「演算加速機構を持つ将来のHPCIシステムのあり方に関する調査研究」アーキテクチャの基本コンセプト

牧野淳一郎
東京工業大学理工学研究科
理学研究流動機構

第 7 回戦略的高性能計算システム開発に関するワークショップ 2012/7/31

概要

- 昨年度のアプリケーション部会等の復習
- 4 つのアーキテクチャイメージ
- 本研究で検討する部分
- おまけ: 今後の方向について

昨年度のアプリケーション部会等の復習

- 去年の今頃に突然つくれという話が発生した。
- アーキテクチャ・コンパイラ・システムソフトウェアのほうは SDHPC 検討グループが横すべり
- アーキテクチャは戦略分野等から人を集めて急拠立ち上げ。9-11 月に集中的にミーティング、検討。

以下 11/15 の合同部会での牧野の報告から抜粋

予備検討の方針(1)

- アーキテクチャ部会での検討では、B/F、メモリ量、ネットワークバンド幅等について非常に狭い範囲しか想定していないように思われた
- 消費電力当り性能は、エクサスケール実現にとって大きな壁である。
- B/F はアプリケーションの効率に大きく影響する一方、消費電力当り性能にも大きな影響をもつ
- メモリ量、ネットワークバンド幅も、大きく変えれば電力に影響する

アプリケーション側で、 B/F、メモリ量、ネットワークバンド幅の必要量をだしておこう

What I learned from Steve Jobs

— Guy Kawasaki

1. Experts are clueless
2. Customers cannot tell you what they need
3. Jump to the next curve
4. The biggest challenges beget best work
5. Design counts
6. Changing your mind is a sign of intelligence
7. ”Value” is different from ”price”
8. A players hire A+ players

(いくつか省略)

予備検討の方針 (2)

- アプリケーション、アルゴリズムにより B/F 等への要求は変わる
- 特に、同じアプリケーションでもアルゴリズムが変われば、またアルゴリズムが同じでも系のサイズ等だけによっても要求は変わる

といった問題があるので、

- 各分野に、重要なアプリケーション (計算法、系のサイズ等含めて) を選定してもらい、それぞれについて要求を見積もってもらう
- それらをいくつかのタイプに分類できるかどうか検討する

という方針を考えた。

アプリケーション

38 アプリケーション

分野	数
1	7
2	13
3	4
4	8
5	7

検討結果

(詳しい話は今日は省略)

- B/F とネットワークバンド幅は関係あり。
- B/F 要求高いがネットワークは弱くていいものはある。逆はない
- ランダムアクセス、非数値計算等、この軸ではよく表現できない要求もある
- 大雑把に数種類にタイプわけできそう

タイプわけの観点

観察:

- $\mathrm{B/F}$ は 0.1 以上の高いものと、桁で小さいものに分かれる
- メモリ要求にも非常に幅がある

注意事項:

- 分野によってはまだ十分な検討が進んでいない
- 分野によっては、そもそもアプリケーション・アルゴリズムの進化が速いために定量的な要求を明確にしにくいところもある

タイプわけの観点

「アプリケーションタイプ」でなくて「アーキテクチャタイプ」としてみた。

- **そのほうが物理的制約をイメージしやすい**
- **アーキテクチャ側との議論もしやすい？**

タイプわけ

以下の 4 タイプ

タイプ	B/F	メモリ量 (1TF)	消費電力 (1EF)	演算性能) (20MW)	バンド (20M
ベースライン	0.1	10-100GB	20MW	1EF	0.1EF
SoC	4	5-10MB	2-5MW	4-10EF	16-40EF
アクセラレータ	0.001	1-10GB	4-10MW	2-5EF	2-5PF
バンド幅重視	1	1TB	120MW	0.15EF	0.15EF

最終レポートではこんな感じ(1)

2．サイエンスロードマップ「アプリケーションからの要求の概要」(9/9)

ネットワークレイテンシ

タンパクMD	1時間ステップがマイクロ秒程度。同期等がこれより十分短い必要あり
格子QCD	大域縮約をマイクロ秒程度
他の多く	もう少し余裕あり

ネットワークバンド幅

格子QCD	隣接ノードとの通信速度が B/F で 0.01 程度
大域FFT	バイセクションバンド幅で性能が決まる。普通の構成では効率 1% 以下 ハードウェアだけでなく、アルゴリズム面からの検討も重要

メモリ容量・バンド幅

地震波動解析 圧縮性流体計算 有限要素解析の 防災・工学応用	100ペタバイト前後のメモリ、高いメモリバンド幅(B/F 0.5 以上)が必要
タンパクMD 格子QCD	メモリ必要量は極めて小さい。大きなバンド幅(B/F 1以上) が必要
大規模粒子系計算 量子化学計算	バンド幅、メモリ量とも比較的要求小さい

ストレージ容量 速度

DNA	シーケンサデータ処理 50EB, 500TB/s 程度が必要
他の多く	1桁程度下の要求

多様な要求

・複数アーキテクチャも視野にいれる必要あり？

・メモリ・ネットワークバンド幅については新しいアルゴリズムの研究開発も重要

最終レポートではこんな感じ(2)

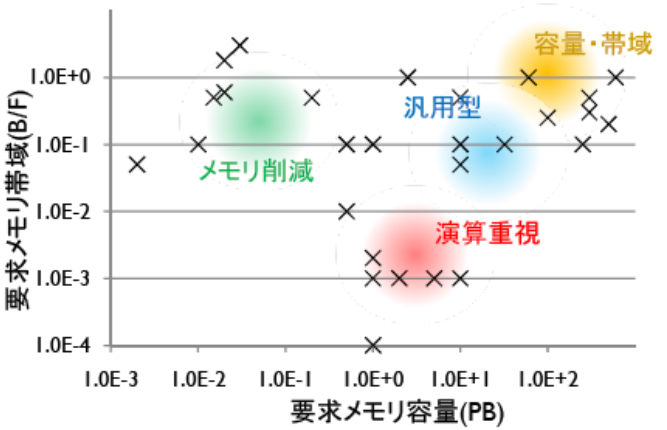
4. ロードマップ達成に向けて(アプリ要求性能)

- サイエンスロードマップ基づいて2018年ごろのアプリケーションに必要な性能を調査
  - 演算性能要求・総メモリ容量・演算性能あたりメモリ帯域・ネットワーク要求を調査

要求性能の解析結果

- 演算性能・メモリ容量・メモリ帯域に関する要求
  - 演算性能は800PFLOPS～2500PFLOPS
  - メモリ容量は10TB～500PBの幅があり、帯域も1000倍程度の差がある
  - 特徴的なもの
    - メモリ容量が少なくても良い: MD・気候・宇宙物理・素粒子物理
    - メモリ帯域が少なくても良い: 量子化学・原子核物理
    - メモリ容量・帯域が両方必要: 構造解析・非圧縮流体解析など
- ネットワークに対する要求　※トポロジに依存する部分もあり継続検討が必要
  - レイテンシ・帯域とも強い要求はないが、性能必要なアプリもあった
    - タンパク質の構造解析などでは1us以下での通信が必要な見込み
    - 物質化学分野ではBisection帯域が必要なアプリもある
    - 1us以下での高速な同期・放送・縮約などが要求されるアプリケーションもあり、専用のハードウェアによるサポートが必要になる可能性もある
- ストレージに対する要求
  - 要求容量に対して他の課題に比べて大きな課題はない
  - 性能要求に対しては今後のストレージデバイス技術に応じて構成方法を検討

- 要求性能をトレンドから予想される性能にマッピングした(図1)
  - サイエンスロードマップの達成には、前スライドの4分類とも、技術トレンドから予想される性能よりも高い数値が要求されている
- ロードマップ達成のためにアプリケーションの特性をさらに詳細化・定量化し、将来のスーパーコンピュータの設計目標を提示していくことが必要である



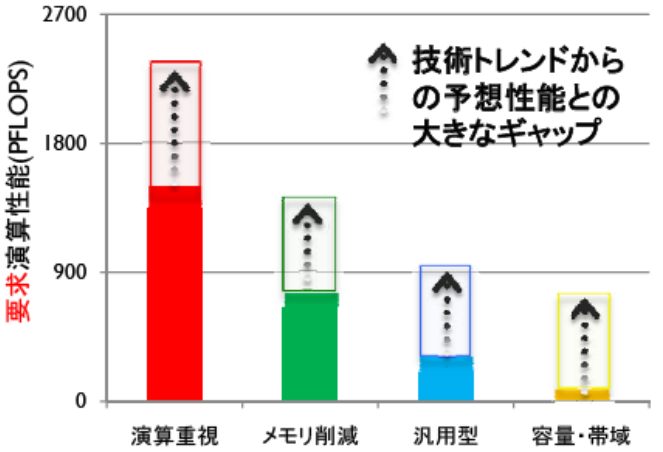


図1. 要求性能と予測性能の相関
上:各アプリの要求するB/F値とメモリ容量
下: 各分類に対する要求性能値と予想性能

要するに

報告書での用語	アプリ部会のつけた名前	具体的イメージ
汎用	ベースライン	「京」の延長
容量・帯域	バンド幅重視	NEC SX-9 の延長
演算重視	アクセラレータ	GRAPE-DR,GPGPU
メモリ削減	SoC	…

注意

- 「汎用」は汎用ではない (「京」でうまく効率がでないアプリケーションはいくらでもある)
- 「容量・帯域」が本当にそのどっちかでも実現できる設計解があるかどうかは自明ではない
- 演算重視は要するに $\mathrm{B/F}$ 要求が低いものを対象にする
- 「メモリ削減」はオンチップメモリないし 3 次元実装でバンド幅を増やすのが本質。小容量になるのは結果

我々のところでは「演算重視」と「メモリ削減」を考える
(=検討するプロセッサアーキテクチャは共通)

演算重視/アクセラレータ

- メモリ帯域が少なくて良いとなったアプリケーションの大半が量子系 (密行列の直交化や対角化が計算量のほとんどを占める)
- 後は大規模な粒子系
- GPGPU 的なものでいいが、アクセラレータ側に外部メモリはあまりなくてもいい (そちらにメモリあるならホストはなんのため?という問題も)
- GRAPE-DR ベース？

メモリ削減/SoC

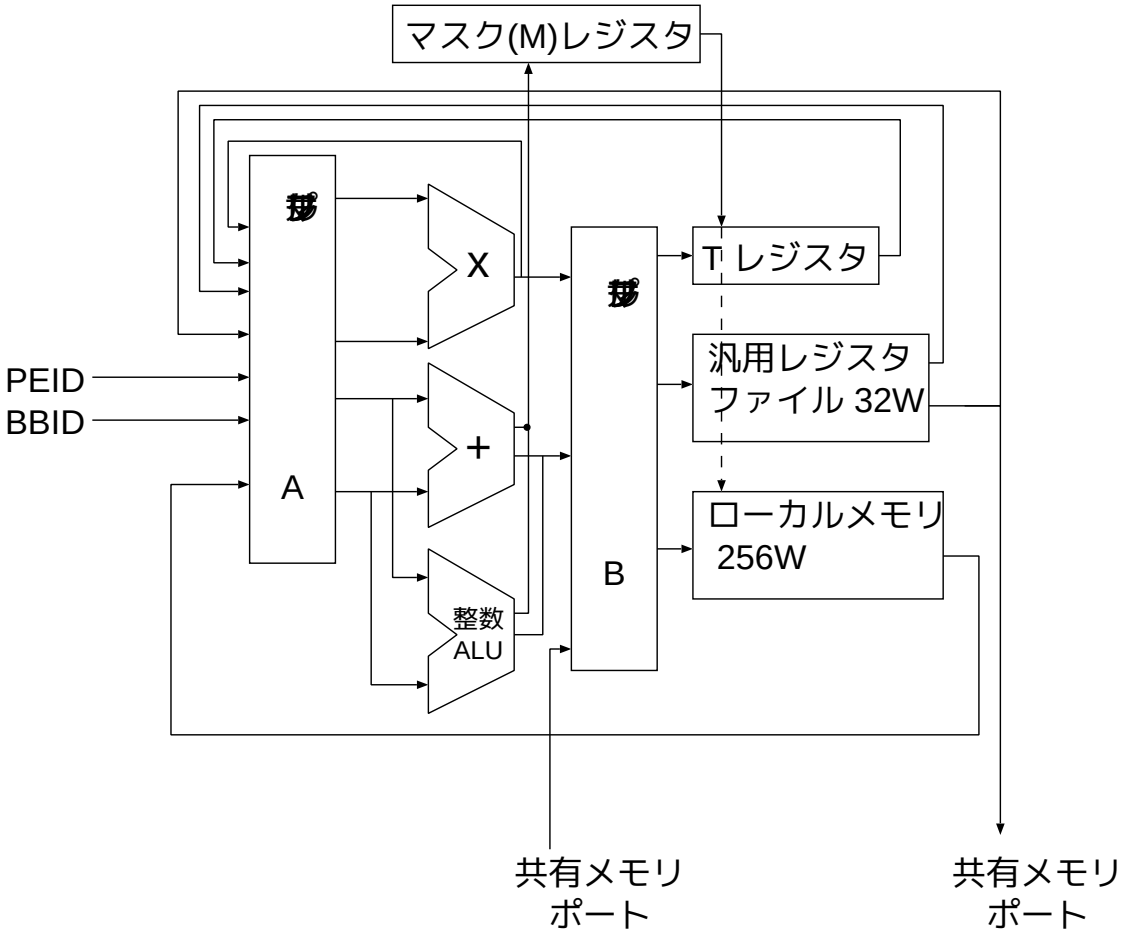
- 想定アプリケーション: 小サイズMD、流体、QCD等
- 大サイズ差分法を out-of-core でできるかどうかは要検討
- 外付けメモリがないか、極端に低バンド幅
- 軽量コアを非常に多数集積して、電力当り性能を上げる
- オンチップメモリに対しては 高B/F
- メモリはチップ当り 1GB以下程度？
- ネットワークは 100GB/s 程度？

もうちょっと具体的なイメージ

- 大雑把には: 70-80 年代の大規模 SIMD マシンを 1 チップ化。Goodyear MPP, CM, MasPar 等
- 例: CM-2。 2048 FPU, トータル 512MB メモリ
- 14nm だと 16384 FPU, 256-512MB メモリくらいが入るかも
- 考えるべきことは
  - コア内部アーキテクチャ
  - コア間接続
  - チップ間接続

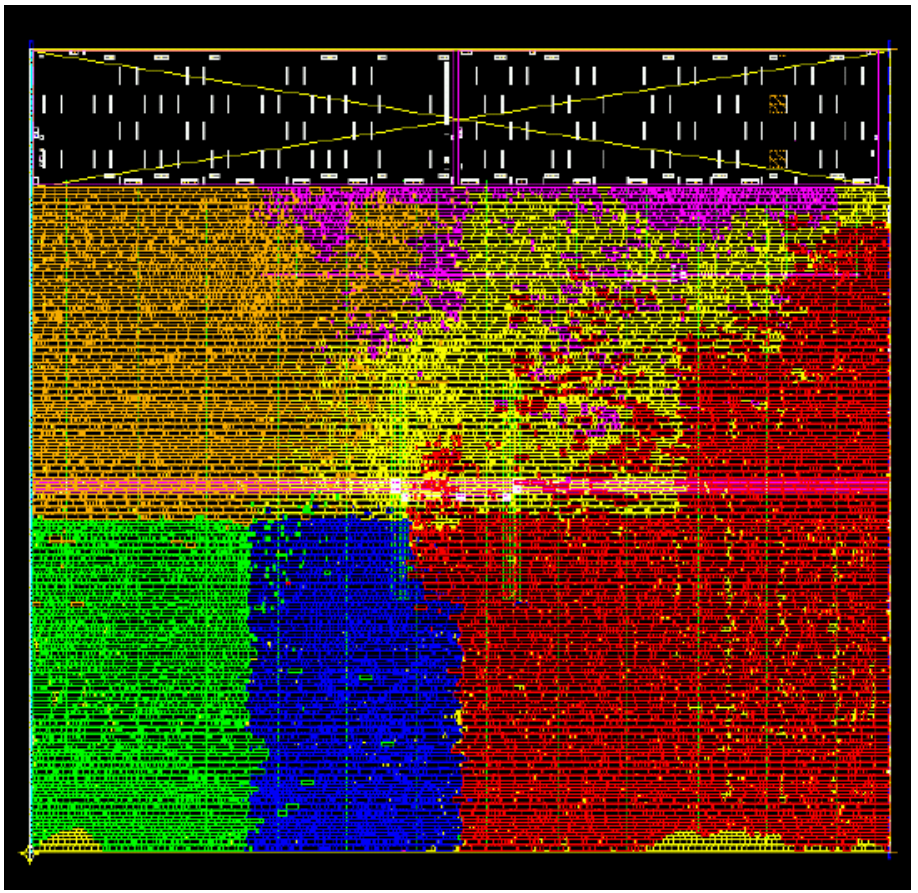
コア内部アーキテクチャ

GRAPE-DR PE の構造



- 浮動小数点演算器
- 整数演算器
- レジスタ
- メモリ 256 語 (今回増やす)

GRAPE-DR PE レイアウト



0.7mm by 0.7mm

Black: Local Memory

Red: Reg. File

Orange: FMUL

Green: FADD

Blue: IALU

他の検討点

- 演算器の方式 (単精度重視か倍精度重視か)、数
- 加算・乗算以外の機能。逆数平方根？もうちょっと汎用の関数評価？
- 制御方式 (牧野の好みは少なくともチップ内は完全 SIMD)

チップ内ネットワーク

考え方 2 つ

- 想定するアプリケーションに必要な通信パターンだけに対応
- 汎用ルーティングネットワーク

使う側の気分としては:

- (アプリケーションの次元での) 隣接通信
- 縮約・放送 (ブロックにわけたものも必要？)

くらいがあれば十分ではないかという気も。任意の 2 コア間とかもちろんあれば便利かもしれないけど、、、

チップ間ネットワーク

- 原理的には、チップ内ネットワークを延長できればよい
- 現実的には、バンド幅、レイテンシが問題
- レイテンシは実装方法を考えたい
- バンド幅については、電力・アプリケーションの要求の両方の検討が必要

今後の方針

- GRAPE-DR を拡張する形でのプロセッサシミュレータ開発、アプリケーションコアの性能評価
- 28nm プロセスでの要素プロセッサのサイズ・性能評価

まとめ

- 「演算加速機構を持つ将来の HPCI システムのあり方に関する調査研究」では、演算重視＋メモリ削減の 2 タイプのアーキテクチャを 1 種類のプロセッサでカバーすることで、伝統的なスカラー超並列やベクトル並列では実現困難である
  - 小規模問題 (強スケーリング) の高速化
  - 非常に高い電力あたり性能

  を両立させる
- 大メモリ・大メモリバンド幅の両方が必要なアプリケーションは対象にしない。
- 小規模なテストチップを試作し、実現性を評価する計画である。

おまけ:「京」開発方針決定経緯から学ぶべきこと

この6月に色々な資料が公開になって今まで謎だったことが明らかになったので、その経緯は今後の方針決定の参考にすべきであろう

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

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。

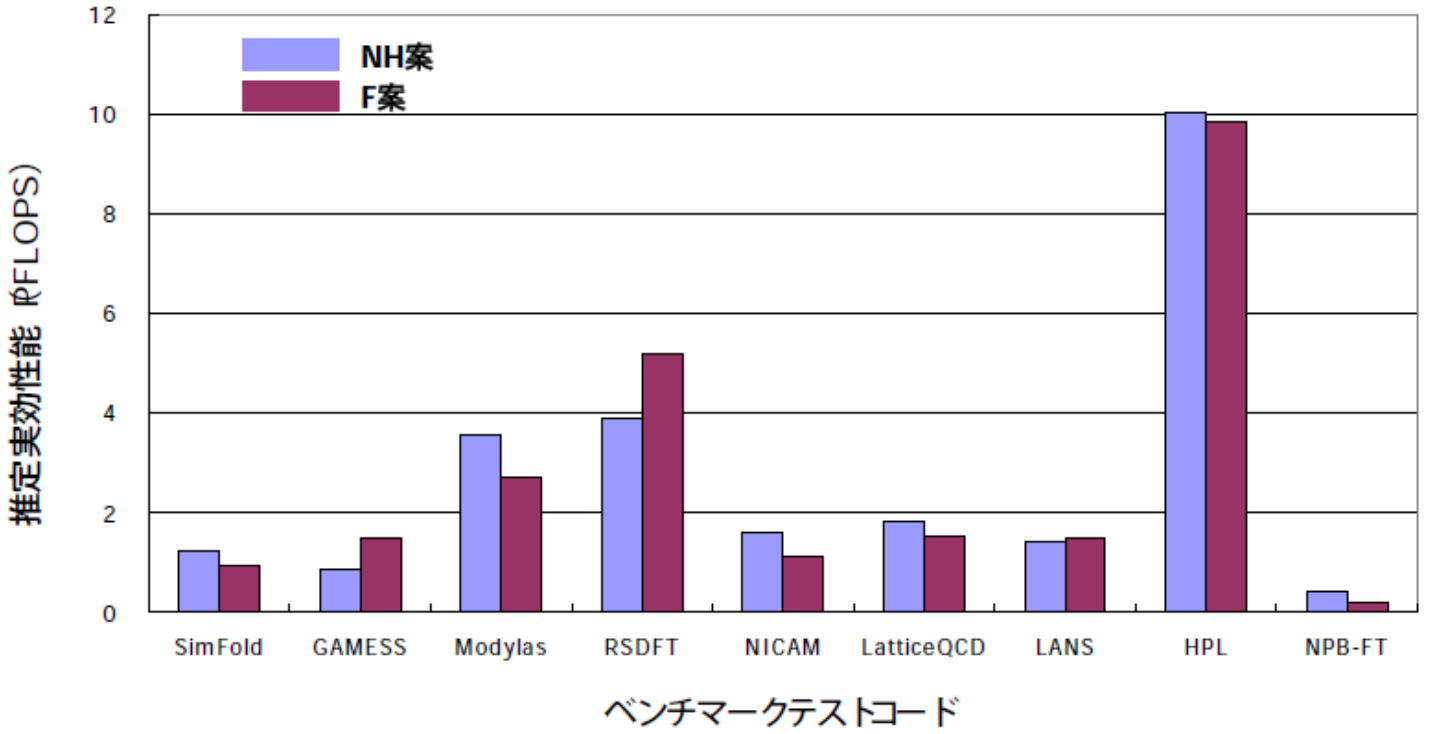
概念設計評価時の判断

- 「汎用」4提案のうち電力・設置面積が大きい2案が何故か採択された
- アクセラレータ案の説明資料には消費電力「-10MW」という謎の表現が。議事録には「10MW以下くらい」とある。
- 実際の提案の数字は公開時注意書きにあるように 0.88-1.7MW

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

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．

NH/F の比較

- ベクトル・スカラーで違うはずだったが、アーキテクチャパラメータはほとんど同じものに収束
- さらに消費電力等も収束
- さらにベンチマーク性能の収束

性能推定

ベンチマーク	NH	F	NAOJ
SimFold	1.3	1.1	2.5
GAMESS	1.0	1.7	2
Modylas	2/6	2.8	2.9
RSDFT	2.9	5.2	4.5
NICAM	1.8	1.4	2.5
LQCD	1.9	1.7	2

性能は PF 単位、ピーク10PF の場合

まあ牧野の見積もりは楽観的かもしらんが、「汎用」でもアクセラレータでもたいして変わらない。

アクセラレータを採用しなかった理由

公式の説明

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

- **目標** $=$ LINPACK 10PF $+$ HPCC 4 **種1位**
- **アクセラレータを採用すれば同じ性能で安いとか消費電力少ない、逆に同じ費用・消費電力でより高い性能が結構な数のアプリケーションででたかもしれないがそれは無視**

アクセラレータを採用しなかった理由

公式の説明

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

- 目標 $=$ LINPACK 10PF $+$ HPCC 4 種1位
- アクセラレータを採用すれば同じ性能で安いとか消費電力少ない、逆に同じ費用・消費電力でより高い性能が結構な数のアプリケーションででたかもしれないがそれは無視

まあ自分の提案だからいうのはアレだけど、こんな決め方では困る。今回はもうちょっと合理的にやって欲しい。

分野1

分子動力学計算

１分子粒度生化学反応拡散系シミュレーション

細胞、細胞間生化学反応・拡散系シミュレーション

循環器系連続体力学・生化学反応計算

次世代シークエンサ解析プログラム

人間の全脳規模に相当する 10^{11} の神経細胞からなる

リアリスティックな神経回路モデル

分野２

第一原理分子動力学法 (実空間基底)
第一原理分子動力学法 (平面波基底)
phase による電子状態計算（＋ＭＤ計算）
実時間密度汎関数法
超高精度電子状態計算
短距離古典分子動力学シミュレーション
長距離古典分子動力学シミュレーション
ナノ構造体電子・電磁場タ?イナミクス法
クラスターアルゴリム量子モンテカルロ法
feram によるリラクサー強誘電体の誘電率の
周波数依存性の分子動力学計算（アレイジョブ）
CLUPAN による相図の計算（アレイジョブ）
phonopy による熱伝導率などの第一原理計算
厳密対角化

分野3

地震動計算

3D 地球モデルに対する理論地震波形計算

地球環境モデル MIROC

全球雲解像大気大循環モデル NICAM

分野4

核融合プラズマ流体解析 GT5D
経路積分分子動力学 PIMD
FMO 計算
流体解析 VCAD
可視化
構造解析 ADVENTURE
流体解析 FFB
流体解析 UPACS

分野5

- 格子 QCD 計算 オーバーラップ型クォーク
- 格子 QCD 計算 クローバー型クォーク
- 核分裂現象シミュレーション
- 軽い原子核の第一原理計算
- 粒子系シミュレーション（大粒子数）
- 粒子系シミュレーション（小粒子数）
- 宇宙流体シミュレーション

分野1

名称	B/F	メモリ量	タイプ
分子動力学計算	0.1	1TB?	ベースライン
１分子粒度生化学反応拡散系シミュレーション	小	3.2TB	ベースライン
細胞、細胞間生化学反応・拡散系シミュレーション	0.1	1TB?	ベースライン
循環器系連続体力学・生化学反応計算	0.1	1TB?	ベースライン
次世代シークエンサ解析プログラム	?	1TB	ベースライン
人間の全脳規模に相当する 10^{11} の神経細胞からなるリアリスティックな神経回路モデル	0.03	15TB	ベースライン

メモリ量はノード性能が 100TF として規格化した

タイプの説明は後で

分野2

名称	B/F	メモリ量	タイプ
第一原理分子動力学法 (実空間基底)	?	100TB?	バンド幅
第一原理分子動力学法 (平面波基底)	?	?	バンド幅
phase による電子状態計算（＋ＭＤ計算）	0.2	100GB	バンド幅
実時間密度汎関数法	0.25	?	バンド幅
超高精度電子状態計算	0.001	1TB	ベースライン
短距離古典分子動力学シミュレーション	0.1	10TB	ベースライン
長距離古典分子動力学シミュレーション	0.1	1TB	ベースライン
ナノ構造体電子・電磁場タ?イナミクス法	0.5	30TB	バンド幅重視
クラスターアルゴリム量子モンテカルロ法	?	10TB	ベースライン
feram によるリラクサー強誘電体の誘電率の	?	100GB	バンド幅重視
周波数依存性の分子動力学計算（アレイジョブ）			
CLUPAN による相図の計算（アレイジョブ）	0.2	100GB	バンド幅
phonopy による熱伝導率などの第一原理計算	0.2	100GB	バンド幅
厳密対角化	12	11TB	バンド幅

分野3

名称	B/F	メモリ量	タイプ	コメント
地震動計算	0.03	250GB	バンド幅	B/F は正し
3D 地球モデルに対する理論地震波形計算	0.2?	50TB	バンド幅	
地球環境モデル MIROC	0.2	32GB	バンド幅	
全球雲解像大気大循環モデル NICAM	0.13	20GB	バンド幅	

分野4

名称	B/F	メモリ量	タイプ	コメント
核融合プラズマ流体解析 GT5D	0.4	10TB	バンド幅	
経路積分分子動力学 PIMD	0.001	500GB	アクセラレータ	
FMO 計算	0.001	1TB	アクセラレータ	
流体解析 VCAD	0.05	30GB	ベースライン	
可視化	0.02	100GB	ベースライン	
構造解析 ADVENTURE	0.25	1TB	バンド幅	
流体解析 FFB	0.5	16TB	バンド幅	
流体解析 UPACS	0.4	200GB	バンド幅	

分野5

名称	B/F	メモリ量	タイプ	コ
格子 QCD 計算 オーバーラップ型クォーク	0.5	20GB	SoC	
格子 QCD 計算 クローバー型クォーク	0.5	12GB	SoC	
核分裂現象シミュレーション	0.01	12GB	アクセラレータ	
軽い原子核の第一原理計算	10^{-5}?	100GB	アクセラレータ	
粒子系シミュレーション（大粒子数）	0.001	100GB	アクセラレータ	
粒子系シミュレーション（小粒子数）	(0.1)	500MB	SoC	
宇宙流体シミュレーション	(0.1)	100MB	SoC	