Vol.2020-HPC-175 No.17
2020/7/31

スーパーコンピュータ「不老」の性能評価

大島 聡史[1,a)]　永井 亨[1]　片桐 孝洋[1]

概要：名古屋大学情報基盤センターでは、2020 年 7 月 1 日に新しいスーパーコンピュータ「不老」のサービスを開始する。本システムは、多様化する利用者の要求に応えるシステムを目指し、特徴の異なる複数のサブシステムと複数のストレージシステムを搭載するシステムとして設計された。特に、Type I サブシステムは「富岳」の商用版である FX1000 を世界に先駆けて本格運用するものであり、さらに大規模なスーパーコンピュータシステムでは初めて大容量光ディスクよるコールドストレージシステムを提供するなど、先進的で挑戦的なシステムとなっている。本稿ではスーパーコンピュータ「不老」の設計について紹介するとともに、Type I サブシステムと Type II サブシステムを中心に、幾つかのプログラムによる性能評価の結果を示す。

1. はじめに

名古屋大学情報基盤センター (以下、当センター) は、名古屋大学 情報連携統括本部に含まれる組織であり、学内外の様々な情報サービスを一元的に扱う組織として 2009 年に従来の情報連携基盤センターに情報メディア教育センターを廃止・統合して発足した。当センターでは、2013 年に導入した FUJITSU Supercomputer PRIMEHPC FX10 システム (以下、FX10 システム)[1] を 2015 年にアップグレードした FUJITSU Supercomputer PRIMEHPC FX100 システム (以下、FX100 システム)[2] と、同様に 2013 年に導入し 2015 年にアップグレードした FUJITSU Server PRIMERGY CX400 システム (以下、CX400 システム) を中心として、全国共同利用施設として国内外に向けて計算サービスを運用してきた。これらのシステムは安定して計算資源の提供を続けてはきたものの、運用期間が当初の予定を超え、またノードあたりの性能も現在主流の機種に劣るものとなってきたことから、2020 年 3 月末日にてサービス提供を終了した。

当センターでは、2018 年頃から本格的に次期システムの仕様について調査検討を開始し、最終的に 2019 年 9 月に仕様書を公開した。その後、2020 年 1 月に富士通株式会社が落札し、2020 年 7 月 1 日のサービス運用開始に向けて準備を進めている[*1]。従来、当センターでは設置されるスーパーコンピュータシステムに愛称をつけてこなかったが、利用者にさらに身近に感じてもらえるシステムとなることを目指して、今回初めて公募を行い愛称を付けた。愛称である「不老」は、名古屋大学の所在地名である愛知県名古屋市千種区不老町に由来し、このシステムによって人類が得た恩恵がその後末永く人類の文明の中に生き続けることを願って命名された。なお、英語名称は処理の流れなどを示す「Flow」であり、「不老」の読みも「フロウ」ではなく「フロー」である。スーパーコンピュータ「不老」は、従来のスーパーコンピュータシステム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、多様化する利用者の需要・要望や社会的ニーズ等に応えるために、特徴の異なる複数の計算サブシステムと 2 種類のストレージシステムを備え、電力モニタリング機構とそれを活用した節電対応運転を行う仕組みを有し、湧水を活用した冷却機構の導入により消費電力を削減するなど、先進的で挑戦的なシステムとなっている。

[1] 名古屋大学 情報基盤センター
[a)] ohshima@cc.nagoya-u.ac.jp

[*1] 本原稿の提出締切は 6 月 29 日であるため、まだサービス運用を開始していない。

2. スーパーコンピュータ「不老」の構成と特徴

2.1 全体構成

スーパーコンピュータ「不老」(以下、“不老” と記す) の全体構成を**図 1** に示す。また、当センター本館地下 1 階にある計算機室に設置された “不老” の写真を**図 2**, **図 3** に示す。

“不老” は、当センターのこれまでの計算需要に限らない様々な需要に応えるために、特徴の異なる複数の計算サブシステムと 2 種類のストレージシステムを中心とした複合的なスーパーコンピュータシステムとなっている。“不老” の主要な計算サブシステムとストレージの主な仕様を**表 1**

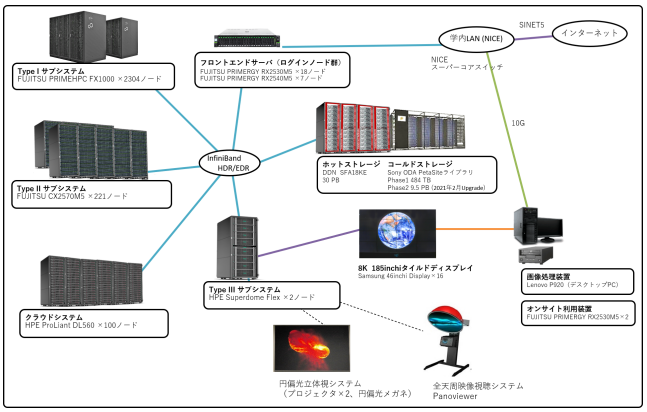

図 1　スーパーコンピュータ「不老」の全体構成

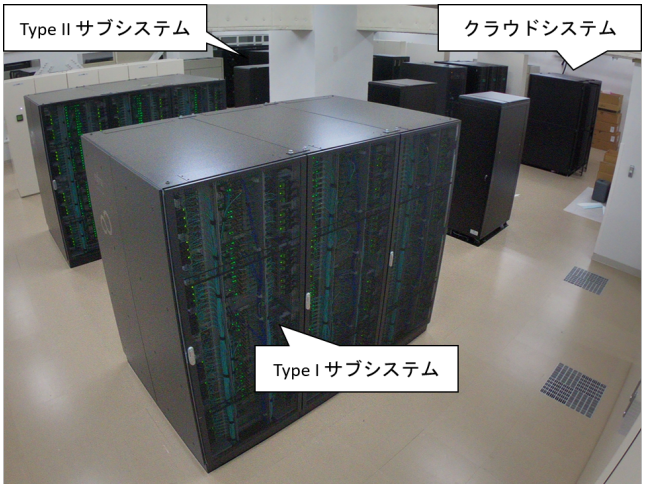

図 2　スーパーコンピュータ「不老」の外観写真 1

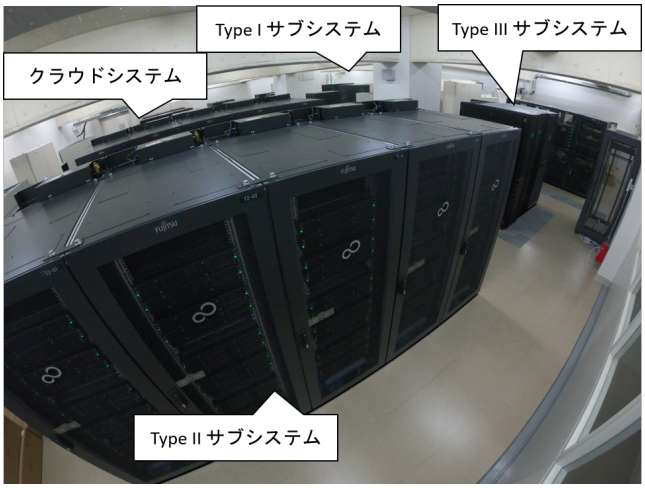

図 3　スーパーコンピュータ「不老」の外観写真 2

および**表 2** に示す。システム全体の理論演算性能 (Type I, II, III, およびクラウドに搭載された CPU と Type II に搭載された GPU の倍精度浮動小数点演算性能の合計値) は 15.88 PFLOPS であり、これは旧システムの全体性能 (約 3.7 PFLOPS) の約 4.3 倍である。各サブシステム間は InfiniBand HDR/EDR にて相互に接続されてはいるものの、複数のサブシステムにまたがって 1 つの計算ジョブを実行することは想定されていない。“不老” の最大消費電力は約 1900kVA である。

2.2　Type I サブシステム

Type I サブシステムは、理化学研究所と富士通社が共同開発している「富岳」[3] の商用版として知られる FUJITSU Supercomputer PRIMEHPC FX1000(以下、FX1000)[4] により構成されている。FX1000 は 1 ノードあたり 48 の計算コアと 2 または 4 のアシスタントコアからなる A64FX CPU (Armv8.2-A SVE アーキテクチャ) を 1 基搭載しており、その理論演算性能 (倍精度浮動小数点演算) は約 3.38 TFLOPS である。また主記憶として 1,024 GB/s という高いバンド幅を有する HBM2 メモリを 32GiB 搭載している。また FX100 は高密度な実装となっているため、わずか 6 ラック (各ラックは一般的な PC サーバラックよりは大型) に 2,304 ノードが搭載され、総理論演算性能は 7.782 PFLOPS と全サブシステム間で最も高い。本サブシステムは水冷方式で冷却されており、サブシステム内のノード間は高速で低遅延な Tofu インターコネクト D によって相互接続されている。

当センターでは従来富士通社の FX10 システムと FX100 システムを運用してきていることから、Type I サブシステムにはその後継機としての役割、すなわち従来のシステム向けのソフトウェアを高性能に実行することへの期待が大きい。さらに、審査があり利用者が限られる「富岳」に対して「不老」は利用資格を持つ者であれば利用できることから、「富岳」における超大規模実行を目指す研究者・開発者にとっての開発・テスト環境として、「富岳」向けに開発された高性能ソフトウェアの利用環境として、また新しいアーキテクチャ向けの研究開発環境としてなど、様々な需要が期待される。

2.3　Type II サブシステム

Type II サブシステムは、FUJITSU Server PRIMERGY CX2570 M5 により構成されている。Type II サブシステムは、各計算ノードに 2 基の Intel Xeon Gold 6230 (20 コア、2.10GHz-3.90GHz、Cascade Lake アーキテクチャ) と 4 基の NVIDIA Tesla V100 (Volta アーキテクチャ、SXM2 接続) を搭載した、いわゆる GPU スパコンである。2020 年 5 月により新しい HPC 向け GPU (Ampare アーキテクチャ) が発表されてはいるものの、現行の最新世代である CPU と GPU を搭載したシステムである。特に高性能な GPU を複数搭載しているため、約 33.89 TFLOPS という高いノードあたりの理論演算性能 (倍精度浮動小数点演算) を備えている。ノード数は 221、合計 10 のラックに搭載されており、総理論演算性能は 7.489 PFLOPS と Type I サブシステムに匹敵する。本サブシステムは水冷方式で冷却されており、サブシステム内のノード間は InfiniBand

表 1 スーパーコンピュータ「不老」の性能諸元 1

Type I サブシステム		機種名: FUJITSU Supercomputer PRIMEHPC FX1000
CPU	型番と数量	FUJITSU A64FX (Armv8.2-A + SVE) × 1 ソケット
	コア数と動作周波数	48 コア+2 アシスタントコア, 2.2 GHz (I/O 兼計算ノードは 48 コア+4 アシスタントコア)
	1CPU あたり理論演算性能	3.3792 TFLOPS (DP), 6.7584 TFLOPS (SP), 13.5168 TFLOPS (HP)
	1 ノードあたりメインメモリ	HBM2 32 GiB
	1CPU あたり理論メモリバンド幅	1,024 GB/s (1CMG=12 コアあたり 256 GB/s, 1CPU=4CMG)
ノード数		2,304
ノード間接続		Tofu インターコネクト D 隣接ノード間 40.8 GB/s × 双方向 (6.8 GB/s/link, 6 link 同時通信可能)
冷却方式		水冷
総理論演算性能		7.782 PFLOPS
総メインメモリ容量		72 TiB

Type II サブシステム		機種名: FUJITSU Server PRIMERGY CX2570 M5
CPU	型番と数量	Intel Xeon Gold 6230 (Cascade Lake) × 2 ソケット
	コア数と動作周波数	20 コア, 2.10 - 3.90 GHz
	1CPU あたり理論演算性能	1.344 TFLOPS (DP), 2.688 TFLOPS (SP)
	1 ノードあたりメインメモリ	DDR4 2933 MHz, 384 GiB (32 GiB × 6 枚 × 2 ソケット)
	1CPU あたり理論メモリバンド幅	140.784 GB/s
GPU	型番と数量	NVIDIA Tesla V100 (Volta), upto 1,530 MHz × 4 枚
	1GPU あたりメモリ	HBM2 32 GB, 900 GB/s
	1GPU あたり理論演算性能	7.8 TFLOPS (DP), 15.7 TFLOPS (SP), 125 TFLOPS (Deep Learning)
	ホストとの接続	PCI-Express Gen.3 ×16 (16GB/s)
	GPU 間の接続	NVLink 2, 1GPU あたり合計 300 GB/s (1GPU から他の 3GPU に対してそれぞれ 50 GB/sec × 双方向)
ノード数		221
ノード間接続		InfiniBand EDR 100Gbps × 2 port
冷却方式		水冷
総理論演算性能		7.489 PFLOPS
総メモリ容量		メインメモリ 82.875 TiB, デバイスメモリ 28.288 TiB
ローカルストレージ		各ノードに NVMe SSD 6.4 TB, 総転送性能 707 GB/s 一部のノードにて BeeGFS, BeeOND, NVMesh による共有ファイルシステムを提供

Type III サブシステム		機種名: HPE Superdome Flex
CPU	型番と数量	Intel Xeon Platinum 8280M (Cascade Lake) × 16 ソケット
	コア数と動作周波数	28 コア, 2.70 - 4.00 GHz
	1CPU あたり理論演算性能	2.4192 TFLOPS (DP), 4.8384 TFLOPS (SP)
	1 ノードあたりメインメモリ	DDR4 2933 MHz, 24 TiB (128 GiB × 16 ソケット)
	1CPU あたり理論メモリバンド幅	1126.272 GB/s
GPU	型番と数量	NVIDIA Quadro RTX6000 (Turing) × 4 枚
	1GPU あたりメモリ	GDDR6, 24 GiB
	ホストとの接続	PCI-Express Gen.3 x16 (16GB/s)
ノード数		2
ノード間接続		InfiniBand EDR 100Gbps
冷却方式		空冷
総理論演算性能		77.414 TFLOPS
総メモリ容量		48 TiB
ローカルストレージ		一方のノードに 102.4 TB SSD を搭載 もう一方のノードに 1008 TB ストレージを接続

表 2　スーパーコンピュータ「不老」の性能諸元 2

クラウドシステム		機種名: HPE ProLiant DL560
CPU	型番と数量	Intel Xeon Gold 6230 (Cascade Lake) × 4 ソケット (Type II サブシステムと同じ CPU のため詳細は省略)
	1 ノードあたりメインメモリ	DDR4 2933 MHz, 384 GiB (32 GiB × 3 枚 × 4 ソケット)
	1CPU あたり理論メモリバンド幅	70.392 GB/s
ノード数		100
ノード間接続		InfiniBand EDR 100Gbps
冷却方式		空冷
総理論演算性能		537.6 TFLOPS
総メインメモリ容量		37.5 TiB

ホットストレージ	
MDS	FUJITSU PRIMERGY RX2540 M5 × 4 台
MDT	FUJITSU ETERNUS AF250 S2 × 1 台
OSS/OST	(DDN SFA18KE × 1 台, DDN SS9012 × 10 台) × 4 セット
ストレージ容量	物理容量 40.32 PB, 実効容量 約 30.44 PB
データ保護	RAID6 (8D+2P), データバックアップなし
転送性能	384 GB/sec (最も転送速度が遅い部分の理論性能)

コールドストレージ	
Phase 1 (2020 年 7 月 1 日から)	
機種名	Sony PetaSite Library
搭載カートリッジ数 / 最大搭載可能カートリッジ数	88 巻 / 88 巻
総物理容量 / 最大搭載可能容量	484 TB / 484 TB
搭載ドライブ数	6 台
Phase 2 (2021 年 2 月 1 日から)	
機種名	Sony PetaSite 拡張型 Library
搭載カートリッジ数 / 最大搭載可能カートリッジ数	1728 巻 / 1980 巻
総物理容量 / 最大搭載可能容量	9.504 PB / 10.89 PB
搭載ドライブ数	20 台

EDR によって相互接続されている。

Type II サブシステムは、当センターでは初の大規模な GPU 搭載システムである。特に多数の Volta アーキテクチャ GPU を搭載していていることから、学内外を問わず機械学習・ビッグデータ・AI といった分野の研究に広く活用されることが期待される。また、それらの研究分野ではしばしば I/O 性能が性能のボトルネックになることが知られているため、Type II サブシステムではその対策として各ノードに NVMe SSD を搭載し、さらにノードを跨いでローカルな共有ファイルシステムを提供する BeeGFS, BeeOND, および NVMesh も利用可能である。加えて、機械学習などの分野で需要の大きいコンテナの利用にも対応する。これらを有効に活用することで、当センターではこれまで十分に対応できていなかった新たな分野の需要にも応えていく予定である。

2.4　Type III サブシステム

Type III サブシステムは、2 ノードの HPE Superdome Flex によって構成されている。Type III サブシステムの各ノードには Intel Xeon Platinum 8280M が 16 基ずつ、NVIDIA Quadro RTX6000 が 4 枚ずつ搭載されている。そのためノードあたりの理論演算性能 (倍精度浮動小数点演算) は約 38.71 TFLOPS と全サブシステムで最も高い。しかし Type III サブシステム最大の特徴はノードあたり 24 TiB という大規模な主記憶を搭載している点であり、大規模なプリ・ポスト処理や可視化処理への活用が期待される。近年のスーパーコンピュータシステムは理論演算性能やメモリ転送性能の向上に比べてノードあたりのメモリ容量があまり増加していない傾向があるが、大容量のメモリを利用したい需要も根強い。そのため当センターでは従来から継続してこのような大規模メモリシステムを提供し、その需要に応えている。本サブシステムは空冷方式で冷却されており、サブシステム内のノード間は InfiniBand EDR によって相互接続されているが、1 ノードは会話型処理、もう 1 ノードはバッチ型処理と異なる利用方式で運用予定である。

Type III サブシステムは当センターの 1 階にある可視化室に設置されている 185 インチ 8K 高精細ディスプレイなど可視化用の設備と組み合わせて使えるよう設計されているため、コンテンツの作成や閲覧などにも活用されることが期待される。

2.5 クラウドシステム

クラウドシステムは、100 ノードの HPE ProLiant DL560 により構成されている。クラウドシステムの CPU は Type II サブシステムと同様の Intel Xeon Gold 6230 であるが、クラウドシステムではノードあたりの搭載数は 4 基である。そのため、ノードあたりの理論演算性能 (倍精度浮動小数点演算) は、CPU だけで比較すれば約 5.38 TFLOPS と Type I, II の各サブシステムよりも高く、スレッド並列化のみ適用されている計算インテンシブなプログラムの高速実行に適している。ただし本サブシステムの冷却方式は空冷方式であるため TurboBoost 機能による性能向上具合が抑えられ、Type II サブシステムの 2 倍のノードあたり CPU 演算性能までは得られないと考えられる。主記憶容量は Type II と同様に 384 GiB である。クラウドシステム内のノード間は InfiniBand EDR によって相互接続されている。

クラウドシステムは、スーパーコンピュータシステムでは一般的であるバッチジョブシステムを介した利用に加えて、Web ベースの予約システムを介した時刻指定実行にも対応したシステムである。時刻指定実行には、指定した時刻になるとバッチジョブが実行される時刻指定バッチジョブ実行と、指定した時間帯にのみ計算ノードに SSH ログインして対話型処理が可能となる時刻指定インタラクティブ実行の二種類の実行方式が用意されている。そのためクラウドシステムには、Type I, II 各サブシステムの混雑緩和とノードあたり CPU 性能の高い x86 CPU 環境の提供に加えて、インタラクティブ処理環境の提供により研究室の PC サーバ/ワークステーションなどの日常的な作業環境を代替するなど、様々な役割が期待されている。

2.6 共有ストレージ

“不老” は 2 種類の大規模な共有ストレージを有している。

ホットストレージは、スーパーコンピュータシステムでは一般的である、多数の HDD によって構成される共有 RAID ストレージである。保存されたデータの情報を保持するメタデータサーバ (MDS)・メタデータストレージ (MDT) は FUJITSU PRIMERGY RX2540 M5 および FUJITSU ETERNUS AF250 S2 により構成されており、データ自体は DDN SFA18KE および DDN SS9012 に保持される。ホットストレージのデータは RAID6(8D+2P) により保護されており、総容量は物理容量 40.32 PB、実効容量約 30.44PB である。従来のストレージは実効容量約 6PB であったため、約 5 倍に増量されている。

ストレージ容量は約 5 倍に増量されたものの、近年はデータ科学研究の発展にともない非常に大容量のデータを生成する研究が増加している。特に、観測機器（センサ）から得られるデータや自動運転に用いる環境データなど、生成されたら変更は行わない（変更してはならない）大容量データや、アクセス頻度は低いが重要性は高い大容量データを扱う需要が急速に増加している。HDD や SSD から構成されるストレージは最大容量にあわせて消費電力も増加し、また記録媒体を取り出して持ち出すことも現実的ではない。そこで “不老” ではこれらの需要に効率よく対応するため、追記書き込みのみ可能で耐久性の高い光ディスク・アーカイブ (Optical Disc Archive) を用いたコールドストレージを、スーパーコンピュータシステムとしては世界に先駆けて導入した。コールドストレージはソニー社の PetaSite ライブラリ/PetaSite 拡張型ライブラリにより構成され、2020 年 7 月の運用開始時点 (Phase1) では総容量 484 TB、2021 年 2 月 (Phase2) には約 9.5 PB に拡張される予定である。

2.7 その他の構成ハードウェア、ソフトウェア

ここまでに述べた装置群に加えて、FUJITSU PRIMERGY RX2530 M5 および RX2540 M5 から構成される合計 25 台のフロントエンドシステム (主にログインノード群として利用)、USB 記憶装置などを持ち込んで “不老” のデータを読み書きする際などに便利なオンサイト利用装置、“不老” と可視化設備に接続されている Windows 端末である画像処理装置、その他の管理ノード群などにより “不老” は構成されている。

“不老” は商用ソフトウェアを含め様々なソフトウェアを提供している。特に、Type I サブシステムは CPU アーキテクチャが他の計算サブシステムとは異なり、また Type II サブシステムは GPU を搭載しているため、ソフトウェアごとに利用が推奨される計算サブシステムが異なる。計算サブシステムとソフトウェアの対応情報など最新の情報については Web ページ [5] にて提供し、継続して更新していく予定である。

2.8 省電力/低消費電力化に向けた取り組み

多くの電力を消費するスーパーコンピュータシステムにとって、消費電力の削減は重要な課題である。当センターでは従来より建物地下に水温 18 度の湧き水が湧いていたため、これを機器の冷却補助に用いることとした。具体的には、屋外に設置されたチラーに対して霧状に吹きかけることでチラーの温度を下げ、チラーの消費電力の削減を行う。これにより、年間 100 万円以上に相当する電気料金の削減が期待されている。さらに、大学全体の電力使用量が大きい夏季の昼間を避けて動作する計算ノード (計算キュー) を

設ける、消費電力が大きい場合に自動的に実行するジョブの数を抑える動的制御機構を設ける、などピーク電力の削減を行う仕組みも用意し、エコなスパコンを目指している。

3. スーパーコンピュータ「不老」の性能

“不老” の各サブシステムを用いて測定した性能を報告する。本稿に掲載している性能値はいずれも実機による性能値である。ただし “不老” はより良いサービス提供のためにシステム設定・パラメタを調整することがあるため、サービス運用開始後には同一の対象問題でも性能が異なることがある。

3.1 HPL ベンチマーク

High Performance Linkpack (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性能を求めるベンチマークである。このベンチマークは、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) の性能に大きく影響を受けることが知られている。また、世界中のスーパーコンピュータシステムの性能を順位付けする TOP500 および Green500[7] の指標としても、本ベンチマークが用いられている。

TOP500 ランキング 2020 年 6 月版において、“不老” の Type I サブシステムは 6,617.8 TFLOPS で 36 位 (2020 年 6 月 28 日現在) にランクインしている。日本国内の順位では、「富岳」、AI Bridging Cloud Infrastructure (ABCI)、Oakforest-PACS、TSUBAME3.0 につづく 5 位である。このときの問題サイズ (Nmax) は 2,621,440、理論演算性能 (Rpeak) に対する測定値 (Rmax) は約 85%である。コンパイラや MPI については Fujitsu Software Technical Computing Suite V4.0L20 に含まれるものを用いた。なお、前回の 2019 年 11 月版における FX100 システムの性能と順位は 2910 TFLOPS で 84 位であった。

“不老” の Type II サブシステムは、導入時期の都合により TOP500 ランキングには掲載されていないが、221 ノード全てを用いた実測値として 4,880 TFLOPS が確認されている。この性能は TOP500 ランキング 2020 年 6 月版の 50 位に相当する。このときの問題サイズ (Nmax) は 1,880,064、理論演算性能 (Rpeak) に対する測定値 (Rmax) は約 65%である。コンパイラや MPI については、Intel Parallel Studio XE Cluster Edition 2018 update4、CUDA10.2、OpenMPI-3.1.3 を用いた。

3.2 HPCG ベンチマーク

HPCG ベンチマーク [8] は、有限要素法から得られる疎行列を対象として共役勾配法 (Conjugate Gradient method, CG 法) を用いて連立一次方程式を解く性能を求めるベンチマークである。このベンチマークは、行列積和演算の性能に大きく左右される HPL ベンチマークよりも現実のアプリケーションに近いものとして提案されたベンチマークである。

HPCG ランキング 2020 年 6 月版において、“不老” の Type I サブシステムは 230.594 TFLOPS で 16 位にランクインしている。日本国内の順位では、「富岳」、AI Bridging Cloud Infrastructure (ABCI)、Oakforest-PACS につづく 4 位である。HPCG は HPL と比べてピーク性能に対する実測性能の比が非常に低いことが知られており、“不老” Type I サブシステムの性能比は 3.0%である。しかしこの性能比は現在のスーパーコンピュータにおいてはとても高い値であり、“不老” よりも上位で 2.0%に到達しているのは HPCG スコア世界一の「富岳」(2.6%) のみである。なお、2019 年 11 月版における FX100 システムの性能と順位は 86.534 TFLOP で 31 位であった。コンパイラや MPI については Fujitsu Software Technical Computing Suite V4.0L20 に含まれるものを用いた。

“不老” の Type II サブシステムについては導入時期の都合により全ノードを利用した測定が行えていない。現時点では 100 ノードを用いたテスト実行 (ランキング登録には 3600 秒以上の実行時間が必要だが、今回はテスト実行として 60 秒だけ動かした) を行った結果として 48254.4 GFLOPS の性能が確認できている。今後システム全体での測定を行う予定であるが、100 ノードの結果から単純に推測すると 100 TFLOPS 程度の性能が期待される。実行には HPCG ベンチマークの Web サイトで配付されている 2019 年 12 月 5 日版の HPCG 3.1 バイナリを用いた。

3.3 GKV ベンチマーク

GKV ベンチマークは、磁場閉じ込め核融合に向けたプラズマ乱流現象の解析のために核融合科学研究所にて開発されたプラズマ乱流解析コード GKV (GyroKinetic Vlasov code) を元に、名古屋大学の片桐・渡邊らがそのカーネル部分を抜き出して作成したベンチマークである。本ベンチマークは以下の 4 つのプログラムにより構成される。

kernel1_fft FFT と MPI_Alltoall に関するベンチマーク。MPI と OpenMP によるハイブリッド並列化コード。

kernel2_intgrl 配列のリダクションに関するベンチマーク。OpenMP によるスレッド並列化コード。

kernel3_diff45 4 次元および 5 次元の有限差分法の演算に関するベンチマーク。OpenMP によるスレッド並列化コード。

kernel4_diff123 1 次元, 2 次元および 3 次元の有限差分法の演算に関するベンチマーク。OpenMP によるスレッド並列化コード。

以上について、Type I サブシステムと Type II サブシステムで実行時間を測定し比較した。プログラムのコンパイルには、Type I サブシステムでは富士通コンパイラ (frtpx (FRT) 4.2.0 20200612) を、Type II サブシステムではイン

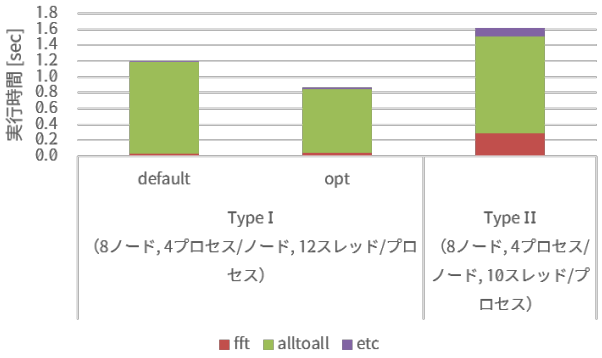

図 4 GKV ベンチマーク kernel1 の実行時間内訳

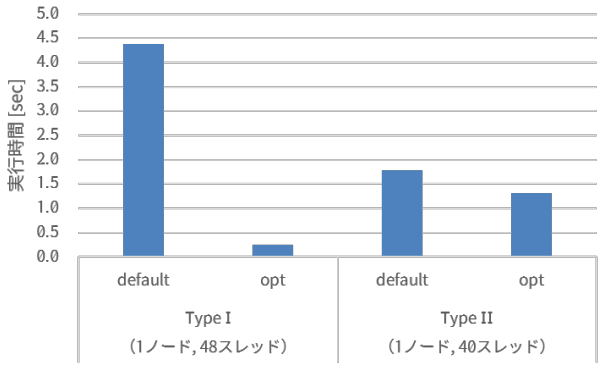

図 5 GKV ベンチマーク kernel2 の実行時間

テルコンパイラ (ifort (IFORT) 19.0.5.281 20190815) を用いた。結果を**図 4** から**図 7** に示す。

kernel 1 (図 4) については FFT と MPI_Alltoall とその他の時間の内訳を示している。Type I サブシステムの 'default' は GKV ベンチマークのコードは改変せずコンパイル時のオプションや実行時のオプションを調整したもの、'opt' はさらに実行時のオプション指定により Alltoall のアルゴリズムを最適化したものである。Type I サブシステムについては`-Kfast -Kparallel -Kopenmp -Knolargepage -Kloop_nofission -Nfjomplib -Nlst=t -lfftw3` オプションを付けてコンパイルし、環境変数 `OMP_STACKSIZE=8m, OMP_NUM_THREADS=12` を付けて実行した。'opt' については mpiexec に引数`--mca coll_select_alltoall_algorithm 101` を追加した。Type II サブシステムについては`-Ofast -xCASCADELAKE -qopenmp -mkl` オプションを付けてコンパイルし、環境変数 `OMP_STACKSIZE=8m, OMP_NUM_THREADS=10, I_MPI_PIN_DOMAIN=omp` を付けて実行した。いずれも 8 ノード実行、各ノード上では 4MPI プロセスを立ち上げ、各プロセスあたりのスレッド数は CPU コア数をプロセス数で割った値 (Type I では 12、Type II では 20) とした。Type II サブシステムに対する Type I サブシステム ('opt') の実行時間比は、FFT については 14.30%、Alltoall については 66.00%と、いずれも Type I サブシステムの性能が Type II サブシステムの性能を大きく上回る結果となった。

kernel 2 (**図 5**) については、'default' は GKV ベンチマークのコードは改変せずコンパイル時のオプションや実行時のオプションを調整したもの、'opt' は Type I サブシステムの性能が最大化されるよう OpenMP 指示文の挿入位置を調整したものである。Type I サブシステムのコンパイルオプションは`-Kfast -Kocl -Kparallel -Kopenmp -Klargepage -Kloop_nofission -Ksimd_nouse_multiple_structures -Kmfunc=2 -Nfjomplib -Nlst=t`、Type II サブシステムのコンパイルオプションは`-Ofast -xCASCADELAKE -qopenmp` を指定した。プログラム実行時には、Type I サブシステムでは `OMP_NUM_THREADS=48, OMP_STACKSIZE=8m, XOS_MMM_L_PAGING_POLICY=demand:demand:demand` を指定し、さらに 'opt' では `numactl --interleave 4-7 --physcpubind 12-59` として計算コアの割り当てを最適化した。Type II サブシステムでは `OMP_STACKSIZE=8m, OMP_NUM_THREADS=40` を指定した。OpenMP 並列化のみ適用されているため、いずれも 1 ノード実行、スレッド数は CPU コア数と同様とした。Type I サブシステムは 'default' から 'opt' にすることで大きく性能が向上、一方 Type II サブシステムも性能は向上したがその度合いは小さかった。'opt' 同士を比較すると、Type I サブシステムの実行時間は Type II サブシステムの 19.57%であり、Type I サブシステムの性能が Type II サブシステムを大きく上回る結果となった。

kernel 3 (**図 6**) は GKV ベンチマークのコードを変更せずそのまま測定した。Type I サブシステムのコンパイルオプションは`-Kfast -Kparallel -Kopenmp -Klargepage -Kloop_nofission -Kprefetch_sequential=soft -Ksimd_nouse_multiple_structures -Nfjomplib -Nlst=t`、Type II サブシステムのコンパイルオプションは`-Ofast -xCASCADELAKE -qopenmp` を指定した。プログラム実行時には、Type I サブシステム, Type II サブシステムともに kernel 2 の 'opt' と同様の指定を行い、1 ノード実行、スレッド数は CPU コア数と同様とした。実行時間を比較すると、Type I サブシステムの実行時間は Type II サブシステムの 15.48%であり、Type I サブシステムの性能が Type II サブシステムを大きく上回る結果となった。

kernel 4 (図 7) については、'default' は GKV ベンチマークのコードは改変せずコンパイル時のオプションや実行時のオプションを調整したもの、'opt' は Type I サブシステムの性能が最大化されるようループの分割や配列の次元入れ替えなどを行ったものである。Type I サブシステムのコンパイルオプションは`-Kfast -Kocl -Kparallel -Kopenmp -Knolargepage -Kloop_nofission -Nfjomplib -Nlst=t`、Type II サブシステムのコンパイルオプションは`-Ofast -xCASCADELAKE`

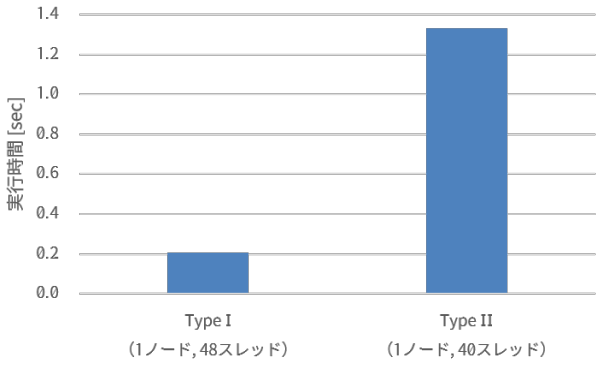

図 6 GKV ベンチマーク kernel3 の実行時間

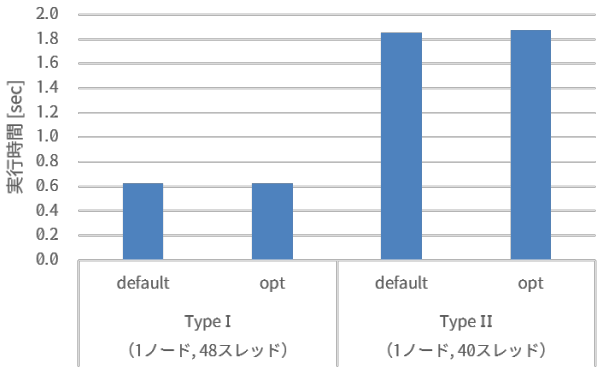

図 7 GKV ベンチマーク kernel4 の実行時間

`-qopenmp` を指定した。プログラム実行時には、Type I サブシステム, Type II サブシステムともに kernel 2 の 'opt' と同様の指定、ただし Type I サブシステムの 'opt' 実行時の numactl 指定は `numactl --membind 4-7 --physcpubind 12-59` に変更した。いずれも 1 ノード実行、スレッド数は CPU コア数と同様とした。'opt' 同士の実行時間を比較すると、Type I サブシステムの実行時間は Type II サブシステムの 33.40%であり、Type I サブシステムの性能が Type II サブシステムを大きく上回る結果となった。

以上のように、GKV ベンチマークについては全体的に Type I サブシステムの性能が Type II サブシステムの性能を大きく上回る結果となった。ただし、ソースコードの修正は Type I サブシステム上での実行性能向上に向けたものしか行っておらず、Type II サブシステム向けに最適化を行う余地がある可能性も残されている。Type I サブシステム向けに指定した様々なコンパイルオプションや実行時オプションについては様々な情報を元にできるだけ性能が高くなるものを選択しているが、当センターとしては、適切なオプションを容易に選択する指標などについて検討し利用者に情報提供していくことも重要だと考えている。

3.4 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り、様々な通信形態に対応する評価プログラムを提供している。今回は原稿執筆時点での最新版である 5.6.3 を用いて、Type I サブシステムと Type II サブシステムの通信性能を評価した。

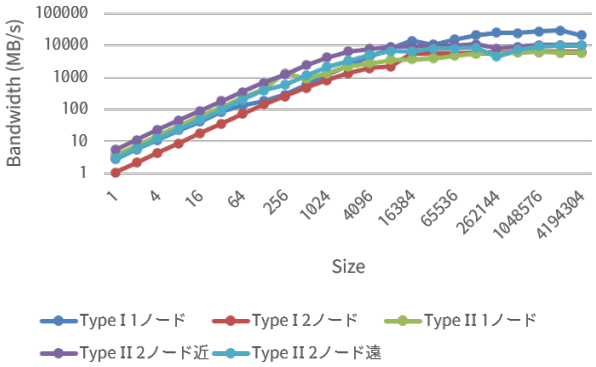

図 8 Bandwidth, 2 プロセス間通信

はじめに、2 プロセス間のバンド幅を測定する osu_bw およびレイテンシを測定する osu_latency の結果を示す。Type I サブシステムについては 1 ノード内の 2 プロセス (1 ノード) および 2 ノードに 1 プロセスずつ (2 ノード) の 2 パターンを測定した。Type II サブシステムについては、ノード内に 2CPU ソケットが存在し、ネットワークカードは片方の CPU ソケット側の PCI-Express にのみ接続されているため、1 ノード内の 2CPU それぞれで 1 プロセスずつ起動した場合 (1 ノード)、2 ノードそれぞれでネットワークカードが接続されている CPU 側に 1 プロセス起動した場合 (2 ノード近)、2 ノードそれぞれでネットワークカードが接続されていない CPU 側に 1 プロセス起動した場合 (2 ノード遠) の 3 パターンを測定した。

測定結果を**図 8** および**図 9** に示す。バンド幅については、小さなサイズの場合に Type II サブシステムの 1 ノードよりも 2 ノード近が速いことや Type I サブシステムの 2 ノードが遅いこと、大きなサイズの場合に Type II サブシステムの 1 ノードが遅い点が目立っている。Type I サブシステムについては、周囲の複数ノードと同時通信できるという Tofu インターコネクト D の特徴がこの問題設定では活かせていないと考えられる。レイテンシについては、ノード内は高速でノード間は時間がかかるというわかりやすい結果が得られているが、Type II サブシステムの 1 ノードについてはサイズが小さい場合に非常に速いがサイズが大きくなると急激に遅くなっている点が目立っている。

つづいて、12 ノードおよび 120 ノードに 1 プロセスずつ配置した際の MPI_Barrier, MPI_Allreduce および MPI_Alltoall 集団通信の性能を示す。Type I サブシステムについては Tofu インターコネクト D の接続が 12 ノード単位になっており、torus 形状と mesh 形状を選ぶことができるため、それぞれについて評価を行った。ただし三次

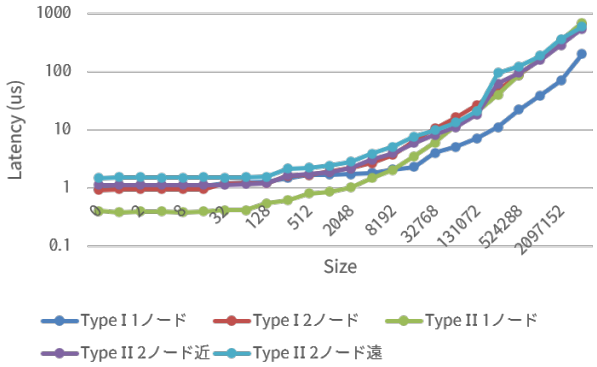

図 9 Latency, 2 プロセス間通信

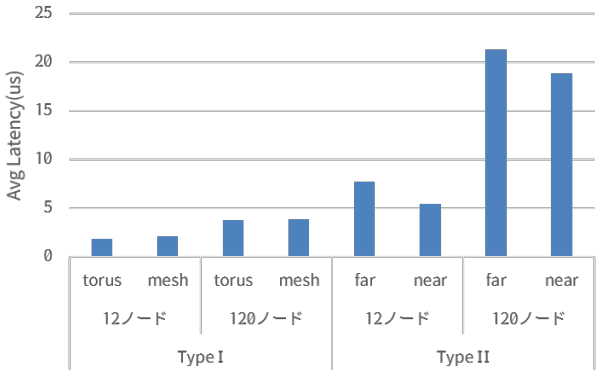

図 10 MPI_Barrier

元のノード指定は行わず、“12:mesh” や “12:torus” といった一次元の指定のみを行った。Type II サブシステムについては、ネットワークカードが接続されている CPU 側にのみ MPI プロセスを配置したもの (near) と、接続されていない側のみに配置したもの (far) を測定して比較した。

測定結果を**図 10** から**図 12** に示す。

MPI_Barrier については、Type I サブシステムが高速に行えており、12 ノードから 120 ノードに増やしてもあまり時間は増加しなかった。torus と mesh の差はほとんどなかった。Type II サブシステムは Type I サブシステムと比べると長い時間がかかっており、12 ノードから 120 ノードに増やした場合の実行時間増加も大きかった。また、MPI プロセスをネットワークカードが接続された側の CPU ソケットに置くか否かは有意な差となり、near の方が far よりも高速に同期を行うことができた。

MPI_Allreduce については、通信サイズが非常に小さい場合には 12 ノード 120 ノードともに Type I サブシステムが高速であるが、それ以外では Type II サブシステムの 12 ノードが高速という結果となった。一方 MPI_Alltoall については、通信サイズが小さめの場合には Type II サブシステムが高速だが、通信サイズが大きな場合には Type I サブシステムの方が高速という結果となった。

集団通信の性能については得意とする問題サイズの影響が強く出ているように見受けられるが、ハードウェアだけではなく MPI ライブラリの特徴が出ている可能性があり、さらに環境変数などの設定によってアルゴリズムの調整ができる余地もあることから、実際のアプリケーションでの利用においては通信パターンに合わせた最適化の余地も大きいと考えられる。

3.5 ストレージ性能

ストレージ性能として、ホットストレージ上で行った IOR ベンチマーク [10] の結果を報告する。

“不老” のホットストレージは全サブシステムから共通にアクセス可能である。ただし、1 台の MDT、4 台の MDS、4 組の OSS/OST によって構成されており、その構成には

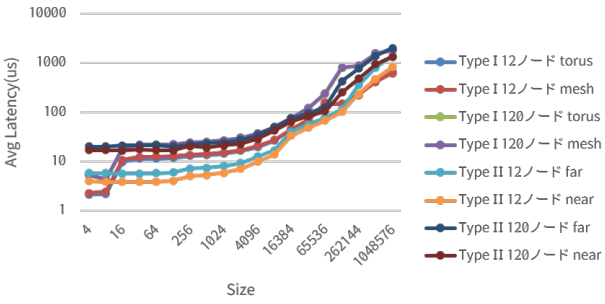

図 11 MPI_Allreduce

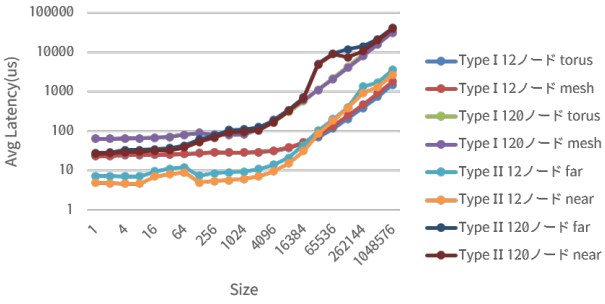

図 12 MPI_Alltoall

幾つか選択肢がある。“不老” では、実際の利用シナリオを想定したストレージの負荷分散とスループット向上や耐障害性などを考慮し、OST_pool とマルチ MDS 機能を用いてアクセスの分散を図っている。構成の概要を**図 13** に示す。/home には一般ユーザのホームディレクトリを配置し、/data/group1 は従来のシステムからの引き継ぎデータや追加で大容量のストレージを要求するユーザ用に利用、/data/group2 は特に I/O 負荷の大きなジョブ (ヘビーユーザ) を隔離するために用いる予定である。

IOR ベンチマークの File Per Process(FPP) 性能評価について、OST の各領域、/data/group1, /data/group2, /home, /data/tmp それぞれで読み書きの性能を測定した。さらに、/data/group1 と/data/group2 については Single Shared File(SSF) 性能についても性能を測定した。各ディレクトリのクライアント数と性能を**表 3** に示す。性能評価

表 3　IOR ベンチマークの結果

操作対象 ディレクトリ	OST クライアント数	File Per Process 性能 write 平均値 [MiB/sec]	File Per Process 性能 read 平均値 [MiB/sec]	Single Shared File 性能 write 平均値 [MiB/sec]	Single Shared File 性能 read 平均値 [MiB/sec]
/data/group1	576	93130.36	106612.15	84112.94	83286.16
/data/group2	288	33770.22	38685.99	32984.10	36369.85
/home	144	17928.59	22037.02		
/data/tmp	144	12697.28	13242.41		
		write+read 合計値	169052.000	write+read 合計値	118381.525

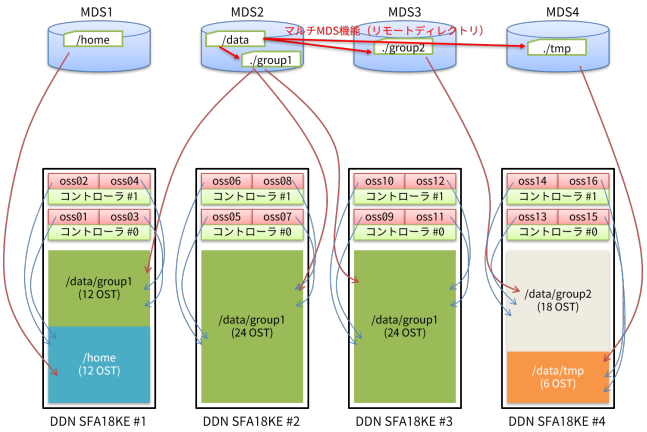

図 13　ストレージの構成

結果から、多数のユーザが同時に利用する/data/group1 領域については読み込みと書き込みそれぞれ 100GiB/sec 程度、少数のヘビーユーザ向けの/data/group2 領域については 35GiB/sec 程度の性能を提供できることが確認できた。

“不老” はグローバルなホットストレージ以外にも、Type II サブシステムや Type III サブシステムに専用のローカルストレージを搭載しており、これらは一般的なバッチジョブから利用することが想定されている。今後はこれらの性能評価や、一般ユーザが容易に使えるような提供方法の検討・マニュアル整備をしていく予定である。

4. おわりに

本稿では、2020 年 7 月 1 日よりサービス開始予定であるスーパーコンピュータ「不老」の設計を紹介し、またベンチマークプログラムにより得られた性能評価結果を示した。スーパーコンピュータ「不老」は、「富岳」型の計算ノードを世界で初めて本格サービス運用するとともに、全国共同利用のスーパーコンピュータではまだ提供数が少ない Tesla V100 GPU を多数提供し、光ディスクによるコールドストレージシステムをスーパーコンピュータでは初めて導入、さらに湧水を用いた冷却システムを導入するなど、非常に先進的で挑戦的なシステムである。安定したサービス運用を提供すると共に、今後も本システムを用いた性能評価結果などを提供し、ユーザによる本システムを用いた研究をサポートするとともに、計算科学・計算機科学・データ科学分野の研究の発展に寄与していきたい。

謝辞　システムの導入と性能評価にご協力いただいた名古屋大学情報統括本部情報基盤課の皆様と富士通株式会社の皆様に感謝します。

参考文献

[1] FUJITSU Supercomputer PRIMEHPC FX10 - 富士通 https://www.fujitsu.com/jp/products/computing/servers/supercomputer/primehpc-fx10/ (accessed 2020-06-26)
[2] FUJITSU Supercomputer PRIMEHPC FX100 - 富士通 https://www.fujitsu.com/jp/products/computing/servers/supercomputer/primehpc-fx100/ (accessed 2020-06-26)
[3] スーパーコンピュータ「富岳」について — 理化学研究所 計算科学研究センター (R-CCS) https://www.r-ccs.riken.jp/jp/fugaku (accessed 2020-06-26)
[4] FUJITSU Supercomputer PRIMEHPC シリーズ : 富士通 https://www.fujitsu.com/jp/products/computing/servers/supercomputer/ (accessed 2020-06-26)
[5] スーパーコンピュータシステム — 名古屋大学 情報連携推進本部 http://www.icts.nagoya-u.ac.jp/ja/sc/ (accessed 2020-06-26)
[6] HPL - A Portable Implementation of the High-Performance Linpack Benchmark for Distributed-Memory Computers https://www.netlib.org/benchmark/hpl/ (accessed 2020-06-20)
[7] Home — TOP500 Supercomputer Sites https://www.top500.org/ (accessed 2020-06-26)
[8] https://www.hpcg-benchmark.org https://www.hpcg-benchmark.org/ (accessed 2020-06-26)
[9] MVAPICH :: Benchmarks https://mvapich.cse.ohio-state.edu/benchmarks/ (accessed 2020-06-28)
[10] hpc/ior: IOR and mdtest https://github.com/hpc/ior (accessed 2020-06-28)