MD シミュレーションと アクセラレータ・専用計算機

牧野淳一郎

天文シミュレーションプロジェクト

理論研究部

国立天文台

話の概要

- はじめに: MD シミュレーション用専用計算機の歴史
  - Delft Molecular Dynamics Processor
  - FASTRUN
  - GRAPE-2A, MD-GRAPE, MDM amd PE
- ANTON
- ANTON とそれまでの「MD 専用計算機」の違い
- 今後の方向

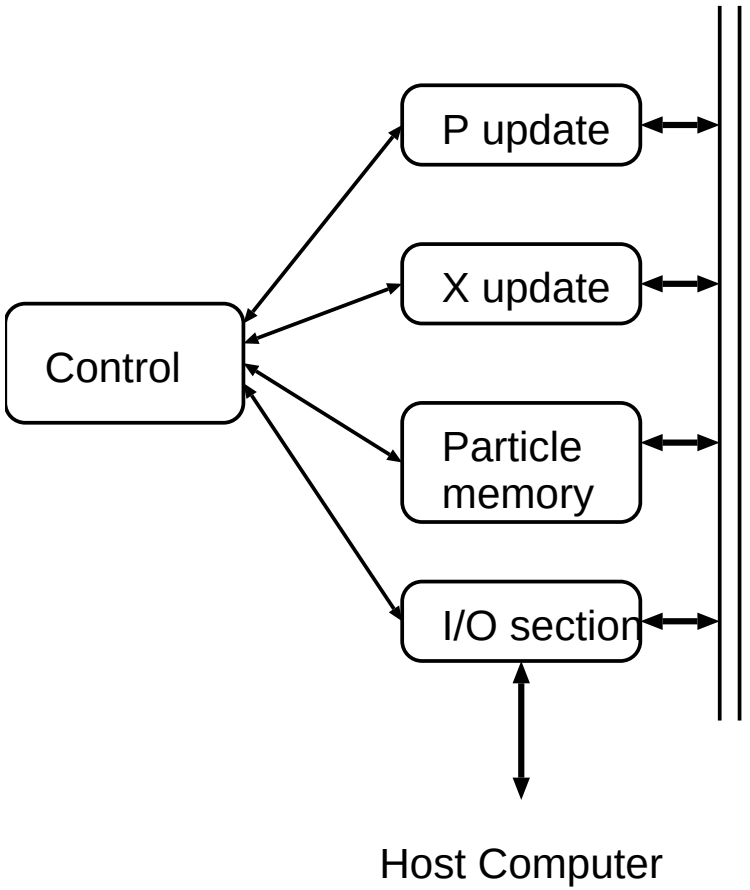
MDシミュレーション用専用計算機の歴史

- MD専用計算機の歴史は結構古い
- Delft Molecular Dynamics Processor: 1980頃完成
- しかし、(分子動力学計算の中だけ見ても)主流になったことはない

歴史をみながら、何故かを考えてみる。

Delft Molecular Dynamics Processor

デルフト工科大学の D. Bakker らが 1980 年頃に完成



LJ ポテンシャルで相互作用する単原子分子の MD 専用機

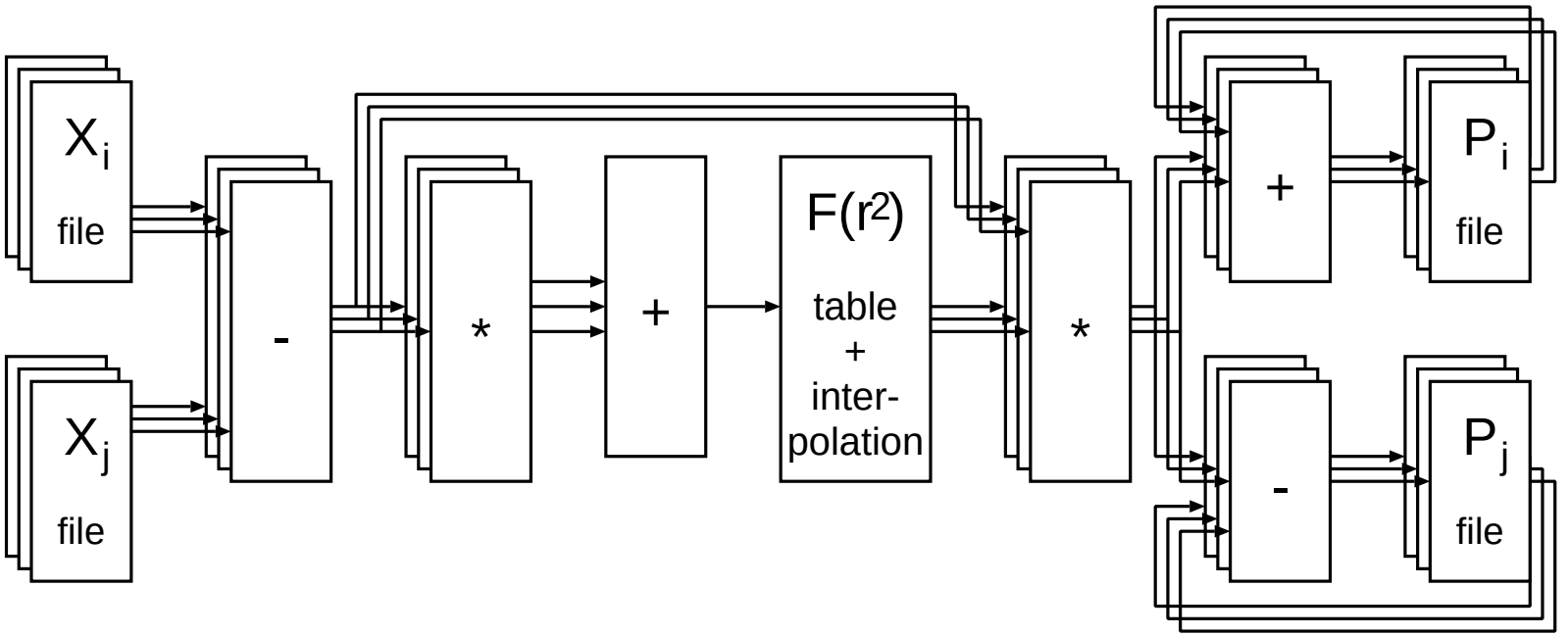
「座標アップデート」(leapfrog での積分)、「運動量アップデート」(加速度計算) それぞれに専用パイプラインプロセッサ

ホストはミニコン、シミュレーションはホストとは独立に走る

CMOS IC を使った 20 枚くらいのラッピング基板で構成。

牧野は 1990 年に見せてもらったことがある

P update

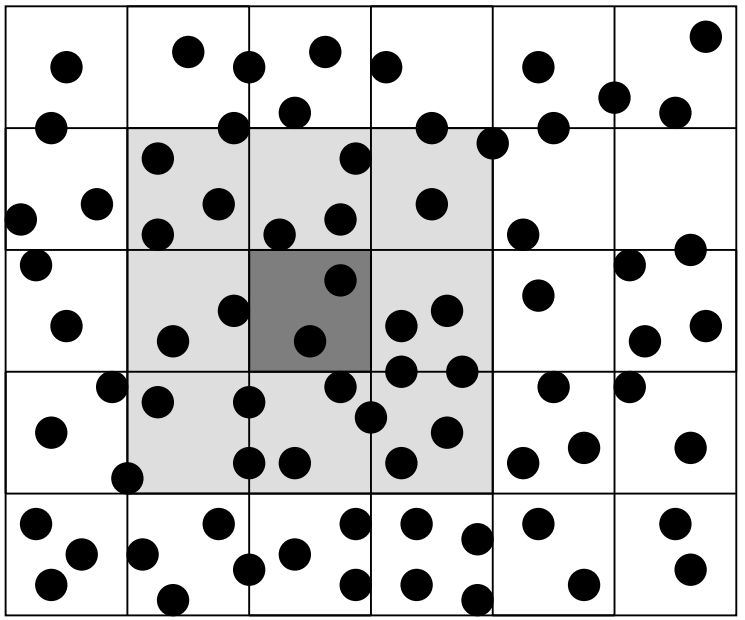


GRAPE と同様、2 粒子の座標をいれて相互作用をだす

P-update パイプラインの詳細

- 入力座標は 24 ビット固定小数点
- 距離の 2 乗は 32 ビットを保持、上位 10 ビットをテーブルにいれて一次補間して力をだす
- linked-list (cell-index) アルゴリズムをハードウェアで実装、隣接セルの粒子間の力を対称性を利用して積算

linked-list アルゴリズム



(単純なバージョン)

- 計算領域を一辺が相互作用到達距離の立方体に分ける
- ある立方体の中の粒子への力は、それを囲む $3^3 = 27$ (自分含む) の箱の粒子の寄与だけ考える

計算量は無駄に多いが、ハードウェアでの実装は (シリアルパイプラインなら) 容易

DMDP をどう評価するか？

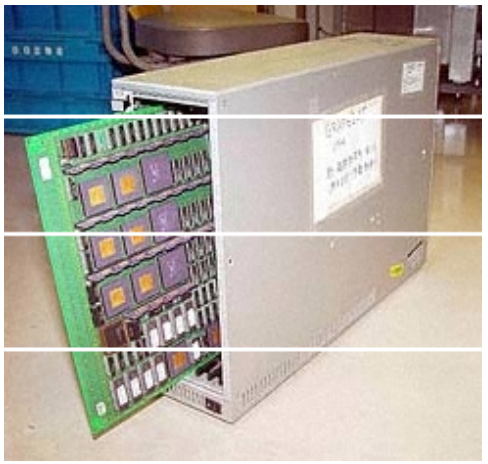
- 性能は素晴らしかった。
  - 性能は 160Mflops 相当: Cray-1 の理論ピーク程度
  - コストは 1000 万円以下くらい (Cray-1 の 1/100 以下)
- どれくらい科学研究に使えたかはあまり資料がない、、、
  - 単原子分子でそんなにできることはない
  - 結果解析もハードウェア追加しないとできない
  - 計算精度がちょっと微妙な気がする

FASTRUN

Fine *et al.* **1991**

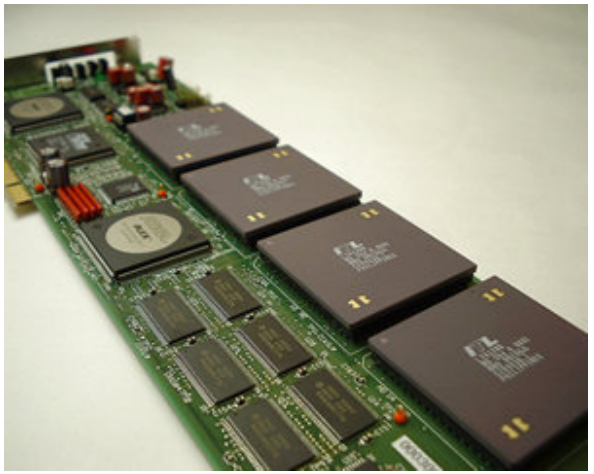
- **コロンビア大学と** BNL **の共同開発**
- **タンパク** MD **用**
- **コスト、速度は** DMDP **と同程度 (10 年たってるのに、、、)**
- **相互作用計算部分のみハードウェア**
- **ネイバーリストをハードウェアで実装した (らしい)**

GRAPE-2A



- 1991 年に開発開始。駒場の杉本グループと、筑波 (当時) の永山グループの共同開発
- DMDP の相互作用パイプラインだけを取り出して汎用計算機につないだようなもの
- 92 年には完成、費用 (人件費以外)100 万くらい、200Mflops
- 当時のワークステーションはまだ 10Mflops くらいだったのでまあ速かった

MD-GRAPE



- 93 年くらいに開発始めた
- 駒場の杉本グループと画像技研 (株) の共同研究。都からの研究費で
- 基本的には GRAPE-2A をカスタム LSI 化、4 チップのせたボード開発。
- エバルド法用の DFT パイプラインにもなる設計 (泰地による)

4 Gflops の性能。

この他に、富士ゼロックス+大正製薬で「MD-Engine」というのも。

MDM と PE

- 戎崎が理研に異動したあと、理研でスタート
- MDM は 2001 年くらいに完成、75Tflops
- PE は 2006 年に完成、1Pflops
- どちらも、基本的には MD-GRAPE の大規模並列化
- どちらも 1 万チップ程度の巨大システム、ホスト計算機も数十台。インフィニバンドで並列化。

ここまで振り返ると、、、

- 分子動力学用専用計算機 (MD-GRAPE とその後継含めて) は性能 (少なくとも価格当りの計算速度) は高い
- でも、どうサイエンスの役に立ったかは色々意見もあるかもしれない

MDM、PE の「問題点」

- 10 万原子では性能でない = 小さい系の長時間計算にはむかない
- Direct Ewald だけなので計算量が粒子数の 1.5 乗で増える (PME をホストでやればいいが、、、)

ANTON

D. E. Shaw の個人研究所「DESRES」が開発。

D. E. Shaw って何者？

Financial Times 2010/3/8 の記事から:
DE Shaw broadens Asian reach

DE Shaw, the $24bn hedge fund founded by mathematician David Shaw, is to open offices in Shanghai and Tokyo as part of an expansion in Asia, according to people familiar with the situation.

The Shanghai office, to house a team of private equity analysts, will increase the group 's presence in the region and mark its first expansion into mainland China. It will focus on acquisition opportunities in China.

モルガン・スタンレーに 1986 年に入る前はコロンビアの計算機科学科のファカルティ。並列計算機 Non-Von の開発を主導。1951 年生まれ。

ARCHITECTURE AND APPLICATIONS OF A HETEROGENEOUS, MASSIVELY PARALLEL MACHINE

David Elliot Shaw

Department of Computer Science, Columbia University, New York, New York 10027

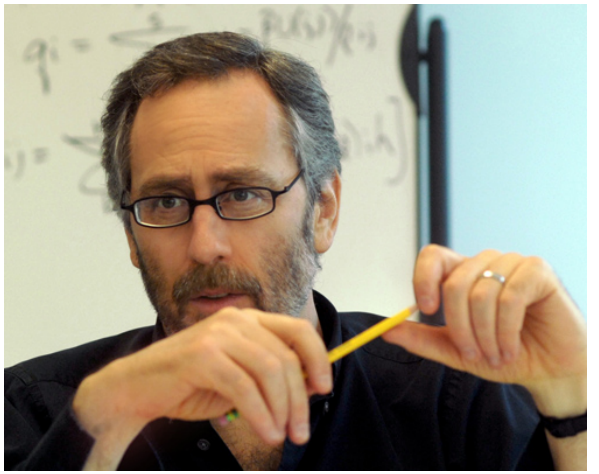
1985 年のレビュー論文

The organization of an experimental, massively parallel machine called NON-VON is described, along with some of the typical artificial intelligence applications for which the machine is intended to provide significant performance and cost/performance advantages over conventional computer systems. The machine incorporates an *active memory*, which is constructed using custom, very large scale integrated (VLSI) chips. Each chip contains a number of simple processing elements and a small number of larger processing elements, each capable of controlling the operation of a subset of the active memory. A simplified, preliminary prototype of the NON-VON architecture is now operational at Columbia University.

Performance projections, derived through detailed analysis and simulation, are summarized for applications in the areas of rule-based inferencing, computer vision, and knowledge base management. The results, most of which are based on benchmarks proposed by other researchers, suggest that NON-VON could improve performance of such tasks by as much as several orders of magnitude, compared to a conventional sequential machine of comparable hardware cost.

D. E. Shaw って何者？

- 元々並列計算機研究者・アーキテクト
- 世界有数のヘッジファンドの創設者
- 個人資産から計算生物学、特にタンパクの機能シミュレーションのための研究所を作った



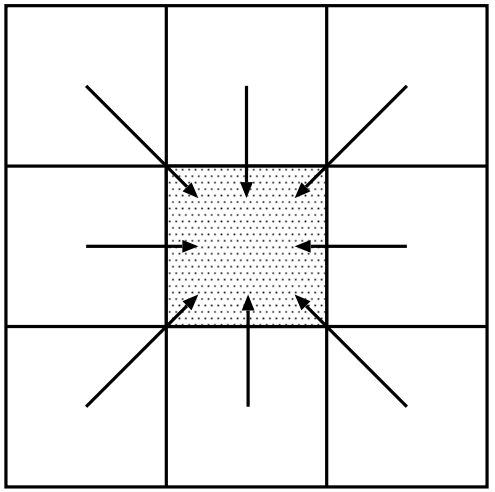
- かなり良い給料と魅力的な仕事でよい研究者を集めた。牧野が知っているところでは Caltech にいた John Salmon (並列ツリーコードで有名) とか

ANTON

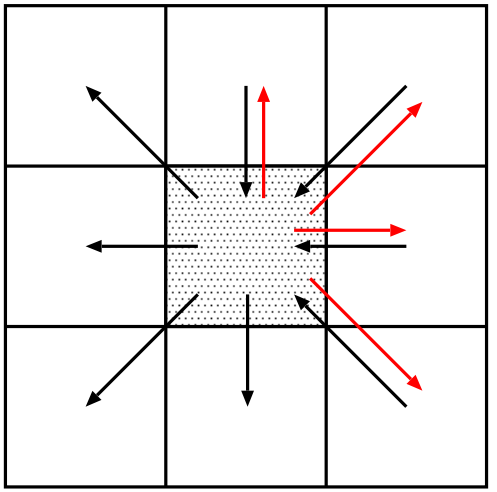
- 数万原子程度のあまり大きくない系で高い性能がでることを目標に設計
- 相互作用計算の新しい並列アルゴリズム (NT 法) を開発、それ用にハードウェアを設計
- 相互作用計算パイプライン+プログラム可能プロセッサ+ネットワークプロセッサを 1 チップに集積

近距離相互作用の並列計算

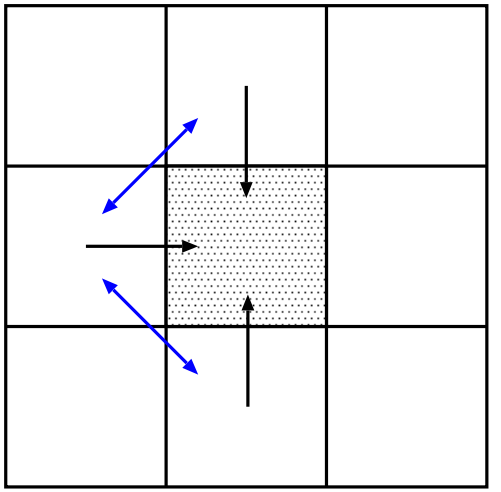
2 次元の場合



対称性を利用しない。周り 8 個から座標をもらって力計算



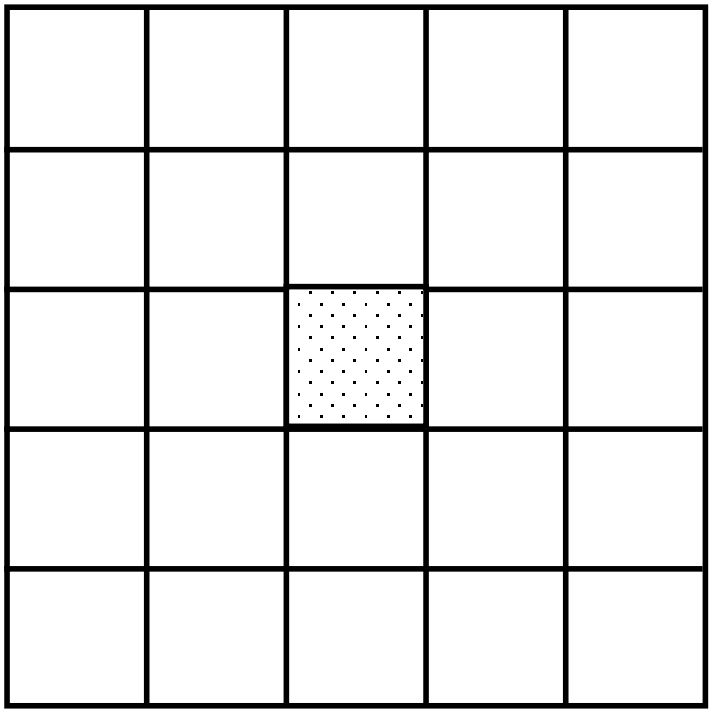
対称性を利用する。4 個からもらって、相手への力は送り返す

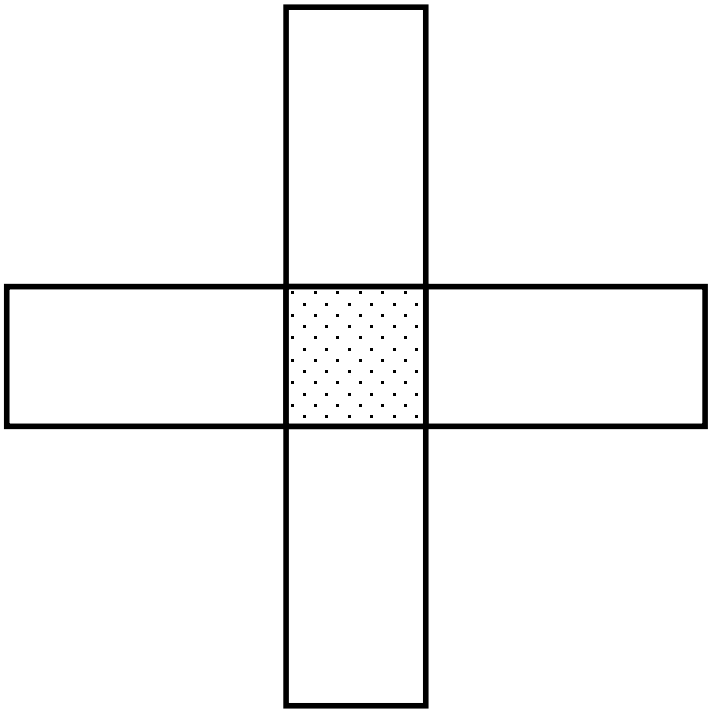


NT 法の考え方。軸方向だけからもらって、斜めの力もついでに計算する

3 次元だと 26 と 6 でだいぶ違う。

基本セルを小さくすると





24 と 6(3 次元だと 124 と 16) で、大きな違い。

漸近的な振る舞い: 問題サイズを固定して、セル数 p が無限大の極限

- 普通の方法: 通信量は一定値に収束
- NT 法:通信量が $p^{-1/2}$ でゼロに収束

ANTON のハードウェアと性能

- 1 チップに DMDP と似たような相互作用パイプライン 32 本、800MHz 動作、1Tflops くらい
- 512 チップを 8^3 トーラスネットワークに接続
- 速度は 500TF くらい、専用機としてはそんなに速くない
- 2 万原子の系で 1 日に 10 マイクロ秒を実現。(1 ステップ数十マイクロ秒)。「京」でできるより 100 倍速い。

何が高速化に効いているか？

- 通信量、通信回数を減らす新しいアルゴリズム
- プログラム可能部分、ネットワークも独自開発することでの通信レイテンシの削減
- その他計算量を減らすための沢山の細かい工夫

ANTON とそれまでの「MD 専用計算機」の違い

- 原子数少ないところで性能をだすことを目標にした
- そのために新しいアルゴリズムを開発した
- 元々汎用並列計算機アーキテクトだったので、まともに複雑なものを作った

これまでの MD 専用計算機では新しいアルゴリズムという話はあまりなかった、、、

今後の方向

- ANTON の真似では勝てないので、やるならアルゴリズムの改良とか新しいアルゴリズムの開発からやらないと駄目。
- ハードウェアの物量がものすごく大事、というわけでもないので、FPGA や構造化 ASIC でもなんかできないとも限らない。プロトタイプは 3 千万円くらいでできる。
- 目標は結構明確: 1 ステップを 1 マイクロ秒あたりまで短くして、1 日にミリ秒くらい動かす。
- 計算速度は Pflops オーダーでいい。

1ステップ1マイクロ秒は可能か？

- 最大の問題は通信回数。
  - 回数という観点では NT 法はあんまり良くない
  - トーラスの上で FFT をしているのも問題
  - SHAKE?
- 相互作用計算のパイプライン遅延も無視できない。

FFT

- サイズは小さい。せいぜい 32^3 で、単精度なら $128\mathrm{KB}$。
- 演算量も大したことはない。$\mathrm{GPU}\ 1$ つでもできる。
- とはいえ。マイクロ秒で通信するなら $256\mathrm{GB/s}$ のグローバルな通信速度が必要。
- 時間方向に予測をいれてデータ圧縮とかできるかもしれない。

これは何か他の役に立つか？

粒子で 10^{10} ステップくらい計算したい話、というのは天文 (惑星関係) には色々ある。

- 土星リング
- 惑星形成

流体でもできたら嬉しい。

もうちょっと広い視点から見ると

「京」のようなペタスケール・エクサスケールシステムの問題

- **実は、全体使って性能がでるような問題は殆どない**
- **これは、コア数が多いから、というよりは通信レイテンシの問題。**
- **通信はノード間だけでもなく、同一チップ内のコア間でもマイクロ秒オーダー**
- **そもそも主記憶アクセスが数百ナノ秒かかる**

改善の方法

- x86 や GPGPU では改善は難しい。プロセッサコアから自分で作るなら改善は簡単だけど、、、
  - コンパイラ、アプリケーション、その他をどうするか？
  - できるけど時間と人が必要
  - コアは買ってきて改造する方法もある (ANTON もそうしている)
- メモリがチップ当り数メガないし数十メガバイトでよければできる。
  - 1000 チップも使えば数十ギガバイト、大抵の計算はできる？

まとめ

- MD シミュレーション用専用計算機の歴史を概観した
- DESRES の ANTON は、比較的小さなタンパクのシミュレーションの高速化のために様々な新しいアイディアをいれて、汎用並列計算機の 100 倍程度を実現した
- もう 1 桁あげるのはできなくもないかも (DESRES でもなんか考えてるとは思います)
- 構造化 ASIC なら 3000 万円くらいで何か作れなくもない