# 国立天文台における
# Cray XT4 の利用

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

Center for Computational Astrophysics

# 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
- 現在の運用状況
- (GRAPE と GRAPE-DR)
- 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。


## 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・測地学の研究が盛んで、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**

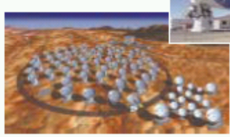
三鷹キャンパスは、国立天文台の本部が置かれ、さまざまなプロジェクト、センター、研究部、事務部が集まっています。

## 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。




## 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

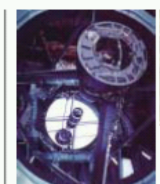
VERA Observatory
：Ishigakizima station



### ■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。

# すばる望遠鏡

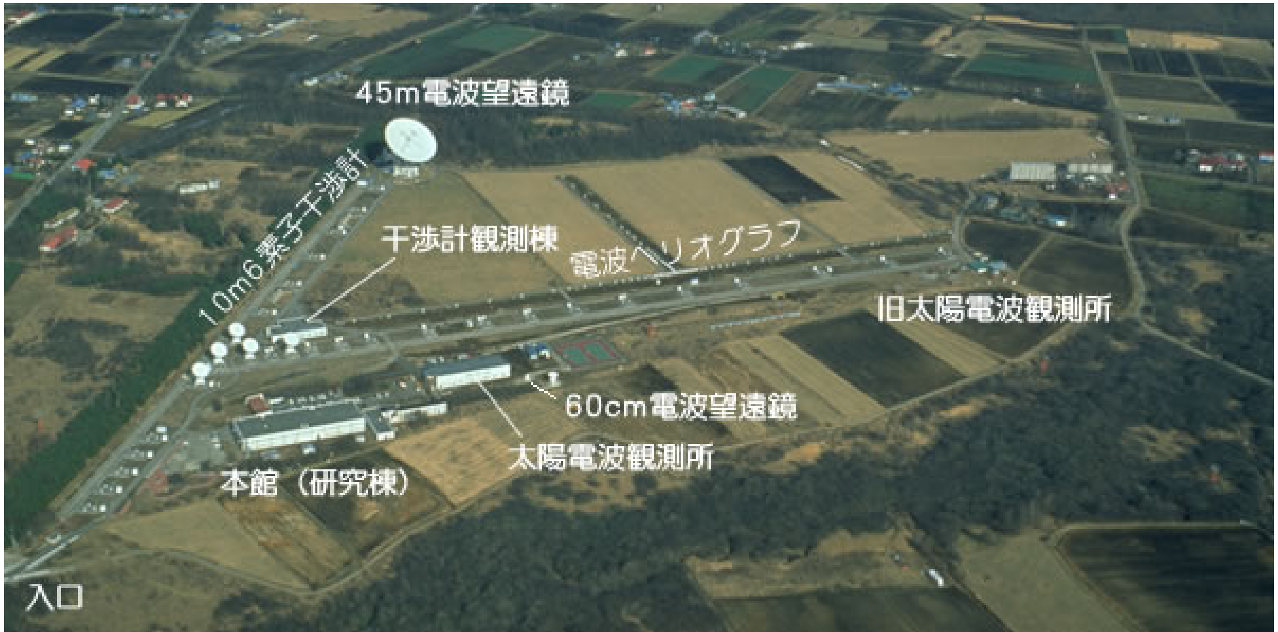
ハワイ・マウナケア山頂

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
(2010 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

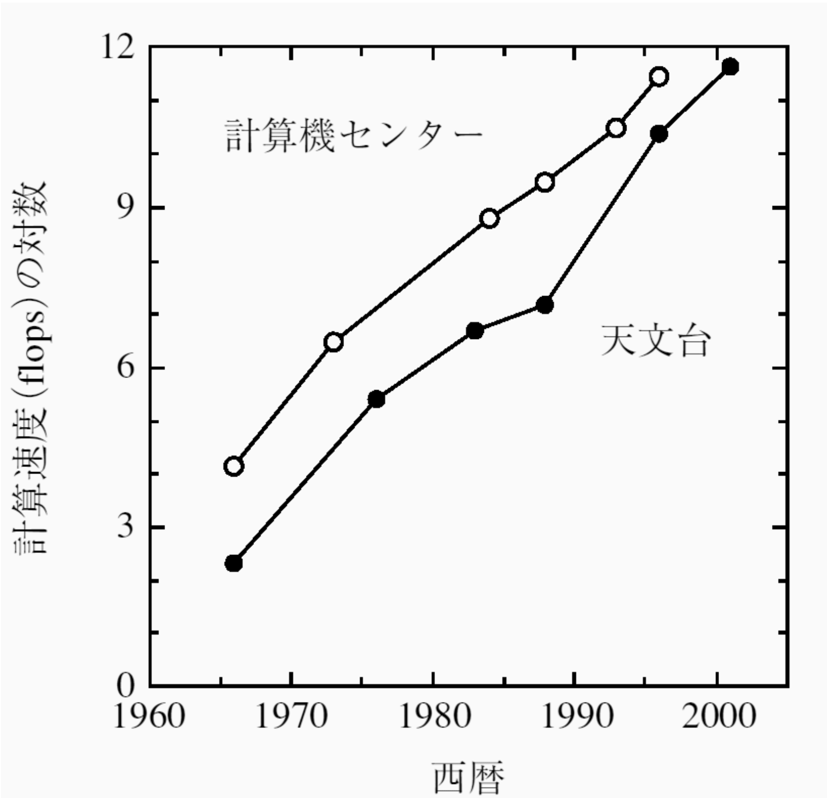
1988 年　天文学データ解析計算センター 発足

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 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：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)

# 資源配分

- プロポーザルベース
- 年 2 回募集 (大規模ユーザー)
- 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)

# 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 試験運用、2008/10/1 本運用

# 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/S 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

# 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

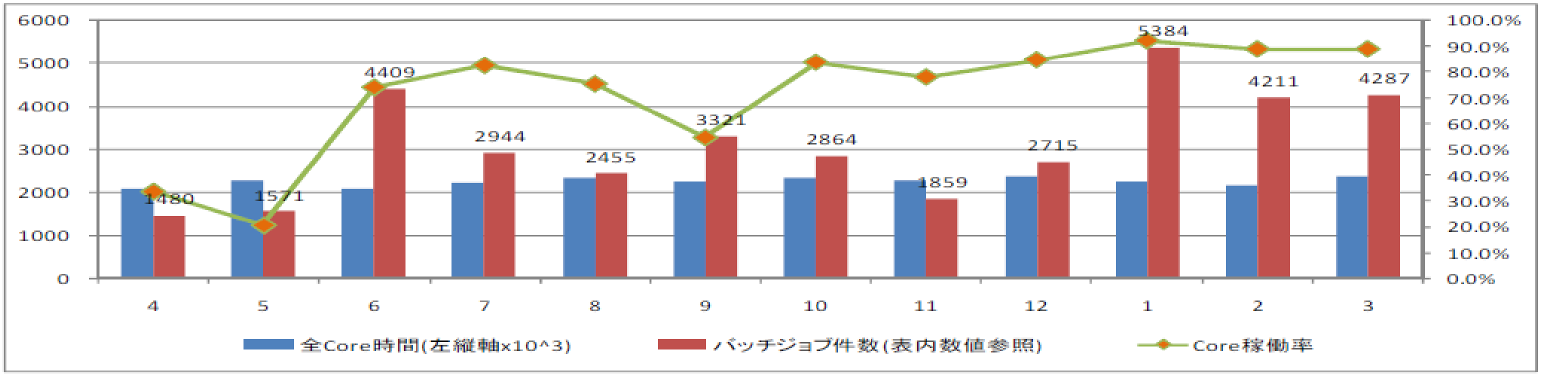
シミュレーション:斎藤貴之
可視化:武田隆顕 (4D2U プロジェクト)

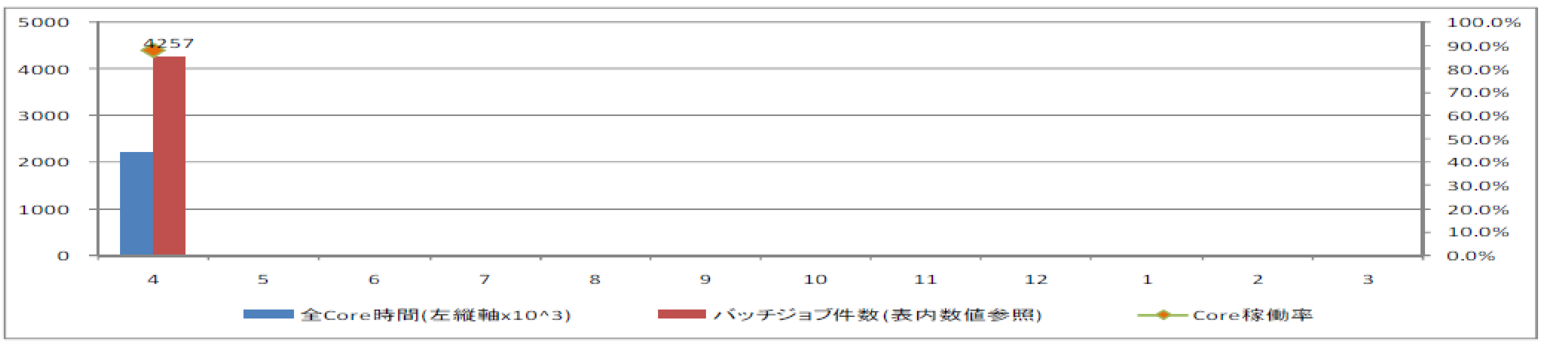
# 天文台以前と天文台以後

以前


以後

# Cray の稼働率統計




- 本運用の 2008/10 から 85% 以上
- キューのパラメータ調整で、極端に待ち時間が長くならないようにする仕掛けを導入(自動的に混雑具合をフィードバック、斎藤による提案・実装)

# 今日の利用状況

```
Compute Processor Allocation Status as of Tue May 26 08:52:21 2009

     C0-0     C1-0     C2-0     C3-0     C4-0     C5-0     C6-0     C7-0     C8-0
  n3 ShShSSSh jjjjjbbb gbbbbccc cccfccfc hhhhffff kkkdddjj ------ii iiiiijjj f-------
  n2  h h   h jjjjjbbb gbbbbccc cccfccfc hhhhffff kkkkddjj ---h--ii iiiiijjj ff------
  n1  j h   h jjjjjbbb gbbbbccc cccffccf dhhhhfff kkkkddjj -------i iiiiijjj ff------
c2n0 SjShSSSh jjjjjbbb gbbbbccc ccccfccf dhhhhfff kkkkddjj -------i iiiiijjj ff------
  n3 SjSjSjSj eeeeejjj jbbbbaag faaccccc fffgghhf dddddkkk lll----- -----hii jfffffff
  n2  j j j j eeeeejjj jbbbbaag faaccccc fffgghhh dddddkkk llll---- -----hii j-ffffff
  n1  j j j j eeeeejjj jbbbbjag faaacccc fffgghhh kddddkkk llll---- -----hii j-ffffff
c1n0 SjSjSjSj eeeeejjj jbbbbjag faaacfcc cfffghhh kddddkkk llll---- ------hi j-ffffff
  n3 SbSbSbSb hhhhheee bfffffjj caaaafff ffffffcc kdkkkkkk kkljjlll iiiiiii- jjjjjjjj
  n2  b b b b hhhhheee bfffffjj caaaafff fffffffc kkkkkkkk kklljlll iiiiiiii jjjjjjjj
  n1  b b b b hhhhhhee beffffjj caaaafff cffffffc kkdkkkkk jklljlll iiiiiiii jjjjjjjj
c0n0 SbSbSbSb hhhhhhee beffffjj caaaagff cffffffc fkdkkkkk jkkljjll iiiiiiii jjjjjjjj
    s01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567 01234567
```

**512 コア x 2, 256 コア x 5, 128 コア x 4, 64 コア x 1**

# 障害状況

- 2009/4 ノード障害 3 件
- 大体こういう感じ。月に 1-3 件程度。当座予備ノードを割り当て、定期メンテナンスの時に交換等で対応。
- 1000 ノード近くある割には壊れない (普通の PC クラスタはもっと壊れる) という印象
- 大規模な障害 (システム全体が落ちるとか) は試験運用中に 1 度のみ。あるユーザーのプログラムで落ちる、、、というような、、、

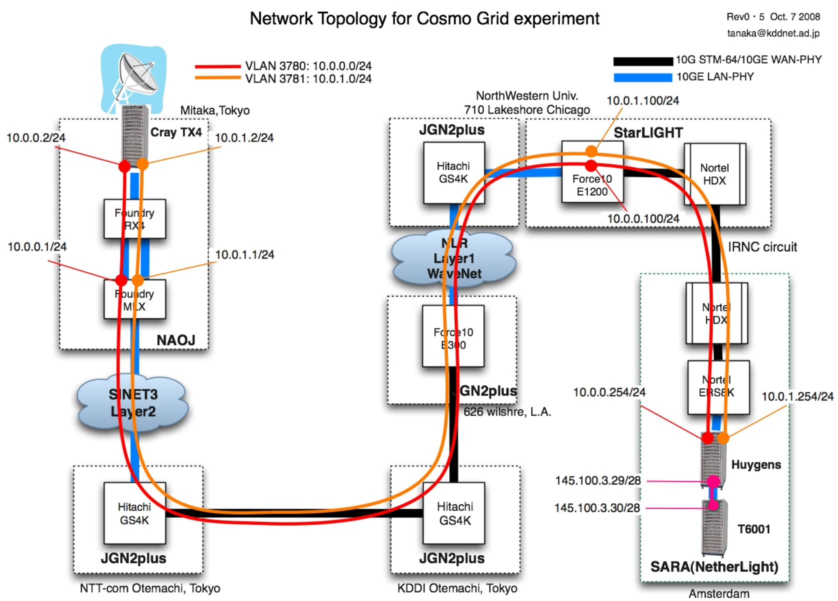
# ユーザー層の変化

- 従来の VPP ユーザーの移行はある程度進みつつある
- 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新規に利用開始が参入
- 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
  * 遅くて細いネットワーク向け最適化済み
  * 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？

# 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)
- 2008 年度に 3 回接続して計算実験

- 遠距離の TCP/IP 通信は簡単ではないが、コネクション数、ソケットパラメータを調整すればバンド幅はでる。
- 技術的可能性は実証出来た

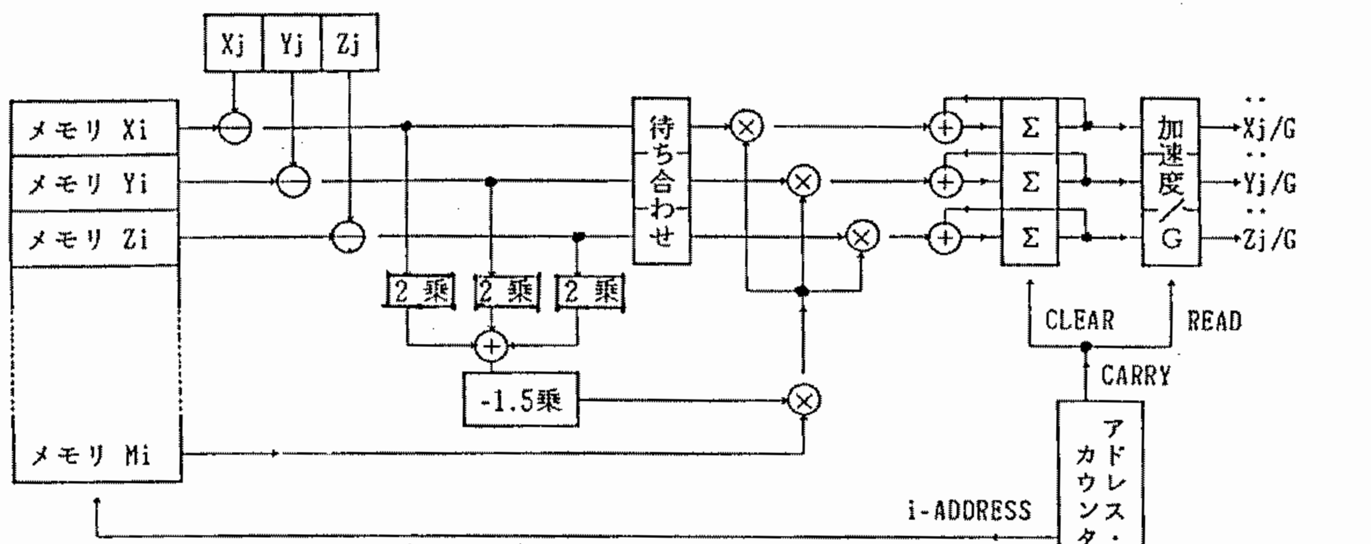
アニメーション

# GRAPE の話

- 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

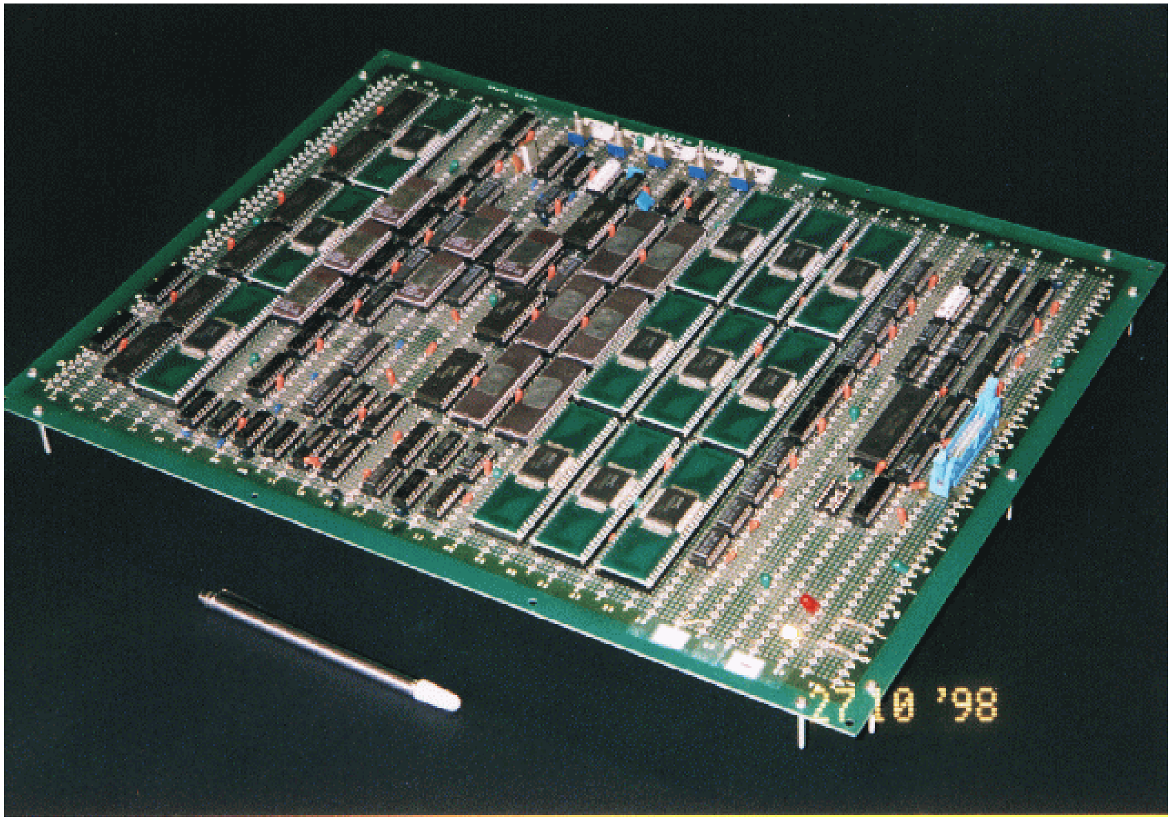



$$\ddot{\mathbf{r}}_j = G\sum_i \frac{M_i}{r_{ij}^3}\mathbf{r}_{ij} \qquad (\mathbf{r}_{ij} = \mathbf{r}_i - \mathbf{r}_j)$$

+, −, ×, 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.

# 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

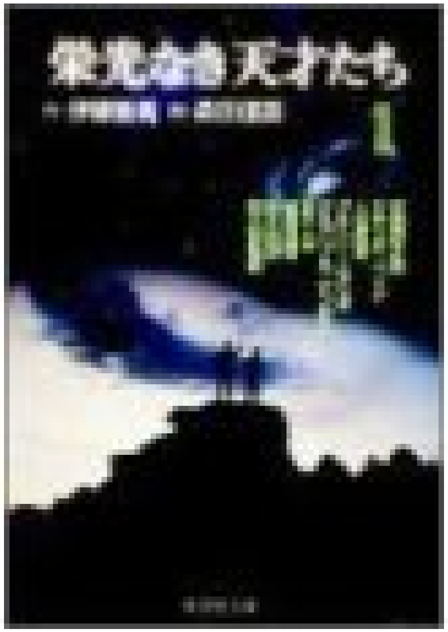
# 開発はどんなふうだったか


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

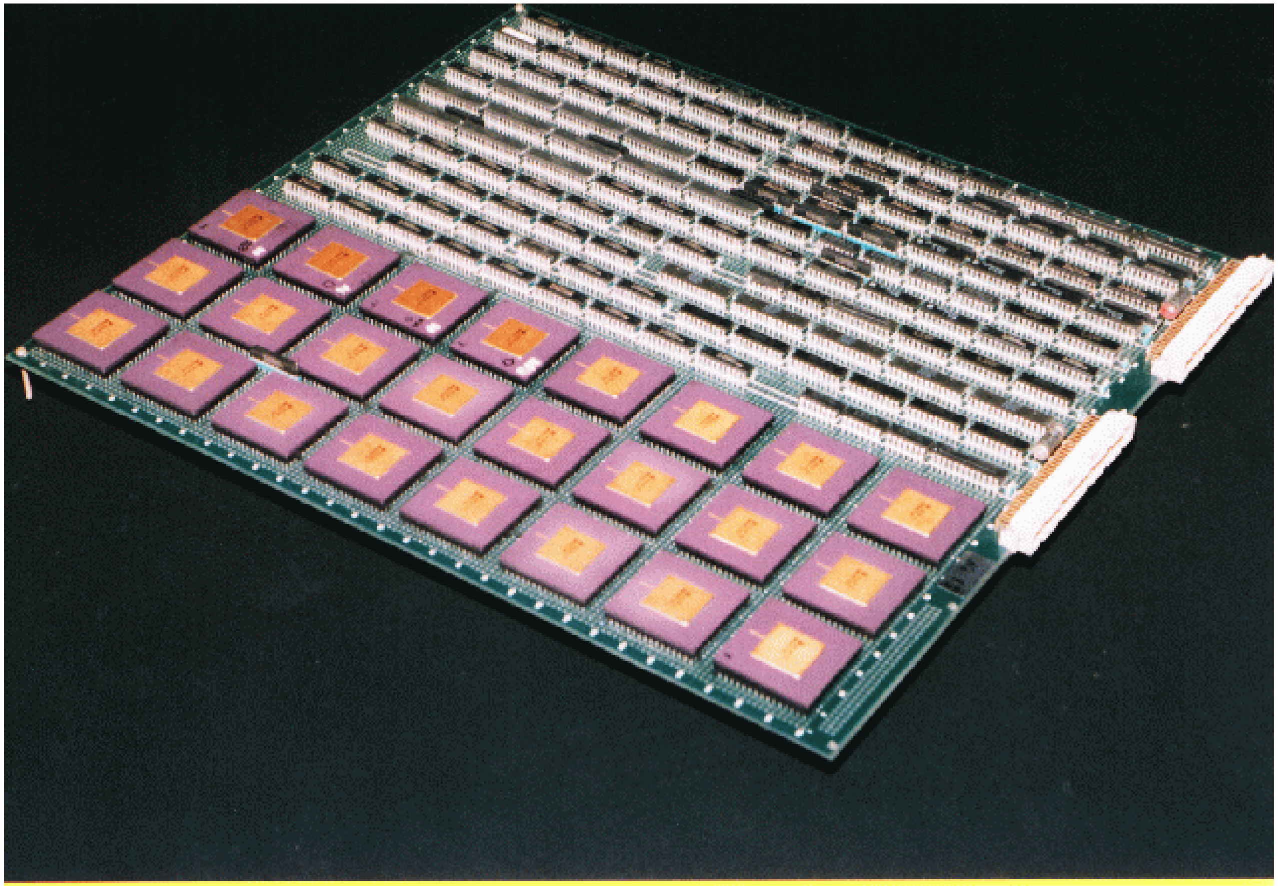



「栄光なき天才たち」の原作者としてのほうが有名という話も。

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-3(1991)

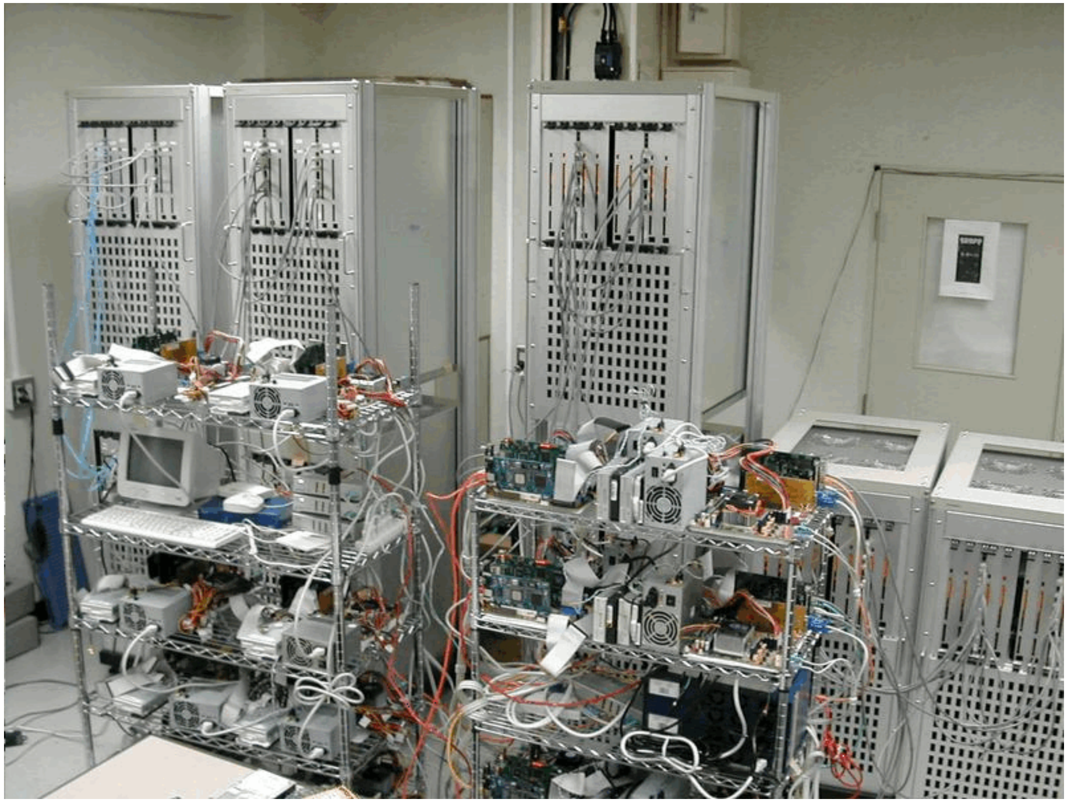


カスタムチップ 24個1ボード、 10MHz 動作、 7.2Gflops

# GRAPE-4(1995)

トータル 1792 チップ、 1.1 Tflops

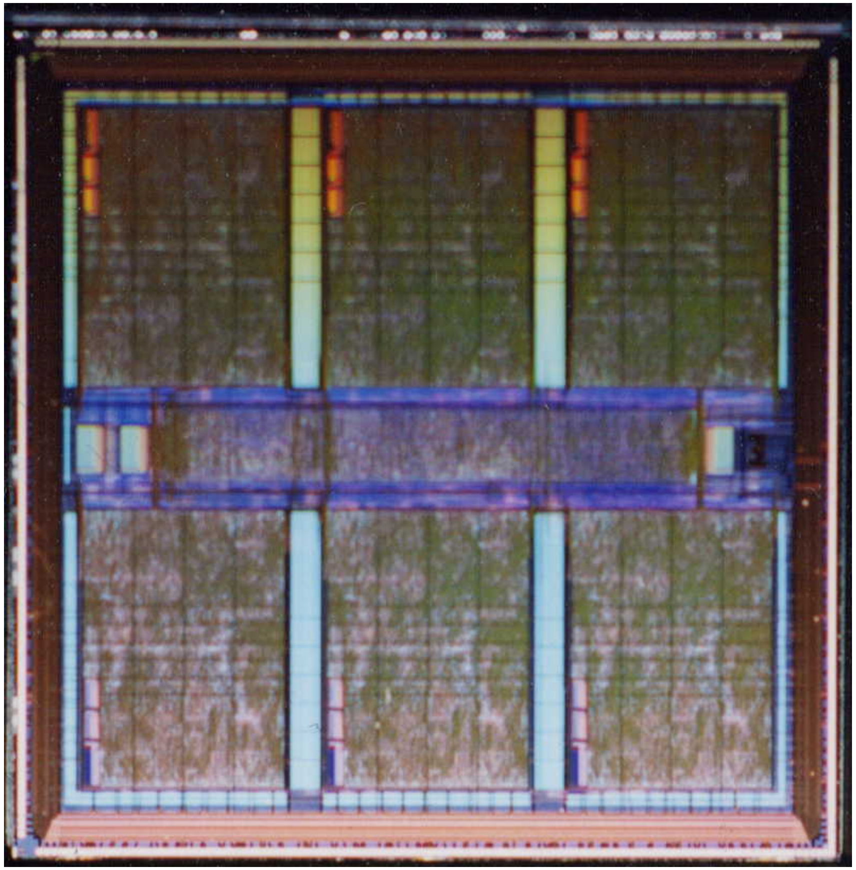
# GRAPE-6(2002)



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

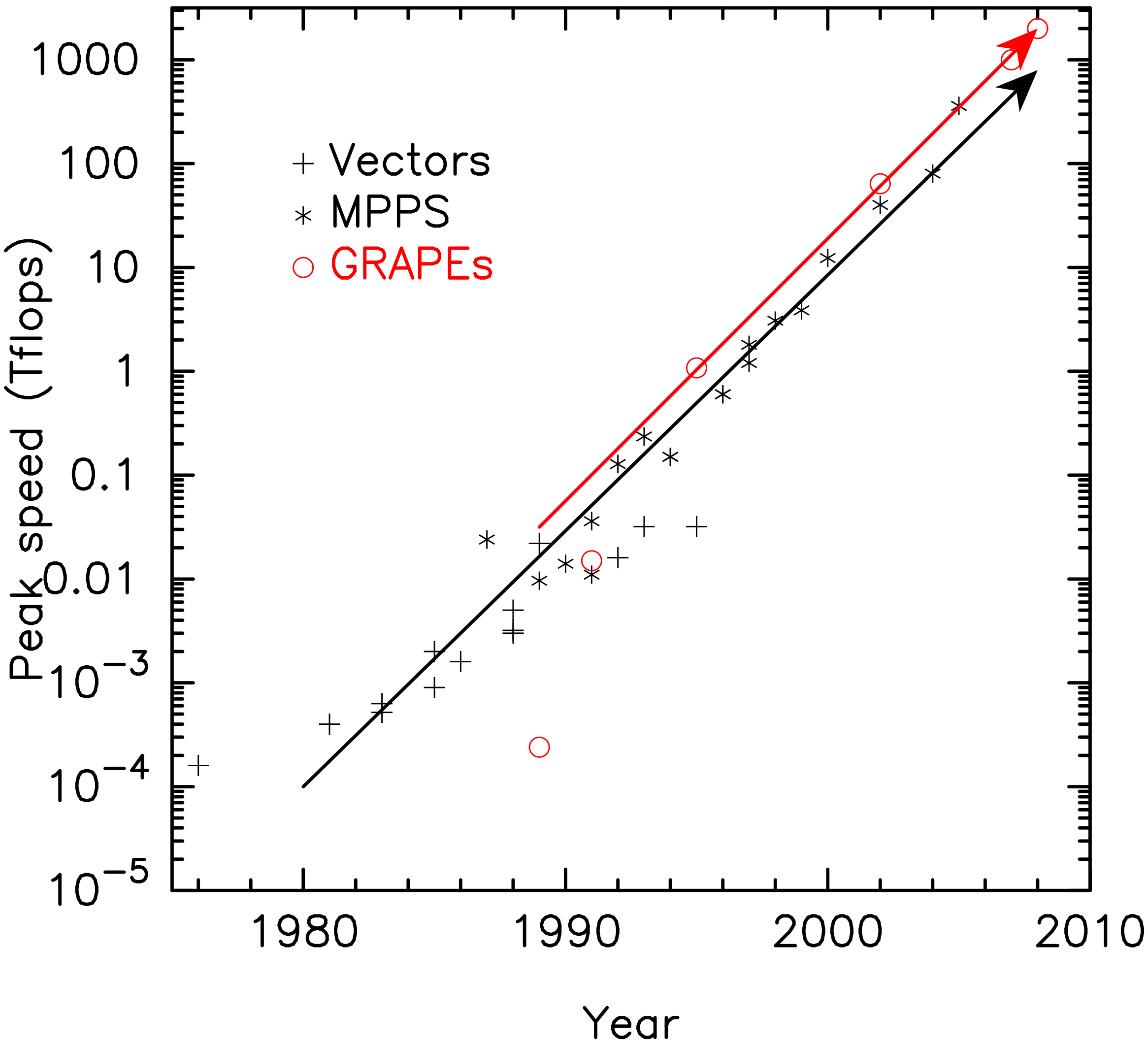


- 0.25 $\mu$m ルール (東芝 TC-240, 1.8M ゲート)
- 90 MHz 動作
- 6 パイプラインを集積
- チップあたり 31 Gflops

# 現在のマイクロプロセッサと比べてみる

| | GRAPE-6 | Intel Xeon 5580 | IBM BG/P |
|---|---|---|---|
| Year | 1999 | 2009 | 2007 |
| Design rule | 250nm | 45nm | 90nm? |
| Clock | 90MHz | 3.2GHz | 850MHz |
| Peak speed | 32.4GF | 51.2GF | 13.6GF |
| Power | 10W | 130 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

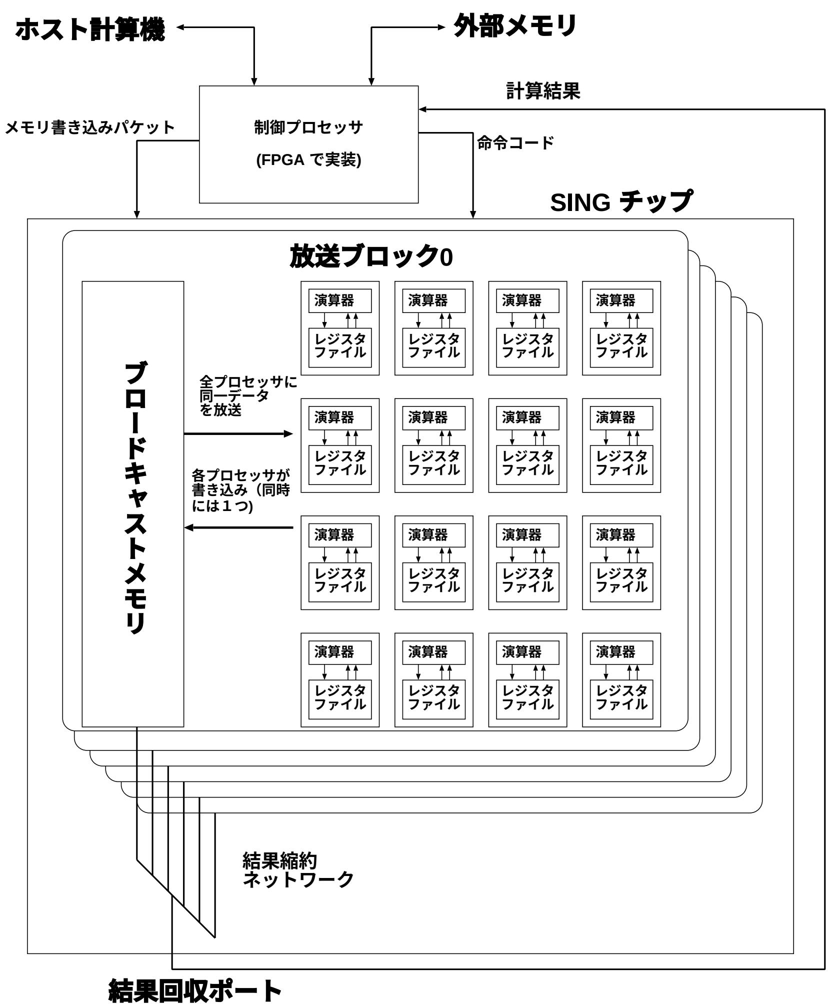
# GRAPE-DR

- カスタム LSI による専用パイプライン:開発コストが高騰

  $\longrightarrow$ 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

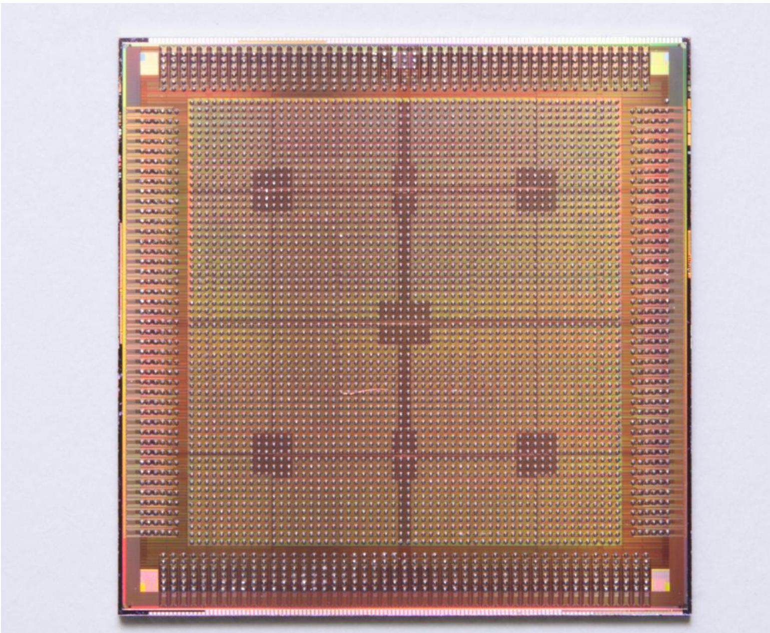
  $\longrightarrow$ 専用パイプラインでなくても効率上げられる

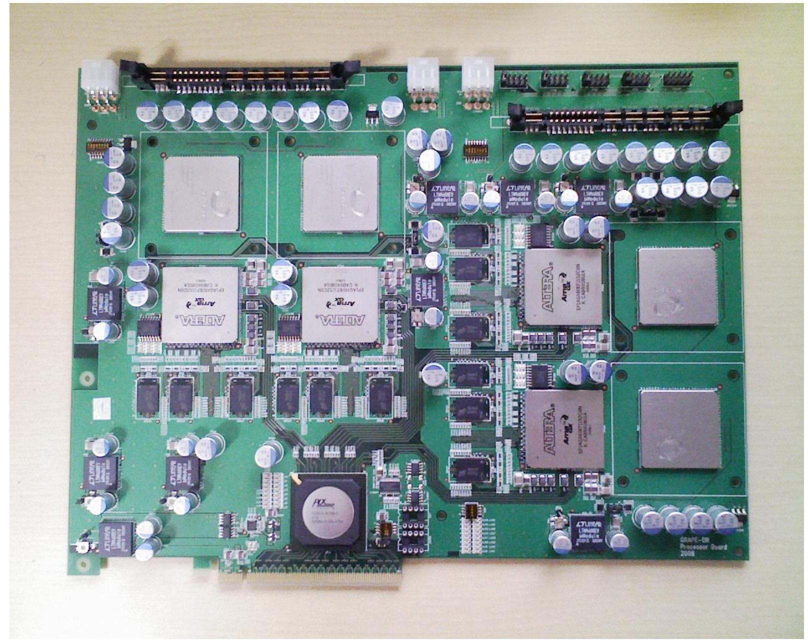
# GRAPE-DR の構成




- 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

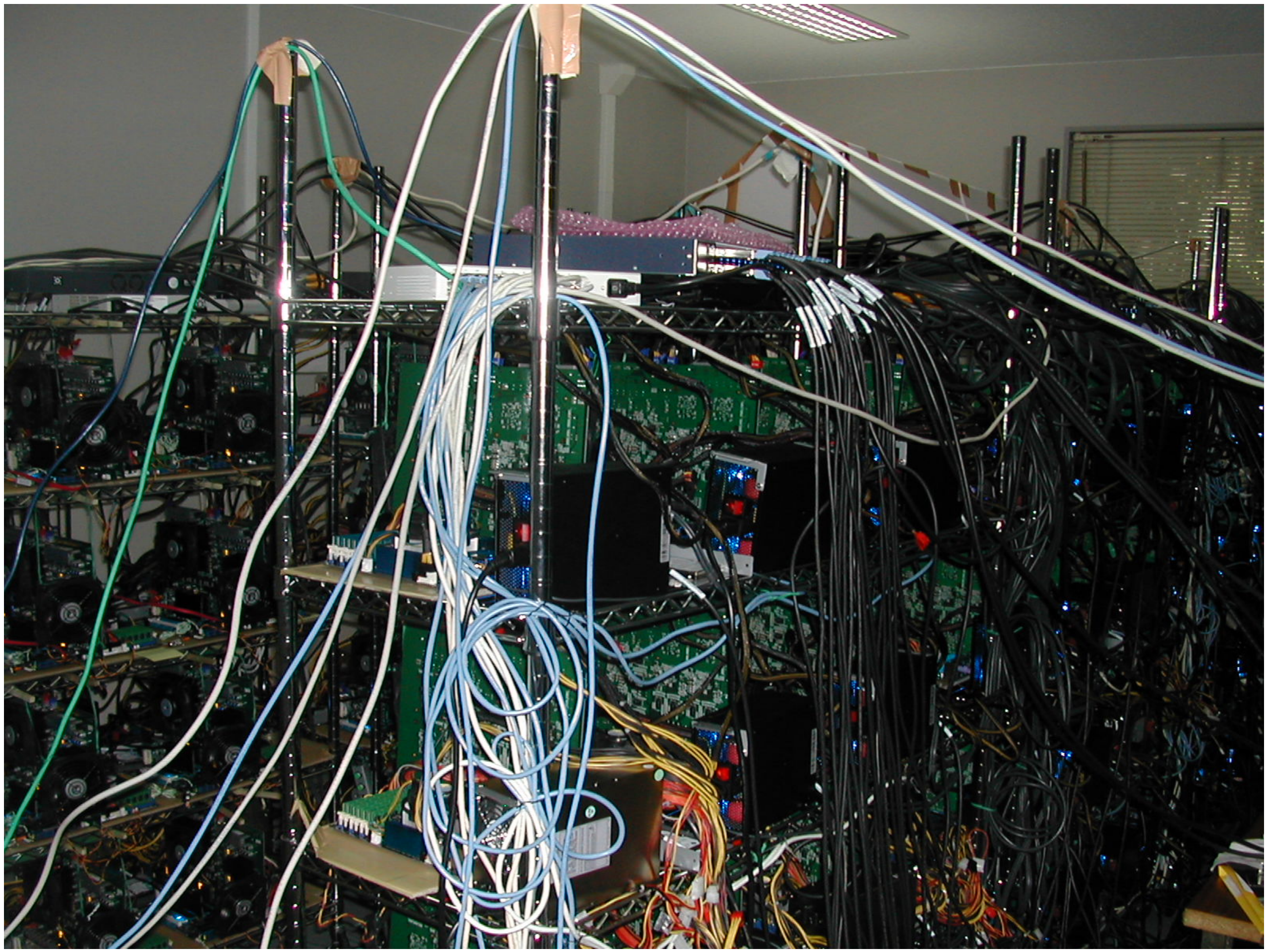
# プロセッサチップとボード





- 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)
- 消費電力 (最大) 250W

# GRAPE-DR クラスタ (世界初公開)



128 ノード (128 カード) のうち半分。稼働中。今年度中には 700 カードくらいまで拡張の予定。

# 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

# まとめ

- 国立天文台では 2008/4 から XT4 (9 キャビネット) を共同利用システムとして運用開始
- システムは非常に安定している。ハード、ソフトともトラブルは少ない
- 従来のベクトル並列システムと互換性はないが、ユーザーは十分多い。
- ユーザー層は若くなったような、、、
- ユーザーの評価は高い。スケーラビリティ、I/O 性能等