資料3-1-1
HPCI計画推進委員会
（第70回）令和8年9月29日

データセンター･クラウド事業者から見る「トークン経済」

Sep. 29th, 2026

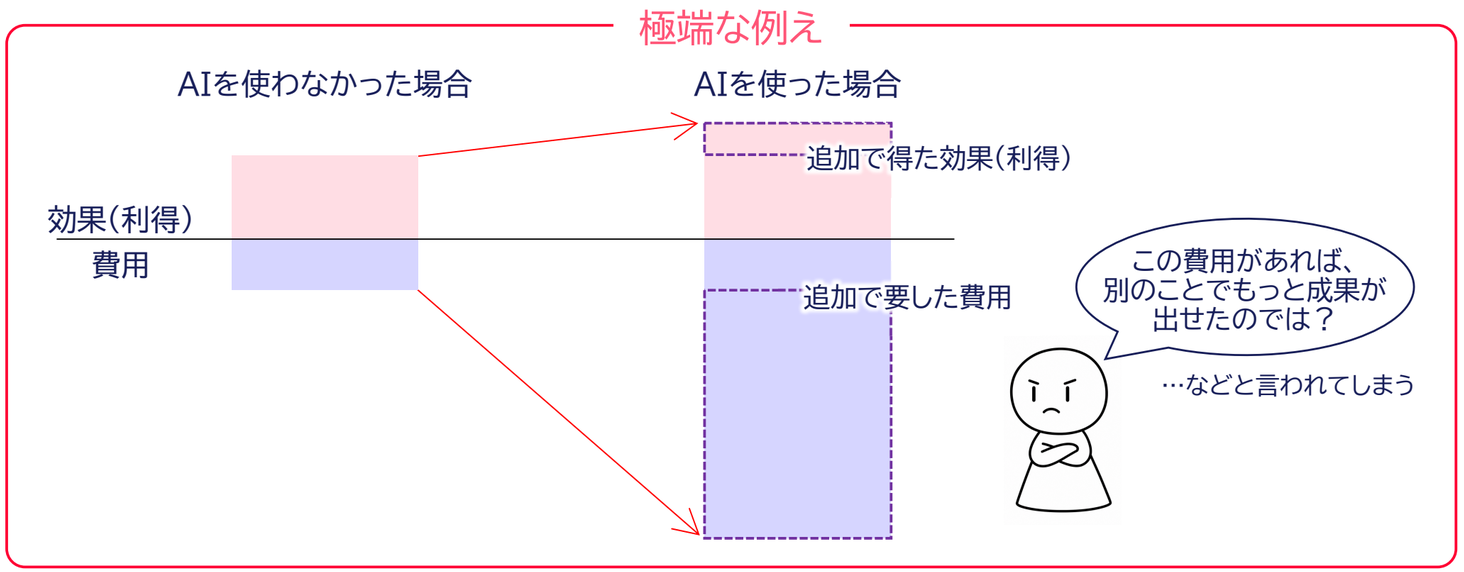
さくらインターネット株式会社
須藤 武文（AI基盤事業推進統括）

SAKURA internet

当日版

AIの経済性

- “AIのための計算資源(基盤)”とは
  - 2026年9月現在、AI学習用の計算資源(基盤)と推論用のそれは大きく異なる
- 学習用の計算資源
  - 雑に言えば、学習はいかに総計算時間を短縮するか /いかに time-to-quality (学習時間あたりの品質 or 品質あたりの学習時間) を向上させるかの勝負
  - そのため、学習用の計算資源はその優劣を比較することが(それなりに)可能
  - 多くの場合、より優れた計算資源はより高性能であり、より高性能な計算資源はより高価となるが、時間を買えるという点において正当化されうる
- 推論用の計算資源
  - 一方で推論、すなわちAIの利用する局面においては「成果創出にどれだけ貢献したか」が問われる
  - 「成果」の定義は無数に存在するため、特定の分野･用途･状況等々で優れた計算資源が、他でも優れているとは限らない(比較が難しい)
  - 創出にかかった費用とのバランス(費用対効果)によって成果が評価される場合も多く、高価な計算資源が常に良いとも限らない



指標の例
- 初回応答時間
- 出力速度
- p99遅延
- 同時実行数
- 要求あたりコスト
- 電力効率

等々…

  - 重視される指標が学習用計算資源と異なることで、システムの構成や運用法も異なり、学習用計算資源をそのまま転用することが最適とはならない

© SAKURA internet Inc.

当日版

データセンター･クラウド事業者から見る「トークン経済」

SAKURA internet

- トークン = AIの計量単位
  - AIが動けばトークンの入出力が生じ、必ず電気が使われることを意味する
- AIは既にあらゆる場面で使われ始めているが、AIへの期待･需要は未だピークを迎えていない
  - より高度なAIはトークンあたりの消費電力が多く、かつより多くの入出力が生じる
  - 一方で顧客の期待は「これ以上のAIの品質向上は不要」という域には達しておらず、トークンの入出力量と電力消費の増大は不可避
- トークンの需要と供給をめぐる経済
  - 「できるだけ品質の高いトークンを、できるだけ安く調達したい」という需要側のニーズに対して、データセンター･クラウド事業者はいかに応えるか
- データセンター･クラウドレイヤでトークンのコストに影響する要素
  - 前提として、AIは本質的にソフトウェア技術であり、ハード･ソフトが同じなら(ほぼ)同品質のトークンを生成可能
  - 事業者間でコスト差が生じやすいもの
    - 土地の取得コスト
    - 電力コスト
    - 自治体による税制優遇
  - 事業者間でコスト差が生じにくいもの
    - 人件費
    - データセンターの建設コスト
    - 計算機などの機材コスト(自前の推論チップなどを用いる場合は別)

© SAKURA internet Inc.

当日版

クローズドモデルとオープンウェイトモデル

SAKURA internet

- 代表的なモデル
  - いわゆるフロンティア級
    - クローズド … GPT-6 Astra, Claude Opus 5.5, Gemini, Grok など
    - オープンウェイト … DeepSeek-V4-Pro, Mistral Large 3 など
  - それ以外
    - クローズド … 各社の廉価･小型モデル
    - オープンウェイト … Llama, Qwen, gpt-oss など多数
- 先端クローズドモデル提供各社はインフラ規模の拡大を追求
  - 先端クローズドモデルを開発した各社によるAIサービスの売上は他を圧倒しており、市場を牽引していることは間違いない
  - 世界最高水準の収益性/将来性を有する各社は、その資金調達力を限界まで行使してDC等の拡張を進めているが、現状は常軌を逸していると言ってよい
    - 現在、資金で解決できない事が各社のブロッカーとなっており、構築は計画よりも遅延しているとされる
    - なお、先端クローズドモデルを提供していてもDC等の物理的なインフラ構築を手掛けていない場合もある
  - 投資金額に見合った収益が得られるのかは不透明だが、十分に手頃な価格でサービスが提供されることへの疑念は高まっている
- オープンウェイトモデルの活用
  - 商用クローズドモデルへの依存から生じる課題への対策として、オープンウェイトモデルの活用に注目が高まっている
  - 例として、米国の大手法律事務所による主な取り組み
    - Kirkland & Ellis …OWMのファインチューニングなども行う自社AI基盤に今後数年で5億ドルを投じる方針 [Reuters, 2026/05/29] ※2025年･売上世界No.1の法律事務所
    - Latham & Watkins …GPUサーバを含めた機材を調達し、OWMのカスタマイズを含めた自社AIインフラを構築 [FT, 2026/09/10] ※2025年･売上世界No.2の法律事務所
    - Morgan & Morgan …3億ドルを投じて自社開発したAIプラットフォーム"MX2"を他の法律事務所向けに販売開始。開発への投資も継続 [Reuters, 2026/09/15]
  - OWMの利用には潜在的なリスクも存在する
    - 英AISIは、危険な能力を持つモデルの場合「持続的かつ不可逆的な悪用リスク」を生み得ると指摘 [参照, 2026/07/17]
- 双方に利点と欠点があり、併用することも選択肢のうち

© SAKURA internet Inc.

当日版

推論用計算資源の動向(中国市場)

※10月7日差替

SAKURA internet

- 政府による戦略の主な柱
  - オープンウェイトモデルにおける優位性の確立
    - Qwen, DeepSeek, MiniMax, Moonshot 等のエコシステムを通じた、オープンウェイトモデルのダウンロード数において世界的な優位性を築く
  - AI導入へのインセンティブの提供
    - 自治体が提供するコンピューティング·バウチャー、税制優遇措置、およびエネルギー補助金による導入促進
  - 標準化の推進
    - 開発プロトコルの統一を図るため、40以上の国家および業界標準を策定・公開
  - “AI Plus”イニシアチブの策定
    - すべての主要な経済分野にAIを組み込むための包括的な指針
- オンプレミスの計算資源(いわゆる「DeepSeekラック」)の流行
  - 企業はハードウェアとソフトウェアがセットアップ済のラックを購入し、自社内で運用する形
    - 米国政府の輸出規制によって最先端プロセッサが入手出来ないことも遠因
  - 恩恵
    - クラウドAIサービスと比べ、データ漏洩の不安が低減する
    - (計算能力の上限までは)使い放題で、従量課金制トークンと比べて費用発生の見通しが明確化する
    - システムの構成を固定することでAIの挙動を自社でコントロールしやすくなる
    - 従業員にとっての魅力にも
  - 副作用
    - 所有者がAIを含めたシステム全体の管理責任およびコストを負うこととなる

大まかな特長

- 全て構築･設定済のターンキーソリューション
  - Deepseek/Qianwen等モデル
  - ハーネス、エージェント環境
  - ネットワークスイッチ/NIC、ストレージ
  - 最大1000ラックまで拡張可能なるものも
- 幅広いxPUの選択肢
  - Nvidia/AMD/Intel/Huawei製GPU
  - Alibaba/Baidu/Bytedance製ASIC
  - Intel/ARM/Risc-V製CPUなど

- 参考資料: AIインフラの最新技術動向2026 – 中国OCPコミュニティから読み解く6つの技術トレンド (2026年7月)
  - https://speakerdeck.com/markunet/aiinhuranozui-xin-ji-shu-dong-xiang-2026-ocp-china

© SAKURA internet Inc.

当日版

トークン経済とHPCI基盤

SAKURA internet

- アカデミアでのAI需要に対する供給設計
  - 言うまでもなくアカデミアにおいてもAIの活用は必須であり、促進施策は既に実行段階にある
  - SPReADの倍率を見れば需要に対して供給が不足しており、トークン供給に関する何らかの手当の必要性は明らか
  - ただし、トークンの量的側面だけでなく質的側面についても考慮は必須であり、その上で適時性と経済性が非常に重要となる
  - これらを勘案したアカデミアに対するトークン供給のロジスティクスを設計する必要がある
- 供給設計におけるHPCI基盤と外部基盤の併用
  - 国内に現存する供給余力と、拡張によって将来的に生まれる供給余力を、量的側面、質的側面、適時性、経済性から整理する
  - 供給元と供給先のアライメントを行う際の基準を整備する
- それはそれとして：推論だけでなく、学習需要への手当も必要？
  - アカデミアにおけるAI活用では、取り組みがモデル開発を伴う場合が考えられる
  - 学習用計算資源についても量的側面、質的側面、適時性、経済性を考慮した提供設計が求められる

© SAKURA internet Inc.

当日版

トークン経済の周辺課題

SAKURA internet

- 本邦が直面するジレンマ
  - 電力コストが安い海外でトークンを生成し、通信で送る方が経済的に有利になりうる = 電力の間接的な輸入（cf. 食料輸入は間接的に水資源輸入である）
  - 安いAIを求めて海外のAI利用が増えれば、海外依存の潜在的リスクが増大する
    - データ主権問題と合わせ、社会･経済活動の見えざる急所を掴まれるおそれが生じる
  - AIに必要な電力の多くを国内で賄おうとした場合にも、
    - AIは、化石燃料を輸入して作る電力を使ってまで国内で賄われるべきか？
    - データセンターが生じさせる大きな負荷変動が電力系統の安定性に与える影響への対処は？
    - 社会･市民生活に影響を及ぼすまでAIの消費電力が増大した時、優先順位を付けられるのか？
- 状況を打開する鍵（下記に限らない）
  - AI処理の電力効率向上 … 半導体、通信、冷却など各要素の改善だけでは不足で、ソフトウェアや運用を含めて組み上げるシステム全体の最適化が必須
  - 需給バランスの調整 … 電力供給の余力をAIが埋める(ワット・ビット連携のような)
  - 脱炭素電源の積極活用 … データセンターが再エネなどをより活用しやすくする工夫
  - トークンの逆輸入 … 電力コストが低く、かつ信頼のおける海外でAI処理(トークン入出力)を行い、日本から利用する
- 多面的な議論が必要
  - 技術開発･人材育成に投資できるか？ / 政策を総合的に連携できるか？ / AIの倫理･正義の議論を整理できるか？ / 外交･安保面の懸念にどう対処するか？
- トークン経済に関する課題や視点は上記に限らず、さらなる議論の掘り起こしと整理が必要

© SAKURA internet Inc.