ダークマターハロー微細構造に関する研究

石山 智明
筑波大学計算科学研究センター
ishiyama@ccs.tsukuba.ac.jp

本研究の目的は、XT4 上で最適化された、宇宙論的状況を含む重力多体シミュレーションのための超並列コード、GreeM (Ishiyama et al. 2009, PASJ) の、数万コアに及ぶシステムであるアテルイ上での実効性能を測定することである。それに加え、XC-A として申請したプロジェクトで予定しているプロダクトランのいくつかを行うことである。

このコードは重力計算のアルゴリズムに TreePM 法を、並列化の領域分割に再帰的多段分割法、ロードバランスの調整に CPU 時間の計測に基づくものを使うことで、XT4 の 1000 コア以上を使っても高い並列化効率を実現してきた。またアテルイに先駆けて、京コンピュータ上でハイブリッド並列化、階層的通信をはじめとする最適化を実装し、京の全システム (82944 ノード、663552CPU コア、ピーク性能 10.6PFlops) を用いた粒子数 2 兆の宇宙論的シミュレーションで、5.8PFlops の実行性能 (55%の対ピーク性能) を達成してきた (Ishiyama et al. 2012, SC12 Gordon Bell Prize) 。

1 並列化効率

粒子数 2048^3、4096^3、6144^3 の 3 種類の宇宙論的 N 体シミュレーションについて、512CPU コアから 24000 コアの flat MPI 環境で、ストロングスケーリングを測定した結果が図 1 である。

必要なメモリ容量の都合上、4096^3 と 6144^3 のシミュレーションは、それぞれ 4096 コア、16384 コアから始まっている。どのシミュレーションも、非常に良いスケーリングを示している。他の計算機環境で数万並列で最適化されたコードであれば、アテルイ上でも良い並列化効率が期待できそうである。

4096^3 のシミュレーションで 16384 並列のところでバンプが存在するが、これには理由がある。シミュレーションでは長距離力を Particle-Mesh 法を用いて計算しているが、ここで大規模な全対通信が発生する。GreeM にはこの全対通信を高速化するための階層的通信アルゴリズム (詳しくは Ishiyama et al. 2012) が実装されている。16384 並列の計算では 1 回の全対通信を実行しているのに対し、24000 並列ではこの全対通信を 2 分割し、通信を高速化している。16384 並列でも 2 分割すればより良いスケーリングが見込め、バンプが消えると予想されるが、比較のため敢えて分割しないで実行した結果を示した。元々、京コンピュータ上のトーラスネットワークに適合するように考案されたアルゴリズムであるが、それとはトポロジが異なるネットワーク上でも有用であることを示す結果となった。

なお計算時間についてであるが、重力計算の精度のパラメータにもよるが、XT4 ではおよそ 80000 粒子/CPU コア/sec 重力計算することが可能であるのに対し、アテルイではおよそ 250000 粒子/CPU コア/sec 計算することが可能であった (ツリー法の見込み角が 0.5 の場合)。単純なコアあたりのピーク性能はアテルイが 2.17 倍高いが、それ以上の速度増加が得られている。CPU やネットワークがより新しくなったことや、コンパイラがより進化した等、色々な要因が考えられるが、XT4 ユーザーからすると、XC30 は後継機として非常に望ましいシステムであると考えられる。

なお京コンピュータ上では、およそ 500000-600000 粒子/**ノード**/sec であり、**1 ノード**あたりのピーク性能は、アテルイの **1 コア**に比べ 6.4 倍である。

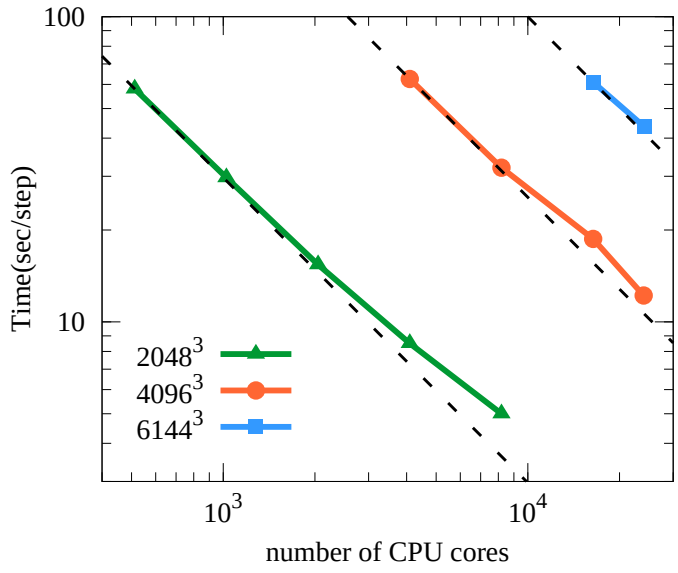

図 1: 並列化効率。横軸が CPU コア数で、縦軸が 1 ステップあたりの計算時間である。

2 プロダクトラン

XC-A として申請したプロジェクトで予定していたプロダクトランのうち 2 つを実行した。シミュレーションには、24000 コアを用い、flat MPI で実行した。結果は解析中であり、科学的成果については現時点で述べられることはない。ここでは計算科学的側面について考察する。

2.1 安定性

実行した 2 ランのうち片方の、各ステップあたりの計算時間の進化をプロットしたのが図 2 である。宇宙論的シミュレーションであることから、シミュレーションが進むにつれて、クラスタリングが進み、重力相互作用演算の数が増加するため、計算時間そのものも徐々に増加していく。図からわかる通り、ステップ毎に多少の浮動はあるが、計算時間は概ね安定

しており、計算時間が徐々に増加する傾向から大きく外れることはない。ネットワーク関連のシステムソフトウェアの最適化が十分でない等で、時たま通信が不安定になるシステムでは、あるステップの計算時間が突然それまでの数倍 (主に通信時間が不安定化することによる) かかることもある。このラン自体は 8 時間程度持続したが、少なくともその範囲内では、アテルイは非常に安定した優れたシステムであると言える。

2.2 IO 性能

1 ランあたり、解析のために 12 のスナップショットを掃き出した。1 スナップショットあたり合計 2TB あり、数 10%程度のばらつきはあるが、概ね 24000 並列に均等分割されている。一切の待ち処理を行うことなく、24000 並列でほぼ同時に書き込み命令を発行し、書き込みが終了するまでの時間を測定したところ、12 回の平均は 165 秒であった。最小で 154 秒、最大でも 183 秒であり、ばらつきは非常に小さい。以上のことから、アテルイ上で IO 処理がボトルネックとなる可能性は極めて低いと考えられる。

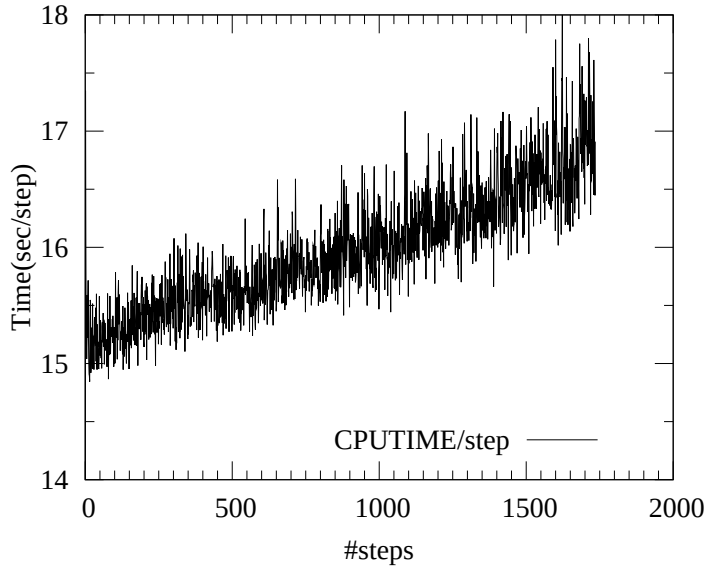

図 2: 各ステップあたりの計算時間の進化。