GRAPE Project の現状

牧野淳一郎

東京大学大学院理学系研究科

講演概要

1. 昨年度の積み残し
2. 今年度の成果
   - GRAPE-6 実効性能評価
   - ゴードン・ベル賞ファイナリスト
   - 東大外での利用
   - 今後の計画
   - 次世代システムの検討
3. まとめ

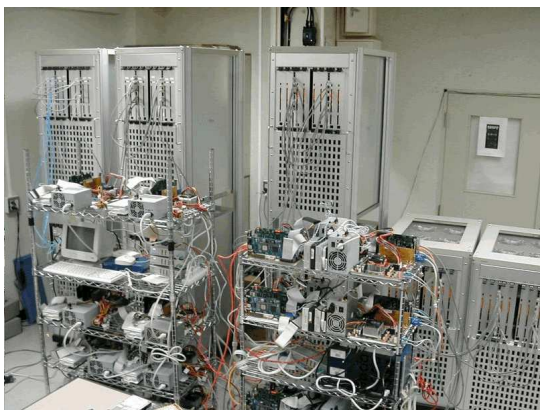
昨年度の積み残し

- 64 Tflops GRAPE-6 の完成

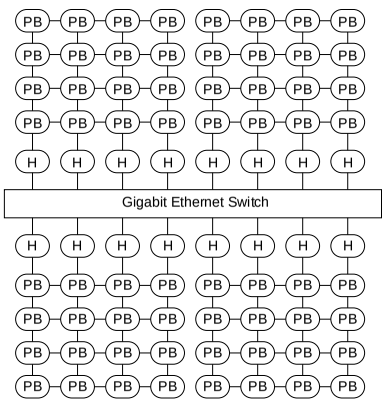
昨年度末時点では 48 Tflops 相当までしか組み上がっていなかったが、2002 年 4 月にプロセッサボード 64 枚、ホスト 16 台、ピーク性能 64 Tflops のシステムが完成した。

- ネットワーク構成
- ホスト計算機
- ホスト間ネットワーク

現在の GRAPE-6



全体ネットワーク構成

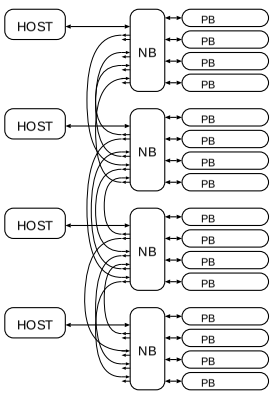


4 クラスタ構成

1 クラスタ = ホスト 4 台+プロセッサボード 16 枚

プロセッサボードは 4×4 の (分割可能な) 2 次元ネットワーク

クラスタ内ネットワーク



ホストはネットワークボードを介してプロセッサボード 4 枚につながる

ネットワークボードは All-to-all 結合 (並列ブロードキャストが可能)

ネットワークトポロジはダイナミックに切換え可能

ホスト計算機

現状:

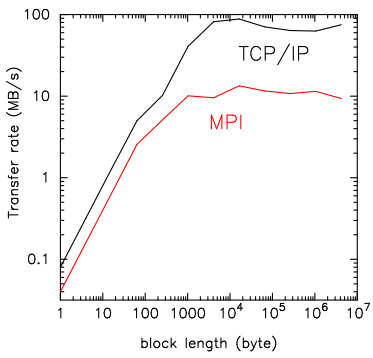
CPU	AMD Athlon XP 1800+
マザーボード	ECS K7S6A (SiS 745)
メモリ	PC2100 1GB
OS	Linux (kernel 2.4.18, Slackware 8.0)
NIC	NS83820 ベース

ホスト間ネットワーク

現状: ギガビットイーサネット。
通信ライブラリ

- 当初は MPICH/pr over TCP/IP を利用
- パフォーマンスに問題が出たので、UNIX ソケットを使う単純なライブラリを実装
- Dual NIC に強化する予定

ネットワーク通信性能



何故 MPI がこれほど遅いかは良く分からない

OS, compiler, MPICH のバージョン等にもよる。

30MB/s 程度でるケースもある。

TCP/IP 自体のレイテンシは 100μs 程度

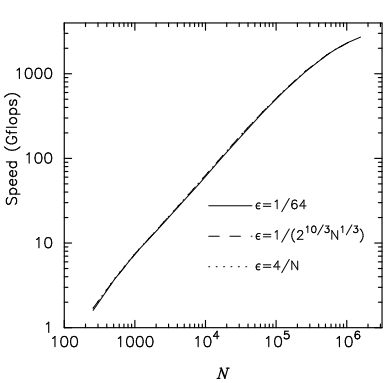
今年度の成果

- GRAPE-6 実効性能評価
- ゴードン・ベル賞ファイナリスト
- 東大外での利用
- 今後の計画
- 次世代システムの検討

GRAPE-6 実効性能評価

- 単一ノード性能
- マルチノード性能
- マルチクラスタ性能

単一ノード性能

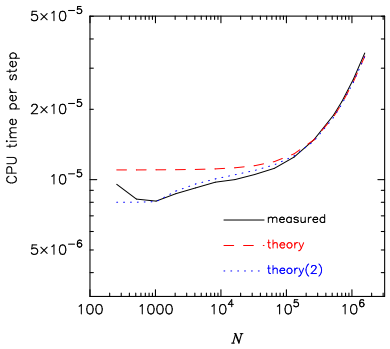


半性能長: 5×10^5 程度

性能はタイムステップの分布にはあまりよらない

通信よりもホストの計算速度がボトルネック

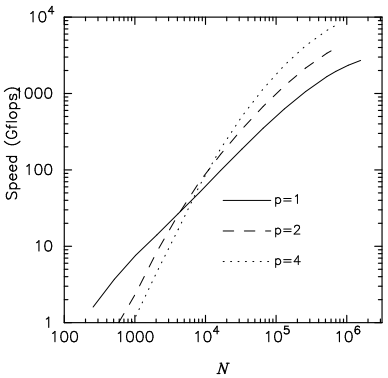
性能モデル



単純な
(ホスト時間)+(通信)+(GRAPE 計算時間)
モデルではあわない

粒子数が大きいところでホストのキャッシュミスの影響が大きい

マルチノード性能

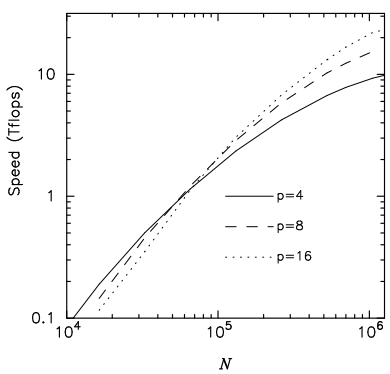


クロスオーバーは
$N = 5,000$ 程度

半性能長は単一ノードの場合とほとんど同じ

通信はネットワークボード間のみで、これはホストからの転送と同時に起こるので完全に隠蔽される。

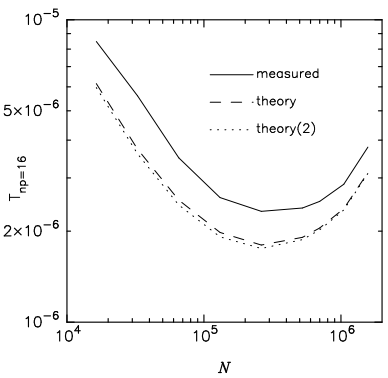
マルチクラスタ性能



クロスオーバーは
$N = 5 \times 10^4$ 程度

ホスト間のデータ交換が発生。隠蔽されない。

マルチクラスタ性能モデル



ホスト間通信が問題

粒子数が少ない: 通信バンド幅よりもレイテンシが効く
粒子数が大きい: バンド幅リミット
有効に使うにはレイテンシの低い通信の実装が重要

ゴードン・ベル賞ファイナリスト

- 1.8 M 粒子の直接計算 (カイパーベルト領域でのランダム運動の成長過程)
- フルシステムで計算
- 実効性能 29.5 Tflops を記録

地球シミュレータの最高速度 (26.6 Tflops) よりも若干速かったのですが、ファイナリストにはなったものの賞はいただけませんでした。

審査委員長の見解:
The committee's charter is to evaluate and reward accomplishment in the field as it relates to advancing real world application execution and to select those valued most highly ...
ということで、性能だけが基準ではないとのこと。

東大外での利用

浜松メトリックス社 (www.metrix.co.jp) による商品版

プロセッサボード 1 枚をホストに直結
ピーク性能 125Gflops — 1Tflops

現在の GRAPE-6 ユーザー

日本
核融合科学研究所、 国立天文台 (11 台、共同利用)、(筑波大学 (8 台))、大阪大学
アメリカ
American Museum of Natural History (5 台), Drexel, Indiana, Michigan, Rutgers
カナダ
McMaster
ドイツ
MPIA Heidelberg
イギリス
Cambridge (Institute of Astronomy)
Durham?

今後の計画

- 天文研究への利用
- FPGA システムの応用
- ホスト更新による性能向上

ホスト計算機更新

性能評価のところでみたように、

現在のシステムではホストの計画速度、ホスト間の通信速度が性能のボトルネック

- ホストの計算性能の向上
- 通信速度の向上

とりあえず、大粒子数での性能向上をめざす

- メインメモリバンド幅
- ネットワークバンド幅

の向上が重要

更新計画

現状:

CPU	AMD Athlon XP 1800+
マザーボード	ECS K7S6A (SiS 745)
メモリ	PC2100 1GB
NIC	NS83820 ベース

現在更新作業中

CPU	Intel P4 2.53GHz (OCed to 2.85GHz)
マザーボード	Iwill P4GB (Intel E7205)
メモリ	Dual PC2700 (300MHz) 1GB
NIC	NS83820 ベース + オンボード

メモリバンド幅 2.25 倍
ネットワークバンド幅 1.5 倍？

次世代システムの検討

- GRAPE-6 の教訓
- 次世代システムの構想

GRAPE-6 の教訓

- 重力多体系の直接計算+独立時間刻みのための専用計算機としては演算方式、ネットワークアーキテクチャはほぼ完成
- ハードウェア開発はかなり大変であった
- 応用範囲があまり広くない
- 商業化はまあまあ成功

ハードウェア開発の大変さ

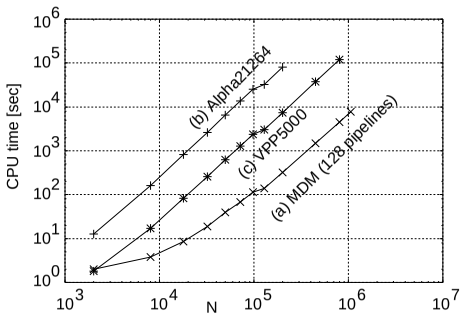
- 専用チップに通信インターフェース、メモリインターフェース、制御回路を集積
  - サイズ的には問題ない
  - 実装も容易にはなる
  - 設計は大変、、、
- GRAPE 側に複雑な (というほどではないが) ネットワークを用意
  - 並列システムで高い性能を実現した
  - 汎用ネットワークの高速化を予見できなかった
  - ネットワーク部分の開発はかなり大変
  - 1-2年で陳腐化 (不要になる)

応用範囲

- 重力計算
  - 直接計算は OK
  - ツリー、FMM、ネイバースキーム等には向かない
- 重力計算以外 — 例:境界要素法
  - ラプラス方程式なら OK
  - ヘルムホルツ問題は扱えない
- 他の相互作用は (近距離力でも) 扱えない (FPGA、、、)

BEM 解析の例

高橋 (2002) 理研 MDM 2 ボードでの解析
ラプラス方程式境界値問題



商業化

GRAPE-3, 5, 6: システムの製造・販売は小さなベンチャー企業に委託
○ 安価
× 海外のユーザーへのサポートが難しい (英語力の問題)

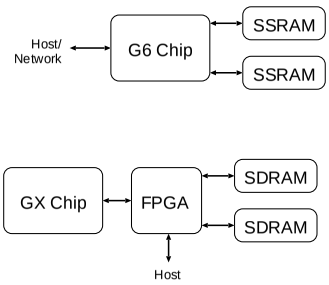
理研 MDGRAPE-2: IBM が製造・販売

次世代システムの構想

目標

- 半導体技術の進歩に見合った速度向上を実現する
- ある程度応用範囲を広げる
- 開発するものを減らす (陳腐化するものは作らない)
- 開発期間を短くする

具体的には、、、



上:現在の GRAPE-6。
メモリインターフェース等をチップに集積

下: 次世代
専用チップは純粋に相互作用計算だけ
大規模 FPGA (1Mゲート程度) に
メモリインターフェース、ホストインターフェース、その他を内蔵

FPGA を持つことのメリット

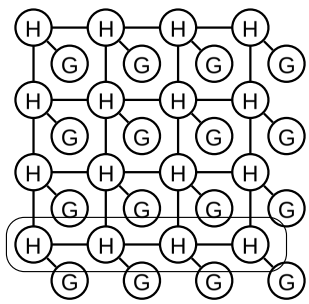
- 高度な (複雑な) アルゴリズムに対応可能
  - ツリーコードのより効率的な実装
  - ネイバースキーム
- 開発の容易化、リスク削減
  - バグの原因となる制御回路、インターフェースは FPGA に任せる
- 多用途化 — 短距離相互作用なら FPGA 側で評価

ネットワークについて

基本方針

- ネットワークハードウェアは汎用のもので済ます
  - 進歩が速い
  - 非常に安価
- ソフトウェアは必要ならチューニング

具体的には



並列化のために必要なネットワークはホスト側で構成

ホスト 1 台に付ける GRAPE は小規模 (4-8 チップ程度、 PCI-X/PCI-Express カード)

まとめ

- GRAPE-6 の重力計算部分は 64 Tflops ピーク性能のハードウェアが完成した。
- 実効性能でも 29.5 Tflops と、ベンチマークではない実際の科学技術計算では世界最高速を達成した。
- 商業化も順調に進み、国内外で利用されている
- 次世代システムの検討も進んでいる