

# TSUBAME4.0: HPC-AI 時代に向けた 東京科学大学のもっとみんなのスパコン

安良岡 由規, 遠藤 敏夫, 野村 哲弘, 渡邊 寿雄, 鶴見 慶

東京科学大学 情報基盤センター

## TSUBAME4.0: More of Everyone's Supercomputer toward HPC-AI Era in Science Tokyo

Yoshinori Yasuraoka, Toshio Endo, Akihiro Nomura, Toshio Watanabe, Kei Tsurumi

Center for Information Infrastructure, Institute of Science Tokyo

### 概要

本稿では、東京科学大学情報基盤センターが運用しているスーパーコンピュータ TSUBAME4.0 の概要を報告する。このシステムは 2024 年 4 月より東京工業大学学術国際情報センター（当時）として運用を開始し、大学統合・改組を経て現在の体制となっている。TSUBAME4.0 は NVIDIA H100 GPU 960 基などを備えており、システムの総演算性能は倍精度で 66.8PFlops, AI において注目されている半精度で 952PFlops にのぼる。高い演算性能に加え、TSUBAME4.0 はこれまでの TSUBAME シリーズの特徴を受け継ぎつつ、学生を含む初学者にもさらに使いやすいスパコンとして、ウェブブラウザからの利用機能などを充実させている。

## 1 はじめに

2024 年 4 月に東京工業大学学術国際情報センター（当時）は、スーパーコンピュータ TSUBAME4.0[1] を稼働開始させた。TSUBAME4.0 は長年にわたり東工大内外の研究・開発・教育を支えてきた TSUBAME スパコンシリーズの最新システムである。TSUBAME4.0 の合計演算性能は、倍精度演算（行列向け）では 66.8 ペタフロップスであり、前世代 TSUBAME3.0[2] の約 5.5 倍である。さらに人工知能（AI）分野で重要とされる半精度演算では 952 ペタフロップスであり、TSUBAME3.0 の約 20 倍にのぼる。TSUBAME4.0 は 2024 年 5 月に発表されたスパコンランキング Top500[3] において、国内 2 位、世界 31 位にランクされた。2024 年 10 月に、東京工業大学と東京医科歯科大学の合併および改組に伴い、東京科学大学情報基盤センターとして TSUBAME4.0 の運用を行っている。

TSUBAME スパコンシリーズは 2006 年導入の TSUBAME1.0 以来、「みんなのスパコン」をキャッチフレーズとしたスパコン利用初心者などへの使いやすさと、GPU アクセラレータを始めとする各時代の先進技術の導入の両立を実現してきた。後者の代

表例は 2008 年 11 月の Top500 ランキングにおいて TSUBAME1.2 が GPU 搭載スパコンとして世界で初めてランクされたことである [4]。また 2024 年 3 月に稼働終了した前世代の TSUBAME3.0 のユーザ数は、学内外の研究者や学生、企業利用者を含み約 4700 人であった。

本稿で解説する TSUBAME4.0 は TSUBAME シリーズの特質を受け継ぎつつ、前述のような性能向上および、Web 利用の充実化などの使いやすさの向上を可能としている。システム構築は日本ヒューレット・パッカード合同会社（以下、HPE）が担当し、主要コンポーネントである 240 台の計算ノード HPE Cray XD665 は、NVIDIA H100 GPU を総計で 960 基、総 CPU コア数 46,080 を備える。また本学大岡山キャンパスに設置された TSUBAME3.0 までと異なり、TSUBAME4.0 はすずかけ台キャンパスに設置されている。図 1 に外観を示す。

## 2 TSUBAME4.0 のシステム

### 2.1 システム概要

TSUBAME4.0 の性能の概要を表 1 に示す。倍精度演算性能の理論値は 66.8PFlops であり、前世代 TSUBAME3.0 の約 5.5 倍となる。なおここでは



図 1 TSUBAME4.0 の外観。デザインコンテストの結果選出されたラックデザインが施されている

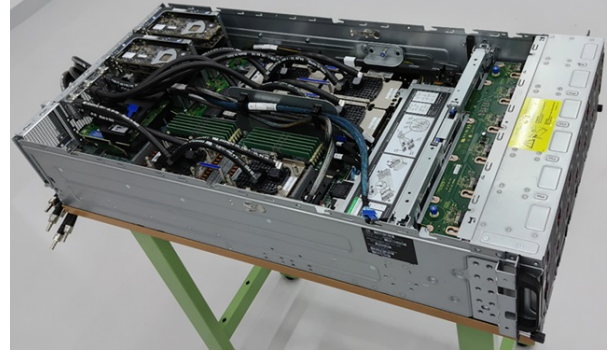


図 2 TSUBAME4.0 の計算ノード HPE Cray XD665 4U サーバ

表 1 TSUBAME4.0 のシステム性能。前世代 TSUBAME3.0 との比較で示す。

	TSUBAME3.0	TSUBAME4.0
総演算性能 倍精度 (FP64)	12PFlops	66.8PFlops (行列演算) 34.7PFlops (汎用演算)
半精度 (FP16)	47PFlops	952PFlops
計算ノード ノード数	HPE/SGI ICE XA 540	HPE Cray XD665 240
共有ストレージ HDD 部容量	DDN 社 SFA 16PByte	HPE ClustreStor E1000 44PByte
Flash 部容量	-	327TByte
インターコネクト	OmniPath (100Gbps)	InfiniBand NDR200 (200Gbps)
トポロジー	Full Bisection Fat Tree	Full Bisection Fat Tree

NVIDIA H100 GPU の Tensor Core による行列演算性能を数えた場合であり、行列以外の汎用演算の性能としては 34.7PFlops となる。深層学習で重要視されている半精度においては演算性能は 952PFlops であり、TSUBAME3.0 の約 20 倍となる。

計算ノードの台数は 240 台であり、台数および後述の GPU/CPU ソケット数においては TSUBAME3.0 時の約 44% となっている。共有ストレージについては、ハードディスク部の実効容量 44PByte であり、TSUBAME3 の約 2.7 倍となっている。また Flash ベースの共有ストレージが新設されており、実効容量は 327TByte である。計算ノード群とストレージを結ぶインターコネクトは InfiniBand NDR200 (計算ノードあたり 4port) であり、フルバイセクションのファットツリートポロジーである。

## 2.2 計算ノード

TSUBAME4.0 の主要部分である計算ノードは、240 台の 4U サーバである HPE Cray XD665 である。そ

の外観を図 2 に、構成を表 2 に示す。各ノードには CPU として 2 基の AMD EPYC 9654 (Genoa 世代, 96 コア) を搭載する。ノードあたりのコア数は TSUBAME3.0 の 28 (14 コア × 2) に比べ 192 と、大きく増えている。

GPU として 4 基の NVIDIA H100 を搭載する。搭載されているのは通常の H100 とはメモリ仕様が異なり、HBM2e メモリを 94GB 搭載、その速度は 2.4TB/s となる。H100 SXM の通常モデルでは HBM3 を 80GB、速度 3.35TB/s であるため、TSUBAME4.0 ではメモリ容量がより大きいアクセス速度が低いというトレードオフがある。AI モデルの大規模化に伴い GPU デバイスメモリの容量への要求が高まっており、その動向に応えた形である。演算性能については GPU あたり倍精度で 67TFlops、半精度で 990TFlops となる。特に半精度性能が TSUBAME3.0 に比較した伸びが大きく、GPU あたりで約 47 倍、システム全体で約 22 倍となる。

## 2.3 システム規模と冷却設備

TSUBAME4.0 システムは東京科学大学すずかけ台キャンパス G4-A 棟に設置されている。この建物は TSUBAME4.0 設置のために改修され、約 200m<sup>2</sup> のマシン室等が整備された。TSUBAME4.0 は 30 基のラックに搭載され、うち 23 基が計算ノード用、5 基がストレージ用、2 基が管理サーバやコアスイッチ等に用いられる。各ラックは Motivair 社の冷却機能付きラックであり、それぞれ下部に CDU (Coolant Distribution Unit) と背面の冷却ドアをもつ。

主な電力系統は 415V 3 相 4 線であり、電力容量は 2MW となる。ほか補助的に 200V 3 相 4 線等を持つ。TSUBAME4.0 のシステムの消費電力は設計上 1840kW だが、通常運用時は約 500~800kW で推移している。

表 2 TSUBAME4.0 計算ノードの構成. 前世代 TSUBAME3.0 との比較で示す.

	TSUBAME3.0 node	TSUBAME4.0 node
CPU	2 × Intel Xeon E5-2680v4 (Broadwell)	2 × AMD EPYC 9654 (Genoa)
CPU あたりコア数	14	96
クロック周波数 (base)	2.4GHz	2.4GHz
メインメモリ	DDR4-2400 256GiB	DDR5-4800 768GiB (24 × 32GiB DIMMs)
GPU (以下, GPU あたり)	4 × NVIDIA Tesla P100 SXM	4 × NVIDIA H100 SXM5 94GB HBM2e
SM 数	56	132
倍精度演算性能	5.3TFlops	67TFlops (行列演算)
半精度演算性能	21.2TFlops	990TFlops
デバイスメモリ	HBM2 16GB	HBM2e 94GB
デバイスメモリ速度	732GB/s	2.4TB/s
Network Interface	4 × OmniPath 100Gbps	4 × InfiniBand NDR200 200Gbps
ローカル SSD	2TB NVMe	1.92TB NVMe

TSUBAME4.0 計算ノードの冷却方法は直接液冷と空冷のハイブリッド方式となる [5]. 最大の発熱源は, 1 つあたり TDP 700W の GPU と 400W の CPU であり, それらは直接液体冷却される. 図 2 にそのための冷媒パイプが見られる. それ以外のメモリモジュール等は空冷であり, そのために各計算ノードは前面にファンを持つ.

計算ノード類からの廃熱 (冷媒, 空気) は図 3 のように処理される. 温度上昇した冷媒はラック CDU に接続され熱交換される. 背面の空気はラックの冷却ドアによって冷却されラック外に排出される. 冷却ドアにも冷媒が通っており, やはり CDU と接続されている. 各 CDU は屋外のチラーと接続され, 熱は最終的に屋外に排出される. チラーから CDU に戻る液温は 2024 年 5 月の時点で 18 °C 程度である.

TSUBAME4.0 の計算ノードラックにおいては, ラックあたりの電力は設計上 55kW, 通常運用時 20~30kW であり, 典型的なデータセンターの 5 倍程度となる. このような高い熱密度のシステムをハイブリッド方式により冷却している.

なお以上の冷却手法は概念的には TSUBAME3.0 の手法 [2] を踏襲したものである. ただし TSUBAME3.0 では屋外機器として冷却塔を用いたフリークーリング (屋内に戻る液温 27 °C 程度) であったの

に対し, TSUBAME4.0 ではコンプレッサーを用いたチラーに依存しており, 冷却の省エネ面では一部後退した形となる.

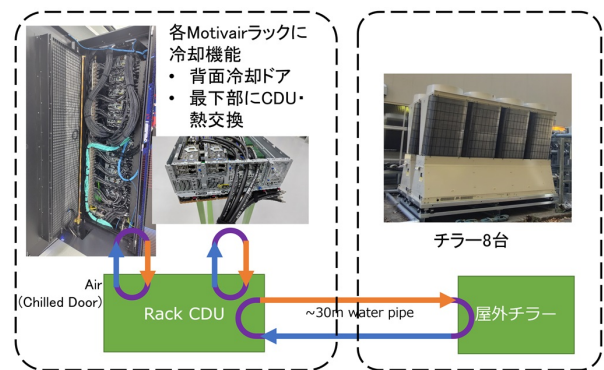


図 3 TSUBAME4.0 の冷却設備の概念図

## 2.4 ベンチマーク

ベンチマークによる TSUBAME4.0 の全体性能測定を 2024 年 4 月に行い, 2024 年 6 月の各種ランキングに登録した. Linpack ベンチマーク測定においては実行プログラムとして HPL-NVIDIA 24.3.0 が用いられ, 216 ノードを用いた実行結果 25.46PFlops(図 4) が Top500 に 31 位としてランクされた. この速度は 216 ノードの理論性能に比べると 42% 程度であり, 他の H100 搭載スパコンが 60% 程度を記録しているのに対し低い. その原因の一つとして導入時テスト・メ

メンテナンスを通した後でもまだノード・GPU 群に速度の低い個体が残ったことが考えられる。

上記の実行中の電力推移を計測し (図 5), その結果を用いて省エネランキング Green500 にも登録した (つまり別途 power-optimized run は行っていない)[6]. このために TSUBAME4.0 に備えられた 5 基の電力計 Schneider PM8240 を用いた. うち 3 基の電力計 (図中の Compute 1-3) は計算ラックノード (CDU 等を含む) を計測する. 4 基目 (Water Facility) は屋外のチラー・ポンプ電力を計測する. 5 基目は部屋エアコン電力を測定するが, その電力は無視できるほどである.

Green500 の規則 [7] において, もっとも厳密な Level 3 に基づく計測を行った. 電力計の精度は 0.2% であり, 積算電力量を 1 秒おきに記録するものであるため, Level 3 計測に用いることができる. 算入する電力としては Compute 1-3 を用い, これらの計測電力の Linpack core phase 期間における平均値は 732kW であった. 電力性能比を  $25.46\text{PFlops} / 732\text{kW} = 34.779\text{GFlops/W}$  として算出し, これは 2024 年 6 月の Green500 において 31 位であった. なおこの期間の Water Facility 平均電力は 89kW, Air Facility 平均電力は 5kW であった.

### 3 利用環境

各計算ノードの OS は RedHat Enterprise Linux 9 である. x86 CPU および Linux OS の組み合わせという, これまでの TSUBAME シリーズと互換性を持ち, また高性能計算システムとしては現時点で最も普及した利用環境と言える. GPU の利用のために CUDA, OpenACC, OpenMP 5.0 対応コンパイラなどが準備されている.

TSUBAME4.0 のハードウェアの特徴の一つは, 計算ノード 1 台の規模が 192 コア +4GPU と, 大きい (ファットノード) ことである. さまざまな利用形態に応じてより適切に計算資源利用を可能とするため, 以下のようなノード分割を, Altair Grid Engine ジョブスケジューラとの連携により行っている.

- 計算ノード群の多くはバッチノードであり, ユーザからの要求に合わせて動的なノード分割を行っている. それぞれの分割は最小 8CPU コアから最大で 1 ノード全体 (192 コア +4GPU) である. ユーザは各分割を排他的に利用可能であり, 他の利用とコア単位で分離される. ただしネットワー

クやメモリのバンド幅の共有のために性能への影響はありうる. 詳細は [8] を参照されたい.

- 計算ノードの一部 (現時点では 2 台) をインタラクティブノードとして運用し, 各ノードを固定的に 8 分割している. 1 分割は 24 コア +0.5GPU (NVIDIA MIG 機能で分割) である. バッチノードと異なり, この分割は複数の利用者によって共有されうる. 性能への影響は上記より大きい一方で, 原則待ち時間がなく利用可能であるという利点がある. ただし分割あたりの最大の同時利用数は決まっており, この制限のために利用開始に失敗することはありうる.

ユーザは上記に述べたような計算ノード (もしくはその分割) を, 以下を含む複数の方法によって利用可能である.

- まずログインノード (GPU がないなど計算ノードと異なる) に SSH ログインし, そこからジョブスケジューラに対して `qsub` コマンドにてジョブ投入する
- ログインノードから `qssh` コマンドでシェルを用いて対話利用する
- Web ブラウザから Open OnDemand システム<sup>\*1</sup>を用いて利用する. 現時点では利用環境として `xfce` (X Window 環境), JupyterLab が用意されている.

スーパーコンピュータの利用に対する敷居を下げるために, TSUBAME3.0 において SSH および公開鍵を介さない Web ブラウザからの利用を, 独自ポータルの開発により実現してきた [9]. 近年, この目的のために Open OnDemand の利用が普及しており [10, 11], TSUBAME4.0 ではそちらを用いている.

TSUBAME は様々な利用分野のユーザによって利用されており, 計算科学ユーザの需要に応える上で Gaussian, AMBER, Material Studio などのアプリケーションソフトウェアが用意されている<sup>\*2</sup>. 近年 AI・機械学習分野の利用も急増しており, その特質の一つは, 各ユーザによって用いるフレームワーク・パッケージおよびそれぞれのバージョンが異なりうるため, 各自準備する場合が多いことである. TSUBAME においては (他の多くのスパコン同様) ユーザが管

<sup>\*1</sup> <https://openondemand.org/>

<sup>\*2</sup> ライセンス上の制限や課金が発生する場合がある. 詳細は <https://www.t4.gsic.titech.ac.jp/applications>

```
=====
T/V          N    NB    P    Q          Time          Gflops (    per GPU)
-----
WCO          3124224 1024   32   27          798.58          2.546e+07 ( 2.946e+04)
-----

||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=  0.000141499787 ..... PASSED
||Ax-b||_oo . . . . . = 0.0000006008384542
||A||_oo . . . . . = 782472.9106226920848712
||x||_oo . . . . . = 15.6451672756239315
||b||_oo . . . . . = 0.4999999086777096
=====
```

図4 Linpack ベンチマークの実行出力 (抜粋)

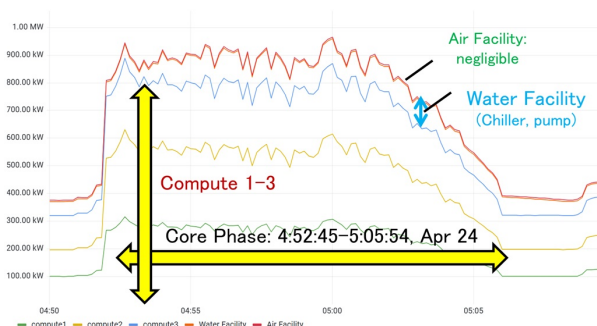


図5 Linpack ベンチマーク時の電力推移. Compute 1-3 の部分 (平均 732kW) を電力性能比算出に用いた.

理者権限を持つことを許さないため, python venv, miniconda などのツールを用いて各自の仮想環境を構築することになる. また, TSUBAME3.0 に引き続き Apptainer(Singularity) コンテナが提供されており, Docker コンテナイメージなどを変換して仮想環境を構築することもできる.

#### 4 TSUBAME4.0 の現状とまとめ

2024 年 4 月に稼働開始した TSUBAME4.0 は, 8 月末時点でユーザアカウント数は 2000 を超えている. また図 6 には TSUBAME3.0, TSUBAME4.0 の月ごとのノード利用率を示す. TSUBAME4.0 では 6~8 月においてノード利用率が 95% 前後と, これまでの同月と比べても高くなっている. これは各ユーザにとっての長い待ち時間を引き起こしており, これまでに述べた動的ノード分割に加え, さらなる方策や研究が必要となっている.

TSUBAME シリーズは東京工業大学のスーパーコンピュータとして知られてきたが, 東京科学大学にお

いても引き続き重要な計算資源インフラとして, 研究・教育・産業分野にて大きく活用される予定である.

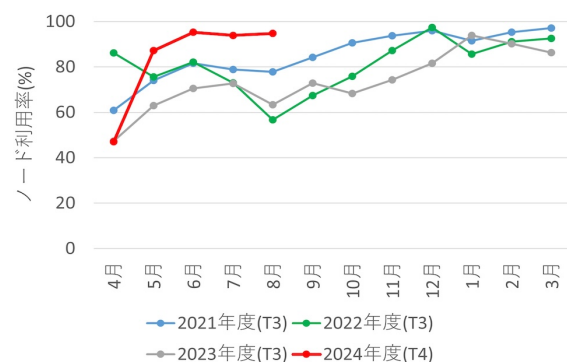


図6 月ごとのノード利用率. 2023 年度までは TSUBAME3.0, 2024 年度は TSUBAME4.0 のものである.

#### 参考文献

- [1] TSUBAME 計算サービス. <https://www.t4.gsic.titech.ac.jp/>.
- [2] 松岡 聡, 遠藤 敏夫, 額田 彰, 三浦 信一, 野村 哲弘, 佐藤 仁, 實本 英之, Drozd Aleksandr. HPC とビッグデータ・AI を融合するグリーン・クラウドスパコン TSUBAME3.0 の概要 . 並列/分散/協調処理に関するサマワークショップ (SWoPP2017), 情報処理学会研究報告, 2017-HPC-160 No.29, 2017.
- [3] The Top500 List. <https://www.top500.org/>.
- [4] Toshio Endo, Akira Nukada, Satoshi Matsuoka and Naoya Maruyama. Linpack Evaluation on a Supercomputer with Heterogeneous Accelerators. In Proceedings of IEEE International Par-

allel and Distributed Processing Symposium (IPDPS 2010), 8 pages, 2010.

- [5] Akihiro Nomura. TSUBAME 4.0 Supercomputer: Introduction of System and Failures during the Installation. HPCInfra Workshop, 2024.
- [6] Toshio Endo, Akihiro Nomura. Experiences with making a power measurement and submission for TSUBAME4.0, Level 3. EE HPC WG Workshop, 2024.
- [7] Energy Efficient High Performance Computing Power Measurement Methodology (version 2.0 RC 1.0), <https://www.top500.org/static/media/uploads/methodology-2.0rc1.pdf>.
- [8] 野村 哲弘, 遠藤 敏夫. スパコン TSUBAME シリーズにおけるリソース分割戦略. 並列/分散/協調処理に関するサマーワークショップ (SWoPP2024), 情報処理学会研究報告, 2024-HPC-195, No.7, 2024.
- [9] 野村 哲弘, 遠藤 敏夫, 三浦 信一, 朝倉 博紀, 越野 俊充, 草間 俊博. TSUBAME3 のインタラクティブ利用の利便性向上にむけた取り組み. 並列/分散/協調処理に関するサマーワークショップ (SWoPP2020), 情報処理学会研究報告, 2020-HPC-175, No. 23, 2020.
- [10] 中尾昌広, 三浦信一, 山本啓二. スーパーコンピュータ「富岳」における HPC クラスタ用 Web ポータル Open OnDemand の導入. 情報処理学会研究報告, 2022-HPC-186, No. 5, 2022.
- [11] 中田秀基, 高宮安仁, 高野了成, 滝澤真一郎, 谷村勇輔. AI 橋渡しクラウドにおける WebUI と量子コンピューティングサービスの導入. 情報処理学会研究報告, 2024-HPC-193, No. 3, 2024.