

作るか買うか～専用計算機に未来はあるか

国立天文台
理論研究部/天文シミュレーションプロジェクト (CfCA)
牧野淳一郎



話の構成

- 牧野の 1995 年の予測
- 2009 年の現実
- 牧野の 2009 年の予測
 - マイクロプロセッサの進化
 - 「専用機」の方向
 - * GRAPE-DR
 - * GPGPU
 - * FPGA
 - * Structured ASIC
- まとめ

話の構成

- 牧野の 1995 年の予測
- 2009 年の現実
- 牧野の 2009 年の予測
 - マイクロプロセッサの進化
 - 「専用機」の方向
 - * GRAPE-DR
 - * GPGPU
 - * FPGA
 - * Structured ASIC
- まとめ

朴さんの話が地べただとすると牧野の話は泥の中

牧野の 1995 年の予測

計算科学と「専用」計算機 パリティ 1996 年 4 月号

grape.mtk.nao.ac.jp/pub/people/makino/papers/parityhtml/parityhtml.html

1. ベクトル並列に将来はない
2. マイクロプロセッサの今後の性能の伸びは遅い
3. 専用アーキテクチャの将来は (誰かが作れば) 明るい

時代背景

- NAL NWT 完成、地球シミュレータの計画はまだ始まってない
- CP-PACS 開発中
- GRAPE-4 できたところ

予測はあたったか？

1. ベクトル並列に将来はない — ○
2. マイクロプロセッサの今後の性能の伸びは遅い
— ○ (14年で100倍程度)
3. 専用アーキテクチャの将来は (誰かが作れば) 明るい
— 誰か作った？
(GRAPE-6, MD-GRAPE, QCDOC(BG/L))

専用アーキテクチャが(牧野が期待したほど)広がらなかった理由

- マイクロプロセッサ (PC) は速くはならなかったけど安くなった
- 専用チップを作るための初期コストがどんどん上がった
 - 新規参入が難しくなった

マイクロプロセッサの価格が さがった理由

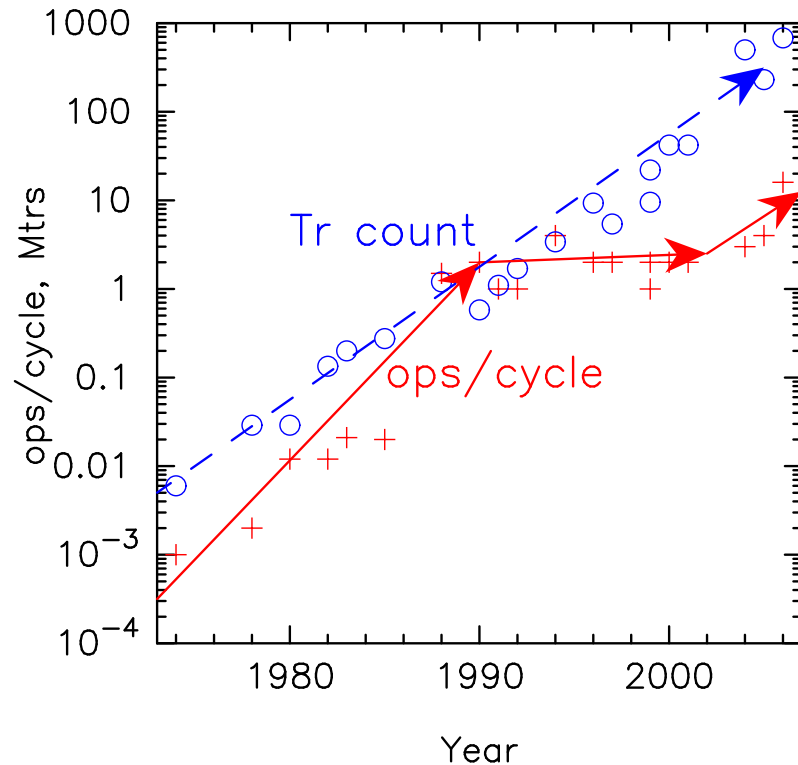
価格がさがったのではなくて、

- 安いものは順調に性能があがった (x86)
- 高いものは性能上がらなくなった (Alpha, SPARC, Power, Itanium, ...)

理由:

- 結局、1つの CPU コアで使い切れる以上のトランジスタが利用可能になったため
- 大量生産と競争の効果: 安価で大量生産されるもののほうが高価なものより速い
 - Intel の場合: デスクトップ用 : Xeon : Itanium

マイクロプロセッサの「進歩」



- 現在のマイクロプロセッサ:演算器の割合は $1/1000$
- 将来にわたって上がらない
- $1/10$ を演算器に使えるば専用機は優位

開発費

1990	1 μ m	1500万円
1997	0.25 μ m	1億円
2004	90nm	3億円以上
2008	45nm	10億円以上

何故上がるか？

- 製造プロセスの複雑化
- 設計の、人手に頼る部分の増大

人手の増大 — GRAPE の場合

年	機械	配置配線の方法、コスト
1990	GRAPE-3	自動
1992	GRAPE-4	ほぼ自動 (エンジニア 1 名、数日)
1997	GRAPE-6	半年、10 名以上 (やり直しあり)
2005	GRAPE-DR	半年、30 名以上 (やり直しなし、上海で)

- こんな調子では専用アーキテクチャなんてありえない
- 元々、配置配線なんて人間がするのがおかしい

QCD 専用機の場合

筑波

アメリカ

ヨーロッパ

90 年代

CP-PACS

QCDSP

APEmille

2000 年代前半 ?

QCDOC

apeNEXT

2000 年代後半 PACS-CS ??

QPACE?

黒:カスタムプロセッサ使うもの

赤:市販プロセッサ使うもの

2009年での将来の方向

- LSI 設計の自動化は退化を続ける
- このため、例えば GRAPE のような応用が狭いものは成り立たなくなる (というか、すでにそうになっている)
- 一方、マイクロプロセッサは今後さらに非効率なものになっていく
- 従って、もうちょっとましなものを作るのはそれほど難しくないが、コストはかかる

マイクロプロセッサの今後の進化

過去 10 年に起こったこと:

- x86 everywhere
- デスクトップ用とサーバー用 CPU の収斂
- 結果的に高性能で安価な CPU が市場に

今後はどうか？

- デスクトップ(ノート)用: 高性能必要でない？
- サーバー用: 高くてもよい？
- 「高くて速い」と「安いけど遅い」に分化？

PC クラスタには悪いニュース？

×安いクラスタ: ソケット当り 10 万円

△高いクラスタ: ソケット当り 50 万-200 万円

今後数年の CPU 開発

- Intel

- サーバー: Nehalem (09, 8 コア), Westmere/Sandy Bridge(09/10)
- デスクトップ: Nehalem (08, 4 コア), Westmere/Sandy Bridge(6 コア)

- AMD

- サーバー: Istanbul (09, 6 コア), Magny Cours (10, 12 コア)
- デスクトップ: Deneb (09, 4 コア), Sao Paolo(10, 6 コア)

デスクトップは速くならない？

どこを「まし」にできるか？

計算機の性能を決める3要因:

- 演算器の数、動作クロック
- メモリバンド幅
- ネットワーク性能

メモリバンド幅: オンチップなら増やせる。チップ外では高価

ネットワーク: 専用化(FPGA, 市販物理層LSI)で結構できる

演算器は？

演算器を増やす？

- GRAPE-DR 的 SIMD 超並列
- GPGPU 的 SIMD 超並列
- MIMD 超並列
- FPGA
- Structured ASIC

未来は予測するものではなくて、作るもの

GRAPE-DR の基本的な考え

- 応用に特化し、多数の演算器を1チップに集積、並列動作させて高い性能を得た専用計算機の特徴を生かす
- しかし広い応用範囲を実現する

そんなことができるか？が問題

多数の演算器を詰め込む方法

境界条件: メモリバンド幅は増やしたくない (システムコストはほぼメモリバンド幅で決まる)

可能な方策

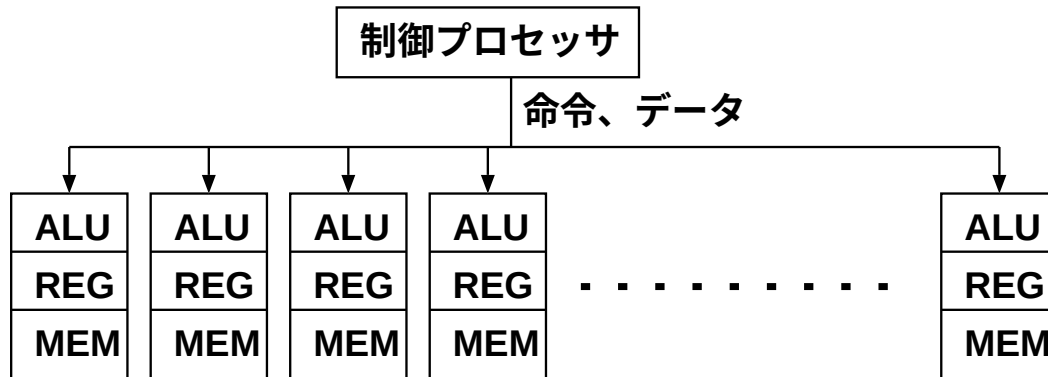
1. GRAPE 的専用パイプラインプロセッサ
2. 再構成可能プロセッサ
3. SIMD 並列プロセッサ

SIMD 並列処理って？

- 古典的 SIMD 並列計算機
- SSE、MMX とかの SIMD 拡張命令
- GRAPE-DR における SIMD

古典的SIMD 並列計算機

Illiac IV, Goodyear MPP, ICL DAP, TMC CM-2,
MASPAR MP-1

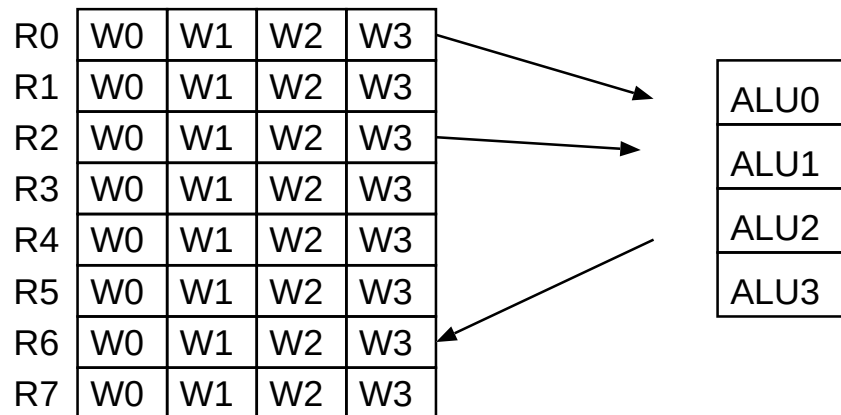


- 1960年代に発生、80年代に絶滅
- 半導体技術の向上に対応できないアーキテクチャ: 計算速度とメモリアクセス速度が比例する必要あり。
- メモリ階層をつける: プロセッサが複雑になりすぎて SIMD の意味が無くなる。

SIMD 拡張命令

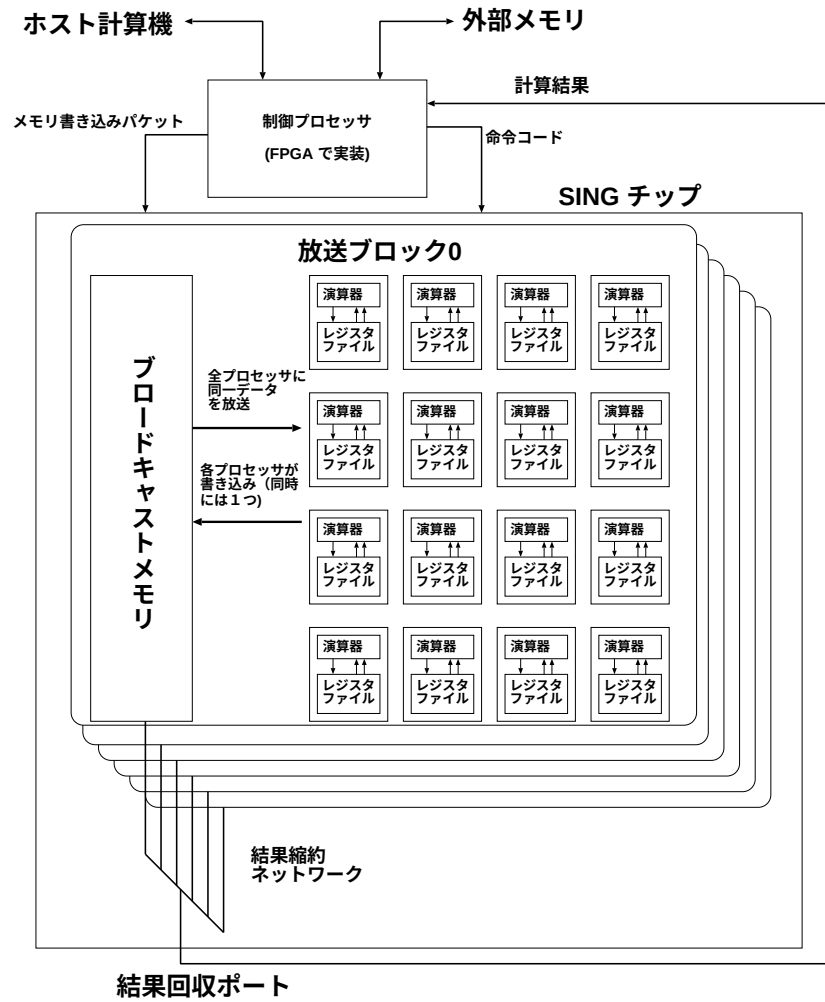
SSE_x が有名

128 ビットなり 64 ビットのデータを 4 語に区切って、それぞれの要素に対する演算を 4 個の演算器で同時に処理



1つのプロセッサの中の話: キャッシュとデータをやりとり
並列度 4 程度が限界? それ以上増やすとキャッシュの速度が追いつかなくなる。

GRAPE-DR における SIMD



- 非常に多数のプロセッサエレメント (PE) を 1 チップに集積
- PE = 演算器 + レジスタファイル (メモリをもたない)
- チップ内に小規模な共有メモリ (PE にデータをブロードキャスト)。これを共有する PE をブロードキャストブロック (BB) と呼ぶ。
- 制御プロセッサ、外部メモリへのインターフェースを持つ

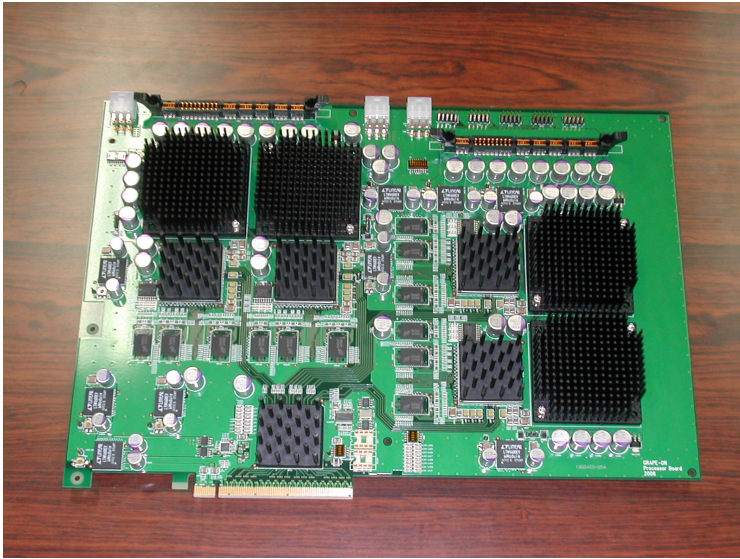
古典的SIMD 並列計算機との違い

- PE のメモリを小さくすることで、外部メモリインターフェースを省略
- 粒子間相互作用等で有用な、総和計算用ネットワーク
- 利点
 - あるトランジスタを全部演算 PE に使える。
- 欠点
 - 沢山メモリがいる計算はできない (QCD は微妙)
 - PE 間的高速ネットワークがいる計算もできない (現行の GRAPE-DR では QCD は無理)

プロセッサチップとプロトタイプボード チップ写真



Processor board

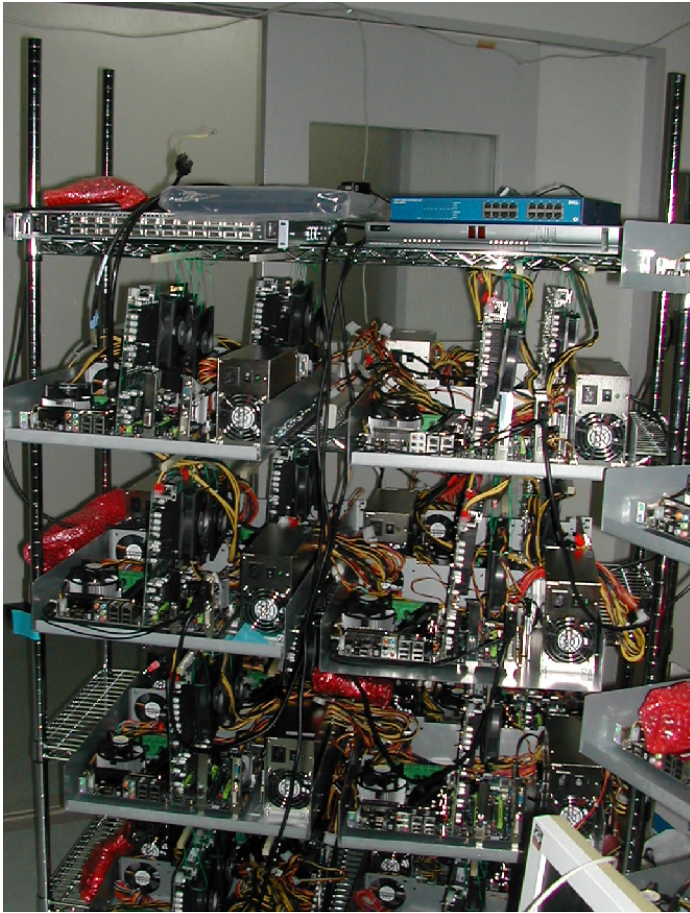


PCIe x16 (Gen 1) interface
Altera Arria GX as DRAM
controller/communication
interface

- Around 250W power consumption
- Not quite running at 500MHz yet...
(FPGA design not optimized yet)
- 900Gflops DP peak
(450MHz clock)
- Available from K&F
Computing Research

GRAPE-DR cluster system

Just to show that the system exists...



Host computer: Intel Core 2
Quad Q6600 with nVidia 780i
chipset

8GB memory

Network: IB (4x DDR)

HPC Linpack passed (not tuned
yet....)

The system and (preliminary)
performance numbers submitted
to TOP500

Major concern: Effective host
memory bandwidth

GDR cluster in 2009

- Nehalem with 3way DDR3 memory should resolve bandwidth problem.
- IB network
- 800T-1P DP peak range.

アプリケーション

「粒子間相互作用」の形ならなんでも。
それ以外:

- 格子・粒子相互作用
- 格子・格子相互作用
- 普通の、依存性のない並列計算 (メモリにはいれば))
- 行列乗算

量子化学計算、2 電子積分を実装中 (理研・名古屋大学)

QCD は？

- QCD 計算はあまりメモリいらないので、このアプローチで OK
- チップ間ネットワークは問題
 - これはどんなアプローチでも問題
 - 1チップの性能が高くと、チップ外の配線総数は減るので他のアプローチより有利

GRAPE-DR 的方向のまとめ

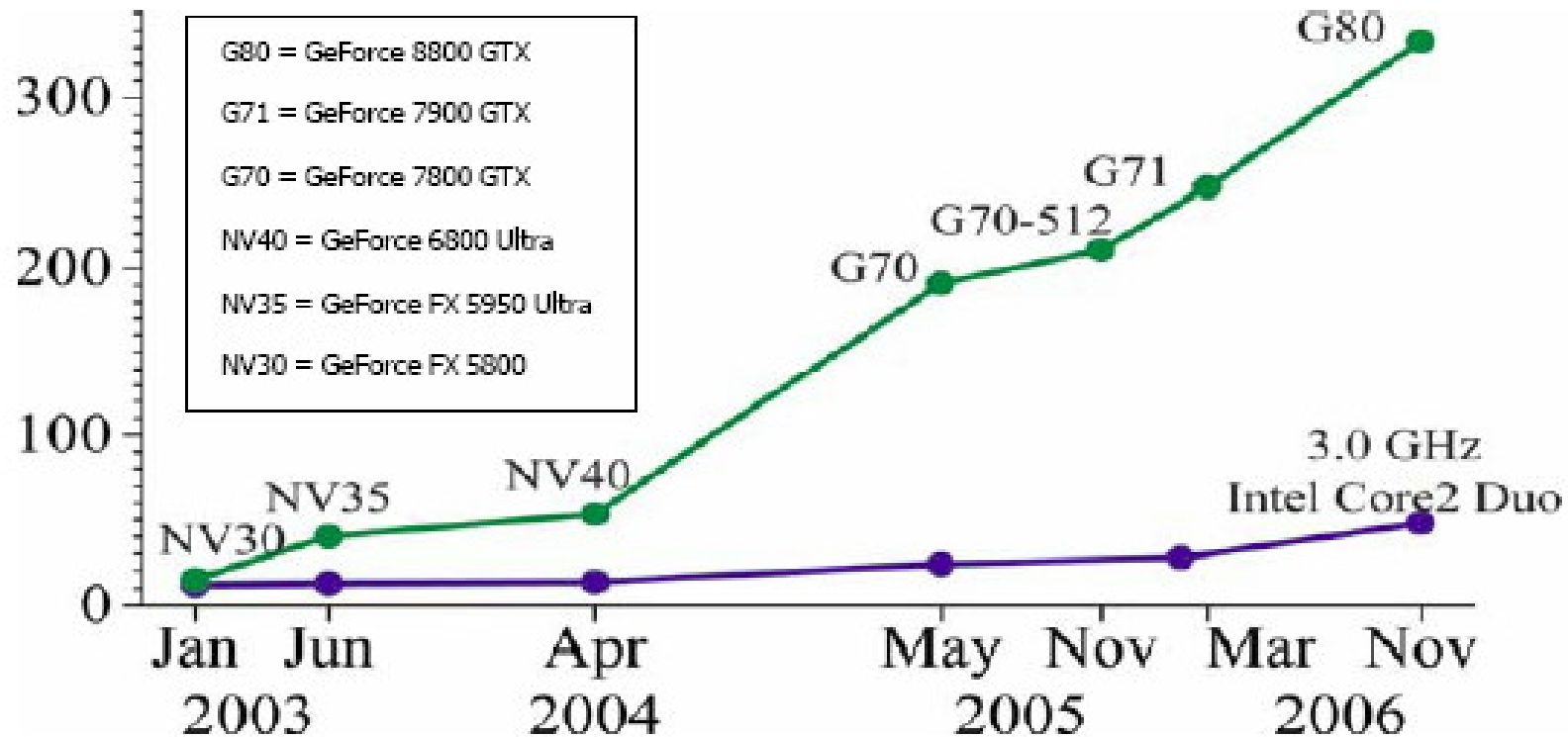
- メモリバンドが不要な計算では高い性能
 - そういうアプリケーションは結構ある
 - もちろん、駄目なものもある
 - アプリケーションの問題かアルゴリズムの問題かの検討が必要
- 開発費を獲得できるかどうかが主な問題

GPGPU 的方向

- グラフィック用プロセッサが、専用パイプラインプロセッサからプログラマブル SIMD プロセッサに変化 (2006, nVidia 8800GTX)
- 単精度浮動小数点演算で 512 Gflops (2006)。現在は 1.2Tflops
- 最近のは倍精度も

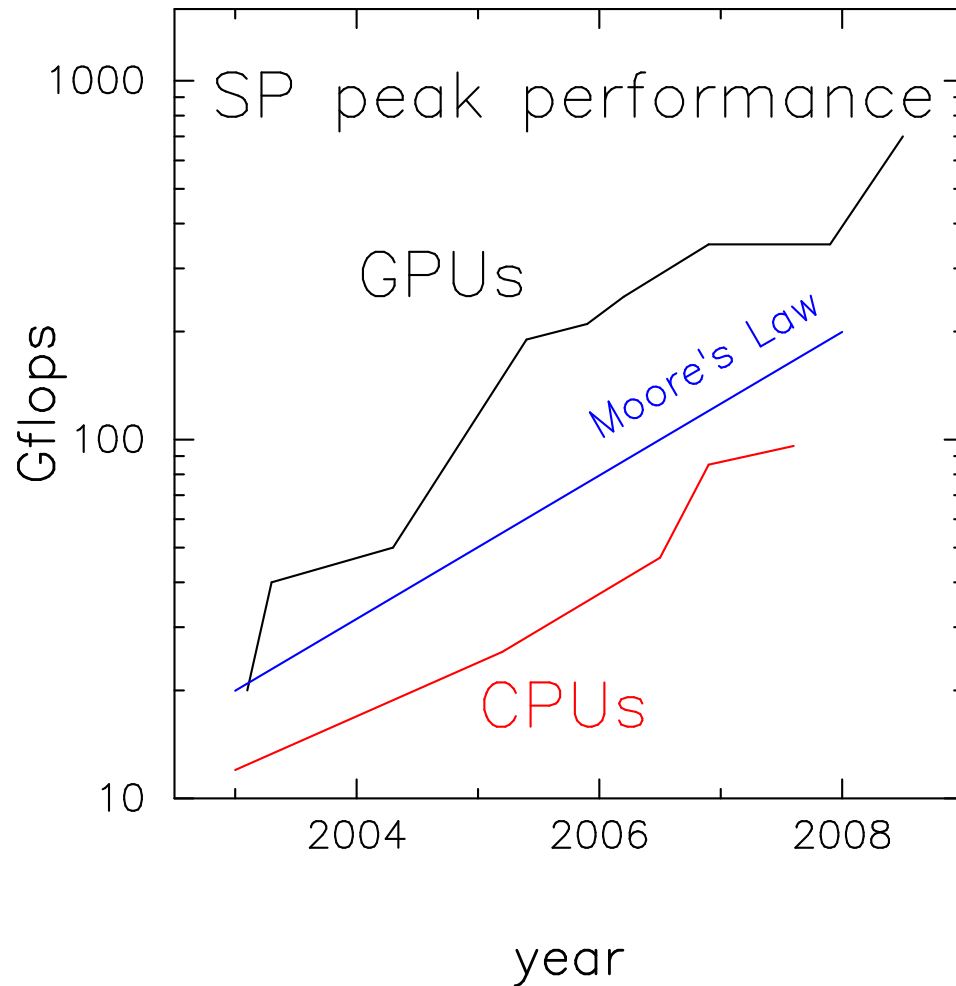
GPGPUs — メーカーさんの見せる図

GFLOPS



「ムーアの法則より速い!」

GPGPUs — 対数グラフでは



- 2005 年以降減速
- 汎用マイクロプロセッサがキャッチアップ
- 倍精度 ??
- メモリバンド幅リミット

GPGPU の将来性

- いわゆる COTS (Commercial off-the-shelf)
- 言い換えると、「誰かが他の目的のために作ったもの」

「他の目的」は指数関数的な性能向上をいつまで必要とするか？(デスクトップCPUと同じ問題)

メーカーさんが GPGPU といっている = 「他の目的」に翳りが見えてきている

やっぱり高くなるかも？(HPC用に売ってるのは現在も高い)

MIMD 超並列

今日は省略

FPGA

- 計算精度低くてよければ大変良い
 - ハードウェア乗算器: 9 ないし 18ビットまで
 - PROGRAPE-4, GRAPE-7: カード 1 枚で Tflops レベルの速度
- チップはメーカーが作ってくれるので絶対ある
- メモリバンド幅の制限があるのは SIMD 並列と同じ

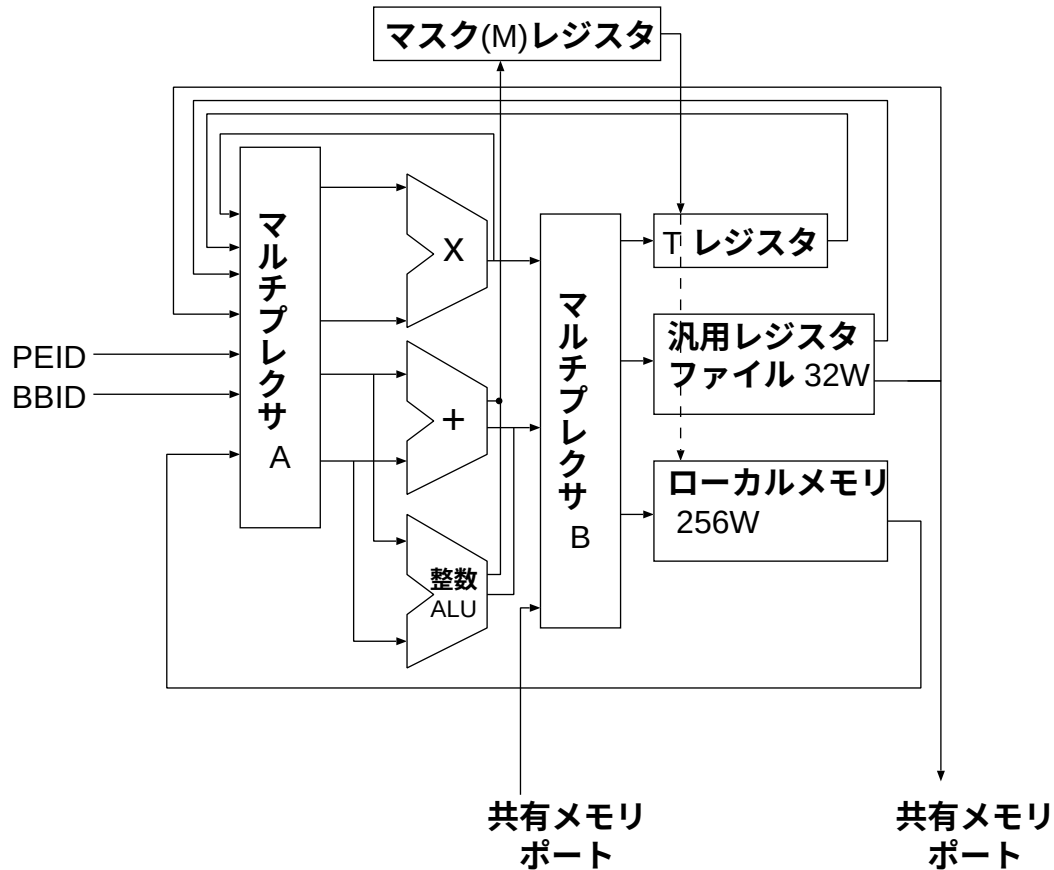
Structured ASIC

- その昔 (90 年代前半くらいまで) のゲートアレイの復活
- FPGA メーカー: プログラマブルな配線等の 1 部をマスクでやることで、ダイサイズ、消費電力を削減 ($1/3$ - $1/4$ に)
- 専業メーカーも (eASIC 等)
- 45nm プロセス製品の場合、フルカスタムに比べて集積度で $1/5$ - $1/10$, 速度は $1/2$ - $1/3$
- 開発コストは $1/100$
- 会津大学・天文台・KEK で実験中

まとめ

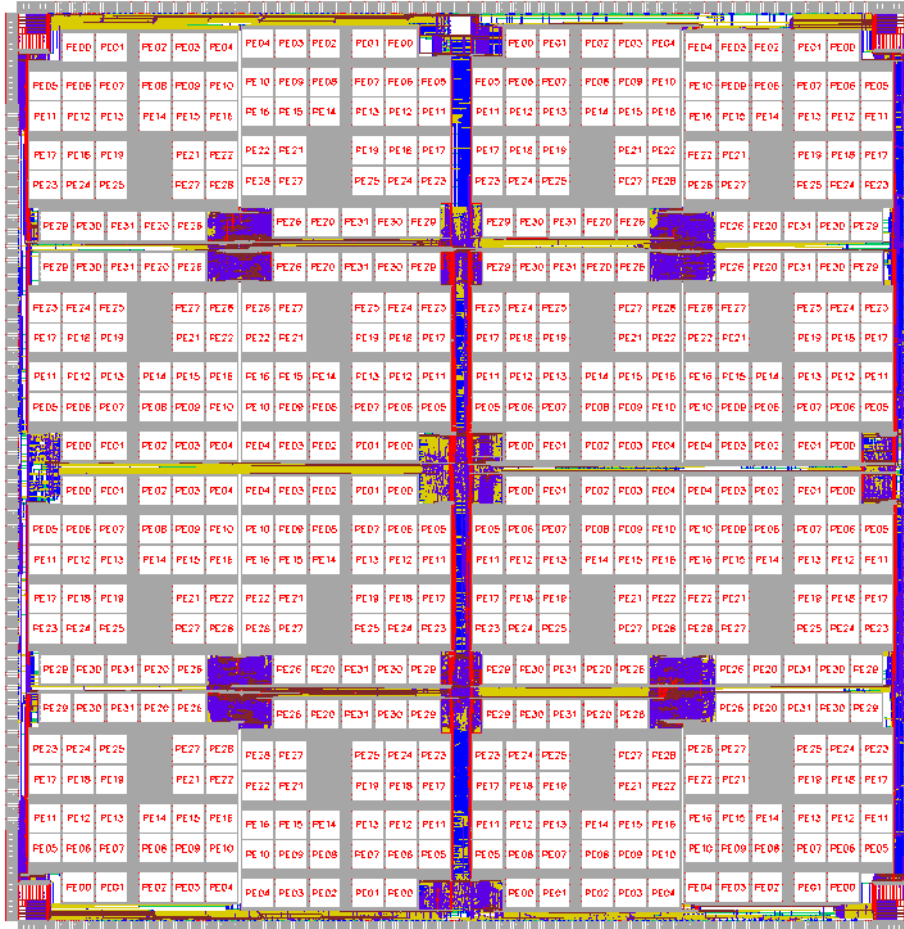
- 汎用マイクロプロセッサに代わるものはでてくるべき時期
- それが専用アーキテクチャかどうかは？
 - FPGA は段々普及する
 - 超並列 SIMD とかは予想よりは作る (作れる) かどうかの問題
 - GPGPU は ??
 - Structured ASIC も可能性あり？
- どういうアプローチでも、メモリバンド幅の制約は回避する必要あり
 - アルゴリズム、実装の問題

PE の構造



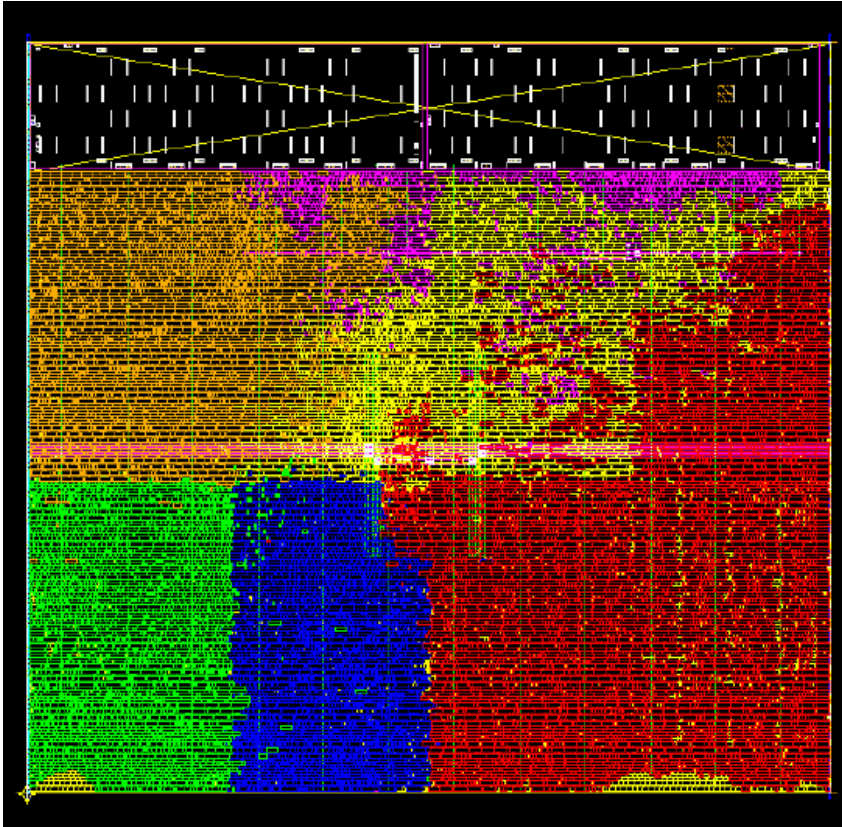
- 浮動小数点演算器
- 整数演算器
- レジスタ
- メモリ 256 語
(K とか M では
ない。)

レイアウト



- 32PE(要素プロセッサ)のブロックが16個
- 空いているところは共有メモリ
- チップサイズは18mm 角

PE レイアウト



0.7mm by 0.7mm

Black: Local Memory

Red: Reg. File

Orange: FMUL

Green: FADD

Blue: IALU

MIMD 超並列

今日は省略

- メーカーやアカデミック計算機科学での研究は一杯ある
- 経験的には、HPC の進歩はそういうところからはでてこないかも、、、
- メモリバンド幅の制限があるのは SIMD 並列と同じ