GRAPE-DR とその先

牧野淳一郎
(東大 (6/1 から国立天文台))
戎崎俊一
(理化学研究所)

今日の話の構成

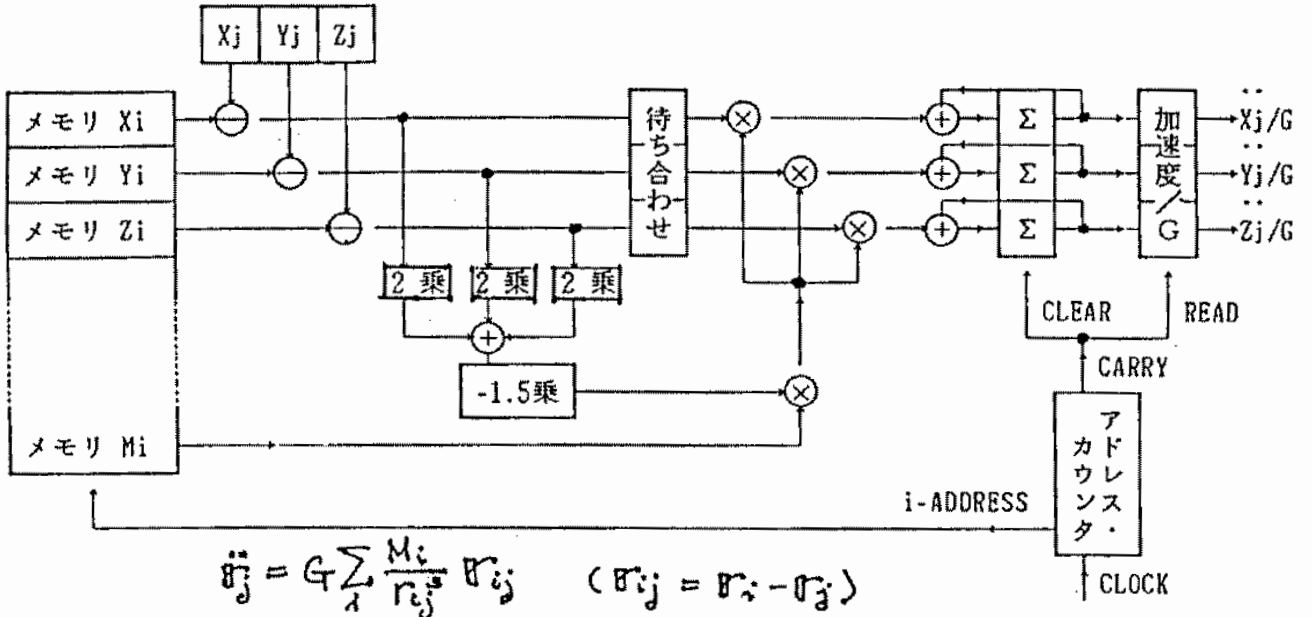
- これまでの GRAPE
- GRAPE-DR の考え方
- 現在の SIMD 超並列チップに共通の問題点
- 解決策
- 性能予測: QCD 計算、多次元 FFT を例に
- まとめ

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



$$\ddot{\vec{r}}_j = G\sum_i \frac{M_i}{r_{ij}^3}\vec{r}_{ij} \qquad (\vec{r}_{ij} = \vec{r}_i - \vec{r}_j)$$

+，－，×，２乗は１operation，　-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．

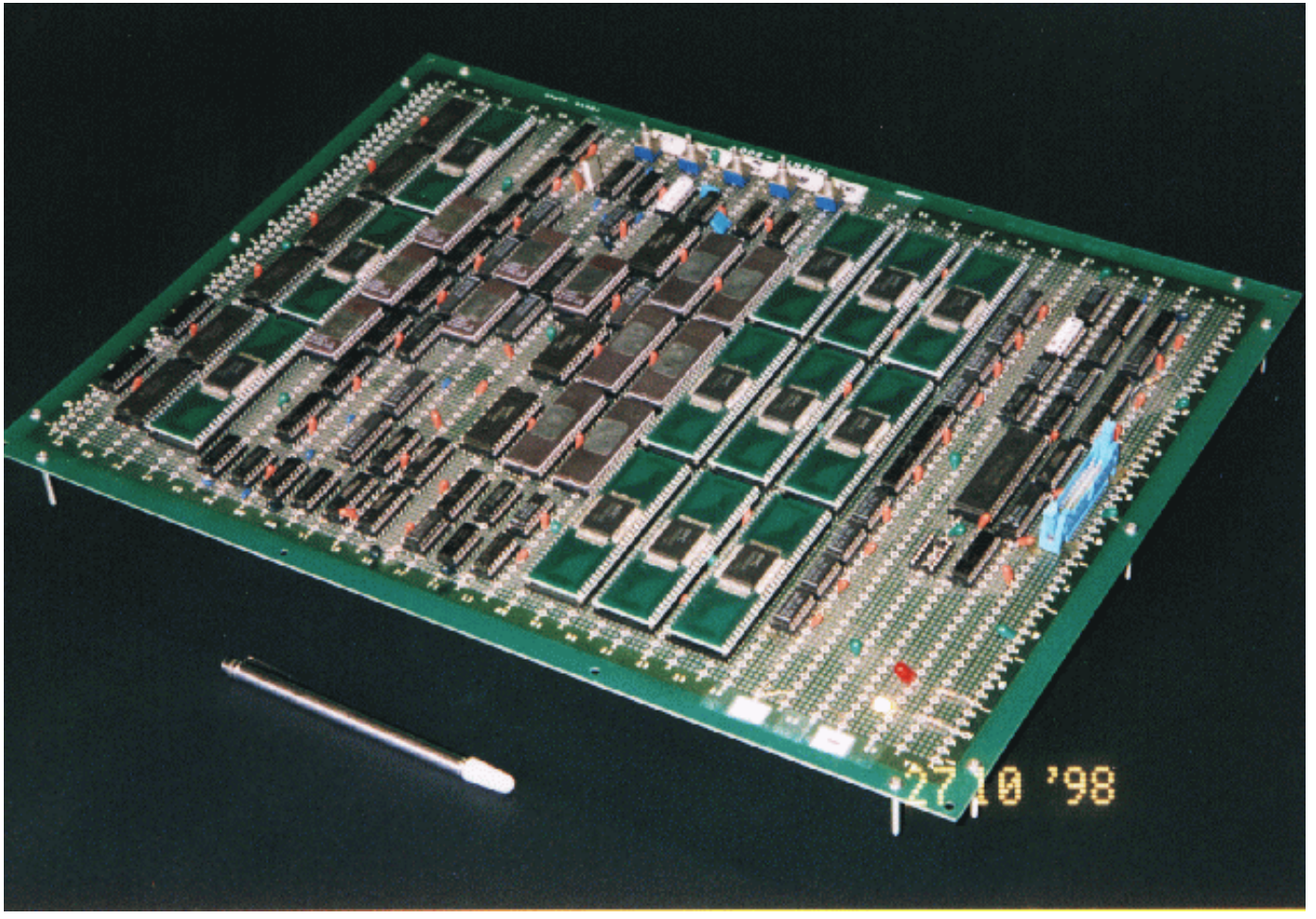
図２．Ｎ体問題のj-体に働く重力加速度を計算する回路の概念図．

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 $0.1\mathrm{m}^3$
- 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)

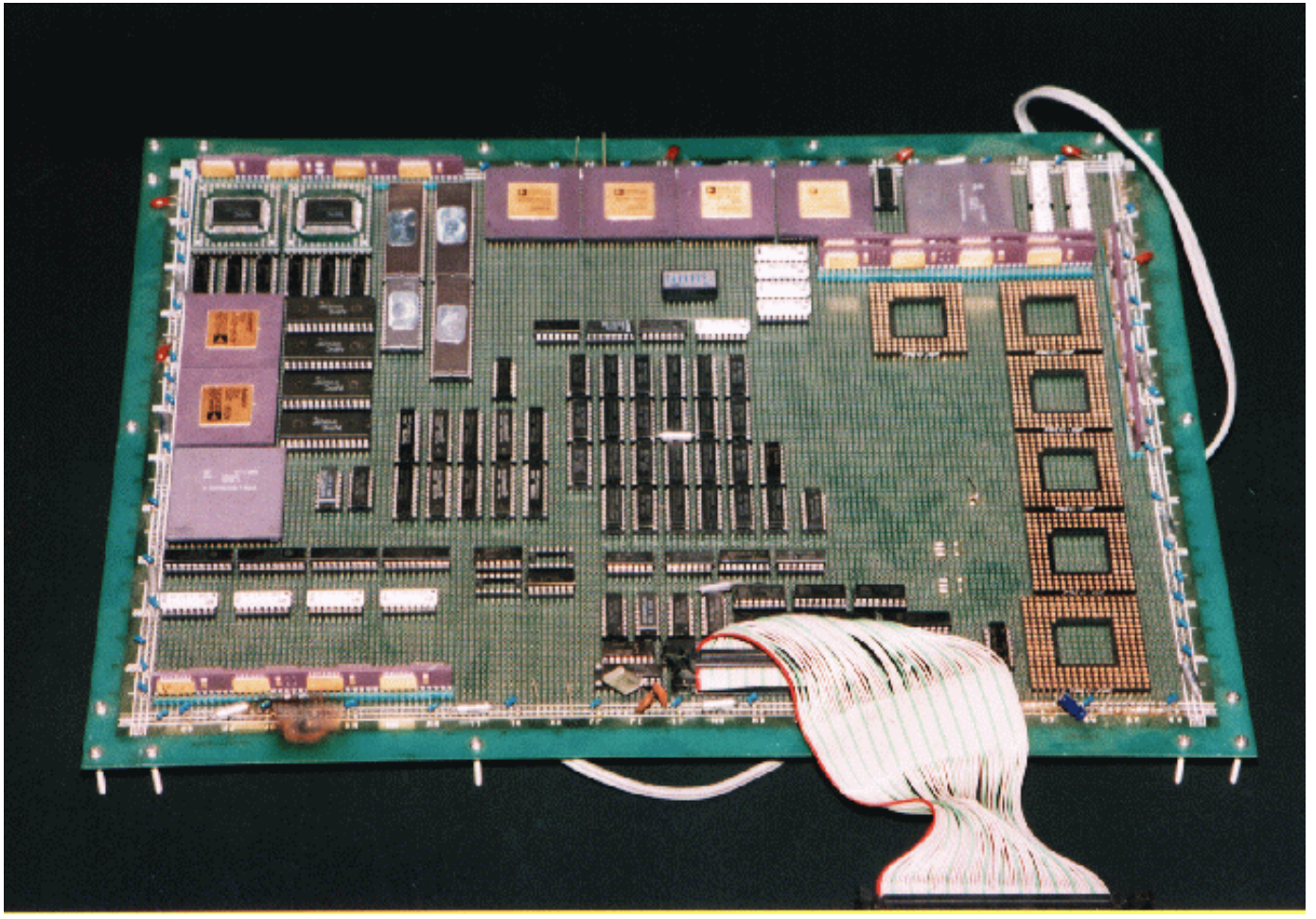


GRAPE-1 の中身

- 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

朴さんには大変お世話になりました。

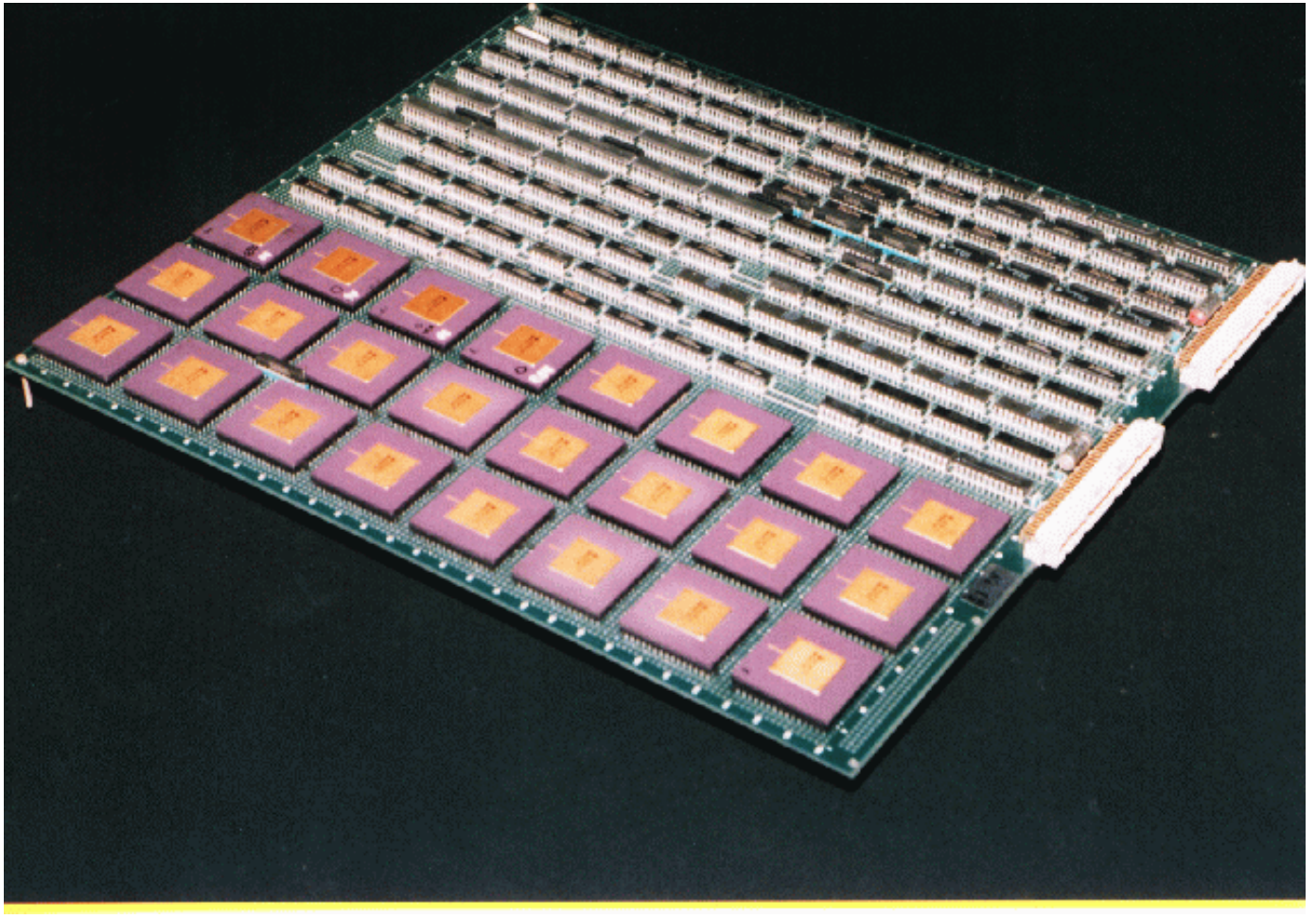
GRAPE-2(1990)



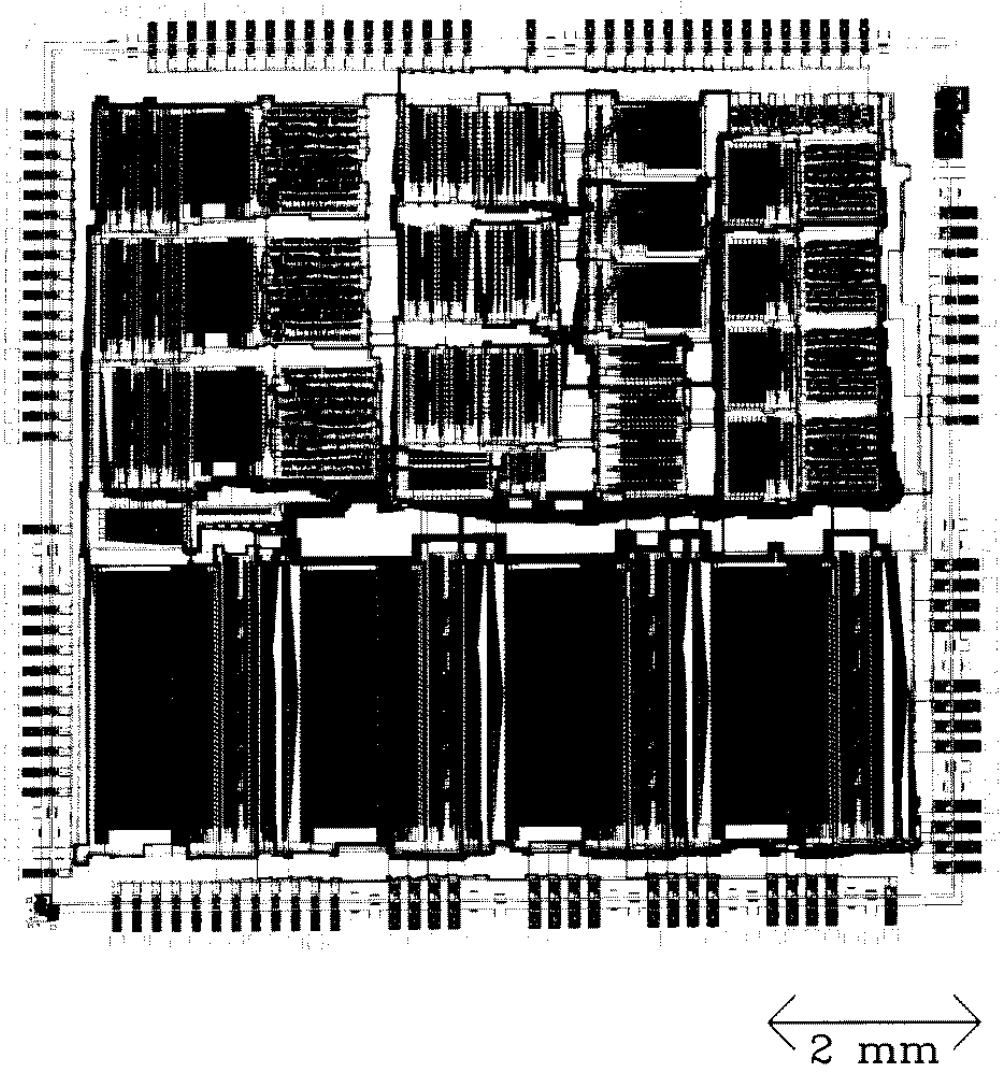
GRAPE-2 の中身

- 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
- 戎崎、伊藤、牧野

GRAPE-3(1991)



GRAPE-3 チップ

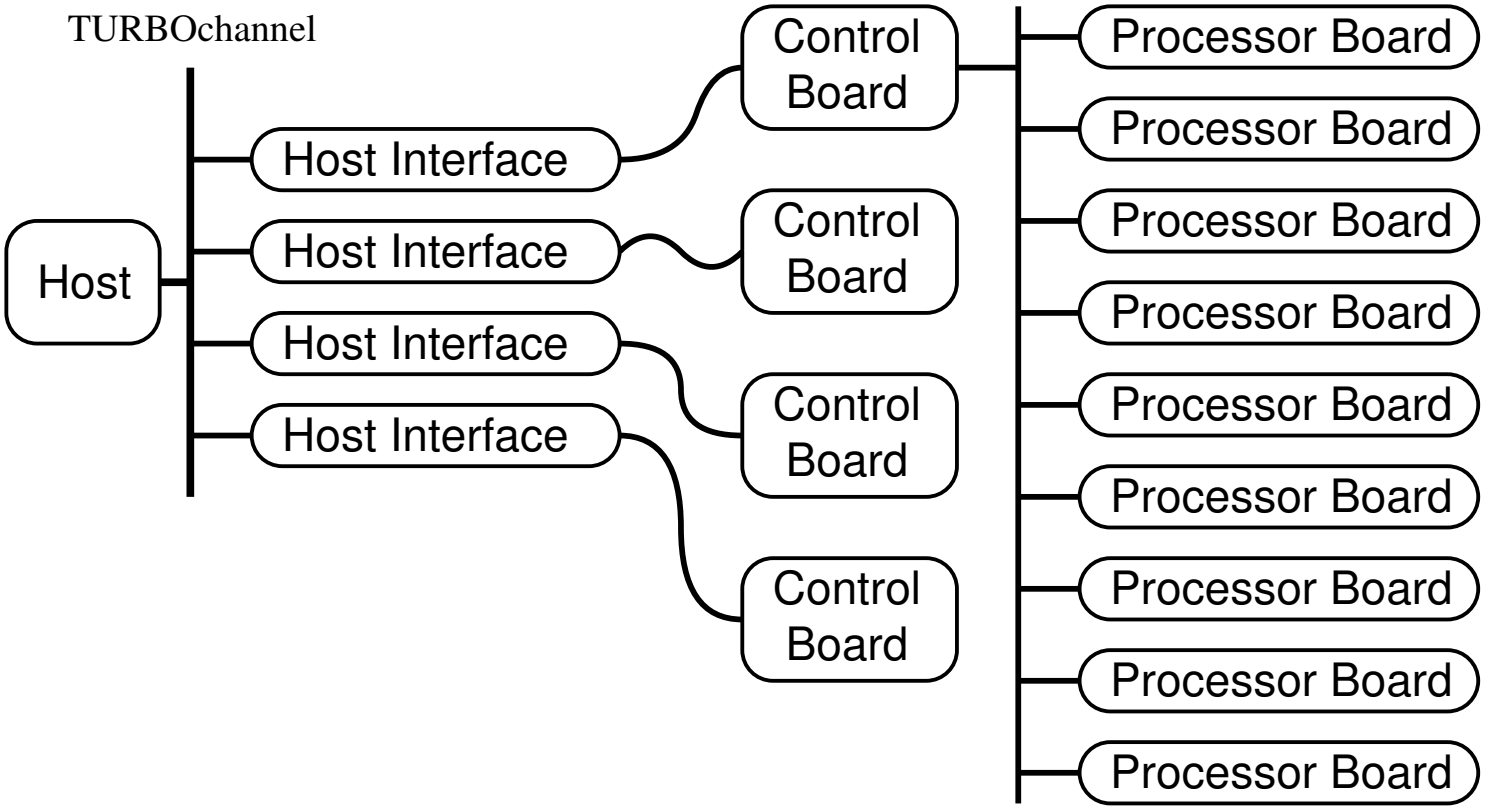


GRAPE-3

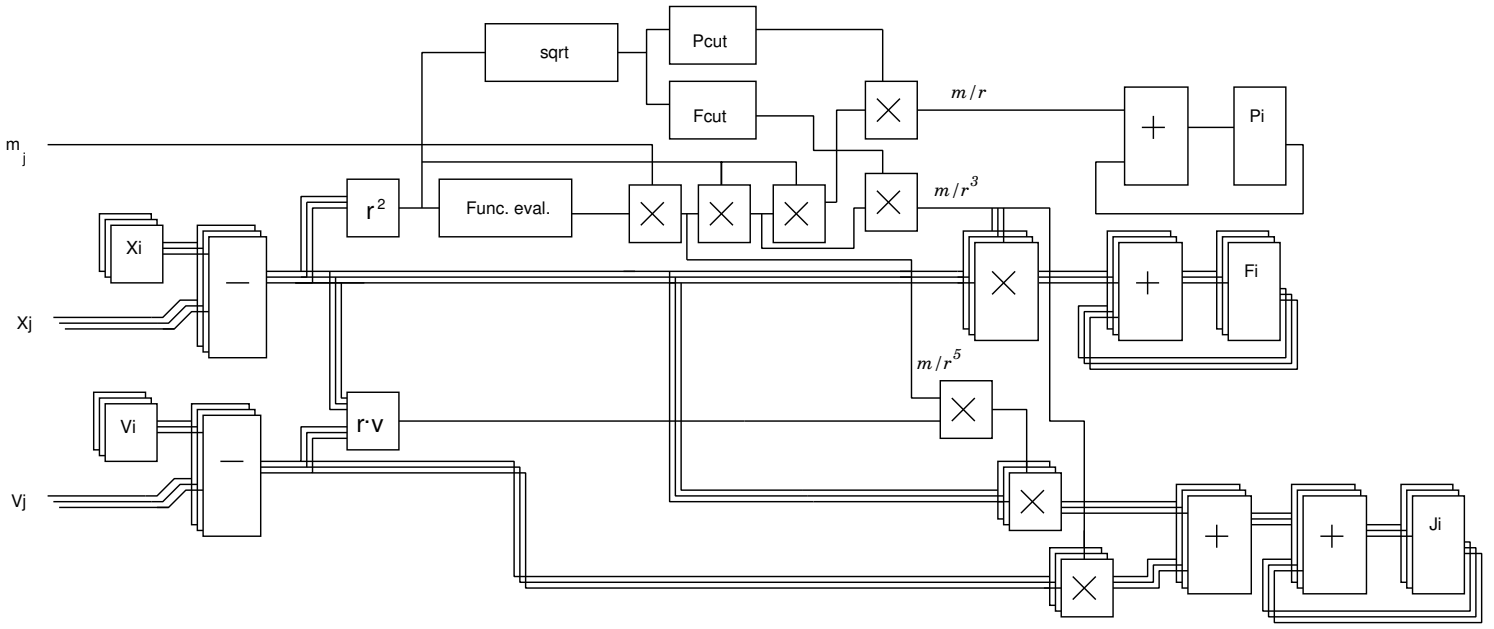
- 仕様決定、シミュレータ (C で記述) は牧野
- 論理設計以降は富士ゼロックス (橋本、富田)
- SCS Genesil で設計
- ファブは NS. 1μm
- ボードは奥村 (現在国立天文台)

GRAPE-4(1995)

GRAPE-4 の構成

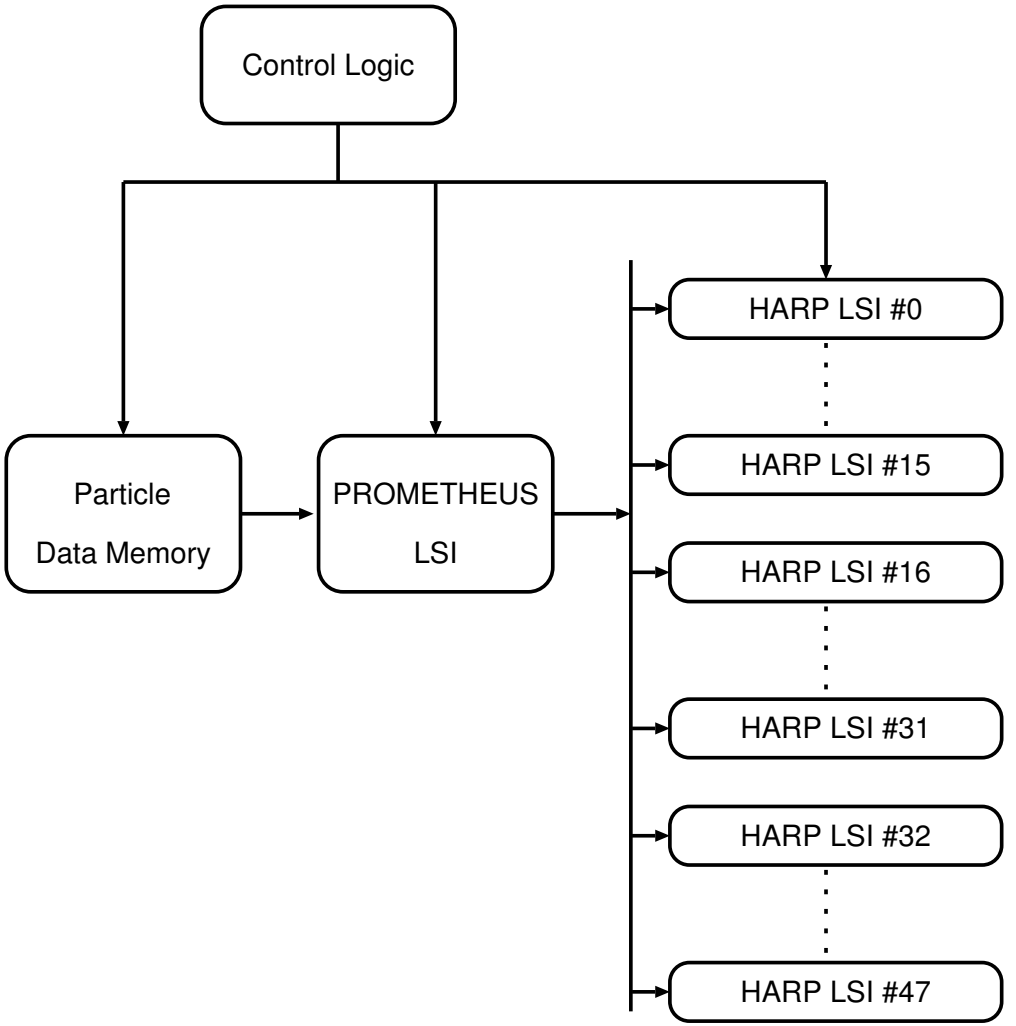

TURBOchannel
Host
Host Interface
Host Interface
Host Interface
Host Interface
Control Board
Control Board
Control Board
Control Board
Processor Board
Processor Board
Processor Board
Processor Board
Processor Board
Processor Board
Processor Board
Processor Board
Processor Board

GRAPE-4 パイプライン



設計:泰地

GRAPE-4 演算ボード



多数の パイプライン LSI がメモリユニットを共有

↓

ボードが単純になり、集積度をあげられる

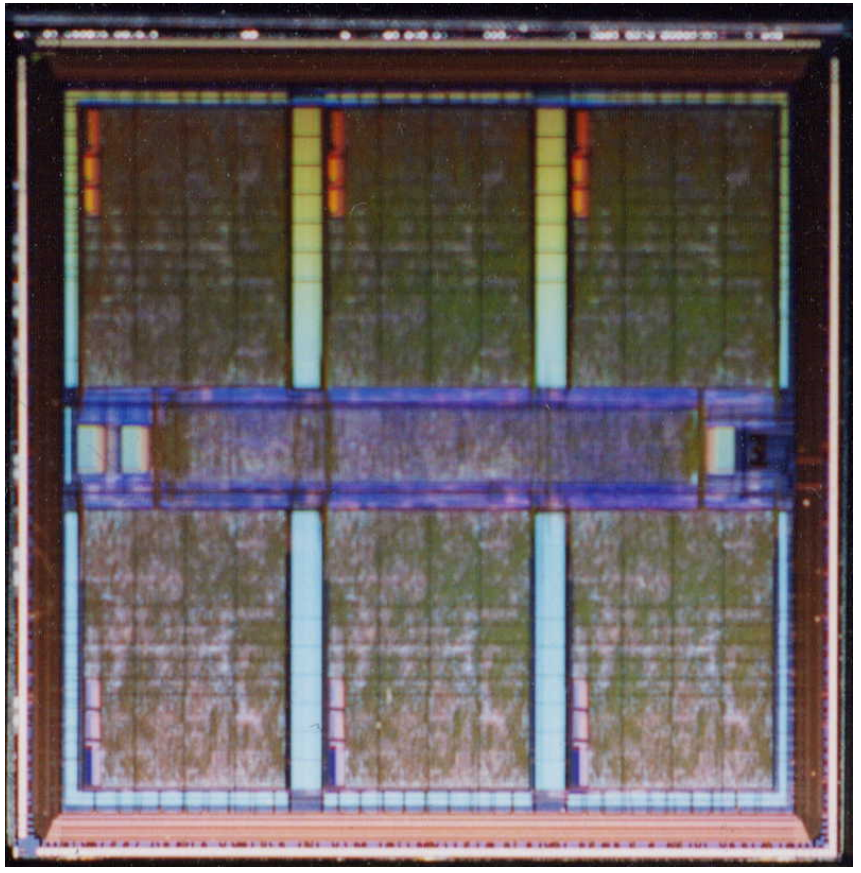
GRAPE-4 概要

- 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 1μm 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

GRAPE-6(2001)

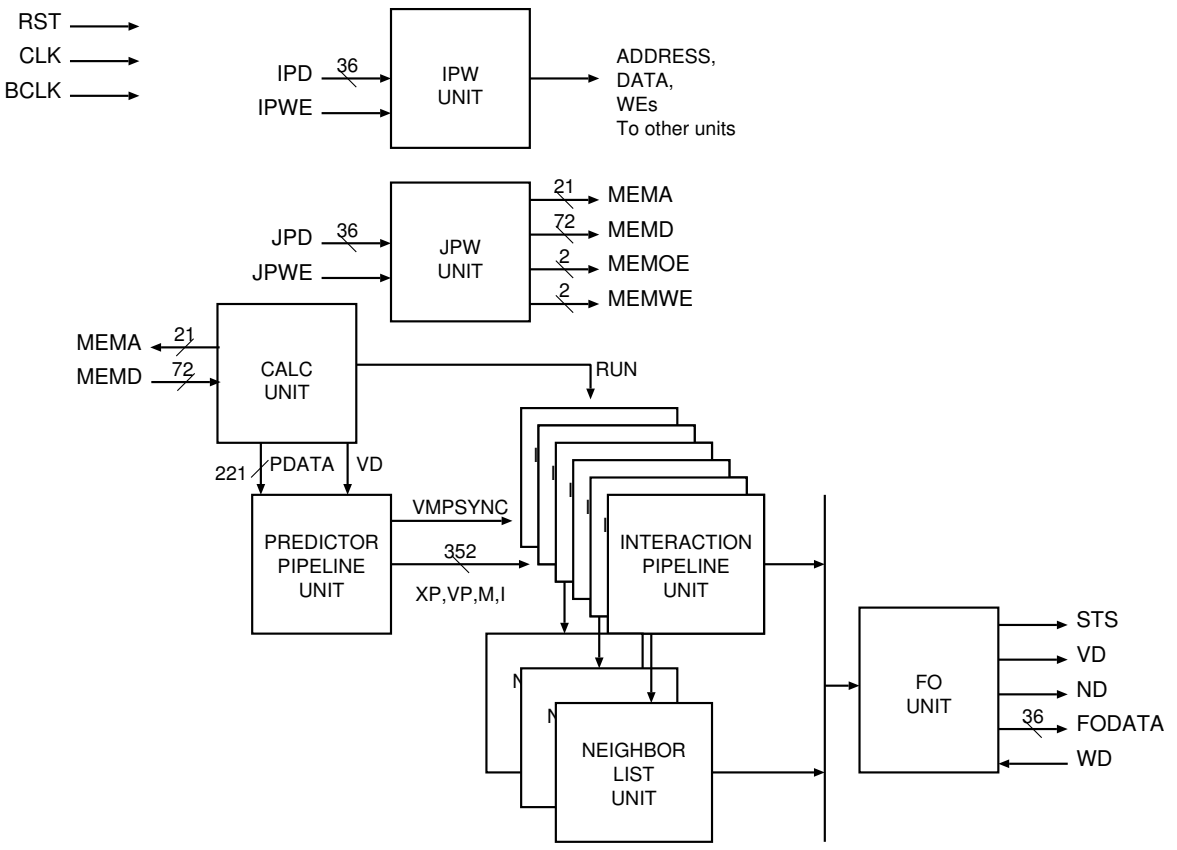
- プロセッサチップ
- プロセッサモジュール
- プロセッサボード
- ネットワークボード
- 全体

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

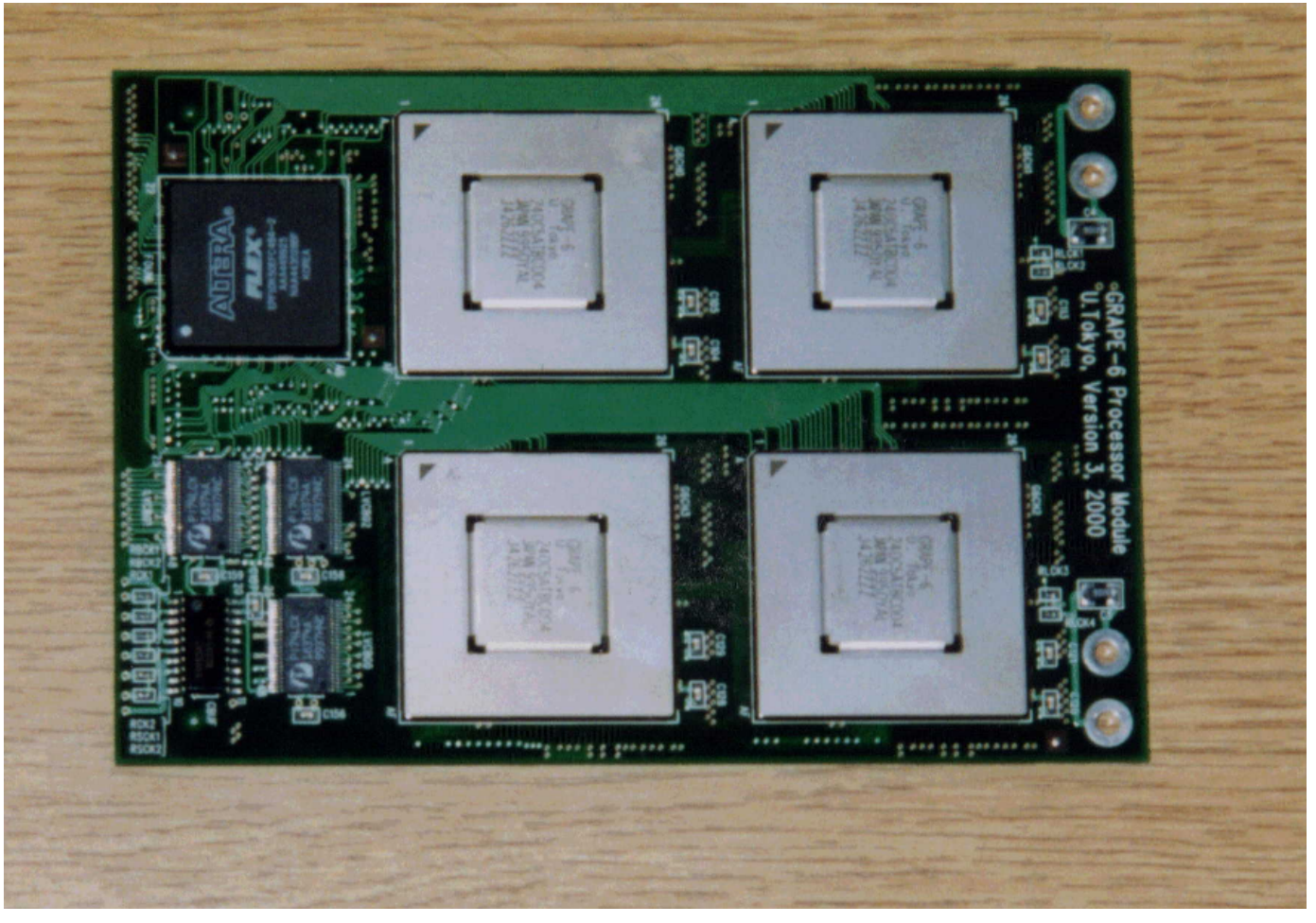
パイプライン LSI詳細



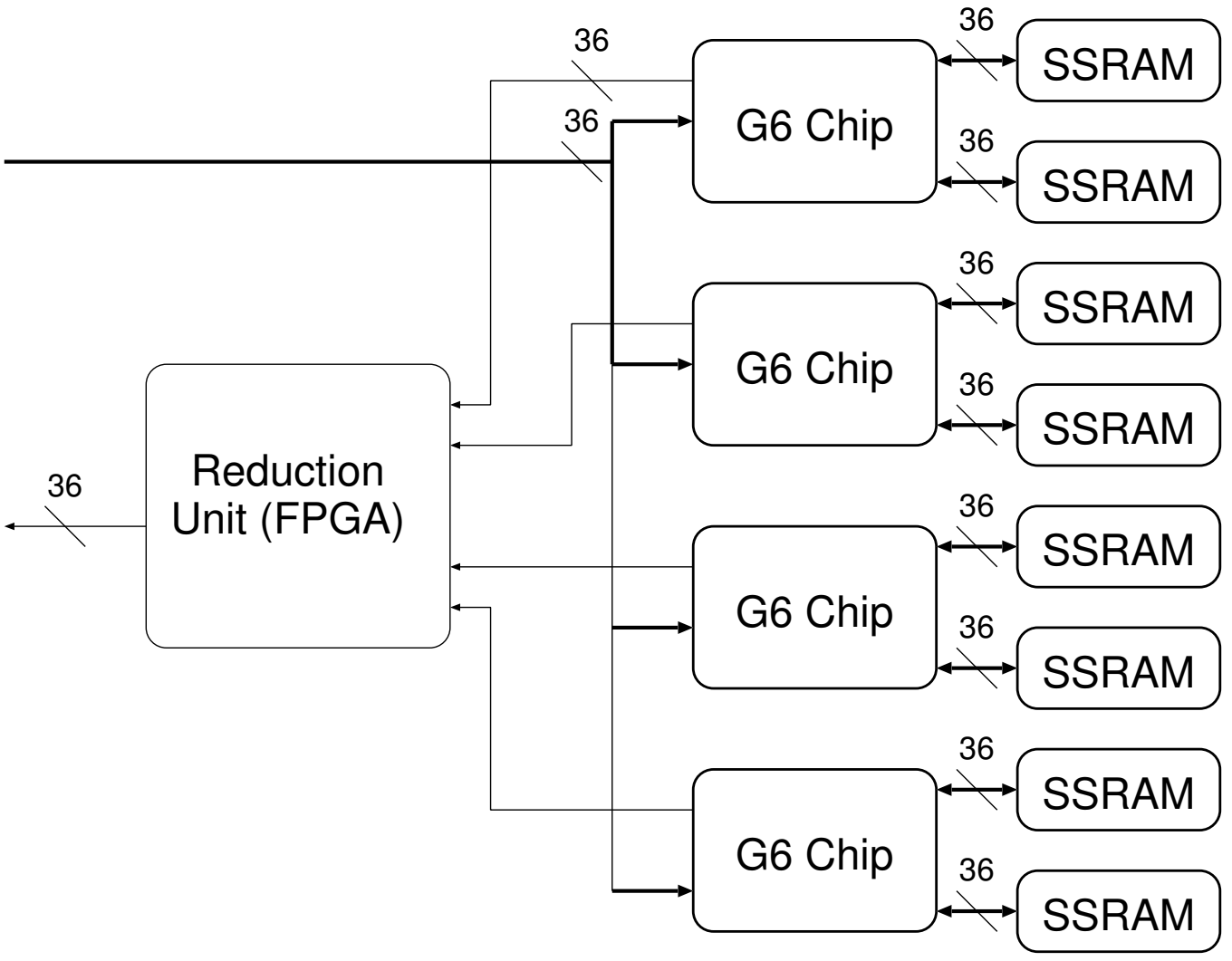
GRAPE-4 プロセッサボードの全機能を集積

- ホスト IF
- メモリ IF
- パイプライン本体
- 制御回路

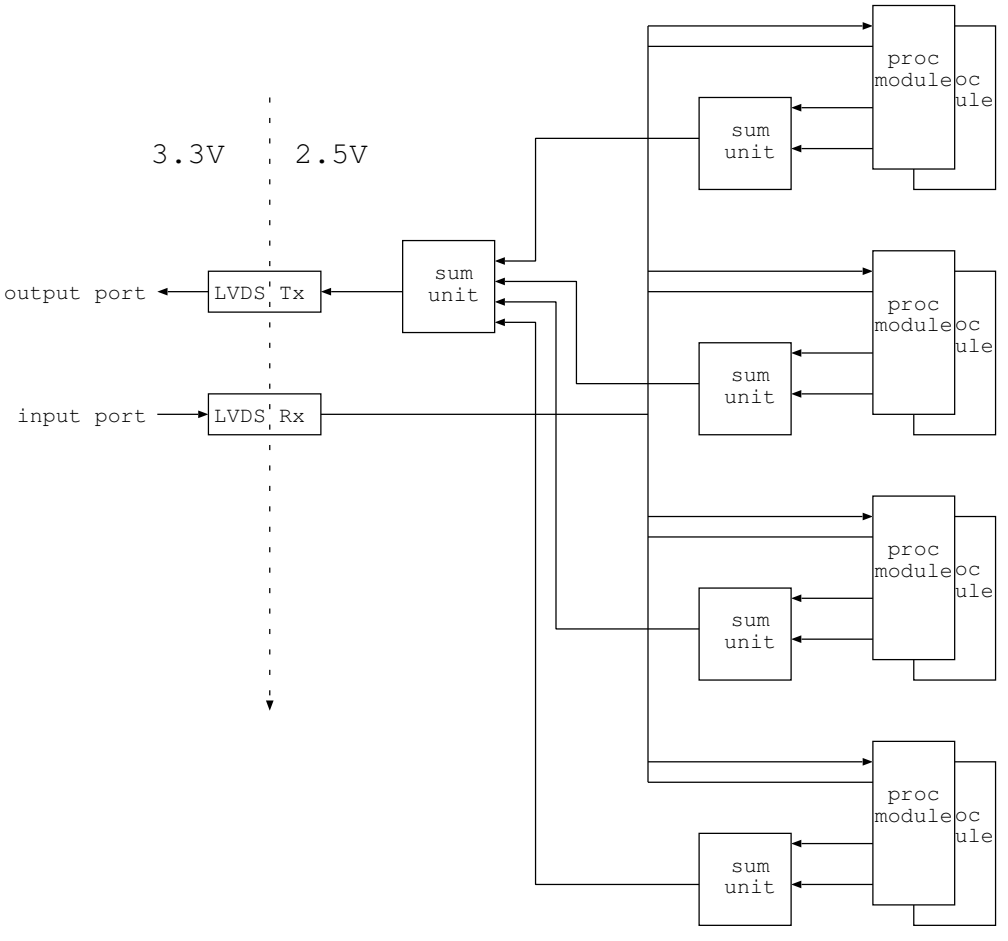
GRAPE-6 processor module



GRAPE-6 processor module

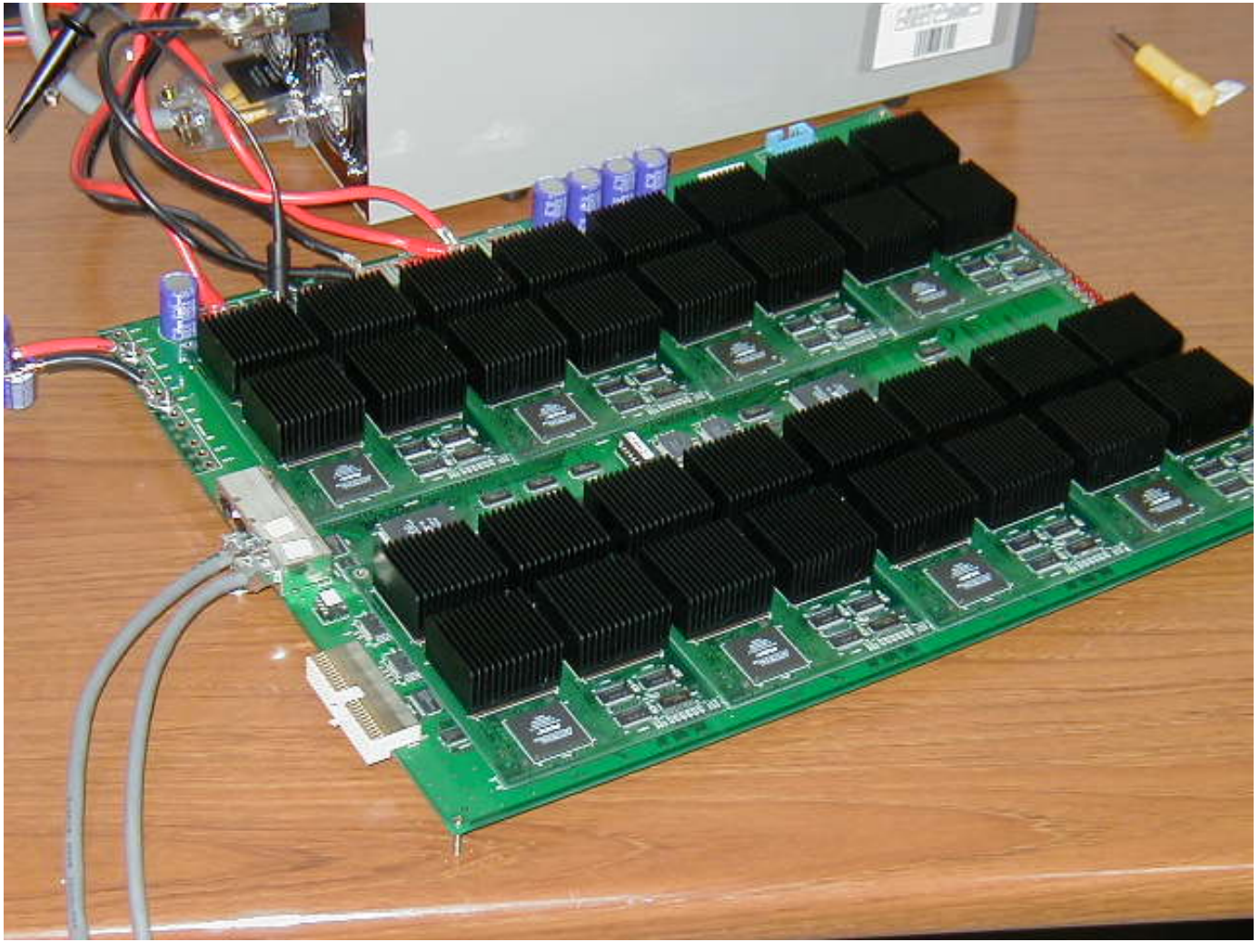


GRAPE-6 Processor board

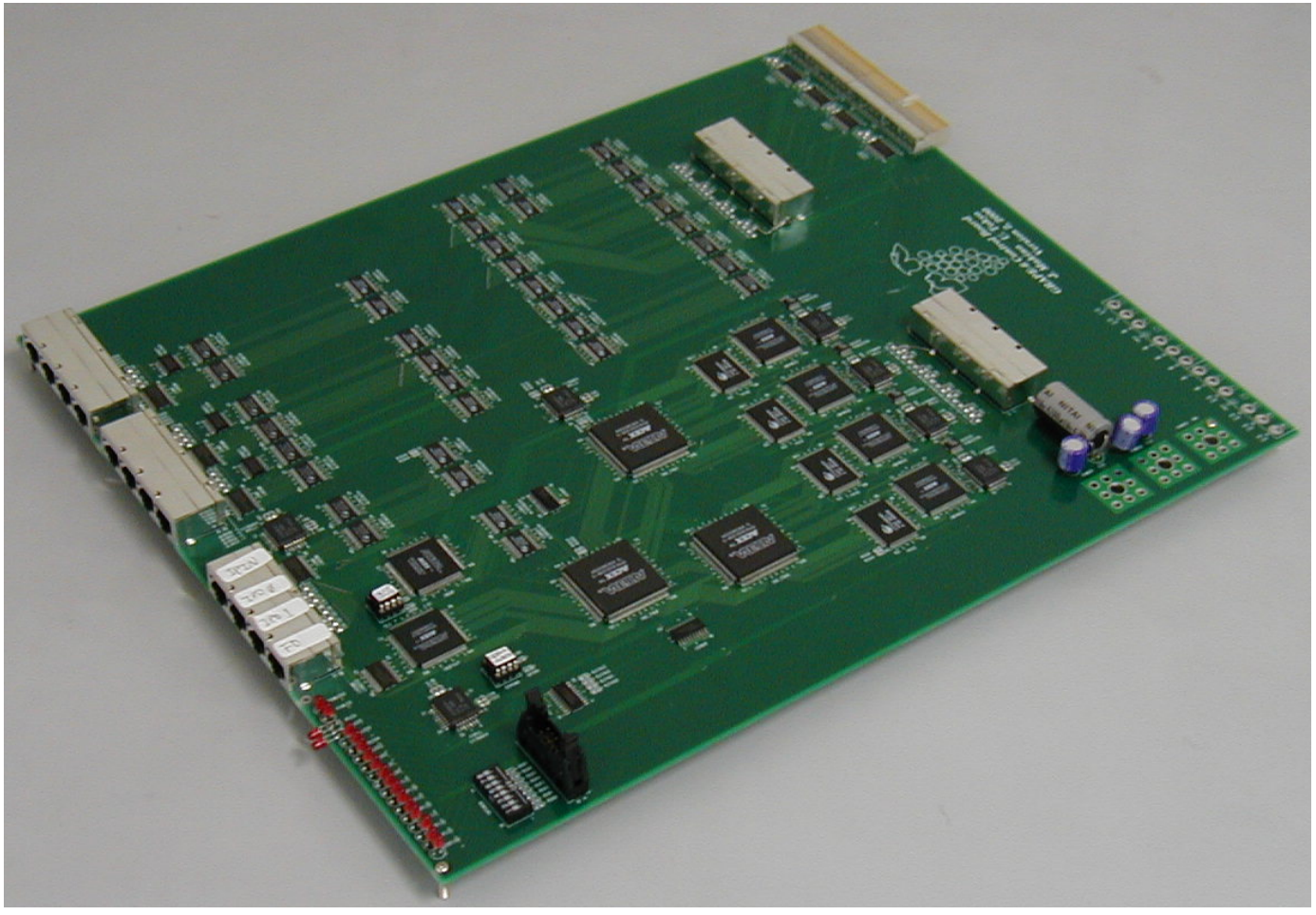


- **ボード 1 枚に 32 チップ**
- **セミシリアル (LVDS) インターフェース (350MHz clock, 4 wires)**

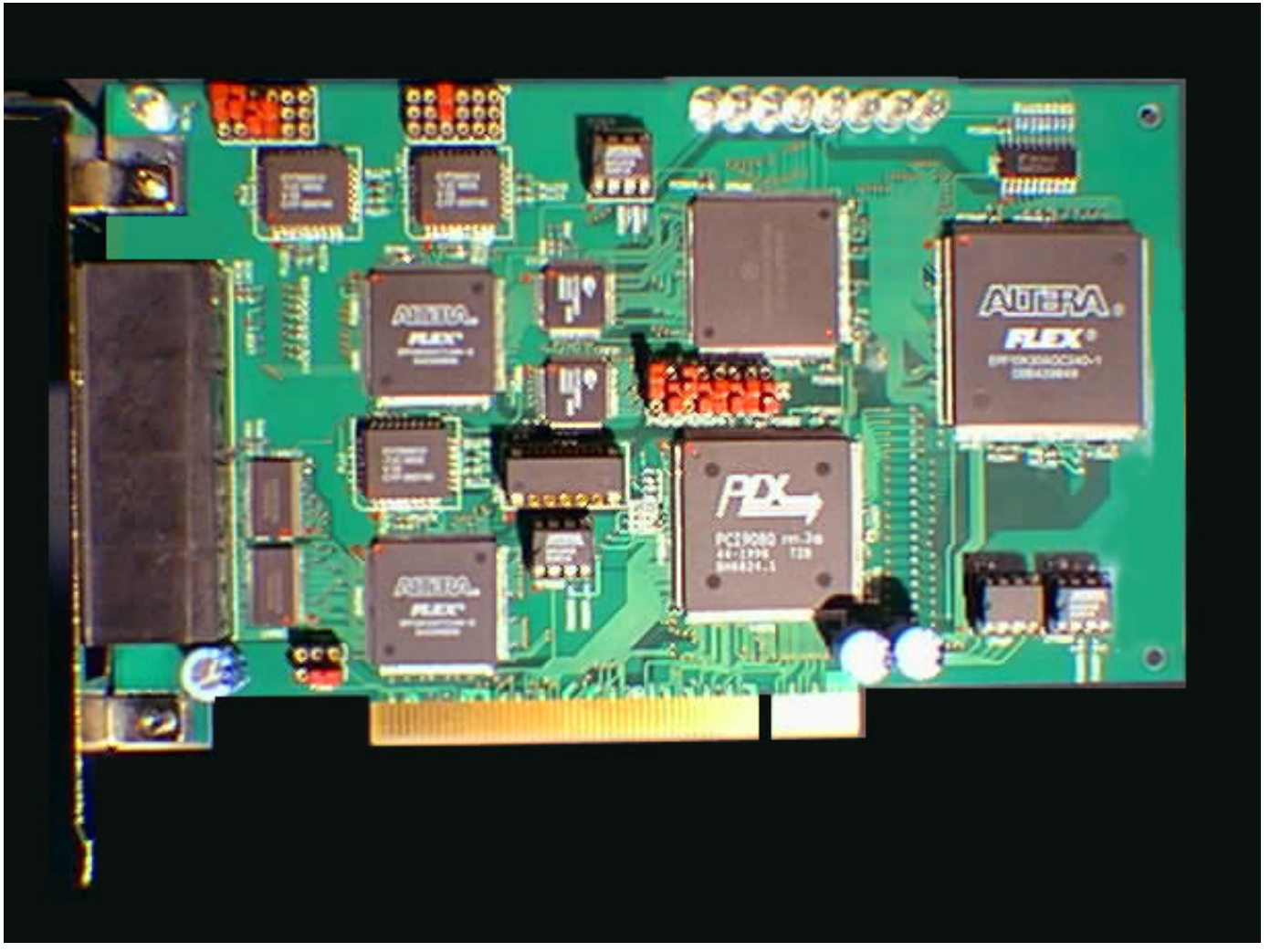
GRAPE-6 processor board



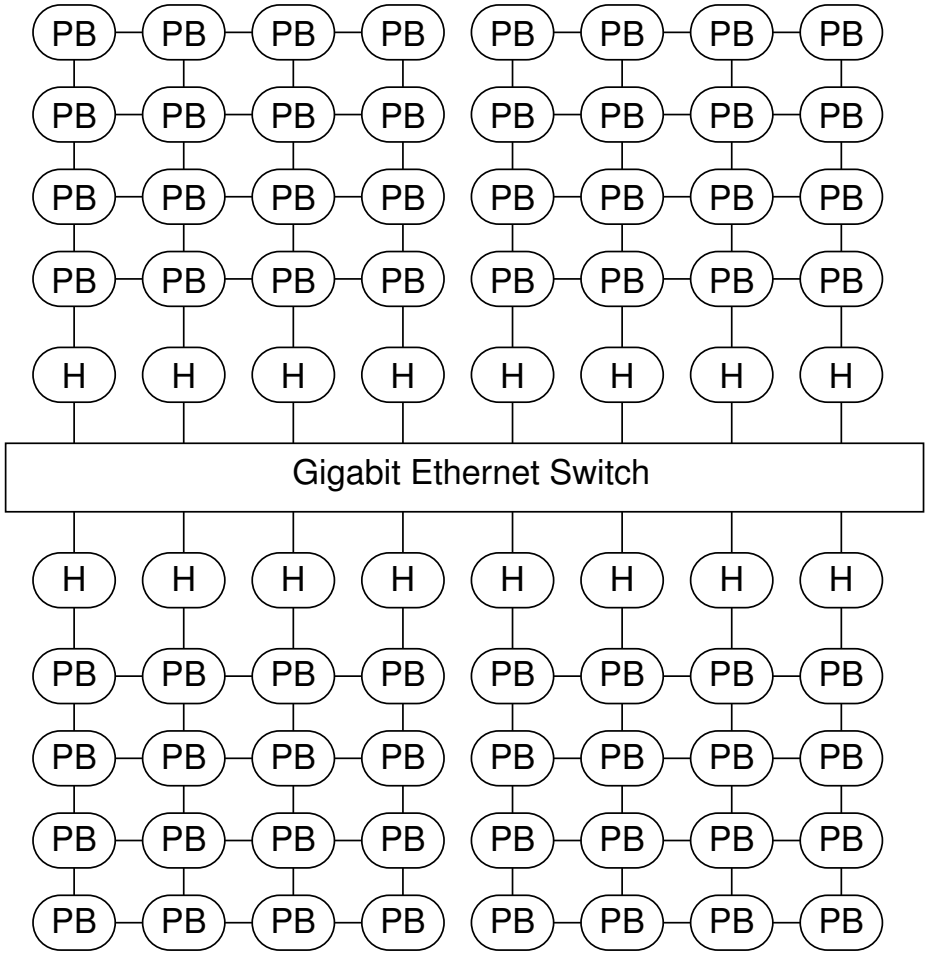
GRAPE-6 network board



GRAPE-6 host interface board



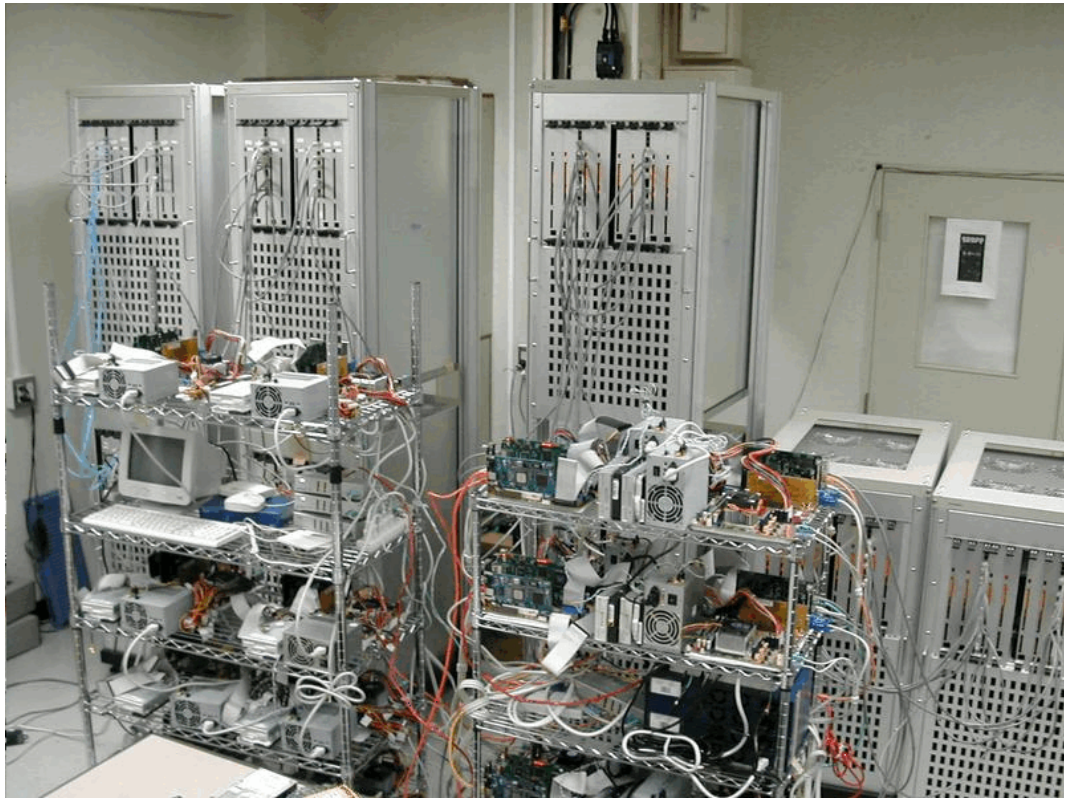
The full 64 Tflops GRAPE-6 system



- 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

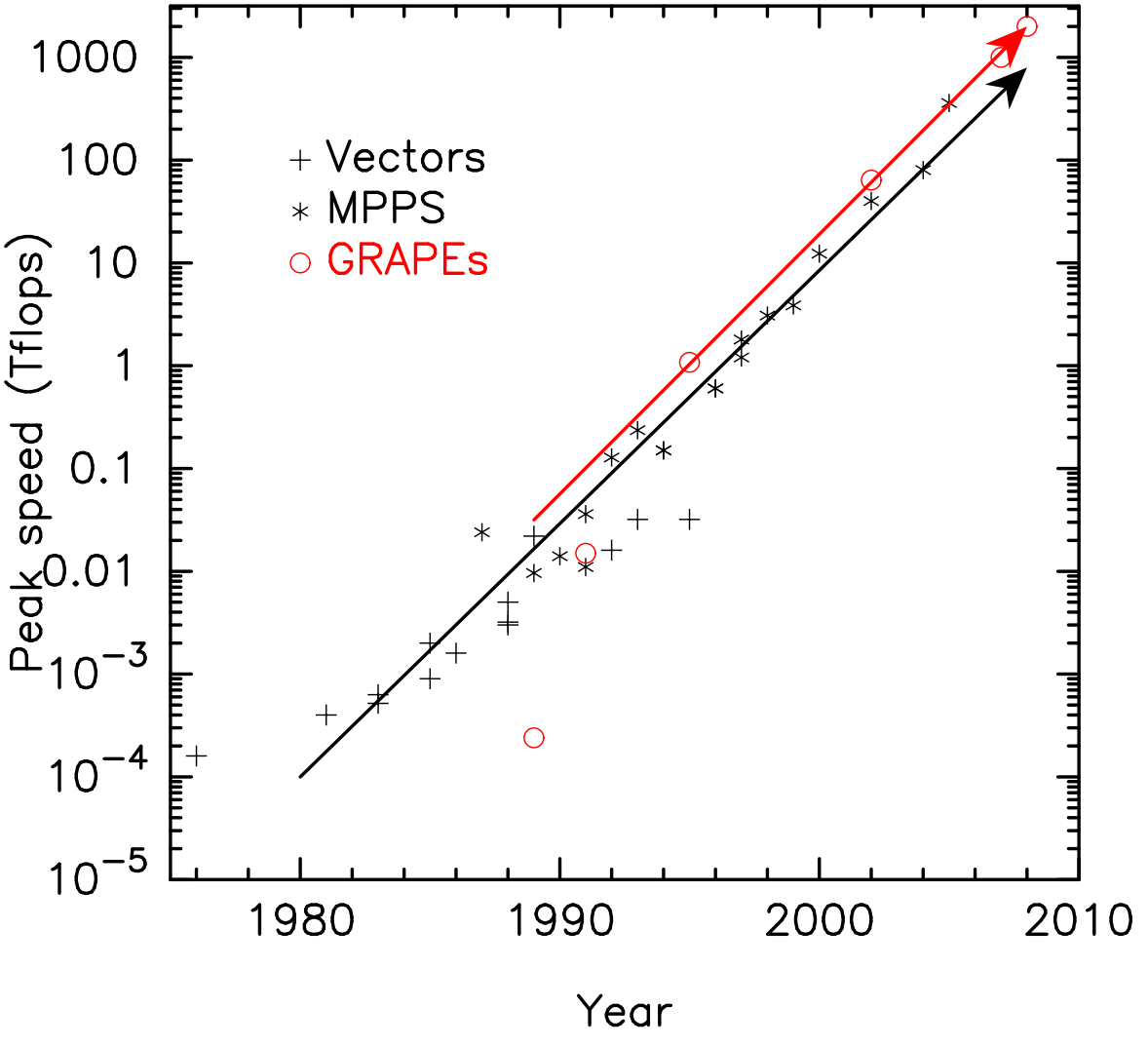
他の GRAPE 型機械

- 1992 GRAPE-2A : 最初の MD 用 GRAPE
  設計:伊藤・福重
- 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

商業版 GRAPE

- GRAPE-3 から商業版
- GRAPE-6 を購入した機関

American Museum of Natural History
Drexel University
Indiana University Rochester Institute of Technology
Rutgers University
Rochester Institute of Technology
University of Michigan
University of California
McMaster University
The United Kingdom The University of Cambridge
University of Edinburgh
Observatoire Astronomique Marseille-Provence(OAMP)
Astronomisches Rechen-Institute (ARI)
Ludwing-Maximillans University
Max-Planck-Institute fur Astronomic

GRAPE-6 を購入した機関 (つづき)

University of Bonn
University of Mannheim
Holland University of Amsterdam
Nanjing Univesity
Citec Co., Ltd
Gunma Astronomical Observatory
Hokkai-Gakuen University
Kansai University
Kyoto University
National Institute for Fusion Science (NIFS)
National Astronomical Observatory of Japan
Osaka University
The University of Tokyo
Tokyo Institute of Technology
University of Tsukuba

約 30 機関、 60Tflops。MDGRAPE-2 も同様

GRAPE-6 の次は？

MDGRAPE-3 の次: MDGRAPE-4, 20Pflops@2010

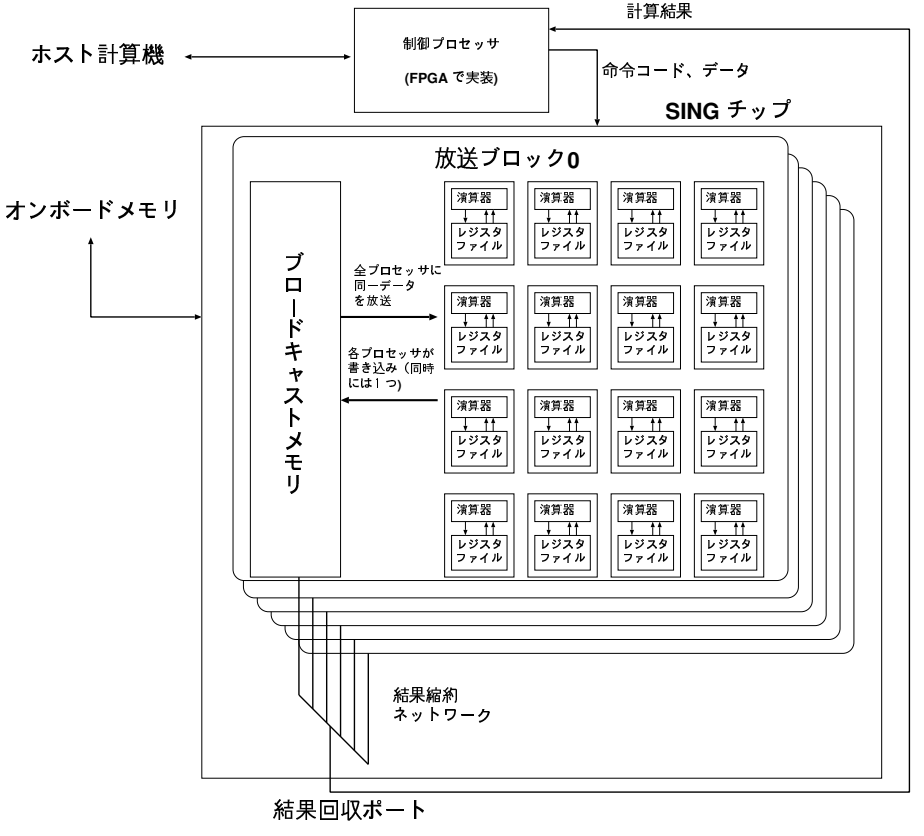
そもそも MDGRAPE-3 にあたるものは? → GRAPE-DR

平木さんが既に詳しく話をしたので簡単に

GRAPE-DR のアプローチ

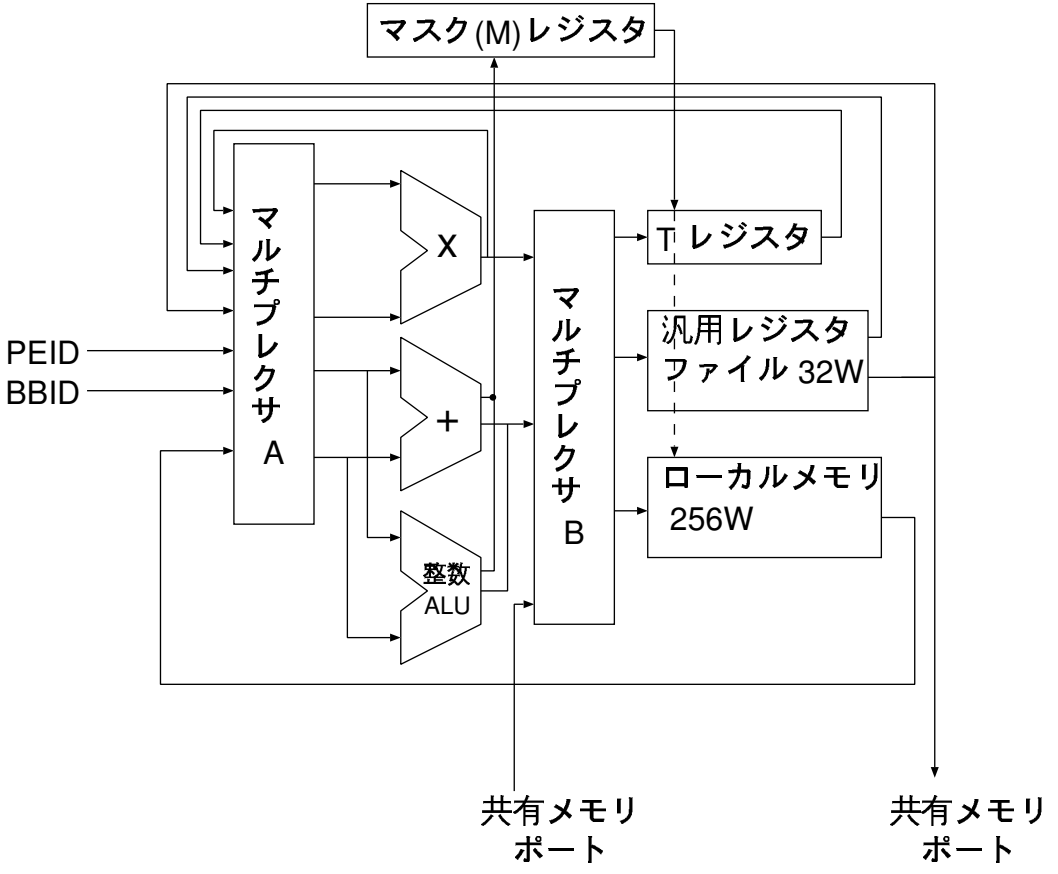
Greatly Reduced Array of Processor Element with Data Reduction

- 沢山の単純なプロセッサを 1 チップに集積
- メモリバンド幅はコストに効かない程度に抑える
- プロセッサを 2 階層でグループ化
- グループにまたがった縮約のためのネットワーク



(従来の GRAPE に比べて) 色々なアプリケーションに使える

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回しか命令ははいらない。
- 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
  アプリケーションのみを対象にする
- 3 つの型に特化
  - 散乱実験型
  - 粒子間相互作用型
  - 密行列型
- 可能ならばアプリケーションを書き換える

散乱実験型

- 多数の PE が、独立にイベントを計算
  - イベント間の相互作用はない、または非常に少ない
    - 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
    - 3 体問題:連星と単独星の遭遇、微惑星同士の遭遇
- “Embarassingly Parallel” とほぼ対応
- 古典的 SIMD 機と同様の振る舞い:
  - Goodyear MPP, ICL DAP, TMC CM-1/2, Maspar MP-1/2
  - 極端に少ないメモリ
  - PE 間通信が遅い
- 計算速度と通信速度の比:
  - 散乱実験の計算がどれだけ複雑かで決まる

粒子間相互作用型

$$f_i = \sum_j f(x_i, x_j)$$

- 他の「粒子」との「相互作用」を縮約。
  - 全ての相互作用を並列に計算可能
  - 同じ「粒子」のための計算結果を高速に縮約する必要
- 計算手順
  - PE に相互作用を受ける粒子をロード
  - 相互作用を及ぼす粒子をロード
  - 計算機終了したら結果を縮約しながら回収
  - 計算速度とチップ外への通信速度の比:
    相互作用を及ぼす粒子数に比例

密行列型

$$c_{ij} = \sum_k a_{ik} b_{kj}$$

- 計算手順
  - 行列が PE に収まるところまで分割。それから
  - 行列 A の部分行列を PE にロード
  - B の1列を分解して各グループにロード
  - 各 PE で B の部分列と A の部分行列の積を計算
  - 計算が終わったものから順次回収。グループ間で合計
- 計算速度・通信速度の比はチップ全体にロードできる行列のサイズに依存
  - メモリサイズの平方根に比例して通信速度を落とせる

計算・通信比のまとめ

- 散乱実験型: アプリケーション依存
- 粒子間相互作用型: 粒子数依存
- 密行列型: オンチップメモリサイズ依存
- 設計におけるトレードオフ:
  - なるべくアプリケーション範囲を広く
    - メモリを多く、バンド幅を広く $\rightarrow$ コスト増
  - コストを圧迫しないようにバランスを考える必要あり
- 実際の設計では密行列型の要求がもっとも厳しい

GRAPE-DR の開発状況

チップ評価ボードができたところ。これから動作チェック

GRAPE-DR の「弱点」

- 非常に通信、メモリ転送の多い計算
  - 疎行列の反復解法
  - 流体、波動方程式の陽解法
  - FFT

要するに、

- ノード内バンド幅
  - オンボードメモリのバンド幅、容量
  - ホストとの転送バンド幅
- ノード間バンド幅

が少ない、ということ。

- プログラム環境は？
- 研究者が作ったハードウェアが実用になるの？

という話は今日はパス。

プログラム環境の話はまた他の誰かが別の機会に

メモリ/ホストバンド幅

実は GRAPE-DR に限った問題ではない。 SIMD 並列で演算器を増やすと必ず起こる。

- ハイエンド GPU
- Clearspeed

どちらも外部メモリ/ホスト通信バンド幅で性能がリミット

GRAPE-DR の場合:
　　縮約ネットワークが有効な問題では高い性能

それ以外の問題では？

メモリバンド幅を増やす

- 単純に外付メモリを速くする
  - 50GB/s くらい(今のハイエンドGPU並)ならできる
  - 計算:メモリ速度で 100:1 くらい
  - FFT・流体・QCDで性能出す: 10:1 が必要
  - 500GB/s は大変。 3Gbit/s の線でも 1000本以上
- チップ積層とかでないと大変

オンチップメモリではどうか？

- 速度はどうにでもなる
- 電気は食うかもしれないけどしれてる
- 容量は問題？
- ファブは？ IBM, N, S, T? 任天堂はN？

eDRAM の使用例

BG/L	130nm	4MB	11mm by 11mm
QCDOC	180nm	4MB	13mm by 13mm
PSP	90nm?	4MB	?
Xbox 360	90nm?	10MB	?

- BG/L の場合: 1.2Mbits/mm^2
- 65nm, 256Mbits なら 50mm^2
  もとが大きなチップなら OK!

32MB 載せると

- 1 チップ 1 Tflops (GRAPE-DR の 4 倍、結構無理)
- 20,000 チップ (全体ピーク 20Pflops)

とすると、メモリ 640 GB — これは結構大きい。

- QCD 計算: 200^4 くらいの格子なら入る？
  - ノード間通信のことは後で
  - チップ内部構成の話は今日は省略
- 流体計算 2000^3 くらい？
  - 圧縮性流体 (陽解法) なら領域分割で大きな計算も可能
  - 陰解法ではポアソン方程式が問題。これも後で

ノード間通信は？

- 流体計算:あまり問題にならない
  チップ内に入るメッシュが大きいため
- 自由度が小さい計算 (QCD 等) では問題

参考: QCDOC のバランス
1 Gflops : 500Mbits/s $\times 8$ (4D トーラス)
$= 0.5$ bytes/flops

1 Tflops の場合でも 0.25 bytes/flops くらい欲しい
(扱う格子のサイズはあまり変わらない)

- APE、筑波の数字はもうちょっと小さい?

チップ 1 つが Tflops とすると、これは 250GB/s。
非現実的？

現実的なネットワーク

今あるものの例:

Cray XT3: 7.6GB/s 3次元トーラス。
ノード当りの売り値 100万くらい

- 計算機メーカーなら4年後にこの5倍くらいできる？
- 我々では無理

XT3 並の 10GB/s 3D トーラスを作ったとすると:

- 通信速度は必要量の 1/4
- QCD 計算での効率は 20% くらいはなんとか
- 3次元 FFT でもその程度の効率。 4000^3 ができる
- 実効性能では 4Pflops

作れそうか？

- チップから信号出すのは可能。 1Gbit/s で 1,000本
- ルータ: 我々には技術、経験ない
  → やるならどこかと共同開発

「V-GRAPE 構想」

GRAPE-DR、 ClearSpeed との違い

- かなり大きなオンチップ DRAM をつける
- メッシュネットワークを組めるだけの信号線を出す
- メッシュネットワークを使う構成では BG/L, XT3 のような高密度実装
- 使わないなら PC クラスタ+加速ボード構成

“Versatile-GRAPE”

大雑把な性能推定

ピーク 20Pflops として

- 重力多体、分子動力学、第一原理計算
  効率 30–50% (6–10 Pflops)
- 陽解法流体、有限要素法
  効率 10–20% (2–4 Pflops)
- FFT、QCD(メッシュネットワーク作った場合)
  効率 10–20% (2–4 Pflops)

本当は、

- 価格あたりの性能
- 消費電力あたりの性能

だけが問題。いわゆる「実行効率」は無意味な指標。

(但し、性能推定には実行効率は有用)

他のアプローチとの比較

筑波大学の将来プラン
48Gflops、 15GB/s、400,000 チップで 20PF

ピーク性能を同じにして比べると

- **チップ数** 1/20
- bisection bandwidth 1/2
- **ノード間接続総本数** 1/5
- **総メモリ** 1/3000

トータルコスト、信頼性、消費電力にかなりの差？

メモリ沢山いる計算: 遅いメモリをうまく使う必要あり

まとめ

- GRAPE-DR 的な1チップ超並列計算機に
  - そこそこの量のオンチップメモリ
  - そこそこの速度のネットワーク

  をつけることで、GRAPE-DR の弱点である
  - 陽解法流体
  - QCD 計算
  - 3D FFT

  でかなり良い実効性能がでるようにできる
- ネットワークはそれでもコストに響く
- ただし、チップ数が多いアプローチより有利
- ネットワーク部分はどこかとの技術協力が必要