

GRAPE-DR の現状 および GRAPE-8 の開発状況

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/理論研究部



2010/10/1 研究室セミナー

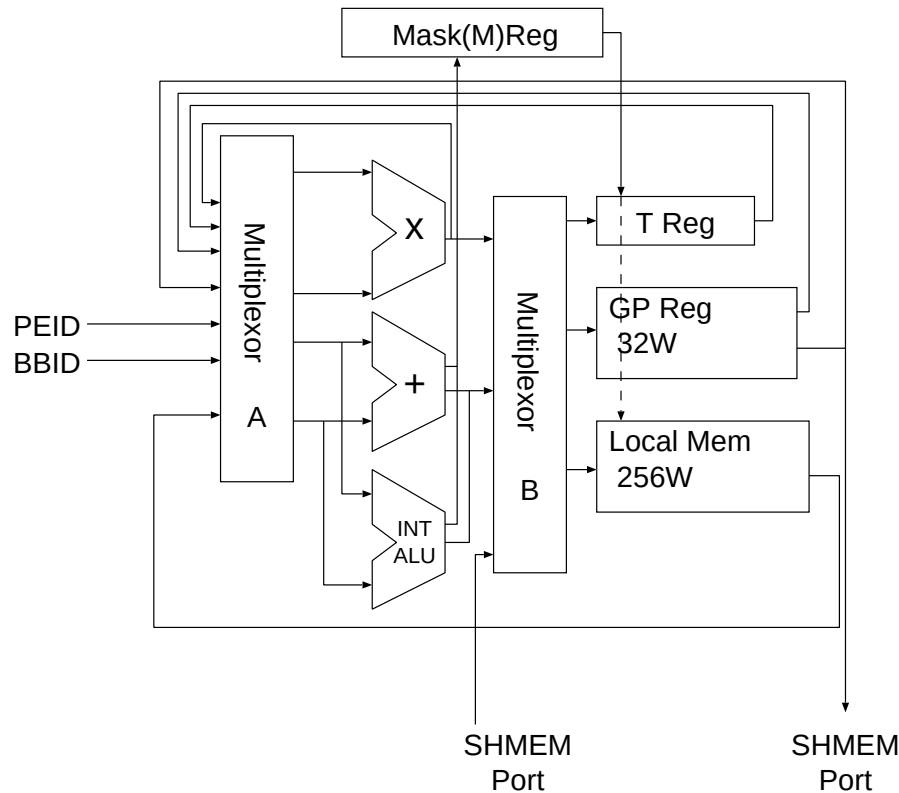
講演概要

- **GRAPE-DR の現状**
 - **GRAPE-DR の概要**
 - **現状の構成**
 - **性能・改良点等**
- **GRAPE-8 (9) の開発状況**
 - **方針**
 - **開発状況**

GRAPE-DR

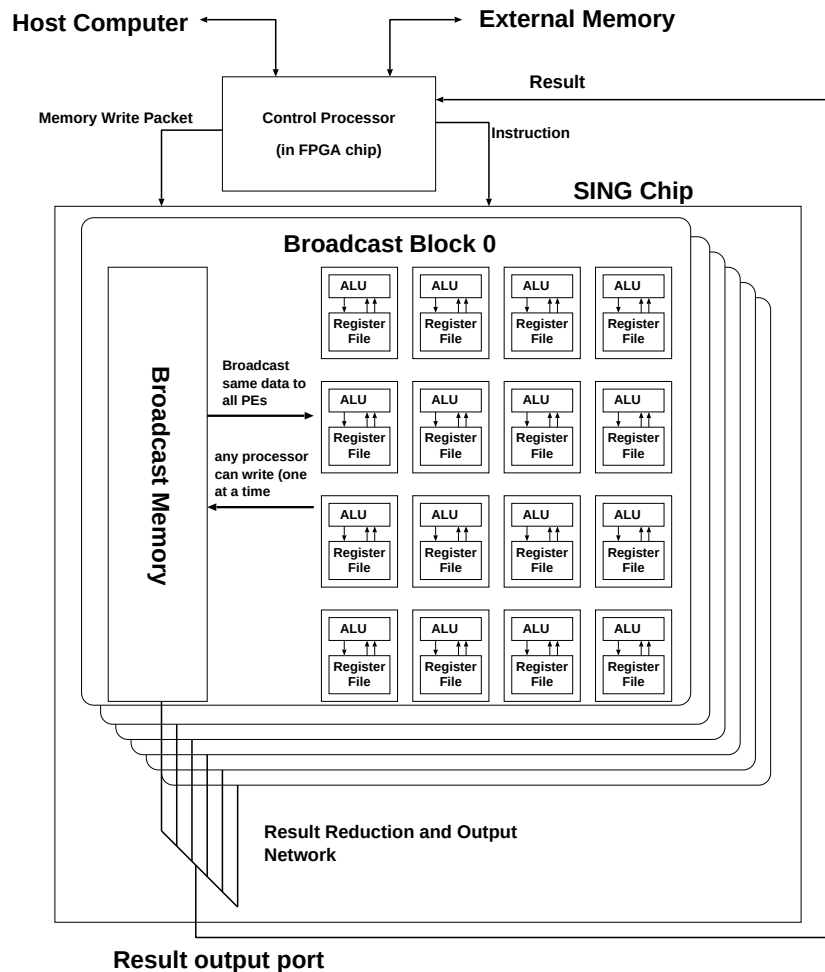
- 従来の GRAPE(重力計算の専用ハードウェア) とは違い
アーキテクチャ
- 「汎用」ということで予算とる
- 非常に単純で小さなメモリだけをもつ演算プロセッサを
多数並べる SIMD 構成
- 何故か最近の GPU と良く似た方向
- 開発予算: 振興調整費 (FY2004-2008)

プロセッサアーキテクチャ



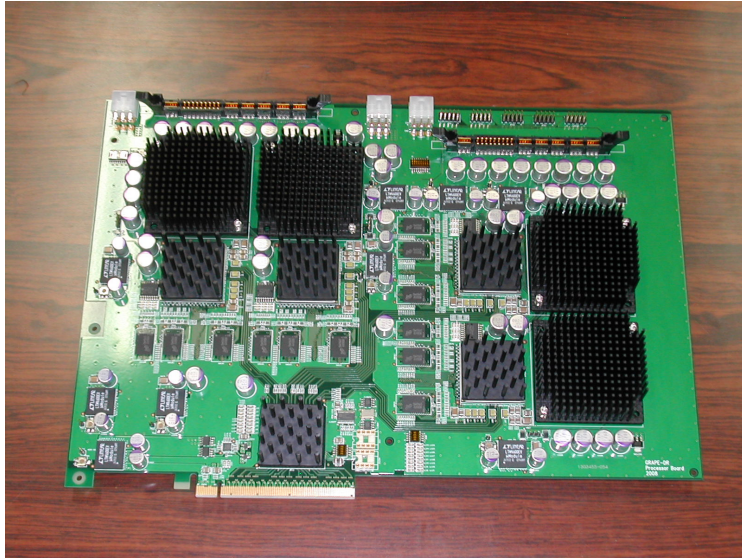
- 倍精度乗算器 (の半分)
- 倍精度加減算器
- 整数・論理演算器
- 32 語のレジスタ
- 256 語のメモリ
- 通信ポート

プロセッサチップ構成



- 32 個 (図では 16) のプロセッサで「放送ブロック」を構成。放送ブロックは 16 個。
- 放送ブロック毎に共有メモリ
- 外からの入力データは全ブロックに放送
- 各ブロックからの出力データを合計等しながら出力可能

プロセッサボード

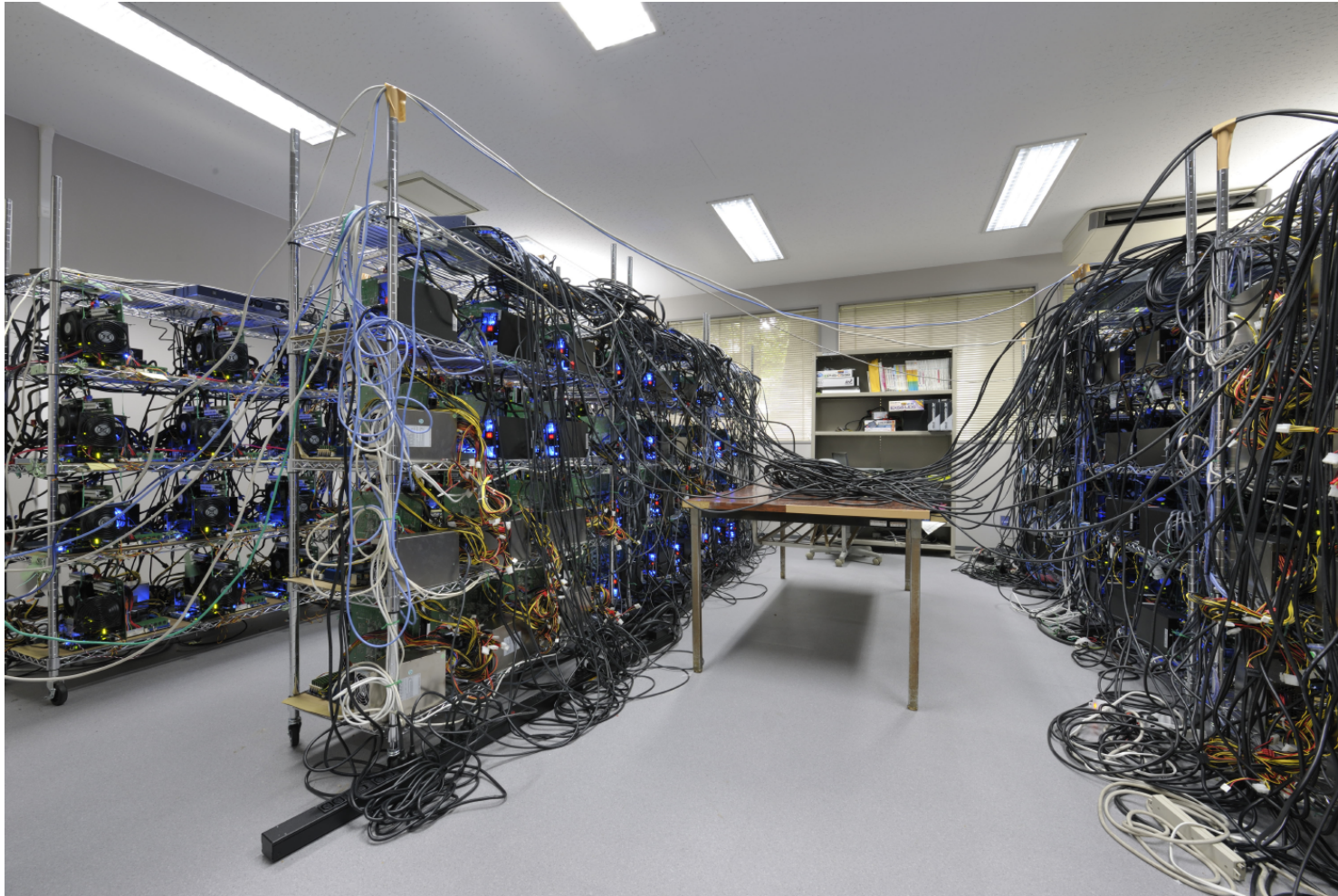


PCIe x16 (Gen 1) インターフェース

Altera Arria GX を DRAM 制御、通信インターフェースに利用。

- 最大電力消費 200W くらい
- 433MHz くらいまでは問題なく動作
- 400 MHz での理論ピーク性能は 819Gflops
- KFCR で販売中

GRAPE-DR クラスタシステム



GRAPE-DR クラスタシステム

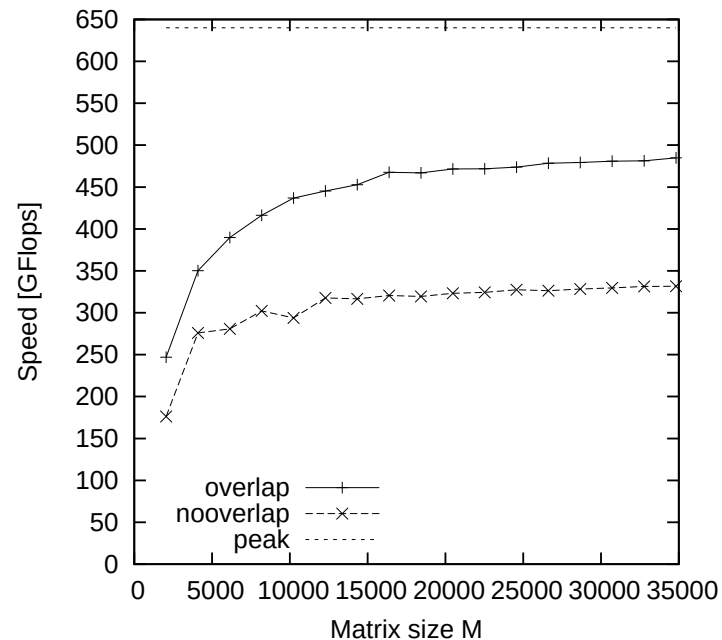
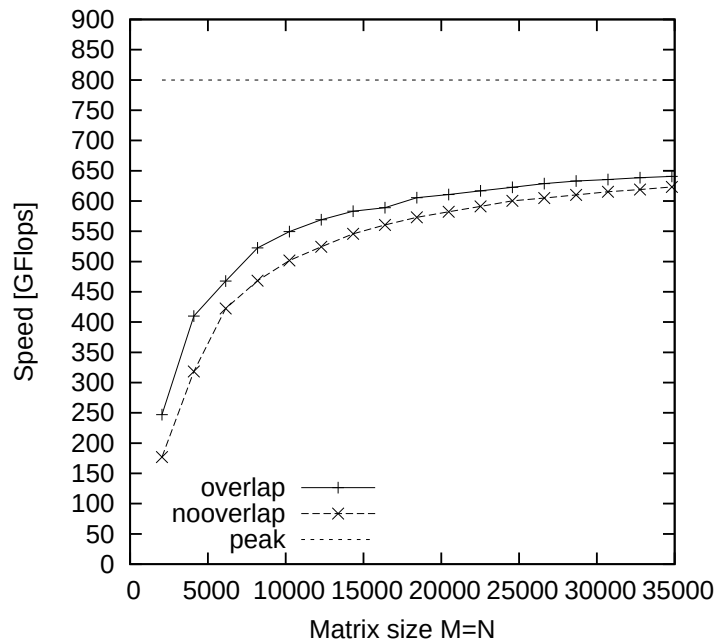
前のスライドの写真とった時の構成:

- 128-ノード, 128-ボード (105Tflops peak)
- ホスト計算機: Intel Core i7+X58 12-24 GB メモリ
- ネットワーク: x4 DDR インフィニバンド
- 現在色々増強中

性能

- 密行列計算
- 重力

行列乗算



M=N, K=2048, 640 Gflops

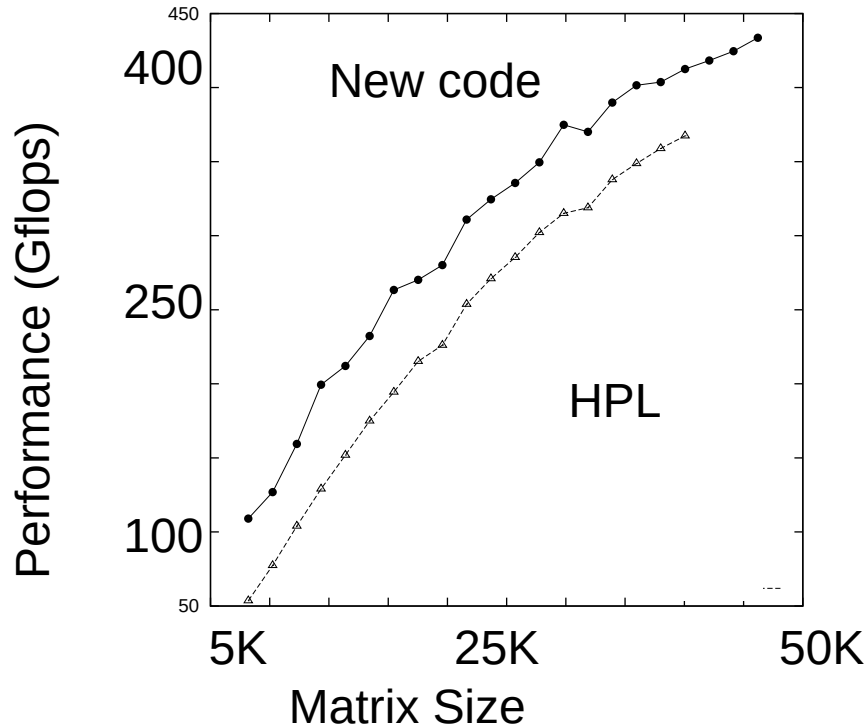
N=K=2048, 450 Gflops

現在シングルカードの性能としては世界一。

NVIDIA Tesla C2050: 300Gflops

AMD Radion 5850: ?? (通信なしで 500Gflops)

LU 分解の性能



行列サイズに対する性能
(Gflops)

上:新規開発コード

下:HPL 1.04a

行列格納順序の変更・処理途中で転置

その他色々新しいアルゴリズム開発

しつこいですが1カードの性能として世界一

Little Green 500, June 2010

Green500 Rank	MFLOPS/W	Site*	Computer*	Total Power (kW)
1	815.43	National Astronomical Observatory of Japan	GRAPE-DR accelerator Cluster, Infiniband	28.67
2	773.38	Forschungszentrum Juelich (FZJ)	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
2	773.38	Universitaet Regensburg	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
2	773.38	Universitaet Wuppertal	QPACE SFB TR Cluster, PowerXCell 8i, 3.2 GHz, 3D-Torus	57.54
5	536.24	Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw	BladeCenter QS22 Cluster, PowerXCell 8i 4.0 Ghz, Infiniband	34.63

電力当りの性能で、しつこいですが、世界一を達成

#2: IBM PowerXCell, #9: NVIDIA Fermi

HPL (並列 LU 分解)

- 1 ノードでやった新しい計算法全部実装
- 縦・横の通信両方を隠蔽 (HPL は横のみ)
- TRSM のオーバーヘッドを減らす新しいアルゴリズム

815Mflops/W: #1 in Little Green500 list

N=240K, 64 ノード: 24Tflops/29KW

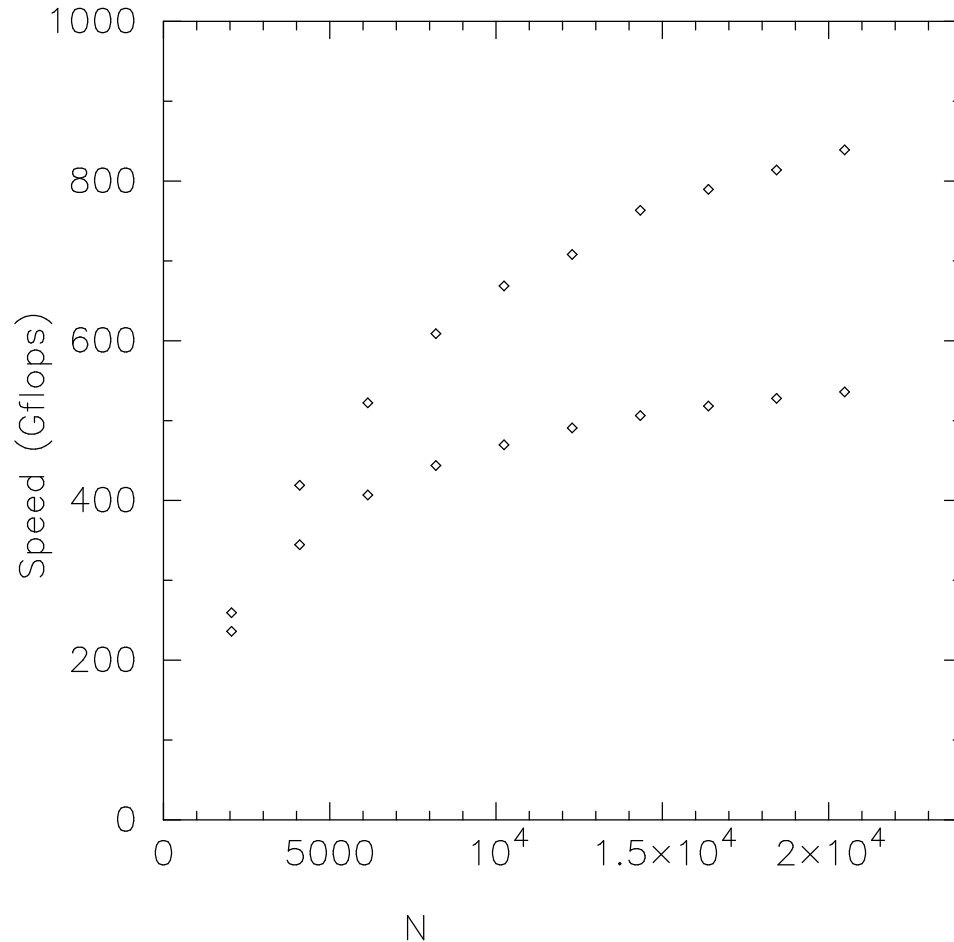
HPL 1.04a の2倍の性能を達成

GRAPE-DR ハードウェア・ライブラリの改良点等

- DMA read がともに動くようにした
 - Ka 君の実装に performance bug があったので、新しい実装にした
 - 起動オーバーヘッドを低減した (non-combined write 1 語で起動。前の実装は 4 語だったと思う)
- 計算起動等のオーバーヘッドも同様に小さくした
- 1 回の計算起動で帰ってくるデータ量を減らした

で、カード 2 枚で動くようにはした。

性能



上: カード 2 枚、下: 1 枚

**行列が十分大きいところ
の性能: カード 2 枚で
1150Gflops くらい**

問題点:

A 行列転送が遅い

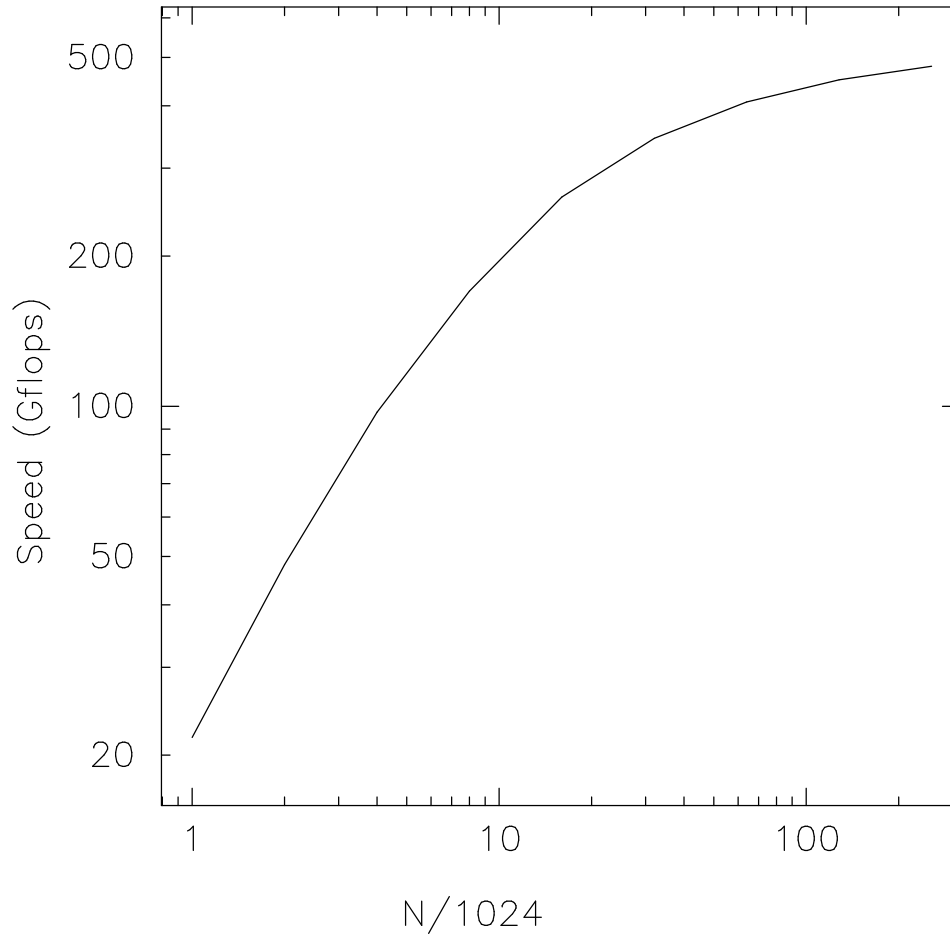
**IB を使って通信が止ま
るようになる、、、**

今後の改良予定

- A 行列の転送時にホストでやっているデータ並べかえを DR チップ側で実行する
- ホストのほうのチューニング。 QPI のクロックとか？

重力 (4次エルミート公式用)

独立時間刻みコードの性能



$N = \infty$ での極限性能:

517 Gflops

Phantom GRAPE on
4core Core2: 27Gflops

程度

GeForce 9800GTX

(OC): 220Gflops 程度

これまでの改良点

- アセンブラが出したコードがそのまま実行できるようにした (今までそうになってなかった、、、)
- i 粒子送るのを速くした
- GRAPE-5 互換版について、1 プロセスで 4 チップ使うインターフェース (シングルスレッドで動作) をつけた
- GRAPE-6 互換版について、1 スレッドで動作するようにした。、また N が小さいところでの性能を KFCR 版よりはあげた。
- 一応ドキュメントを書いた

<http://grape.mtk.nao.ac.jp/pub/people/makino/software/gdrlib/>

今後の改良点

- i 粒子送るのもうちちょっと速くする
- G6 互換版で j 粒子を送るのがまだ遅いのを直す
- 計算カーネルのアセンブラステップ数が多いのを短いのに入れ替える

GRAPE-8(9?)

GRAPE-DR の次をどうするか？

- GRAPE-DR 的な、プログラマブルなものの次世代機
- 従来の GRAPE 的専用機
 - － カスタムチップを作るのは無理
 - － Structured ASIC (後述) を想定

一昨年に前者を特別推進、後者を基盤 (A) で申請して、前者はヒアリングでこけたので現在後者を開発中。

Structured ASIC

カスタム LSI と FPGA (論理回路変更可能な LSI) の中間

	ASIC	FPGA	Str. A.
初期費用 (10 億円)	1	0.001	0.02
回路規模 (1 千万ゲート)	20	1	1
ゲート単価 (円/1000 ゲート)	0.3	10	2

良くいうと、 ASIC より初期費用安くて FPGA より量産費用安い

上の数字の例では、予算 2400 万から 12 億の間ではベストソリューション

そもそも作る意味はあるか？という問題

- GPGPU を買ってくるほうが安くないか？
 - 安い GPU 買ってくるならそう
 - 電気代いれるとそうでもない。
電力当り性能は 20 倍以上
- 重力だけで役に立つか？
 - 無衝突系だけだと？
 - 衝突系なら役に立つ

どんなものを作る計画か？

9/23 現在の話

- 相対精度 9 ビットの「低精度型」
- 64 パイプライン/チップ、400MHz 動作、1Tflops、10W(この辺の数字はまだ微妙に流動的)
- ボードとしては、 10Tflops、100W が目標

2012 年の GPU: 1.5Tflops(DP), 300W?

(NVIDIA 社長の講演@GTC2010)

衝突系に使える？

ツリーコードと直接計算のハイブリッドなら「多分」大丈夫

PPPT (Particle-Particle, Particle-Tree), Oshino et al. (in prep).

$$H = H_{Hard} + H_{Soft},$$
$$H_{Hard} = \sum_{i=1}^N \left(\frac{p_i^2}{2m_i} - \frac{Gm_i m_{\odot}}{r_i} \right) - \sum_{i < j}^N \frac{Gm_i m_j}{r_{ij}} W(r_{ij}),$$
$$H_{Soft} = \sum_{i < j}^N \frac{Gm_i m_j}{r_{ij}} (1 - W(r_{ij})),$$

- P³M スキームと同様に近距離力と遠距離力を分割
- 遠距離力はツリー、近距離力は独立時間刻み直接計算

PPPT テストラン

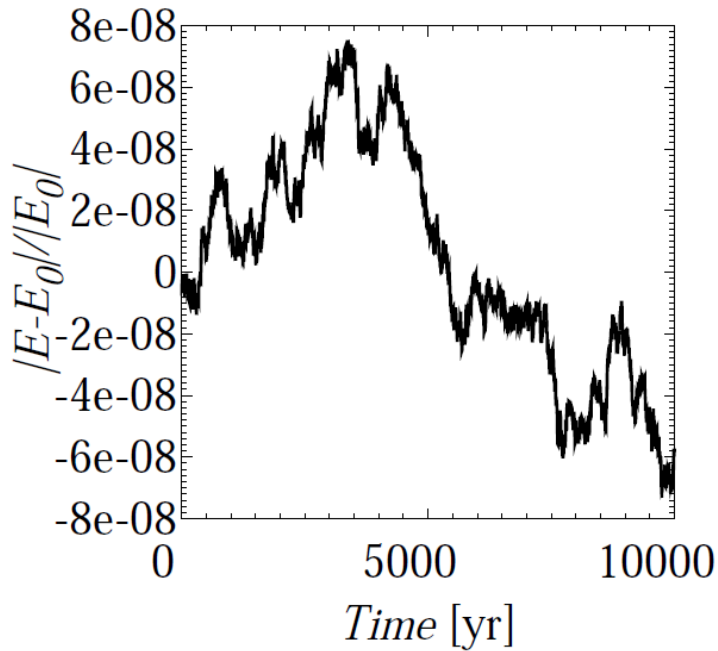
地球領域、微惑星 10^4 粒子

ツリーコードの力の誤差が主要因

エラーがランダムウォークする (累積しない)

惑星形成では十分な精度

星団とかはまだテストはしていない。原理的には問題ないはず。



まとめ

- GRAPE-DR は大体出来て、性能もできるようになった。
- カード1枚の性能は 4 コア CPU の
15(DGEMM)-20(重力) 倍
- 次世代は低精度重力専用だが、ハイブリッドスキームで
衝突系もやる
- 目標は GRAPE-DR カードの 10 倍の速度、半分以下の
電力