専用計算機による大規模シミュレーション
— GRAPE を例として

牧野淳一郎

東京大学大学院理学系研究科

講演概要

1. 重力多体問題とはどんなものか
2. どうやって計算するのか
3. なぜ専用計算機を考えるのか
4. GRAPE の基本的考え
5. GRAPE アーキテクチャの発展
6. 次世代プロジェクト— GRAPE-DR
7. 専用計算機開発の今後
8. まとめ

重力多体問題とはどんなものか

支配方程式:

$$\frac{d^2 x_i}{dt^2} = \sum_{j \neq i} -\frac{G m_j r_{ij}}{r_{ij}^3}$$

- ある粒子（星、銀河、、、）は、系の中の他のすべての粒子からの重力を受けて運動する。
- 太陽系、星団、銀河、銀河団、宇宙の大規模構造などの基本方程式
- 電荷の正負がある他はプラズマと同じ

星団の進化

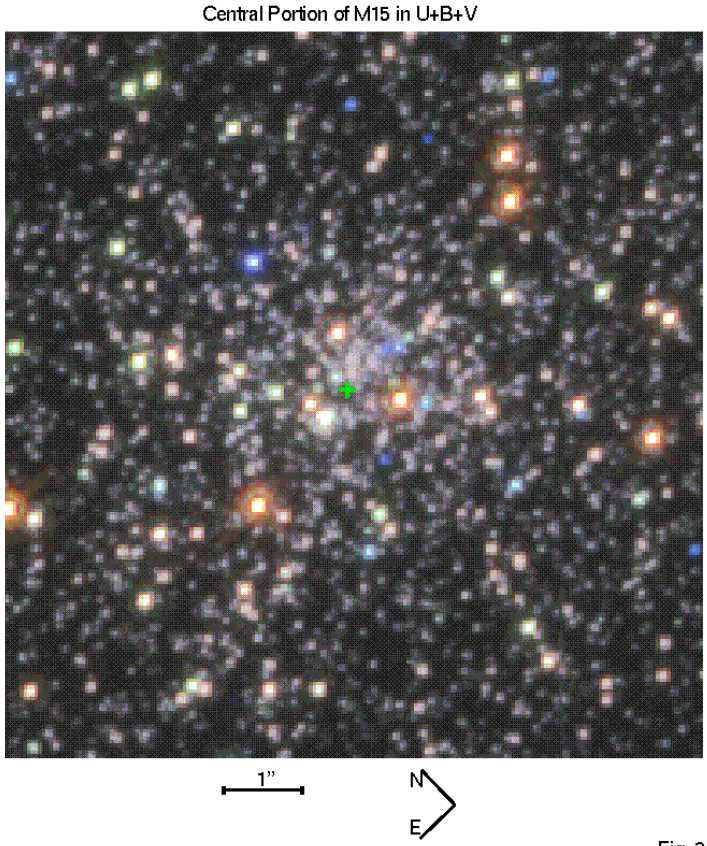
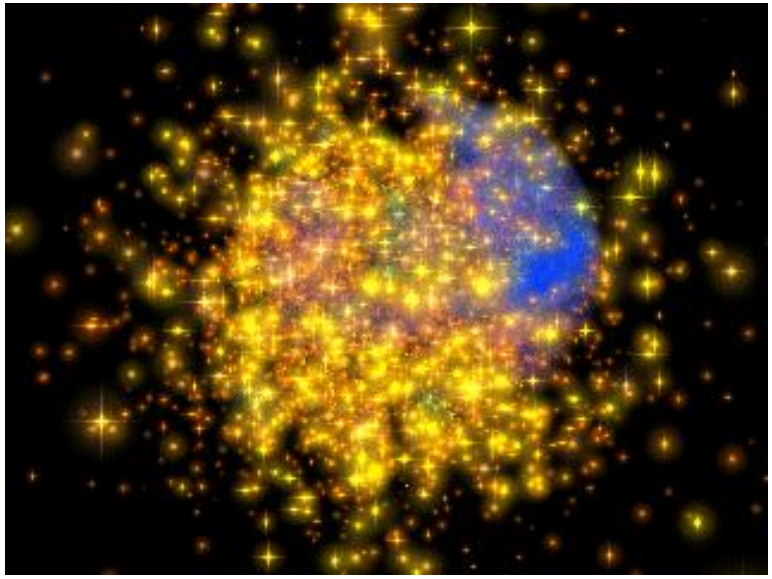


Fig. 2



観測

（球状星団 M15 の中心部）

シミュレーションの例

最近の話題の例

- 球状星団の中心にブラックホールはあるか？
- 「中間質量」ブラックホールはどうやってできたか？

観測の解釈に、シミュレーションが重要な役割を果たす。

- 熱力学的な進化が重力の非線型性とカップルして、自律的な構造形成を起こす。
- 普通の意味で実験ができない

計算機シミュレーションが理論と観測をつなぐ。

どうやって計算するか？

- 基本的には力任せ: 一つの粒子は他のすべての粒子からの重力を受ける
- 計算量を減らす方針は2種類
  - 時間方向: 粒子毎に時間ステップを変えて、「独立」に積分
  - 空間方向: ツリー法、高速多重極法
  - 両方を同時に: 難しい

(普通の) プラズマとの違い

- 引力だけ

  – デバイ遮蔽がない (デバイ長 $=$ 系の大きさ)

- 要素数が (比較的) 小さい
  プラズマ: $> 10^{23}$ 重力: $10^{2\sim10}$

$$\downarrow$$

- 2 体緩和が系の進化 (熱力学的な進化) を決める
- 熱平衡状態が存在しない

数値計算法への要求

- 2 体の重力散乱を「正確に」追う必要がある
- 空間構造の発達 (中心密度が 10 桁上がるとか) を追う必要がある
- 積分時間が長い (典型的な粒子で例えば 10^5 軌道、中心付近の粒子はもう何桁か多い)

↓

- 空間、時間的に適応的な時間刻み (粒子毎にバラバラに時間刻みを変える)
- 比較的高次の積分公式
- 粒子間相互作用の高空間分解能、高精度な計算

なぜ専用計算機を考えるのか

基本的には、汎用計算機に比べて、「速く、安く」できる
（かもしれない）から。
なぜ「速く、安い」か？

- 問題自体の特性
- 技術的な要因
- 歴史的、経済的な要因

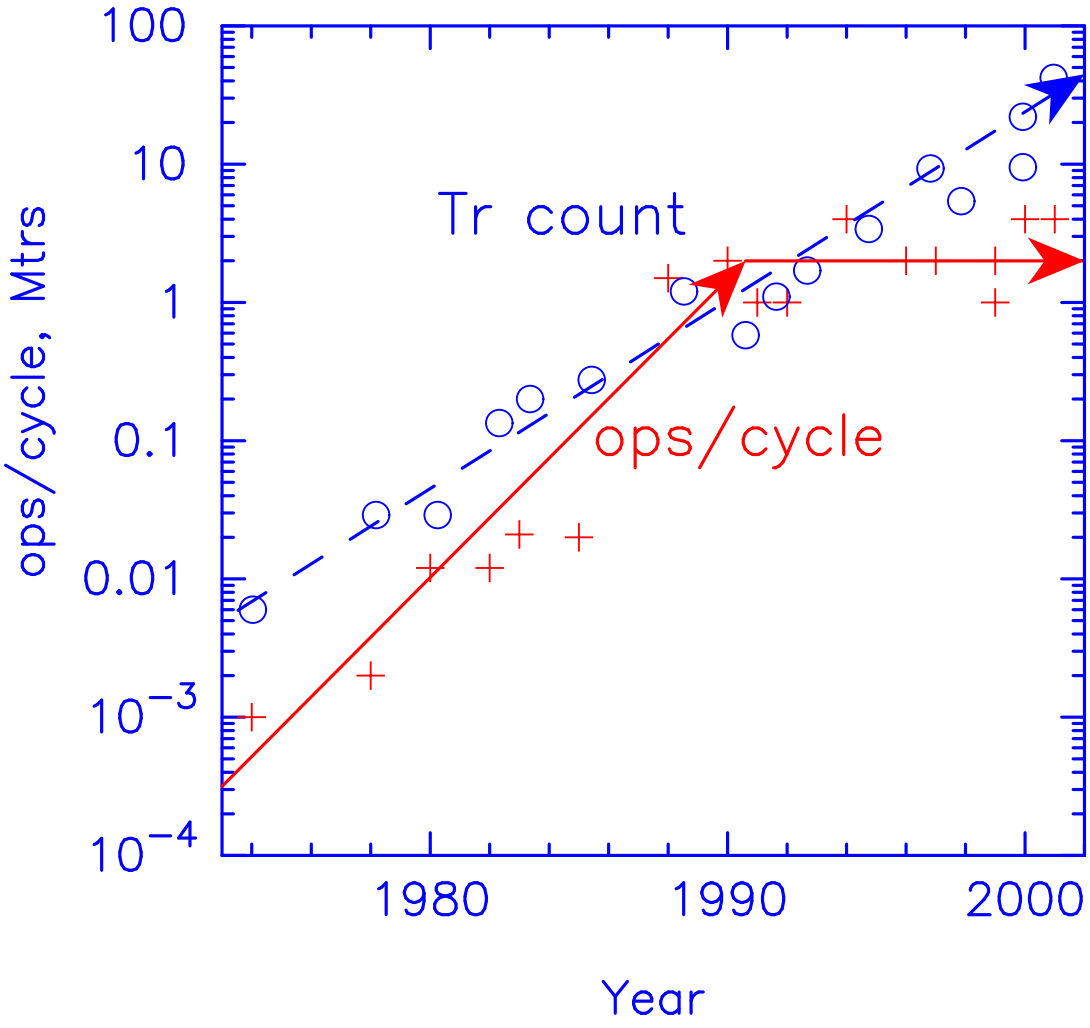
問題自体の特性

粒子系シミュレーションの特徴:一つの粒子が多数の粒子と相互作用する

- 計算量が多い（メモリ必要量に比べて）
- 計算が比較的単純な繰り返しである
- 通信パターンが規則的である

（独立時間刻み、ツリー法では細かい考慮が必要にはなる）

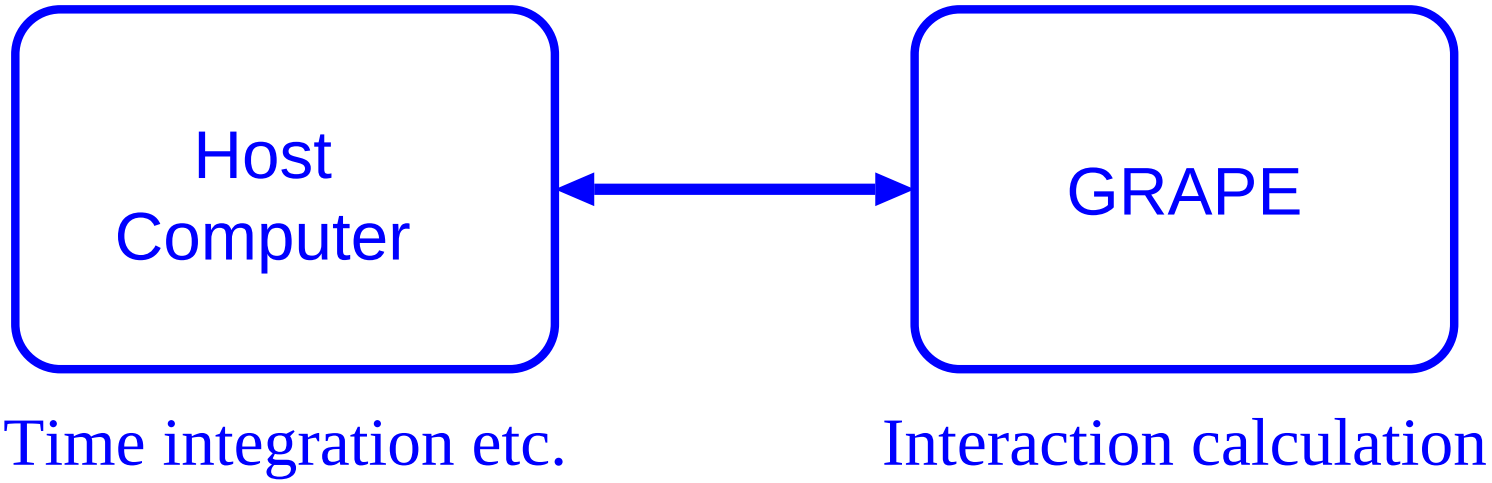
技術的な要因：マイクロプロセッサの「進化」



代表的なマイクロプロセッサの、サイクル当たりの浮動小数点演算数

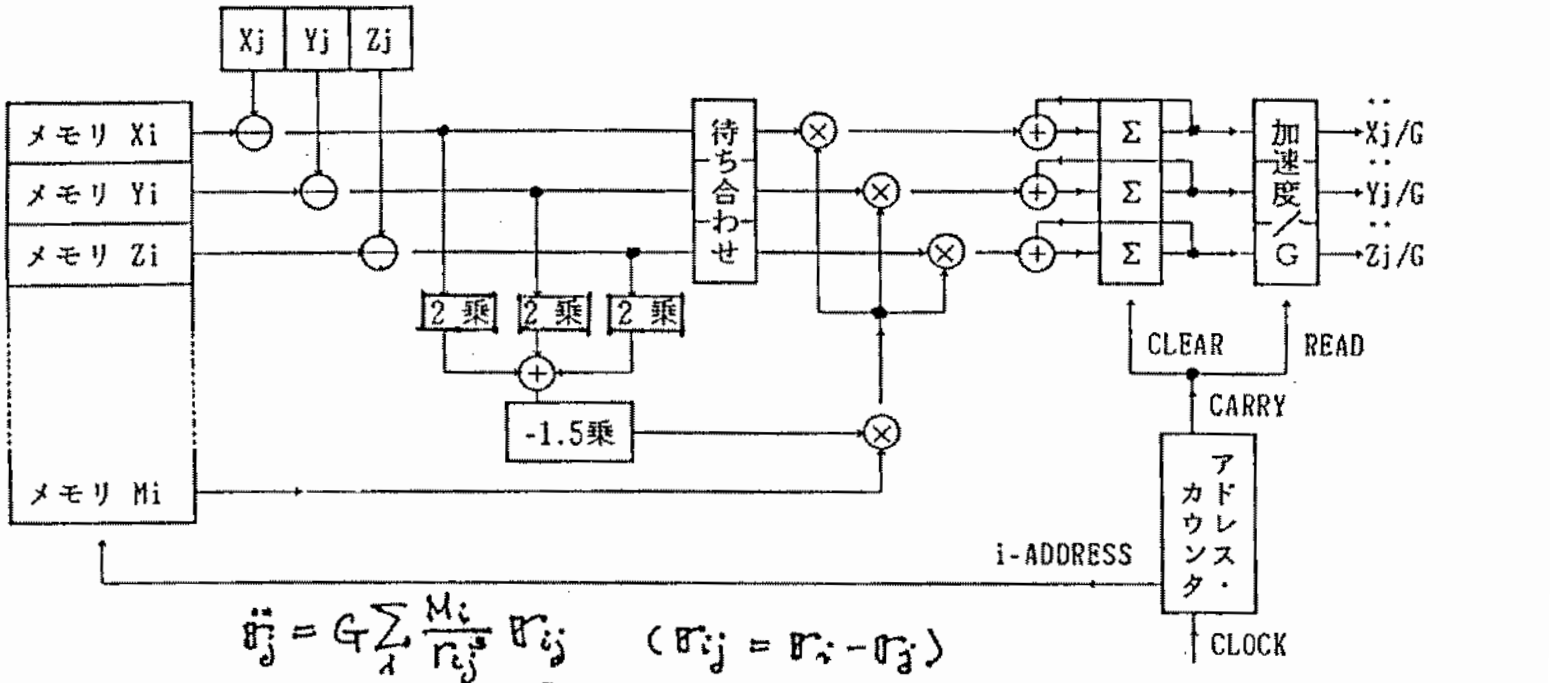
1 を超えてから、ほとんど止まっている ＝ 汎用の並列処理は難しい

GRAPE の基本的考え



専用ハード: 相互作用の計算
汎用ホスト: 他のすべての計算

GRAPE パイプライン



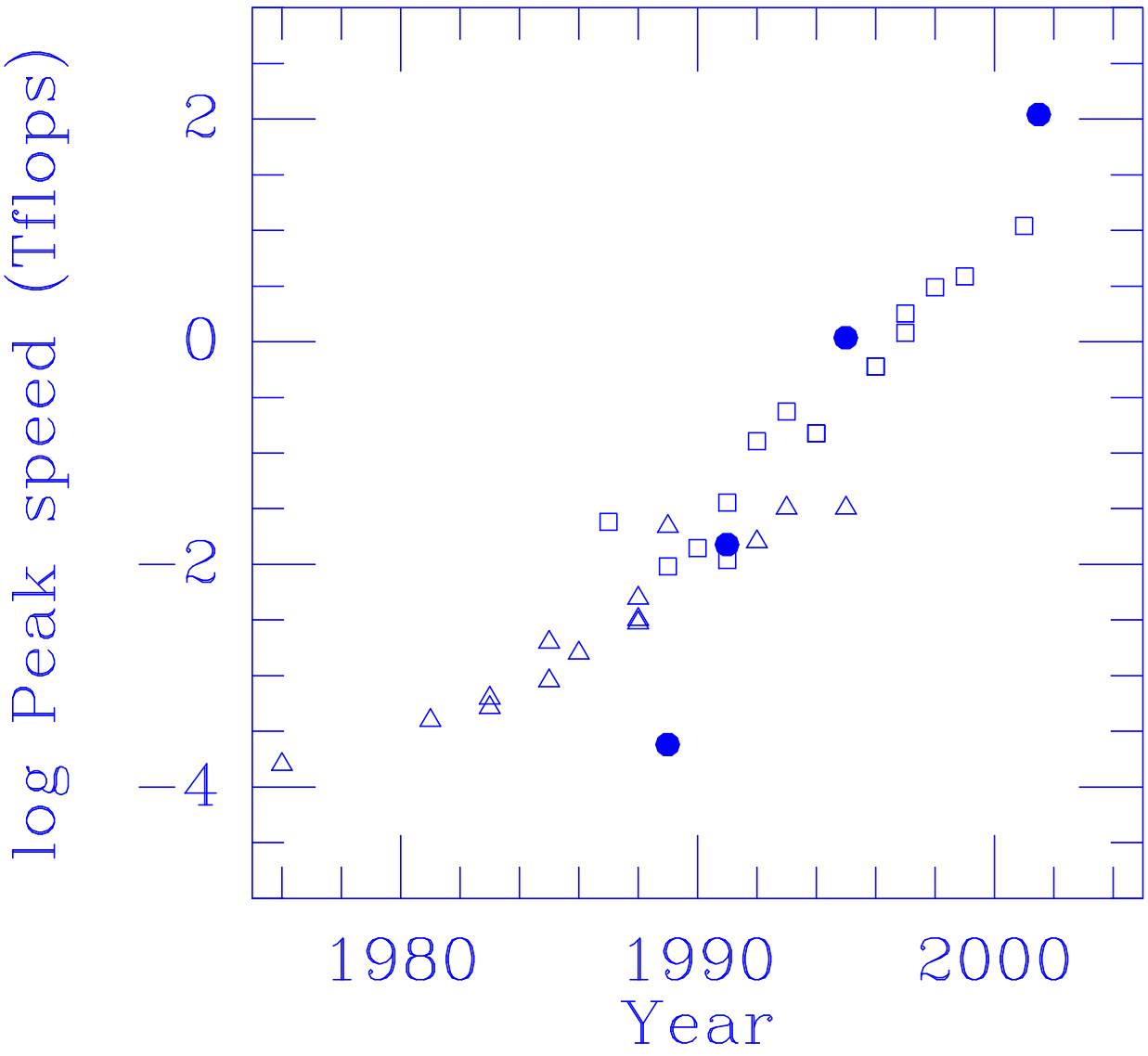
$$\ddot{\vec{r}}_j = G\sum_i \frac{M_i}{r_{ij}^3}\vec{r}_{ij} \qquad (\vec{r}_{ij} = \vec{r}_i - \vec{r}_j)$$

+，－，×，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．

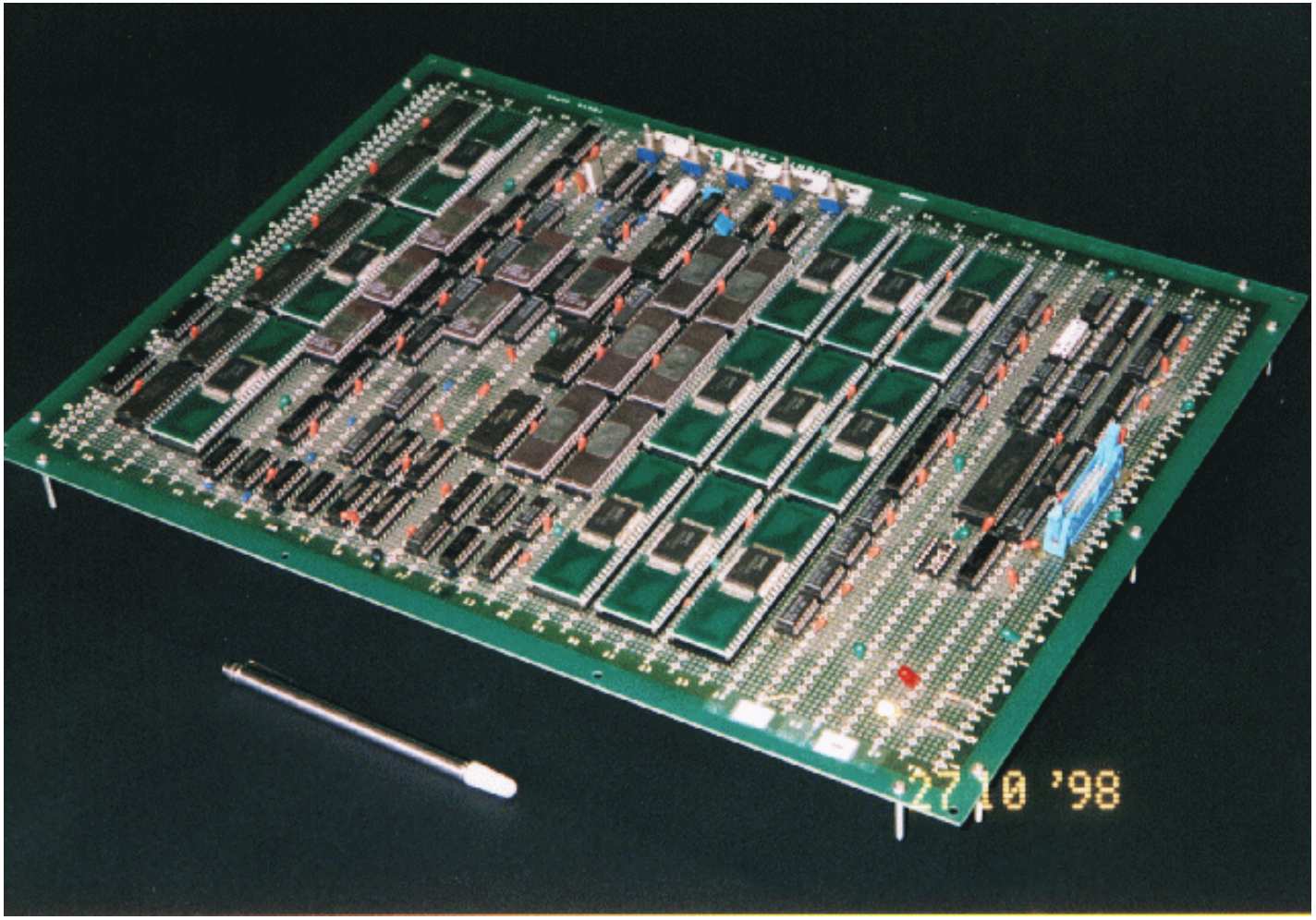
図２．N体問題のj-体に働く重力加速度を計算する回路の概念図．

(近田 1988)

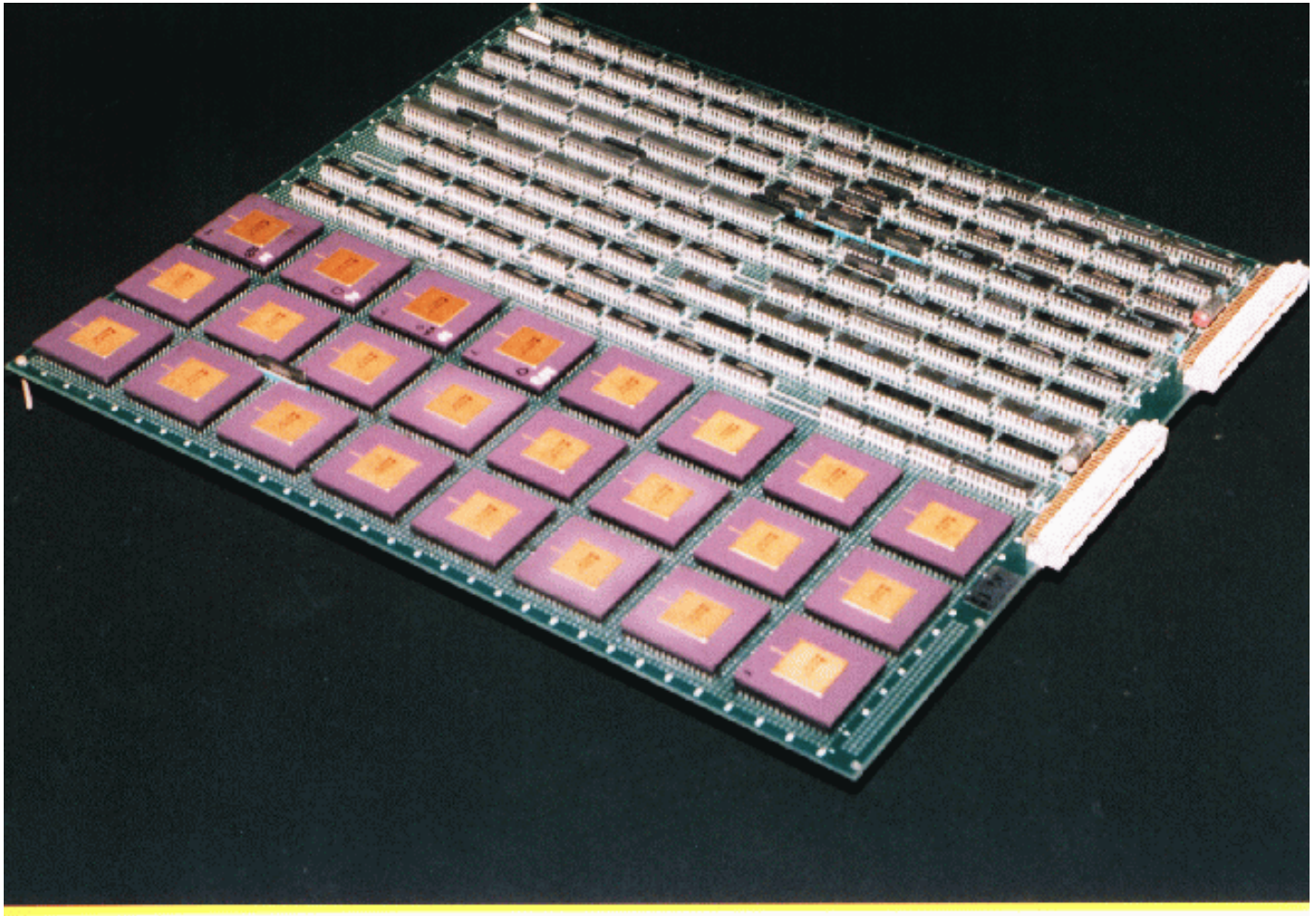
計算速度の発展



GRAPE-1 (1989)



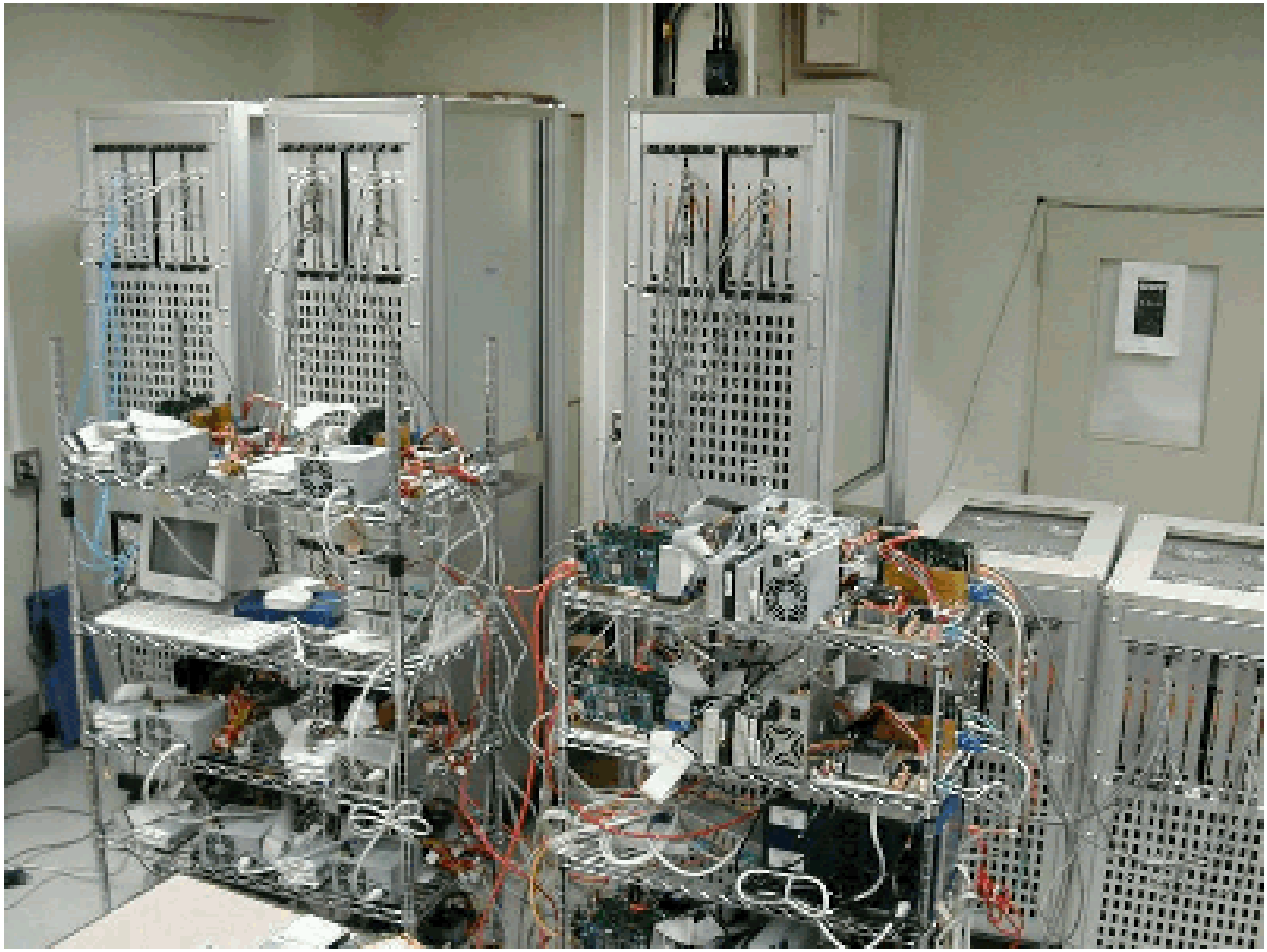
GRAPE-3 (1991)



GRAPE-6

- 1997年度からの5年計画 (総予算約5億円)、（旧）文部省（学振）未来開拓学術研究推進事業、「計算科学」部門の1プロジェクト
- ピーク性能 31 Gflops の専用 LSI を 2048 個並列動作させ、 63 Tflops のピーク性能を実現。
- IEEE ゴードンベル賞を 2000, 2001, 2003 の3回受賞。(地球シミュレータは 2002-2004)
- 総開発費: 約 5 億円

GRAPE-6 full system



MDM

- 理研の戎崎のグループが開発
- 主にタンパク分子の古典 MD 計算を対象
- クーロン力：エヴァルド法。波数空間と実空間に別々の専用ハード
- 分子間力：直接計算。クーロン力実空間と共用
- 波数空間（WINE-2）：46 Tflops
- 実空間（MDGRAPE-2）：25 Tflops
- 2001 年 3 月に完成。現在「実用計算で世界最高速」

GRAPE-DR

GRAPE-DR (GRAPE Data Reservoir): 次世代 GRAPE プロジェクト

- 目標ピーク性能: 2 Pflops
- ターゲットアプリケーション: 重力、それ以外の古典粒子系、密行列計算等
- 完成予定年度: 2008

- 背景
- GRAPE-DR の考え方
- アーキテクチャ

専用計算機の「限界」

開発費の高騰

1990	1μm	1500万円
1997	0.25μm	1億円
2004	90nm	3億円以上？

ある程度広い応用を持つものでないと難しい

FPGA、「リコンフィギャラブルデバイス」は？

トランジスタ利用効率で大きく劣る

↓

データ語長が短い応用でないと高性能は難しい

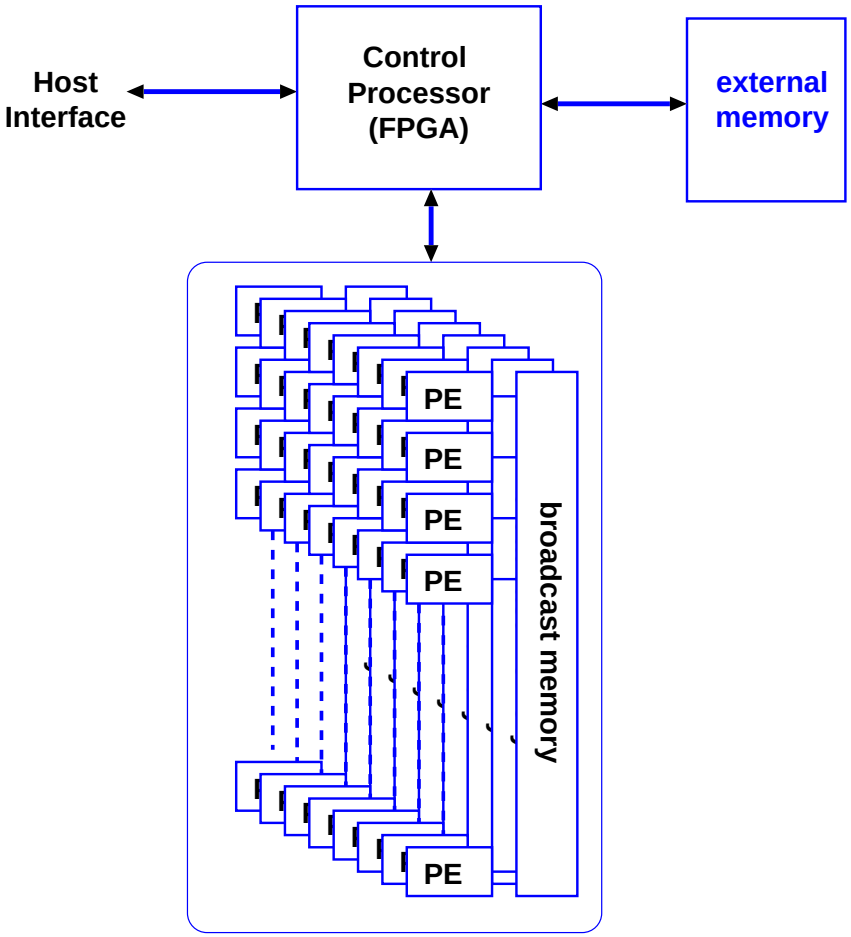
別のアプローチ？

- 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
- 「重力だけ、天文だけ」と言われないようにする

↓

GRAPE-DR アーキテクチャ

GRAPE-DR の アーキテクチャ



- 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)
- PE はプログラムによって並列動作する
- 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ユニット (BU) 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

重力計算以外の応用

- 分子動力学・SPH 等の粒子法計算
- 密行列計算 (Linpack, LU 分解、固有値計算)
- 境界要素法: ポアソン方程式、ヘルムホルツ問題、、、
- ルジャンドル展開による極座標流体計算
- 分子軌道法での 2 電子積分

専用計算機の今後

専用アーキテクチャ：汎用計算機に比べた相対的優位はだんだん大きくなる

ムーアの法則でトランジスタは増える

汎用計算機では増えたトランジスタの全部を演算器には使ってない
専用計算機は全部使える (使えるように作る)

ムーアの法則が成り立たなくなる頃 ???

汎用計算機の今後 ???

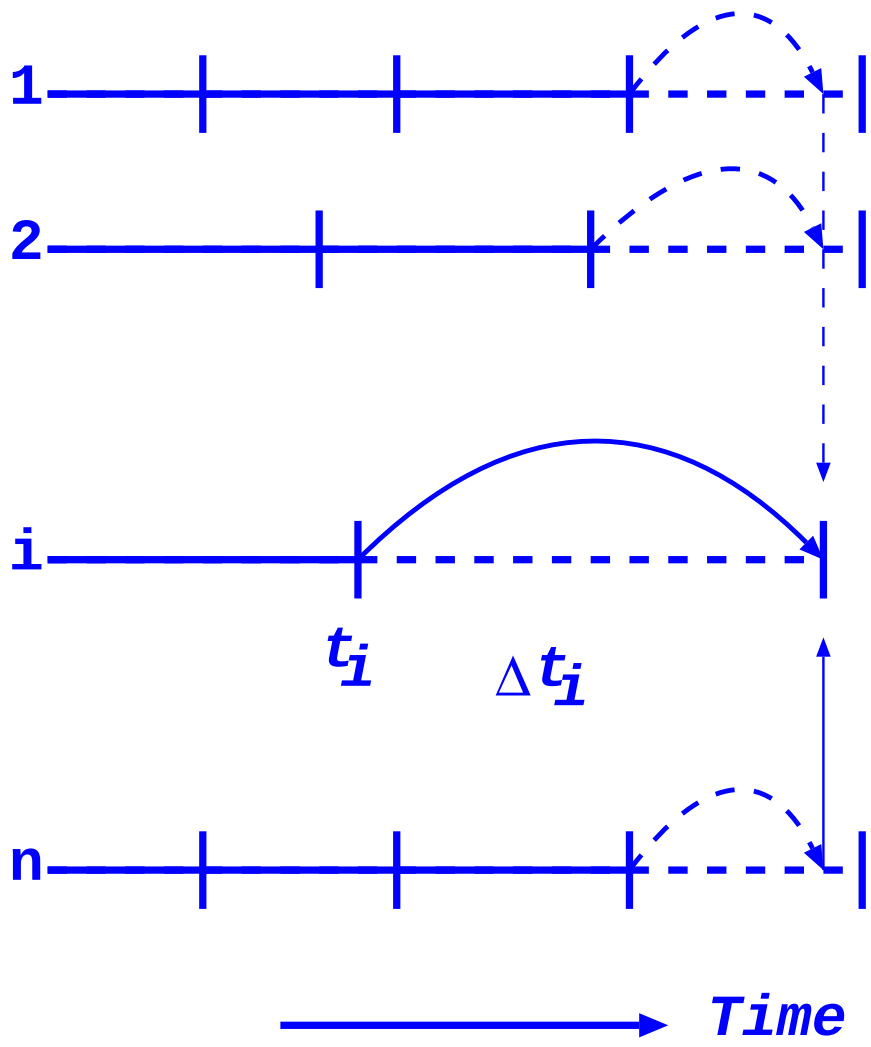
本当はこっちが問題。
GRAPE: 汎用計算機につける「アクセラレータ」
5 年後、10 年後の（科学技術計算用）汎用計算機はどんなものか？（そんなものがあるのか？）

x86 アーキテクチャで科学技術計算をしていていいのか？
ベクタパラレルに未来はあるのか？

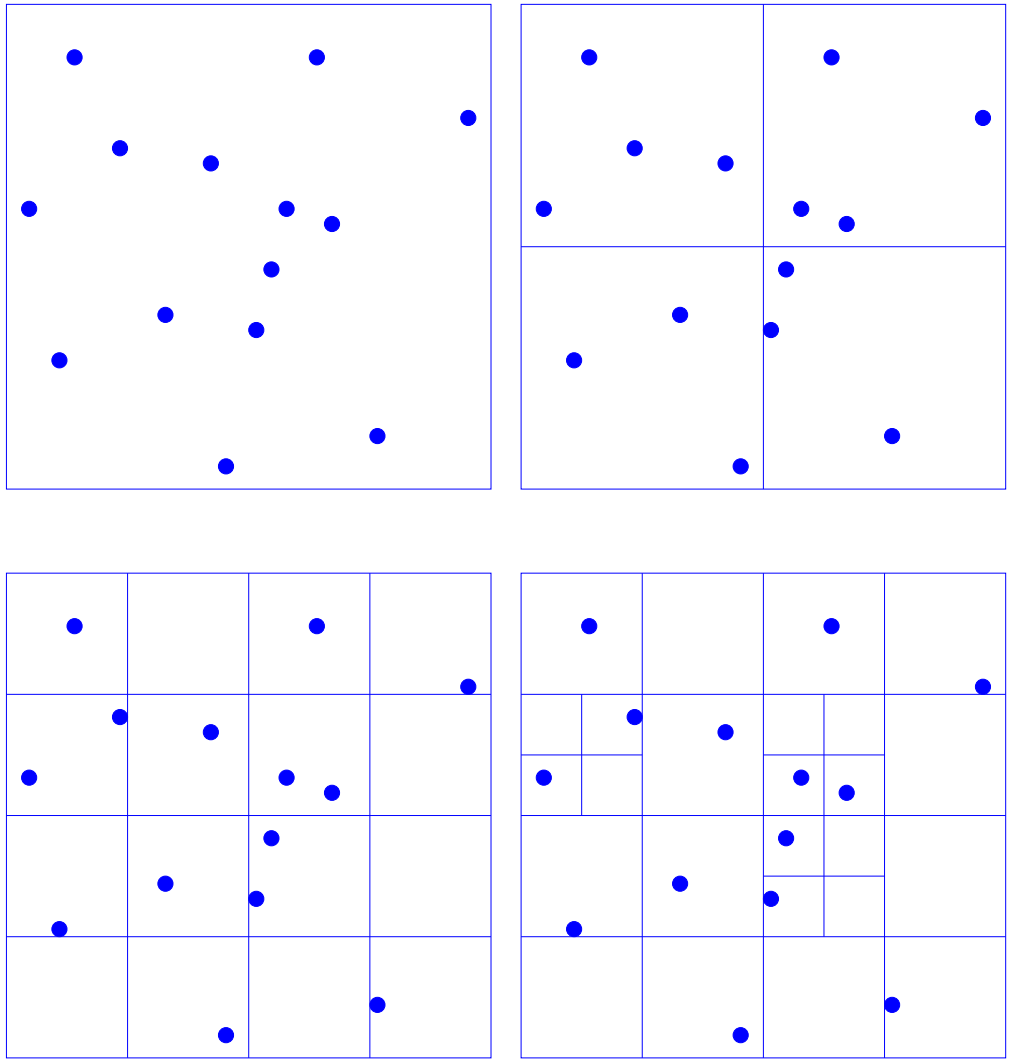
まとめ

- GRAPE では、もっとも演算量の多い相互作用の計算だけを専用化し、他の計算は汎用計算機に回すことで高い性能と柔軟性を両立させている。
- この分担のためにハード開発は比較的容易である。
- このような分担をうまく機能させるためには、採用するアルゴリズムがそういうふうにできている必要がある。可能ならアルゴリズムの変更も考えるべき。
- 5-10 年先でも GRAPE の方法は多分有効
- GRAPE-DR では 重力多体系外に応用範囲を広げる

独立時間刻み



ツリー法



GRAPE における並列化

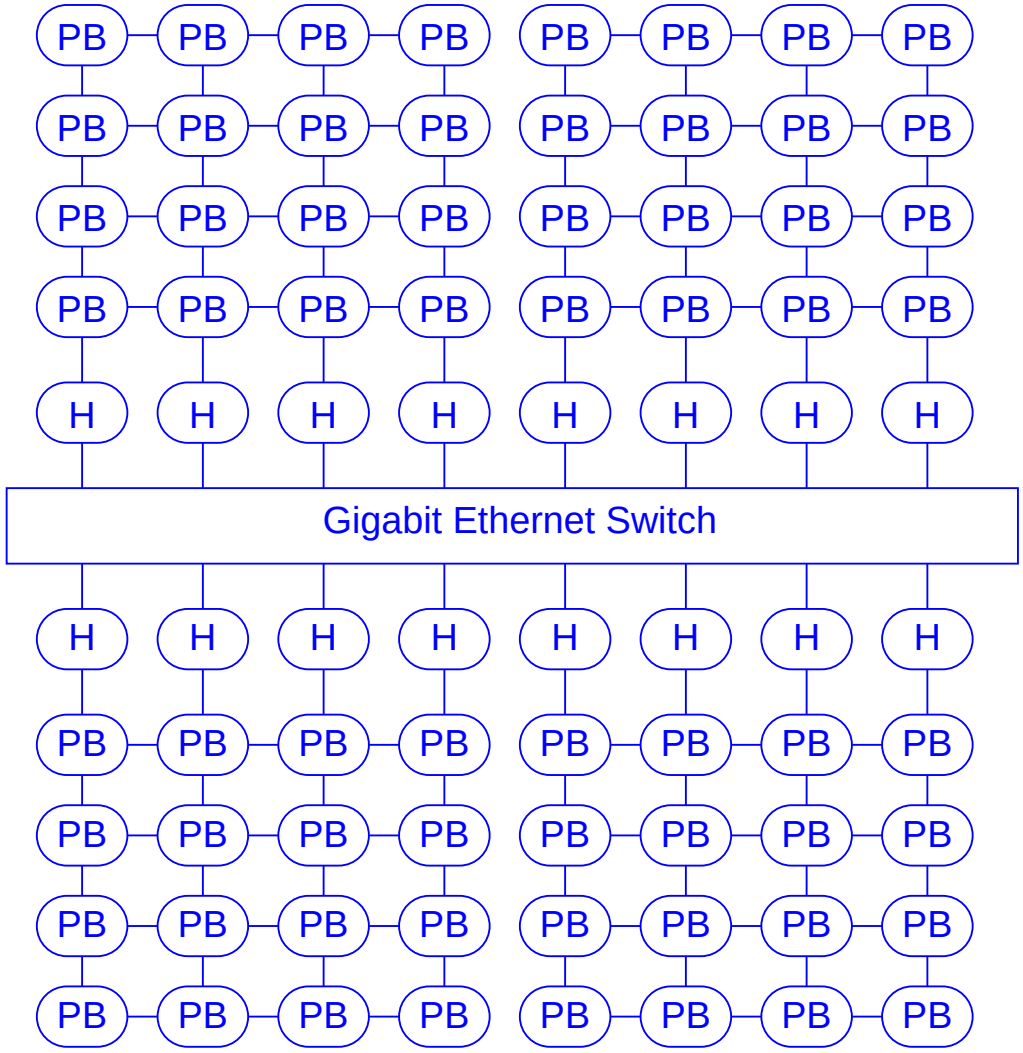
パイプライン 1 本が速くても、並列化できなければたいした意味はない。
ホストは 1 台として、 GRAPE の側の並列化を考える。
方針は 2 つ

- 複数の粒子への力を並列に計算する (i 並列)
- 複数の粒子からの力を並列に計算する (j 並列)

GRAPE アーキテクチャの発展

1989 GRAPE-1 低精度、EPROM で演算
1990 GRAPE-2 高精度、浮動小数点演算 LSI
1991 GRAPE-3 低精度、カスタム LSI
1995 GRAPE-4 高精度、カスタム LSI 、超並列
1998 GRAPE-5 低精度、複数パイプラインを集積
2001 GRAPE-6 高精度、複数パイプラインを集積

全体構造



- ホスト PC 4台とプロセッサボード 16 枚で1クラスタを構成
- 4クラスタで全体マシン。ピーク 64 Tflops

GRAPE における並列性

する計算:

$$a_i = \sum_j f(r_i, r_j, m_j)$$

i についても j についても並列 (j は総和が必要)

1. **パイプライン演算 (j 並列)**
2. **物理/仮想マルチパイプライン (i 並列)**
3. **複数チップ (i 並列/j 並列)**

「パイプラインであること」はあまり本質的ではない？

「GRAPE として」使う

最も単純には

- 全ての PE が、自分の粒子への、同じ粒子からの力を計算
- 力を及ぼすほうの粒子データは外部メモリから供給

現実問題としては

- 違うブロードキャストユニットの同じ位置の PE には同じ粒子データを書く
- 各ブロードキャスト
- 力を及ぼすほうの粒子データはブロードキャストユニット毎に違うものにする。

つまり、 ブロードキャストユニット内で i 並列、ブロードキャストユニット間で j 並列とする。

普通の計算機を作るのに比べると

普通の計算機では

- 対象となる問題、アルゴリズム、要求精度についての知識

「必要ない」（本当？）

- それ以外

「まあまあ」ではまあまあの計算機しか出来ない。

他の誰かと同じようなことをすれば済むという面もある（1番になれるか？という問題はある）

あんまりうまくいかないのはなぜ？

どれが成功でどれが失敗かというのはさておき、、、
基本的な理由は 2 つ。

1. 出来たものが実は使えなかった。
2. 出来た時には速度が汎用計算機より速くなかった。

理由 (1) は専用計算機に固有だが、例は少ない（公表されていないのがあるのかもしれないが）
理由 (2) は普通の計算機と同じ。開発に時間がかかるとムーアの法則の分相対的に遅くなる。

専用計算機開発の損得

15 年ばかりやってきてどうだったか？ということ。

もともとターゲットとした問題（天文の多体問題）に対しては、予想以上の成果をあげたといって嘘にはならないと思う。

GRAPE は 100 枚以上のコピーが世界中の 20 以上の研究機関で使われ、惑星形成や星団の進化のシミュレーションのほとんど、また銀河形成や宇宙論シミュレーションについてもかなりの割合が GRAPE を使って行なわれている。

専用計算機開発の損得

GRAPE のようなものを「うまく」作るのは結構大変。

- 対象となる問題
- アルゴリズム・要求精度
- 計算機アーキテクチャ
- 論理設計
- 実装
- OS、デバイスドライバ等のソフトウェア

の全般についてまあまあの理解が統合されていないと全体設計が出来ない。

専用計算機は難しい？

「成功」といえるようなものはそんなに多くない。
多体系に話を限るとアプローチは2つ。どちらも結構いろいろある。

専用パイプライン

- DMDP (Delft)
- FASTRUN
- GRAPE
- MD-Engine
- MDM

並列計算機

- Digital Orrery
- Transputer-based projects...
- HaMM
- 他にも沢山

上手くいかない基本的な理由: できた時には速度が大して速くない。

開発期間の問題

大抵の人は（私を含めて）見積もりが甘いというのは確か。

本質的な問題：
専用計算機の場合、開発期間が多少（1-2年）延びたくらいで意味がなくなるようなプロジェクトは、そもそもあまり意味がない（出来てからのハードウェアの寿命がどうせ短い）。
大雑把には、計画時点で

- 価格性能比が 1000 倍 — 問題なし
- 価格性能比が 100 倍 — ちょっと苦しい
- 価格性能比が 10 倍 — やめたほうがいい

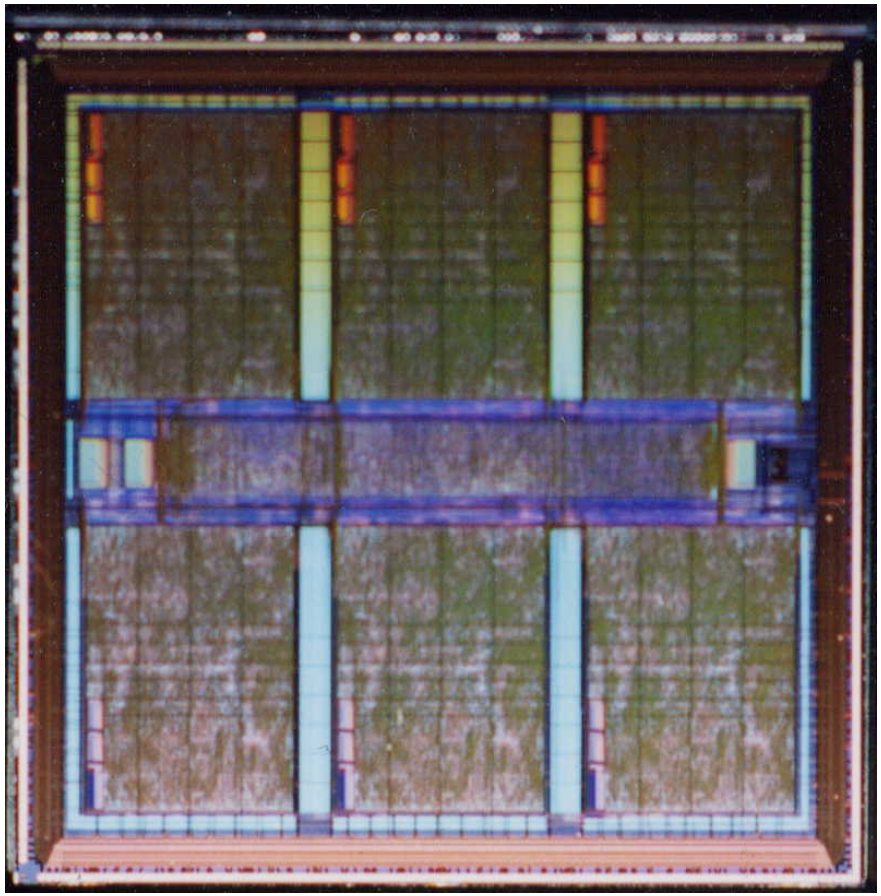
GRAPE の場合

- 実／仮想マルチパイプラインによるチップ外通信速度の削減
- ハードワイヤドパイプラインによる最小コストでの演算器間接続
- 同じくハードワイヤドのため内部メモリ不要

通常 on-chip multiprocessor で起きる問題をほぼ完全に回避している。
基本的にはそういうことが出来る問題だから。
逆にいえば、そういうことが出来る問題（アルゴリズム）でないと専用計算機はうまくいかない。

GRAPE-DR は本当に上手くいくか？

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