TSUBAME 共同利用　平成30年度　学術利用　成果報告書

利用課題名　スパコンのメモリ階層を活用したスケーラブル・大規模計算
英文：Scalable and Large Scale Computation Harnessing Memory Hierarchy on Supercomputers

遠藤敏夫
Toshio Endo

産業技術総合研究所 RWBC-OIL／東京工業大学学術国際情報センター
RWBC-OIL, AIST / GSIC, Tokyo Institute of Technology
https://unit.aist.go.jp/rwbc-oil/

邦文抄録
本課題ではスパコンのメモリ容量の限界を超えるような超大規模演算の実現を目的とし、ソフトウェアの研究開発を行った。この目的のために、TSUBAME3.0 が備える高速 NVMe SSD を含めたメモリ階層を用いた。このメモリ階層の活用をアプリケーション開発者から容易に行えることを目的として、Partitioned global address space(PGAS)モデルに基づきつつ、SSD の大容量を利用可能なミドルウェア vGASNet の実装改良・大規模評価を行った。TSUBAME3.0 の 32 台の計算ノードを利用した実験を、マイクロベンチマーク・ステンシル計算を対象に行った。それにより、SSD アクセスのオーバヘッドの抑制および、ノード数が多い場合に提案する協調キャッシュ方式がスケーラビリティを大きく改善することを実証した。

英文抄録
We have promoted software research and development, whose objective is to realize extreme big data computation that exceeds memory capacity of supercomputers. For this purpose, we used memory hierarchy including high-performance NVMe SSDs. Towards easier use of the hierarchy, we have implemented and evaluated vGASNet, which is a middleware to support usage of large capacity of SSDs. The evaluation has been done with 32nodes of TSUBAME3.0, using microbenchmarks, matrix computation and stencil computation. The evaluation has shown that vGASNet reduces overhead for SSD accesses, and our cooperative caching method improves scalability largely.

Keywords: SSD, Memory hierarchy, PGAS, Caching algorithm

背景と目的

スパコンにおける超高速・超大規模な演算の実現は、特にビッグデータ時代・人工知能時代と呼ばれる近年特に重要であり。学術的にも重要な目的とされている。本課題では、スパコンのメモリ(DRAM)容量を超えるような並列計算をアプリケーション開発者が容易に記述できるようなミドルウェアの実現を目的とする。具体的には Partitioned global address space(PGAS)モデルに基づきつつ、SSD の大容量を利用可能なミドルウェア vGASNet の研究開発を行ってきた[1,2]。

TSUBAME3.0 の 32 台の計算ノードを利用した実験を、vGASNet 上で記述されたマイクロベンチマーク・行列演算・ステンシル計算を対象に行った。それにより、SSD アクセスのオーバヘッドの抑制および、ノード数が多い場合に提案する協調キャッシュ方式がスケーラビリティを大きく改善することを実証した。

vGASNet の概要

vGASNet の設計は、代表的な PGAS 実装である GASNet をベースに行われた。GASNet では、アプリケーションは複数の計算ノードに分散したプロセス群からなる。GASNet 上では、2 つのメモリ領域であるグローバル領域とローカル領域が提供される。グローバル領域は概念的にプロセスたちにまたがった単一のものであるが、プロセスは直接計算のためにアクセスはできない。各プロセスは、get または put オペレーションにより、グローバル領域と、自プロセスのローカル領域にデータコピーを行うことができ、ローカル領域を直接の計算対象とする。

GASNet および PGAS モデルは、MPI 等よりも分散

並列プログラミングが容易なツールとして期待されている一方、計算ノード数が増えた際のスケーラビリティが(特に専用の集団通信処理を使わない場合に)低い傾向にあるという課題が指摘されている。また、MPI と同様に、各計算ノードの DRAM の合計容量を超える計算に、それ自身が対応するものではない。

それに対して vGASNet では、DRAM 合計容量を超えるグローバル領域を、計算ノードの SSD を用いて実現することが主眼である。TSUBAME3.0 では各ノードに 2TB（DRAM の 8 倍)、2GB/s 程度の高速 NVMe SSD が搭載されており、これを活用する。

技術的な課題は以下の通りである。

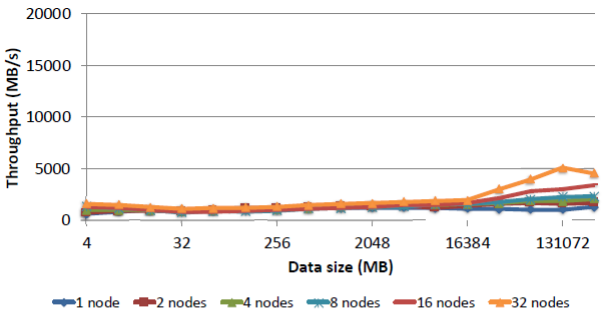
- 高速 SSD と言っても、DRAM (TSUBAME3.0 ではノードあたり 150GB/s)に比べればアクセスバンド幅は 2 桁近く悪い。そのアクセスオーバヘッドを抑制する必要がある。
- 計算に参加するノード数が増えた際に、たとえば単一のグローバルアドレスに多数ノードからのアクセス(put/get)が集中した場合に 1 ノードの SSD 帯幅・ネットワーク帯幅がボトルネックとなりうる。

そこで vGASNet は以下のような方策をとる。

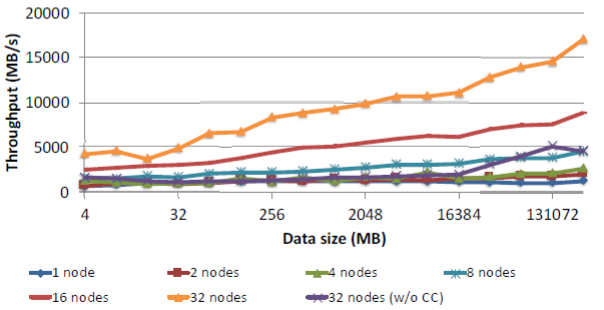
- DRAM 容量の一部を、SSD のキャッシュとして確保し、(アプリケーションの局所性が良好な場合に)SSD のアクセス頻度を削減する。
- グローバル領域の、あるアドレスのデータは、複数ノードの DRAM キャッシュ上に複製が存在しうるとする。新たなノードが同じアドレスをアクセスしようとするとき、キャッシュを持つ別ノードがデータ転送の役割を担うことができる。これを「協調キャッシング方式」と呼ぶ。
- 上記の場合、複数キャッシュ間で一貫性を保つ必要があり、新しい一貫性プロトコルである MOESI-F を用いる。

結果および考察

TSUBAME3.0 の 32 台の計算ノードを利用した実験を、vGASNet 上で記述されたマイクロベンチマーク・ステンシル計算を対象に行った。

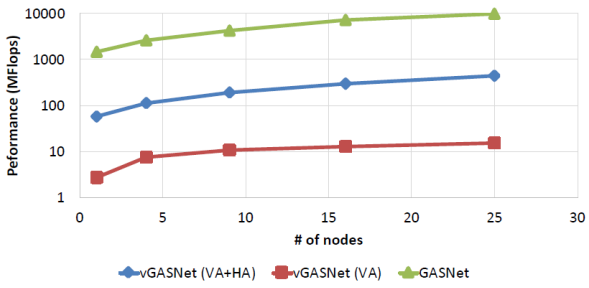


Sequential access throughput on T3 without the horizontal approach



Sequential access throughput on T3 with the horizontal approach

上図は、1～32 ノード上において、各プロセスがグローバル領域上の単一配列を逐次にスキャンするという、マイクロベンチマークの性能を表す。上のグラフは協調キャッシングを使わない場合、下のグラフは使う場合である。後者では明らかにアクセススループットが良く、特に 32 ノードの場合は 3 倍以上高速となっている。この結果は協調キャッシング手法の効果を示している。



Weak scale performance of the stencil program on T3

次に、アプリケーションにより近い例として、並列ステンシル計算の性能(それぞれ weak scale)を示す。横軸はノード数である。青のラインは、計算対象配列が DRAM 合計容量を超える場合の vGASNet 上の性能を示す。赤のラインは、協調キャッシングを使わない場合である。これらの比較から、このケースについては、協調キャッシングが速度性能に 10 倍以上の影響を持つ。なお、緑のラインは計算対象が DRAM におさまる、

小規模(そのため高速に計算できる)の場合をオリジナル GASNet 上で実行した場合である。理想的には大規模計算が、小規模計算に近い速度で可能であると望ましいが、まだ 20 倍ほどの差があることが分かる。しかしその差はハードウェアの性能差(100 倍弱)より抑制されており、キャッシュの効果が表れているといえる。

まとめ、今後の課題

DRAM メモリ容量を超える大規模データ計算を容易に可能とする vGASNet について述べ、その TSUBAME3.0 上の性能について報告した。NVMe SSD の大容量を活用しつつ、アクセスオーバヘッド等を大幅に抑制するキャッシング機構を中心に評価した。

将来に向けて、インメモリ(DRAM 内)の場合に近いアプリケーション速度性能を、大規模計算においても実現する課題があげられる。そのために、我々が研究してきた局所性向上アルゴリズム(ステンシル計算の場合は再帰的時間ブロッキング[3]など)との統合などを計画している。

文献

[1] Ryo Matsumiya, Toshio Endo. Scalable RMA-based Communication Library Featuring Node-local NVMs. In proceedings of 2018 IEEE High Performance Extreme Computing Conference (HPEC '18). Sep 2018.

[2] Ryo Matsumiya. Integration of Non-volatile Memory into One-sided Communication for Extreme Big Data Applications, PhD Thesis, Tokyo Institute of Technology, Jan 2019.

[3] Toshio Endo. Applying Recursive Temporal Blocking for Stencil Computations to Deeper Memory Hierarchy. In proceedings of the 7th IEEE Non-Volatile Memory Systems and Applications Symposium (NVMSA 2018). Aug 2018.