梅村さんと GRAPE と私

牧野淳一郎

研究会「宇宙天体形成史〜初代星から生命の起源まで〜」2023/9/2

梅村さん、定年おめでとうございます!!!

なにがめでたいかって、無事に定年まできたことです — Prof_hrk

と、いわないといけないことはこれだけなので、もうおしまいにしてもいいわけですが、もうちょっと何か。

話の構成

- 梅村さんと GRAPE と私
- その後の梅村さんと私
- その後の GRAPE と今とこれから

梅村さんとGRAPEと私

1 ☐ 1990Natur.345...33S 1990/05cited: 292

A special-purpose computer for gravitational many-body problems

Sugimoto, Daiichiro; Chikada, Yoshihiro; Makino, Junichiro *and 3 more*

2 ☐ 1991PASJ...43..841F 1991/12cited: 34

GRAPE-1A: Special-Purpose Computer for N-body Simulation with a Tree Code

Fukushige, Toshiyuki; Ito, Tomoyoshi; Makino, Junichiro *and 3 more*

3 ☐ 1993PASJ...45..311U 1993/06cited: 27

Smoothed Particle Hydrodynamics with GRAPE-1A

Umemura, Masayuki; Fukushige, Toshiyuki; Makino, Junichiro *and 4 more*

実は共著論文は3本しかない

(もう 1 つ情報処理学会論文賞もらったのがあったような気がする須佐さんの話の HMCS の)

最初の論文:

A special-purpose computer for gravitational many-body problems

Daiichiro Sugimoto*, Yoshihiro Chikada†, Junichiro Makino*, Tomoyoshi Ito*, Toshikazu Ebisuzaki* & Masayuki Umemura‡

* Department of Earth Science and Astronomy, College of Arts and Sciences, University of Tokyo, 3-8-1 Komaba, Meguro-ku, Tokyo 153, Japan
† Nobeyama Radio Observatory, Minamimaki-mura, Minamisaku-gun, Nagano 384-13, Japan
‡ National Astronomical Observatory, Mitaka, Tokyo 181, Japan

A processor has been constructed using a 'pipeline' architecture to simulate many-body systems with long-range forces. It has a speed equivalent to 120 megaflops, and the architecture can be readily parallelized to make teraflop machines a feasible possibility. The machine can be adapted to study molecular dynamics, plasma dynamics and astrophysical hydrodynamics with only minor modifications.

(very-large-scale integration) chips. Our special-purpose computer is constructed relatively easily using such chips. The problem is to develop an architecture that allows both ease of construction and sufficient flexibility to simulate a wide variety of physical systems. The N-body problem is in fact well suited to these apparently conflicting demands. The simplicity is reflected in the computer's cost effectiveness; our prototype machine GRAPE-1 achieved a performance comparable to the CRAY-XMP/1 at a cost 10,000 times lower.

The pipeline architecture

The N-body problem is described by the equation

GRAPE の最初の論文

- もちろん、コンセプトと、実際にハードウェア作って速いのができたよ、がメイン。
- GRAPE-1: いわゆる「20 万円でできたコンピュータ」
- 杉本さんが、拡張するコンセプトを 2 つ書いた
  - 衝突系 (球状星団) 用 (独立時間刻み)
  - 重力+SPH 用
    ここが梅村さんの話
  - 時期的には Hernquist and Katz の TREESPH のちょっとあと
  - 当時の梅村さんの SPH は観山さん系列？天文台理論部が池内、観山、梅村だった頃。

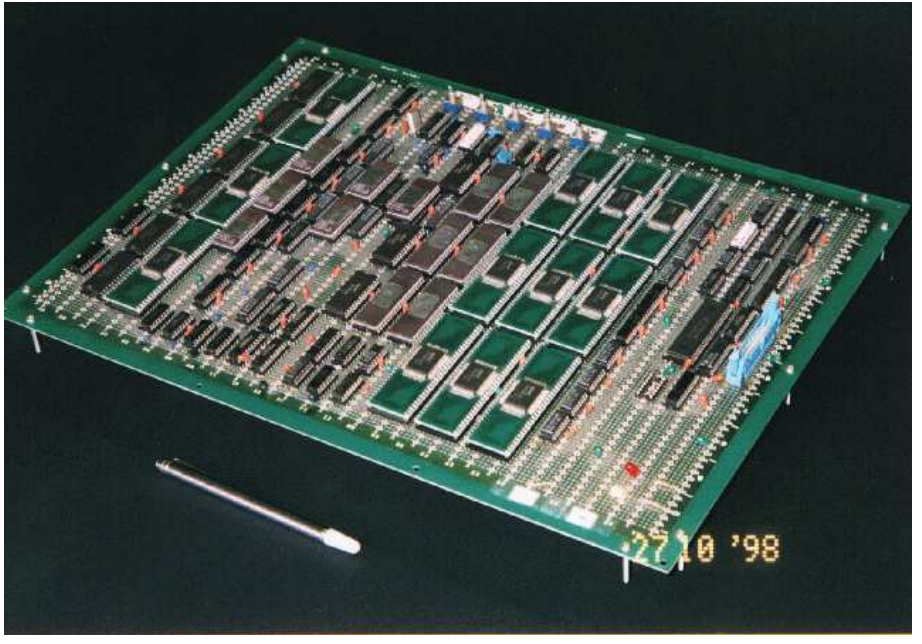
Nature 論文から

Short-range forces need be calculated only for pairs of particles lying within the sphere of the force's effective range. Moreover, the form of the force law depends on the problem considered. Therefore, it is better calculated in the programmable host machine. To do this, however, we still have to identify which particles are close neighbours. This involves constructing a list of neighbour particles, which requires a calculation of $O(N^2)$ operations. But the GRAPE hardware determines inter-particle distances during the calculation of the long-range forces, and can therefore generate such a list (with only small changes to the GRAPE hardware) and send it to the host. We call such a machine GRAPE(nl).

GRAPE での重力+SPH 計算(当時の考え)

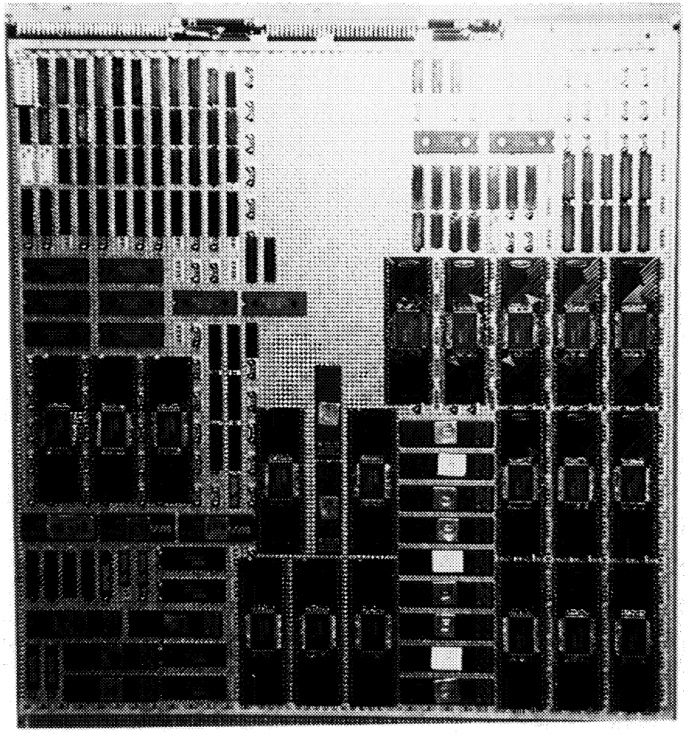
- 重力は $O(N^2)$ だけど GRAPE ですごく速くなる (tree + GRAPE もぼちぼちやってましたが GRAPE-1 は使えない)
- SPH 計算は近傍粒子だけなので $O(N)$ だけど、近傍粒子を見つけるのは単純な方法では $O(N^2)$ でこっちが大変
- GRAPE はどうせ $O(N^2)$ で粒子間距離全部計算するから、それにコンパレータとカウンタとメモリつければ近傍粒子リストがつくれる。演算パイプラインよりハードウェアは小さい。
- GRAPE から帰ってくるデータは何倍かに増えるけど、まあ許容範囲

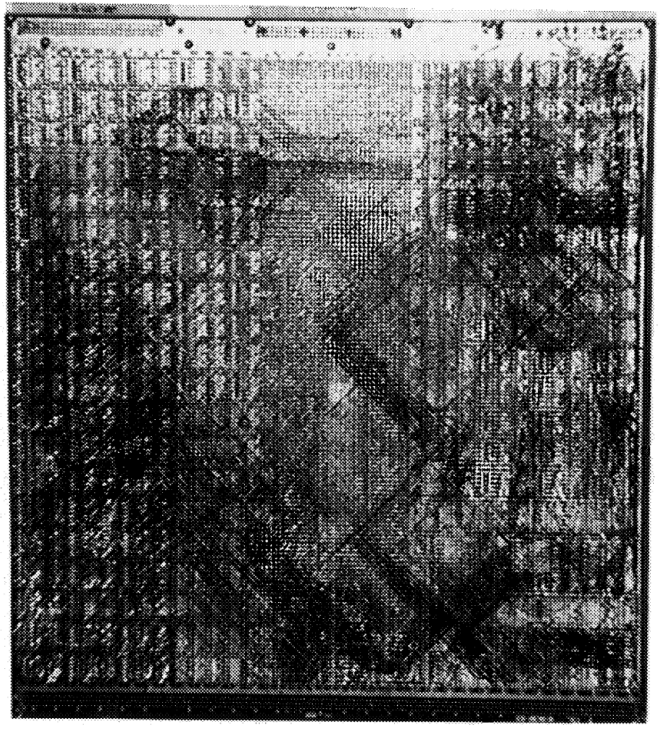
GRAPE-1(1989)



当時 M1 の伊藤君 (2023 年現在千葉大工学研究院長) が作った。

GRAPE-1A(1990)



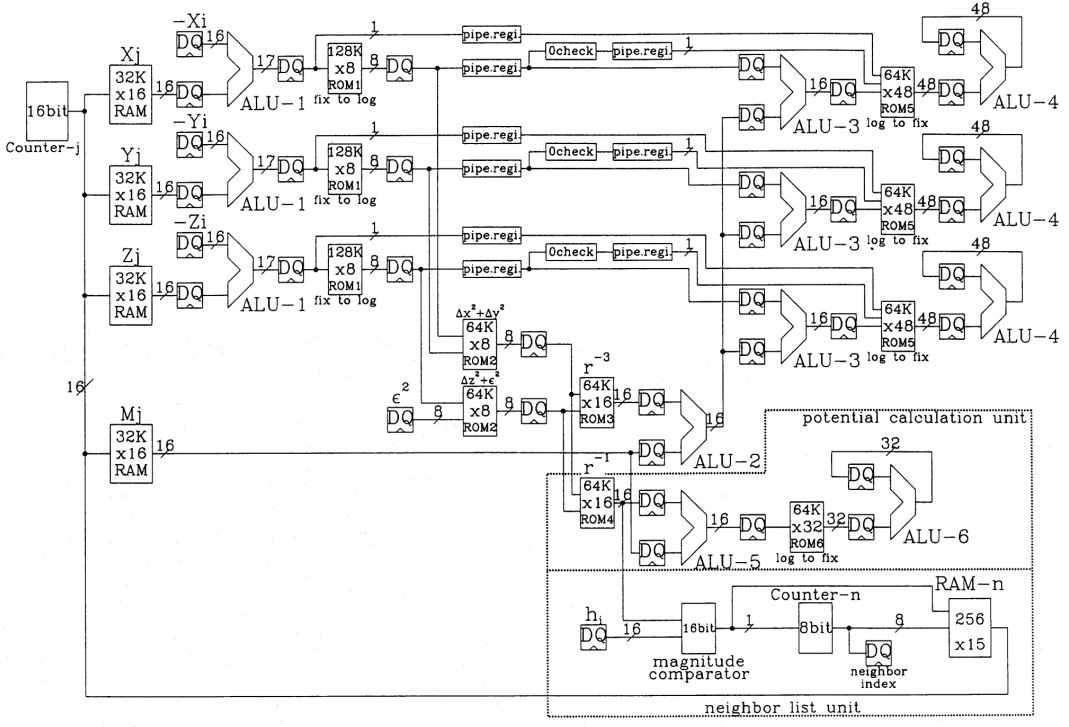


論文 (PASJ 1991, 43, 841) から

1と1Aの違い

- インターフェースが GP-IB から VME に変わって 100 倍くらい速くなった
- 粒子に質量がついた
- 1ビット精度ふやした
- エネルギー計算、近接粒子リストハードウェアを追加

GRAPE-1A パイプライン



ちなみに

- パイプラインとかの大体の設計は牧野がやったかも
- 詳細設計は福重君 (B4) と伊藤君。配線は福重君
- 1 号機は梅村さんの科研費で部品調達。三鷹に設置
- 2 号機は野本さんがお金だして大野君 (現在理研) が製作
- 論文によると SPH コードで VP-200 より速かった。

GRAPE-1A 使った計算の論文:
Early Cosmic Formation of Massive Black Holes, Umemura, Loeb, Turner, 1993 とか。(須佐さん、大向さんの話でも登場)

その後の梅村さんと私

- 梅村さんは92年に在外でプリンストンに
- 私なんかでプリンストンにいった時に梅村さんと確か「サクラ」で寿司食べたような気が。
- 梅村さんは帰国してわりとすぐに筑波に異動。輻射流体の研究グループを創始。
- 東大 GRAPE グループは92年から96年まで GRAPE-4 の開発とそれ使った計算。
- 97 年から未来開拓「計算科学」次世代超並列計算機開発を、筑波大では「連続体向け超並列計算機の開発」、東大では「多粒子系向け超並列計算機の開発」 – GRAPE-6 から FIRST(特別推進) に。

その後の梅村さんと私(続き)

- その後、2009あたりから「京」の戦略分野、ポスト「京」萌芽的課題で宇宙・惑星分野の提案を梅村さん、松元さん、牧野他で取りまとめ。富岳成果創出(I)では梅村さん手をひいて、今年からのIIでは大須賀さんが代表。

その後の梅村さんと私(続き)

- その後、2009 あたりから 「京」の戦略分野、ポスト「京」萌芽的課題で宇宙・惑星分野の提案を梅村さん、松元さん、牧野他で取りまとめ。富岳成果創出 (I) では梅村さんは手をひいて、今年からの II では大須賀さんが代表。

30 年以上にわたって、大変お世話になりました。ありがとうございます。

その後の GRAPE と今とこれから

- GRAPE-6 以降、専用計算機で大きなシステム、というのはできてない
- これは、カスタム LSI 作る費用があまりに巨額になったから。90 年代から 2023 年の現在まで、10 年で 6 倍くらい。30 年間で 200 倍。
- 90 年代前半は 2000 万くらいだったのが今は 40 億とか。
- じゃあ「汎用」で効率よいの作れないか、ということで牧野は東大の平木さんと GRAPE-DR (振興調整費)。一応 2010 年に Little Green 500 で 1 位とかになるくらいの成果はあり。
- その流れで 「京」の時にシステム提案、富岳では筑波大朴さん、佐藤さんのグループでアクセラレータ検討、富岳開発では副プロジェクトリーダー、ポスト富岳検討も 1 グループ代表とかを、、、
- PFN と一緒に深層学習向けプロセッサも。

というわけで、ポスト富岳検討の話をちょっと。

計算機アーキテクチャの世代交代とそれが起こる理由

あまり聞いたことがない人が多いかもしれないが「ベルの法則」(ゴードン・ベル賞のベル) というのがある:

Roughly every decade a new, lower priced computer class forms based on a new programming platform, network, and interface resulting in new usage and the establishment of a new industry.

スパコンだと

- 1976 年まで: スカラー計算機
- 1976 年から 1992 年まで: ベクトル (共有メモリ並列含む) 計算機
- 1993 年から 2008 年まで: (マルチコア含む) マイクロプロセッサの分散メモリ並列計算
- 2008 年から: GPU アーキテクチャのアクセラレータとマイクロプロセッサの組合せ

大体 15 年。GPU の次はまだでてきてない。
なお、これは 1988 年に近田さんがいったことと同じ。

近田さんのお言葉

〔昨日，今日，明日〕

先ず，Ｍ１，Ｍ２の若い人達に話す時の定石の昔話から始めます．約２０年前の１９６０年代の末，三鷹の天文台に口径６ｍのミリ波望遠鏡が作られ，後の野辺山の望遠鏡の雛形が生まれた時，その制御とデータ収集はOKITAC4300と言うミニコンが受け持ちました．そして今，６ｍ鏡は水沢に運ばれてＶＬＢＩに使うために化粧直し中．今度の制御用計算機は日電のパソコンPC9801VMです．表１に見るように，この２０年間に記憶容量も演算速度も約４桁良くなっています．つまり，同じ予算でも１０年たてば１００倍の仕事をさせられるようになったと言うことです．これは天文台と言うローカルな世界だけのことではありません．事実，図１〔１〕の並列計算機の並列度の推移にみるように１０年で１００倍というのはかなりよいestimateです．但し，これは重要なことなのですが，同一の構造，同一の使われ方，つまり同一の製品系列の中での進展はこんなに早くありません．例えば野辺山の汎用計算機についてこの４年間の価格／性能比を見るとせいぜい２倍しか良くなっていません．つまり計算機（の技術）の使い方を常に見直して，自分の問題に適した構造の機械を使う（作る）ようにし，かつその上で解き易いように問題を立て直すことがないと「１０年で１００倍」を生かした人に比べ１０倍分は立ち遅れることになります．

アーキテクチャの変化が必要な理由

基本的な理由: デバイスの質的、あるいは量的な変化によって設計の暗黙の前提が成り立たなくなる。

- スカラー計算機の暗黙の前提: 使えるトランジスタの数は完全パイプラインの乗算器をつくれるほどない。主記憶は磁気コアでトランジスタよりずっと遅い
- ベクトル計算機の暗黙の前提: 1 チップにはいるトランジスタの数はプロセッサをつくれるほどではない。メモリは SRAM や DRAM でプロセッサの論理並みに速いが容量は小さいので沢山使う (=メモリバンド幅は高い)
- マイクロプロセッサシステムの暗黙の前提: 1 チップにある程度のコアははいるが、そんなに滅茶苦茶沢山ではなく、キャッシュコヒーレンシを維持してもそれほどハードウェアのオーバーヘッドは大きくない

この辺の暗黙の前提は、もちろん、それを前提としないことによって性能を飛躍にあげたものがでてこないと多くの人には理解されない。

では GPU 設計の暗黙の前提はなにか?

GPU が メニーコア CPU に比べて電力・面積性能的に有利な理由:
キャッシュコヒーレンシは捨てて、 x86 との互換性も無視して非常に沢山の演算コアを詰め込んだこと。
(逆にいうと、Intel の「GPU」が決して成功しないのは、この2つを投げ捨てられないから)

「牧野の私見としては」GPU の制約になっているものは以下の2つ

- 階層キャッシュまでのデータの (チップ内) 横方向の移動
- 外部メモリまでのデータの (チップ外) 横方向の移動

何故横方向のデータ移動が制約になるか？

配線遅延・配線消費電力の問題
微細化しても、「単位長さ辺りのキャパシタンス」はかわらない、ところが、「単位長さあたりの抵抗」は配線幅の二乗に反比例して増える:

- **配線長がスケールしても配線遅延は「デザインルールに反比例して」増える**
- **配線長一定だと遅延は二乗で増える。電力は減らない**

数字をいれてみると

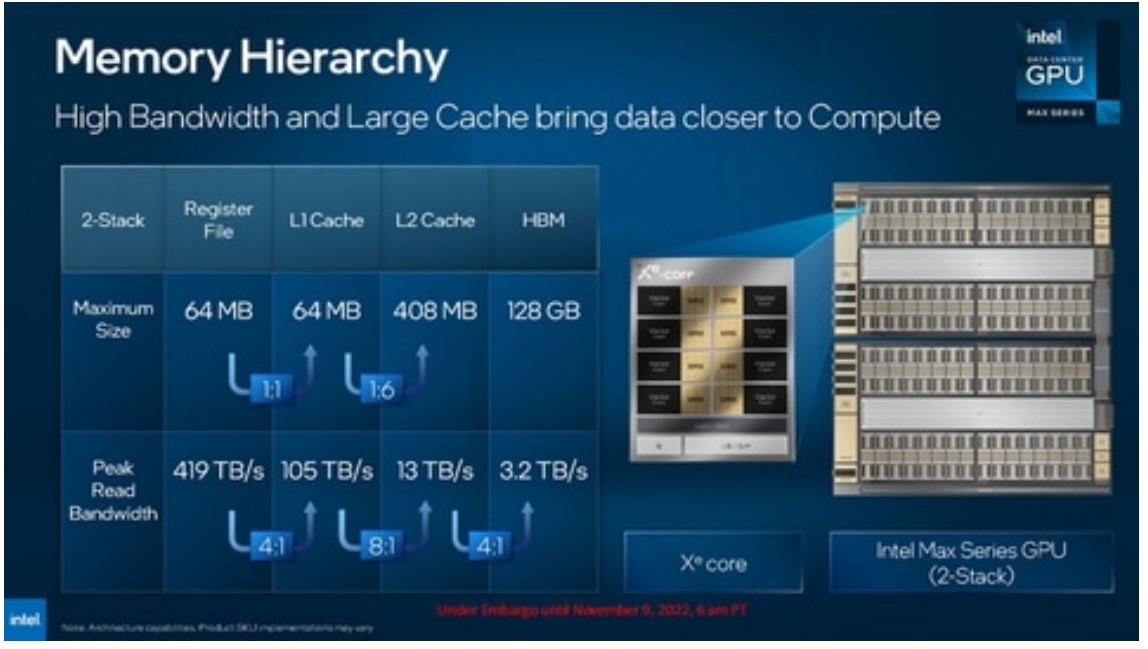
- 10cm ($10^5 \mu\mathrm{m}$) の配線は 20pF のキャパシタンスがある。これは微細化してもかわらない (線と平面の間のキャパシタンス) 電圧 1V だと、この線は 10pJ/bit を消費する。
- DDR5 メモリは 20pJ/bit くらい。電圧 1.3V とか。。
- LPDDR5 と GDDR6x: 10 pJ/bit くらい。配線が DDR5 のモジュールより短くて電圧も低い。
- HBMx: 3-4 pJ/bit。概ね 25mm くらいまで配線短くなっている。

3pJ/bit は 64 ビットワード移動に 0.2nJ なので、HBM でも 5GW 移動すると 1J 消費する。 B/F=4 の計算機作ると 10GF/W しかでない。 0.1 にしても DRAM メモリだけで 400GF/W までしかいかないのでキャッシュの分とかいれると 100GF/W あたり。

つまり

GPU の次のアーキテクチャに移行するためには、「データの横方向の移動」を限りなく小さくする必要がある。

Intel GPU の例



B/F の値
レジスタ: 8? (普通 16 いる)
L1: 2
L2: 0.25
メモリ 0.06
A464fx に比べても数字が小さい
それでも電力消費が大きい原因になっている

必要な方向

- 基本的には、DRAM をプロセッサダイの上(下かも)につむ。HBM でやってることと変わりはない。
- 但し、 HBM では DRAM ダイの中心に集中しているのは配線を、もうちょっと細かい単位でだす必要がある(横方向の移動を減らす)
- プロセッサアーキテクチャ自体も、演算器の物理的に近くに分散したメモリがある形に移行する必要がある。
- 90 年代の SIMD 超並列計算機(CM-2、MasPar とか)に類似。
- ポスト富岳 FS の神戸大学チームはこういう方向を検討している

まとめ

- 30 年以上にわたって大変お世話になりました。ありがとうございます。
- 梅村さんが始めた、GRAPE を重力計算以外に使う、というアプローチは、色々形を変えながら今も続けてます。
- ポスト富岳では、GPU の「次」のアーキテクチャが必要になります。今日はその話をちょっとしました。