次期システム導入に向けた GPU へのプログラム移行支援の取組

中張 遼太郎 1)，山田 新 1)，前田 光教 1)，佐藤 孝明 1)，山崎 一哉 2)，三木 洋平 2)，
下川辺 隆史 2)，住元 真司 2)，塙 敏博 2)，中島 研吾 2)

1）東京大学 情報システム部情報基盤課

2）東京大学 情報基盤センター

nakahari@cc.u-tokyo.ac.jp

Initiatives to Support Program Porting to GPUs towards the Introduction of the Next Generation Supercomputer System

Ryotaro Nakahari 1), Hajime Yamada 1), Mitsunori Maeda 1), Takaaki Satoh 1), Kazuya Yamazaki 2), Yohei Miki 2), Takashi Shimokawabe 2), Shinji Sumimoto 2), Toshihiro Hanawa 2), Kengo Nakajima 2)

1) Information Technology Group, Information Systems Department, The University of Tokyo
2) Information Technology Center, The University of Tokyo

概要

東京大学情報基盤センターと筑波大学計算科学研究センターが共同で運営する最先端共同 HPC 基盤施設（JCAHPC）では 2025 年 1 月運用開始を目指して次期システム（OFP-II）の導入を進めている．OFP-II は GPU 搭載ノードを中心としたシステムとする予定であり，これまでの汎用 CPU を中心としたシステム利用者のプログラム移行が大きな課題である．本稿では，GPU へのプログラム移行支援の取組と成果に関して報告する．

1. はじめに

東京大学情報基盤センター（以下，本センターと記す）では，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/H/L スーパーコンピュータシステム）[1]，大規模超並列スーパーコンピュータシステム（Oakbridge-CX スーパーコンピュータシステム）[2]，「計算・データ・学習」融合スーパーコンピュータシステム（Wisteria/BDEC-01 スーパーコンピュータシステム）[3]，本センターと筑波大学計算科学研究センターが共同で運営する最先端共同 HPC 基盤施設（以下，JCAHPC と記す）[4]が導入したメニーコア型スーパーコンピュータシステム（Oakforest-PACS スーパーコンピュータシステム）[5]など，多数のシステムを運用してきた．現在は，2025 年 1 月運用開始を目指して Oakforest-PACS の後継システム（以下，OFP-II と記す）の導入を JCAHPC として進めている．また，2023 年 11 月運用開始予定で OFP-II のプロトタイプとして Wisteria-Mercury という小型システムの導入も進めている．スーパーコンピュータへの性能要求とともに省電力・脱炭素化という昨今の状況を考慮すると GPU 等の演算加速装置の導入は不可避と考え，OFP-II は GPU 搭載ノードを中心としたシステムとなる想定である．

これまで運用してきたシステムは汎用 CPU を中心としたシステムが多く，GPU 搭載ノードを中心としたシステムの導入に際しては，システム利用者の CPU 対応プログラムを利用者自身で GPU へ移行することが大きな課題となる．上記を踏まえて，本センターおよび JCAHPC では，GPU へのプログラム移行を支援する取組を 2022 年 11 月より実施してきた．本稿では，GPU へのプログラム移行支援の取組と成果を報告する．

2. スーパーコンピュータシステム運用状況

2.1 運用スーパーコンピュータシステム概要

直近 5 年度の間に運用してきたスーパーコンピュータシステムの概要を表 1 に示す．各スーパーコンピュータシステムの特長としては下記があげられる．

表 1 スーパーコンピュータシステム概要

システム名	総理論演算性能	総ノード数	総主記憶容量	運用開始時期
Reedbush-U/H/L	3.36PFLOPS (U:0.51, H:1.42, L:1.44)	604 (U:420, H:120, L:64)	151TB (U:105, H:30, L:16)	2016 年 7 月
Oakforest-PACS	25.00PFLOPS	8208	897TB	2016 年 12 月
Oakbridge-CX	6.61PFLOPS	1368	282TB	2019 年 7 月
Wisteria/BDEC-01	33.10 PFLOPS (Odyssey:25.9, Aquarius:7.2)	7725 (Odyssey:7680, Aquarius:45)	286TB (Odyssey:258, Aquarius:38)	2021 年 5 月

- Reedbush-U/H/L
  - Reedbush-H/L は本センターで初めての GPU 搭載ノードで構成されたシステム
- Oakforest-PACS
  - メニーコア型プロセッサを搭載
  - 国内 2 位，世界第 16 位の演算能力（2019 年 6 月 TOP500 より）
- Oakbridge-CX
  - 一部ノードに SSD 搭載
- Wisteria/BDEC-01
  - 「富岳」と同じ計算ノード群（Odyssey）と GPU 搭載ノード群(Aquarius)で構成

2.2 システム利用者分析

本センターで運用してきたスーパーコンピュータシステムは多数の利用者を有し，東京大学に限らず幅広い機関の方に利用されている．直近 5 年度の総利用者数と所属機関別利用者数の推移を図 1 に示す．当該年度に複数のスーパーコンピュータシステムを運用している場合は重複を除いて合算した値である．利用者は約 3,000 名であり，東京大学以外の利用者が 55%以上を占めている．国内企業所属の利用者は約 10%を占めており，国外機関所属の利用者は割合としては約 3%と少ないが各年度約 100 名が利用している．

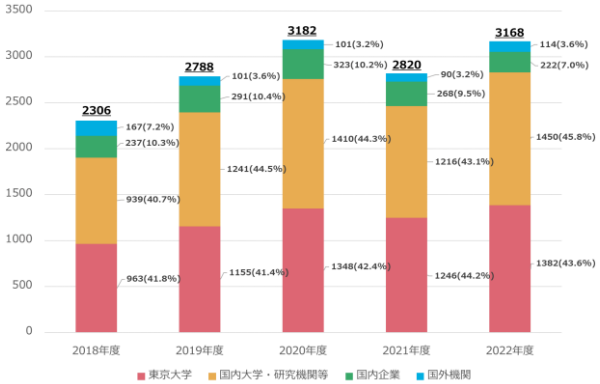

図 1 総利用者数と所属機関別利用者数の推移

2.3 システム利用状況

直近 5 年度の間に運用してきたスーパーコンピュータシステムの利用率推移を図 2 に示す．全体として，年度の始まりである 4 月は利用率が落ち込み，年度末の 1 月から 3 月にかけて利用率が上昇する傾向がある．汎用 CPU 機である Reedbush-U, Oakforest-PACS, Oakbridge-CX については利用率が 80%を超える時期もあり，多く利用されている．一方で，本センターで初めての GPU 搭載機である Reedbush-H, Reedbush-L については運用開始当初の利用は低調であるものの、徐々に利用率が上昇し，年度末には 80%を超える月もある．2021 年 5 月に運用を開始した Wisteria/BDEC-01 は利用率が徐々に上昇しており，60%を超える月もある．

各スーパーコンピュータシステム利用グループの研究分野を，プロジェクト名称とプロジェクト内容に基づいて 13 分野に分類し，スーパーコンピュータシステムごとに利用状況を集計した結果を図 3 に示す．年度の途中で運用を終了したシステムは運用終了の前年度，年度末に運用を終了したシステムは最終運用年度，現在も運用中のシステムは 2022 年度を対象に集計を実施した．

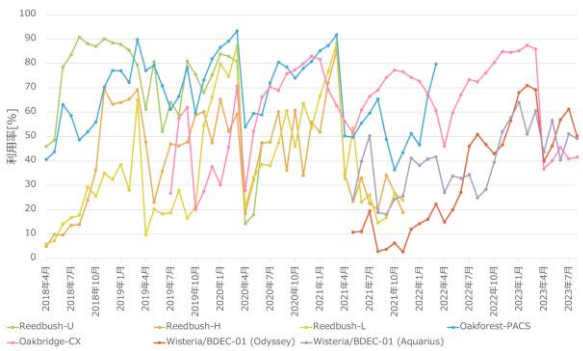

図 2 スーパーコンピュータ利用率推移

それぞれの特色として，相対的に小規模な CPU 機である Reedbush-U，Oakbridge-CX においては工学・ものつくり系や材料科学系の分野での利用が多いのに対して，大規模な CPU 機である Oakforest-PACS，Wisteria/BDEC-01 Odyssey では，豊富な計算資源が必要となる地球科学・宇宙科学系，エネルギー・物理学系の利用が大勢を占めた．一方で，GPU を搭載する Reedbush-H/L，Wisteria/BDEC-01 Aquarius は，材料科学系，情報科学：AI 系，生物科学・生体力学系の利用割合が高い．CPU 機と GPU 機で利用分野が大きく異なる要因として，Deep Learning に関連する研究が活発化したこと，材料科学系で利用の多い GROMACS[6]が GPU に対応していたこと，AlphaFold[7] のオープンソース化等が考えられる．

実行ジョブにおける利用ノード数，GPU 数の利用者ごとの最頻値を当該利用者の主な利用ノード数，GPU 数と推定し，2022 年度の主な利用ノード数，GPU 数の分布を図 4，図 5，図 6 に示す．いずれも，最小の利用単位である 1 ノードもしくは 1GPU を主に利用する利用者が大半を占めている．Wisteria/BDEC-01 Odyssey では 12 ノードの利用者が 1 ノードの利用者に次いで多いが，インターコネクトである Tofu インターコネクト D[8]における最小の構成単位が 12 ノードであることに起因すると推測される．

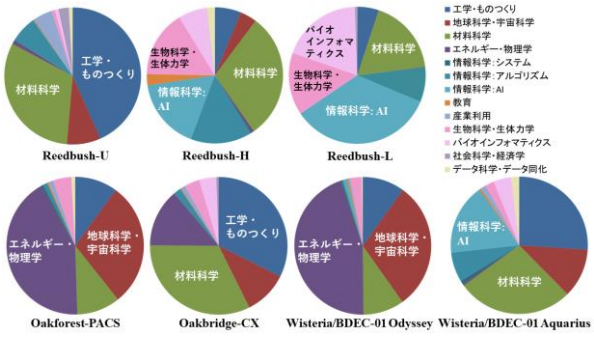

図 3 スーパーコンピュータシステム別利用分野

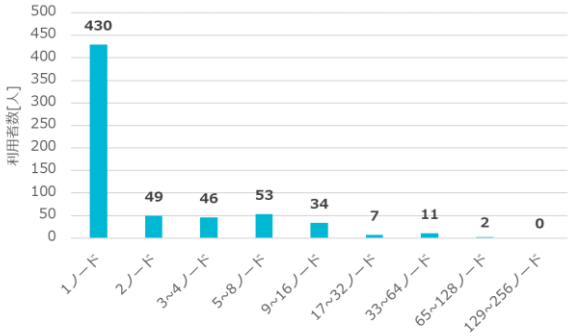

図 4 Oakbridge-CX 2022 年度利用ノード数分布

Wisteria/BDEC-01 Aquarius では GPU 単位(1，2，4GPU)で指定するジョブキューとノード単位で指定するジョブキューの大きく 2 種の不具合の可能性が考えられるため，実行時間が 1 分以下のジョブも集計対象外とした．

GPU 利用率が 80%以上と GPU を有効に活用しているジョブは全体の約 31%と高い割合を占める一方で，GPU 利用率が 0%(0.05%未満)のジョブは全体の約 19%，10%以下のジョブでは全体の約 36%と同様に割合が高い．ジョブキューの種別で分類すると，GPU 単位のジョブキューでは GPU 類が存在するため，ノード単位での最小値である 1 ノードに相当する 8GPU の利用者が多いと推測される．

GPU 機における実行ジョブについて，GPU の活用状況を推定するため，Wisteria/BDEC-01 Aquarius における 2022 年度実行ジョブの GPU 利用率分布を図 7 に示す．ジョブ実行以外の時間を含む可能性が高いため，インタラクティブジョブは集計対象外とし，デバッグ目的であることやプログラム利用率 80%以上のジョブが約 45%と大半を占めているが，ノード単位のジョブキューでは GPU 利用率 10%以下のジョブが約 63%と傾向が異なる．

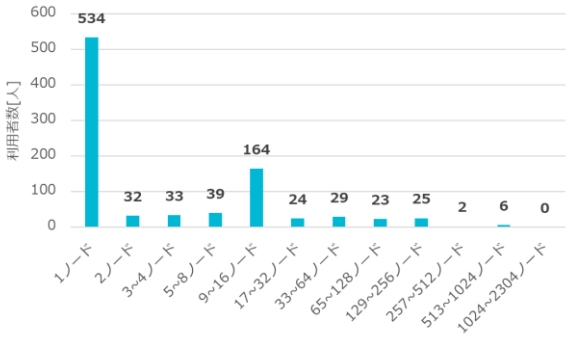

図 5 Wisteria/BDEC-01 Odyssey
2022 年度利用ノード数分布

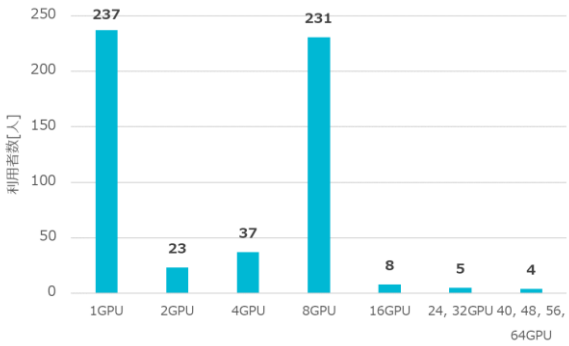

図 6 Wisteria/BDEC-01 Aquarius
2022 年度利用 GPU 数分布

GPU 単位のジョブキューでは明示的に利用 GPU 数を指定するため，利用者が想定している利用 GPU 数と割り当て GPU 数が一致するが，ノード単位のジョブキューでは GPU 数の指定は行わず 1 ノードにつき 8GPU が自動的に割り当てられるため，利用者の誤解により利用予定の GPU 数を上回る GPU が割り当てられ，GPU 利用率が低下した可能性が考えられる．また，GPU 単位のジョブキューでは，1GPU あたりの割り当て CPU コア数とメモリ容量が 1 ノード利用時の 8 分の 1 であるため，CPU コア数等の必要性によりノード単位のジョブキューを選択した可能性も考えられる．その他の要因として，Wisteria/BDEC-01 では CPU 機である Odyssey のプロセッサが FUJITSU Processor A64FX であるため，インテル社製プロセッサ環境を目的に Aquarius において GPU を利用しないジョブを実行している可能性が考えられる．

3. GPU へのプログラム移行支援

GPU 搭載ノードを中心としたシステムとなる想定である OFP-II に，約 3000 名の利用者が円滑に移行するためにはプログラムの GPU 移行が課題となる．2.3 節で述べたように，現在運用している GPU 搭載機である Wisteria/BDEC-01 Aquarius の利用率は徐々に上昇しているが，GPU 利用率が 0% のジョブの割合も高いことから GPU を利用しているジョブに限定した実際の利用率は低くなり，CPU 機と比較すると利用は低調である．また，CPU 機と GPU 機では利用分野が大きく異なるため，GPU への移行は途上であると推定される．

上記を踏まえて，本センターおよび JCAHPC では，GPU へのプログラム移行を支援する取組を 2022 年 11 月より実施してきた．利用者自身で移行できるように支援することを基本としているが，利用者の多いコミュニティコード等を対象に GPU 移行の特別サポートも実施している．

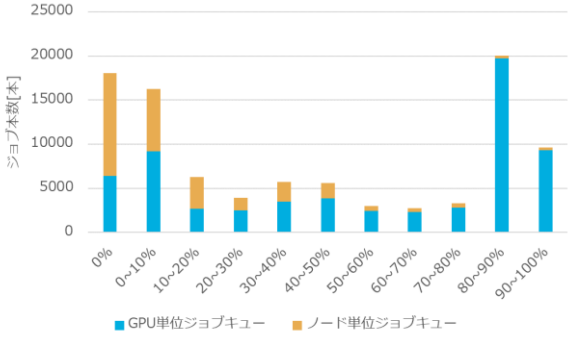

図 7 GPU 利用率分布

3.1 GPU 移行ポータルサイト

既存の CPU プログラムを GPU へ移行する方法は複数存在し，CPU プログラムの実装方法や GPU 移行後の要求性能，投じることのできる開発コスト等により選択すべき方法が異なる．利用者自身で GPU への移行方法を選択し，移行を実践できるように GPU 移行に関連する情報を集約した GPU 移行ポータルサイト[9]（図 8）を公開した．GPU 移行を支援する各種イベント情報も掲載している．

3.2 GPU 移行相談会

GPU 移行に関する様々な疑問を JCAHPC の研究者やエヌビディア合同会社の技術者と直接相談することができる場として，GPU 移行相談会を 2022 年 12 月から月に 1 回オンラインで開催している．本センターのスーパーコンピュータシステム利用者に限定せず誰でも参加可能であり，参加費は無料である．幅広い相談内容を受け付けており，様子を知るためだけの参加も歓迎している．

参加者数を図 9 に示す．2023 年 6 月は GPU 移行相談会を開催していない．平均参加者数は約 5 名であり，以下内容のような相談が行われた．

- 複数 GPU 利用時に演算性能が向上しない
- GPU 移行の進め方
- GPU で高速化するための具体例
- GPU 移行時に生じたエラーの原因究明

図 8 GPU 移行ポータルサイト

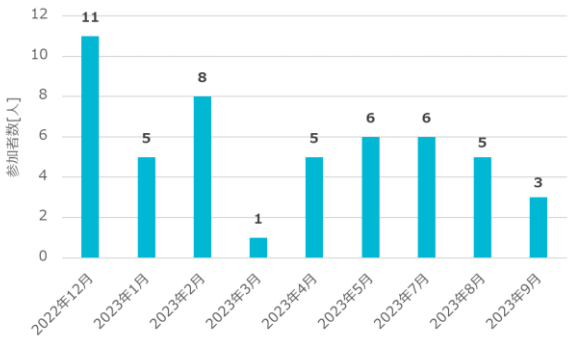

図 9 GPU 移行相談会参加者数

3.3 GPU 移行を支援する講習会

本センターでは，全国のスーパーコンピュータ利用者, および利用を検討している新規ユーザ（企業の技術者・研究者を含む）を対象とした，スーパーコンピュータを用いた実習付きの「お試しアカウント付き並列プログラミング講習会」（以下，講習会と記す）を定期的に実施している[10]. 現在は各年度に約 20 回開催しており，1 回あたり平均 10 名以上が参加している．これまでにも GPU プログラミングに関連する講習会は複数種類開催してきたが，GPU 移行を支援するために，2022 年 12 月以降は以下の講習会を新設するとともに既存の GPU プログラミング関連講習会の開催頻度を増やした.

- MPI+OpenMP で並列化された Fortran プログラムの GPU への移行手法[11]
- UTokyo N-Ways to GPU Programming Bootcamp [12]
- OpenMP で並列化された C++プログラムの GPU 移植手法[13]

3.4 GPU ミニキャンプ

既存の CPU シミュレーションコードを GPU 化する方や, 既存の単体 GPU コードを複数 GPU コードにする方などを対象に，「GPU ミニキャンプ」[14]を 2019 年度より年度に 1, 2 回開催しており 2023 年 9 月までに合計 8 回開催した．GPU ミニキャンプでは参加者がコードやデータセットを持ち込み，各自のペースで GPU 化や GPU 利用効率向上などを実践し，JCAHPC の研究者や企業の技術者で構成されるメンターと随時相談することができる．通常の講習会とは異なりチームで参加することが可能であり，ジョブの実行時間や 1 ジョブの利用可能ノード数等の制限も通常の講習会より大幅に緩和している.

2022 年度からは年度あたりの開催回数を増やし, 参加者に付与されるアカウントの利用期間を延長することで, より GPU 移行に取り組みやすくした. また，第 1 回以降は新型コロナウイルス感染症対策のためにオンラインで開催してきたが，対面での実施希望に応えて 2023 年 7 月の「第 8 回 GPU ミニキャンプ」[14]は初めてハイブリッドで開催した.

3.5 GPU 化特別サポート

多数の利用者を有するコミュニティコード等を対象に，JCAHPC の研究者やエヌビディア合同会社の技術者による個別の調査や最適化をサポートする「GPU 化特別サポート」を実施している. コード量が多い場合には外部委託も実施できるよう予算も準備した．2022 年 11 月頃より徐々にサポートを開始し, 現在は合計 17 グループを対象に実施しており，その内 15 グループが Fortran のコードであり，2 グループが C, C++のコードである．利用分野は，工学系が 3 グループ，生物物理学系が 3 グループ，物理系が 3 グループ，地球科学系が 8 グループである．GPU 移行に際してアルゴリズムやデータ構造の変更が必要になる場合もあり，対象グループで将来も継続して管理ができるよう，GPU 移行相談会や GPU 移行を支援する講習会，GPU ミニキャンプへの積極的な参加も勧めている.

4. GPU 移行支援による効果

利用者による GPU の利用状況を，GPU へのプログラム移行支援前後で比較を行い，GPU 移行支援による効果を考察する．GPU 移行支援の取組は 2022 年 11 月より実施しているが，実施から時間が経過してから効果が出始めると想定されるため, 2022 年度を GPU 移行支援前，2023 年度（2023 年 8 月まで）を GPU 移行支援後と便宜的に区切り集計した.

4.1 利用者全体の利用状況変化

GPU 移行相談会や講習会に参加していない利用者についても GPU 移行ポータルサイト等を参照して各自で GPU 移行を実施していることは考えられるため，全利用者を対象とした比較を行う．利用者ごとに実行ジョブ本数が最も多いシステムを当該利用者の主な利用システムと推定し，2022 年度と 2023 年度における主な利用システムの利用者数分布を表 2 に示す. 2022 年度もしくは 2023 年度（2023 年 8 月まで）にジョブを実行した利用者 2,241 名の内，一方の年度でのみジョブを実行していた 1,681 名は集計対象外とした．いずれのシステムにおいても約 80%の利用者は GPU 移行支援前後で主な利用システムに変化はないが，主な利用システムが Oakbridge-CX, Wisteria/BDEC-01 Odyssey から Wisteria/BDEC-01 Aquarius に変化した利用者は合計 40 名となり，GPU 環境への移行が進んでいることを示唆している.

表 2 GPU 移行支援前後における主な利用システムの利用者数分布（括弧内は GPU 移行支援者の値）

2022 年度の主な利用システム	2023 年度の主な利用システム		
	Oakbridge-CX	Wisteria/BDEC-01 Odyssey	Wisteria/BDEC-01 Aquarius
Oakbridge-CX	154 (4)	31 (3)	**17 (3)**
Wisteria/BDEC-01 Odyssey	2 (0)	188 (22)	**23 (0)**
Wisteria/BDEC-01 Aquarius	4 (0)	14 (2)	127 (13)

表 3 主な利用 GPU 数増減の利用者数分布
（括弧内は GPU 移行支援者の値）

	利用者数[人]	割合[%]
増加	28 (4)	16.3 (21.1)
減少	18 (1)	10.5 (5.2)
変動なし	126 (14)	73.2 (73.7)
合計	173 (19)	-

表 4 平均 GPU 利用率増減の利用者数分布
（括弧内は GPU 移行支援者の値）

	利用者数[人]	割合[%]
増加	33 (4)	19.2 (21.1)
減少	38 (5)	22.1 (26.3)
変動なし	102 (10)	58.7 (52.6)
合計	173 (19)	-

既に Wisteria/BDEC-01 Aquarius でジョブを実行している利用者について，2022 年度と 2023 年度の主な利用 GPU 数増減の利用者数分布を表 3 に示す．2022 年度もしくは 2023 年度にジョブを実行した利用者 756 名の内，一方の年度でのみジョブを実行していた 583 名は集計対象外とした．主な利用 GPU 数が増加した利用者は 28 名であり，減少した利用者数 18 名を上回り，GPU の利用規模の拡大を示唆している．同様の利用者を対象に 1 ジョブあたりの平均 GPU 利用率増減の利用者数分布を表 4 に示す．増減幅が 10%未満の場合は変動なしとして集計した. 平均 GPU 利用率が 10%以上増加した利用者は 33 名であり，10%以上減少した利用者数 38 名を下回り，平均 GPU 利用率向上の傾向は確認できなかった.

4.2 GPU 移行支援者の利用状況変化

GPU 移行相談会や講習会等, GPU 移行支援を受けている利用者に限定して 4.1 節と同様に利用状況の変化を分析した結果が表 2，表 3，表 4 の括弧内の値である．GPU 移行支援を受けている利用者であるため, 利用者全体の結果と比較して GPU 環境への移行等がより進んでいることが示唆されると期待されたが，GPU 移行支援前後における主な利用システムの利用者数分布では移行が進んでいないことが示唆された．GPU の利用状況については大きな傾向の違いは確認できない．比較を行うために，2022 年度と 2023 年度ともにジョブを実行している利用者のみを対象としているため，2023 年度にまだジョブを実行していない利用者の動向により結果は変動すると考えられる.

5. おわりに

本稿では，GPU 搭載ノードを中心としたスーパーコンピュータシステムとする予定である OFP-II の 2025 年 1 月運用開始に向けて，本センターのスーパーコンピュータシステム利用者約 3000 名が円滑に次期システムに移行するために本センターおよび JCAHPC が 2022 年 11 月より実施してきた GPU へのプログラム移行を支援する取組を報告した．GPU 移行支援の取組前後で GPU 環境への移行が進んでいることが示唆されたが，約 3000 名の利用者の GPU 移行はまだ途上である.

より多くの利用者が円滑に次期システムへ移行できるよう, 今後も現在の GPU 移行支援の取組を続けるとともに，利用者の利用状況も定期的に分析を行い充実した支援を提供できることを目指して尽力する.

謝辞　本論文の作成に際して，様々なご助言をいただいた情報基盤課，情報基盤センターの関係者の皆様に深く感謝いたします．また，JCAHPC を共同で運営している筑波大学計算科学研究センターの関係者の皆様，エヌビディア合同会社の皆様には GPU へのプログラム移行支援の取組にご尽力賜り，深く感謝いたします．この場をお借りして厚くお礼を申し上げます.

参考文献

[1] Reedbush スーパーコンピュータシステム, https://www.cc.u-tokyo.ac.jp/supercomputer/reedbush/service/

[2] Oakbridge-CX スーパーコンピュータシステム, https://www.cc.u-tokyo.ac.jp/supercomputer/obcx/service/

[3] Wisteria/BDEC-01 スーパーコンピュータシステム，https://www.cc.u-tokyo.ac.jp/supercomputer/wisteria/service/

[4] 最先端共同 HPC 基盤施設 (JCAHPC), https://jcahpc.jp/

[5] Oakforest-PACS スーパーコンピュータシステム，https://www.cc.u-tokyo.ac.jp/supercomputer/ofp/service/

[6] GROMACS，https://www.gromacs.org/

[7] AlphaFold Protein Structure Database, https://alphafold.ebi.ac.uk/

[8] Y. Ajima, T. Kawashima, T. Okamoto, N. Shida, K. Hirai, T. Shimizu, S. Hiramoto, Y. Ikeda, T. Yoshikawa, K. Uchida, and T. Inoue, "The Tofu Interconnect D", 2018 IEEE International Conference on Cluster Computing, pp. 646-654, 2018

[9] GPU 移行ポータルサイト, https://jcahpc.github.io/gpu_porting/

[10] 講習会, https://www.cc.u-tokyo.ac.jp/events/lectures/

[11] 星野哲也，第 196 回お試しアカウント付き並列プログラミング講習会「MPI+OpenMP で並列化された Fortran プログラムの GPU への移行手法」, 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ニュース，Vol.25，No.2，pp.67-68，2023.

[12] 下川辺隆史，第 207 回お試しアカウント付き並列プログラミング講習会「UTokyo N-Ways to GPU Programming Bootcamp」, 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ニュース，Vol.25，No.5，pp.35-36，2023.

[13] 三木洋平，第 208 回お試しアカウント付き並列プログラミング講習会「OpenMP で並列化された C++プログラムの GPU 移植手法」，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ニュース，Vol.25，No.5，pp.37-38，2023.

[14] 下川辺隆史，第 210 回お試しアカウント付き並列プログラミング講習会「第 8 回 GPU ミニキャンプ」，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ニュース, Vol.25, No.5, pp.39-41, 2023.