ABCI
AI Bridging Cloud Infrastructure

AIST
Create the Future, Collaborate Together

生成AIの研究開発を支援する計算基盤としてのABCIの課題と展望

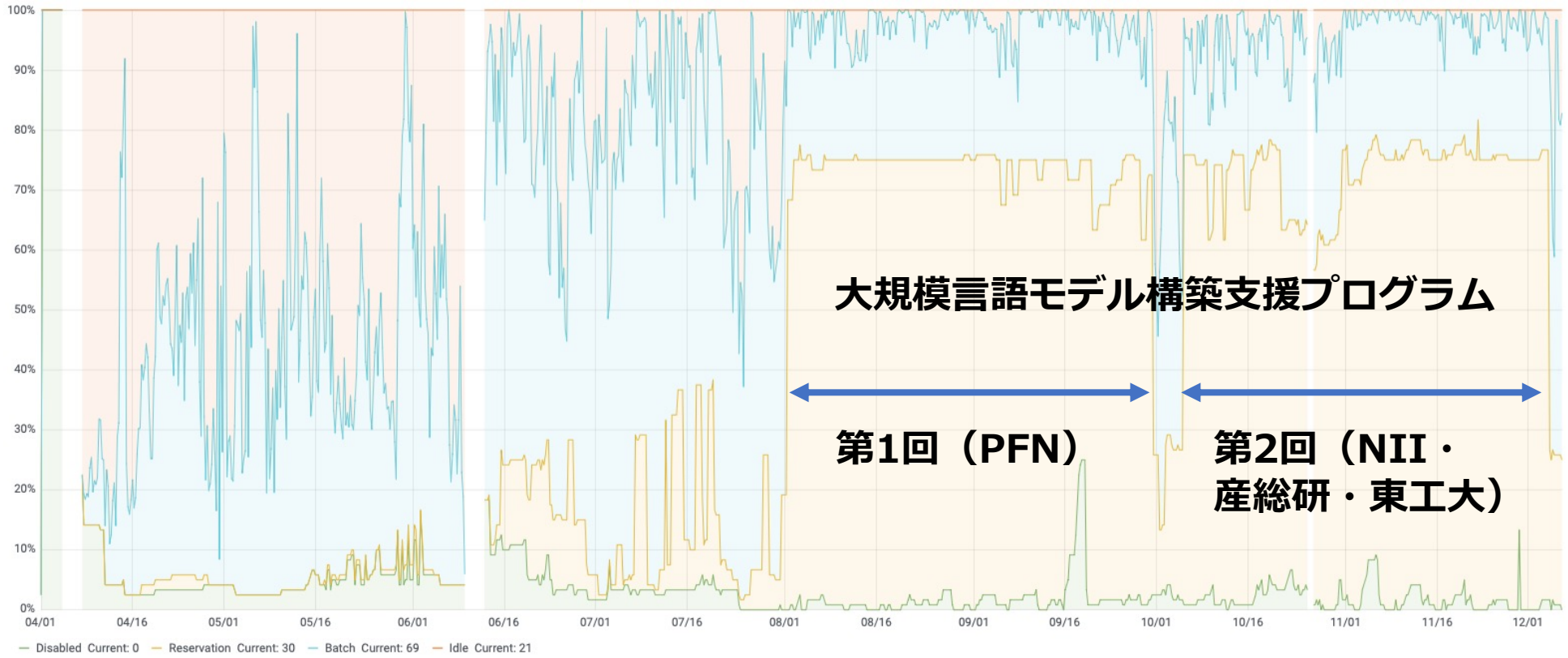
国立研究開発法人 産業技術総合研究所
高野 了成

2023-12-08 第23回PCクラスタシンポジウム「HPC基盤技術と生成AI」

ABCIの現状（1/2）

- 生成AI需要の高まりを踏まえ、大規模言語モデル構築支援プログラムを開始した結果、計算ノード(A)の利用率が100%近傍である状態が恒常化。
- ジョブが実行されないという問い合わせが増える。

計算ノード(A)の利用率



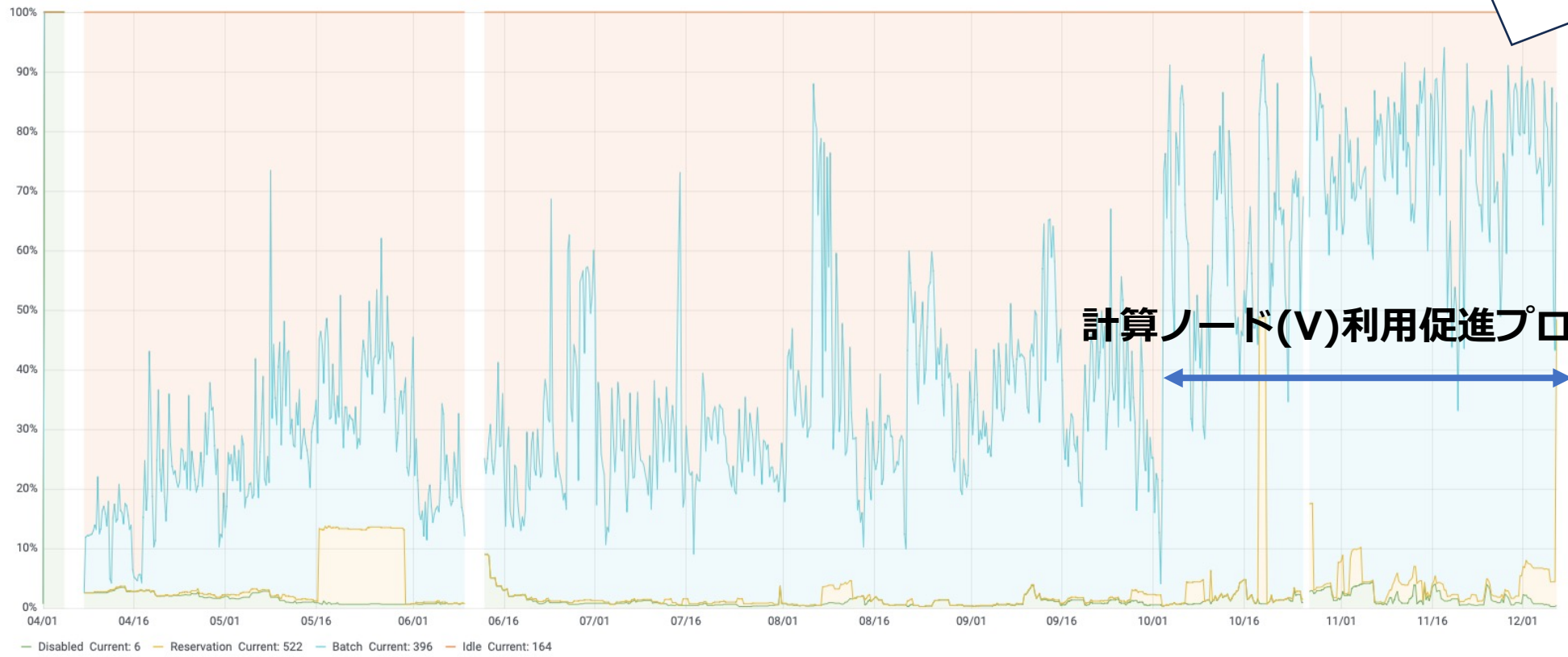
2023/4/1　　8/1　　12/1

ABCIの現状（2/2）

- さらに計算ノード(V)の利用率も90%近くまで上昇。

11/29 CUDA Quantum Workshopの開催中に、インタラクティブジョブが起動しない。。。

計算ノード(V)の利用率

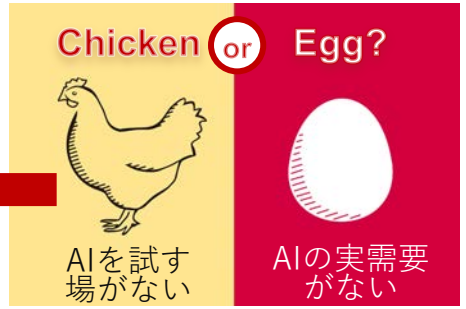


計算ノード(V)利用促進プログラム

2023/4/1　　10/1　　12/1

世界最大級・超省電力・オープンＡＩインフラストラクチャABCI

省電力AIスパコン研究



産業界のAI導入が進まない

ABCI AI Bridging Cloud Infrastructure

- 産総研は2000年代初頭からスパコン研究開発を継続
- 特に近年AIの計算需要に応え、同時に運用・管理コストの圧縮を実現する省電力AIスパコンに取り組む
- 一方で、当時AIについては産業界の関心の高さの割に導入が進んでいないという課題があった

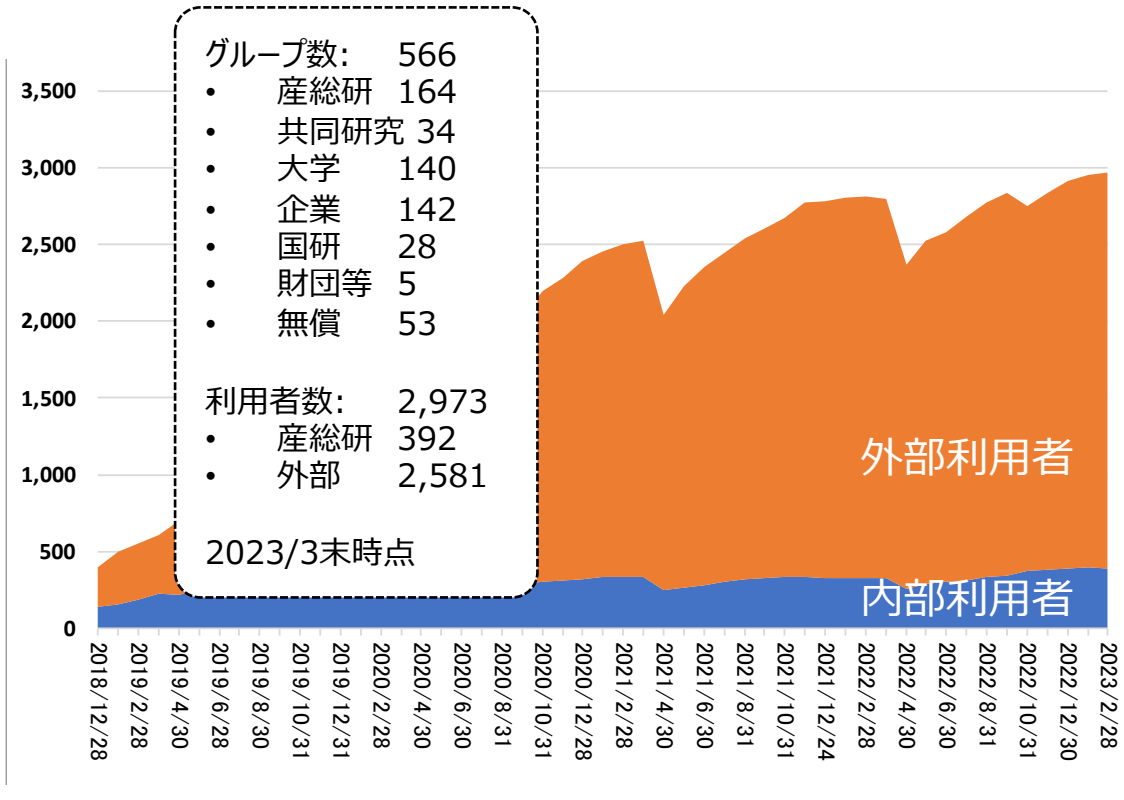
- AI橋渡しクラウド（ABCI）：膨大な計算とデータを要するAIをすみやかに試す「場」を提供し、我が国における産学官によるＡＩ研究開発と社会実装を加速するオープンイノベーションプラットフォーム
- 高い計算能力を活用した人工知能技術の研究開発・実証、社会実装の推進、ＡＩ分野の最重要課題への挑戦が目的

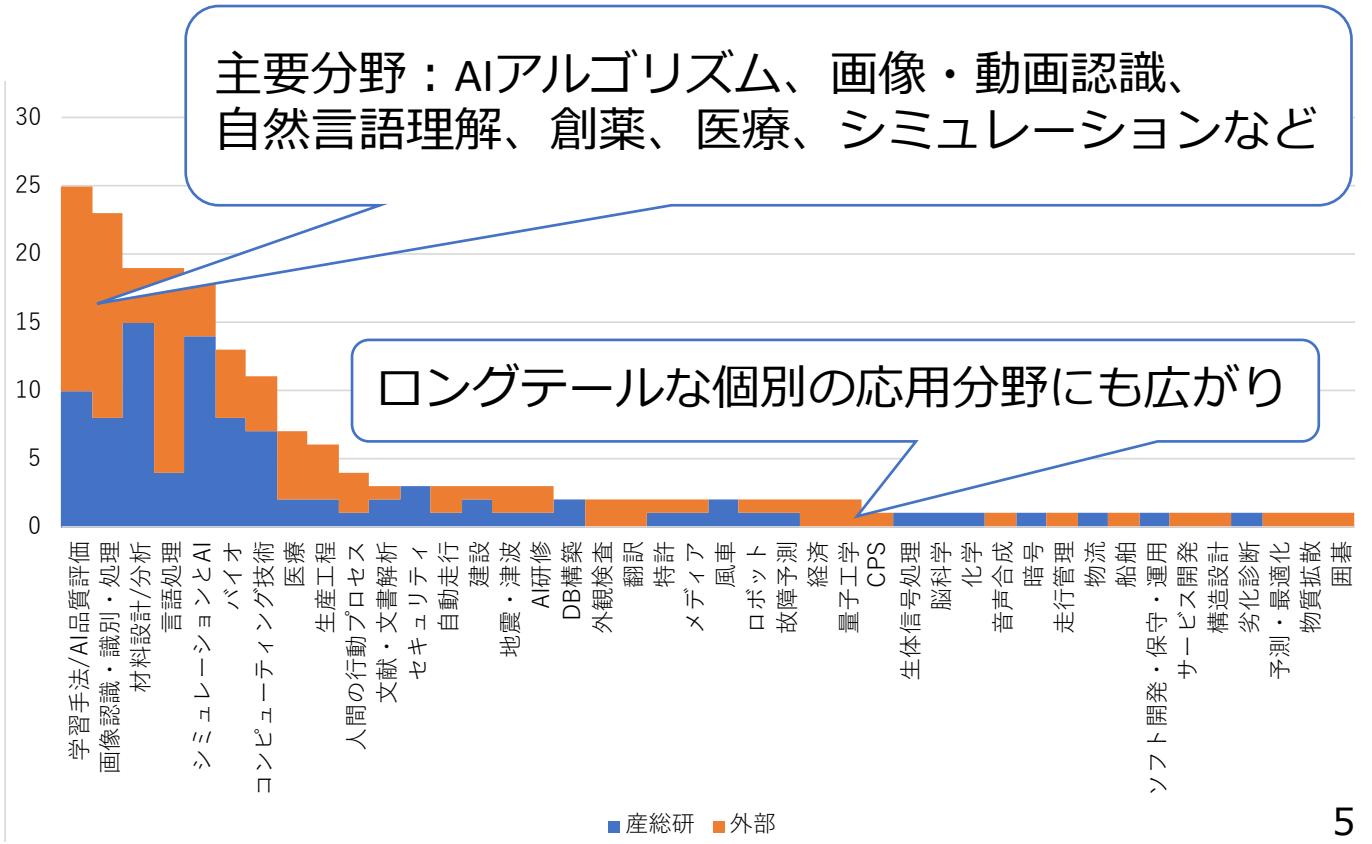
2018年8月1日　運用開始
2021年5月10日 2.0運用開始

- 経産省「人工知能に関するグローバル研究拠点整備事業」（2016年）の一環として整備
- 経産省「人工知能に関する橋渡しインフラ拡張」により2021年にアップグレードを実現

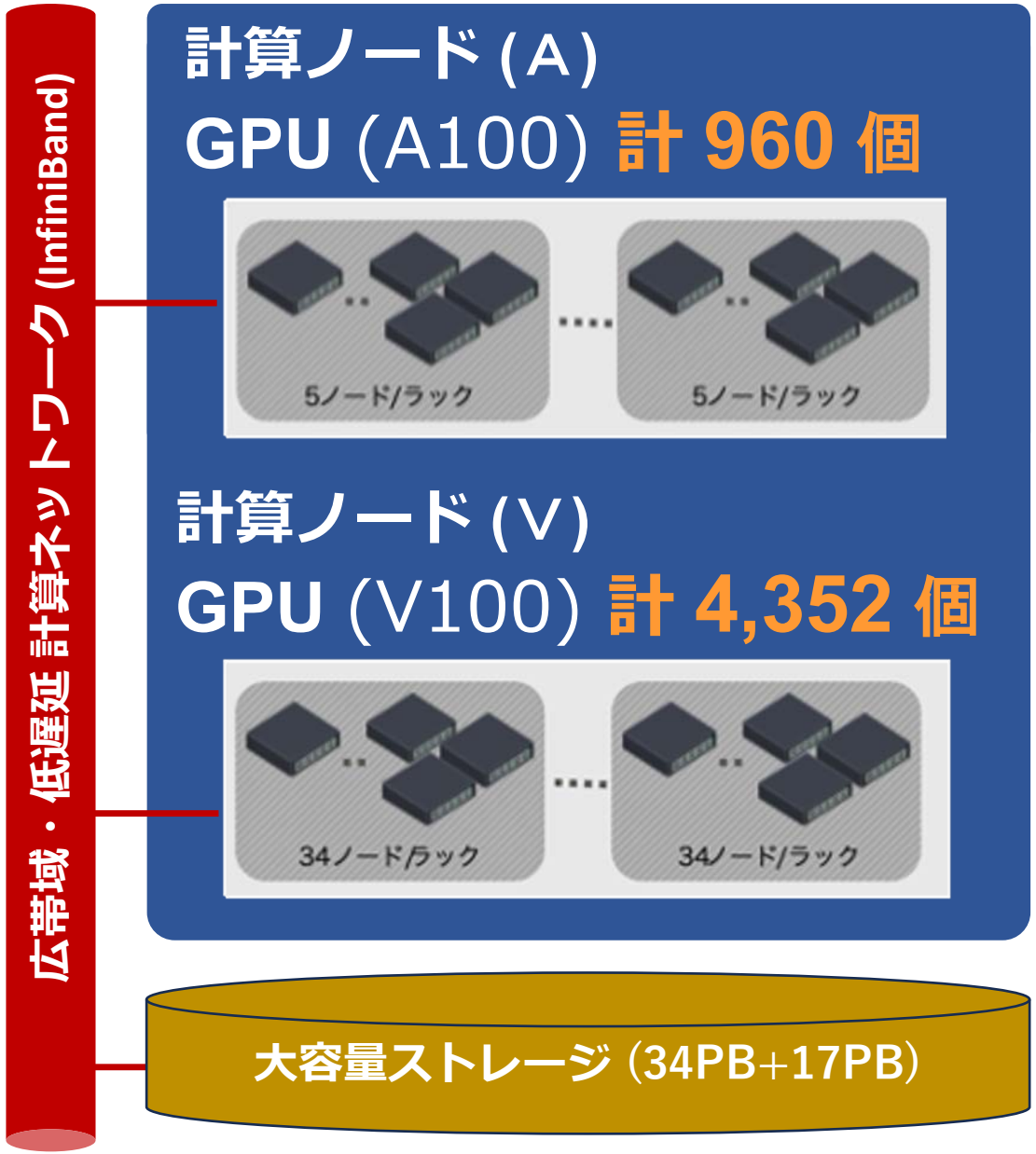
ABCIの利用者数・主な利用分野

- 2018年8月の運用開始以来、利用者数は右肩上がりで増加
- 2023年3月現在の利用者数は約3000人 (うち外部利用が約87%)
- AIスタートアップから総合電機メーカーまで利用が拡大し、世界に伍する研究が可能になるとともに、我が国のAI研究全体を支える重要基盤に成長





ＡＩに特化したハードウェア・ソフトウェア構成と先進的取組み

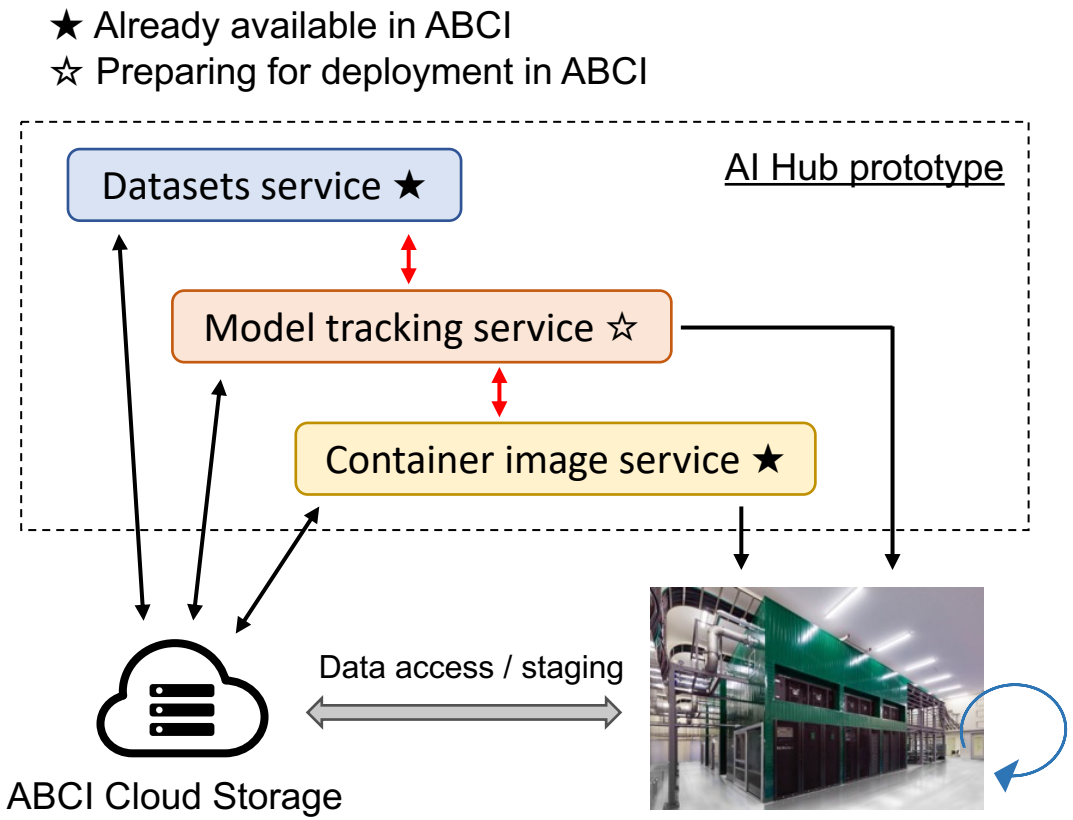


- 多数の計算ノード・GPUを使った**大量データの高速な並列処理に特化したハードウェア構成**
  - 高速なAI計算に不可欠な高性能GPU
  - 広帯域・低遅延な計算ノード間通信
  - 大容量・広帯域・高可用性ストレージ
- AI研究開発を効率化し、**新規ユーザの敷居を下げるソフトウェア構成**と先進的な取り組み
  - 最先端のミドルウェアや並列化コンパイラ、最新のGPU向け開発環境や深層学習のツール類を提供
  - ABCIに最適化された「コンテナ」開発により分散並列学習などの最先端技術をユーザが簡単に利用可能に
  - データと学習済みモデル、実行環境をひとまとめにしてAI開発を容易にするABCI上のサービス「AIハブ」の一部機能がすでに提供（ABCIデータセット）

産総研がこれまで培ったスパコン構築・運用のノウハウが、日本を代表する「高性能」で「使いやすい」 AIスパコンの実現を可能にした

ABCIにおけるAIハブの研究開発

- **AI 資源**：膨大な学習用データ、大規模汎用学習済みモデル、学習処理系のコンテナイメージ、それらを活用する Jupyter Notebook レシピ等
- **AI ハブ**：大規模な汎用学習済みモデルの再利用を促進し、適応的に転移学習等に利用可能するデータ・ソフトウェアエコシステムを実現するサービス



Datasets service (https://datasets.abci.ai)

- データセットの公開・共有を支援するカタログサービス
- NEDO AI 2.0プロジェクトの成果11件以上がすでに登録済み

Model tracking service

- 学習済みモデルの共有・再利用を促進するツールセット
  - MLflowをベースにモデル構築の記録・共有だけではなく、派生モデルの開発・転移学習への応用を支援
  - Develop derived models and/or apply to transfer learning
- データセットとコンテナイメージサービスと連携

Container image service

- Singularity Enterpriseを利用した、ABCI利用者向けのコンテナイメージライブラリ、リモートビルドサービス

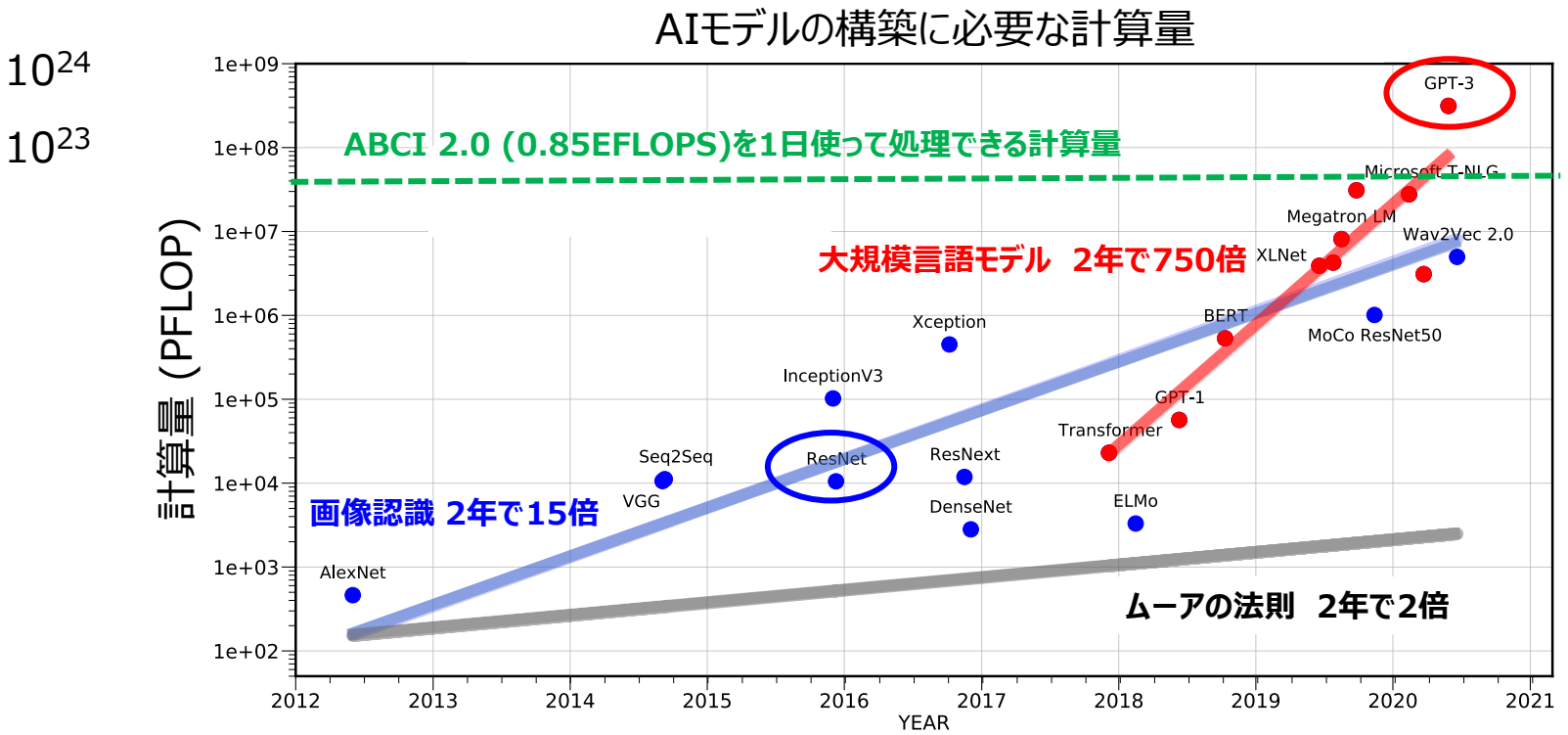
生成AIに振り回された一年

- 2022/11/30 OpenAIがChatGPTを発表
- 2022/02/02 ChatGPT公開後約2ヶ月で1億ユーザを突破
- 2023/02/03 自民党AIの進化と実装に関するプロジェクトチーム（第1回）
- 2023/03/14 ChatGPT PlusでGPT-4が利用可能に
- 2023/05/09 基盤モデルに関する3国研協議（第1回）（理研、産総研、NII）
- 2023/05/24 富岳政策対応枠 利用開始（東工大、東北大、富士通、理研）
- 2023/06/16 経済安保法に基づく認定供給確保計画（クラウドプログラム）にさくらインターネットが認定
- 2023/07/06-14 大規模言語モデル分散学習ハッカソン
- 2023/07/12 大規模言語モデル構築支援プログラム公募開始（第1回）
- 2023/07/18 MetaがLLaMA2のソースコードを公開
- 2023/10/06 シンABCI資料招請公示
- 2023/10/23 PCCC AI/機械学習技術部会「大規模言語モデルハンズオン」
- 2023/11/10 NEDOポスト5G「競争力ある生成AI基盤モデルの開発」公募開始
- 2023/11/13 西村大臣と米国のAI／半導体企業7社の幹部と意見交換会
- 2023/12/04 岸田総理とNVIDIAジェンスン・ファンCEOの面談

（参考）横田先生のHPCI成果報告会資料
https://www.hpci-office.jp/application/files/2217/0063/4061/20231026_Special_lecture.pdf

基盤モデル構築に必要な計算量とABCIの性能のギャップの広がり

- ABCIの運用開始時（2018年）のAI研究の中心は画像認識系。
- 大規模言語モデルの登場以降、モデルの大規模化に伴い、モデル構築に必要な計算性能が急激に上がっている。
  この規模の学習を定常的に処理するには現状の国内計算インフラでは圧倒的に性能不足であり、大幅増強が必要。
  - ABCI2.0を使った理論性能：ResNet（画像識別）の学習に13秒、GPT-3の学習に4日。
  - GPT-3を1日で学習するには3.4EFLOPSの性能が必要。



Source: https://github.com/amirgholami/ai_and_memory_wall

注：すべての見積もりは半精度で計算している。

生成AI開発力強化に向けた計算資源の確保（令和5年度補正予算）

- **生成AIの開発・活用には、大規模な計算資源（スパコン）とデータが必要**。世界的に、十分な計算資源を確保できる希少なプレイヤーのみが競争力あるAIを開発できている状況。将来の国の競争力を左右することになる**AI用計算資源の確保等に対して集中的に支援**。

- **圧倒的に不足するAI用計算資源の国内整備【1,566億円】**

国内最大は産総研の0.8EFLOPS規模。拡充に向け、経済安保基金を活用し、計算整備への補助を決定。

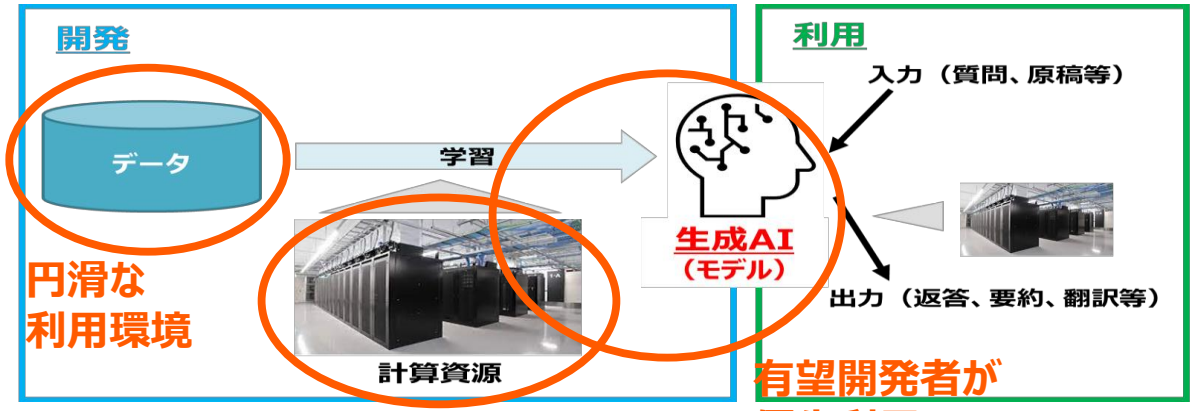
→ 引き続き圧倒的に不足しており、**民間への補助を拡充【1,166億円（経済安保基金）】**するとともに、**産総研の計算資源も4.25EFLOPS※に拡充【400億円（産総研施設設備費補助金の内数）】**。

※生成AI利用時の計算では、最大8.5EFLOPSの計算性能が発揮される。

- **AI開発の加速支援【290億円】（ポスト５G基金）**

AI開発に意欲と能力を持つスタートアップ等は存在するが、計算資源やデータの確保等が課題。

→ **有望なスタートアップ等に対して計算資源の利用を一定期間補助**し、開発を加速。
AIの性能向上・活用促進には、WEB上のデータに加え、企業等が保有するデータの活用が重要。情報漏洩や規制面等での**課題解決に向けたデータ提供者とAI開発者の連携を実証**。



（出典）経済産業省「半導体・デジタル産業戦略の現状と今後」（2023/11/29）

科学研究向け生成AIモデルの開発・共用

～ Artificial General Intelligence for Science of Transformative Research Innovation Platform（TRIP-AGIS） ～

令和6年度要求・要望額　85億円
（新規）
※運営費交付金中の推計額

文部科学省

- **特定科学分野（ドメイン）に強みを有する研究機関と連携体制を構築**し、基盤モデルを活用して、**科学研究データを追加学習（マルチモーダル化）**等することで、**ドメイン指向の科学研究向け生成AIモデル（科学基盤モデル）を開発**
- **開発した科学研究向け生成AIモデルの利用を産学に広く開放**することで、多様な分野における**科学研究の革新（科学研究サイクルの飛躍的加速、科学研究の探索空間の拡大）**をねらう。

AIに関する暫定的な論点整理
（令和5年5月26日、AI戦略会議）

【AI開発力】

- AIの研究成果がAI以外の分野の研究開発の加速に寄与することもほぼ確実である。
- 生成AIによって世界の変革がもたらされようとしている中、可及的速やかに生成AIに関する基盤的な研究力・開発力を国内に醸成することが重要である。
- 世界からトップ人材が集まり切磋琢磨できる研究・人材育成環境の構築や産学官の基盤開発力の強化を進めていくことが期待される。

良質なデータ

- トレーニングやファインチューニング、インストラクションなどに必要なデータを良質な形で整備
- データを蓄積する関係研究機関と連携・共同開発
- 特定科学分野：まずは、
  生命・医科学分野　（例：薬物等による動的変化・遺伝子変異による差異予測向け）
  材料・物性科学分野（例：新奇材料の物性予測向け）など

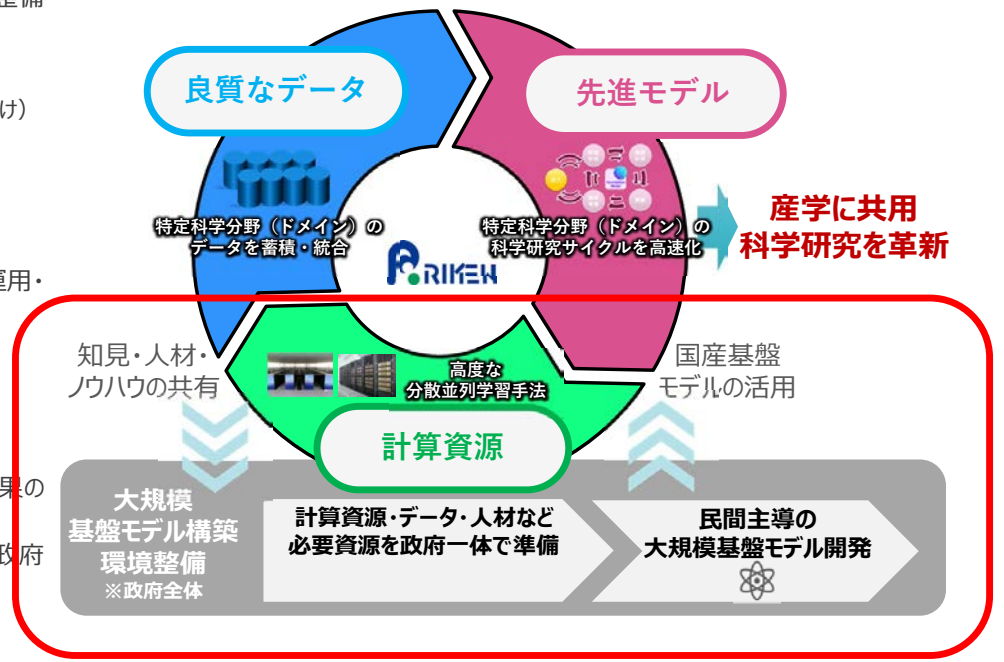
先進モデル

- 基盤モデルを活用し、特定科学分野（ドメイン）指向の科学基盤モデルを開発・運用・共用
- 並行して、マルチモーダルデータを読込・学習・生成するために必要な研究開発

計算資源

- スパコン「富岳」の大規模言語モデル分散並列学習手法の開発（実施中）、成果の活用
- 試行錯誤を繰り返して、小規模モデルから徐々に大規模化し、大規模計算時は政府全体として整備する計算資源を活用
- 並行して、「高速」、「セキュア」、「エコ」を実現する革新的な計算資源の研究開発

“科学研究向け生成AIモデル”による研究革新



※科学基盤モデル：　基盤モデル（言語・画像等）に科学研究データ（論文、リアルタイムな実験・シミュレーションデータ等）を追加学習、推論等させ、特定の科学研究分野（ドメイン）向けに調整した基盤モデルのこと

（担当：研究振興局基礎・基盤研究課）

（出典）文部科学省
「令和6年度概算要求のポイント」

生成AIに係る研究開発ニーズの高まりを踏まえた取り組み（1/2）

現行ABCIによるLLM支援

- **「大規模言語モデル構築支援プログラム」**
  - 最大で計算ノード(A)の半分を最大60日間占有利用する機会を提供する公募型プログラム
  - 第1回：**PFN**（8月～9月）　※成果であるPLaMo 13Bが公開
  - 第2回：NII・産総研 ・**東工大**（10月～11月）、Elyza （12月～1月）
- **「ABCIグランドチャレンジ」**
  - 成果公開を前提に計算資源を無償提供
  - 第1回（5月）　：A, V-Large, V-Week　※東北大がV-Weekを利用したLLM課題に採択
  - 第2回（10月）：V-Large, V-Medium　※第2回からAクラス募集を中止
  - 第3回（12月）：V-Large, V-Medium
  - 第4回（1～2月）　：V-Week
- **計算ノード(A)の事前予約制限の変更**
  - 1予約あたりの最大予約ノード数：18→30 （5月18日）
  - 1予約あたりの最大ノード時間積：6912
- **計算ノード(V)利用促進プログラム（10月2日～1月9日）**
  - 課金係数を約1/2に変更。例）rt_F 1.5ポイント/時間→0.75ポイント/時間

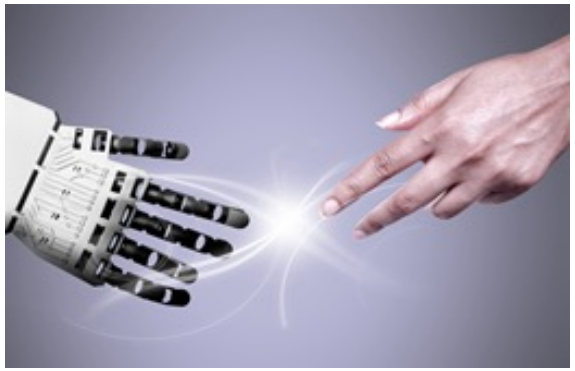
生成AIに係る研究開発ニーズの高まりを踏まえた取り組み（2/2）

現行ABCIによるLLM支援

- **大規模言語モデル分散学習ハッカソン（7月6日～14日）**
  - ハッシュタグ #ABCILLM
  - https://github.com/ohtaman/abci-examples/
- **PCCC AI/機械学習技術部会 第5回ワークショップ「大規模言語モデルハンズオン」（10月23日）**
- **ABCIでLLMを動かそう！セミナー（全4回。12月12日～）**
  - https://aitconsortium.doorkeeper.jp/events/166286
  - 第1回申込み受付中

ABCI LLMハッカソン開催概要（2023年7月6日~14日）

ABCIを活用した大規模言語モデルの学習（ファインチューニング）を実践



- ABCI利用者・ABCI利用を検討中の方またはチーム（1チーム3人程度）を対象に、ABCIを活用した大規模言語モデルの学習（ファインチューニング）を実践する場を提供

- 参加者が訓練用のデータセットを持ち込み、既存の言語モデルを単一GPU、マルチGPU、またマルチノードで訓練（ファインチューニング）。最終日には訓練されたモデルの出力や、ハッカソンで得た学び等を共有

- ABCIスタッフ（産総研職員）に加えて、深層学習や自然言語処理のスペシャリスト及びGPUのエンジニアがチューターとして参加し、モデルの改善や分散処理等をサポート

- 主催
  - 産業技術総合研究所 情報・人間工学領域
- 共催
  - エヌビディア合同会社
  - ユビー株式会社
  - PCクラスタコンソーシアム　AI・機械学習技術部会
  - TensorFlow User Group

- 1ノードの詳細
  - NVIDIA Tesla V100 (SXM2) x 4
  - Intel Xeon Gold6148 (2.4GHz/20cores) x2
  - 384GiB Memory
  - 1.6TB NVMeSSD

ABCI LLMハッカソン

LLMの学習に関して実践的なノウハウを獲得し、その学びを共有できた他、ABCIのLLMへの取り組みに対する認知向上にも貢献

- 27チームから参加申し込みがあり、先着順に20チームを採択した。
- 開催者が用意したサンプルプログラムを用いたファインチューニングを行うことで、ABCIの使い方に慣れてもらい、その後、各チームが設定したテーマを実施した。
- 参加チームからの声
  - マルチノード分散学習を実際に動かすことで実践的な知識を得ることができた。
  - LLM学習に必要なハード・ソフトのスタックに爆速入門できた。
  - ハイパーパラメータの調整やデータの重要性に気づいた。
  - Slackで質問をしたらすぐに回答してくれて、とても勉強になった。サンプルプログラムもとてもわかりやすかった。実務者の交流やコミュニティが重要。
  - スパコン利用に敷居を感じていたが、ABCIは使いやすくて感動した。
  - 第２回の開催に期待。
- ABCIポイント使用量は平均で670（上限1000）。9チームが残ポイントが100以下とほぼ使い切り。
- Twitter等SNSでのアピールも推奨した。10日間でのユニークビューが113万回とABCI自体やABCIのLLMへの取り組みに対する認知向上に貢献できたのではないか。
- Githubやブログ等で成果を報告、または報告予定のチームもあり。

実世界基盤モデル開発向け大規模クラウド基盤（シンABCI）

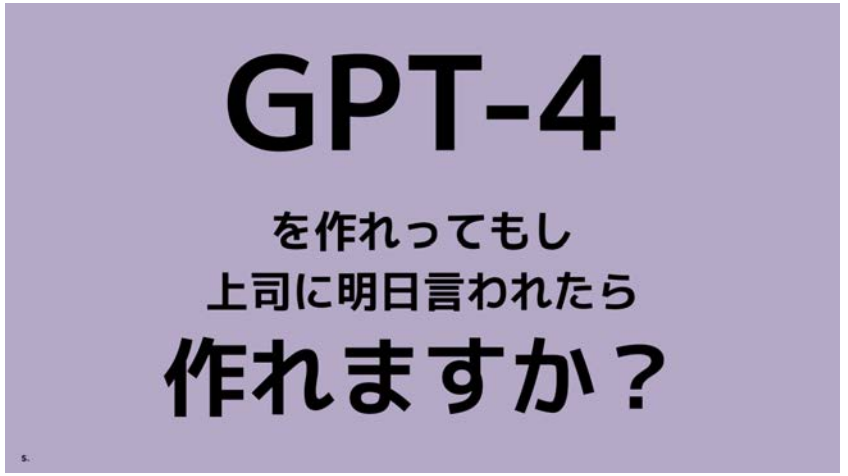
- ポストChatGPT時代の計算基盤としてシンABCIを整備
- 実世界に臨機応変に対応できるAI技術を実現するためのマルチモーダルな基盤モデルや透明性の高い基盤モデルの開発などの様に、研究開発目的でかつ公共性の高い基盤モデルの開発等に計算基盤を優先的に提供する
- 現在のABCIをリプレースして、2024年12月末に導入完了予定。乞うご期待！

ABCIの反省・課題

- 必要なときにタイムリーに資源を提供することができなかった（国から予算を取り、スパコン調達すると2年のリードタイム）。継続的な拡張のロードマップがなく、代替策を提示することもできなかった。
- バッチスケジューラは多くのユーザに公平な資源アクセスを提供するという意味で優れているが、（少数の）大口利用者に資源を切り出すような使い方には向かず、アドホックな対応をせざるを得ない。
- データのライフサイクルがシステムのライフサイクルよりも長い。データ移行のコストや継続的な保管を保証できない。
- セキュリティ要件への対応が不十分であり、機微なデータを預かることが難しい。今後は経済安保の観点からセキュリティクリアランスへの対応などが求められる可能性がある。

今後の展望

- **ABCI規模でしかできない大規模基盤モデル開発**
  - とりあえず基盤モデル作ってみましたという話が一巡すると、プレイヤは収斂されるのではないか。
  - 国研を中心に政策的に重要な課題に優先的に資源を配分する。
- **汎用的な基盤モデルをもとにした個別AIモデル開発の促進**
  - 事前学習よりもファインチューニングの需要が多くなるのではないか。
  - ファインチューニングの構築を容易にする技術や、安全にデータを持ち込める仕組みやポリシ・制度の整備が必要である。
- **AIハブの強化**
  - 生成AIのグローバル市場規模は2030年までに約14兆円にまで拡大（年平均成長率35.6%）[1]。その大半は推論と考えられる。
  - ABCI上の学習だけではなく、民間クラウド等でのサービス（推論）も含めたMLOpsを実現する技術やビジネスモデルが必要である。



LLMの開発は難しい？簡単？Stability AIの現場から - Stability AI 秋葉さん @ W&B Fully Connected Tokyo

[1]総務省、令和5年度版情報通信白書

最後に「AI橋渡しクラウド」と言うけど、

- いつまで橋渡ししてるの？　出口のモデルはどう考えているの。
- どこがクラウドなの？単なるGPUスパコンでしょう。

  という問いに改めてちゃんと向き合う時期が来ているのかもしれない。

- 機能性・利便性・持続性等を踏まえて、クラウドとの連携・使い分けの再整理が必要なのではないか。

- 現時点で解があるわけではないので、続きはパネルにて議論！

https://abci.ai/