

GRAPE-DR と次世代スーパーコン ピュータ開発と天文台の関係

牧野淳一郎

今日の話の構成

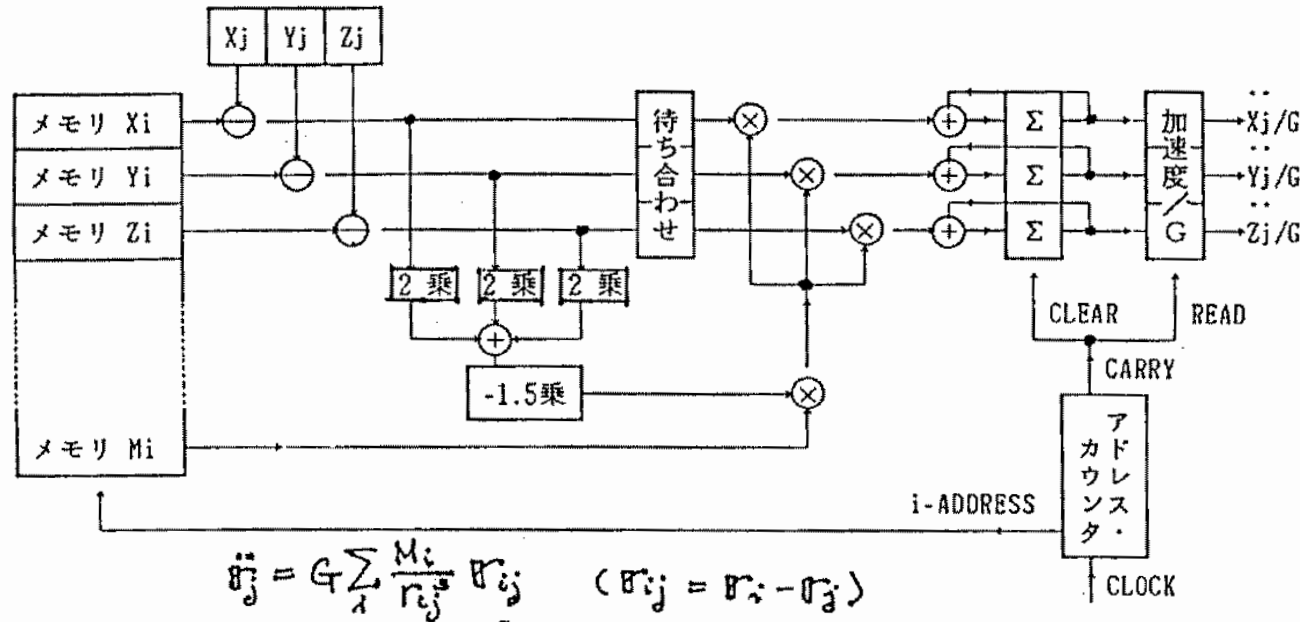
- GRAPE-DR について
- 次世代スーパーコンピューター開発
- 天文台になんか関係あるか？

GRAPE の考え方

- **重力多体問題 (強結合プラズマ、分子動力学): 粒子間相互作用の計算が計算量のほとんど全部**
- **効率のよい計算法 (Barnes-Hut tree, FMM, Particle-Mesh Ewald(PPPM) ...): 粒子間相互作用の計算を速くするだけでかなり加速できる**
- **そこだけ速くする電子回路を作る (「計算機」というようなものではない)**

近田提案

1988 年、天文・天体物理夏の学校



+, -, ×, 2乗は1 operation, -1.5乗は多項式近似でやるとして10operation 位に相当する。
 総計24operation.

各operationの後にはレジスタがあって、全体がpipelineになっているものとする。

「待ち合わせ」は2乗してMと掛け算する間の時間ズレを補正するためのFIFO (First-In First-Out memory).

「Σ」は足し込み用のレジスタ。N回足した後結果を右のレジスタに転送する。

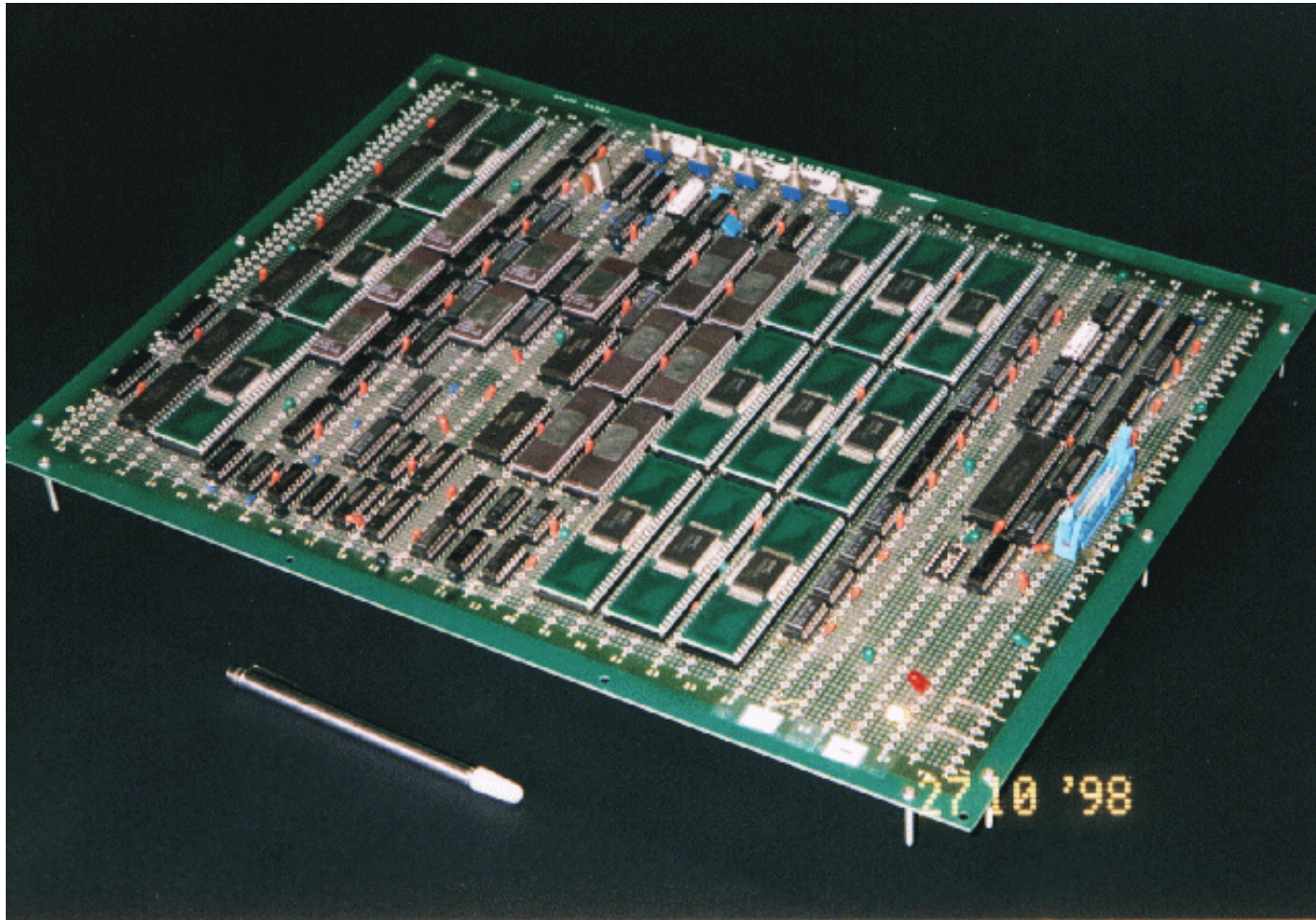
図2. N体問題のj-体に働く重力加速度を計算する回路の概念図。

近田さんによる見積もり

- 32 ビット固定小数点
- IC 200 個
- 体積 0.1m^3
- コスト 400 万

「但し、近田電子製作所の見積もりは甘いという声もあることを付け加えます」

GRAPE-1(1989)

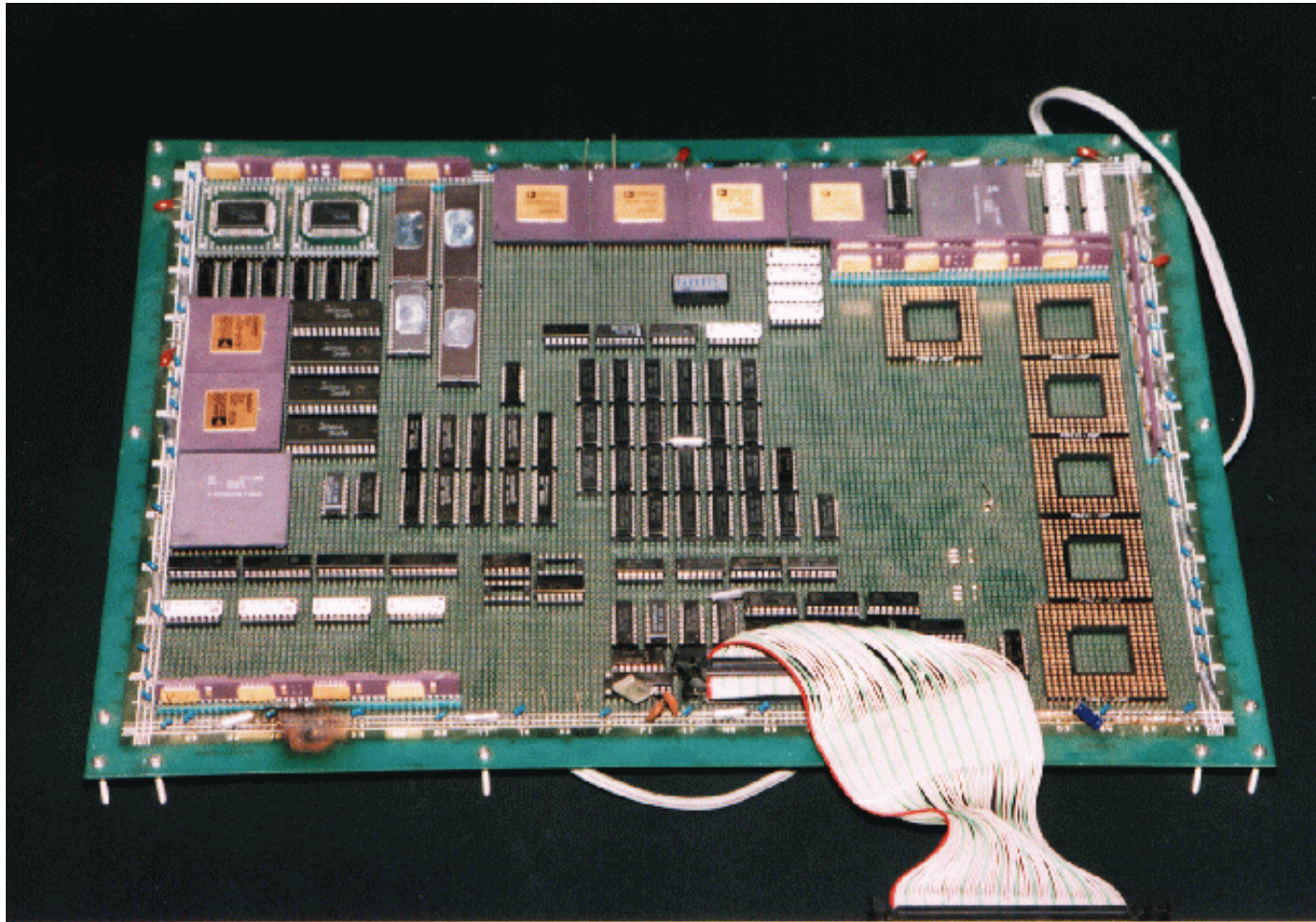


GRAPE-1 の中身

- 演算毎に語長指定。固定 16—対数 8—固定 32—固定 48
- IC 100 個
- コスト 20 万
- 240Mflops 相当
- 伊藤、牧野、戎崎

近田さんの他、慶応の川合研の朴さんに色々教えてもらった。

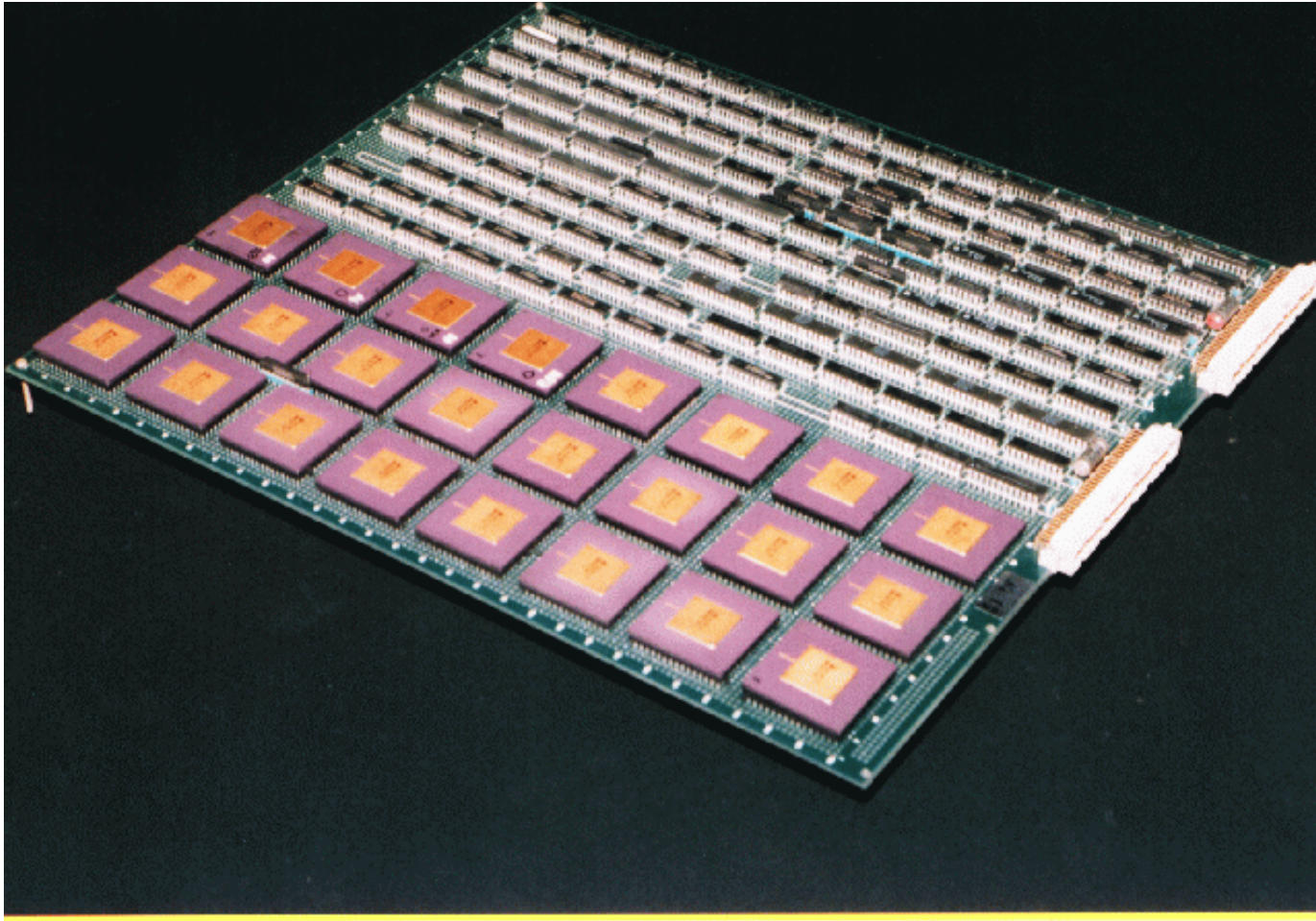
GRAPE-2(1990)



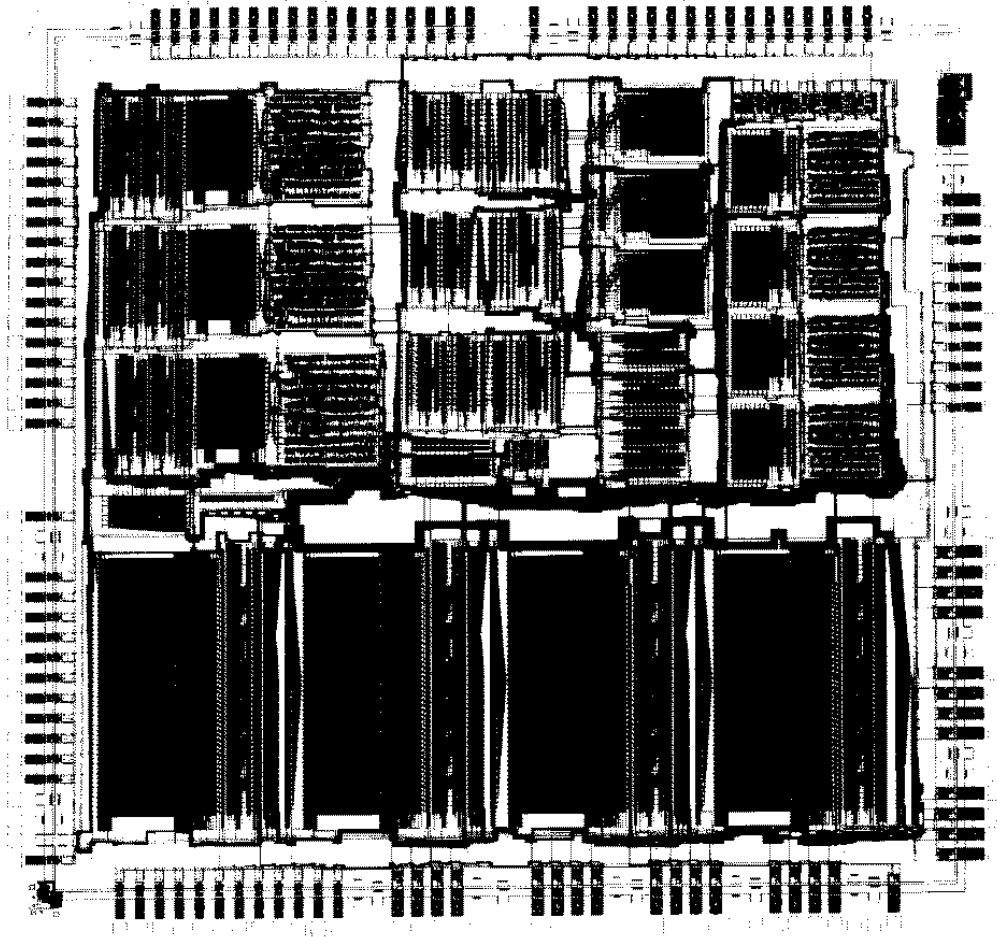
GRAPE-2 の中身

- 8ビット演算とかは止めて普通に浮動小数点演算 (倍精度は最初と最後だけ)
- 40Mflops
- 天文台から開発費をいただいた
- 戎崎、伊藤、牧野

GRAPE-3(1991)



GRAPE-3 チップ



2 mm

GRAPE-3

- 仕様決定、シミュレータ (C で記述) は牧野
- 論理設計以降は富士ゼロックス (橋本、富田)
- SCS Genesil で設計
- ファブは NS. $1\mu\text{m}$
- ボードは奥村 (現在国立天文台)
- 科研費と山田科学振興財団？

GRAPE-4(1995)



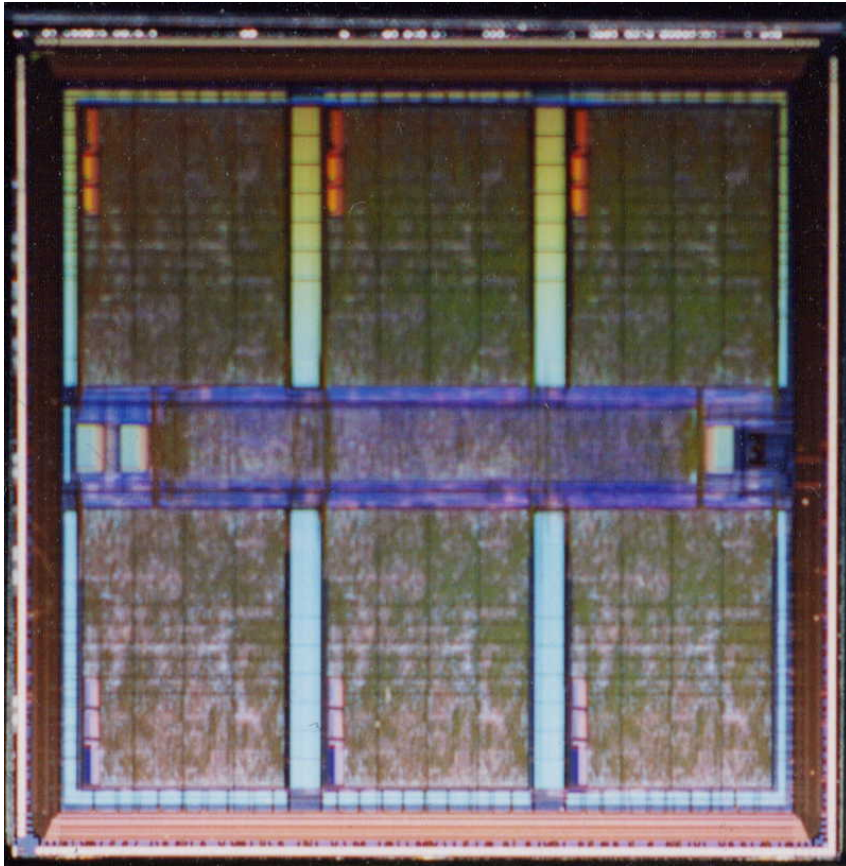
GRAPE-4 概要

- 48 個プロセッサチップがのったボード 36 枚
- 1 チップ 20 演算、32 MHz 動作で 640 Mflops
- LSI logic 1 μ m スタンダードセル
- プロセッサチップは泰地、「予測子」チップ牧野
- 特別推進研究。3 億円弱

GRAPE-6(2001)

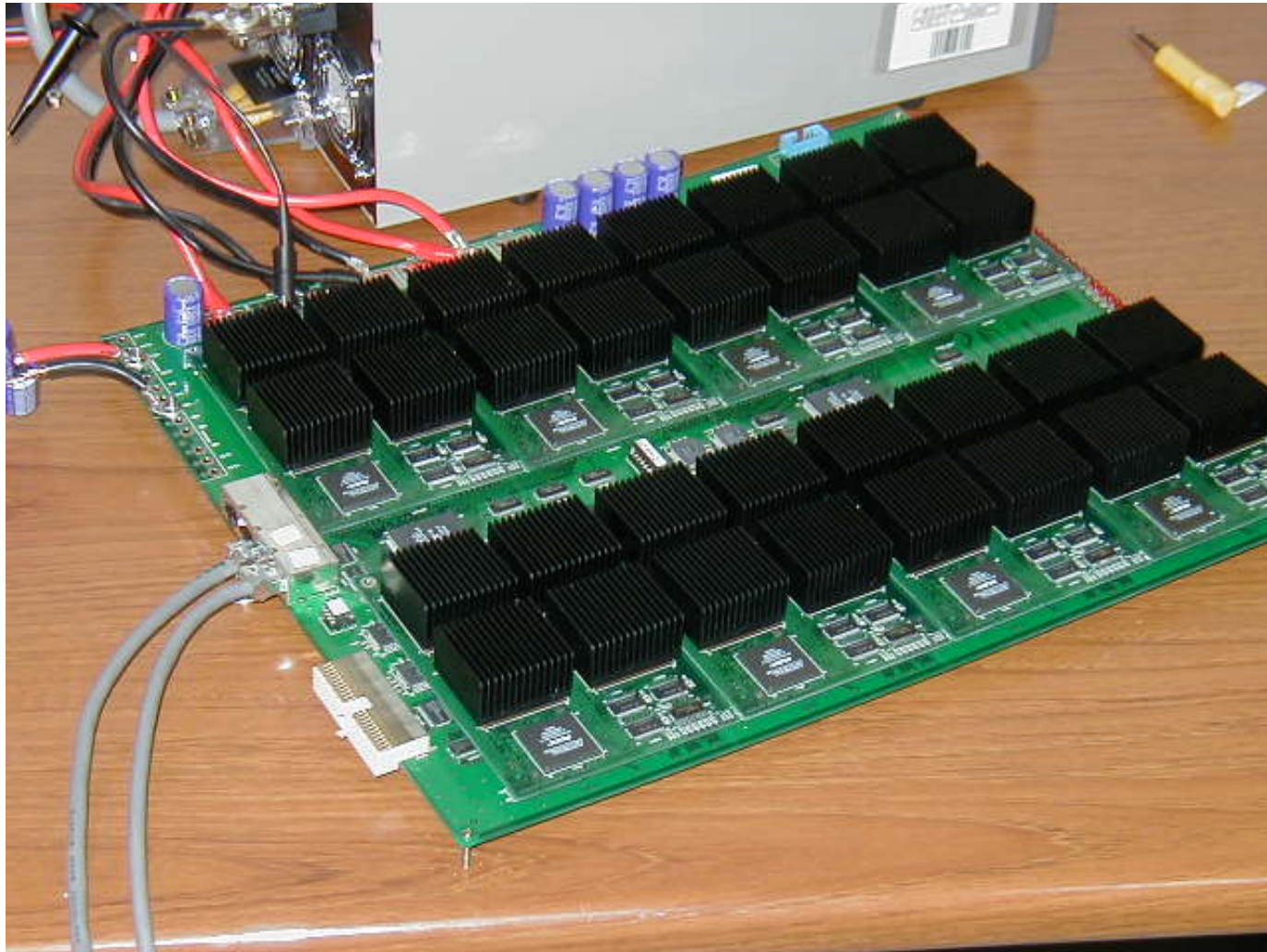
- プロセッサチップ
- プロセッサモジュール
- プロセッサボード
- ネットワークボード
- 全体

パイプライン LSI

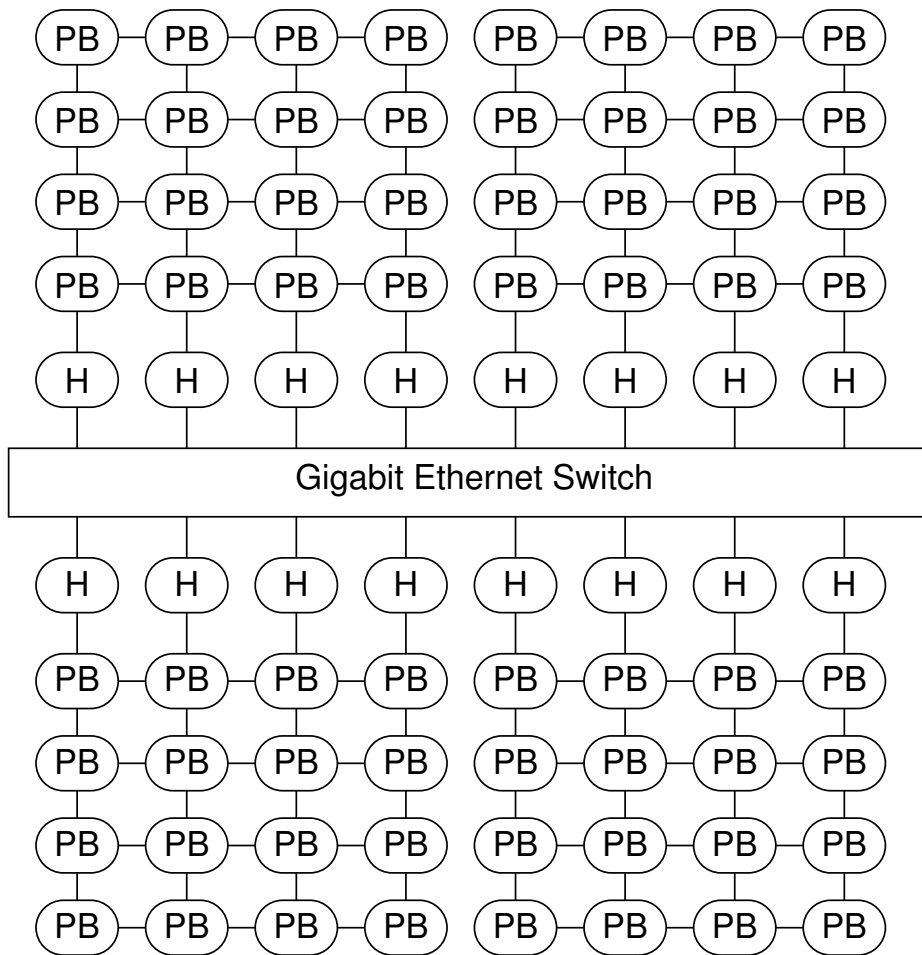


- 0.25 μm ルール
(東芝 TC-240, 1.8M
ゲート)
- 90 MHz 動作
- 6 パイプラインを集積
- チップあたり 31 Gflops

GRAPE-6 processor board



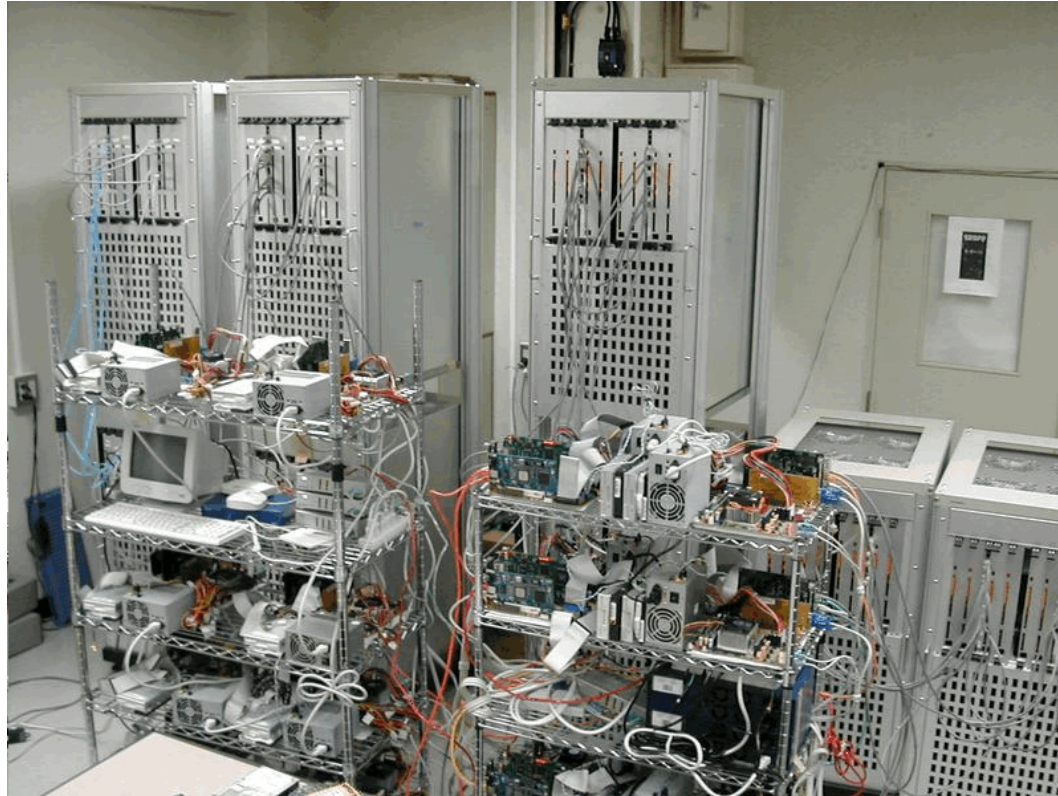
The full 64 Tflops GRAPE-6 system



- **ホスト 4 台、プロセッサボード 16 枚のブロック**
- **4 ブロックで全体システムを構成。結合は GbE**

専用ネットワークと汎用ネットワークの組み合わせ。

The 64-Tflops GRAPE-6 system



2002年現在の 64
Tflops システム

4 ブロック

16 ホスト

64 プロセッサボード

GRAPE-6 の開発体制

- 予算は「未来開拓」。研究費バブルの頃、、、
- 筑波大学と分担して「次世代超並列計算機開発」
- 東大分予算およそ 5 億
- 東大分代表は牧野。LSI は牧野・福重で設計。ボードは一部大学院生 (古賀君)

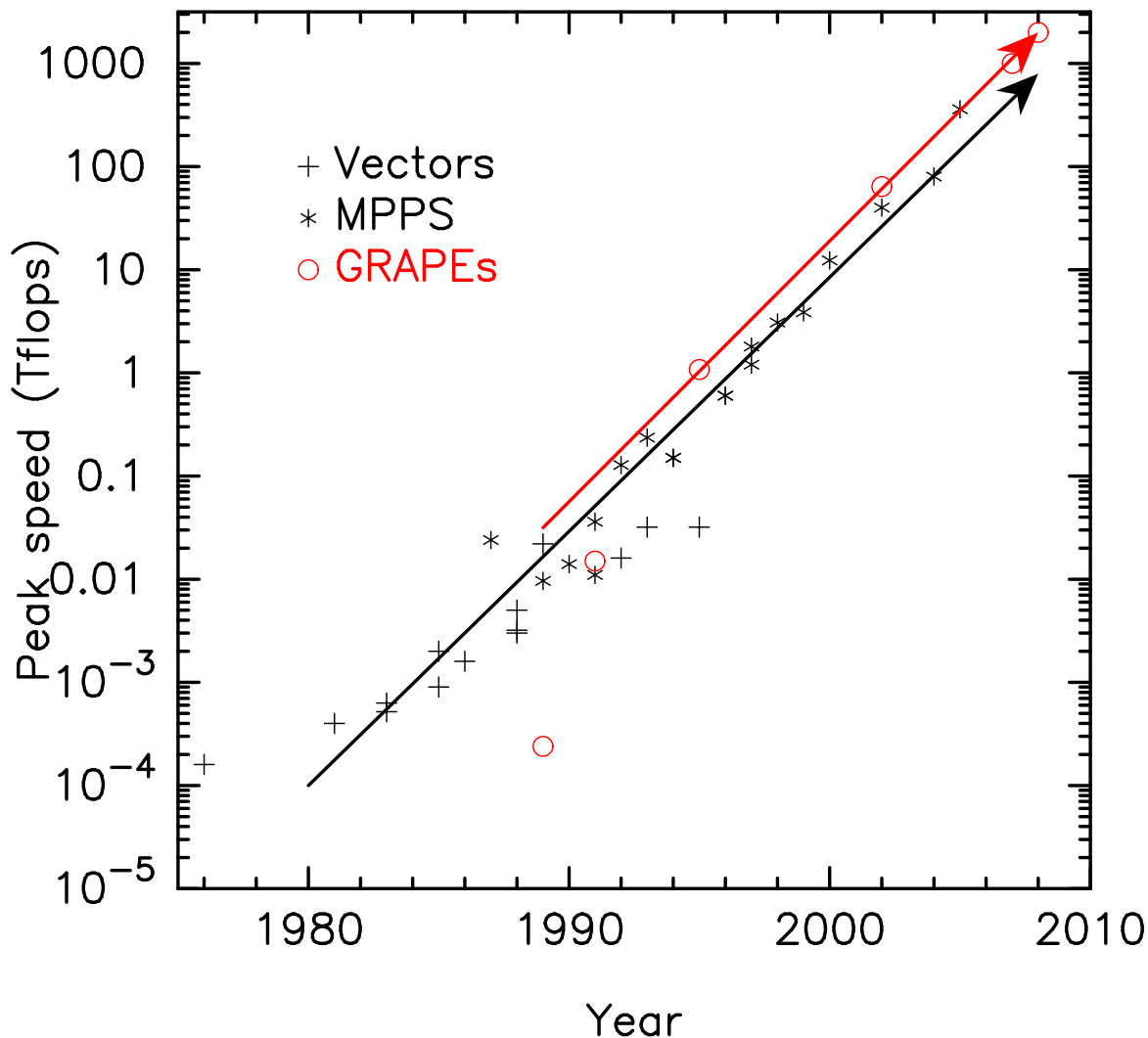
他の GRAPE 型機械

- 1991 GRAPE-1A : GRAPE-1 の細かい改良。設計: 福重
- 1992 GRAPE-2A : 最初の MD 用 GRAPE 設計:伊藤・福重
- 1992 HARP-1: 小久保君設計。 Hermite 公式用
- 1993 GRAPE-3A: 商業版
- 1996 MD-GRAPE: 最初の MD 用 GRAPE チップ
チップ設計:泰地
- 2001 MDM (MDG2): 理研・戎崎グループによる。 75T
- 2006 PE (MDG3): 理研・泰地グループによる。 1P
- 1992- MD-Engine: 富士ゼロックス、大正製薬(現在 NEC)

他の GRAPE 関係専用計算機

- 1991 DREAM: 大規模偏微分方程式計算「専用」計算機。
ハードディスクを主記憶に、という発想。
ディスク 1 台の試作程度まで。大野、牧野、戎崎。
- 1993 ZEBRA : Radiocity 法専用計算機。成見、戎崎
- 1995 General: LU 分解専用計算機。市販チップで構成。
清木、戎崎、泰地、牧野。
- 2002? MACE: LU 分解専用計算機。泰地、戎崎

ピーク性能の進歩



GRAPE-4 以
降、完成した時
点で世界最高速
を実現

GRAPE によるサイエンス

- 惑星形成、Young and Massive cluster、球状星団、BH ありの銀河中心
 - GRAPE を使った計算が世界的に主流
 - 直接計算なので GRAPE がもっとも有効 (100-1000 倍)
- 銀河形成、宇宙論
 - GRAPE が主流というわけでもないかも
 - SPH が重い、Tree だと GRAPE が圧倒的に速いというわけではない (10-30 倍程度)

GRAPE-DR 計画とは何か？

「基本的には」次期 GRAPE 計画

- 2004年度から5年計画
- 目標ピーク性能: 2 Petaflops
- チップ数 4096
- 単体チップ性能 0.5Tflops

と、これだけなら今までの GRAPE が速くなっただけ。
実際のアーキテクチャ: 今までの GRAPE とは全然違う

- なぜ違うか
- それで何ができるか

「次期 GRAPE」の実際的な問題

天文だけ (しかも理論だけ (しかも N 体だけ)) でもらうにはチップ開発コストが大き過ぎる

チップ開発費

1990 $1\mu\text{m}$ 1500 万円

1997 $0.25\mu\text{m}$ 1 億円

2004 90nm 3 億円以上？

2006 65nm 10 億円以上

ある程度広い応用を持つものでないと予算獲得が難しい

ALMA でも関連器は FPGA、、、

ではどうするか

1. やめる
2. 安くあげる方法を考える
3. なんかお金を取る方法を考える

GRAPE-DR では (3) を選択

基本的な考え

- チップに演算器を 2000 個くらい入れる
- それを (GRAPE が得意なタイプの問題に対しては) ある程度のプログラム可能性をもった形で使う。GRAPE のようなハードワイヤードなパイプラインにはしない。

もうちょっとそれらしく言うと:

- 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

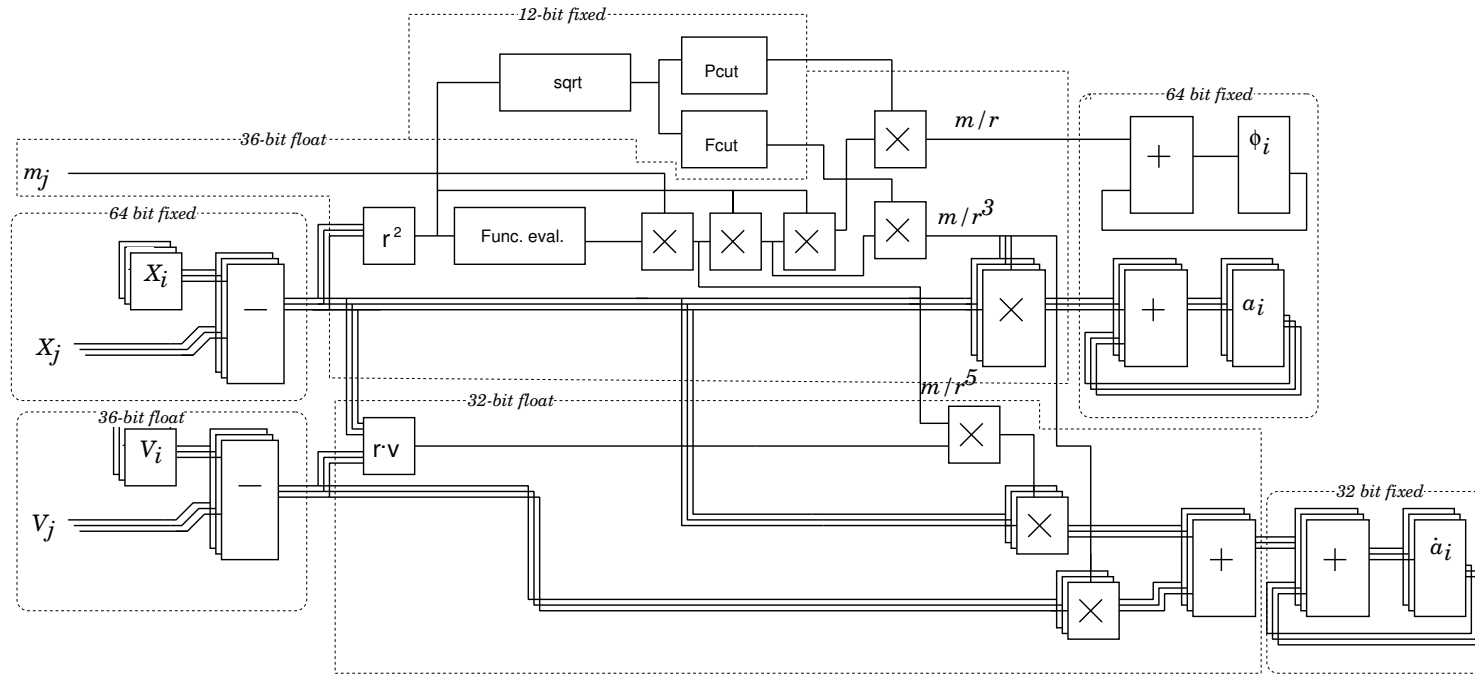
境界条件: メモリバンド幅は増やしたくない(システムコストはほぼメモリバンド幅で決まる)

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

GRAPE的専用プロセッサ

これでよければ別に何も考えることはない。



再構成可能プロセッサ

基本的に、配線、論理素子の中身を書き換えるもの。

FPGA ベース

- 任意のロジックを実現可能
- 集積度、速度は大きなペナルティがある
- 精度が低くてもいい応用には向く

いわゆる「動的再構成可能プロセッサ」

IPFlex DAP/DNA 等

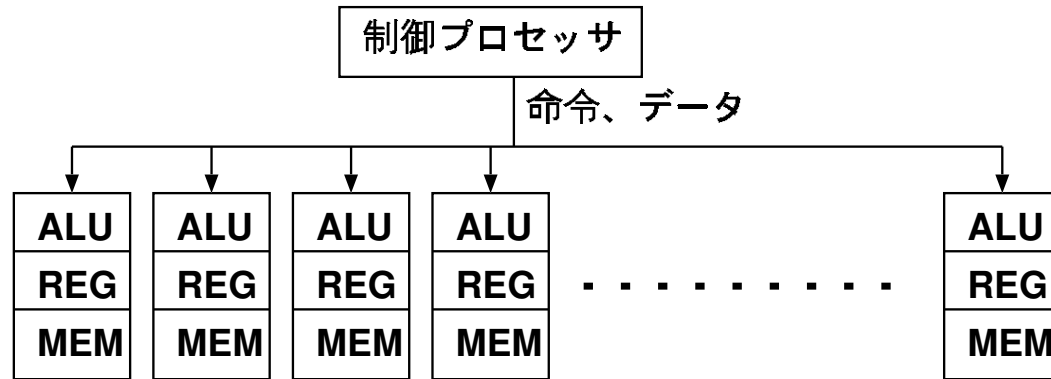
- 8-32 ビットの単純な ALU を多数集積
- その間をプログラマブルな配線でつなぐ
- 集積度、速度はやはり大きなペナルティ

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

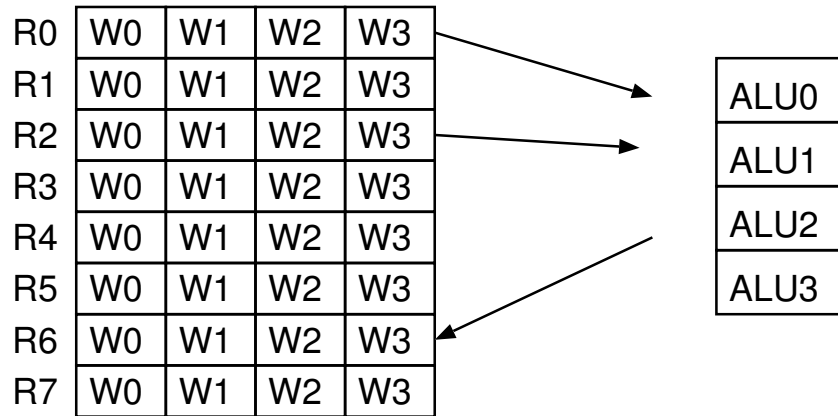


1960年代に発生、80年代に絶滅
半導体技術の向上に対応できないアーキテクチャ: 計算速度と
メモリアクセス速度が比例する必要あり。
メモリ階層をつける: プロセッサが複雑になりすぎて SIMD
の意味が無くなる。

SIMD 拡張命令

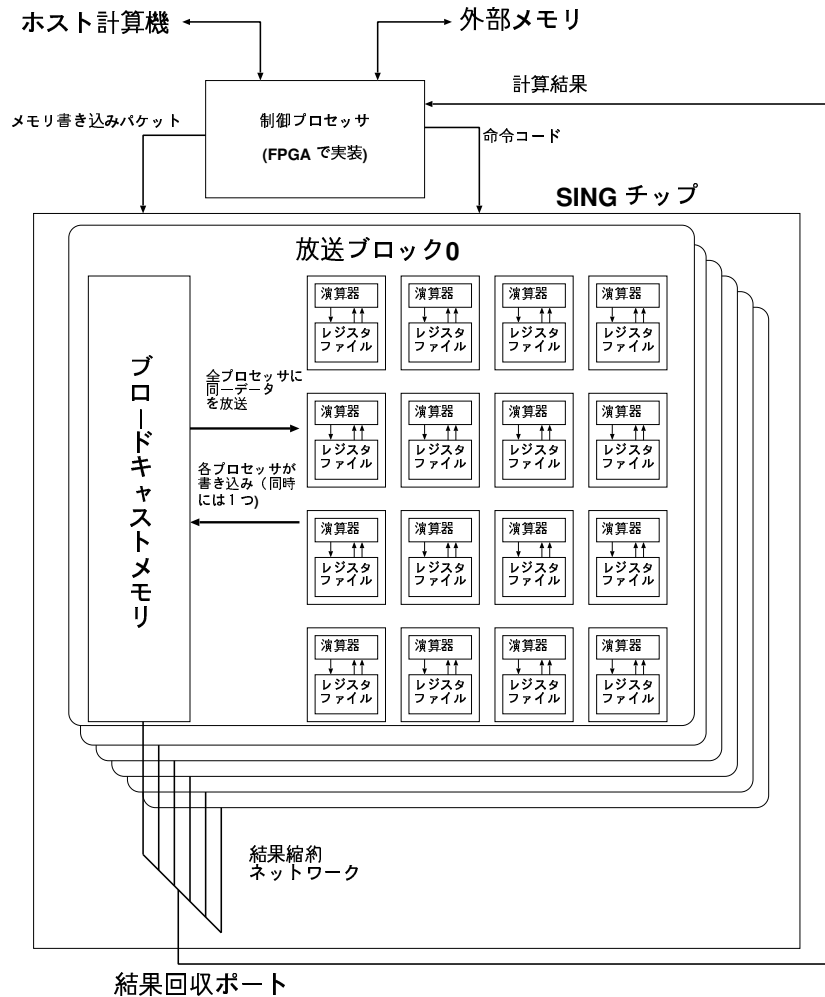
Pentium III, 4 が有名

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



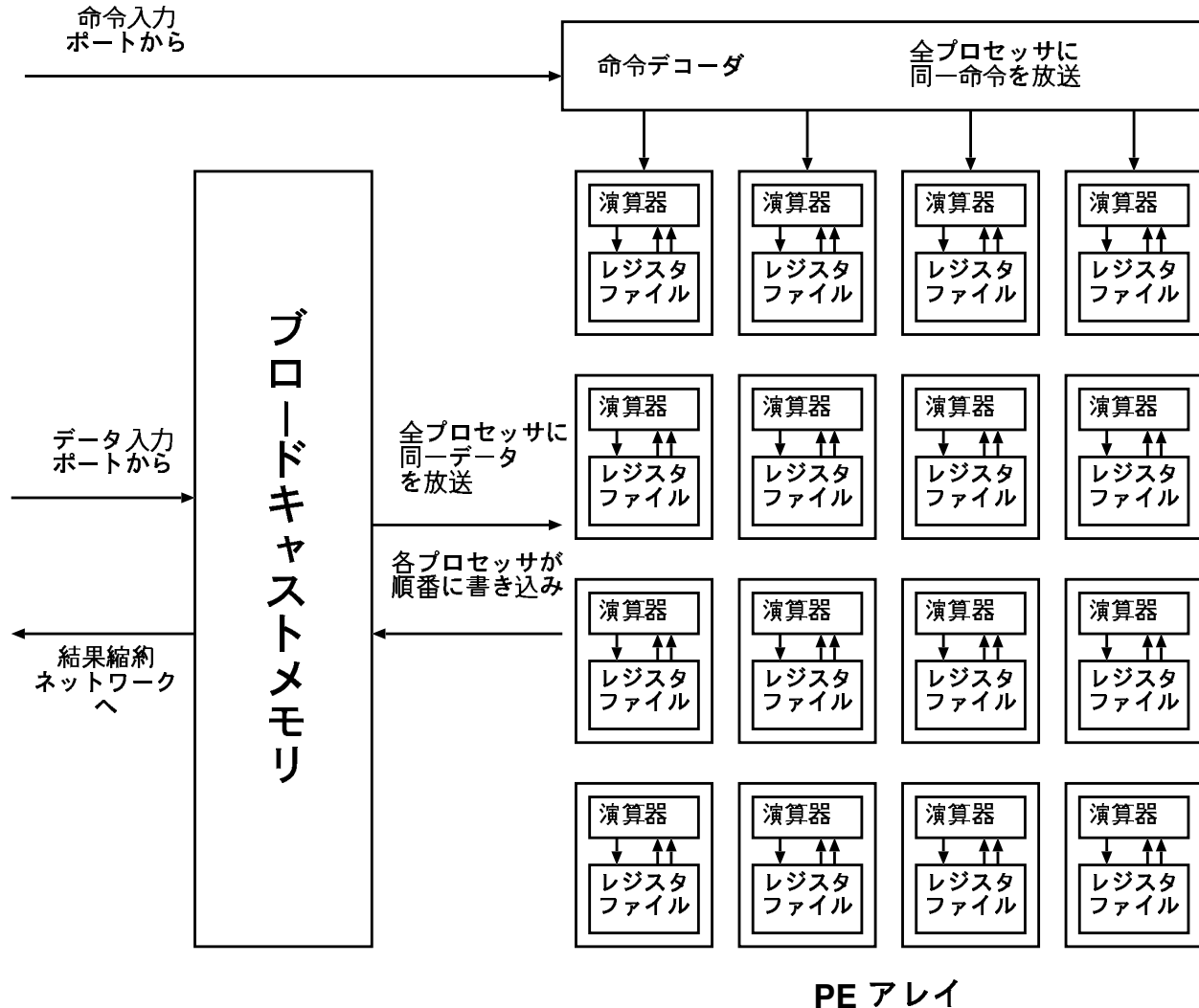
1つのプロセッサの中の話: キャッシュとデータをやりとり
並列度 4-8 程度が限界? それ以上増やすとキャッシュの速度が追いつかなくなる。

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

ブロードキャストブロック



各プロセッサは同じデータをブロードキャストメモリから受け取る

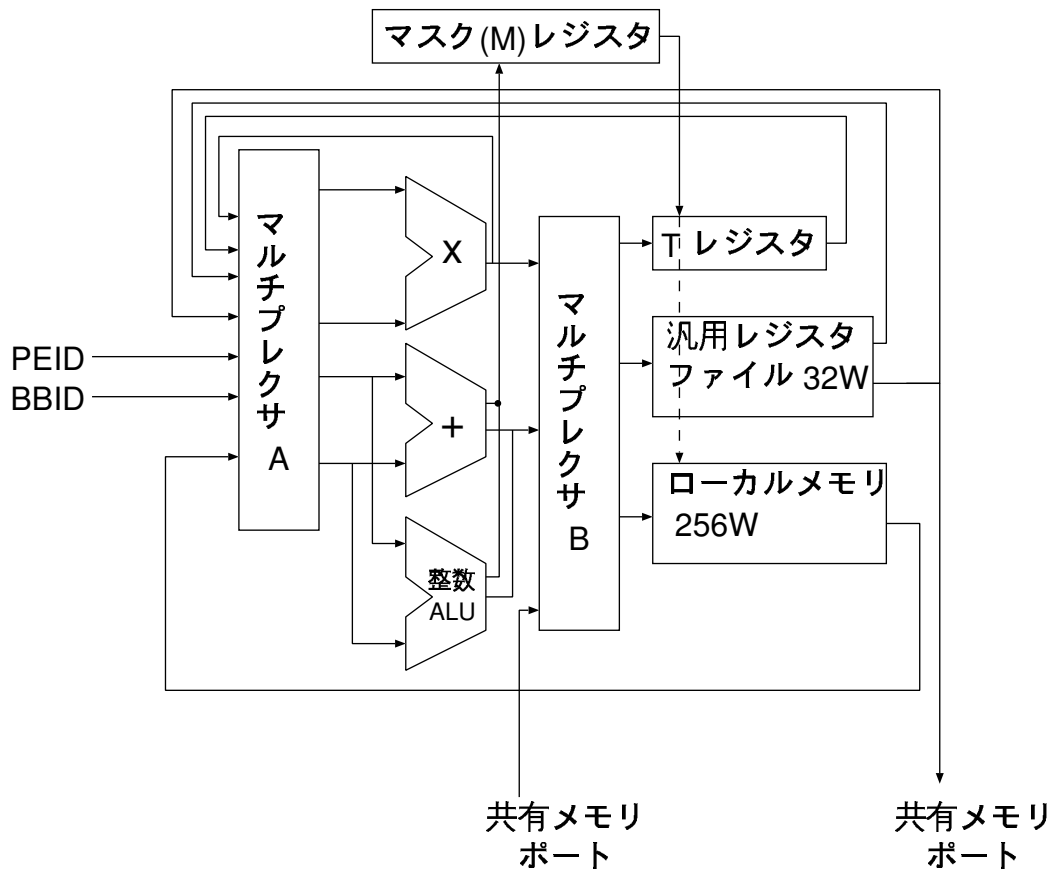
このブロックがチップ内には沢山ある。答はブロックにまたがって合計することもできる。

ちなみに

GRAPE-DR

Greatly
Reduced
Array of
Processor
Elements with
Data
Reduction

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。

PEの詳細

データ形式

単精度浮動小数点: 36 ビット (符号 1、指数 11、仮数 24)

倍精度浮動小数点: 72 ビット (符号 1、指数 11、仮数 60)

36/72 ビット固定小数点数

演算命令

乗算は単精度のみ (倍精度のための部分積をサポート) 倍精度

乗算を 2 サイクルでするために 25×50 ビットの乗算器

整数演算、加減算は倍精度のみ (メモリ/レジスタからの読出し/格納時に単・倍変換ができる)

特殊な浮動小数点命令: 仮数を正規化しないまま演算を続ける。これにより、演算順序によらないで結果が同じになることを保証する (GRAPE-6 の積算と同様)

普通に正規化もできる (こっちがデフォルト)

PEの詳細(続き)

- パイプラインは8ステージ。
- 基本命令は4データに対するベクトル命令。4サイクルに1回しか命令ははまらない。
- T レジスタのみ直前の命令の実行結果を利用可能。
- T レジスタはアドレスレジスタになる(間接アクセス)

サポートする命令等は基本的には昔の SIMD 計算機、例えば CM-2, MasPar MP-1 なんかとあまり変わらない。但し、PE がはるかに強力になっている。

アプリケーションに対する考え方

- Memory Wall が問題にならないようなアプリケーションでは性能がでる。
- そうでないものはそこそこに。
- 可能ならばアプリケーションを書き換える

性能がでる例

- 小規模な問題を独立に沢山解く
 - レイトレース計算：光学部品（レンズ、導光版）設計
 - 放射線伝播のモンテカルロ計算：
検出器設計？エアシャワー？
 - 3体問題:連星と単独星の遭遇、微惑星同士の遭遇
 - 量子化学計算での 2 電子積分
- GRAPE 的な粒子間相互作用計算
 - SPH とかも
 - 分子動力学も
- 密行列計算
 - 結構需要はある
 - Top 500 に載る

検討が必要な例

- 陽解法流体計算
 - 複雑で計算量が多い方法ならそれなりに
- PIC(PM) でのプラズマシミュレーション
 - 核融合研の知り合いと検討始めた
 - 高次補間 (3次元3次補間) だと charge assignment とかで結構性能でるかも

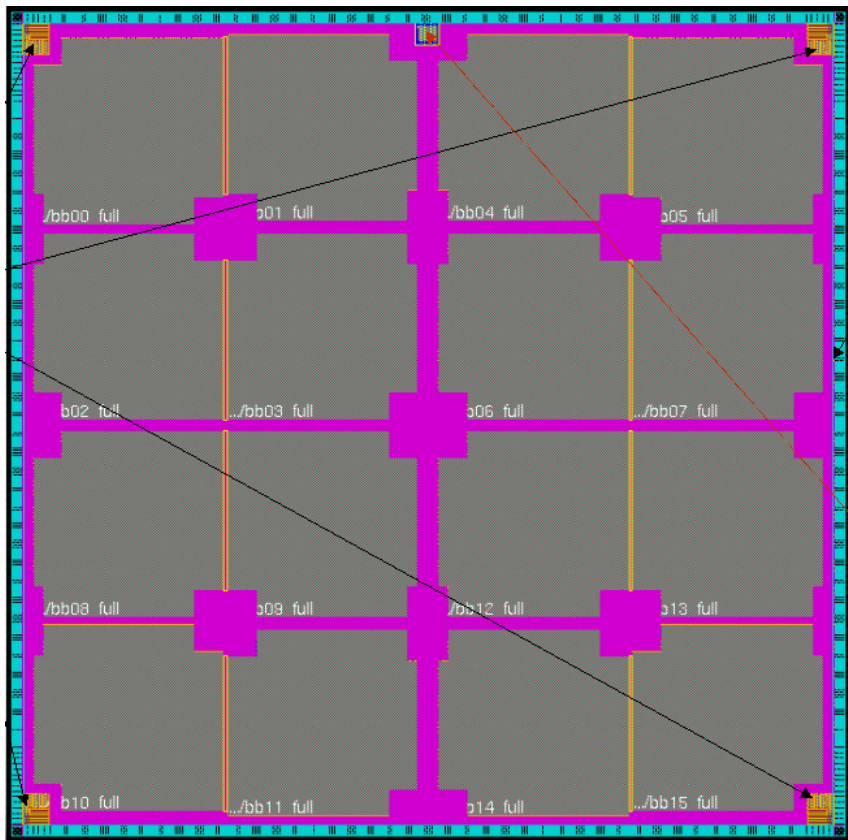
性能がでない例

- 多次元 FFT
- 低次陽解法な格子コード

GRAPE-DR の開発体制

- 振興調整費。代表は東大の平木さん (情報理工学研究科)
- 予算 年 3 億、5 年計画 (現在 3 年目)
- ネットワーク関係のハード、ソフト開発と GRAPE-DR プロセッサ開発の 2 つが柱。
- プロセッサの仕様策定は牧野が中心。シミュレータとかも。
- 物理設計は日本 IBM の名村さん (GRAPE-6 の時と同じ人)
- ボードは福重君
- ボードの FPGA の中身は藤野君 (M2)

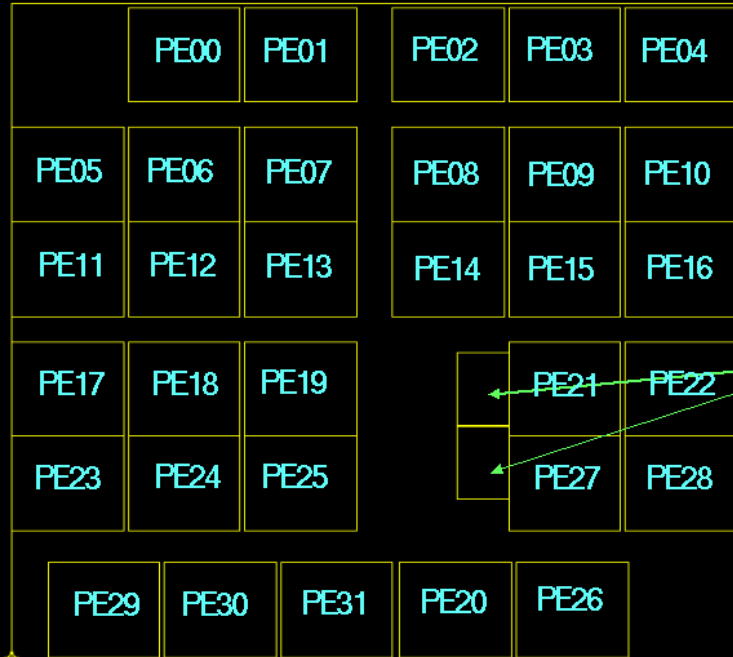
チップ全体フロアプラン



- 特におかしいところなし
- 4.5 億トランジスタ
- サイズは大きい。
17mm 角

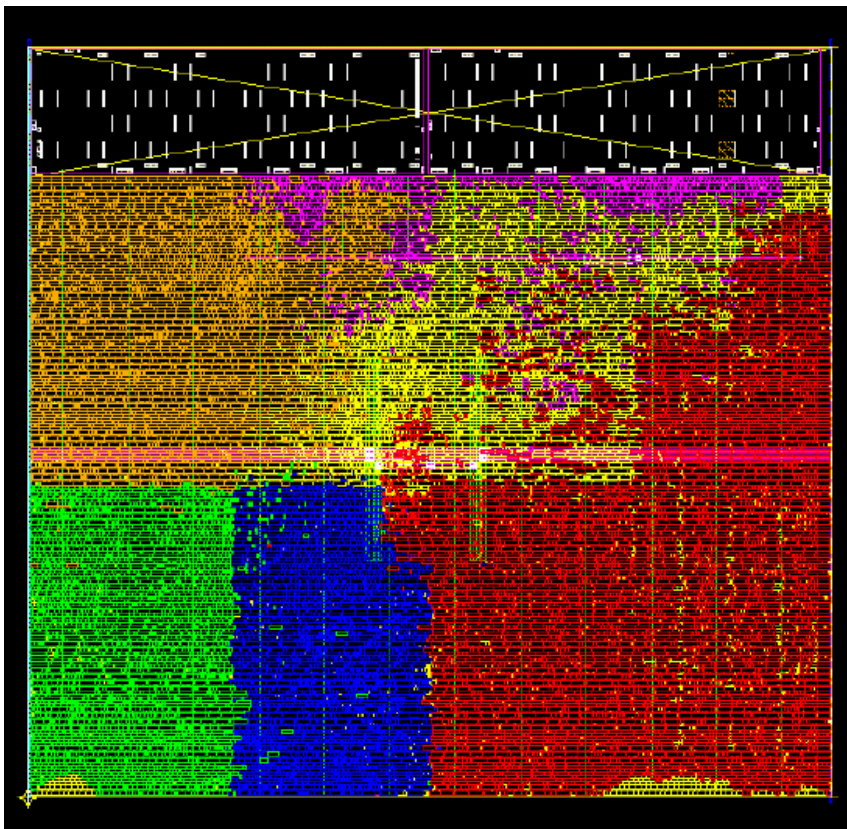
放送ブロックフロアプラン

block Size : 4007.64 um x 4039.56 um



- 特におかしいところなし

PE フロアプラン



Module Name

grf0

fmul

fadd

alu

lm

Others

Color

red

orange

green

blue

magenta

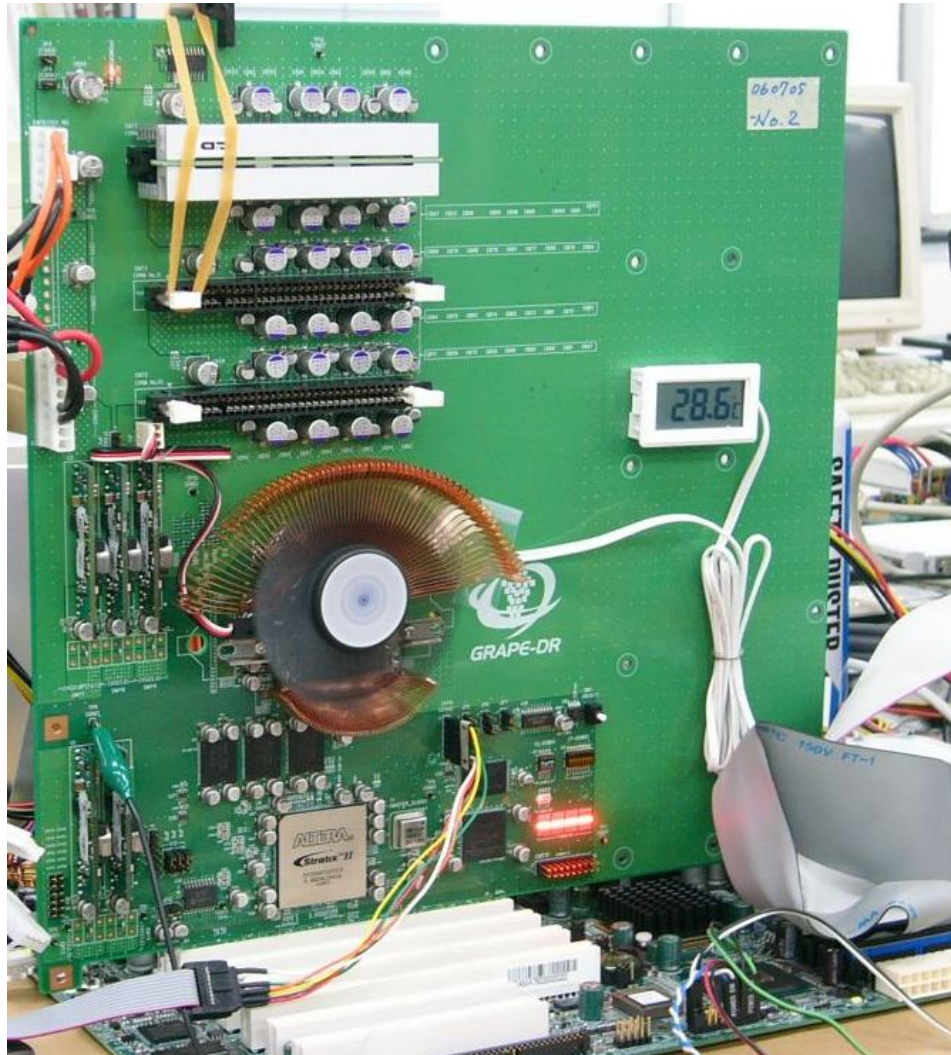
yellow

GRAPE-DR の開発状況



シミュレーションデータと同じものを供給して同じ答がでるところまで確認。
(これとは別ボードで) 500MHz 動作も確認、消費電力 25-50W 程度。

GRAPE-DR 別ボード



- こっちが「プロジェクト公式」
- 中身は殆ど同じ
- 何故か大きい
- 500MHz 動作まで確認済

原始的なコンパイラ

(中里 2006)

```
/VARI  xi, yi, zi, e2;  
/VARJ  xj, yj, zj, mj;  
/VARF  fx, fy, fz;  
dx = xi - xj;  
dy = yi - yj;  
dz = zi - zj;  
r2 = dx*dx + dy*dy + dz*dz + e2;  
r3i= powm32(r2);  
ff = mj*r3i;  
fx += ff*dx;  
fy += ff*dy;  
fz += ff*dz;
```

これから GRAPE 並のことをするアセンブラ、インターフェースライブラリを生成。
基本的なアイデアは PGR (FPGA を使った PROGRAPE 用コンパイラ、濱田 D 論 2006) と同様

次世代スーパーコンピューター プロジェクト

- 2006 年度から始まっている。6or7 年計画
- 今年度は調査費 35 億。どこに消えたか不明
- 来年度 87 億。総合科学技術会議の攻撃はかわしたらしい？

お金の話はともかく

一体何を、何のために作るのか？が問題
なんだけど、これが実は良くわからない。
「政治案件」という噂もあるらしい。

2006/10 現在での目標:
汎用並列計算機で LINPACK 10 Pflops

プロジェクト目標の変遷

- 2005 年始め:文部科学省の下の情報科学委員会の下のなんとか WG の報告がでた。
 - 複合型、ベクトル 0.5P、スカラ 1-2P、「準汎用」10P
 - ターゲットはナノ、バイオ
- 2005 年秋:総合科学技術会議 (CSTP) と調整
 - 複合型、ベクトル 1P, スカラ 3P、「準汎用」>10P、LINPACK 10P
 - CSTP からは「来年までにまともな案をもってこい」
- 2006 年春:ハードウェア構成案募集「ピーク 10P」
 - 各メーカー、GRAPE-DR 関係、筑波関係、他
 - ベンチマーク評価をすることになった。そのベンチマークは「アプリケーション検討部会」(台長もメンバー)で選ばれた

プロジェクト目標の変遷 (2)

- 2006 年 7 月: ベンチマーク結果でた
- 2006 年 9 月: 来年度の方針決定「LINPACK 10P」
 - 非公開なので私は何も結果は知らない
 - アプリケーション検討部会では回収資料だったという噂
 - $N+H$, F の案
 - GRAPE-DR 後継については交渉中

結局どうなってるのか？

- 日経産業新聞には「迷走」と書かれた
- 12月まで本当に何も決めないらしい
- GRAPE-DR の次は開発費くらいはでるのかもしれない
(まだ不透明)

本質的な問題点は何か

- 2012年なら、多分パソコンは 30 万円で 100Gflops である。
- 従って、10 Pflops なら 300 億でいい。
- ネットワークとか色々してもなかなか 2 倍にはならない
- 何故 1000 億もかけるのか？

参考：天文台のリプレースの目標は x 億で来年度末に 20TF。
4 年違うのでまあこんなものかな、みたいな。

お金がかかる理由:開発費

- 生産量が少ないので開発費の分が大きい
- 開発費が少ないのでハードウェアがあんまり効率良くない

まとめ

- GRAPE-DR のチップは出来た。
- 次世代スーパーコンピュータプロジェクトは迷走している
- 上手くいけば次世代スーパーコンピュータプロジェクトのお金で GRAPE-DR の次をやる。
- 天文台を沢山お金が通るかも。(通るだけ)