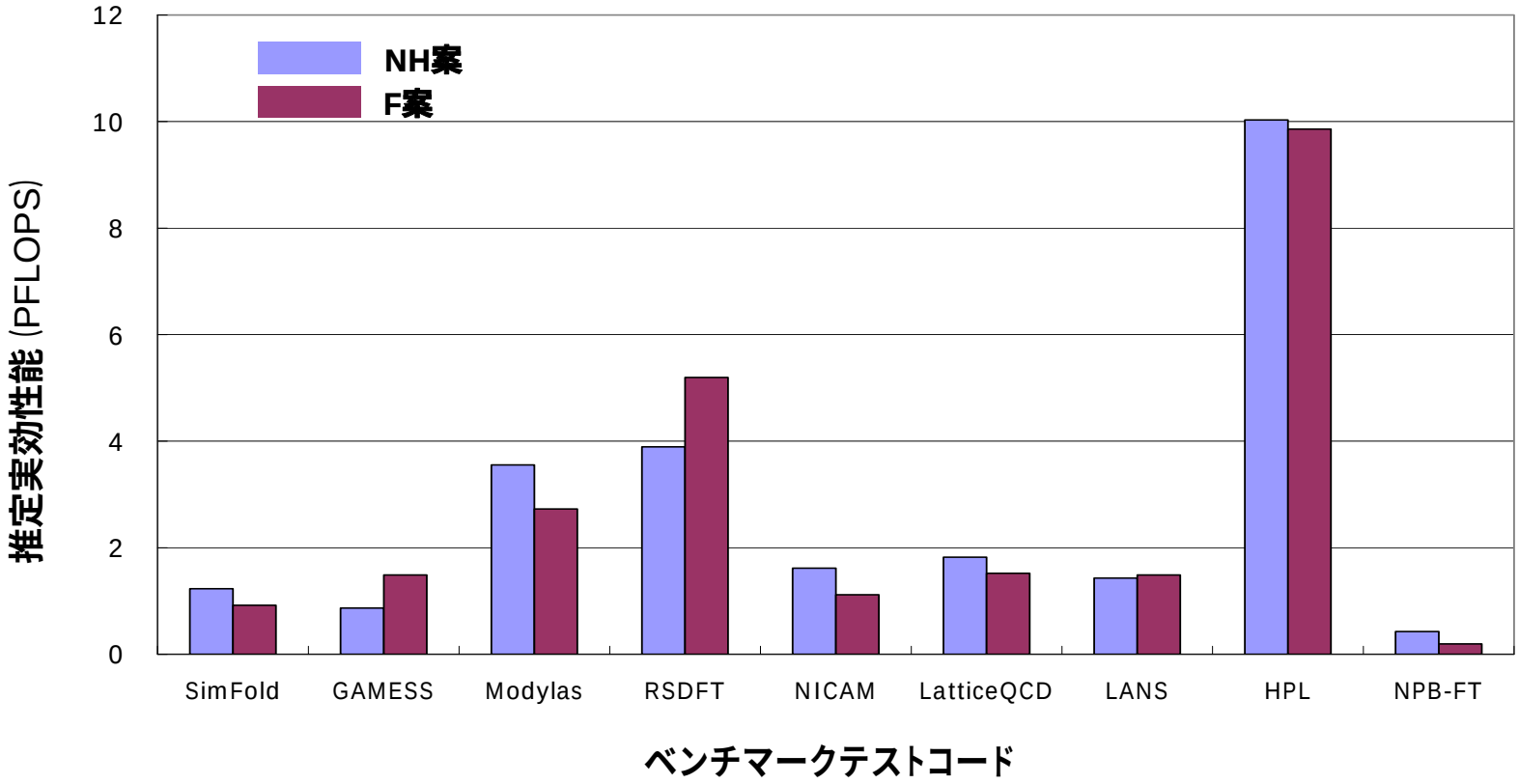
- **計算ノードの並列度は大差なし（2：1）**
  - NH案　：約4万計算ノード（演算コア：約16万）
  - F案　：約8万計算ノード（演算コア：約66万）

- **設計思想の違い**
  - **演算加速機構**
    - NH案　：ベクトル型　⇒ 演算器拡張性重視
    - F案　：SIMD型　⇒ 汎用性，柔軟性重視
  - **計算ノード間ネットワーク**
    - NH案　：Fat Tree　⇒ 汎用性重視
    - F案　：3Dトーラス　⇒ 次々世代を見据えた拡張性重視

ベンチマーク・テストによる性能評価について

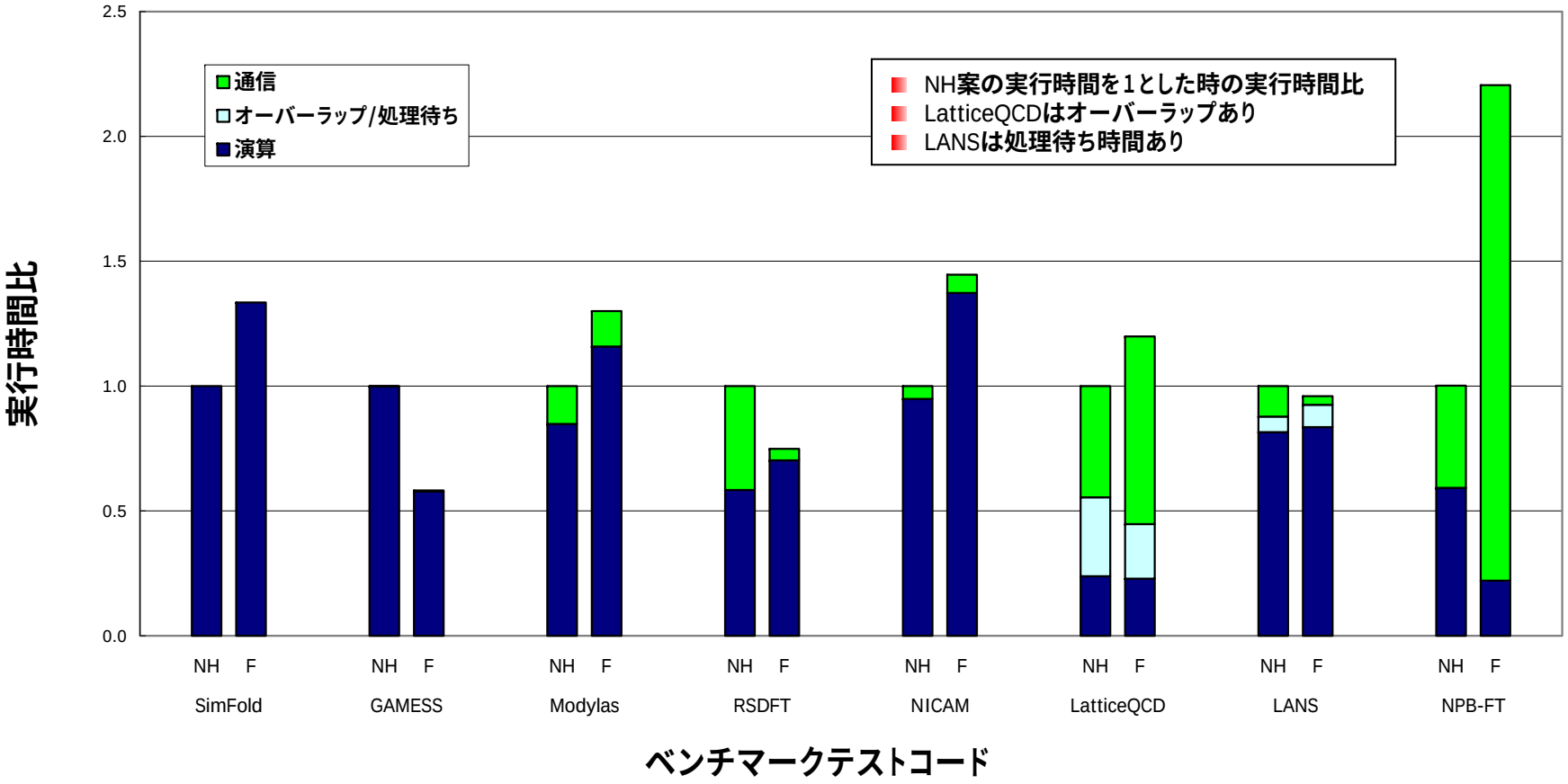
- ベンチマーク・テスト・プログラム（21本）の実行時間を推定
- 特に，ベンチマーク・テスト・プログラム（9本）について，詳細に評価
  - ターゲット・アプリケーションから7本のベンチマーク・テスト
    - SimFold，GAMESS，Modylas，RSDFT，NICAM，LatticeQCD，LANS
  - HPL（High Performance Linpack），NPB-FT
- 推定方法は，両者独自の手法を採用
  - 実機での計測値から推定
  - 新たなアーキテクチャ部分は，机上で分析，評価

ベンチマーク･テストによる性能予測 (詳細9本)



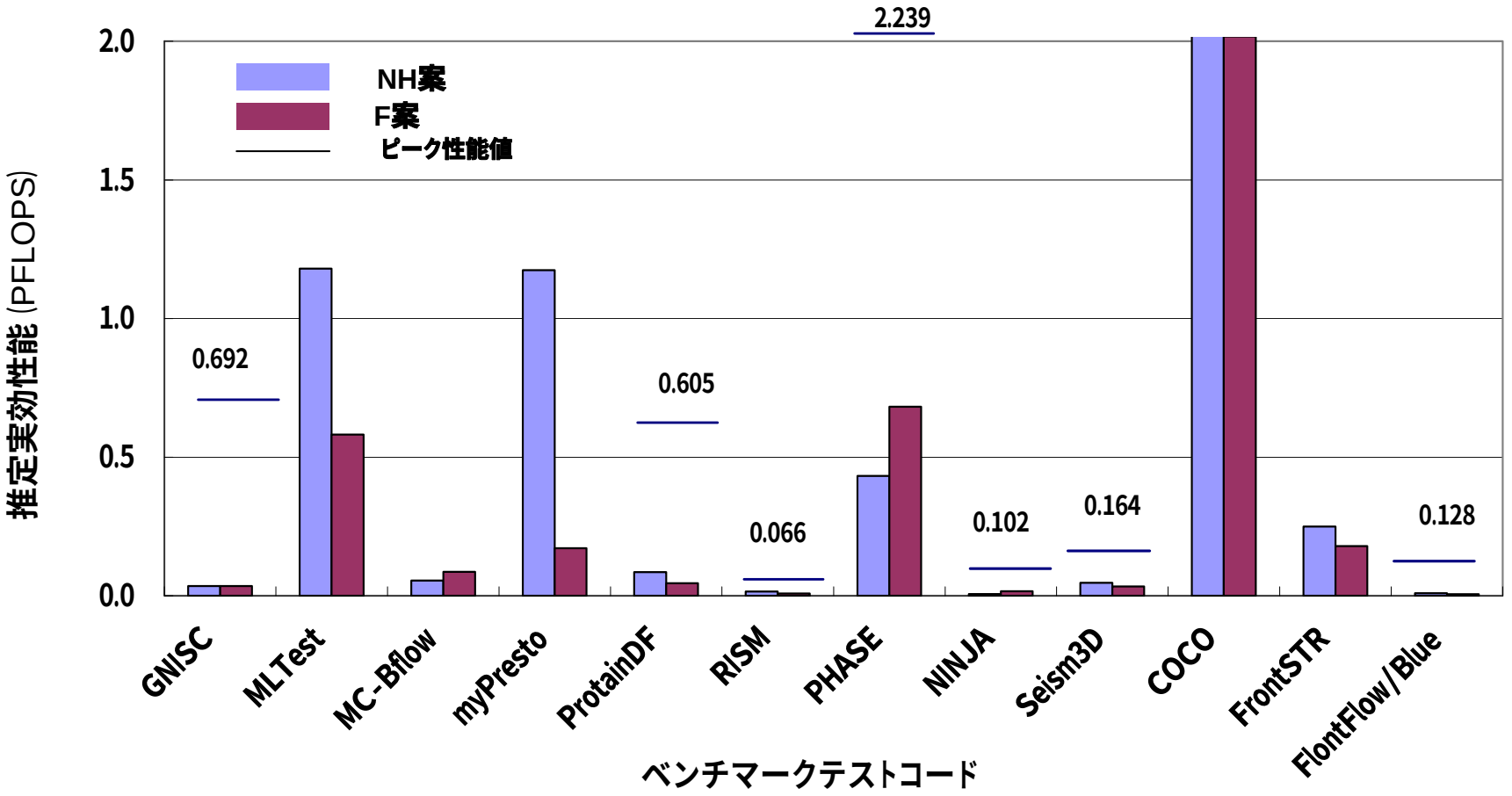
- ターゲット･アプリケーションから7本のベンチマーク･テスト，及びHPL，NPB-FTについて，実効性能を推定．
- いずれのベンチマーク･テストもほぼ同等の性能．

ベンチマークテストによる性能予測 (実行時間比：詳細9本)



- RSDFT及びNPB-FTは，通信時間の差が大きい．
  - ネットワーク・トポロジの違いが影響している．

ベンチマーク・テストによる性能予測 (他12本)



- 各BMTの最大並列数からピーク性能を設定し，その範囲内で性能予測を実施．
- チューニング等に差がある (分析中) ．

概念設計中間報告の評価結果

- **両提案に対する評価**
  - **概念設計の要求仕様（ピーク性能10PFLOPS以上，メモリ容量2.5PB以上，消費電力30MW以下，設置面積3,200㎡以下など）を満足．**
  - **ベンチマーク・テスト（BMT）による性能推定結果，電力性能比等はほぼ同等．**

- **CPUに対する評価**
  - **F案は，既存スカラプロセッサと親和性が高く，より幅広い技術展開が可能．**
  - **NH案は，ベクトルプロセッサの課題を解決し，高い演算性能を容易に達成．**

- **ネットワークに対する評価**
  - **F案の新規性・将来性は評価できるが，汎用性，運用性，実績などに優れたNH案を採用すべき．**

システム構成案検討の考え方

- 概念設計の評価結果を踏まえ，以下の2つのケースを検討中．
  - 2者のいずれかを選択（2者択一）．
  - 2者の案をベースに共同開発．少なくとも以下の項目を満たすことが条件．
    - 共同開発のシステム構成の方が単独開発のシステムより，性能が上がること．
    - 共同開発により，将来の我が国のスパコン開発の技術力，国際競争力，ビジネス展開力等の向上に一層貢献すること．
    - 開発予算の範囲内で，共同開発システムが構築できること．

- 2者のシステム構成により，目標性能達成の見込みが確認できたため，アクセラレータの採用は考慮しない．