Vol.2017-HPC-162 No.7
2017/12/18

スーパーコンピュータシステムITOの性能評価

大島 聡史[1,a)]　南里 豪志[1]　渡部 善隆[1]　天野 浩文[1]　小野 謙二[1]

概要：九州大学情報基盤研究開発センターではスーパーコンピュータシステム “ITO” を導入し、2017 年 10 月より一部システムによる試験運用を開始、2018 年 1 月より全システムによるサービス提供を予定している。本システムは最新の CPU や GPU を搭載していることに加えて、オープンデータの活用やパブリッククラウドサービスとの連携を考慮した挑戦的なシステムである。本稿では ITO の設計を紹介し、既に試験運用を開始しているバックエンドサブシステム B を用いて測定した性能評価の結果を示す。

1. はじめに

九州大学情報基盤研究開発センター（以下、当センター）ではこれまで数年間にわたって高性能アプリケーションサーバシステム (HITACHI HA8000 および SR16000)、高性能演算サーバシステム (FUJITSU PRIMERGY CX400)、およびスーパーコンピュータシステム (FUJITSU PRIMEHPC FX10) を運用してきた。当センターは、箱崎キャンパスにて 2007 年に従来の情報基盤センターと事務局情報企画課の統合による改組により設置された、研究開発と全国共同利用計算サービスを担当する部局である。情報基盤センターとしての設置システムも含め、従来のスーパーコンピュータシステム群は全て箱崎キャンパスに設置してきた。一方、現在九州大学では全学的なキャンパス移転を進めており、当センターも 2016 年 10 月に箱崎キャンパスから伊都キャンパスへ移転した。これに伴い、新システムは伊都キャンパスに設置する第一号のシステムとなるものとして調達を進めてきた。

今回当センターに導入されたスーパーコンピュータシステムは、伊都キャンパスに設置される初めてのスーパーコンピュータシステムであることから **ITO** と名付けられた。ITO は従来のスーパーコンピュータシステム群の利用者に引き続き計算資源を提供するという役割に加えて、利用者からの要望や社会的ニーズ等に答えるべく、AI（人工知能, 機械学習）、ビッグデータ、データサイエンスなどの新しい分野の研究及びこれらを活用した研究に対応した研究基盤の提供を目指して仕様策定されたものである。さらに詳細な電力モニタリング機構や本格的なクラウド連携の仕組みを導入し、従来にはない新しいスーパーコンピューティングの方向性や利用者層・課題の拡大に向けたインフラの提供を目指している。

ITO は 2017 年 10 月に一部システムの試験運用 (無償提供) を始めており、2018 年 1 月に全システムによる有償サービスの提供開始を予定している。本稿ではバックエンドサブシステム B を用いて実施した性能評価の結果を報告する。なおバックエンドサブシステム B は試験運用中のシステムであり、動作の安定化や性能向上のためにシステム設定の変更やドライバの更新等を度々実施している。したがって、本稿の内容と有償サービス提供開始後とでは得られる性能やその傾向に違いが生じる可能性がある。

2. ITO システムの紹介

2.1 概要

ITO の全体構成を**図 1** に示す。ITO は、2000 ノードの計算ノードにより構成される**バックエンドサブシステム A**、ノードあたり 4 基の GPU を搭載した 128 台の計算ノードにより構成される**バックエンドサブシステム B**、インタラクティブな用途のために用意された**フロントエンドサブシステム** (**基本フロントエンドノード群**および**大容量フロントエンドノード群**)、システム全体で共有される**ストレージサブシステム**により構成されており、これらが 100Gbps の InfiniBand EDR によって接続されている。

ITO システムの納入業者は富士通株式会社である。ただしフロントエンドシステムなど一部の構成要素については富士通株式会社以外による製品も活用されている。**図 2** に ITO システムの外観 (稼働中のバックエンドサブシステム B の写真) を示す。**表 1** には ITO システムの仕様 (性能諸元) を示す。ITO システムの冷却システムは、バックエンドサブシステム A/B のみ水冷、その他は空冷である。全システムは情報基盤研究開発センター内に設置されている。

[1] 九州大学 情報基盤研究開発センター
[a)] ohshima@cc.kyushu-u.ac.jp

表 1　ITO システムの性能諸元

バックエンドサブシステム A (2000 ノード)		
CPU	型番と数量	Intel Xeon Gold 6154 (Skylake-SP) 18 コア 3.0 - 3.7 GHz × 2
	1CPU あたり理論演算性能	1.728 TFLOPS
	メインメモリ	DDR4 2666 MHz, 192 GiB (96 GiB/CPU)
	1CPU ソケットあたり理論メモリバンド幅	127 GB/s
ローカルストレージ		1TB HDD × 2, 0.8TB SSD (一部ノードのみ)
ノード間接続		InfiniBand EDR 100Gbps

バックエンドサブシステム B (128 ノード)		
CPU	型番と数量	Intel Xeon Gold 6140 (Skylake-SP) 18 コア 2.3 - 3.7 GHz × 2
	1CPU あたり理論演算性能	1.3248 TFLOPS
	メインメモリ	DDR4 2666 MHz, 384 GiB (192 GiB/CPU)
	1CPU ソケットあたり理論メモリバンド幅	127 GB/s
GPU	型番と数量	NVIDIA Tesla P100 (Pascal), 1189 - 1328 MHz × 4
	1GPU あたりメモリ	HBM2 16 GB, 732 GB/s
	1GPU あたり理論演算性能	5.3 TFLOPS (DP)
	ホストとの接続	PCI-Express Gen.3 x16 (16GB/s)
	GPU 間の接続	NVLink 2 (20GB/sec)
ローカルストレージ		1TB HDD × 2, 0.8TB SSD
ノード間接続		InfiniBand EDR 100Gbps (2port)

基本フロントエンドノード群 (160 ノード)		
CPU	型番と数量	Intel Xeon Gold 6140 (Skylake-SP) 18 コア 2.3 - 3.7 GHz × 2
	1CPU あたり理論演算性能	1.3248 TFLOPS
	メインメモリ	DDR4 2666 MHz, 384 GiB (192 GiB/CPU)
	1CPU ソケットあたり理論メモリバンド幅	127 GB/s
GPU	型番と数量	NVIDIA Quadro P4000 (Pascal) × 1
	1GPU あたりメモリ	GDDR5, 8 GiB
	ホストとの接続	PCI-Express Gen.3 x16 (16GB/s)
ローカルストレージ		2TB HDD × 2
ノード間接続		InfiniBand EDR 100Gbps (2port)

大容量フロントエンドノード群 (4 ノード)		
CPU	型番と数量	Intel Xeon E7-8880 v4 (Broadwell-EP) 22 コア 2.2 - 3.3 GHz × 16
	1CPU あたり理論演算性能	774.4 GFLOPS
	メインメモリ	DDR4 1600 MHz, 12 TiB (0.75 TiB/CPU)
	1CPU ソケットあたり理論メモリバンド幅	51.2 GB/s
GPU	型番と数量	NVIDIA Quadro M4000 (Maxwell) × 1
	1GPU あたりメモリ	GDDR5, 8 GiB
	ホストとの接続	PCI-Express Gen.3 x16 (16GB/s)
ローカルストレージ		2TB HDD × 2
ノード間接続		InfiniBand EDR 100Gbps (4port)

ストレージサブシステム	
OSS	PRIMERGY RX2540 M2, 16 台 (4 台 ×4 セット)
OST	DDN 社製 SFA14KX, 4 台
MDS	PRIMERGY RX2540 M2, 6 台
MDT	ETERNUS DX600 S3, 3 台
ストレージ容量	24.64 PByte
バンド幅	バックエンドサブシステム A に対して 100GB/sec 以上 バックエンドサブシステム B およびフロントエンドサーバ群に対して 30GB/sec 以上 A,B, フロント全体に対して 120GB/sec 以上

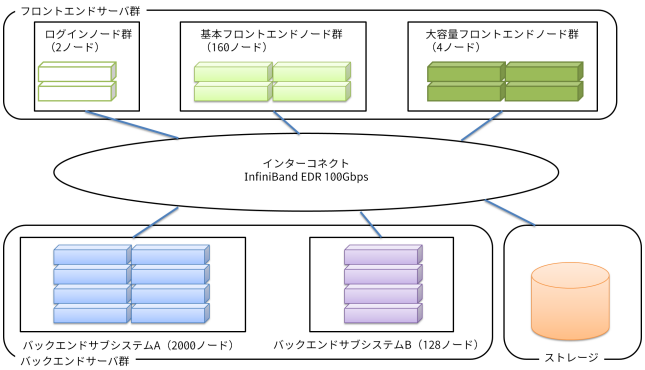


図 1　ITO システムの全体構成

図 2　ITO システム (バックエンドサブシステム B) の外観

2.2　バックエンドサブシステム A

バックエンドサブシステム A は 2000 ノードの計算ノードにより構成されている。各計算ノードには 18 コアからなる Intel Xeon Gold 6154 (Skylake-SP) が 2 ソケット、メインメモリとして DDR4 メモリが 192 GiB 搭載されている。計算ノードあたりの理論演算性能は 3.456 TFLOPS、メモリバンド幅は 255 GB/sec である。バックエンドサブシステム A 内の全計算ノードは InfiniBand EDR により Full Bisection Bandwidth Fat Tree で接続されており、合計で 6.912 PFLOPS および 510 TB/sec の性能を持つ。

2.3　バックエンドサブシステム B

バックエンドサブシステム B は 128 ノードの計算ノードにより構成されている。各計算ノードには 18 コアからなる Intel Xeon Gold 6140 (Skylake-SP) が 2 ソケット、メインメモリとして DDR4 メモリが 384GiB 搭載されている。さらにアクセラレータとして NVIDIA Tesla P100 (Pascal) が 4 基搭載されており、GPU 間は NVLink により相互に接続されている。計算ノードあたりの理論演算性能は 23.8496 TFLOPS(CPU のみで 2.6496 TFLOPS、GPU のみで 21.2 TFLOPS)、メモリバンド幅は、3183 GB/sec (CPU のみで 255 GB/sec, GPU のみで 2928 GB/sec) である。バックエンドサブシステム B 内の全計算ノードは InfiniBand EDR により Full Bisection Bandwidth Fat Tree で接続されており、合計で 3.052 PFLOPS および 407 TB/sec の性能を持つ。

2.4　フロントエンドサブシステム

フロントエンドサブシステムは、ログインノードと、160 ノードの基本フロントエンドノード群および 4 ノードの大容量フロントエンドノード群によって構成されている。バックエンドサブシステム A/B がバッチ処理にて運用されるのに対して、フロントエンドノード群はリソースを事前に予約して利用するインタラクティブなシステムであり、仮想マシンまたはベアメタルマシンとして様々な用途に活用されることが期待されている。

基本フロントエンドノード群は 18 コアからなる Intel Xeon Gold 6140 (Skylake-SP) を 2 ソケット、メインメモリとして DDR4 メモリを 384GiB、可視化処理等のために NVIDIA Quadro P4000 (Pascal) を 1 基搭載した HPE DL380 Gen 10 により構成されている。

大容量フロントエンドノード群は 22 コアからなる Intel Xeon E7-8990 v4 (Broadwell-EP) を 16 ソケット、メインメモリとして DDR4 メモリを 12TiB、可視化処理等の為に NVIDIA Quadro M4000 (Maxwell) を 1 基搭載した SGI UV300 により構成されており、特に大容量のメモリを活用した大規模プリ処理/ポスト処理への活用が想定されている。

2.5　ストレージサブシステム

ITO システムには実効容量 24.64PB の共有ファイルシステムが搭載されている。ストレージサブシステムは、バックエンドサブシステム A から 100GB/sec 以上、バックエンドサブシステム B およびフロントエンドサブシステムからそれぞれ 30GB/sec 以上、全システムから同時にアクセスした場合でも 120GB/sec 以上の性能でアクセス可能な共有ファイルシステムである。ストレージのフォーマットは FEFS である。

ITO システムの備える共有ファイルシステムは単一のストレージサブシステムのみであり、ログインノードでの作業からバッチ処理およびインタラクティブ処理まで常に同じストレージを参照することができる。一方、各計算ノードにはローカルな HDD(SATA 接続) が 1TB または 2TB 搭載されており、さらにサブシステム A の一部 (256 ノード) およびサブシステム B の全ノードにはノードあたり 0.8TB の SSD(SATA 接続) も搭載されている。これらのローカルストレージはおもにジョブ実行中の一時的な作業領域としての活用が期待されている。

2.6　ソフトウェア、その他

システム納入業者である富士通株式会社製のコンパイ

ラやライブラリ、ジョブ管理システムが利用可能である。さらに Intel コンパイラや PGI コンパイラ、様々なサードパーティー製及びオープンソースのソフトウェアやライブラリがインストールされている。詳細情報は Web ページにて提供し、随時更新しているので参考にされたい [1]。

3. 性能評価

本章では計算ノードに対する性能評価として STREAM、HPL、HPCG を、通信性能に対する評価として OSU Micro-Benchmarks を、ファイルシステムに対する性能評価として IOR および mdtest を行った結果を示す。さらに、より実アプリケーションに近いプログラムを実行した際の性能を評価するため、GeoFEM-Cube-OMP/CG ベンチマークの結果を示す。

3.1 評価環境

本章の性能評価はすべてバックエンドサブシステム B にて実施した。バックエンドサブシステム B の基本的なハードウェア構成は 2.3 節で述べたとおりである。NUMA 構成については、CPU1 ソケットあたり 1 つの Sub Numa Cluster (SNC)、計算ノード 1 ノードあたりは 2 つの SNC である。CPU と GPU の配置については、nvidia-smi topo –matrix の結果から GPU0 と GPU1 が CPU ソケット 0 に、GPU2 と GPU3 が CPU ソケット 1 に接続されていることが確認できている。また GPU 間の NVLink 接続については、CUDA の p2pBandwidthLatencyTest の結果から GPU0 と GPU1 の間および GPU2 と GPU3 の間が高速である (NVLink 接続が 2 本になっている) ことが確認できている。

また、性能比較対象として東京大学情報基盤センターに設置されている Reedbush-U/H/L(以下では RB-U/H/L と称する)[2] を用いる。RB-U/H にて性能評価を行った結果は、実測値および参考文献 [3],[4] にて公開されている情報を参照している。ITO と RB-U/H/L の主な違いについては**表 2** の通りであり、CPU については ITO が 1 世代新しく、GPU はいずれも同じ型番である。

3.2 STREAM ベンチマーク

STREAM ベンチマーク [5] はメモリ転送性能を測定することができるベンチマークプログラムであり、

Copy 配列のコピー
Scale 配列のスカラ倍
Add 2 つの配列の要素同士の加算
Triad Scale と Add の組み合わせ

の 4 種の測定が用意されている。STREAM ベンチマークは Fortran 版と C 版が提供されており、いずれも OpenMP による並列化が施されている。今回は C 版を用いて 1 ノードのみの STREAM 性能を測定した。コンパイラとしては icc 17.0.4 を使用し、主なコンパイルオプションとしては-O3 -no-prec-div -fp-model fast=2 -xHost -qopenmp -mcmodel=medium -qopt-streaming-stores=always を指定、問題サイズは 400,000,000 とした。

スレッド数と性能の関係を**図 3** に示す。1 ソケット使用時の性能 (a) については環境変数 KMP_AFFINITY=granularity=fine,compact 指定を行い、2 ソケット使用時の性能 (b) については環境変数 KMP_AFFINITY=granularity=fine,scatter 指定を行い測定した。いずれの測定についても numactl コマンドで-l を指定することにより常に近くのメモリを参照させている。実行結果から、1 ソケット使用時には 11 スレッド程度用いた時点で性能が頭打ちとなり、Copy と Scale で約 90GB/sec、Add と Triad で約 82GB/sec の性能が得られた。最も高い性能が得られたのは 13 スレッド Scale の 91.9GB/sec であり、理論メモリバンド幅 127GB/sec に対して 72.3%である。2 ソケット使用時には、20 スレッド程度用いた時点で性能が頭打ちとなり、Copy, Scale, Add で約 180GB/sec、Triad では約 165GB/sec であった。最も高い性能が得られたのは 26 スレッド Copy の 182.2GB/sec であり、理論メモリバンド幅 255GB/sec に対して最大 71.4%である。利用スレッド数が少ない場合には Copy と Scale に比べて Add と Triad の方が高速であるのにスレッド数が増えると逆転することや、Add と Triad の性能の上下幅が Copy と Scale と比べて大きいことも確認できたが、その理由は判明していない。

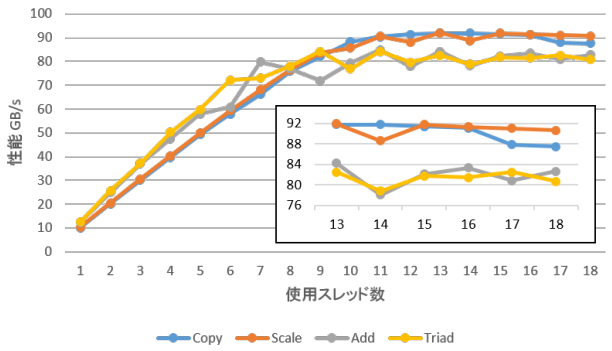
さらに、STREAM ベンチマークのソースコードに OpenACC の指示文を挿入して GPU 向けのプログラムを作成し、1GPU 上で実行して性能を測定した。コンパイラとしては pgcc 17.7 を使用し、主なコンパイルオプションとしては-acc -ta=tesla,cc60 -tp=haswell を指定、問題サイズは 100,000,000 とした。

実行結果を**表 3** に示す。測定は 5 回実施し、各測定項目ごとに最大性能のものを選択している。Copy と Scale は約 515GB/sec、Add と Triad は約 545GB/sec の性能であった。最も高い性能が得られたのは Triad の 547.8GB/sec であり、理論メモリバンド幅 732GB/sec に対して 74.8%の性能が得られた。

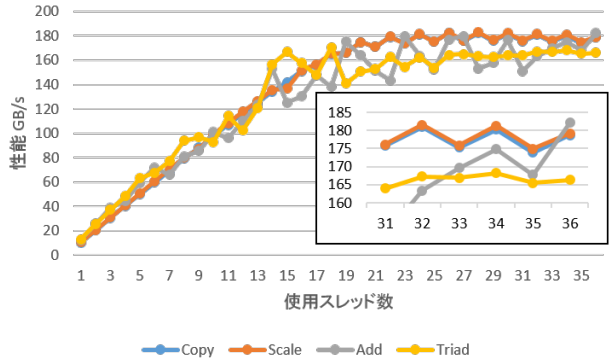
RB と ITO の性能を比較すると、RB-U にて CPU2 ソケット使用時の最大性能が 130.5GB/sec(Add) と報告されているのに対して ITO では 182.2GB/sec(Copy) であり、1.39 倍の性能が得られている。1CPU あたりの理論性能の差 (1.65 倍) と比べると低い倍率ではあるが、メモリクロックの向上およびメモリチャンネル数の増加の効果は大きい。GPU の性能については、同一の型番ということもあり、両者にほとんど違いはなかった。

表 2　ITO と RB-U/H/L の比較 (計算ノード 1 ノードあたり)

	RB-U/H/L	ITO (サブシステム B)	性能差
CPU 型番と数量	Xeon E5-2695 v4 ×2 (18 コア, 2.10-3.30GHz, Broadwell-EP)	Xeon Gold 6140 ×2 (18 コア, 2.30-3.70GHz, Skylake-SP)	
1CPU あたり理論演算性能 (DP)	604.8GF	1324.8GF	× 2.19
メインメモリ種別と容量	DDR4-2400 128GiB (1CPU あたり 2 チャンネル)	DDR4-2666 384GiB (1CPU あたり 3 チャンネル)	
1CPU あたりメモリ理論転送性能	76.8GB/s	127GB/s	× 1.65
CPU-CPU 間接続	QPI 9.6GT/s ×2	UPI 10.4GT/s ×3	
GPU 型番と数量	Tesla P100 ×2 (RB-H) Tesla P100 ×4 (RB-L)	Tesla P100 ×4	
1GPU あたり理論演算性能 (DP)	5.3TF	5.3TF	× 1.0
ノード間接続	IB EDR 100Gps (RB-U) IB FDR 56Gbps ×2 (RB-H) IB EDR 100Gbps ×2 (RB-L)	IB EDR 100Gbps ×2	



(a) 1 ソケット使用時の性能



(b) 2 ソケット使用時の性能

図 3　STREAM ベンチマーク (CPU) の結果

表 3　STREAM ベンチマーク (GPU) の結果

測定項目	性能 GB/sec
Copy	516.4
Scale	515.6
Add	544.0
Triad	547.8

3.3　HPL ベンチマーク

High Performance Linpack (HPL) ベンチマーク [6] は LU 分解により連立一次方程式の求解を行うベンチマークであり、倍精度浮動小数点データに対する行列積和演算 (Level-3 BLAS DGEMM) の性能が性能に大きな影響を及ぼすことが知られている。またスーパーコンピュータシステムの性能をランキング付けする TOP500/Green500[7] で利用されていることでもよく知られている。今回は試験運用期間中の測定ということもあり、計算ノード単体から最大で 16 ノードまでの小規模な測定を実施した結果を示す。

表 4　HPL ベンチマーク (CPU) の結果

ノード数	N	NB	P	Q	性能 (TF)	ピーク性能比 (%)
1	38400	384	1	1	1.76	66.4 (%)
2	268800	384	1	2	3.41	64.3 (%)
4	268800	384	2	2	6.41	60.5 (%)

CPU のみを用いて実施した HPL の結果を**表 4** に示す。プログラム (実行ファイル) は Intel コンパイラ 2017.4.196 に含まれている MKL によって提供されているコンパイル済 HPL(xhpl_intel64_dynamic) を用いた。MPI についても同コンパイラに含まれている IntelMPI を用いた。測定の結果、1 ノードでは 1.76TFLOPS、4 ノードでは 6.41TFLOPS の性能が得られており、それぞれ理論演算性能に対して 66.4%および 60.5%の性能であった。今回の測定では問題サイズやプロセスの配置などの最適化が十分に行えておらず、さらなる最適化によってより高い性能が得られることを期待している。

GPU を用いて実施した HPL の結果を**表 5** に示す。プログラム (実行ファイル) は、NVIDIA 社により提供された P100 向けコンパイル済 HPL バージョン 2.13.17 のうち、OpenMPI 1.10.2 向けにコンパイルされたものを用いた。測定の結果、1 ノードでは 15.19TFLOPS、16 ノードでは 167.1TFLOPS の性能が得られた。それぞれ理論演算性能に対して 63.5%および 43.7%の性能であった。使用するノード数を増やすごとに性能比が低下しており、問題の分割方法や MPI プロセスの配置には改善の余地がありそうな結果となった。

表 5　HPL ベンチマーク (GPU) の結果

ノード数 * GPU 数	N	NB	P	Q	性能 (TP)	ピーク性能比 (%)
1 * 1	38400	384	1	1	4.036	60.9
1 * 2	67200	384	2	1	7.641	57.6
1 * 4	76800	384	2	2	15.19	63.5
2 * 4	130176	384	4	2	28.53	59.8
4 * 4	192000	384	4	4	51.19	54.3
8 * 4	260352	384	8	4	93.50	49.8
16 * 4	384000	384	8	8	167.1	43.7

RB と ITO の性能を比較すると、RB-U では 1 ノードあたり CPU のみで 1149.6GF (95.0%)、RB-H では 2GPU も含めて 1 ノードあたり 10.04TF (85.0%) の性能が報告されている。ITO では CPU のみで 1.76TF (60.5%)、2GPU を含めて 7.641TF (57.6%)、4GPU を含めて 15.19TF (63.5%) となっており、ITO における HPL 実行については最適化の余地が大きいと考えられる。

3.4 HPCG ベンチマーク

HPCG ベンチマーク [8] は HPL ベンチマークよりも実アプリケーションに近いベンチマークとして提案されているベンチマークであり、有限要素法から得られる疎行列を対象として共益勾配法 (Conjugate Gradient, CG 法) を用いて連立一次方程式を解く部分の演算性能を求めるものである。実行の結果を**表 6** および**表 7** に示す。CPU, GPU ともに問題サイズは $nx = ny = nz = 256$ である。

CPU 向けの HPCG ベンチマーク測定は、Intel コンパイラ 2017.4.196 に含まれている MKL によって提供されているコンパイル済 HPCG を用いて行った。対応する HPCG のバージョンは 2.4 である。1 ノードあたり 1 プロセスよりも 2 プロセス (ソケットごとに 1 プロセス) の方が数%程度高い性能が得られ、また AVX512 環境向けに作成されたバイナリよりも AVX2 環境向けに作成されたバイナリの方が数%程度高い性能が得られたことから、AVX2 向けに作成されたものを用いてノードあたり 2 プロセスで実行した結果のみを示す。なお AVX2 環境向けに作成されたバイナリの方が高速である理由としては、Skylake-SP CPU は AVX 未使用時よりも AVX2 使用時、AVX2 使用時よりも AVX512 使用時に計算コアの動作周波数が低くなるためであると考えられる。

GPU 向けの HPCG ベンチマーク測定は、HPCG の Web サイトにて配布されている NVIDIA GPU 向けの HPCG 3.1 Binary (dated Oct 8, 2017) を用いて行った。得られた性能については、CPU との性能比較のため HPCG 3.0 基準と HPCG 2.4 基準のデータを併記する。

測定の結果、CPU のみの性能は 1 ノードで 32.8GF、16 ノードで 505GF、ピーク性能比はそれぞれ 1.23%および 1.19%であった。GPU を用いた場合は 1 ノード 4GPU で 394.4GF、16 ノード 4GPU で 5470.1GF、ピーク性能比はそれぞれ 1.65%および 1.44%であった。使用しているノード数が少ない割に複数ノード使用時の性能の低下具合が目立っており、プロセス配置の見直しになどによる最適化の余地があると考えられる。

表 6　HPCG ベンチマーク (CPU) の結果

ノード数	性能 (GF)	ピーク性能比 (%)
1	32.82	1.23
2	66.83	1.26
4	134.64	1.27
8	259.38	1.22
16	505.20	1.19

表 7　HPCG ベンチマーク (GPU) の結果

ノード数 * GPU 数	性能 (GF)	性能 (GF) (HPCG 2.4 相当)	ピーク性能比 (%)
1 * 2	200.04	202.73	1.67
1 * 4	394.49	398.44	1.65
2 * 4	746.50	755.04	1.56
4 * 4	1463.57	1479.39	1.53
8 * 4	2891.07	2922.94	1.51
16 * 4	5470.15	5529.53	1.44

(ピーク性能比は HPCG 3.0 のスコアに対して算出)

RB と ITO の性能を比較すると、RB-U では 1 ノードあたり CPU のみで 21.9GF (3.6%)、RB-H では 2GPU も含めて 1 ノードあたり 226.2GF (1.9%) の性能が報告されている。ITO では CPU のみで 32.82GF (1.23%)、2GPU を含めて 200.04GF (1.67%)、4GPU を含めて 394.49GF (1.65%) となっている。CPU の性能については、ITO は RB-U と比べてピーク性能比で大きく劣るものの得られた性能自体は 1.49 倍と高い。GPU の性能については、ITO は同じ 2GPU で RB-H に劣っており、HPL 同様に実行時のパラメタ等を見直す余地がありそうである。

3.5 OSU Micro-Benchmarks

OSU Micro-Benchmarks はオハイオ州立大学にて公開されている [9] 通信性能評価用のベンチマークである。今回は 1 ノード内の 2 プロセスによる通信性能と、2 ノード間に 1 プロセスずつ配置した場合の通信性能を測定した。CPU と GPU の対応付けについては、MPI プロセスは常にプロセスが存在する CPU ソケットから近い GPU(2GPU のうちのいずれか) を制御している。

図 4 は、1 ノード内の 2MPI プロセスが通信を行った場合の性能を示している。2 プロセスがともに CPU ソケット 0 上に配置されている場合 (0-0) と 2 プロセスがともに CPU ソケット 1 上に配置されている場合 (1-1) はほぼ同様の性能傾向を示しており、最大で 10GB/sec 超の性能を得られている。2 プロセスが CPU ソケットを跨いで配置さ

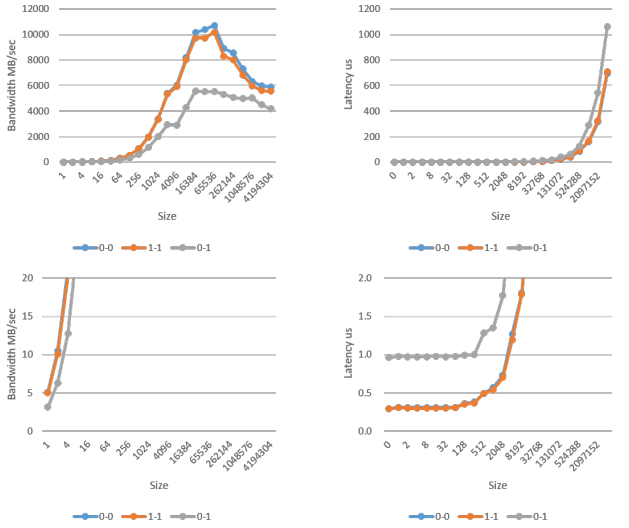


図 4　1 ノード内の CPU 間通信性能 (OpenMPI 1.10.7, 下段のグラフは上段のグラフの部分拡大版, 凡例は CPU ソケット番号)

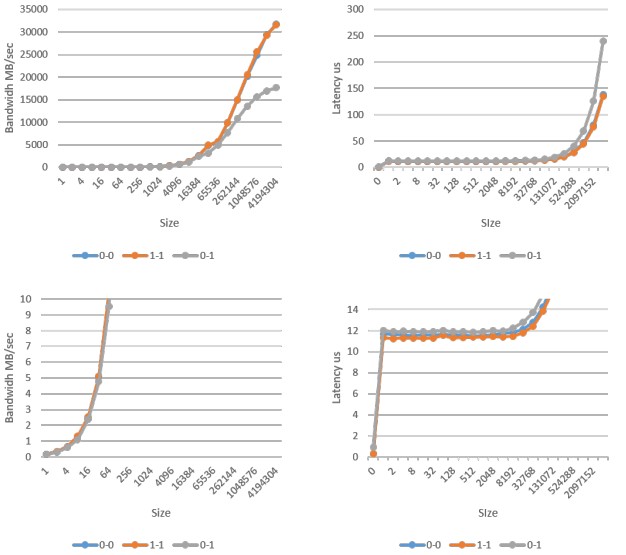


図 5　1 ノード内の GPU 間通信性能 (OpenMPI 1.10.7, 下段のグラフは上段のグラフの部分拡大版, 凡例は CPU ソケット番号)

れている場合 (0-1) には大幅に性能が低下しており、最大で 5.5GB/sec 程度である。この性能は CPU ソケット間を繋ぐ UPI の性能に律速されていると考えられるが、GT/s 性能しか明確にされておらず、理論性能に対する実測性能の差は不明である。レイテンシについても同一ソケット内では最小で 0.3us 程度まで低下しているのに対して、ソケット間では 1us 程度要していることがわかる。

図 5 は、1 ノード内の 2MPI プロセスが GPU 上のメモリを用いて通信を行った場合の性能を示している。具体的には OpenMPI に–mca btl_openib_want_cuda_gdr 1 を与え、ベンチマークプログラムの引数に D D を与えて性能を測定している。実行結果から、CPU と同様に同一 CPU ソケットに接続された GPU メモリによる通信 (0-0 および 1-1) とソケットをまたいだ GPU メモリによる通信 (0-1) には性能の隔たりがあり、最大転送性能は前者が 31.6GB/sec に対して後者が 17.6GB/sec であった。いずれも最大性能 16GB/sec の PCI-Express(Gen.3) ではなく、NVLink(20GB/sec、前者は 2 本) によって通信が行われていることがわかる。

図 6 は、2 ノード間の 2MPI プロセスが通信を行った場合の性能を示している。結果から CPU ソケット 1 同士の通信がバンド幅もレイテンシも最も優れており、CPU ソケット 0 同士が最も劣っていることがわかる。これは各ノード上の NIC が CPU ソケット 1 側に接続されていることを裏付ける結果であると言える。

図 7 は、2 ノード間の 2MPI プロセスが GPU 上のメモリを用いて通信を行った場合の性能を示している。ホストメモリを用いて通信を行ったときと同様に CPU ソケット 1 上の GPU メモリ同士で通信を行ったときの性能が最も優れており、CPU ソケット 0 上の GPU メモリ同士は最も劣っている。

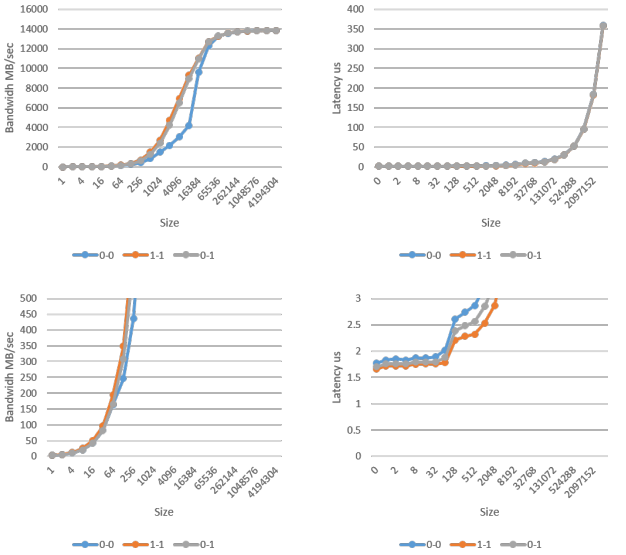


図 6　2 ノード間の CPU 間通信性能 (OpenMPI 1.10.7, 下段のグラフは上段のグラフの部分拡大版, 凡例は CPU ソケット番号)

3.6　IOR ベンチマーク

IOR ベンチマークは Los Alamos National Lab (LANL) が公開している I/O ベンチマーク [10] であり、ブロック入出力のスループットを計測するものである。今回は 1 プロセスあたり 1 ノードに割り当て、プロセスごとに異なるファイルに対する読み書きの性能を測定した。

表 8 に測定結果を示す。POSIX と MPIIO に大きな性能差は見受けられないが、ジョブ実行状況の都合により高い並列度での評価ができておらず、より高い並列度での性能評価が必要であると言える。

3.7　mdtest ベンチマーク

mdtest ベンチマークは IOR ベンチマークとともに Los

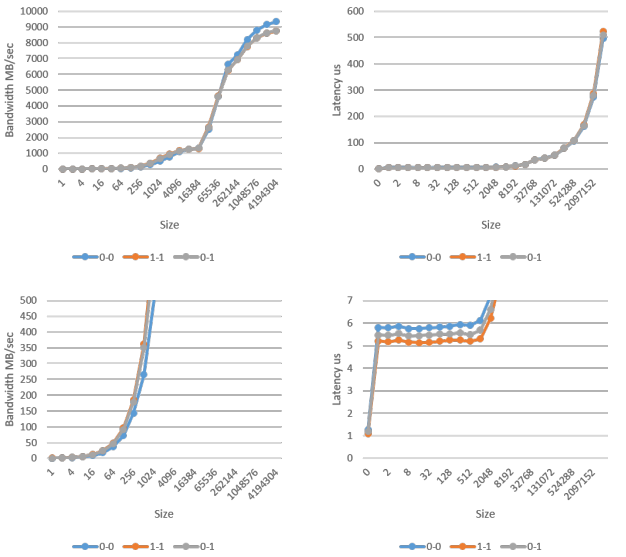


図 7　2 ノード間の GPU 間通信性能 (OpenMPI 1.10.7, 下段のグラフは上段のグラフの部分拡大版, 凡例は CPU ソケット番号)

表 8　IOR ベンチマークの結果 (MB/sec)

I/O 方式:POSIX				
R/W	1 ノード	2 ノード	4 ノード	8 ノード
Write	1006.05	1957.61	3870.06	7710.11
Read	730.78	1424.61	2700.58	5448.13
I/O 方式:MPIIO				
R/W	1 ノード	2 ノード	4 ノード	8 ノード
Write	988.49	1976.00	3936.60	7685.96
Read	748.05	1405.91	2777.77	5251.23

表 9　mdtest ベンチマークの結果 (ops/sec)

対象ファイル	ノード数	File creation	File stat	File read	File removal
単一	1	4439	2259	6785	6576
	2	8755	5479	13259	13531
	4	16005	10268	24018	17072
独立	1	4412	2430	6667	6400
	2	8716	5041	12287	8148
	4	14841	10436	22174	10858

Alamos National Lab (LANL) が公開している I/O ベンチマークであり、各種のファイル/ディレクトリ操作の性能を計測するものである。今回は 1 プロセスあたり 1 ノードに割り当て、ファイル操作の性能を測定した。

表 9 に測定結果を示す。性能値にはベンチマーク出力結果のうち Mean の値 (小数値切り捨て) を示している。単一ファイルへの処理の方がプロセスごとに独立したファイルへの処理よりも若干高速であるが、ジョブ実行状況の都合により高い並列度での評価ができておらず、より高い並列度での性能評価が必要であると言える。

3.8　GeoFEM-Cube-OMP/CG ベンチマーク

GeoFEM-Cube-OMP/CG は並列有限要素法プラットフォーム「GeoFEM」を元に整備したベンチマークプログラムである。対象問題は、一様な物性を有する単純形状 (Cube 型) を対象とした三次元弾性静解析問題である。係数行列が対象正定な疎行列であることから、ブロック対角化 (BlockDiagonalization) による前処理を適用した共役勾配法 (Conjugate Gradient, CG 法) によって連立一次方程式を解いている。本ベンチマークではこの CG 法部分の性能を計測している。オリジナルの GeoFEM は Fortran90 で書かれており、Flat MPI 版と OpenMPI/MPI ハイブリッド版が存在するが、本ベンチマークは後者を元に OpenMP のみで並列化を行ったものである。プログラムの詳細については参考文献 [11] を参照されたい。なお本ベンチマークは NUMA 環境向けの最適化は適用されていない。

性能の測定は RB-U の計算ノードと ITO のバックエンドサブシステム B の各 1 ノードにて行った。コンパイラは両者ともに ifort 17.0.4 を用いた。主なコンパイルオプションは、RB-U では-O3 -no-prec-div -fp-model fast=2 -xCORE-AVX2 -qopenmp -mcmodel=medium -align array32byte 、ITO では-O3 -no-prec-div -fp-model fast=2 -xCORE-AVX2 -qopenmp -mcmodel=medium -align array64byte と-O3 -no-prec-div -fp-model fast=2 -xCORE-AVX512 -qopenmp -mcmodel=medium -align array64byte の 2 種類を比較した。使用 CPU ソケット数 1 の際には環境変数 KMP_AFFINITY に compact を指定し、使用 CPU ソケット数 2 の際には環境変数 KMP_AFFINITY に scatter を指定し、さらに常に numactl -l を用いてスレッドとメモリを近い配置にした。問題サイズ Nx=Ny=Nz=129(2,146,689 節点、2,097,152 要素、6,440,067 自由度) の問題をそれぞれ実施したところ、**表 10** に示す性能が得られた。ITO の AVX2 と AVX512 については、わずかながら AVX2 の方が高い性能が得られた。これは、Skylake-SP が通常時より AVX2 使用時、AVX2 使用時より AVX512 使用時の動作周波数が低く、AVX512 命令の実行率が低いときやメモリ律速の場合などには性能が低下してしまうことが原因である。RB-U と ITO の性能を比較すると、ITO(AVX2) は RB-U の 1.48 倍 (1 ソケット) および 1.80 倍 (2 ソケット) の性能を発揮している。両者の理論性能には演算性能で 2.19 倍、メモリ性能で 1.65 倍の差があるが、メモリ性能の差に近い性能差が得られており、妥当な性能だと考えられる。1 ソケット使用時よりも 2 ソケット使用時に性能差が広がる点については、本プログラムが NUMA 環境向けの最適化が適用されていないために CPU ソケット間の通信性能の差が影響したと考えている。

4.　おわりに

本稿では、2017 年 10 月より一部運用を開始したスーパーコンピュータシステム ITO のバックエンドサブシステム B を用いた性能評価の結果を示した。ITO の搭載する

表 10 Geo-FEM-Cube-OMP/CG ベンチマークの結果

実行環境	使用 CPU ソケット数	実行時間 (秒)	反復回数	1 反復時間 (秒)
RB-U	1	1.08e+2	1305	8.29e-2
	2	6.75e+1	1305	5.16e-2
ITO (AVX2)	1	7.31e+1	1305	5.60e-2
	2	3.72e+1	1305	2.85e-2
ITO (AVX512)	1	7.33e+1	1305	5.61e-2
	2	3.78e+1	1305	2.89e-2

最新の Skylake-SP CPU は旧世代の CPU よりも大幅に演算性能が向上しているものの、SIMD 長が長いことなどから高い実行効率を得る難しさも感じられる性能評価結果となった。メモリ転送性能の向上は疎行列ソルバーを始めとした様々なプログラムにて恩恵が得られると期待される。各ノードに 4 基搭載された GPU は、最適化研究も進んでおり対応するアプリケーションも増えているものの、MPI を用いた多ノードでの実行についてはまだ難しさがあるため、今後も利用技術の調査と ITO システム利用者への適切な情報提供に努めていく予定である。

今回の性能評価は、システム利用開始からの時間が短くノード数の少ない部分運用環境下における測定のため、十分な最適化が行えていない評価項目も目立つ結果となった。今後、2000 ノード 4000 ソケットの Skylake-SP を搭載したバックエンドサブシステム A が稼働開始した後には、1000 ノード規模や全系を用いた性能評価やプログラム最適化技術の研究、さらにフロントエンドサブシステムの活用などについても取り組んでいきたい。

謝辞 ベンチマークプログラムや Reedbush システム上での性能評価についての情報をご提供いただいた東京大学情報基盤センターの中島研吾教授と塙敏博准教授および NVIDIA Japan の皆様に感謝します。

参考文献

[1] ソフトウェア — 九州大学情報基盤研究開発センター https://www.cc.kyushu-u.ac.jp/scp/software/ (accessed 2017-11-22).
[2] Reedbush スーパーコンピュータシステム [東京大学情報基盤センタースーパーコンピューティング部門], https://www.cc.u-tokyo.ac.jp/system/reedbush/ (accessed 2017-11-24).
[3] 塙敏博, 中島研吾, 大島聡史, 伊田明宏, 星野哲也, 田浦健次朗: データ解析・シミュレーション融合スーパーコンピュータシステム Reedbush-U の性能評価, 情報処理学会研究報告 (HPC-156), 9 月 8 日発行 (Vol.2016-HPC-156), pp.1-10 (2016).
[4] 塙敏博, 星野哲也, 中島研吾, 大島聡史, 伊田明弘: GPU 搭載スーパーコンピュータ Reedbush-H の性能評価, 情報処理学会 研究報告 (HPC-159), 4 月 10 日発行 (Vol.2017-HPC-159), pp.1–6 (2016).
[5] John D. McCalpin, “STREAM: Sustainable Memory Bandwidth in High Performance Computers”, http://www.cs.virginia.edu/stream/ (1991-2007).
[6] HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers http://www.netlib.org/benchmark/hpl/ (accessed 2017-11-22).
[7] Home — TOP500 Supercomputer Sites https://www.top500.org/ (accessed 2017-11-22).
[8] HPCG http://www.hpcg-benchmark.org/ (accessed 2017-11-22).
[9] MVAPICH :: Benchmarks http://mvapich.cse.ohio-state.edu/benchmarks/ (accessed 2017-11-24).
[10] IOR-LANL/ior: IOR and mdtest https://github.com/IOR-LANL/ior (accessed 2017-11-24).
[11] 中島研吾: T2K オープンスパコン (東大) チューニング連載講座番外編 Hybrid 並列プログラミングモデルの評価 (I), http://www.cc.u-tokyo.ac.jp/support/press/news/VOL11/No4/200907tuning.pdf (accessed 2017-11-24).