Vol.2019-HPC-168 No.13
2019/3/6

各種計算科学アプリケーションにおける NEC SX-Aurora TSUBASA システムの性能評価（２）

西川武志[†1]

概要： 前回は 1 ベクトルホストに 1 または 2 ベクトルエンジンを搭載した NEC SX-Aurora TSUBASA システムの各種計算科学アプリケーション（線形演算,分子動力学法,流体計算）での性能評価を行い報告した．今回は 1 ベクトルホストに 8 ベクトルエンジンを搭載した NEC SX-Aurora TSUBASA A300-8 システムにおいて VE 上のプロセス間の MPI 基本通信テスト（OSU MPI Latency Test）, 姫野ベンチマークテスト, 分子動力学法（MDCORE）, FDTD (Finite Difference Time Domain method)法（OpenFDTD）の性能評価を行ったので報告する．

1 ベクトルホストに 8 ベクトルエンジンを搭載したシステムの OpenFDTD の性能スケーラビリティは良好であったが,絶対性能では Xeon E5-2698 v4 や Tesla P100 の半分から 1/6 の性能であった．しかしながら OpenFDTD はベクトル化率向上余地があることも判明した..

キーワード：インセンティブ設計，計算センター運用，運用統計，並列度向上

1. はじめに

前回の報告[1]では DGEMM, 姫野ベンチマーク[2], 嶋ベンチマーク, MDCORE のいずれでも SX-Aurora TSUBASA の VE は良好な性能を Xeon CPU と比較して示した.

今回の報告では NEC から 1 ベクトルホストに 8 ベクトルエンジンを搭載した NEC SX-Aurora TSUBASA A300-8 システムを遠隔利用できる環境を提供されたので VE 上のプロセス間での MPI 基本通信テスト（OSU MPI Latency Test[3]）を行いノード間接続が Infiniband FDR 接続の FOCUS スパコン V システムとの比較も行った． 前回の報告で評価した姫野ベンチマークテスト, 分子動力学法（MDCORE）に加えて株式会社 EEM が開発公開しているオープンソースの FDTD (Finite Difference Time Domain method)法プログラム OpenFDTD[4]の性能評価を行った.

2. 性能評価対象システム

FOCUS スパコンシステムの概要については前回の報告[1]で述べているが, 今回, 性能評価の対象とした A, F, H, V の各システムの基本仕様を再度示す.

2.1 FOCUS スパコン A, D, F, H, V システム概要

(1) **A システム（224 ノード）**

高並列化環境（40Gbps QDR-Iinfiniband 接続）

CPU：Xeon L5640（Westmere-EP） 2.26 GHz 6 コア×2 108GFLOPS, RAM：48GB, HDD：500GB

(2) **F システム（60+2 ノード）**

高並列化環境（56Gbps FDR-Iinfiniband 接続）

CPU：Xeon E5-2698 v4（Broadwell） 2.2 GHz 20 コア×2 1152GFLOPS, RAM：128GB, HDD：6000GB

2 ノードには PCI 版 NVIDIA Tesla P100 をそれぞれ 1 基搭載

(3) **H システム（136 ノード）**

高密度高並列化環境（34 ノード/3U シャーシ, シャーシ間 40Gbps Ethernet ×16 シャーシ内ノード間 10Gbps Ethernet ×2 接続）

CPU：Xeon D-1541（Broadwell） 2.1 GHz 8 コア×1 205GFLOPS, RAM：64GB, SSD：512GB

(4) **V システム（2 ノード）**

NEC SX-Aurora Tsubasa A300-2 ベクトルエンジン環境（56Gbps FDR-Iinfiniband 接続）

ベクトルホスト（VH）: Xeon Gold 6148（Skylake） 2.4 GHz 20 コア×1 1024GFLOPS, RAM：96GB, HDD：240GB

PCI-Express 接続でベクトルエンジン（VE）NEC SX-Aurora TSUBASA Type 10B（周波数 1.4GHz 8 コア 2.15TFLOPS, メモリ帯域 1.22TB/s, HBM2 メモリ 48GB）をノードあたり 1 基搭載

2.2 NEC SX-Aurora TSUBASA 試用システム

今回の性能評価では NEC 提供リモート環境で VE を 2 基搭載した A300-2 を 1 台, VE を 8 基搭載した A300-8 を 1 台利用した．システムの概要は以下の通りである．

(1) **A300-2（VE 2 基搭載）**

NEC SX-Aurora TSUBASA A300-2 ベクトルエンジン環境

CPU：Xeon Gold 6148（Skylake） 2.4 GHz 20 コア×1 1024GFLOPS, RAM：96GB, HDD：240GB

PCI-Express 接続で NEC SX-Aurora TSUBASA Type 10B（周波数 1.4GHz 8 コア 2.15TFLOPS, メモリ帯域 1.22TB/s, HBM2 メモリ 48GB）をノードあたり 1 基搭載

(2) **A300-8（VE 8 基搭載）**

NEC SX-Aurora TSUBASA A300-8 ベクトルエンジン環境（100Gbps EDR-Iinfiniband 接続）

CPU：Xeon Gold 6148（Skylake） 2.4 GHz 20 コア×2 2048GFLOPS, RAM：192GB, HDD：240GB

PCI-Express 接続で NEC SX-Aurora TSUBASA Type 10B（周波数 1.4GHz 8 コア 2.15TFLOPS, メモリ帯域 1.22TB/s, HBM2 メモリ 48GB）をノードあたり 8 基搭載

†1(公財)計算科学振興財団
Foundation for Computational Science

3. 性能評価と考察

性能評価においては Xeon CPU はインテル Parallel Studio XE 2018.0.3.222 に含まれる Fortran, Intel MPI, Intel Math Kernel Library(MKL)により作成した実行モジュールを用い, NEC SX-Aurora TSUBASA の VE は NEC SX-Aurora TSUBASA Fortran コンパイラ 1.6.0, NEC SX-Aurora TSUBASA C/C++コンパイラ 1.6.0, NEC MPI 1.3.0（A300-2）, NEC MPI 1.4.0（A300-8）, NEC Numeric Library Collection の Version 1.0 を用いた.

3.1 通信遅延時間評価

NEC SX-Aurora TSUBASA システムの VE 上のプロセス間通信遅延時間の評価に osu-micro-benchmarks-5.5 の OSU MPI Latency Test を使用した. 性能測定結果を図 1 に示す.

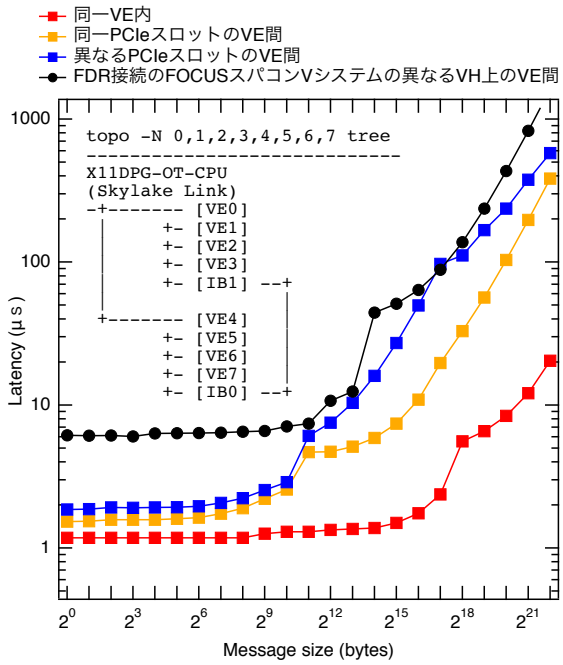

図 1 MPI 2 プロセス間通信による通信遅延時間評価

最低遅延時間は, 同一 VE 内（on same VE）では 1.3 μs±0.3 μs, 同一 PCIe スロットの VE 間（on same PCIe slot）では 1.7 μs±0.3 μs, 異なる PCIe スロットの VE 間（on same PCIe slot）では 2.1 μs±0.3 μs, FDR 接続の FOCUS スパコン V システムの異なる VH 上の VE 間では 6.4 μs±0.4 μs であった.

3.2 姫野ベンチマーク（HimenoBMT）

今回は 2001 年 11 月 26 日作成の姫野ベンチマーク Version 3.0, OpenMP に加えて MPI 対応の単精度版を用いて評価を行なった. 各 NEC SX-Aurora TSUBASA システムに対する姫野ベンチマーク（L: 512x256x256）の性能測定結果と FOCUS スパコン A システムに対する相対性能を表 2 に示す.

表 2 姫野ベンチマーク（L: 512x256x256）の性能

システム	スレッド数もしくはプロセス数	GFLOPS	相対性能 A=1
OMP A: L5640（Westmere）	12	8	1
MPI A: L5640	12	13	1.7
MPI A: L5640	24	25	3.1
OMP VE: Type10B	8	285	36
MPI 8VE/1VH (1proc/VE)	8	246	31
MPI 8VE/1VH (2proc/VE)	16	393	49
MPI 8VE/1VH (3proc/VE)	24	453	57
MPI 8VE/1VH (4proc/VE)	32	514	64
MPI 8VE/1VH (5proc/VE)	40	541	68
MPI 8VE/1VH (6proc/VE)	48	602	75
MPI 8VE/1VH (7proc/VE)	56	531	66
MPI 8VE/1VH (8proc/VE)	64	587	73

FOCUS スパコン A システムでは 1 ノード 12 スレッドの OpenMP 版よりも 1 ノード 12 プロセスのフラット MPI 版の方が 1.7 倍の性能を示した. 一方, NEC SX-Aurora TSUBASA システムでは 1VE に関しては OpenMP 版が MPI 版の約 1.2 倍の性能を示した. A300-8 の 1VH 上での複数 VE を使用した MPI 版の性能は 1VE あたり 6 プロセス実行の全 48 プロセス実行までは性能が向上するがそれを超えると飽和した.

3.3 MDCORE ベンチマーク

筆者が分子化学研究所、産業技術総合研究所、東京工業大学でのスーパーコンピュータシステム調達でのベンチマーク用に作成した 3 次元周期境界条件系の古典分子動力学法プログラム（粒子数 N=64 千, 相互作用数 N(N-1)/2+26*N*N=1.1e11）を用いて性能評価を行なった結果を表 3 に示す.

FOCUS スパコン V システム 2 ノードならびに 2VE/1VH の A300-2 上では 2VE 実行時に 8 プロセス, 全体の半分のベクトルコア利用で性能が飽和した. ただし 2VE/1VH の

A300-2 上での 8 プロセス実行時の性能は 2VE/2VH の FOCUS スパコン V システム 2 ノードより約 1.2 倍高速であった. 8VE/1VH の A300-8 上では 24 プロセス(3proc/VE)まで性能は向上した.

表 3 MDCORE（3 次元周期境界条件分子動力学法, 粒子数 N=64 千, 相互作用数 N(N-1)/2+26*N*N=1.1e11）の性能

システム	プロセス数	GFLOPS	相対性能 A=1
A: L5640	12	4.8	1
F: E5-2698 v4	40	53.4	11
H: D-1541	8	4.1	0.86
VH: Gold 6148	20	21	4.4
1VE(Type B)/1VH	8	111	23
2VE/2VH	4	110	23
2VE/2VH	8	172	36
2VE/2VH	12	171	36
2VE/2VH	16	148	31
2VE/1VH	4	123	26
2VE/1VH	8	199	41
2VE/1VH	12	198	41
2VE/1VH	16	173	36
8VE/1VH(1proc/VE)	8	211	44
8VE/1VH(2proc/VE)	16	322	67
8VE/1VH(3proc/VE)	24	369	77
8VE/1VH(4proc/VE)	32	355	74
8VE/1VH(5proc/VE)	40	303	63
8VE/1VH(6proc/VE)	48	309	64
8VE/1VH(7proc/VE)	56	294	61
8VE/1VH(8proc/VE)	64	192	40

粒子数 64 千と比較的小さい系のため少ないプロセス数で性能が飽和してしまったと考えられる.

飽和時の 8VE/1VH24 プロセスのピーク性能時で平均ベクトル長 224, 平均ベクトル化率 99.4%のため性能向上の余地は少ないと考えられる.

3.4 OpenFDTD

株式会社 EEM によってオープンソース FDTD (Finite Difference Time Domain method)[5][6][7]法シミュレータ OpenFDTD が公開されており, FOCUS スパコンシステム上で複数の課題で利用されているため今回の性能評価の対象とした. 当該プログラムは SIMD, OpenMP, MPI, CUDA, XeonPhi 向けに並列化・高速化されており[8], 配布ソースコードでは SIMD+OpenMP+MPI の組み合わせで高速化されているが, 今回の性能評価では OpenMP+MPI の組み合わせが予備調査で FOCUS スパコン A, D, F, H の Intel CPU システムにおいて性能が向上しなかったため, SIMD（AVX）+MPI の組み合わせの高速化を適用した実行バイナリを作成して評価した. CUDA によって高速化されたコードについては FOCUS スパコン F システムの Tesla P100 を 1 基 1 ノードに搭載したものでも性能評価した.

SX-Aurora TSUBASA システムに対しても MPI のみでの高速化が最高性能示したためそれのみを評価した.

性能評価には付属の 3 つのベンチマークテストのうち最大サイズ(セル数 400^3)必要メモリ 1920MB の benchmark3 を用いた. 表 4 に OpenFDTD benchmark3（セル数 400^3）の性能測定結果を示す.

表 4 OpenFDTD benchmark3（セル数 400^3）の性能

システム	プロセス数	経過時間（秒）	相対性能 A=1
A: L5640	12	492	1
F: E5-2698 v4	40	130	3.8
Tesla P100 on F	3584	42	12
H: D-1541	8	567	0.87
8VE/1VH(8proc/VE)	8	248	2.0
8VE/1VH(8proc/VE)	16	145	3.4
8VE/1VH(8proc/VE)	24	99	5.0
8VE/1VH(8proc/VE)	32	73	6.7
8VE/1VH(8proc/VE)	40	55	8.9
8VE/1VH(8proc/VE)	48	50	9.8
8VE/1VH(8proc/VE)	56	45	10.9
8VE/1VH(8proc/VE)	64	41	12.0

系のサイズが 400^3 と大きいため並列性能は飽和しておらず, 8VE/1VH64 プロセス並列のピーク性能時で平均ベクトル長 191, 平均ベクトル化率 77.1%のため性能向上の余地は大きい. 配布ソースコードを改変しない状態の 8VE システムの性能が CUDA 化したソースコードの Tesla P100 での実行性能とほぼ同一となったためシステム利用の利便性は高いと言える.

OpenFDTD コードのプロファイル解析すると電界と磁界の x,y,z 成分を更新する 6 つのルーチンがほぼ均等に実行時間の 15%ずつ 90%を占めており、残り 9%を Mur の一次吸収境界条件を磁界の接線成分 x,y,z に各成分に適用するものが占めており,さらに残りの 0.55%が電磁界の平均化ルーチンでいずれも高並列・高ベクトル化が可能な見通しが立てられた. 現在のコードでベクトル化を阻害している要因として C 言語で書かれた各関数の値の引き渡しに全てグローバル変数が利用されていることが推定される. 今後グローバル変数の利用を無くし自動ベクトル化コンパイラが

ベクトル化を適用しやすいように書き換えて性能評価を行う予定である.

4. まとめ

姫野ベンチマーク，OpenFDTD のいずれでも 1VH に 8VE を搭載した SX-Aurora TSUBASA の並列性能向上性は良好であった．現状のベクトル化チューニングが不十分な OpenFDTD では CUDA 版, Xeon CPU 利用版に絶対性能でもコストパフォーマンスでも Intel CPU を採用したシステムに劣っているが，コードのプロファイル解析から性能向上の余地が十分にあると考えられる.

5. 今後の課題

今後は，より多くの応用プログラムや MPI を用いた並列化効率の測定を様々な問題サイズに対して行う必要がある.

謝辞 評価に利用したベンチマークプログラムを公開してくださっている各位，評価のためベクトルエンジン Type 10B を 2 基搭載した SX-Aurora TSUBASA A300-2 および 8 基搭載した SX-Aurora TSUBASA A300-8 を試用させてくださったＮＥＣグローバルＰＦ本部関係者各位ならびに FOCUS スーパーコンピュータシステムの運用や利用者の開拓に尽力されている計算科学振興財団の同僚と利用してくださっている利用者各位に，謹んで感謝の意を表する.

参考文献

[1] 各種計算科学アプリケーションにおける NEC SX-Aurora TSUBASA システムの性能評価（１），西川 武志,研究報告ハイパフォーマンスコンピューティング（HPC），2018-HPC-167(17), 1-4 (2018-12-10).
[2] 姫野ベンチマーク, http://accc.riken.jp/supercom/documents/himenobmt/
[3] Point-to-Point MPI Benchmarks, osu_latency - Latency Test, http://mvapich.cse.ohio-state.edu/benchmarks/
[4] 株式会社 EEM, OpenFDTD, http://www.e-em.co.jp/OpenFDTD/
[5] 宇野亨「FDTD 法による電磁界およびアンテナ解析」コロナ社,1998
[6] 株式会社 EEM, "EEM-FDM" http://www.e-em.co.jp/fdm/eem_fdm.htm
[7]「EEM-FDM 理論説明書」 http://www.e-em.co.jp/doc/fdm_theory.pdf
[8] 株式会社 EEM、"高速化プログラミング入門" http://www.e-em.co.jp/tutorial/
[9] SX-Aurora TSUBASA におけるプロセス間通信の性能評価, 塩月 信智, 江川 隆輔, 滝沢 寛之， 研究報告ハイパフォーマンスコンピューティング（HPC）, 2018-HPC-165 (21),1-6 (2018-07-23).