

エクサフロップスとその先を考える — 作業部会での検討と私見

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

アクセラレーション研究会 2 2012/9/6

欲しかったのは

FX-78-2 GUNDAM



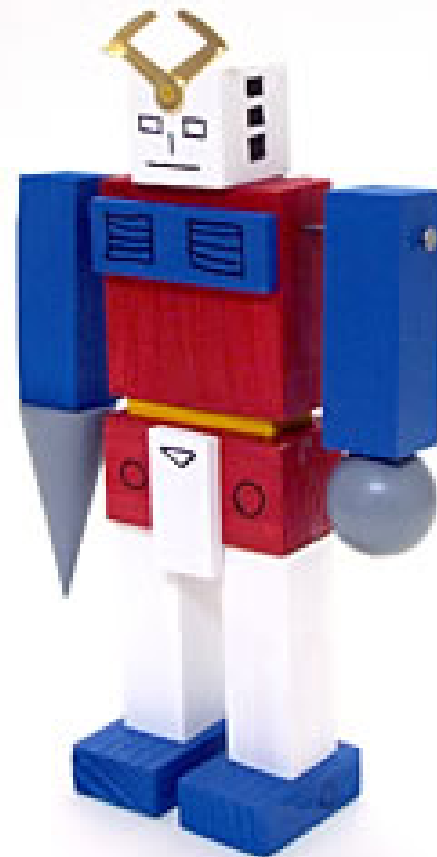
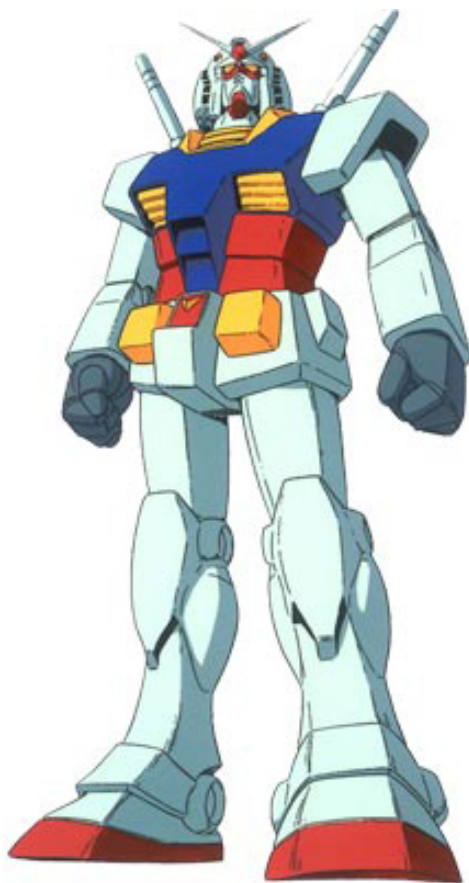
届いたのは



(<http://www.zariganiworks.co.jp/korejanairobo/>)

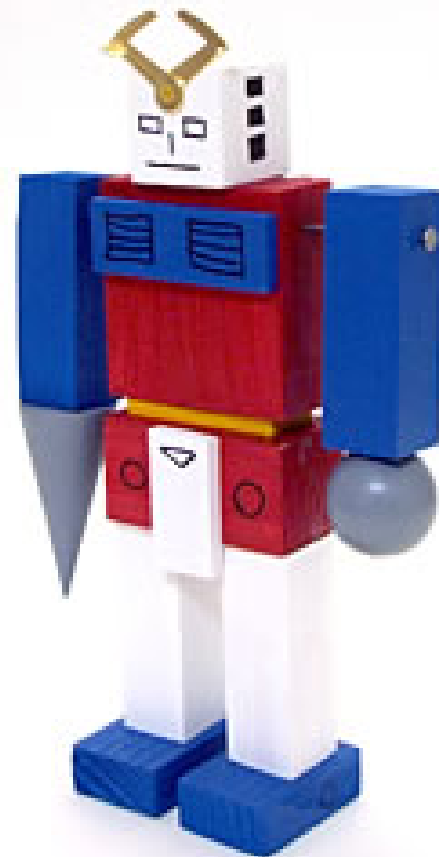
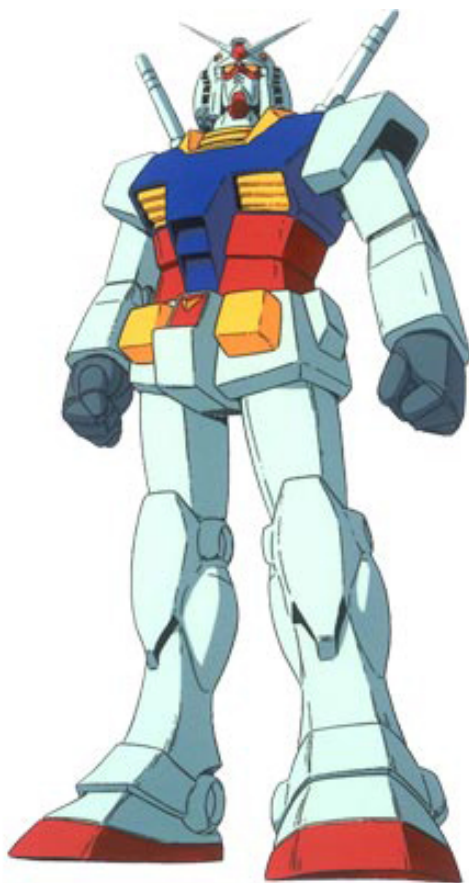
何か違う、、、

FX-78-2 GUNDAM



どうしてこうなった、、、

FX-78-2 GUNDAM



とならないために、、、

概要

- 「京」プロジェクトのデザインと概念設計
- 現在進行中のプロセス
 - － 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
- おまけ: エクサの先、ゼッタ、ヨッタ

「京」プロジェクトのデザインと概念設計

- 大体の時系列
- ターゲットアプリケーション
- 「概念構築に関する共同研究」

大体の時系列

- 2004 年くらいから文科省の下の情報科学技術委員会、計算科学技術推進ワーキンググループで議論
- 2005 年の「中間報告」
 - － 8 分野からの要求をまとめた(ことになっている)
 - － ベクトル 2PF + スカラー 4PF + 「特定処理計算加速機」20PF
- 2006 年度「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」実質的にはデザインコンペ。NFH の他、東大(+天文台)、九大、筑波大等も参加

時系列続き

- 2005 年から 2006 年にかけて CSTP 評価委員会からボロクソに言われる。目標、アーキテクチャを文句がでないように色々変更。
- 2006 年初め (だったと思う) 開発実施本部設置、ヘッド:渡辺 (NEC OB)
- 2007 年夏: ベクトル (?PF) + スカラー (?PF) 合計 10PF 以上、と決定
- 2009 年春: ベクトル担当の N が撤退
- 2009 年 11 月 13 日 (金) 仕分け。「2 位じゃいけないんですか？」
- 2010 年 10 月。プロトタイプ機が Green500 4 位にランクイン
- 2011 年 6 月。8 割完成で Top500 1 位
- 2011 年 11 月。全ノード動作で Top500 1 位。

ターゲットアプリケーション

<http://www.nsc.riken.jp/target-application/target-application.htm>

名前がでたのは 2007/4 だが選定は 2006/4 までになされていた:

- 1 Cavitation 有限差分法によるキャビテーション流れの非定常計算
- 2 COCO 超高解像度海洋大循環モデル
- 3 FrontFlow/Blue Large Eddy Simulation (LES) に基づく非定常流体解析
- 4 FrontSTR 有限要素法による構造計算
- 5 GAMESS FMO分子軌道法計算
- 6 GNISC 遺伝子発現実験データからの遺伝子ネットワークの推定
- 7 LANS 航空・宇宙機解析における圧縮性流体計算
- 8 LatticeQCD 格子QCDシミュレーションによる素粒子・原子核研究
- 9 MC-Bflow 血流解析シミュレーション
- 10 MLTest オーダーメイド医療実現のための統計的有意差の検証

ターゲットアプリケーション(2)

- 11 Modylas 高並列汎用分子動力学計算ソフトウェア
- 12 NICAM 全球雲解像大気大循環モデル
- 13 NINJA/ASURA 天体の起源を探る超大規模重力多体シミュレーション
- 14 Octa 粗視化分子動力学計算
- 15 PHASE 平面波展開第一原理分子動力学解析
- 16 ProteinDF 巨大タンパク質系の第一原理分子動力学計算
- 17 RISM/3D-RISM 溶液内タンパク質の電子状態の 3D-RISM/FMO 法による解析
- 18 RSDFT 実空間第一原理分子動力学計算
- 19 Seism3D 地震波伝播・強震動シミュレーション
- 20 sievgene/myPresto タンパク質・薬物ドッキングシミュレーション
- 21 SimFold タンパク質立体構造の予測

但し

実際に性能評価に使われたのは HPL, FFT と以下の7個

- SimFold, Modylas: 古典MD
- GAMESS, RSDFT: 量子化学
- LANS, NICAM: 流体
- LQCD: 4次元格子でのCG反復

この時点で、

- したい計算
- 高効率でできる計算
- 高効率でできるがオーバースペックな計算

の乖離

大雑把なアプリケーションの特性

- 古典MD、量子化学: メモリバンド幅もネットワークもありたくない
- 流体: メモリバンド欲しい。ネットワークはそこそこ
- QCD: メモリ量たくない。メモリバンド幅もネットワークも欲しい

これからわかること:

- 一台で全て満たそうとすると帯に短し襷に長しになる
- 低レイテンシ要求があるアプリケーションははいってない

「京」のデザインポイント

- CPU:128Gflops, 64GB/s, 16GB
- ICC: リンク 5GB/s x 2 x 10, CPU 20GB/s, レイテンシ:論文に書いてない
- 最近の計算機にしては B/F 重視
- 最近の計算機にしてはネットワーク重視
- CPU, ICC 別チップ

BG/Q との比較

	「京」	BG/Q	比率
ピーク速度 (GF)	128	204.8	—
メモリバンド幅 (GB/s)	64	42.6	2.4
ネットワーク (GB/s)	5	2	4
電力あたり性能 (GF/W)	0.85	2.1	(1/2.5)

演算性能あたり、「京」は BG/Q の 2.4 倍のメモリバンド幅、4 倍のネットワークバンド幅を持ち、2.5 倍電気を食う。

消費電力 $\propto$ メモリバンド幅?

どうしてこうなったか？

形式的な理由： 要求仕様にそう書いてあった

- HPL 10PF
- 電力 30MW 以下
- アクセラレータ付けるならやはり 10PF
- この範囲でアプリケーションの性能を上げること

アプリケーションの側からは

欲しいもの:

俺のコードが速く走る計算機をよこせ!

但し: 「俺のコード」には温度差あり

- 必要ならアセンブラでも
- コンパイラの出力見るのは基本だよね?
- プロファイラ使ったことある
- OpenMP という言葉聞いたことがある
- 並列化ってコンパイラがしてくれるんでしょ?

最近の問題

アセンブラで書くとかくらいではどうにもならないことがある

- **メモリ階層への対応**
- **分散メモリ並列化**

アルゴリズム、データ構造どちらも根本から見直す必要あり

1 ステップもどって考える

何故分散メモリ＋深いメモリ階層なのか？

チップ内の処理能力の増加に比べてチップ間データ転送能力が上がらないから

これは、言い換えると、現在の典型的マイクロプロセッサの設計では

- 演算能力はデータ転送能力に比べて相対的に安い
 - これは製造コストだけでなく電力コストも
- ということ

さらに言い換えると

- 「京」より演算性能を大幅に上げるのは、メモリバンド幅やネットワークバンド幅をあげなくてもいいなら電力・コストを大きくは増やさなくてもできたはず
- 逆に、HPL 10PF だけならもっと安価にもできたかもしれない
- メモリバンド幅を増やすなら全く別の作りかたもあったかも

万能輸送機



こんなのが欲しかったのかも？



現在進行中のプロセス

- 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

昨年度のアプリケーション部会

- 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

前提条件

- 消費電力は 20MW 程度にする
- 想定完成時期は 2019 あたり？
- 想定テクノロジーは 11nm あたり？

何が問題か？

- 消費電力 — CMOS スケーリングが 130nm で終わり、フィーチャーサイズが下がっても駆動電圧下がらなくなった
- 2000 年以前: $P \propto L^{-3}$, 2000 年以後: $P \propto L^{-1}$
- 同一面積、同一クロックで、消費電力が「上がる」
- 2000 年までと同じ方法では 10 年で 100 倍は無理

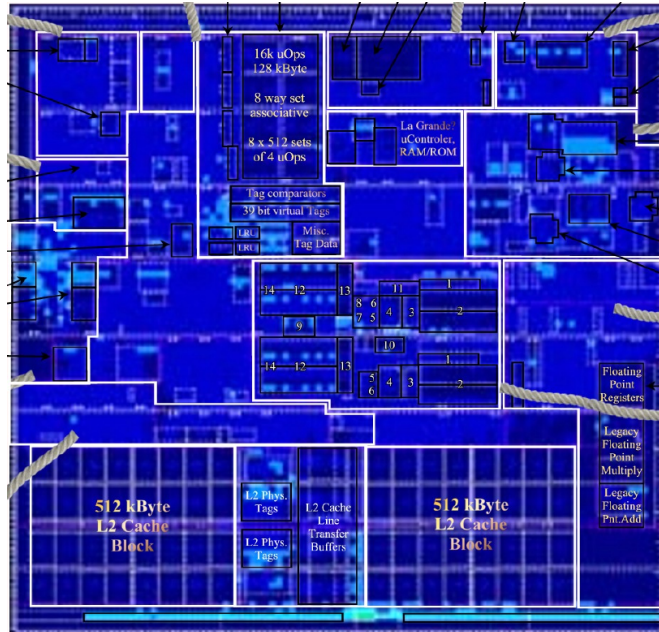
今までなんとかかなっているように見えるのは何故か？

- クロック落としている：2002年あたり：3GHz。今のGPUだと 1GHz あたり
- コアあたりの演算器数増やしている。 SSE2: 1, AVX: 4
- 相対的なメモリバンド幅落としている。 BF 1 $\rightarrow$ 0.2
- 実は電力増えてるがまだ計算機調達費用よりは安い

色々限界

とはいえ

計算ってどれくらい電気食うものか？は作りかたで色々



Intel Prescott (2004)

GRAPE-DR core (2006)

90nm, **7.6 GF**, > 100W? 90nm, 0.4 Gfops, 0.1W

同じ世代の半導体技術で、GRAPE-DR は Prescott P4
の 50 倍のエネルギー効率

上には上がある

プロセッサ	性能 Gflops	電力あたり GF/W	B/F
NEC SX-8	16	0.01?	4
Intel P4	7.6	0.07	0.5
GRAPE-DR	410	4.0	0.01

どれも 90nm、ほぼ B/F に比例、、、

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でも要求は変わる

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。

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:

- 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)	バンド幅 (20MW)
ベースライン	0.1	10-100GB	20MW	1EF	0.1EB/s
SoC	4	5-10MB	2-5MW	4-10EF	16-40EB/s
アクセラレータ	0.001	1-10GB	4-10MW	2-5EF	2-5PB/s
バンド幅重視	1	1TB	120MW	0.15EF	0.15EB/s

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に基づいて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

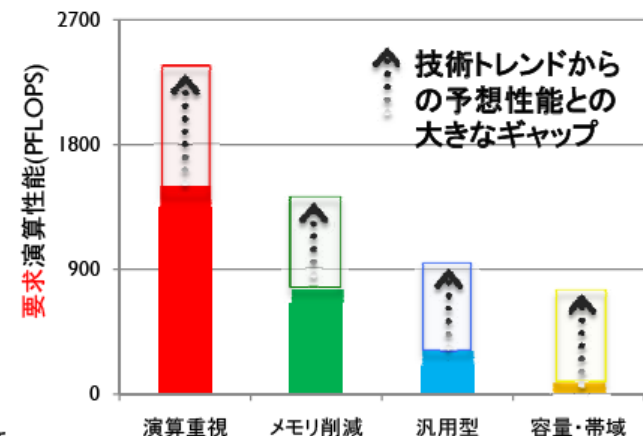
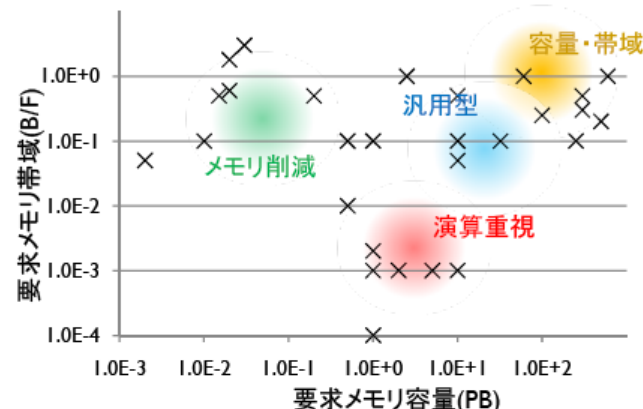


図1. 要求性能と予測性能の相関

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 容量・帯域 演算重視 メモリ削減	ベースライン バンド幅重視 アクセラレータ SoC	「京」の延長 NEC SX-9 の延長 GRAPE-DR,GPGPU ...

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」のどちらかでも実現できる設計解はいまのところ知られてない
- 演算重視は要するに 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

汎用/ベースライン

- 「京」の延長
- B/F 0.1 ~ 0.2?
- 富士通さん頑張って

演算重視 / アクセラレータ

- メモリ帯域が少なくても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?

容量・帯域/バンド幅重視

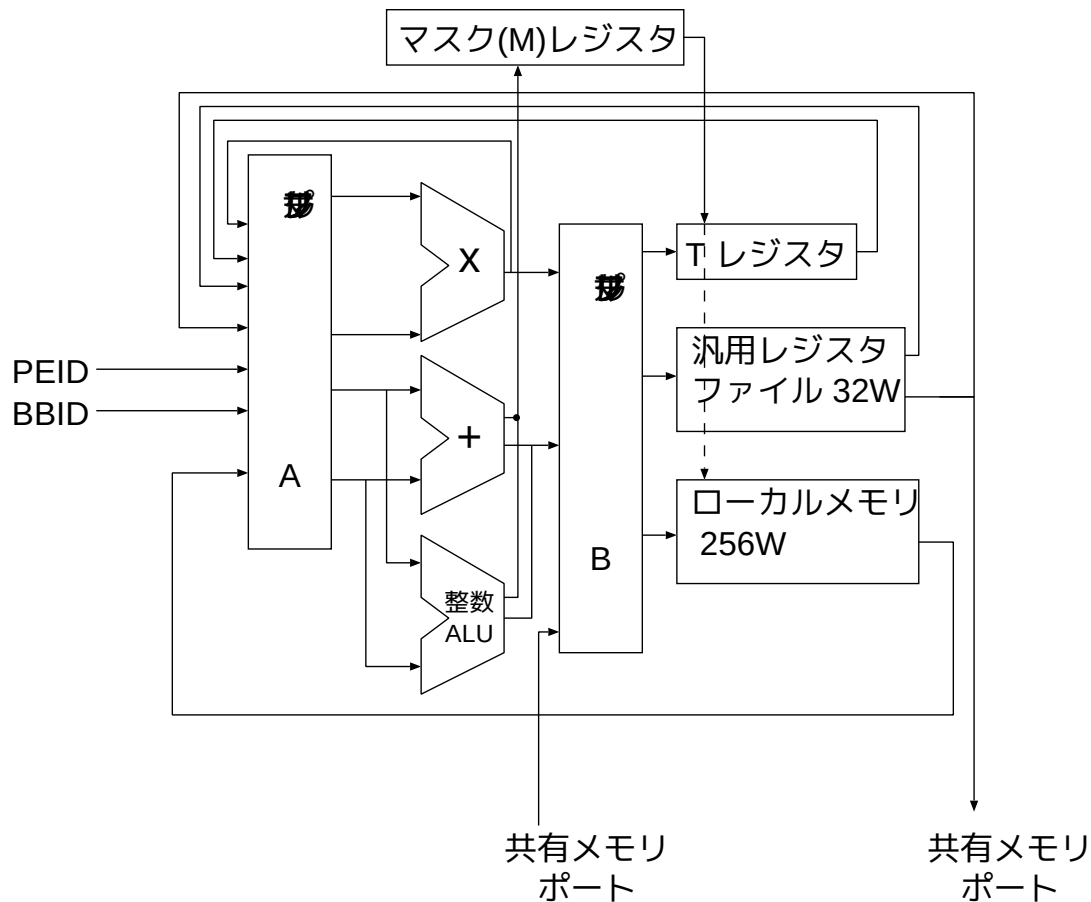
- SX-9 の延長
- NEC さん頑張って、、、

演算重視・メモリ削減の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

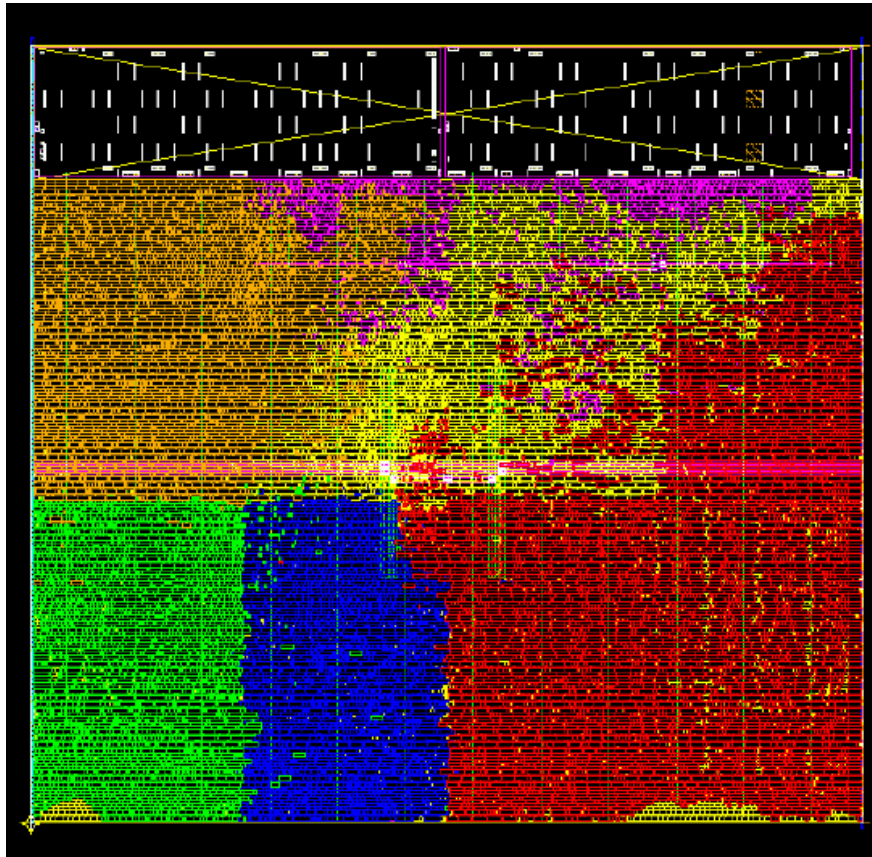
コア内部アーキテクチャ

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

今後の方向

- 今年度、来年度 feasibility study
- その後？5年くらいで開発？

最悪シナリオ

- 3つのFS チームが「競争」に
- 同じようなデザインになる
- 1つ作るだけになる

(まあその、どこかで聞いたような話)

前回のキーポイント：概念構築のためのなんとか

- 概念構築といいつつすでに枠が決まっていた
 - HPL 10PF
 - 電力 30MW 以下
 - アクセラレータ付けるならやはり 10PF
 - この範囲でアプリケーションの性能を上げること
- その結果、各社の提案が収斂した

以下、2012/6 に公開になった当時の会議資料から

(2012/6 公開)

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

(2012/6 公開)

アーキテクチャ案の概要 (汎用システム) 【月末時点】

- 基本仕様 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 m ² 以上	1000-2000 m ²	3000 m ² 以上	1000-2000 m ²

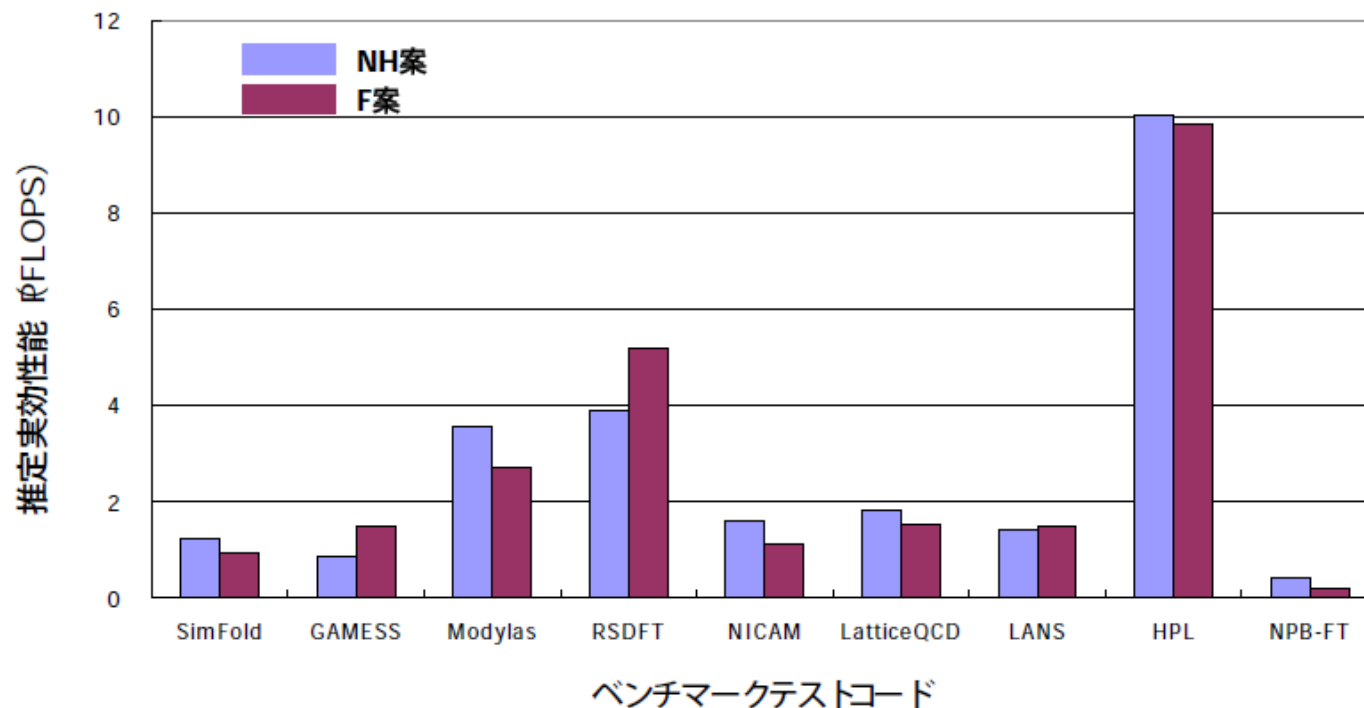
(2012/6 公開)

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 (6: 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	スカラーレジスタ 128本
CPUチップ 計算 ノート	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

(2012/6 公開)

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斂
- さらに消費電力等も収斂
- さらにベンチマーク性能も収斂

おまけ：アクセラレータについては？

アーキテクチャ案の概要 (アクセラレータ) 【月末時点】

■ 基本仕様

- 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

- 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

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
- アプリケーションのことはすっかり忘れられた、、、

教訓的なもの

- プロジェクトは(実現可能なら)設定した目標通りのものが実現される
- 「プログラムは思った通りではなく、書いた通りに動く」
と同じ
- なので、Top500 1 位とか HPCC 1 位とかを目標にしてアプリケーションのことを忘れられては困る
- 目標をちゃんと決めよう

おまけ:エクサの先

牧野の印象: ちゃんと作る方向にいけばエクサは悪くない

- エクサは(アクセラレータやオンチップメモリタイプなら)それほど厳しくない
- 特にオンチップメモリですむアプリケーションはおそらく膨大。
- さらに100倍はどうか？

まずはデバイスの問題。4nm とかその先？ CNT? これは先があるとする、使い道とソフトウェアの問題

使い道とソフトウェア

- もちろん、使えるような計算機かどうかが第一、、、
- ネジ一本一本からシステム全体を、というのが夢物語ではなくなる
- ソフトウェアをどうするかは大問題。連成問題的アプローチでは無理では？