参考資料2

AI-for-Science ロードマップ

暫定版
(2024年3月8日時点)

第 0.1.0 版

2024 年 3 月 8 日

第 1 章

概要

1.1 AI for Science: 科学研究における AI の活用

近年、スーパーコンピュータの性能の目覚ましい向上と共に、人工知能（AI）は科学技術の進展とその応用範囲において、世界中で革新的な変革をもたらした。この進化は、日本においても、科学研究の多岐にわたる分野での発展の可能性を示唆している。我々は、AI 技術を科学研究に応用し、これらの領域における新たな課題の解決と未来のイノベーションの推進に向けた取り組みを AI for Science ロードマップとしてまとめ、ロードマップに基づく将来必要とされるスーパーコンピュータの AI 計算性能の推定、AI ガバナンスに関する課題を明らかにすることで、日本の将来の AI の基礎及び応用研究の方向性を示す。

画像・音声・言語データや科学データ等の蓄積と利活用、これらビッグデータの活用による AI の科学研究への応用によるデータ駆動型アプローチの優位性の発見、スーパーコンピュータの性能向上により、AI 技術を科学分野で応用する “AI for Science” 研究がこれまでより遥かに速いスピードで広がり、さらに加速している。これらの進展は、科学研究における新たなパラダイムを生み、より複雑な科学的問題へのアプローチ方法を根本から変える可能性を秘めている。このため日本国内において現在の AI for Science 研究の動向や技術的課題を明らかにし、今後の AI for Science 研究を強化し、国際連携を促進するための AI for Science ロードマップ の作成が重要となる。

このため、HPCIC 計算科学ロードマップで対象とされている 11 分野（素粒子・原子核、ナノサイエンス・デバイス、エネルギー・資源、生命科学、創薬・医療、設計・製造、社会科学、脳科学、地震・津波、気象・気候、宇宙・天文）において AI の応用研究を行っている研究者を集め、各分野における科学的背景、AI 活用の必要性、AI の活用例や国内外の動向、AI for Science に向けたロードマップの作成を行った。その結果、基盤モデルを活用した生成 AI の利用が不可欠であることを再認識した。

生成 AI を活用し AI for Science を推進する上で、スーパーコンピュータの役割は極めて重要となる。大規模言語モデルに代表される基盤モデルを高速に学習するためには、膨大な学習データセットを高速に処理する必要があるため、スーパーコンピュータのような大規模な計算機の利用が事実上必須となっている。今後も基盤モデルのサイズは大規模化の一途を辿り、そのモデルを学習するのに必要な計算性能に対する需要はこれまで以上に高くなる。このため、AI for Science ロードマップに基づき、将来必要とされるスーパーコンピュータの AI 計算性能の推定し、その需要に基づき次世代スーパーコンピュータを開発することが重要となる。このため、我々はいくつかの既存研究の文献をもとに、将来必要とされる要求実効 AI 性能の推定を行った。その結果、2030 年頃には 43.4 〜 86.8 EFLOPS 以上の実効 AI 性能が必要になる結果となった。

このような AI の急速な発展は、科学の様々な分野で革新をもたらしている。AI 技術の進歩は人類にとっ

て大きな恩恵をもたらす一方で、ハルシネーション、プライバシー侵害、セキュリティリスクなど、新たな課題も引き起こす。これらの課題に対処するため、効果的な AI ガバナンスの枠組みの確立が急務となっている。AI ガバナンスとは、AI 技術の開発と利用を適正に管理・監督するための方針やルール、メカニズムの総称である。これには、AI の倫理的な使用を保証するガイドラインの策定、AI によって生じるリスクの評価と管理、そして AI 技術の利用に関する透明性と説明責任の確保が含まれる。しかし、AI 技術の急速な発展に伴い、これらのガバナンスメカニズムを策定し、実施することは容易ではなく、幾つかの課題がある。AI for Science ロードマップでは、国内外における健全な社会実装のための AI ガバナンスについて紹介する。

AI for Science ロードマップは、日本における科学技術の進歩と国際競争力の向上を牽引する重要な指針として機能することが期待される。このロードマップが提案する先進的な基盤モデルの活用は、特に科学研究に革命をもたらす可能性が高く、2030 年頃に 43.4 〜 86.8 EFLOPS 以上の実効 AI 性能を持つスーパーコンピュータが必須であるとの推定は、日本における次世代計算基盤に対する投資方針とその開発戦略の策定に役立つと期待する。AI 技術の進歩がもたらす利益とリスクのバランスを適切に取ることで、科学だけでなく社会全体に対して肯定的な影響を与える持続可能な発展が可能となる。AI ガバナンスに対する取り組みは、このバランスを達成するための鍵となり、適正な方針とルールの下で、AI 技術のポテンシャルを最大限に活かしつつ、課題に対処することが重要である。科学研究の未来と日本のテクノロジーのリーダーシップを確固たるものとするためには、ロードマップで提示された方向性を具体化し、国際的に AI for Science 研究を牽引することが求められる。

1.3 次世代計算基盤開発で要求される AI 性能

1.3.1 2030 年頃の次世代計算基盤で要求される実効 AI 性能

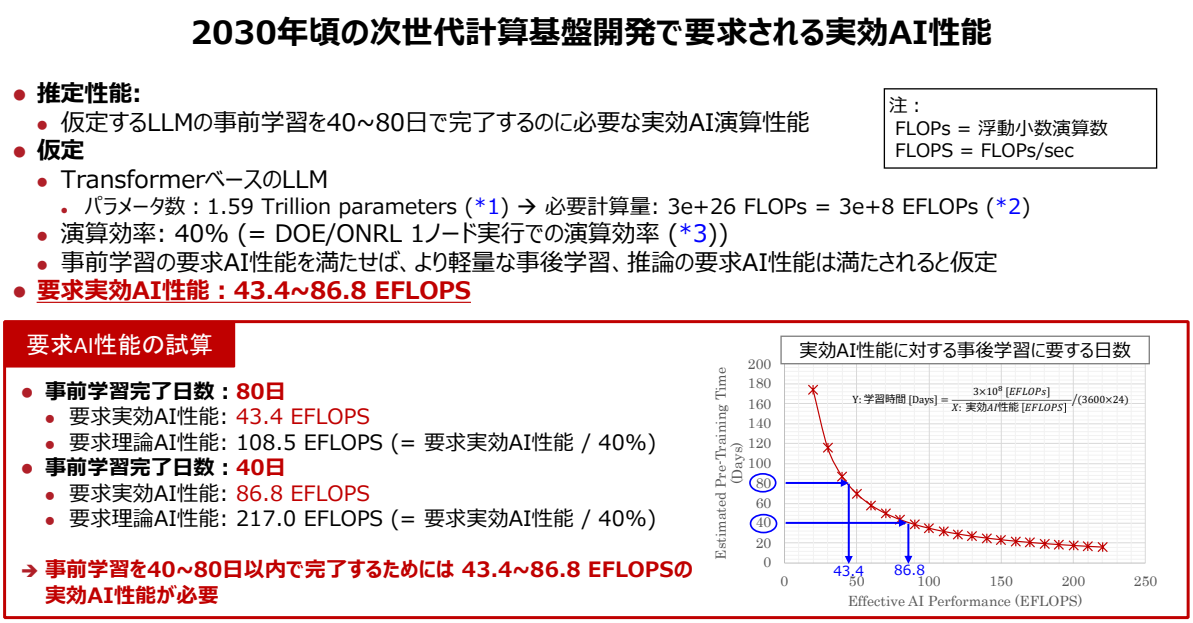

図 1.1 2030 年頃の次世代計算基盤開発で要求される実効 AI 性能

次世代計算基盤開発で要求される AI 性能について推定する（図 1.1）。推定する性能は、大規模言語モデル (LLM) の事前学習を 40 80 日で完了するのに必要な実効 AI 演算性能と定義する。我々が対象とする LLM は、1.59 兆 (1.59×10^{12}) のパラメータ数を有する Transformer ベースのモデルとする (詳細は、1.3.2 章を参照)。また、この対象とする LLM の事前学習に必要な計算量を $3 \times 10^{26} FLOPs (= 3 \times 10^{8}) EFLOPs$ と推定する (詳細は、1.3.3 章を参照)。理論 AI 性能に対する実効 AI 演算の演算効率を 40% とした (詳細は、

1.3.4 を参照)。また、事前学習の要求 AI 性能を満たせば、より軽量な事後学習、推論の要求 AI 性能は満たされると仮定する。

以上の仮定から、図 1.1 のグラフに示される通りの実効 AI 性能に対する事後学習に要する日数を見積もることが可能である。この見積もりから事前学習を 80 日で完了するためには要求理論 AI 性能が 108.5 EFLOPS、要求実効 AI 性能が 43.4 EFLOPS、事前学習を 40 日で完了するためには要求理論 AI 性能が 217.0 EFLOPS、要求実効 AI 性能が 86.8 EFLOPS であると推定することができる。以上の推定結果から、次世代計算基盤開発で要求される実効 AI 性能の目標値として 43.4 ~ 86.8 EFLOPS 以上とするのが妥当であると考えられる。

1.3.2 2030 年頃に利活用される LLM のパラメータ数の予測

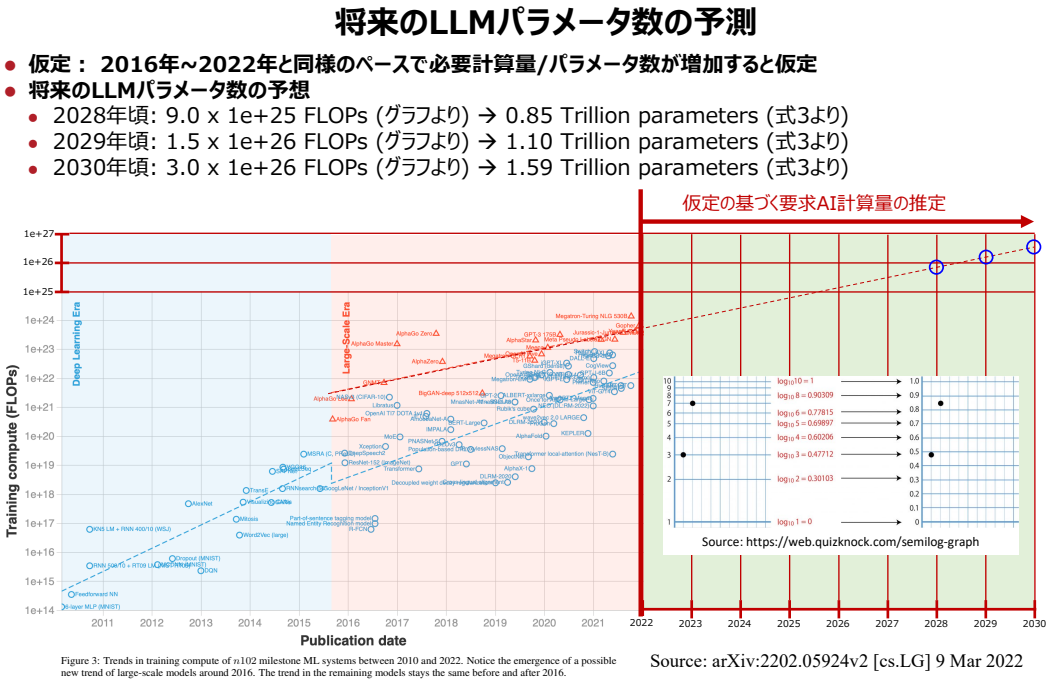

Figure 3: Trends in training compute of n102 milestone ML systems between 2010 and 2022. Notice the emergence of a possible new trend of large-scale models around 2016. The trend in the remaining models stays the same before and after 2016.

Source: arXiv:2202.05924v2 [cs.LG] 9 Mar 2022

図 1.2 将来の LLM パラメータ数の推定

2030 年頃に次世代計算基盤開発で要求される AI 性能の推定には、2030 年頃に研究開発される LLM のパラメータ数を推定する必要がある。ここでは、Sevilla らの研究 [*1]において作成された図 1.2 に示されるグラフに基づいて推定する。このグラフは、論文発表日 (Publication data) に対して、その論文で提案された LLM を学習するのに必要な計算量 (FLOPs) の関係を表すグラフである。しかし、この調査では 2022 年までに発表されたモデルに留まる。このため我々は 2016 年～2022 年の間に発表された LLM と同様のペースで必要な計算量が増加すると仮定し、2028 年～2023 年に必要な計算量を推定した。その結果、2028 年頃には 9.0×10^{25} $FLOPs$、2029 年頃には 1.5×10^{26} $FLOPs$、2030 年頃には 3.0×10^{26} $FLOPs$ の計算量が必要とされる LLM が登場すると推定した。1.3.3 章に示される関係式 (式 3) より、この必要な計算量はそれぞれ

[*1] Jaime Sevilla, Lennart Heim, Anson Ho, Tamay Besiroglu, Marius Hobbhahn, Pablo Villalobos, "Compute Trends Across Three Eras of Machine Learning", arXiv:2202.05924v2 [cs.LG] 9 Mar 2022

0.85 兆、1.10 兆、1.59 兆のパラメータ数をもつ LLM に必要かつ十分な計算量に相当する。以上の推定から、2030 年頃に利活用される LLM のパラメータ数は 1.59 兆個であると推定する。

1.3.3 LLM のパラメータ数に対する必要な計算量 (FLOPs) の推定

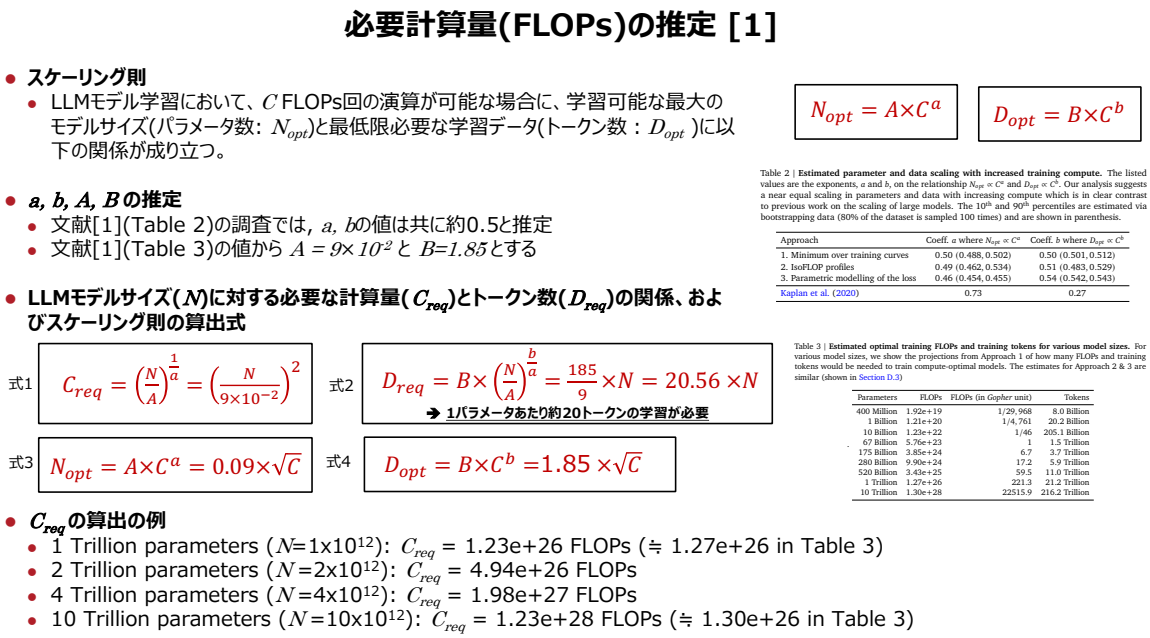

Table 2 | **Estimated parameter and data scaling with increased training compute.** The listed values are the exponents, a and b, on the relationship $N_{opt} \propto C^a$ and $D_{opt} \propto C^b$. Our analysis suggests a near equal scaling in parameters and data with increasing compute which is in clear contrast to previous work on the scaling of large models. The 10th and 90th percentiles are estimated via bootstrapping data (80% of the dataset is sampled 100 times) and are shown in parenthesis.

Approach	Coeff. a where $N_{opt} \propto C^a$	Coeff. b where $D_{opt} \propto C^b$
1. Minimum over training curves	0.50 (0.488, 0.502)	0.50 (0.501, 0.512)
2. IsoFLOP profiles	0.49 (0.462, 0.534)	0.51 (0.483, 0.529)
3. Parametric modelling of the loss	0.46 (0.454, 0.455)	0.54 (0.542, 0.543)
Kaplan et al. (2020)	0.73	0.27

Table 3 | **Estimated optimal training FLOPs and training tokens for various model sizes.** For various model sizes, we show the projections from Approach 1 of how many FLOPs and training tokens would be needed to train compute-optimal models. The estimates for Approach 2 & 3 are similar (shown in Section D.3)

Parameters	FLOPs	FLOPs (in *Gopher* unit)	Tokens
400 Million	1.92e+19	1/29,968	8.0 Billion
1 Billion	1.21e+20	1/4,761	20.2 Billion
10 Billion	1.23e+22	1/46	205.1 Billion
67 Billion	5.76e+23	1	1.5 Trillion
175 Billion	3.85e+24	6.7	3.7 Trillion
280 Billion	9.90e+24	17.2	5.9 Trillion
520 Billion	3.43e+25	59.5	11.0 Trillion
1 Trillion	1.27e+26	221.3	21.2 Trillion
10 Trillion	1.30e+28	22515.9	216.2 Trillion

図 1.3　LLM のパラメータ数に対する必要な計算量 (FLOPs) の推定

必要な計算量の見積もりには、Hoffmann らの研究 [*2] で発表されたスケーリング則に基づいて推定する (図 1.3)。スケーリング則とは、LLM の学習において C $FLOPs$ 回の演算が可能な場合、学習可能な最大の LLM のモデルサイズ (パラメータ数: N_{opt}) と最低限必要な学習データ (トークン数：D_{opt}) に対して、$N_{opt} \propto C^a$、$D_{opt} \propto C^b$ の関係が成り立つことを示している。Hoffmann らの研究では実際に a 及び b が 0.5 になると実験的に示している。一方、比例定数に関しては文献中の Table 3 の値から推定することで、$N_{opt} = 9.0 \times 10^{-2} \times C^{0.5}$、$D_{opt} = 1.85 \times C^{0.5}$ の関係式を導くことができる。この式を変形させることで図 1.3 の式 1〜4 を導出することができる。式 1〜2 は、モデルサイズが N の LLM の学習に必要な計算量 (C_{req}) と学習に必要なトークン数 (D_{req}) の関係式となる。特に式 2 から 1 パラメータあたり約 20 トークンの学習が必要であることがわかる。

暫定版
(2024年3月8日時点)

DOE/ORNL Frontierでの演算効率 [1]

- **Training performance in 1 GCD (= 191.5 TFLOPS) in the sentence**
  - 1.76B: 77 TFLOPS per GCD (**Efficiency: 40.2%**)
- **Training performance in 385 GCDs (= 73.7 PFLOPS) in Figure 6**
  - FORGE-S (1.44B): 63 TFLOPS per GCD (Efficiency: 32.6%)
  - FORGE-M (13B): 63 TFLOPS per GCD (Efficiency: 32.6%)
- **Training performance in 1,024 GCDs (= 196.1 PFLOPS) in Figure 9**
  - FORGE-S (1.44B): Projected 60 PFLOPS (Efficiency: 30.6%)
  - FORGE-M (13B): 47 PFLOPS (Efficiency: 24.0%)
  - FORGE-L (25.6B): Projected 42 PFLOPS (Efficiency: 21.4%)
  - 175B: 32 PFLOPS (Efficiency: 16.3%)

➔ **将来AI向けハードウェア、ソフトウェア、アルゴリズムの高度化により演算効率が大規模モデル&大規模実行において 40% となると仮定**

Note that on a single device (1 GCD), the performance will be higher, and about 77 TFLOPS is obtained for a 1.76B parameter model of similar architecture as FORGE-S (the difference is that the hidden size is increased to 2304). This indicates that it is possible to achieve around $\alpha = 40\%$ of the MI250X peak performance for this model.

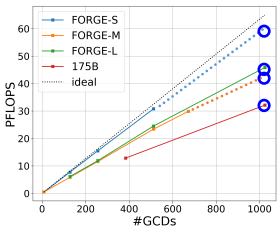

Figure 9: Scalability of training FORGE models and 175B parameters up to 2048 GCDs on Frontier. (left) Aggregated performance in PFLOPS. (right) TFLOPS-per-GCD and time for training FORGE-S4 with a batch size of 16.8M. The per GCD batch size is adjusted accordingly.

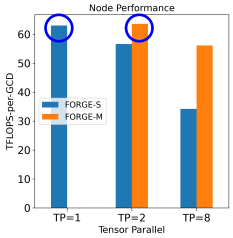

Figure 6: Assessing computation performance (TFLOPS-per-GCD) by training FORGE-S and FORGE-M at different tensor parallel (TP) levels on a single node, and training GPT-175B on 385 GCDs with various parallelism techniques ([TP, PP, DP]).

[1] Junqi Yin and Sajal Dash and Feiyi Wang and Mallikarjun Shankar, "FORGE: Pre-Training Open Foundation Models for Science", SC23, Nov., 2023, DOI¥:10.1145/3581784.3613215

図 1.4　DOE/RNL Frontier での LLM 事前学習の演算効率

1.3.4　LLM 学習の演算効率の推定

これには Yin らの研究 [*3] で得られた実効効率 40% を採用した（図 1.4）。この研究によると、現時点では 40% の演算効率を達成することは挑戦的な課題となっているが、将来 AI 向けハードウェア、ソフトウェア、アルゴリズムの高度化により演算効率が大規模モデル&大規模実行において 40% となると仮定する。

[*2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, Laurent Sifre, "Training Compute-Optimal Large Language Models", arXiv:2203.15556v1 [cs.CL] 29 Mar 2022

[*3] Junqi Yin and Sajal Dash and Feiyi Wang and Mallikarjun Shankar, "FORGE Pre-Training Open Foundation Models for Science", SC23, Nov., 2023, DOI 10.1145/3581784.3613215