2 生成 AI とハイブリッド計算基盤

技術潮流最先端での挑戦：生成 AI ×ハイブリッド計算基盤の取り組み

イノベーションセンター テクノロジー部門では、世の中の先進技術要素の目利き・β版開発の推進、他組織への技術支援、高度エンジニア育成に取り組んでいる。ここではオンプレミスとパブリッククラウドを包含したハイブリッドな計算基盤の技術開発、および、計算基盤を活用した生成 AI の取り組みや将来展望について紹介する。

顧客提案と社内検証を加速させるハイブリッド計算基盤への探究

クラウド事業者やハードウェアベンダの技術革新、OSS の成熟や多機能化に伴い、AI や IoT のユースケースに資する「モダンな」計算基盤の最適解も変わり続けている。イノベーションセンターのテクノロジー部門では、そうした技術潮流を追いかけ、パブリッククラウド・エッジ・オンプレミスを組み合わせたハイブリッドな計算基盤の技術開発や社内支援に取り組んでいる。

まず、パブリッククラウド環境に関しては、協業パートナーである Microsoft Azure、AWS、Google Cloud 各社の機能調査や、自社商材との組み合わせ検証を行なっている。特にデータ所在やレイテンシに厳しい要件があるお客様への提案に活かす目的で、パブリッククラウドの機能をエッジへ延伸して利用可能とする機能・製品群に着目しており、Azure Stack ファミリーや AWS Outposts ファミリー、Google Distributed Cloud Edge などの各社製品を日本国内初導入事例も含めていち早く自社データセンターに導入してきた。エッジとクラウドとの接続では Flexible InterConnect のような自社 NW 相互接続サービスを活用し、自社アプリとも組み合わせて実証。Node-AI on AWS Outposts のソリューションリリースや、Interop Tokyo 2023 ShowNet や docomo Open House'24 での 5G 出展を通じてお客様に技術の良さを理解いただくことができた。

NTT コミュニケーションズ株式会社
イノベーションセンター　テクノロジー部門
（左から）担当課長 **張 暁晶** 氏
担当課長 **岩瀬 義昌** 氏

そして、パブリッククラウド上に多数あるデータ利活用のソリューションについても、新たな提案モデルを開拓する目的で Azure Synapse、Snowflake、Databricks などの各種ツールに関して、比較や組み合せ、最新機能の調査検証にも取り組んでいる。

最後に、オンプレミス環境に関しては、最新アーキテクチャの GPU サーバ（NVIDIA H100 や A100 等）や高速ストレージを組み合わせて、Kubernetes クラスタとして運用している。利用者の利便性を考慮し、アプリのデプロイに限らず、分析環境やコンテナレジストリーの周辺機

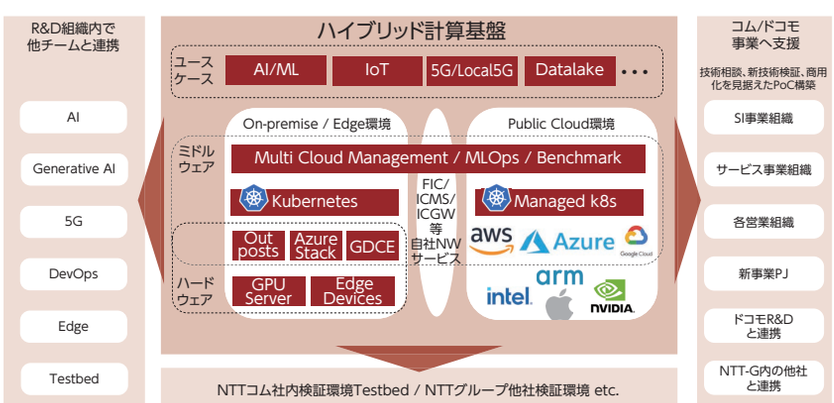

図 1　ハイブリッド計算基盤に関連する取り組み

能も充実させている。また、パブリッククラウド上のマネージドk8sとの間のマルチクラスタ管理など、実運用を通して得た知見を、社内のk8sクラスタ運用者への啓蒙活動に活かしている。

ここまで紹介した計算基盤は全て社内検証環境Testbedを通して社内ユーザへ提供しており、プリセールス、アプリ開発者、AIリサーチャーにとって気軽に試せる場として機能している。

イノベーションセンターの生成AIの取り組み

2022年11月のChatGPTの登場以来、生成AIという言葉が注目を浴びている。生成AIとは、テキスト生成だけでなく、画像や音声などの多様なメディアを出力可能なAIのことを指す。

生成AIの中でも、ChatGPTが当初から対応していたテキストを生成するための機械学習モデルである大規模言語モデル（LLM, Large Language Model）は、各社がオープンなモデルを競って公開するなど、目にする機会が増えている。

このように注目を集めるLLMだが、実際のビジネスで活用する場合には気をつけるべき課題がたくさんある。本記事では2つほど課題と、その課題への取り組みについて紹介する。

1つ目は「最新情報や機密情報への対応」である。LLMは学習時点の情報だけが蓄積されており、新たな情報や特定の企業内情報には対応できない。この問題に対処するために、外部データを取得し入力プロンプトに混ぜるRAG（Retrieval Augmented Generation）という手法や、LLMに追加で事前学習を行う方法（ファインチューニングの一つ）が使用される。どちらの方式にもメリット・デメリットがあり、イノベーションセンターでは、前で紹介した計算基盤を活用して両方式の開発・検証を進めている。ビジネス要件に応じて最適なものを選択できるようにするためだ。

2つ目は「LLM自体の評価」である。LLMの評価には様々な側面がある。トークンの生成速度もあれば、出力された結果の内容の精度といった面もある。特に内容の精度は従来のシステム開発と異なり、テストや評価が困難だ。従来のシステム開発では、「このリクエストにはこのレスポンスを返す」といったようにテストの入力と出力が明確に対となることが多かった。だが、LLMの出力は（設定次第ではあるが）一意に決まらない。そのため、従来とは異なる評価を進める必要がある。

そこでLLMに対してさまざまな評価手法が提案されている。本記事では、昨今LLMの性能評価に利用されるLLM-as-a-judgeを紹介する。LLM-as-a-judgeを簡単に言えば、「LLMの評価自体をLLMに実施させる」方法である。この方式の具体的な実装の1つであるRakudaベンチマーク※1では、1）LLMを2つ選んで出力を生成、2）2つの出力をLLM（例えばGPT-4）が比べて、Win/Draw/Loseのいずれかを判別、という順で評価をする。これらの情報を活用すると、LLMがN個あったとしても、どのモデルが優れているかランキングで分かるようになる。

実際にイノベーションセンターでも、Rakudaベンチマークを使ってLLMの評価を実施しているが、実際の検証ではもう一工夫を加えている。Rakudaベンチマークが従来備えているプロンプトは、汎用的な質問が大半である。実際に仕事や業務ドメインで活用する場合は、汎用的なものではなく、業務に特化したプロンプトを活用して評価を進める必要がある。そうしなければ、実際のビジネス要件では、どのLLMが最適なLLMか判断できないためだ。この対応として、社内で実証実験を進めるLLM付加価値基盤（社内用のChatGPTに近いもの）で実際に入力されるプロンプトを参考に、評価用のデータセットを作成して評価を進めている。評価のためのベンチマーク実行にもGPUインフラが必要なので前述の計算機基盤を利用している。

NTTグループで独自開発したNTT版LLMであるtsuzumiの活用に向けて

ここまでで述べてきた取り組みで利用するLLMには、NTTグループで独自に開発したtsuzumi※2や他のオープンなLLMの両方を対象として開発・検証を進めている。

我々は研究所の開発物を社会実装するミッションを担っており、tsuzumiはさまざまな顧客要件へ対応できる高い可能性を秘めていることから社内の期待は高い。

だが、実際に活用する場合は、要件ごとにさまざまなチューニングなどが必要になることを予想しており、最適手法の提案・実装に向けて引き続き、研究開発を進めていく。

※1: https://yuzuai.jp/benchmark
※2: https://www.rd.ntt/research/LLM_tsuzumi.html