ポストペタ時代のソフトウェア技術

高橋大介
筑波大学大学院システム情報工学研究科

ペタスケールシステム

- 2010年6月のTop500において，以下の3システムがPFlopsの大台を突破している.
  1. Jaguar（Cray XT5-HE Opteron Six Core 2.6GHz）: 1.759 PFlops（224,162 Cores）
  2. Nebulae（Dawning TC3600 Blade，Inel X5650, NVidia Tesla C2050 GPU）：1.271 PFlops（120,640 Cores）
  3. Roadrunner（BladeCenter QS22/LS21 Cluster, PowerXCell 8i 3.2GHz / Opteron Dual Core 1.8GHz）: 1.042 PFlops (122,400 Cores)

エクサスケールシステムの出現時期は？

- 1964年：最初のMFlopsシステム
  - CDC 6600（RISCの元祖）
- 1983年：最初のGFlopsシステム
  - NEC SX-2（Vector）
- 1997年6月：最初のLinpack TFlopsシステム
  - Intel ASCI Red（Parallel）
- 2008年11月：最初のLinpack PFlopsシステム
  - IBM Roadrunner（Heterogeneous）
- 2018年11月？：最初のLinpack EFlopsシステム

ポストペタスケールシステム上のアルゴリズムにおける問題点

- 並列度
  - 次世代スパコン「京」では，64万コア（8万ノード）以上
  - MPIとOpenMPのハイブリッド実行を行ったとしても，MPIプロセス数が8万個以上になる.
- 演算精度
  - 倍精度で十分か？
  - 4倍精度演算も視野に入れる必要がある.
  - 精度保証も必要になってくる.

ポストペタスケールシステム上のシステムソフトウェアにおける問題点

- ノード間通信におけるバンド幅は物量投入で解決できるが，レイテンシを小さくするのは難しい．
- MPIプロセス数が多くなるに従って，ノード間通信に用いるバッファの領域も増大する．
  - 数十万MPIプロセスでは，GBのオーダーになることもあり得る．
- OSジッタ（各種デーモンやタイマ割り込み等）の影響が無視できなくなる．
  - 特に，同期処理（バリアやリダクション）処理を頻繁に行うアプリケーションでは，スケーラビリティが低下する原因になる．

ポストペタスケールシステムにおける性能チューニング

- 1990年代までのベクトルプロセッサでは，性能チューニングに必要なパラメータはそれほど多くなかった.
  - ベクトル長，アンローリング段数
  - 性能チューニングのコストの一部を「高価なハードウェア」で補っていたと考えることもできる.
- しかし現在では，アーキテクチャの階層が増えていることから，性能チューニングに必要なパラメータが増えている.
  - キャッシュブロッキングサイズ
  - ノード間通信のメッセージサイズやノード間通信アルゴリズムの選択

ポストペタ時代のソフトウェア（1/2）

- 各ノードが均一なアーキテクチャとなっているホモジニアス構成は，次世代スパコン「京」やBlueGene/Qが最後になってしまうかも知れない．
- ポストペタスケールシステムでは，アクセラレータを搭載したノードからなる，ヘテロジニアス構成が主流になると予想される．
- 性能チューニングに必要なパラメータが，さらに増える．
- 自動チューニングにより，「ある程度以上の性能を保証する」ための性能チューニングのコストを削減することが重要．

ポストペタ時代のソフトウェア（2/2）

- 2010年10月28日に発表された，中国NUDTの「Tianhe-1A」におけるLinpackの性能は2.507 PFlopsとなっているが，理論ピーク演算性能（約4.7 PFlops）に対する実行効率は約53%程度になっている.
- Linpackよりも実行効率の良いアプリケーションとしてはN体問題やEmbarrassing Parallelくらいしか考えられない.
- つまり，大多数のアプリケーションでは演算器の半分以下しか使われていないことになる.
- 今後は（実行効率が悪いのを覚悟で）ノード数を無理やり増やす方向にならざるを得ない.

ポストペタ時代に向けて（1/2）

- これまで，並列スパコンでは1コア当たりのメモリ容量がほぼ一定になるように，コア数が増えてきている.
- しかし，O(N)のデータに対してO(NlogN)やO(N^2)の演算量を必要とするアルゴリズムでは，Weak Scalingの場合，計算時間の増大が無視できなくなってきている.
  - Linpackでは1日以上の計算時間が必要な場合もある.
- つまり，同一時間内に計算できるデータ量の増加が今後は緩やかになる可能性がある.
  - 「Weak Scaling」よりも「Strong Scaling」で計算する機会が増えると考えられる.

ポストペタ時代に向けて（2/2）

- これまでのソフトウェアでは，いかにして「演算量を減らすか」ということに重点が置かれてきた.
- しかし，ポストペタ時代におけるソフトウェアではスケーラビリティが最も重要な指標になる.
- 演算量やメモリ使用量が2倍になっても，スケーラビリティが3倍改善されれば，十分元が取れる.
- 発想の転換が必要？