専用計算機の将来——GRAPEの挑戦

牧野淳一郎 (東京大学)

平成 14 年 10 月 4 日

1　はじめに

本稿では、我々が開発してきた多体問題専用計算機 GRAPE (GRAvity PipE) に代表される、特定問題向け専用計算機について、その現状と将来の方向をまとめたい。なお、1999 年 5 月号にも「GRAPE プロジェクト」という題で書かせてもらっているので、一般的な背景やその時点までのプロジェクトの歴史についてはそちらも参照していただけるとありがたい。

2　GRAPE-6

まず、完成したばかりの GRAPE-6 のことから話を始めよう。GRAPE-6 (図 1) は重力多体問題に特化した専用計算機システムであり、名目ピーク性能で 64 Tflops と原稿執筆時点で科学技術計算用計算機として世界最高速を実現した。開発のための総予算は 5 億円強、開発期間は 5 年間である。

まずは同程度の性能をもつ汎用スーパーコンピューターと比較してみることにする。現時点で単体の汎用計算機として最高速を実現しているのは今年 3 月に完成した地球シミュレータであり、ピーク速度は 40 Tflops である。開発費は 400 億程度と言われている。アメリカで開発中の ASCI Q もピーク性能、コストは同程度である。また、文部科学省はグリッドコンピューティングによって高いピーク性能を実現するプロジェクトを計画中であり、これは予算 700 億、完成は 5 年後で 速度は 300 Tflops と報道されている。同じ価格の計算機の性能は 5 年で 10 倍というのが普通なので、700 億、5 年後なら 700Tflops はいかないと計算があわないような気もするが、これはソフトウェア等の開発費もみこんでいるのではないかと想像される。

いずれにしても、これら汎用スーパーコンピューター (あるいは超並列/グリッド) と比べた場合、 GRAPE-6 は 100-300 倍程度高い価格性能比を実現していることがわかる。

少し別の観点から考えてみよう。現時点でもっとも価格性能比が高い汎用計算機を入手しようとするなら、最適な解はインテル P4 か AMD Athlon XP の最高クロックよりもちょっと遅いものを使ったパソコン 1 台である。例えば 1.6 GHz の Athlon XP (1900+) であれば、名目ピークとしては 3.2 Gflops の機械が 8 万円程度で購入可能であろう。まあ、理論的には 2 演算が並列に実行できることもあるとはいえ、実際には滅多にそんなことはないのでピーク 1.6 Gflops と数えておく。これは 1 Tflops 当り 5000 万になり、 64 Tflops だと 32 億円ということになる。

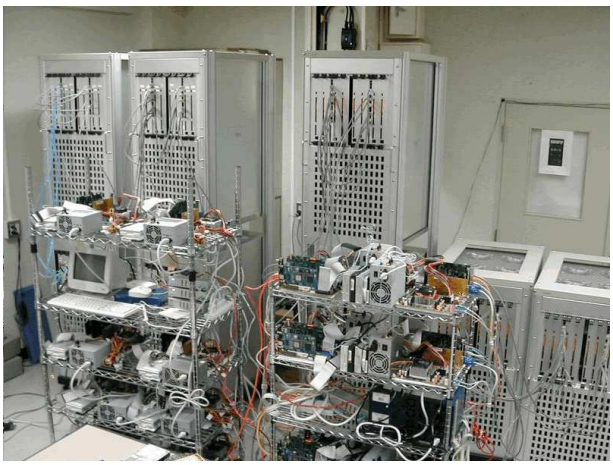

図 1: GRAPE-6。

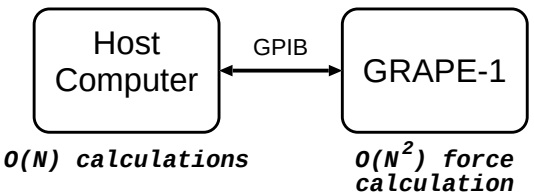

図 2: GRAPE の基本的構成

これは ASCI Q や地球シミュレータにくらべて 1 桁程度安いが、それでもまだ GRAPE-6 に比べると 1 桁程度高くなっている。パソコン 1 台と地球シミュレータの名目の価格性能比の 10 倍の違いに意味があるかどうかは後で議論するが、いずれにしても GRAPE-6 の価格性能比が卓越していることは御理解いただけよう。

では、なぜこのような大きな差があるのだろうか。

3 GRAPE アーキテクチャ

なぜ差があるのかを理解するためには、まずハードウェアの構成の違いをみておく必要がある。図 2 と図 3 に GRAPE の基本的な考え方を示す。 GRAPE が対象にする問題は、以下の運動方程式で表現される自己重力多体系である。

$$\frac{d^2\mathbf{x}_i}{dt^2} = -\sum_{i\neq j} \frac{Gm_j(\mathbf{x}_j - \mathbf{x}_i)}{|\mathbf{x}_j - \mathbf{x}_i|^3} \tag{1}$$

ここで $\mathbf{x}_i$、 m_i は粒子 i の位置と質量、 t は時刻、 G は重力定数である。右辺の和は i 以外の系のすべての粒子についてとる。

これは、単に系のすべての粒子が他のすべての粒子からの重力を受けて運動するというだけである。球状星団を始めとする星団、銀河、銀河団などは基本的にこういうものとして理解できる。太陽系での惑星形成や惑星の回りの衛星、リングについても基本的には上の式で表現され、

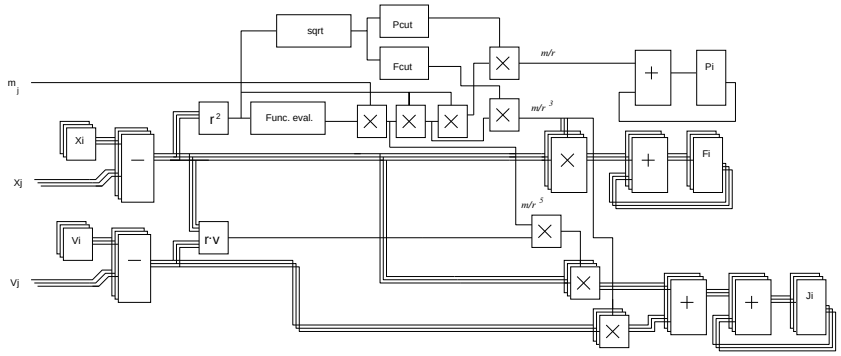

図 3: GRAPE パイプライン

これに物理的に衝突した粒子が跳ね返ったりあるいは合体する過程をとりいれて計算することになる。

このような系がどう振舞うかを調べるには、上の運動方程式を数値的に積分すること、つまりシミュレーション (普通 N 体シミュレーションと呼ばれる) がもっとも直接的で強力な道具になる。なお、原子一つ一つの運動を解くことで物質のマクロな性質や有機高分子の振舞いを調べる分子動力学法も基本的には同じような多粒子系を扱う。

このような多粒子系のシミュレーションの特徴は、計算量が極めて大きく、しかもそのほとんどが粒子間相互作用の計算に費やされることである。重力やクーロン力をまともに評価すれば計算量は粒子数 N の 2 乗になる。時間方向のステップ数も多く、問題によっては粒子数に比例して必要なステップ数も増える。さらに、特に重力の場合にはそれが引力であるために 2 つの粒子が無制限に近づいたり、自己重力不安定や自己重力と熱輸送がカップルして起こる重力熱力学的不安定で部分的に密度が高く粒子の運動が速い領域が形成される。このような場合には、粒子の必要に応じて独立に時間刻みを変えないと計算量が膨大になる。

時間刻みが一定、あるいは全粒子で共通でいいような場合には、バーンズ・ハットのツリー法 [4] や高速多重極法 [5] によって計算量のオーダーを N^2 から $N \log N$ や N に減らせる場合もある。しかしこの場合でも粒子間相互作用の計算がもっとも大変であることには変わりがない。

そこで我々が考えた (最初の提案は国立天文台の近田によるものである) のは、粒子間相互作用の計算だけを高速に行うハードウェアを作って、それを普通の計算機につなげて使うことである。このやり方には、いろいろな利点がある。

もっとも重要なものは、専用機側のトランジスタ利用効率を非常に高くできることである。最初にあげたようなマイクロプロセッサは演算器を 1-2 個しかもたない。これは、トランジスタ数が 3 千万を超えるということを考えると驚くべき少なさである。というのは、64 ビットの浮動小数点演算器を実現するだけならば必要なトランジスタ数は 10-20 万程度であり、計算上は 100 個以上の演算器を 1 チップに収められるからである。

これに対して GRAPE-6 用に開発した LSI では、 0.25μm と 2 世代前の技術を使っているにもかかわらず、演算器 60 個以上からなる演算パイプラインを 6 本集積し、約 400 個の演算器を 1 チップに収めた。これにより、動作クロックが 90 MHz と低くてもピーク性能で 31 Gflops と高い性能が実現できた。また消費電力も 15W 程度と低く、演算速度当りではマイクロプロセッ

サにくらべて 2 桁程度良くなっている。

粒子間相互作用の、多数の粒子が多数の粒子と相互作用するという特性を考えると、専用化した時にこのような高い効率が実現できることは特に不思議なことではない。この特性のためにまず非常に高い並列度があり、さらに必要なメモリバンド幅が小さいからである。

基本的なパイプラインの使い方としては、一本が一つの粒子への力を計算する。この場合には、クロック毎に力を及ぼす粒子一つのデータが必要になる。しかし、複数パイプラインがあればそれらは違う粒子への力を計算することでメモリからのデータは共有できる。さらに、1 本のパイプラインを時分割して仮想的に複数のパイプラインとして扱うことで、必要なメモリバンド幅をさらに下げることができる。 GRAPE-6 の場合は、仮想的な多重度を 8 とすることでメモリバンド幅を 720MB/s まで落としている。

さらに、多数のチップを並列に使うのも、同じ粒子への別の粒子からの力を計算させる (結果を回収する時に合計する)、あるいは別の粒子への力を計算させることで容易に実現できる。合計等も、すべてハードウェアで実行できるのでネットワークの段数が増えても遅延の増加はマイクロセコンド以下であり、大規模なネットワークを構成してもさして通信レイテンシが増えることはない。

図に GRAPE-6 のハードウェア構成を示す。1 ボードに 32 チップを載せ、これらは FPGA で構成された加算器のツリーでホストにつながる。さらに 4 ボードからの結果がネットワークボードで合計されてホストに返される。このネットワークボードは、互いにデータを交換して 4 台のホストからのデータを受け取ることができ、これにより 4 台のホストが粒子データを分散して持っていても GRAPE の側ではデータを重複させ、各ホストの粒子への力をホスト間通信なしに計算する。 64 Tflops システムは、これを 8 ホストに拡張することもできたが、分割して使う場合のことを考えて 4 ホストからなるクラスタを 4 組作り、つごう 16 台のホストをギガビットイーサネットで結合する構成にした。このために、特にレイテンシの面で不利ではあるが、64Tflops マシン全体を使いたいような大規模な計算ではそれほど大きな問題というわけではない。

4 専用機の利点 — 汎用機の欠点？

前節では、 GRAPE アーキテクチャの基本的な原理と、それが最新の GRAPE-6 ではどのように実現されているかを簡単にみた。汎用計算機と比べた時の本質的な利点は、専用化してやることを固定しているために、チップ上のトランジスタのほとんど全てを演算器に使い、それらを全て並列に動作させることが可能であるということである。

これは、言い換えると現在の汎用計算機のアーキテクチャには、多数の演算器を 1 チップに集積することも、それらをを並列に使うことも難しい、あるいは少なくともそういうことになっているということである。実際、既に述べたようにインテルや AMD の最新のプロセッサで演算器の数は 1-2 個であり、マイクロプロセッサを設計・製造する側は多数の演算器をいれても商品価値にはさしてつながらないと考えていることは間違いない。

この判断は、商品として大量にマイクロプロセッサを売るという観点からは正しいものであろう。絶対性能が重要でないというわけではないが、過去のソフトウェア資産が使えることなどの

要求を無視できないからである。

しかし、それと同時に、仮に「商品として大量に売る」という観点を無視したところで、インテル等のマイクロプロセッサよりも高い性能が得られるかといえば、理論的にはともかく現実的には現在そういうものが存在しているとはいいがたい。既にみたように、汎用とはいえ科学技術計算に特化しているはずの地球シミュレータでも、そのへんのマイクロプロセッサにくらべて価格性能比ではむしろ悪いからである。

もちろん、地球シミュレータ、あるいはそれに代表される (といっても他にはほとんど存在しなくなったが) ベクトルプロセッサについては、その演算能力当りのコストがマイクロプロセッサに比べて高いことには明確な理由がある。端的には、 これはメインメモリのバンド幅が高いことによっている。例えば 2GHz の P4 の場合、主記憶はのバンド幅は高価と言われる DRDRAM を使った場合で 3.2GB/s であり、もちろん連続アドレスのアクセスの場合にしか (まあ、少なくともキャッシュラインサイズ単位以上ないと) これに近い数字も実現されない。 これに対し伝統的なベクトルプロセッサのデザインでは、演算器一つ当り最低 1 クロック 1 語、可能であれば 2 語の転送を、連続アドレスだけでなく一定ストライドアクセス (多次元配列の外側の添字を回すような場合) にも可能にしている。このため、バンド幅だけでなくメモリアクセスレイテンシをほぼ隠蔽できるだけの数のメモリバンクが必要になり、これがコストを引き上げている。単に必要なバンド幅を供給するだけでも容易なことではない。2GHz のプロセッサでは 1 演算としても 16GB/s と現在の 5 倍であり、 800 Mb/s の信号線が 160 本必要になる。

ベクトルプロセッサではこの大きなメモリバンド幅と単純なメモリ階層が、実用プログラムで高い実効性能を実現することを容易にしていることはいまさら詳しく議論するまでもないと思う。

しかし、他方では、その高いコストに見合っただけの実効性能が得られているかというと、必ずしもそうは考えてないユーザーの割合が増えてきていることは確かであり、そのために大規模な計算機はほとんど PC クラスタや 4-8 CPU 程度の SMP マシンのクラスタに置き換わってきている。

この辺りの、科学技術計算に使われる計算機はどんなものか、あるいはどんなものであるべきかという話はなかなか難しい。難しい理由の大きなものは、現在の解が歴史的なものであることであろう。「なぜ現在使っているものがこういうものであるか」、「なぜもっとましなものがないか」を理解することはできなくはないが、、、というような抽象論をしてもしょうがない。以下、もう少し具体的に考えてみよう。

5 GRAPE における「成功」

GRAPE プロジェクトは 1988 年に始まり、それから 6 世代、10 種以上のバリエーションがあるハードウェアを開発してきた。 GRAPE-3 以降は商品化もされており、現在、世界中に広がる 30 以上の研究機関で 100 枚を超えるボードが稼働している。さらに、天文学の他、分子動力学用の類似のハードウェアも、GRAPE-3 を東大と共同開発した富士ゼロックスがその後独自に開発した MD-Engine、その改良形の MD-Engine II の他、東大が中心に開発した MD-GRAPE (商品版は画像技研が開発)、その後継であり、理研で開発されたピーク性能 75 Tflops に達する

MDM、さらには計画中のその後継機と発展してきている。

天文の分野についていえば、 GRAPE はそれなりの成功を収めてきたといってもまあそんなに嘘ということもないであろう。 GRAPE-4(1995 年、1.08Tflops) および GRAPE-6 によって 2 世代にわたって曲がりなりにも世界最高速の計算機を、汎用計算の価格の 1/100 程度の総開発費 (人件費がはいってないとかいいたい人もいるかもしれないが、4,6 のどちらも大学でハードウェア開発にかかわったのは学生を含めて数名であり、どう多めに数えても人件費は大した額にはならない) で実現したのである。また、単に大きな機械が出来たというだけでなく、GRAPE を使った研究も太陽系形成から宇宙の大規模構造の形成まで様々な領域で成果をあげていると思う。

汎用マイクロプロセッサの発展が現在のコースの延長上にある限り、トランジスタの演算に使われる割合はますます下がっていくであろう。これに対し、粒子系シミュレーションの計算のうち、相互作用計算に専用化したハードウェアを作るという GRAPE のアプローチでは、基本的に全てのトランジスタを演算器に使い続けることが出来る。従って、 GRAPE のような専用計算機の優位性は、新しいハードウェアの開発を続けていく限り当面の間揺らぐことはない。

6 専用計算機の今後

さて、それでは、 GRAPE にとって次の「挑戦」というべきものはなんであろうか？

多粒子系以外の問題に適用範囲を広げるという方向についてはどうであろうか？

これはあまりに多様な可能性がある。基本的に、ある程度計算法が確立した問題で、なおかつメモリ必要量ではなく計算量が問題ならば、専用化によって利益がえられないとは考えにくい。しかしながら、実際の問題に適用した時にどれだけの利点があるか、あるいはそもそもどんなアーキテクチャを考えるべきかというのは、問題毎に違っているであろう。

例えば、地球シミュレータの主なターゲットであろう気候モデルのシミュレーションを考えてみよう。対象を全球モデル、すなわち「地球シミュレータ」の名前の通り地球全体の気候を一度に計算する場合を考える。計算法としては従来の延長でいくならば球面上で解く関数をルジャンドル変換する方法になる。ルジャンドル変換はフーリエ変換のような高速アルゴリズムが存在しているわけではなく (高速多重極法の考え方を利用した方法などが研究されているが、まだ実用とはいいがたいようである)、計算量が球面上の格子点の数の 2 乗に比例する。これに対して、シミュレーションのそれ以外の部分の計算量は基本的には格子点数にしか比例しないので、地球シミュレータで扱うような高計算時間のほとんどがルジャンドル変換に費やされることになると考えられる。

このように明確に計算量のほとんどが一つの計算に費やされるならば、これを専用化するのは特に難しいことではない。基本的にはルジャンドル変換は基底関数との内積をとるだけなので、基底関数の発生と内積計算に専用化したハードウェアを設計するのはそれほど難しいことではない。

だが、ではそういったハードウェアを作るのが現実的な解といえるかどうかは難しいところであろう。全球モデルは主なターゲットではあるかもしれないが唯一の目標ではないであろう。さらに、海洋・大気結合モデルとかいろいろややこしいことを考える時にはルジャンドル変換が

もっとも大変かどうか分からない。

まあ、当たり前のことではあるが、専用化を考えるには、問題の特性を理解し、さらに使うアルゴリズムを問題の特性と計算機ハードウェア技術の特性の両面から最適化できる必要がある。GRAPE が対象とする長距離相互作用多粒子系にはまあまあうまくそういうことができていると考えられるが、これはまあ我々にそれなりの経験の蓄積があったからであり、良く知らない問題に手を出してうまくいく可能性はそんなにない。実際にその分野の専門家が自分で作ろうという時に、何か助けになるとかそれくらいのことはできるかもしれないが、それ以上のことはなかなか難しい。

興味を持つ方がいるかもしれないので、どういう場合に専用計算機が役に立つか、またそういうものを作るのはどうれくらい大変かということについて以下私見をまとめておきたい。

まず、当然のことではあるが、いかに普通の計算機の効率が悪いとかそういう問題があるにしても、それでも大して時間がかからないような問題ならばそのために何か作るというのはあまり意味がない。

ここで時間がかかるというのが具体的にはどれくらいのことをさしているかは状況によるわけで一般的な指針というのはありえない。計算が 10 倍とか 100 倍速くなることによって得られる利益が、計算機を作ることに自分が投入する手間その他に見合うかどうかというのが究極的な判断規準ということになる。

手間はもちろん主観的なコストも入るわけで、あなたが電子回路をいじったり低レベルのハードウェアを操作するプログラムを書いたりすることに全く興味も関心もなければ主観的なコストは非常に高いだろうし、逆にそういうことが嫌いというわけではなければもちろんコストは相対的には低いということになろう。

極めてラフな話をするなら、例えば論文 1 本のために必要な計算がパソコン 1 台で 1 ヶ月でできてしまうなら、それがもっと速くなるからといってあんまり意味はないかもしれない。でも、あなたにライバルがいて、そっちは 2 週間で同じ計算ができちゃうなら、1 週間で同じ計算が出来れば嬉しいだろう。とはいえ、計算機は 1 週間ではできないので、計算機が出来た時にまだ意味があるかという観点で考えないといけない。この時に考慮するべき重要なポイントは、普通の計算機も (当面は) 時間がたてば速くなるということである。

なお、上の、例えばパソコン 1 台で 1 ヶ月というような計算は、多くの場合それで十分だからそうしているのではなくそれ以上時間を掛けては研究にならないからそれくらいでできることをやっているわけである。「計算機がもっと速くなったらなにができるか」ということを考える想像力も大切かもしれない。

というわけで、速くなればいいことがある場合については、次に考えるべきことは速くできるかどうかである。 これは具体的な問題をみないとわからないわけだが、比較的単純な繰り返し計算が計算量の大部分を占めていれば、その部分を速くすればいいし、それは原理的には難しくない。そうでない場合は難しい。

さて、具体的にハードウェアにすることを考えるわけだが、とりあえずパソコンにつないで、ハードウェアがやってくれること以外の計算はパソコン側でやることを考える。この場合、重要なことはデータ転送速度が十分かどうかである。現在、普通のパソコンで利用可能な拡張バスは 32bit 33 MHz の PCI バスであり、DMA バースト転送を使った場合の理論ピーク速度が

133MB/s である。最近のパソコンは実際にこれにかなり近い速度がでる。これはしかし主記憶の速度に比べても 1/10 以下に過ぎないものであり、非常に遅い。現在は PCI-X が普及の初期であり、こちらは 1GB/s 程度が実現可能だが、その代わりカードの設計はそれなりに厄介になる。原理的には、 GRAPE のようなホストと専用機で仕事を分担するような方式では、データ転送量に比べて計算量が十分に大きい必要がある。これは、物理系のシミュレーションであれば、基本的に近接相互作用にはむかないということである。

転送速度が十分であれば、実際にハードウェアを作ることを考えることができる。もっとも、現在ではいきなりカスタム LSI を作るのは莫大な費用もかかり、失敗した時のリスクが大きいしそもそもそれだけの研究費を取ってくるのもなかなか難しいので、もうちょっと安くできる方法がいい。これは、現時点では基本的には FPGA (Field Programmable Gate Array) を使うということになる。FPGA は書き換え可能なメモリによって論理ゲートとゲート間接続をプログラム可能にした LSI である。カスタム LSI に比べて速度は遅く、集積度も小さいが、それでも 100 万ゲートを超えるものが利用可能になってきており演算精度によるが 100 を超える演算器が 1 チップに集積できる。これが 50 MHz で動作すれば 5 Gflops が実現でき、現時点のマイクロプロセッサに比べて高速なものが実現できる。 PCI カード上に FPGA とメモリを載せたボードはいろいろなものが商品化されているので、それらを使った小規模な実験から始めるのが現実的な解であろう。大学等であれば FPGA 設計のためのソフトウェアも FPGA メーカーから無償、または極めて安価に提供されている。

我々も、重力以外の相互作用計算に FPGA を使うシステムを開発中である。基本的に粒子間相互作用をターゲットに、演算パイプラインの設計等をある程度自動的に行うためのソフトウェアも作りつつある。

FPGA によって少なくともコスト的には小規模な専用計算機の開発は多くの人の手の届くものになったにもかかわらず、なかなか科学技術計算への応用は進んでいないように思う。上にみたように、専用計算機が有効であるためには様々な条件があることは確かだが、そういった条件を満たす問題は決して少なくはない。これを読んで、やってみようと思う人がでれば幸いである。

参考文献

[1] Applegate J. H., Douglas M. R., Gursel Y., Huner P., Seitz C. L., and Sussman G. J. (1985) A digital orrery. *IEEE Transactions on Computers* C-34: 822–831.

[2] 杉本大一郎 (1994) 専用計算機によるシミュレーション. 朝倉書店, 東京.

[3] 牧野淳一郎 (1998) 高速多重極展開法とツリー法—多体シミュレーションのための高速算法, 応用数理 8: 277–287.

[4] Barnes J. and Hut P. (1986) A hiearchical o(nlogn) force calculation algorithm. *Nature* 324: 446–449.

[5] Greengard L. and Rokhlin V. (December 1987) A fast algorithm for particle simulations. *Journal of Computational Physics* 73: 325–348.

[6] Hut P. and Makino J. (1999) Astrophysics on the grape family of special-purpose computers. *Science* 283: 501–505.

[7] Hachisu I. and Sugimoto D. (1978) Gravothermal catastrophe and negative specific head of self-gravitating systems. *Progess of Theoretical Physics* 60: 123–135.

[8] Makino J. and Taiji M. (1998) *Scientific Simulations with Special-Purpose Computers — The GRAPE Systems*. John Wiley and Sons, Chichester.

[9] Sussman G. J. and Wisdom J. (1992) Chaotic evolution of the solar system. *Science* 257: 56–62.