CBRC Newsletter 36
http://www.cbrc.jp/

Big dataと節電

津田　宏治
(Koji Tsuda)
機械学習研究班　班長

新世代シークエンサの到来によって、バイオインフォマティクスでも、数十から数百ギガバイト単位の処理は当たり前のものとなってきました。このような大量のデータとそれに伴う計算処理の負荷の急激な増加は、ウエブ、センサーデータなど、あらゆる科学技術、工学の分野で見られるものです。近年、このような現象は、Big Dataという言葉で表されることが多くなってきました[1]。それにより、従来の少量のデータに対して高度な処理を施すような贅沢はできなくなり、比較的簡単な処理をいかに高速に行うかという点に、機械学習、データマイニングなどのデータ工学の力点は移動しています。また、データ量が増加することによって、高度な統計的推論が不要になり、単純な最近傍法、線形推定等の方法でも、同等の精度が得られる例が多く報告されています (Data Beats Algorithm)。このような状況下で成功しているのは、Hadoop, MapReduceなどの枠組みを用いて計算を、多数のコアに分散させるというアプローチです。このような状況では、如何に多くのコアを用意し、実際に稼働できるかが、研究や事業の勝敗を決します。Beijing Genomics Institute (BGI)や、Googleでは、多額の資金を集め大規模なデータセンターを用意し、新興国において多数の人員を雇用して、非常に大規模な演算を行うことにより圧倒的なアドバンテージを得ています。

一方、このようなアプローチの限界も見えてきました。大きな問題は電力使用量です。現在のCPUは、毎時数百ワットの電力を必要とし、また、それを非常に狭い面積に集中させるため、多大な熱を発生します。そのため、データセンターでは大規模な空調を要します。

日本においては、東日本大震災のため、大幅な節電が求められ、CBRCでもスーパーコンピュータの停止などの対策を取らざるを得なくなっています。一方、実は震災が仮に無かったとしても、資源の逼迫による電力の危機というのは、すぐそこに迫っていたのでした。例えば、中国においては、特に災害がないにも関わらず、資源価格高騰により、今夏は計画停電を実施する予定のようです。このような状況では、データ処理のために、数千コアを用いる計算をすることは、今後社会的に認められにくいと思います。

これまで、デバイスの電力消費量には大きな関心が払われてきましたが、アルゴリズムの電力使用量という概念は、あまり一般的ではありませんでした。しかし、例えば10倍高速なアルゴリズムを用いれば、大まかにいって、電力使用量は1/10で済むわけですから、デバイスの改良よりも大幅な改善幅が期待できます。これまで、アルゴリズムの高速化は、職人的というか、感覚的に行われていた部分が大きかったですが、これをシステマティックに進めていく必要があります。これは容易なことではないですが、大きな足がかりとなるのは、Space-Time Trade-offを利用することです。低速でメモリをあまり使用しないアルゴリズムがあるとすると、それを、高速で多くのメモリを使うものに置き換えられる場合があります。直感的には、中間的な解が書いてあるカンニングペーパーを用意するのと同じです。メモリも、ある程度電力を消費しますが、CPUほど狭い面積に集中させないため、熱の問題は限定的です。また、圧縮データ構造を用いることによって、メモリの増加を抑えることもできます。

普通に考えれば、データ量が増加しているにもかかわらず、電力が存分に使えないのですから、日本の計算科学は、かなり不利な状態になっているのは間違いないでしょう。しかし、個々人が工夫することによって、結果的に、電力消費を抑えながら大規模データを扱う技術を飛躍的に発展させる契機にできるのではないかと思っています。

[1] G. Bell et al., Science, 323, 1297-1298.