「不完璧」なデータセンターとスーパーコンピュータを目指そう

国立情報学研究所　アーキテクチャ科学研究系 准教授　鯉渕（こいぶち）　道紘（みちひろ）

1. はじめに

1.1　計算の質の変化

スーパーコンピュータ（以後、スパコンと呼ぶ）とデータセンターにおいて、先進的な大規模アプリケーションの主流が、物理法則などの理論に基づく厳密さが要求される古典的な大規模計算から、ビッグデータ解析、人工知能、脳などの大まかに判断するというコンピュータが苦手とする領域に変化しつつある。

この変化は、コンピュータの設計に大きな影響を与えつつある。具体的には、許容誤差を若干大きくすることで計算の精度を落とし消費電力を削減、ハードウェアのスループットを向上させるApproximate Computingが注目されている。概算については、多くのディープラーニング系の計算をプロセッサの倍精度演算ではなく、半精度演算で行っても結果の大勢に影響しないことが報告されている。

従来、計算と情報の表現の精度はソフトウェア、ハードウェアの設計において悩みの種であった。コンピュータは数値を近似して表現（例：数0.110進は0.0001100［1100］2進で丸め）し、複数のプロセッサがハードウェアレベルで非決定的な順序により共有変数にアクセスするため、計算結果の潜在的誤差を完全に除去することが難しい。さらに厄介なことは、コンピュータのソフトエラーである。ソフトエラーとは、ハードウェアの故障による恒久的なものではなく、メモリに格納されているデータなどの一部のビットが、反転（0↔1）してしまう不良が非決定的に発生する不具合である。現在のスパコン、データセンターのコンピュータには様々なソフトエラーを検出訂正する機構が搭載されているが、現実的なハードウェア/ソフトウェアコストで、この不良から完璧に回復することを保証することは難しい。つまり、アプリケーションの実行が強制終了とならず、しかし、アプリケーションの計算結果が変造されることが起こる。この従来のコンピュータでは悩みの種を放置することが、Approximate Computingでは性能向上の糧となる。

1.2　ムーアの法則に従ったコンピュータの性能向上の終焉

コンピュータシステムの性能を向上させるためには、(1) プロセッサ単体の性能向上と (2) より多くのプロセッサを相互接続して1つのシステムを構成する並列化という2つの方向性がある。これまでの大規模コンピュータシステムは、この2つの方向性をうまく組み合わせることで急激な成長を達成してきた。例えば、スパコンは10年で1,000倍近い性能向上を達成している。しかし、(1) については、ムーアの法則が終焉し、コンピュータ機器の単純な性能向上が見込めなくなる時代が約10年後に迫る危機に直面している[1]。(2) についても並列化/巨大化の限界が見えている。例えば、最新のスパコンは、数百万プロセッサコア規模、消費電力が数百万ワット、そのネットワーク配線が1,000kmに達する。スパコンをさらに100倍大きくすることは難しいであろう。さらに、数百万並列で動作し、性能向上が見込める*1大規模アプリケーションの種類は限定されるであろう。よって、コンピュータの高速化を進めるためには、従来とは違う設計手法の確立が急務である。

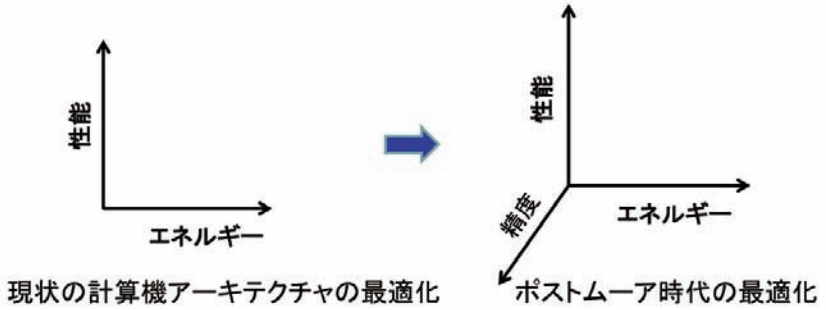
1.3　本稿の狙い

本稿では、今後先進的なアプリケーションが要求する計算結果の精度が従来と比べて緩和されることを利用し、従来のムーアの法則に頼らずともコンピュータの性能向上が実現できることを示す。つまり、Approximate Computingの探求である。コンピュータを設計する上で、従来は電力と性能の2軸を最適化していたが、将来は、加えて精度という3軸で最適化することになる（図1）。すなわち、今後、いい加減さ（時々計算を間違える不完璧さ）を許容することで、コンピュータとネットワークの性能が大幅に向上することが可能である。

著者らは、この視点を特にネットワークに向けたい。現状、スパコンとデータセンターのネットワークではソフトエラーについて標準規格があり、厳密に守られているため

*1　並列処理の分野では、N台のプロセッサでアプリケーションがN倍高速に動作する性能向上を「強スケーリング」、同様にN倍大きな問題の計算が完了する性能向上を「弱スケーリング」と呼び、性能向上の指標としている。ここでは、どちらかを満たせばよいという意図で用いている。

	今	将来
大規模アプリケーション	科学技術演算	ビッグデータ解析、AI/脳
計算機アーキテクチャ	**性能/電力効率探求**	**加えて、いい加減さを探求**
回路設計	ムーアの法則による性能/電力性能比向上	チップ単体の性能向上が困難



■図1. 大規模コンピュータ設計とアプリケーションの推移

Approximate Computingの考え方に基づく研究開発は見られない。例えば、イーサネットの規格では10^{-12}のビット誤り率[*2]を定めている。スパコンで頻繁に用いられるInfiniBandにいたっては10^{-19}で動作している。つまり、その高信頼性を確保するために、多大なコストを払い、また、自らが性能限界を作っているとも言える。著者らは、この標準規格から逸脱することで数倍～10倍の通信帯域の増加と、大幅な通信遅延の低下を見込んでいる。

2. Approximateネットワーク：多少の誤りを放置しよう

最近のスパコンとデータセンターのネットワークは、光ケーブルと電気スイッチを用いて構成される。リンク帯域向上の需要が著しいため、光通信チャネルの変調フォーマットとして、スペクトル効率の高い直角位相振幅変調（QAM）などの高度なフォーマットの使用が見込まれる。しかしこの場合、ビット距離が近くなるため信号対雑音比耐性が低下し、FEC（Forward Error Correction）によるエラー検出訂正を導入せざるを得なくなる。報告[2]によると、データセンターを対象とした25Gbps光通信において、FECの導入により、1リンク通過あたり100ナノ秒の遅延を見積もっている。この見積は、誤りを訂正するための処理遅延がネットワークの通信遅延の支配的要因になることを警鐘している。大規模アプリケーションの実行では並列処理のための通信機構が性能向上の鍵となるため。誤り検出、訂正が致命的にシステム全体の性能低下を招くことが生じる[3]。

そこで、著者らはデータの価値と伝送の確実性を比例させることで大幅な高帯域課と低遅延化を実現するApproximateネットワークを提案する。すなわち、誤差を許容し、誤り率を可変化することで、物理限界に迫る高密度な情報伝送を目指している。スパコンではこの「いい加減さ」を許容することで図2に示したApproximateネットワークによる性能向上が可能である（詳細な議論、解析は［3］に任せる）。

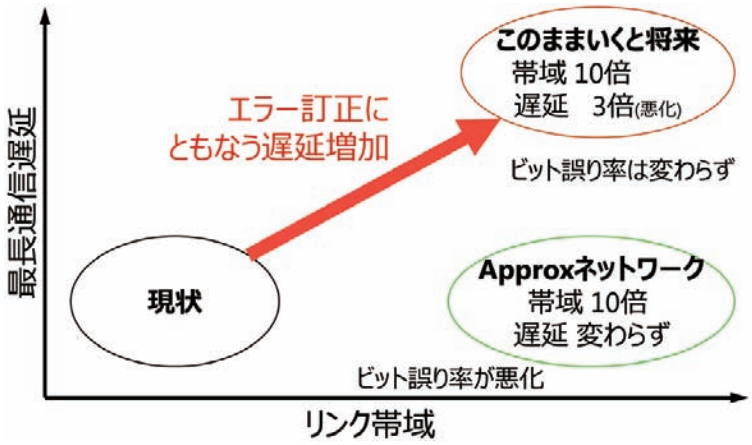
3. 多少の誤り放置によるアプリケーション性能の向上

Approximateネットワーク上のアプリケーションは、計算の精度について重い責任を持つことになる。ここでは、そのアプリケーション設計の2つの方策を述べる。

3.1 完全誤り放置型

1つ目の方策は、確信犯的にソフトエラーを放置することである。ソフトエラーが生じた場合でもそのまま計算処理を続行する。

*2 平たく述べると、誤ったデータの受信確率



■図2. スパコンの通信遅延、リンク帯域、ビット誤り率の関係

著者らは文献 [3] において、スパコンで頻繁に用いられる2種類（フーリエ変換、共役勾配法）、ビッグデータ処理で頻繁に用いられるK-平均クラスタリングアルゴリズムを対象に解析と拡張を行い、Approximateネットワークの有効性を示した。具体的には、ビット誤り率が10^{-5}と極めて悪いネットワークにおいて、共役勾配法では一部の通信において、浮動小数点数値を表現する64ビットのうち、上位16ビットのみを保護することで正しい解が得られた。つまり、下位48ビットは誤りを放置することが可能である。同様にして、フーリエ変換、K-平均クラスタリングアルゴリズムについても十分な精度の解が得られた。そして、スパコンのシミュレーション結果より、Approximateネットワークを用いることにより、最大3倍のアプリケーションの性能向上が達成できることが報告されている[3]。

3.2 Algorithm-Based Fault Tolerance（ABFT）

2つ目の方策は、アプリケーションによる「（誤りの）気付き」に期待する方法である。多くのアプリケーションでは、実行中に計算途中のデータが取り得る値かどうか、簡単な検算で判別することが可能である。そこで、検算結果から、許容誤差を越えた場合は、その計算を途中からやり直す。この方策は、従来ハードウェアが担っていた耐故障技術（チェックポイントやエラー検出/訂正）を用いずに、アプリケーションのアルゴリズムによって信頼性を担当することから、アルゴリズムに基づく耐故障技術（ABFT）と呼ばれる。

ABFTは、Approximateネットワークが提供するビット誤り率よりも高い処理精度がアプリケーションに必要な場合、必須と言える。理想的には、ABFTでは何度でも検算し、再実行すれば有限時間内に必要な精度の解が得られる。

完全誤り放置型と同様に、Approximateネットワークは、元の精度の高いネットワークと比べて、共役勾配法と行列計算について倍近い高速化を達成することが報告されている[3]。

なお、2つの方策のどちらが良いのか?という議論については、ケースバイケースであり、現時点で統一的な見解を著者らは得ていない。

4. その他の議論

4.1 アナログコンピュータの可能性

計算の精度を落とすという発想は、直感的にアナログコンピュータの復活を彷彿させる。事実、ニューラルネットワーク処理の一部の演算を、100MHzなどの低速で動作する特殊なアナログアクセラレータを用いることで、現状比数百倍の高速化と電力性能比の向上を達成する研究などが、コンピュータアーキテクチャ分野の研究をリードするトップ国際会議ISCA（International Symposium on Computer Architecture）、MI-CRO（International Symposium on Microarchitecture）などで近年発表され、注目を浴びている。

「デジタル処理」は0と1の間に十分にマージンがあるように閾値を決め、回路のノイズの影響を抑えることで安定的に高信頼な計算を可能としている。ただし、このマージンを削減することで伝送効率を向上させることができる。このマージンを小さくするにしたがって、アナログ処理に近い特徴が現れる。そして、「アナログ処理」は理想的には無駄なく信号処理ができるという点で（信頼性は低くとも）実行効率が高いと言える。

しかし、著者らは、アナログ回路がApproximate Computingを実現する中心的な役割を担うとは現時点では考えていない。これは、「アナログ回路を多数接続した大規模コンピュータを正しく制御できるのか?」あるいは「そもそも、設計段階で大規模なアナログ回路の検証を十分に行うことができるのか?」とまだ課題があると考えているためである。なお、著者らが提案するApproximateネットワークは、高効率のデジタル多値変調を用いるが、アナログコンピューティングではないことを申し添える。

4.2 限界

前章で述べた通り、Approximate Computingは、アプリケーションのデータ処理の精度を落とすことで性能向上を実現する。つまり、銀行オンライン処理、企業の業務基幹系処理、果ては宇宙ロケットの軌道計算など絶対に誤りが生じてはいけないコンピュータ処理に向かない。理想的にはABFTを用いることでこれらの処理系をApproximate Computingに用いることは可能であるが、誤りをその都度完璧に検出し、正しい結果が得 られるまで再実行することになるため、効率が悪い。あくまで著者らは、Approximate Computingとしてビッグデータ解析、人工知能、脳などの、大まかに判断するという先進的な大規模アプリケーションでの利用を想定している。

5. おわりに

ムーアの法則+αによるコンピュータの性能向上（例えばスーパーコンピュータは10年で1,000倍弱）が数十年続いてきた結果、その性能向上の継続が他分野、他業種にも知れ渡るマイルストーンであり続け、社会的要請になっている。

しかし、ムーアの法則の終焉が近づき、従来の設計方法では、コンピュータの性能向上の継続が困難となる可能性が高い。そこで、本稿では先進的アプリケーションの質的な変化に注目し、コンピュータ性能の成長戦略を示した。この質的変化とは、先進的な大規模アプリケーションの主流が、物理法則などの理論に基づく厳密さが要求される古典的な大規模計算から、ビッグデータ解析、人工知能、脳などの、大まかに判断するというコンピュータが苦手とする領域へ進むというものである。

このアプリケーションの変化により、コンピュータを設計する上で従来は電力と性能の2軸を最適化していたが、将来は、加えて精度という3軸で最適化することになる、現状、データセンターやスパコンのネットワークでは、ソフトエラーの発生確率、つまりビット誤り率について標準規格があり、厳密に守られている。現状のこの精度に関する標準規格から逸脱することで数倍～10倍の通信帯域の増加と、大幅な通信遅延の低下が見込め、その結果、フーリエ変換やK-平均クラスタリングアルゴリズムの実行速度が2～3倍向上する。

本稿では大規模アプリケーションの計算の精度に焦点を当て議論を行った。一方、世の中の大規模アプリケーションには入力データ自体の精度がそもそも低いものが多く存在し、計算の精度に関わらず複数の実行解を許容するものが散見する。例えば、センサーデータやノイズを多数含む観測データを入力にする処理はその典型である。著者らは、今後、Approximateネットワークの有効性を、これらのアプリケーションに対しても提示していく予定である。

謝辞

情報通信研究機構の藤原一毅主任研究員には、本稿の初期検討において多くの有益な助言をいただいた。本研究の一部はJST CREST、科研費16H02816、総務省SCOPE 152103004による。

文献

[1] “ポストムーアに向けた計算機科学・計算科学の新展開シンポジウム、” http://www.cspp.cc.u-tokyo.ac.jp/p-moore-201512/.

[2] M. Andrewartha, B. Booth, and C. Roth, “Feasibility and Rationale for 3m no-FEC server and switch DAC,” http://www.ieee802.org/3/by/public/Sept15/andrewartha_3by_01a_0915.pdf, Sept. 2015.

[3] D. Fujiki, K. Ishii, I. Fujiwara, H. Matsutani, H. Amano, H. Casanova, and M. Koibuchi, “High-bandwidth low-latency approximate interconnection networks,” The Inter-national Symposium on High-Performance Computer Ar-chitecture (HPCA), 12ページ, Feb 2017.