国立天文台における
大規模シミュレーション

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

概要

- 国立天文台の紹介
- 天文シミュレーションプロジェクトの紹介
- 計算機の歴史
- 2008/4 以前の構成
- 2008/4 からの構成
- 現在の運用体制
- ユーザー層の変化
- GRAPE との関係
- アプリケーション例
  - ダークマター構造形成シミュレーション
  - 銀河形成シミュレーション
- 今後の方向

国立天文台の紹介

- 日本の天文学のナショナルセンター
- 地上観測施設
  - すばる望遠鏡
  - 野辺山宇宙電波観測所
  - その他、岡山観測所等
- JAXA と共同で宇宙からの観測も
- 暦、理科年表等の歴史的な機能
- 理論研究・シミュレーション研究による天文学

国立天文台の研究施設

国立天文台の研究・観測施設は日本各地にとどまらず、すばる望遠鏡や建設中のALMA（アルマ）のように海外にも進出しています。天文学の観測では、可視光、赤外線、電波、重力波などの観測手段と、太陽とそれ以外の宇宙などの観測対象に応じて、最適の観測条件と環境とが必要とされるからです。

野辺山エリア

野辺山宇宙電波観測所
Nobeyama Radio Observatory

日本の電波天文学を世界のトップレベルに押し上げた観測施設です。写真の45m電波望遠鏡（右）は、ミリ波で世界最大の望遠鏡で、新たな星間分子の発見や原始惑星系の回転ガス円盤の発見など、数々の画期的な成果を挙げています。

野辺山太陽電波観測所
Nobeyama Solar Radio Observatory

太陽面爆発を高精度で観測する干渉計システム「電波ヘリオグラフ」（写真下）で、太陽活動のモニターを行っています。

乗鞍岳エリア

太陽観測所・乗鞍コロナ観測所
Solar Observatory:Norikura Solar Observatory

北アルプス・乗鞍山系摩利支天岳（海抜2876m）の頂上に位置する太陽観測所です。太陽物理現象の精密な観測を行うために3台のコロナグラフが設置されています。

岡山エリア

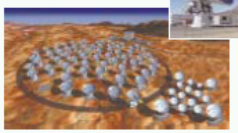
岡山天体物理観測所
Okayama Astrophysical Observatory

国内最大級の口径188cmの反射望遠鏡を中心に、銀河・恒星・太陽系天体などの光学赤外線観測研究の国内の拠点となっています。また、赤外線分光装置や赤外線超広視野カメラなど、宇宙を見る新しい目を次々と開発しています。

チリ・エリア

ALMA（アタカマ大型ミリ波サブミリ波干渉計）
ALMA (Atacama Large Millimeter/submillimeter Array) Project

ALMA（アルマ）は、日米欧が共同でチリの標高5000mの高原に建設中の巨大な電波望遠鏡群（イラスト）で、国立天文台が現在総力を挙げて取り組む大型プロジェクトです。



水沢エリア

水沢観測所
Mizusawa Astrogeodynamic Observatory

旧緯度観測所として長い歴史をもつ施設です。位置天文学・測地学の研究が進んで、日本の標準時を決める天文保時室などがあります。

江刺地球潮汐観測施設

レーザー光線を利用して地面の伸縮の変化を測るレーザー歪計です。潮汐による地球の微妙な変形をモニターします。

VERA観測所・水沢観測局
VERA Observatory：Mizusawa station

銀河系の3次元地図を作成するVERA観測局のひとつです。

三鷹エリア（本部）

三鷹キャンパス

三鷹キャンパスは、国立天文台の本部が置かれ、さまざまなプロジェクト、センター、研究部、事務部が集まっています。

VERA・4局
（水沢局を含む）

銀河系の3次元地図を作成します。

VERA観測所・入来観測局
VERA Observatory
：Iriki station

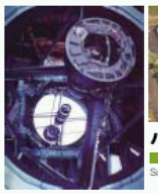
VERA観測所・石垣島観測局
VERA Observatory
：Ishigakijima station

VERA観測所・小笠原観測局
VERA Observatory
：Ogasawara station

銀河系の3次元地図を作成するVERA観測局のひとつです。

ハワイ・エリア

ハワイ観測所
Subaru Telescope



すばる望遠鏡

ハワイ島のマウナケア山頂（標高4200m）に設置された口径8.2mの世界最大級の可視・赤外線望遠鏡です。平成12年度から本格的な観測を始め、現在、世界最先端の研究成果を挙げつづけています。

ヒロ・オフィス（写真右上）

ハワイ島ヒロ市にあるハワイ観測所の本部です。「すばる望遠鏡」による観測研究の拠点となっています。

アメリカ合衆国
ハワイ州ハワイ島

すばる望遠鏡

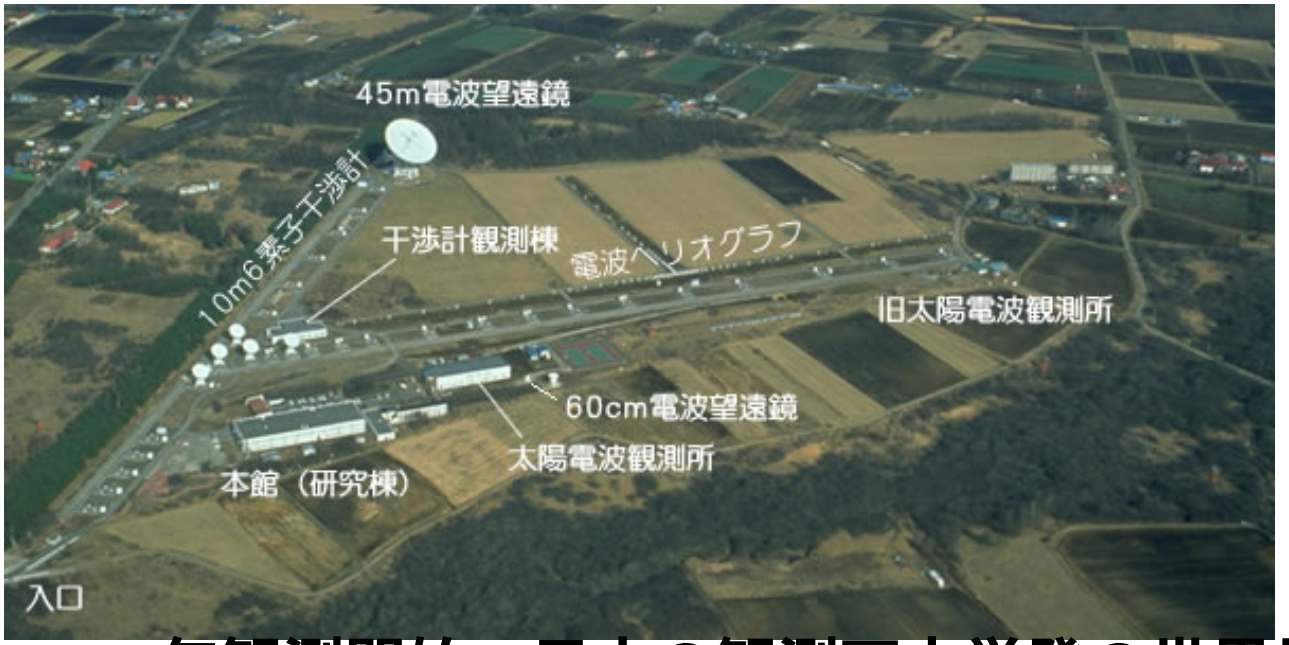
ハワイ・マウナケア山頂

主鏡直径 8.2m，世界最大級

大視野の主焦点カメラ (30分角)、ハッブル望遠鏡の100倍の視野

現在のところ、最遠方の銀河 (QSO) の10個のうち9個を発見

野辺山宇宙電波観測所



1982年観測開始。日本の観測天文学発の世界最先端装置
後継:日米欧共同プロジェクト ALMA 望遠鏡
(2009観測開始？)

その他主要プロジェクト

- ひので衛星
- 重力波観測装置 TAMA300
- VERA VLBI(超長基線電波干渉計) による銀河の構造・運動の観測
- VSOP-2 衛星による VLBI

天文シミュレーションプロジェクト

2 つの役割:

- 国立天文台の中の、理論・シミュレーション天文学研究グループ
- 国内外の天文学研究者のための計算機センター

歴史

1965年　人工衛星国内計算施設 発足

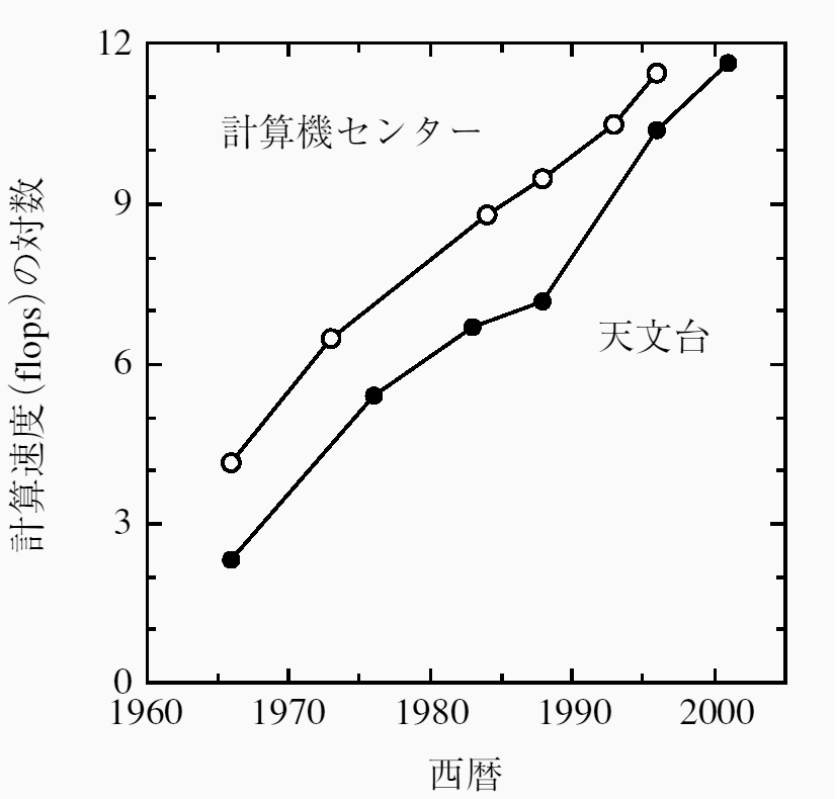
1988年　天文学データ解析計算センター 発足

同年　理論天文学研究系 発足

2006年　天文学データ解析計算センターを天文シミュレーションプロジェクトと天文データセンターに分離

「天文学理論・シミュレーションのための計算センター」として世界的にもユニークな位置を占める。

計算速度の進化



計算速度の対数

$(12 = 1\ \mathrm{Tflops})$

白◯: 東大センター

黒●: 国立天文台

大体東大の $1/10$ — $1/100$

OKITAC, UNIVAC **の他は**

富士通

VPP300/16R

1996 年運用開始
国立天文台最初のベクトル並列機 (1988 くらいに野辺山に VP-200)
32 Gflops

2001-2007 のシステム

- VPP5000/60 600 Gflops
- GRAPE-5 (+GRAPE-6 (+GRAPE-7)) $>$ 10 Tflops
- WS/PC Cluster (Opteron 250 20CPUs, 2004) 96Gflops

VPP5000/60

60 ノード、600Gflops, 960GB メモリ, 12 TB ディスク, 60TB テープ

並列環境 t: MPI, VPP-Fortran

Up to 48 nodes/job

GRAPE hardwares

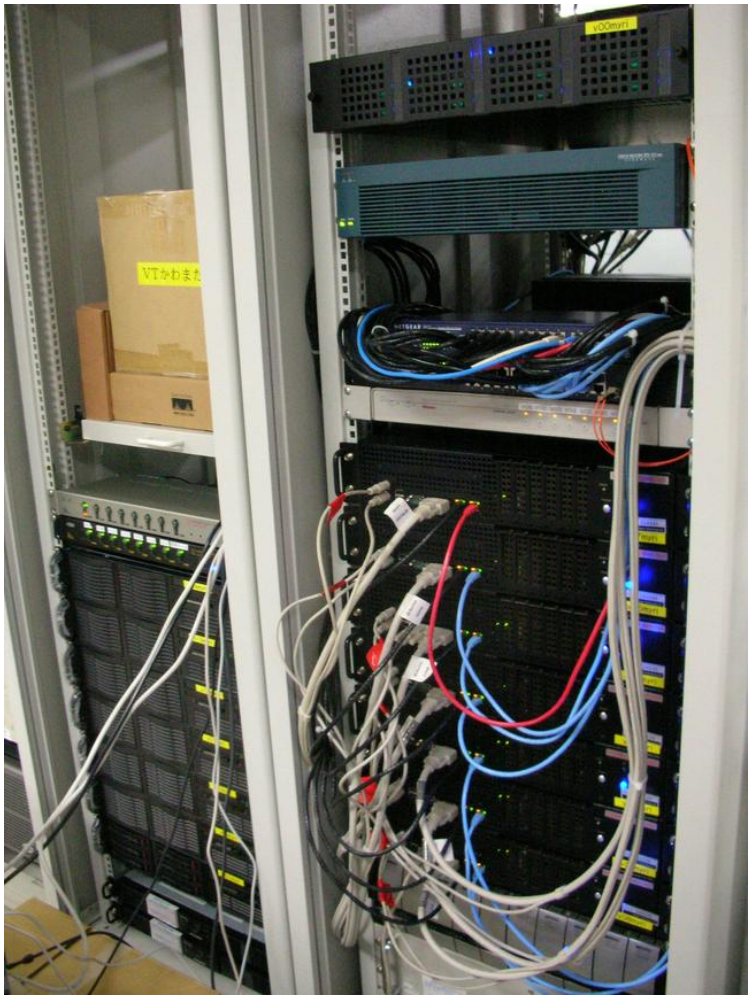
GRAPE-5 16 nodes (640Gflops peak)

GRAPE-6 8 nodes (8Tflops peak)

GRAPE-7 16 nodes (10Tflops peak)

No parallel job queue (yet)

PC Cluster



10-node, Dual-Opteron cluster (2004∼)

9.6Gflops/node

単一CPU での長時間計算に利用。

古典的「計算機センター」機能

結構利用者は多い

資源配分

- プロポーザルベース (VPP, GRAPE)
- 年2回募集 (大規模ユーザー: A, B カテゴリ)
- 小規模ユーザーは随時応募可能
- 課金等ない (大学センターとの大きな違い)

VPP users (2007)

Category A
Masahiro Machida
M. Shibata
M. Noguchi
K. Sumiyoshi
H. Yahagi
T. Inoue
T. Matsumoto
S. Hirose
K. Sugimoto
T. Sano
N. Ishitsu
Mami Machida
Y. Sekiguchi
H. Isobe
Y. Uryu

Category B
M. Takizawa
T. Kudo
Y. Aikawa
K. Nakazato
D. Shiota
T. Kato
M. Hayashi
T.Tsuribe
Y. Masada
C. H. Baek
S. Suzuki
N. Asai
K. Tomisaka
E. Asano
M. Tanaka
H. Hanayama
K. Nishida

GRAPE users (2007)

Category A

Y. Funato
M. Chiba
H. Daisaka
S. Inoue
A. Tanikawa
C. Kobayashi
H. Matsui
T. Ishiyama
J. Baba
T. Saitoh
M. Fujii
M. Iwasawa
T. Muranushi

Category B

J. Daisaka
E. Ardi
T. Tatekawa
S. Ida
T. Takeda
T. Wada
M. Ogihara
O. Iguchi
H. Genda
S. An
T. Tsuribe

現行システム

- 2006 年度前半に主な要求仕様を決定
- 2007 年 6 月入札公告
- 2007 年 9 月開札
- 2008/4/1 から試験運用中

要求仕様の概要

- 「ベクトル型並列計算機」 1.6Tflops 以上
  - HPF かなにかで 256GB 以上のメモリを使えるプログラムが走ること。
  - メモリバンド幅 (STREAM 実測)2TB 以上
- 「スカラー型並列計算機」 20Tflops 以上
  - ノード当りメモリバンド幅、ノード間通信バンド幅、レイテンシ、バイセクションバンド幅等色々規定
- ディスク容量、I/O 速度はそこそこ
- 消費電力上限 300KVA (国立天文台三鷹キャンパスの受電施設の制約、、、)

要求仕様の精神

- ベクトル型
  - 従来の vpp ユーザーのスムーズな移行
  - 若干の性能向上
  - FFT 等高速な大域通信が必要なアプリケーション
- スカラ型
  - 大規模計算
  - 地球シミュレータその他で既に MPI 並列化されたアプリケーション
  - 価格当りの性能高いこと
  - スケーラビリティ良いこと

ベクトル並列計算機

- 別にベクトルでなくてもよい
- 別に物理共有メモリでなくてもよい
- そんなにメモリサイズ要求は厳しくない
- メモリバンド幅要求はきつい

NEC SX の他、 SGI Altix, IBM Power 富士通 PQ とかでも物量があれば、、、

スカラ並列計算機

- 納入する機械と同程度のコア数でのベンチマークを要求
  - 世界最大級の機械を、という調達ではない
  - マシンの半分程度以上を単一ジョブで使うことがありえる
- ベンチマークは VPP, ES 上で開発された流体コード主体
  - どちらかというと x86 マシンには不利 (メモリバンド幅要求が高い)

落札したシステム

- **ベクトル部分:** NEC SX-9 16CPU+4CPU
- **スカラー部分:** Cray XT4 9キャビネット (812演算ノード、28.6TF)

国立天文台の Cray XT4

フロントパネル CG とデータ提供者

Cray になった理由

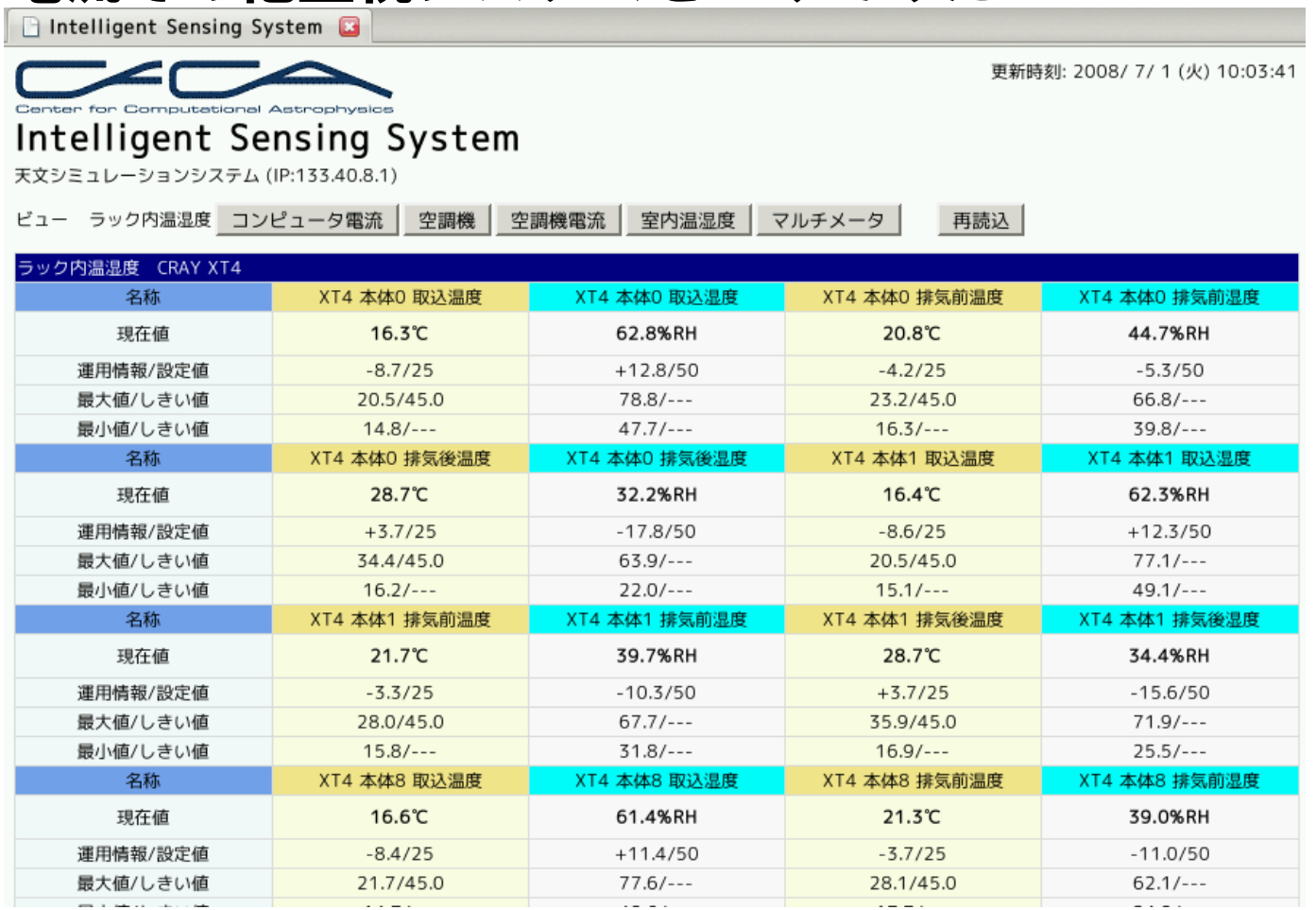
- 価格
- 消費電力
- 数千コアまでで実際に性能がでていること

現在の運用体制

- 9 月までは「試験運用」
- ジョブキューの設定、スケジューラの設定等を調整
- ファイルシステムの運用形態も色々試行
- 超大規模計算の、マシン安定性も兼ねた実行

空調

電流その他監視システムをつけてみた



ラック内温湿度 CRAY XT4

名称	XT4 本体0 取込温度	XT4 本体0 取込湿度	XT4 本体0 排気前温度	XT4 本体0 排気前湿度
現在値	16.3℃	62.8%RH	20.8℃	44.7%RH
運用情報/設定値	-8.7/25	+12.8/50	-4.2/25	-5.3/50
最大値/しきい値	20.5/45.0	78.8/---	23.2/45.0	66.8/---
最小値/しきい値	14.8/---	47.7/---	16.3/---	39.8/---
名称	XT4 本体0 排気後温度	XT4 本体0 排気後湿度	XT4 本体1 取込温度	XT4 本体1 取込湿度
現在値	28.7℃	32.2%RH	16.4℃	62.3%RH
運用情報/設定値	+3.7/25	-17.8/50	-8.6/25	+12.3/50
最大値/しきい値	34.4/45.0	63.9/---	20.5/45.0	77.1/---
最小値/しきい値	16.2/---	22.0/---	15.1/---	49.1/---
名称	XT4 本体1 排気前温度	XT4 本体1 排気前湿度	XT4 本体1 排気後温度	XT4 本体1 排気後湿度
現在値	21.7℃	39.7%RH	28.7℃	34.4%RH
運用情報/設定値	-3.3/25	-10.3/50	+3.7/25	-15.6/50
最大値/しきい値	28.0/45.0	67.7/---	35.9/45.0	71.9/---
最小値/しきい値	15.8/---	31.8/---	16.9/---	25.5/---
名称	XT4 本体8 取込温度	XT4 本体8 取込湿度	XT4 本体8 排気前温度	XT4 本体8 排気前湿度
現在値	16.6℃	61.4%RH	21.3℃	39.0%RH
運用情報/設定値	-8.4/25	+11.4/50	-3.7/25	-11.0/50
最大値/しきい値	21.7/45.0	77.6/---	28.1/45.0	62.1/---

運用の感触

- 想定していたよりもずっと安定している
- 全体が落ちたのは 1 度だけ (ソフトウェア障害、対応済)
- CPU の障害はちょっとまだあり (B2 コアだし、、、先週末に対策)
- 6 月からはほぼ 100% 稼働
- ユーザーの評判は大変良い
  - PC クラスタに比べてコア数ずっと高いところまで性能でる
  - I/O が速い

ユーザー層の変化

(試験期間での、予備的なもの。今後変化する可能性大)

- 従来の VPP ユーザーの移行はまだ進んでいない
- これまで PC クラスタ、筑波 PACS 等スカラ並列機を使っていた研究者が参入
- GRAPE ユーザも利用
- ユーザー若返り

GRAPE との関係

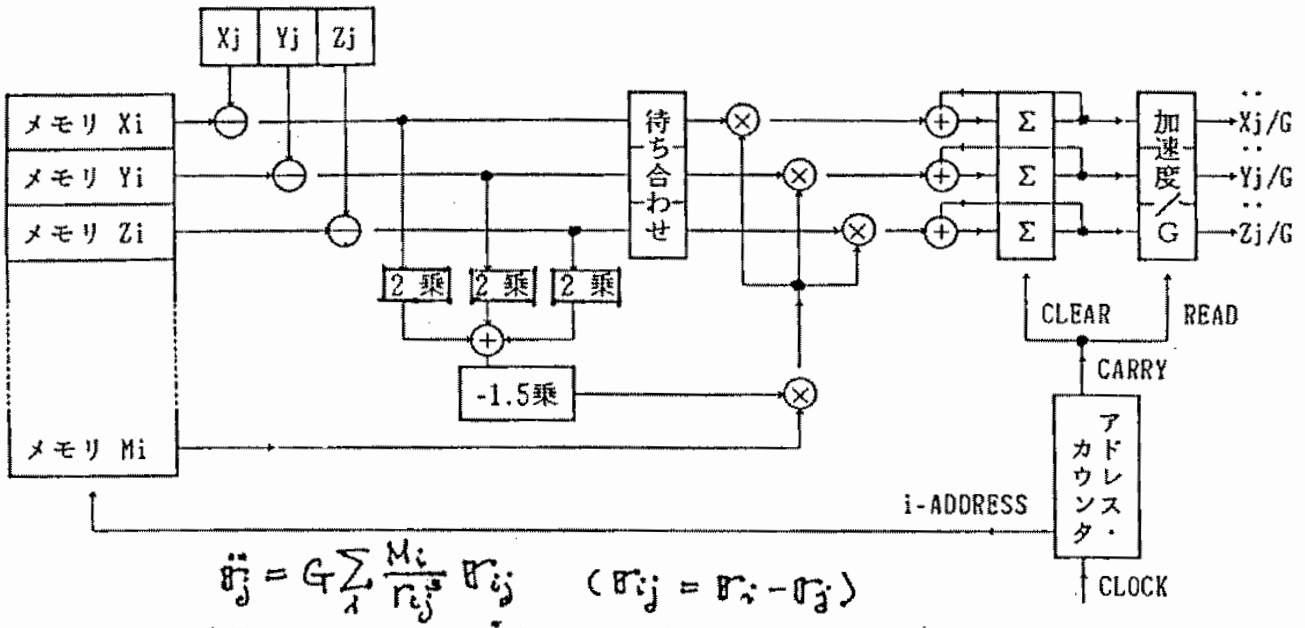
- GRAPE 用で MPI 並列化されたコードは XT4 で性能・スケーラビリティ非常に良い
  - あまりメモリバンド幅を必要としないアプリケーション
  - GRAPE 向けチューニングは並列計算機一般に有効
- 現行の GRAPE-6/7 に比べて XT4 の性能は高い
- GRAPE-DR の運用が開始すればそっちに移動？

GRAPE の考え方

- 重力多体問題: 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

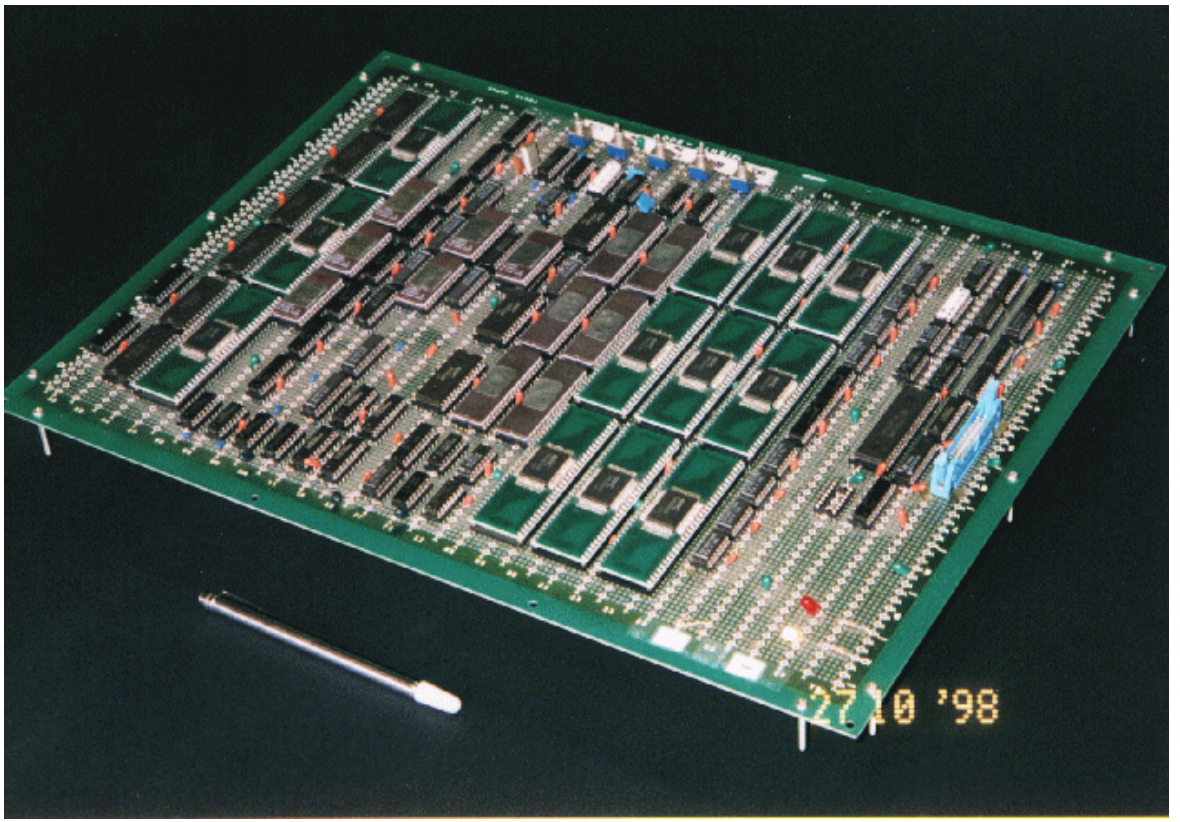


$$\ddot{\vec{r}}_j = G\sum_i \frac{M_i}{r_{ij}^3}\vec{r}_{ij} \qquad (\vec{r}_{ij} = \vec{r}_i - \vec{r}_j)$$

+，－，×，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．

GRAPE-1(1989)



演算毎に語長指定。固定 16–対数 8–固定 32–固定 48
240Mflops 相当

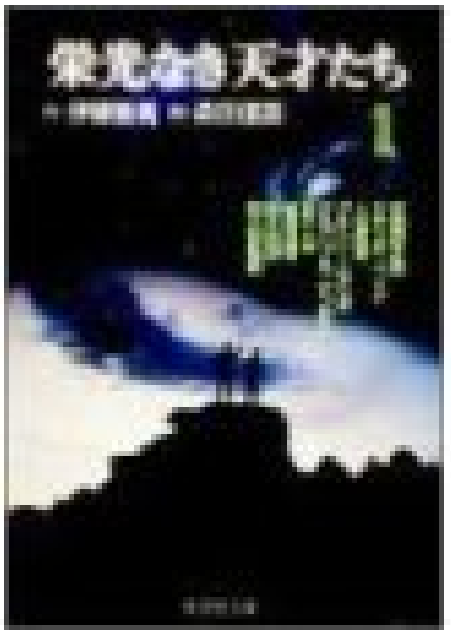
開発はどんなふうだったか

ハードウェアは私が知らないうちにできていた (伊藤君が作った) ので良く知らない。

ホストの GPIB インターフェースの性能がでない (特に Sony NEWS を使った時に) のでなかなか大変だった。

最終的にはユーザープロセスから GPIB 制御チップをいじり回すプログラムを書いた。

伊藤君

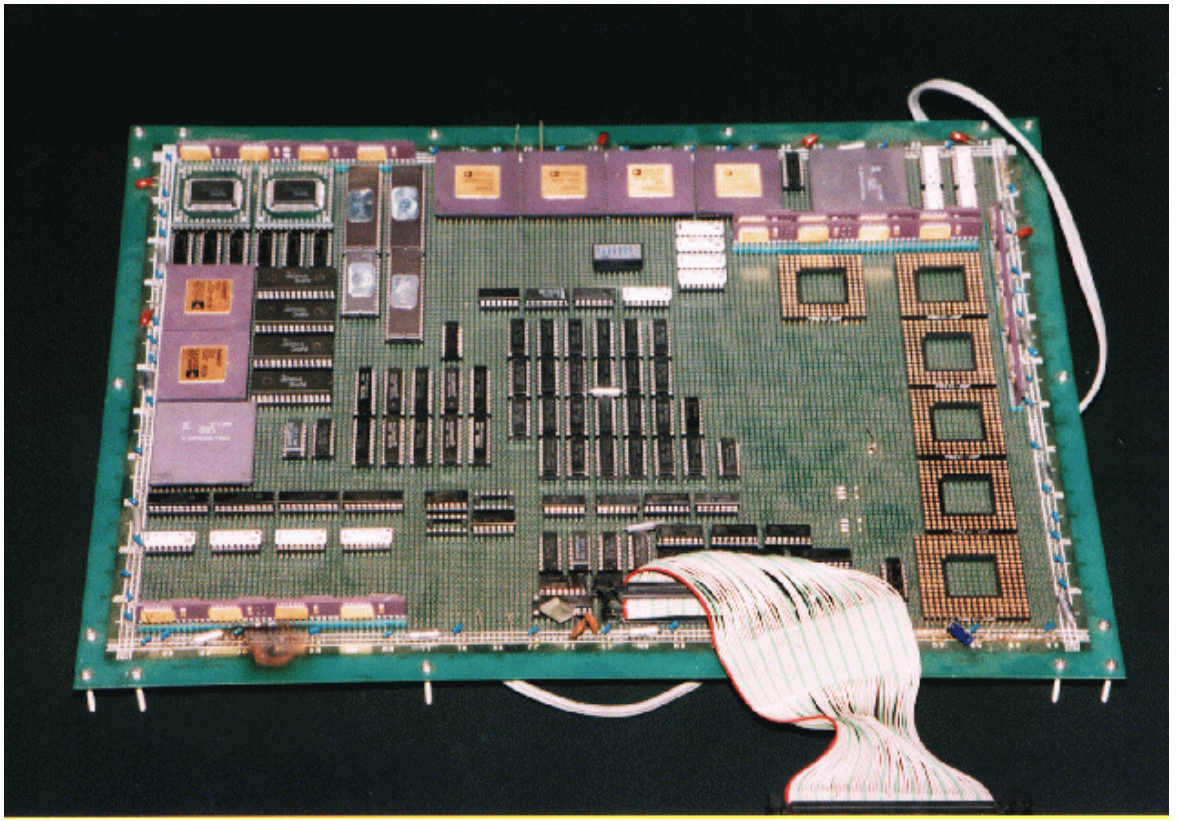


「栄光なき天才たち」の原作者としてのほうが有名という話も。

学部の頃から GRAPE-2 の開発の後くらいまで原作者をしていた。

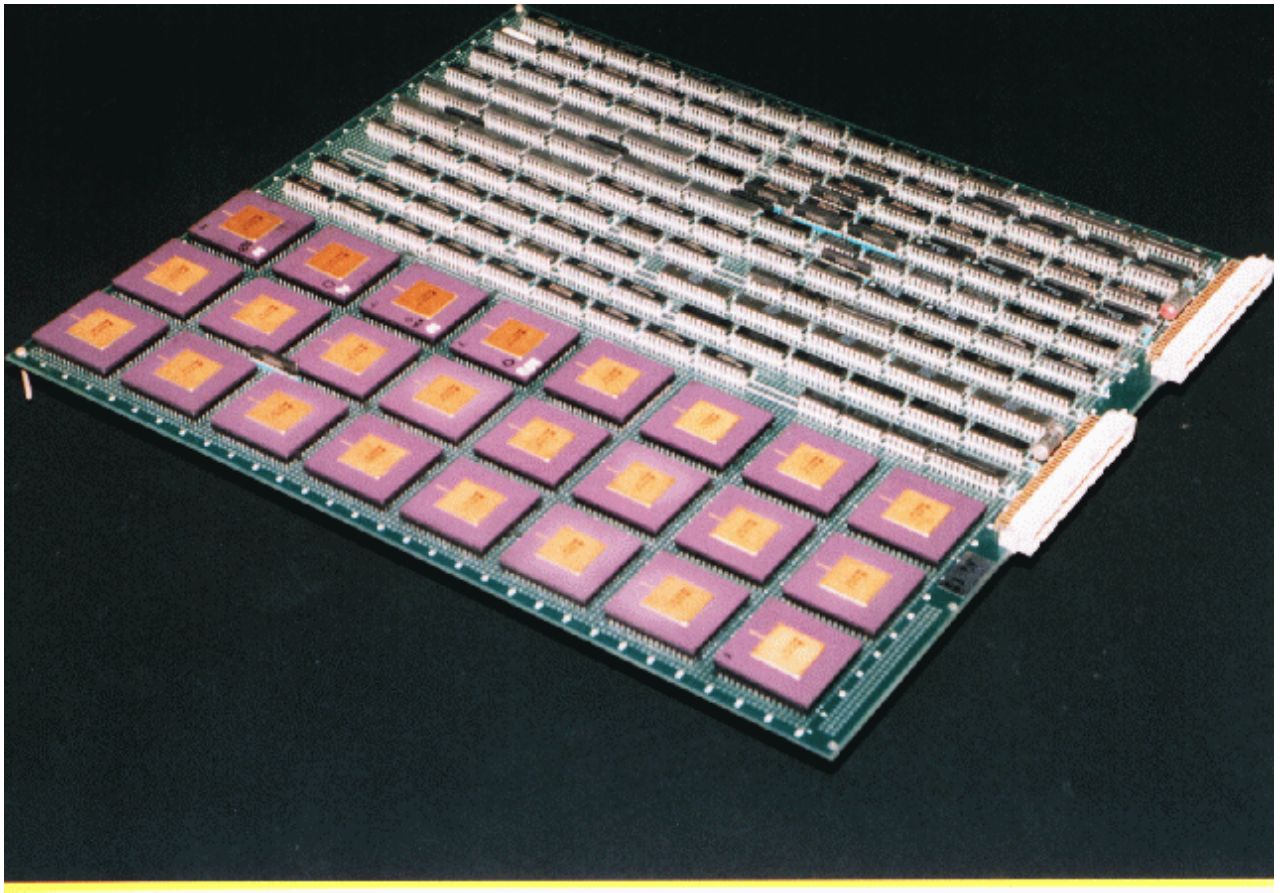
現在千葉大教授

GRAPE-2(1990)



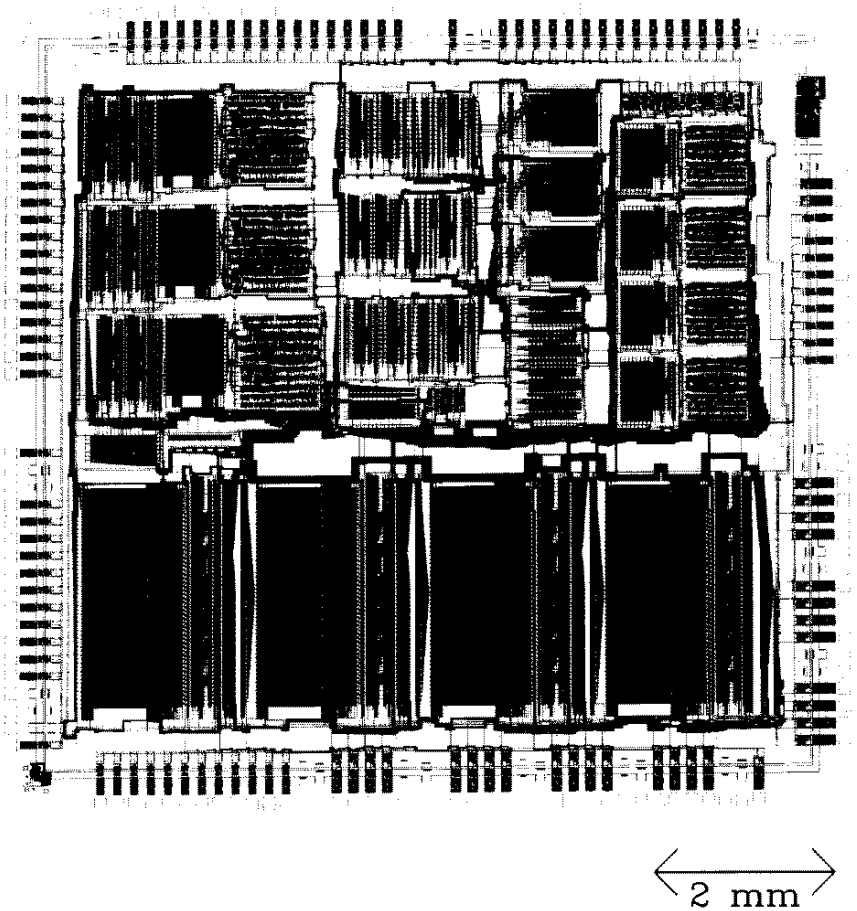
8ビット演算とかは止めて普通に浮動小数点演算 (倍精度は最初と最後だけ)
40Mflops

GRAPE-3(1991)



カスタムチップ 24個1ボード、 10MHz 動作、 7.2Gflops

GRAPE-3 チップ



1μm プロセス

11万トランジスタ

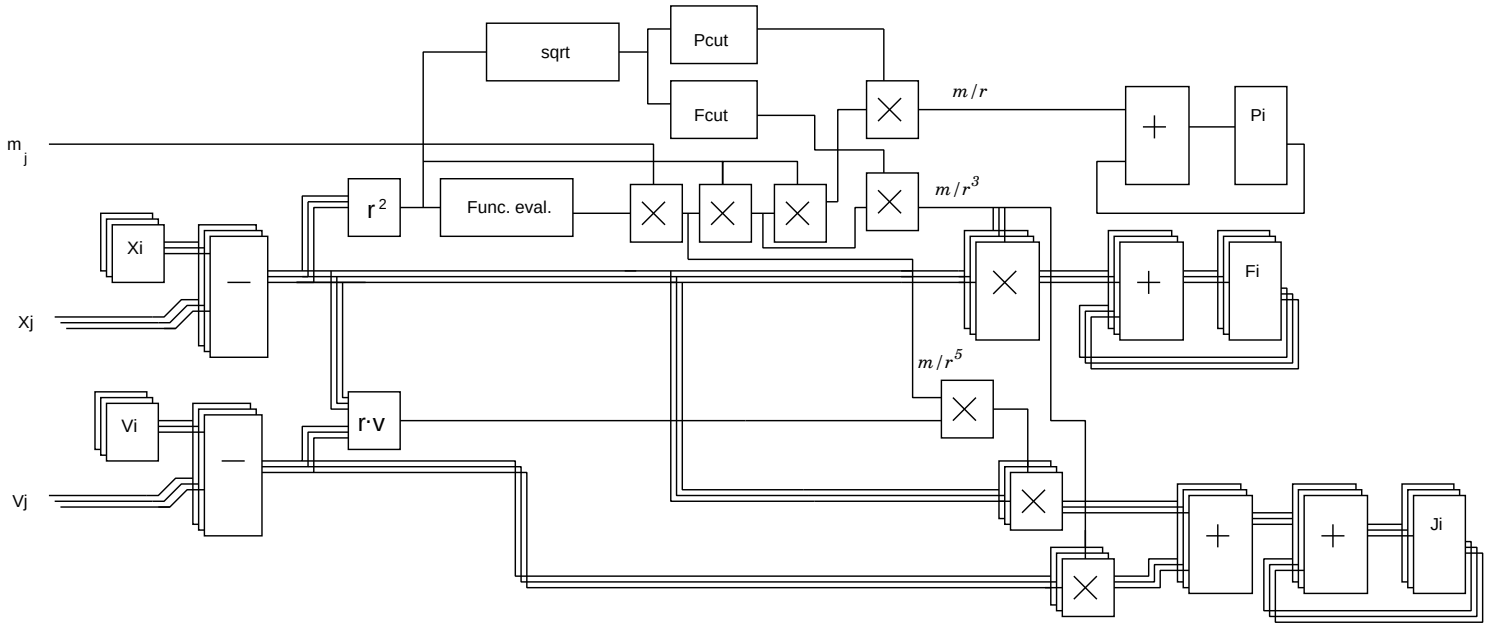
20 MHz クロック動作

600 Mflops 相当

GRAPE-4(1995)

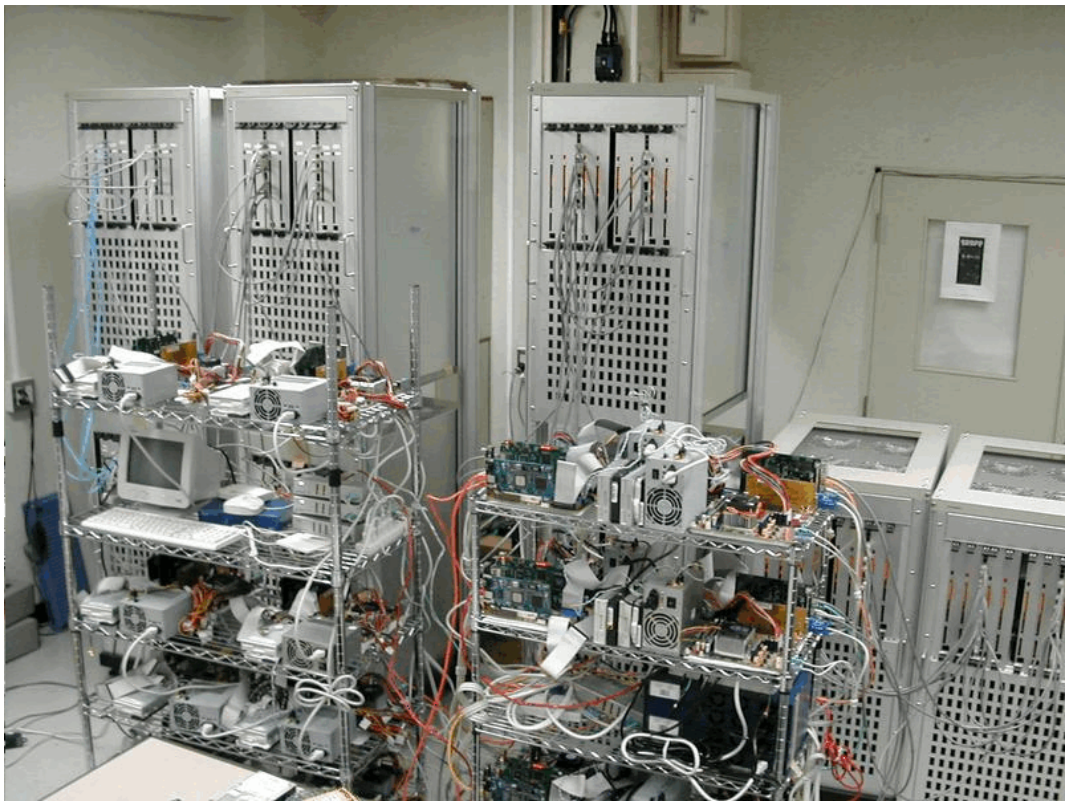
トータル 1792 チップ、 1.1 Tflops

GRAPE-4 パイプライン



1μm プロセス、10万ゲート(40万トランジスタ)、640Mflops

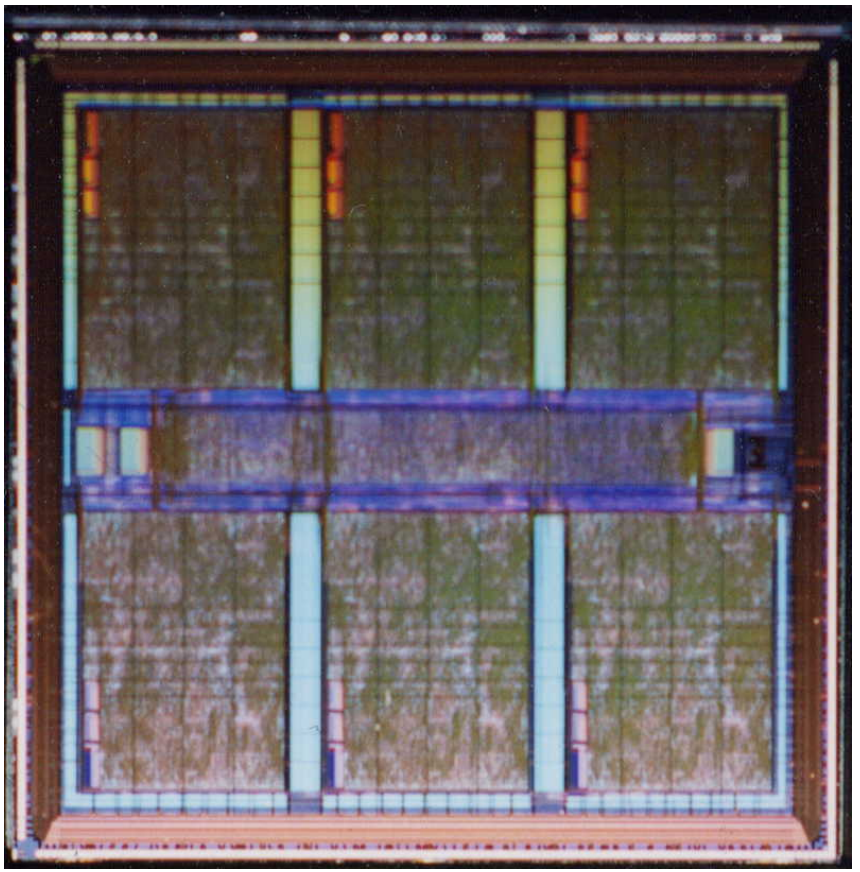
GRAPE-6(2002)



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

パイプライン LSI

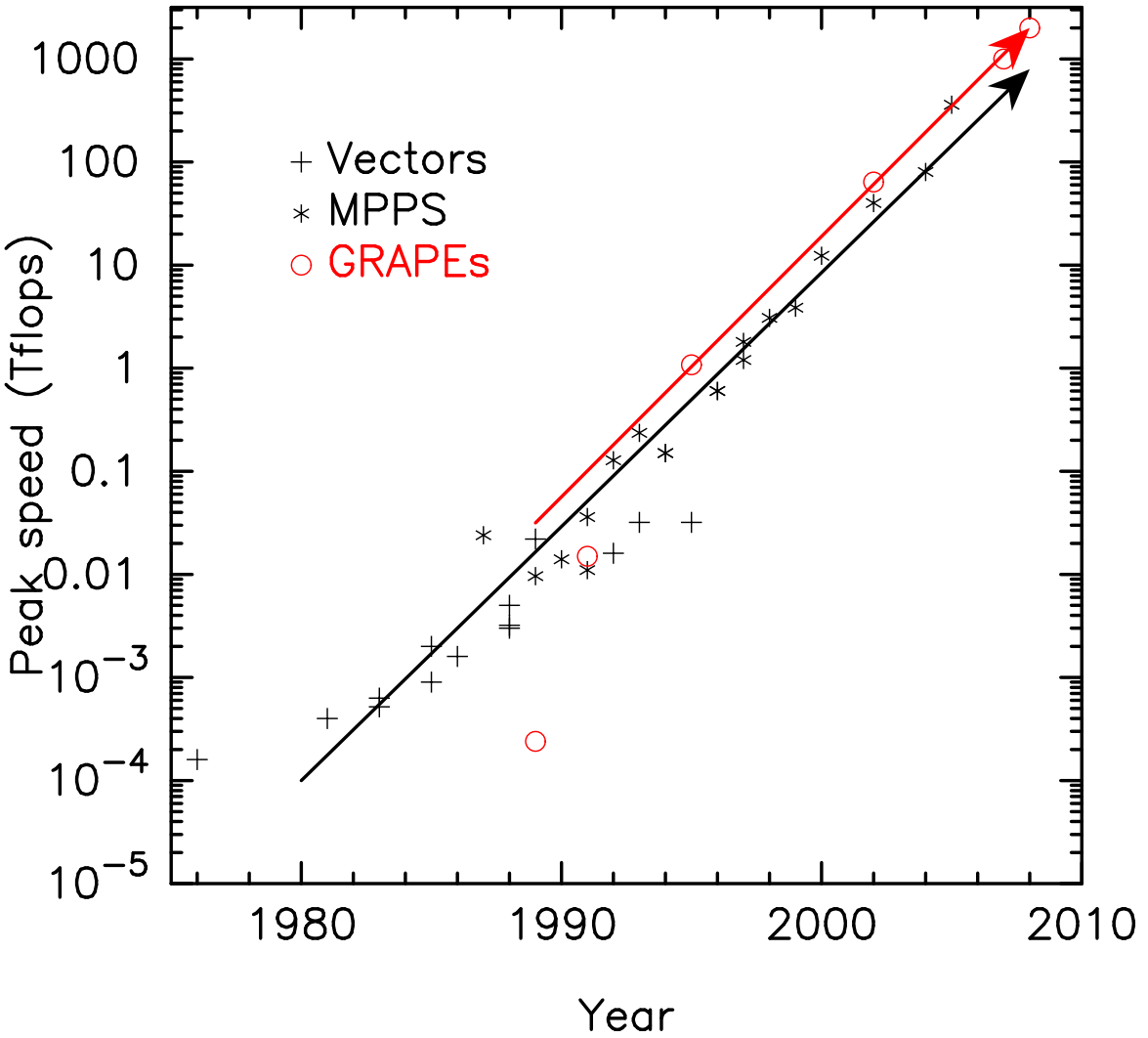


- 0.25 μm ルール
  (東芝　TC-240，　1.8M ゲート)
- 90 MHz 動作
- 6 パイプラインを集積
- チップあたり 31 Gflops

現在のマイクロプロセッサと比べてみる

	GRAPE-6	Intel Xeon 5365	IBM BG/P
Year	1999	2006	2007
Design rule	250nm	65nm	90nm
Clock	90MHz	3GHz	850MHz
Peak speed	32.4GF	48GF	13.6GF
Power	10W	120 W	15W?
Perf/W	3.24GF/W	0.4 GF/W	0.91GF/W

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-DR

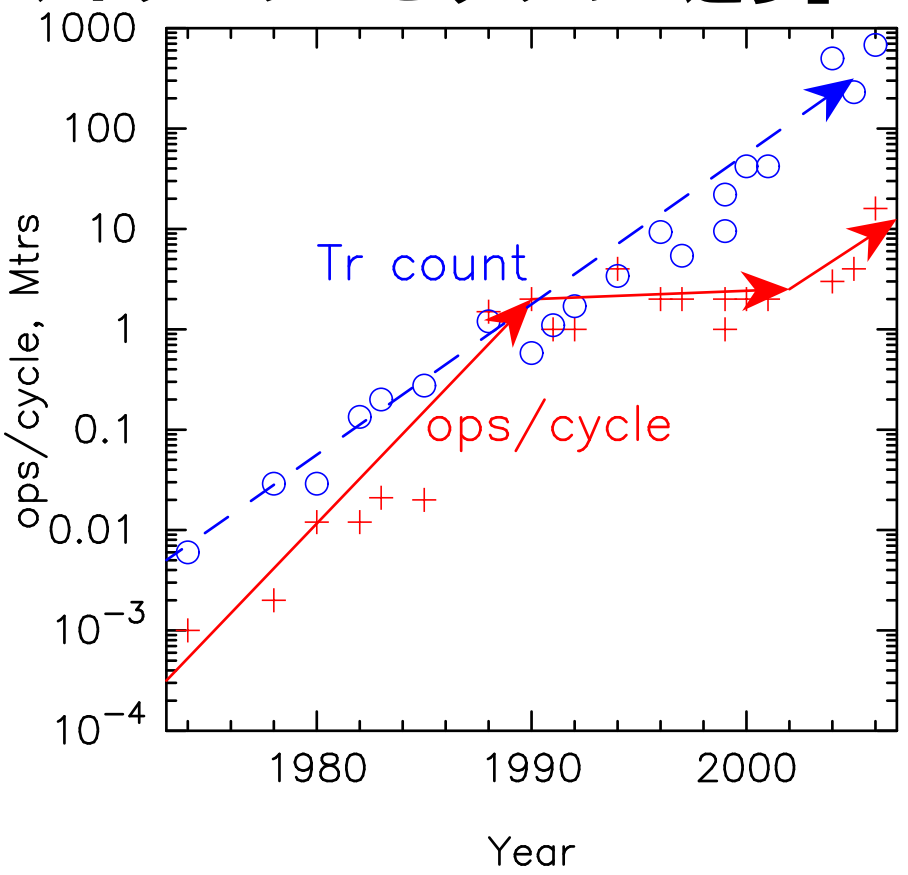
- カスタム LSI による専用パイプライン:開発コストが高騰

  → 応用範囲を広げる必要あり
- 汎用マイクロプロセッサは効率が低下中

  → 専用パイプラインでなくても効率上げられる
- 2008 年度で、 1Pflops/15 億円

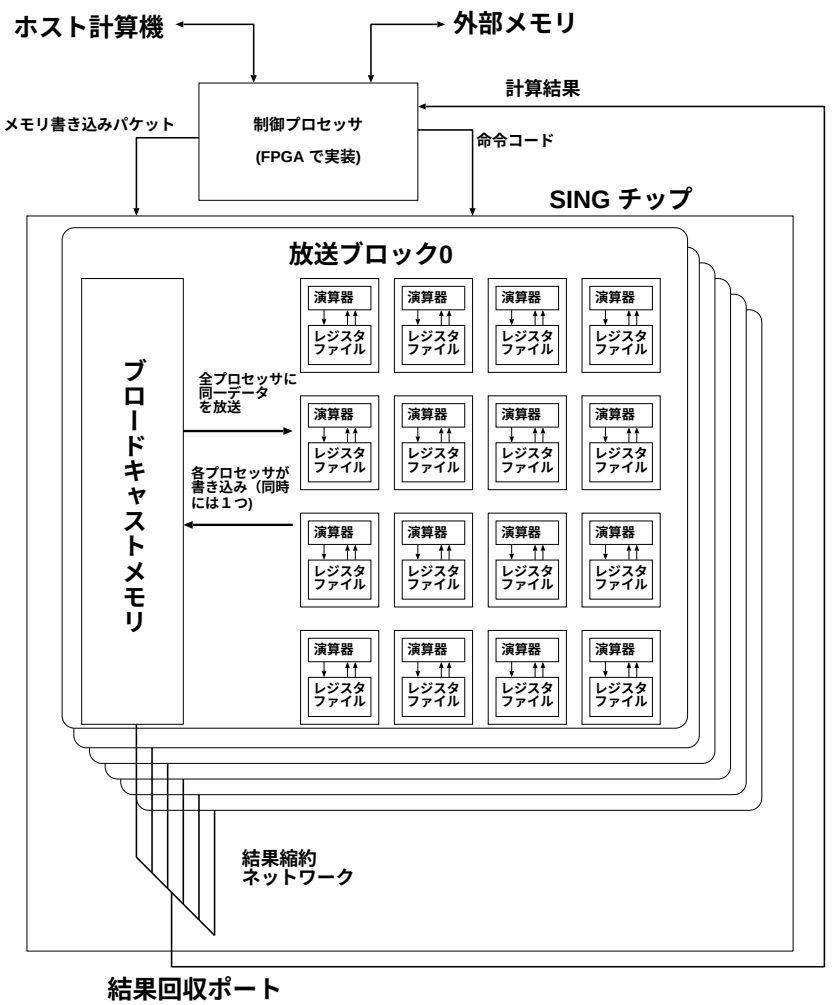
トランジスタは沢山ある

マイクロプロセッサの「進歩」



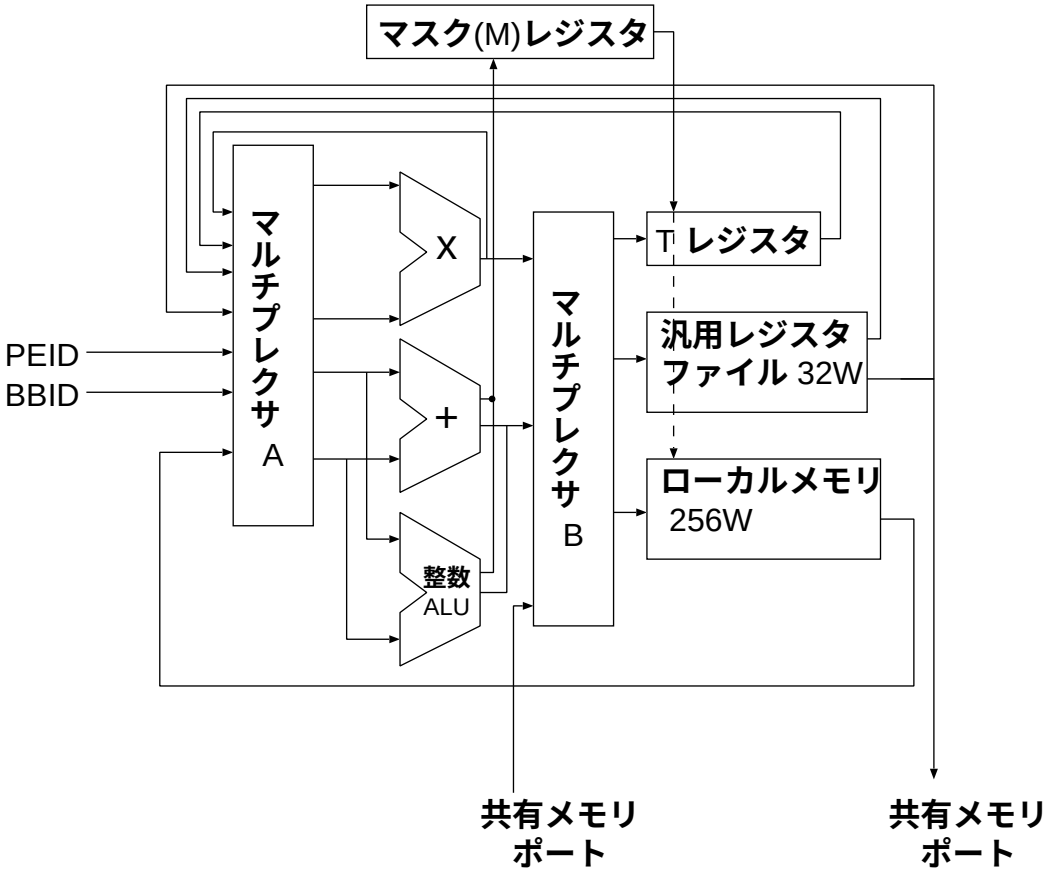
- トランジスタ: 15 年で 1000 倍
- 演算器の数: 同じ期間に 8 倍
- 100 倍分がどこかに消えた
- 最も良かった時でもチップ上の演算器の割合は 5% くらい

GRAPE-DR の構成



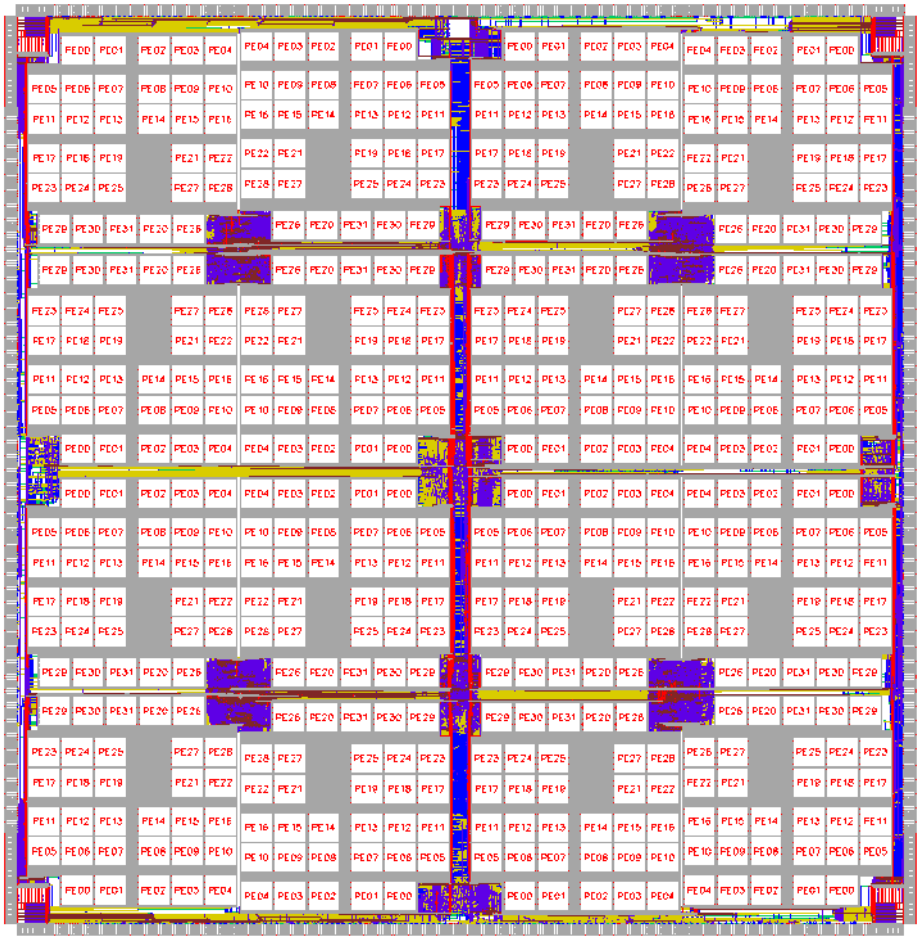
- 非常に多数のプロセッサエレメント (PE) を 1 チップに集積
- PE = 演算器 + レジスタファイル (メモリをもたない)
- チップ内に小規模な共有メモリ (PE にデータをブロードキャスト)。これを共有する PE をブロードキャストブロック (BB) と呼ぶ。
- 制御プロセッサ、外部メモリへのインターフェースを持つ

PE の構造



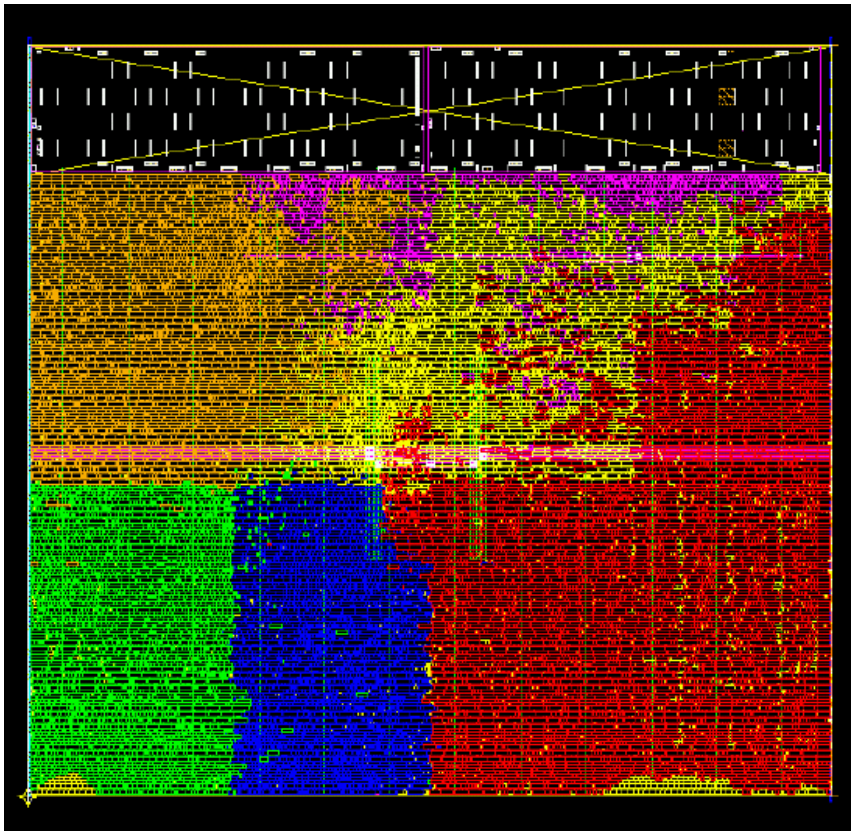
- 浮動小数点演算器
- 整数演算器
- レジスタ
- メモリ 256 語 (K とか M ではない。)

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

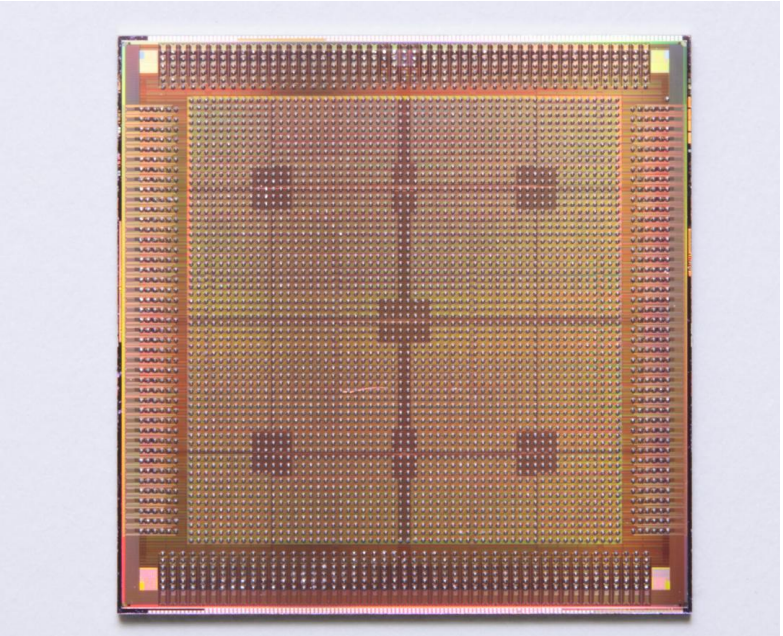
Red: Reg. File

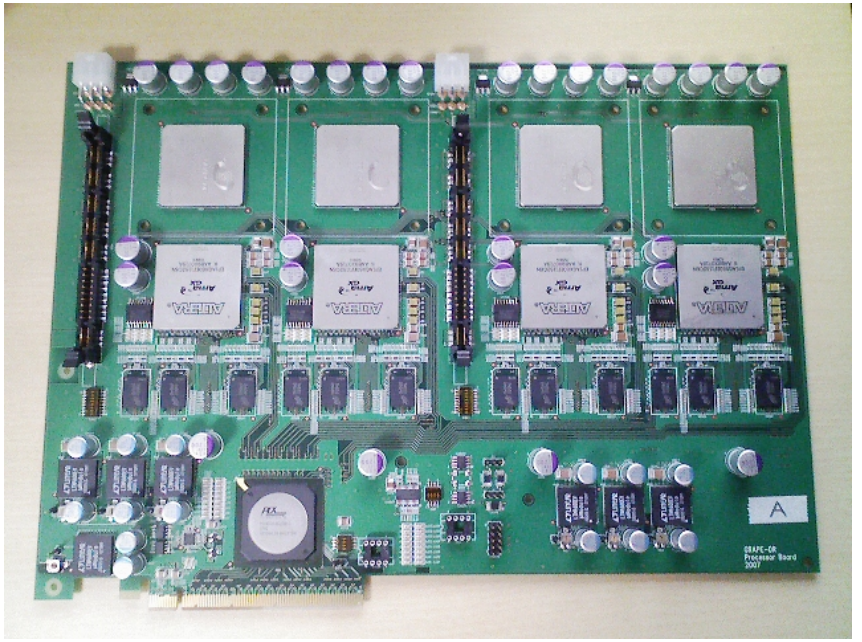
Orange: FMUL

Green: FADD

Blue: IALU

プロセッサチップとボード

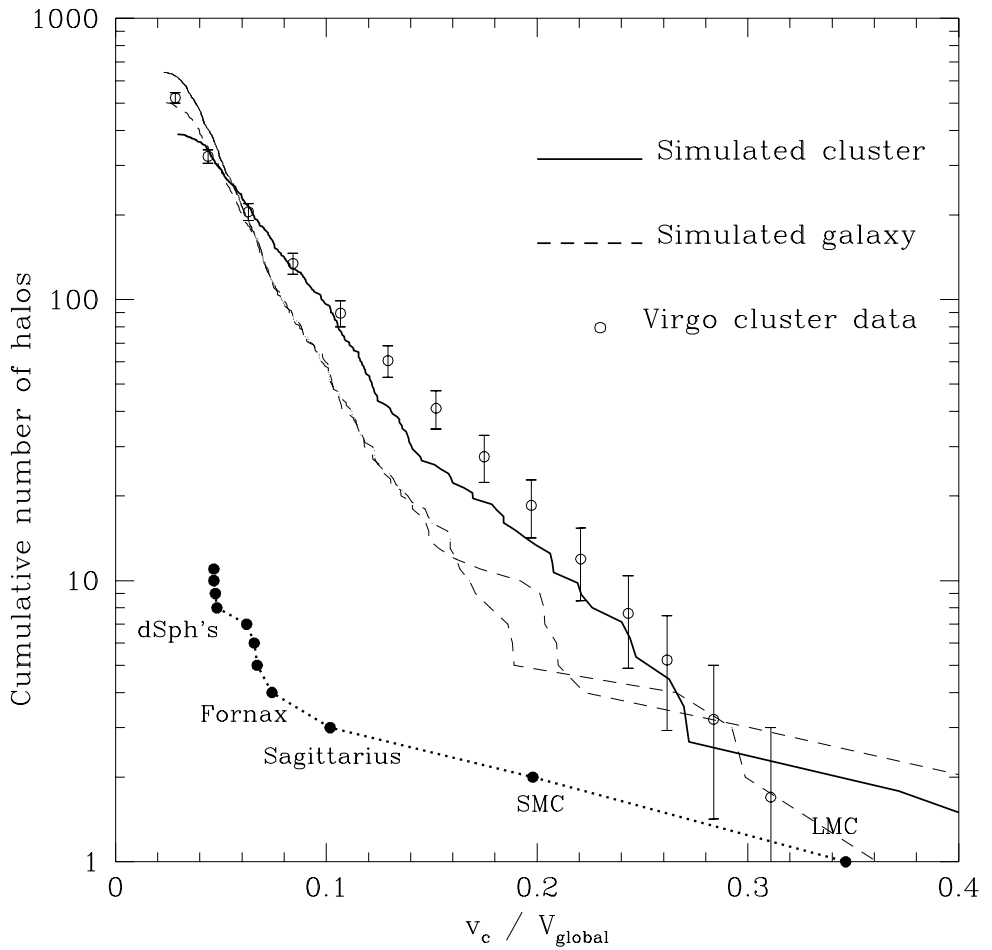




- PCI-Express カード (16 レーン、通信速度 2GB/s)
- 4 GRAPE-DR チップ
- 理論ピークスピード 1TP(DP), 2TF(SP)

アプリケーション例:大規模構造形成

考えた問題



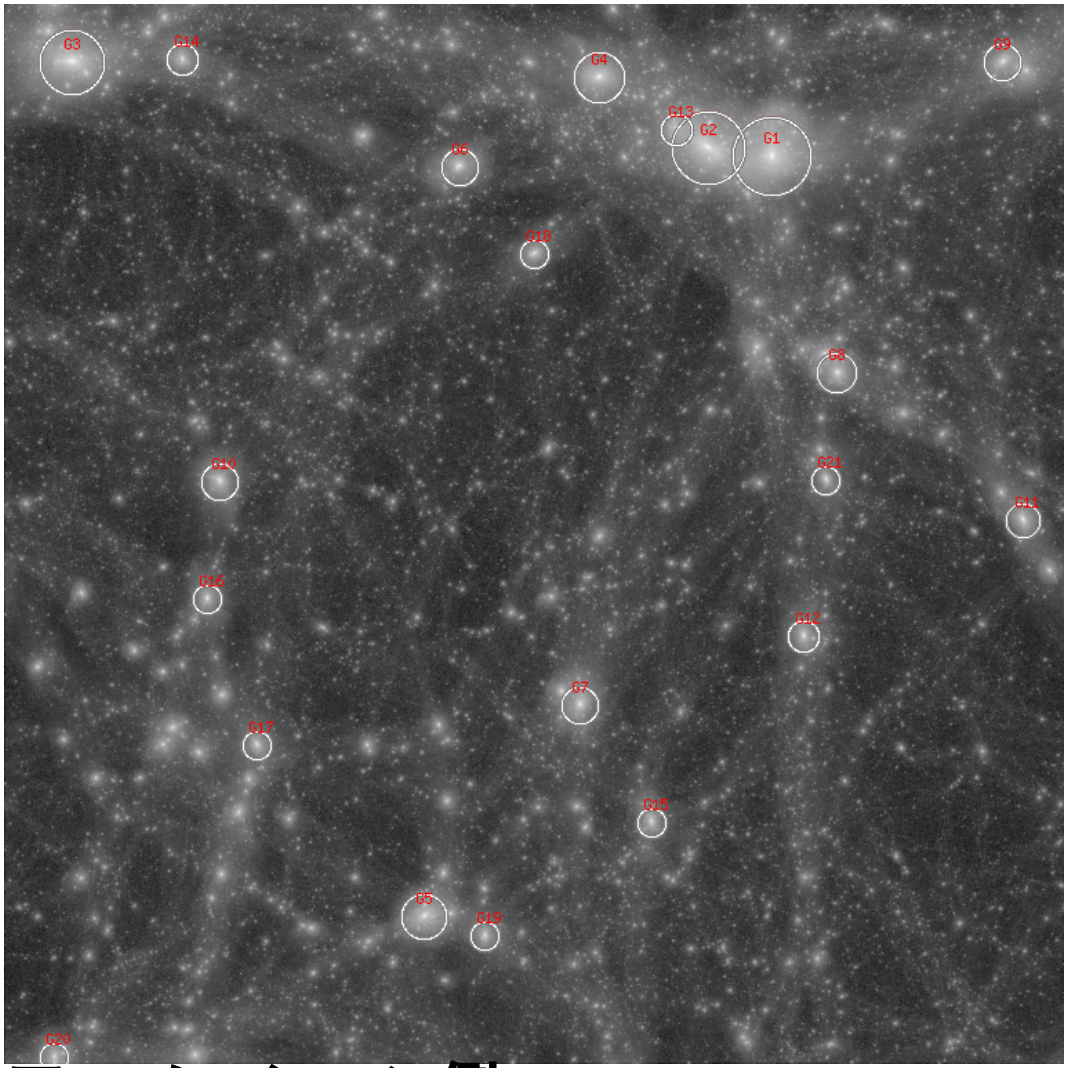
Moore et al 1999

- 銀河サイズの暗黒物質ハローには、小さいハローができすぎる。
- そんな多数の矮小銀河は見つかってない
- 何故？

我々の計算

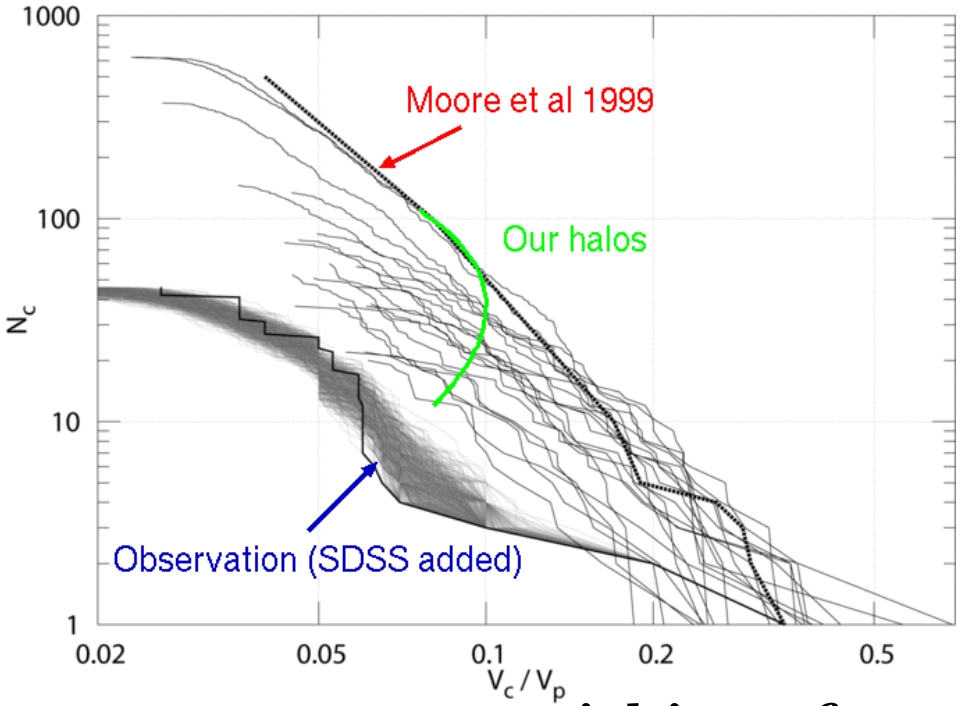
- ある領域での全部のハローを、無バイアスで「観測」
- GRAPE-6A クラスタ/PC クラスタ with IB/XT4
- 512^3 particles — 2048^3

512^3 計算結果



アニメーション例

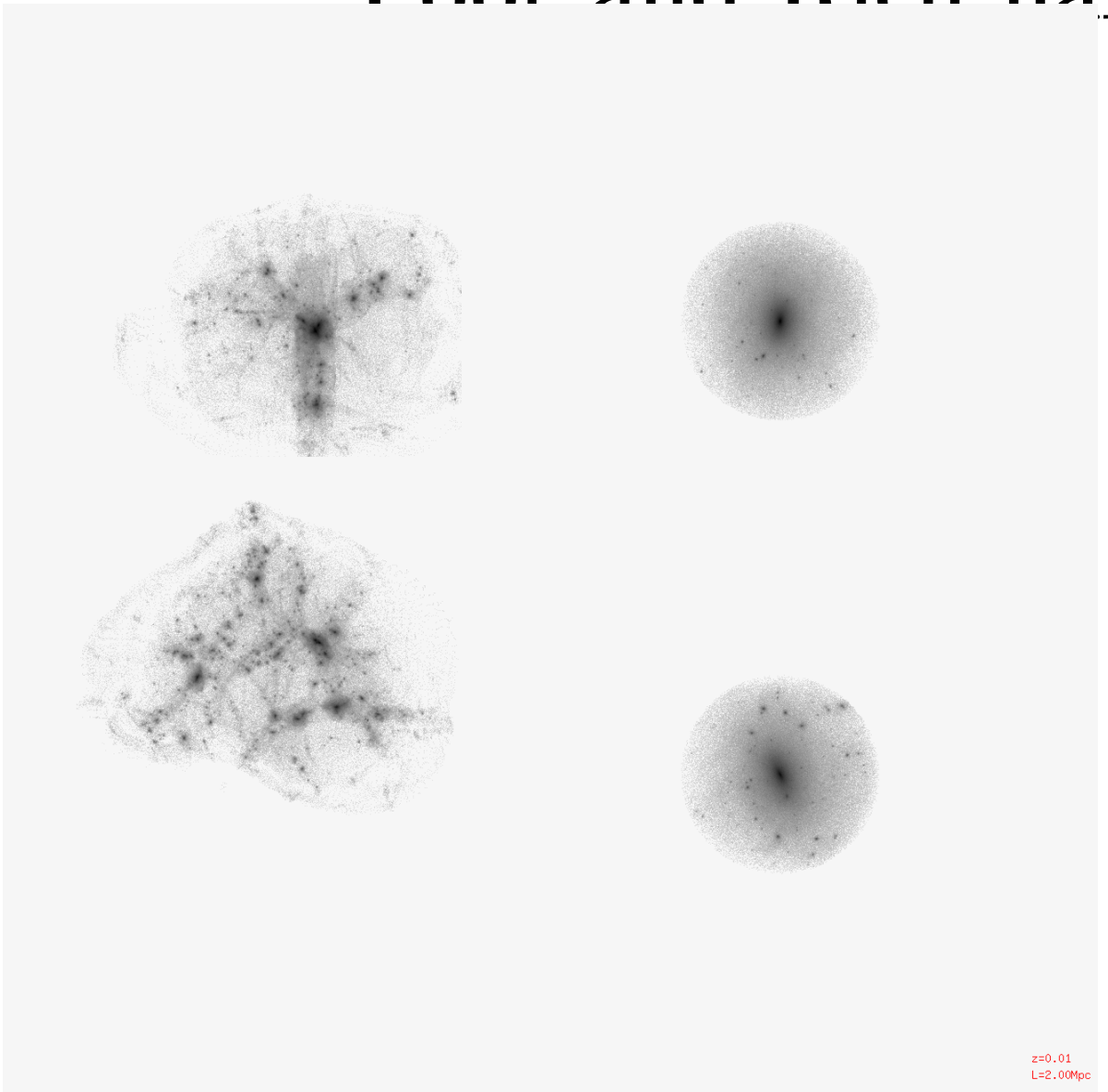
Result



- Large variation in number of subhalos
- The richest ones agree with Moore's result The poorest ones are within a factor of two with observations

= Dark CDM subhalos are not necessary

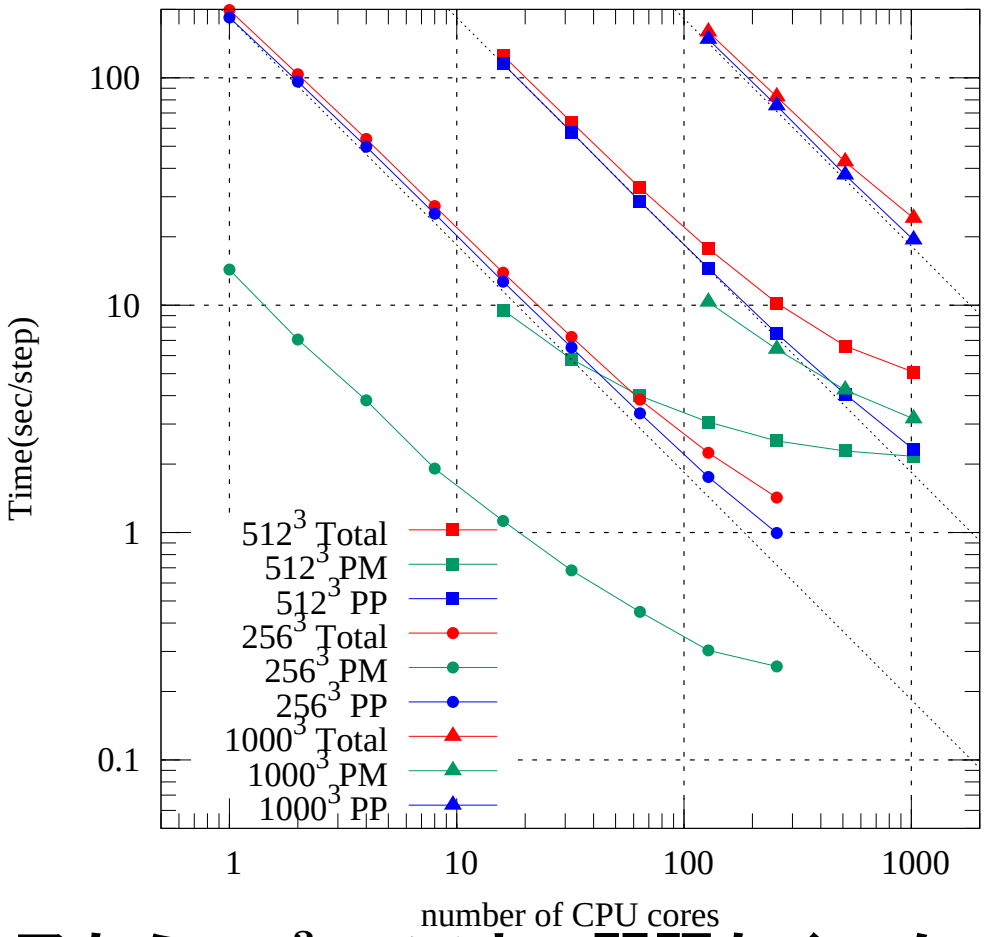
Poor and Rich halos



A poor halo at z=3 (left) and 0 (right)

A rich halo at z=3 (left) and 0 (right)

性能



10^9 粒子なら 10^3 コアまで問題なくスケールする。

銀河形成シミュレーション

アニメーション例
アニメーション例
まだあまりスケーラビリティはよくない (128 コア程度)
アルゴリズム改良中。目処はたっている。

グリッド計算実験

- NAOJ XT4 とアムステルダム大学の Power6 システムを 10G でつないで並列計算
- そのための異機種並列コード開発 (一応終了)
- ネットワーク接続実験: 2008/5/21
  - アムステルダム側は準備 (Power6 自体、、、) が間に合ってない
  - XT4 の2個の 10GbE インターフェースを使って、大陸間折り返し実験
  - コンスタントに 6Gbps (PCI-X インターフェースの限界) を実現
- 技術的可能性は実証出来た

今後の方向

- 計算センターとしての当面の必要事項
  - 大容量ファイルサーバ
  - 流体計算ユーザーの XT4 への移行
  - 粒子系計算ユーザーの GRAPE-DR への移行
- 次期システムに向けた検討・開発
  - 価格性能比の追求、電力効率の向上
    - GRAPE-DR 後継
    - GPGPU 等
    - その他