通信ビルエッジを活用したGPU・データレイクの技術開発

NTT東日本デジタル革新本部デジタルデザイン部では，ネットワーク事業推進本部高度化推進部と連携し，当社が保有する通信ビルの特徴を活かしたサービスの実現に向けて，映像AI（人工知能）解析技術，データレイク技術の確立に取り組んでいます．映像AI解析では，AI推論のコアとなるGPU（Graphics Processing Unit）を複数のユーザでシェアし，低コストに利用可能とする技術，およびAIアプリケーションの開発，試験から商用サービスへの適用を確実かつスピーディに実現するコンテナベースの運用手法を確立するとともに，多種多様なAIサービスをプラットフォーム上で提供するためのサービサー向けAPI（Application Programming Interface）の開発に取り組んでいます．また，データレイクでは，カメラ映像や医療系情報等の地域に蓄積されたさまざまなデータをセキュアに蓄積し，利用者が目的に応じて必要な情報を取り出し，AI等を用いて解析することができる技術の確立をめざしています．

技術開発の背景

近年，膨大な数のIoT（Internet of Things）デバイスやカメラ等の大容量データをAI（人工知能）技術を活用して解析することにより，さまざまな価値を生み出す利用形態が広がっています．これらの利用において，データをクラウド上で処理する場合，クラウドまでの通信コスト，遅延，セキュリティ・安心感の課題が想定されます．また，データを端末やオンプレミスなどのユーザ拠点で処理する場合，初期投資や保守の課題があります．そこで，データの処理や蓄積の機能をクラウドとユーザ拠点の間に位置する「通信ビルエッジ」に配置することにより，これらの課題の解決が期待できます（**図1**）．

一方で，国内の地域を見渡すと，人手不足の解決やニューノーマルの模索などのさまざまな社会課題があります．そこで通信ビルエッジを活用し，地域のさまざまな課題の解決をめざす「REIWAプロジェクト」[*1]において，PoC（Proof of Concept）を通じた技術の確立とユースケースの開拓

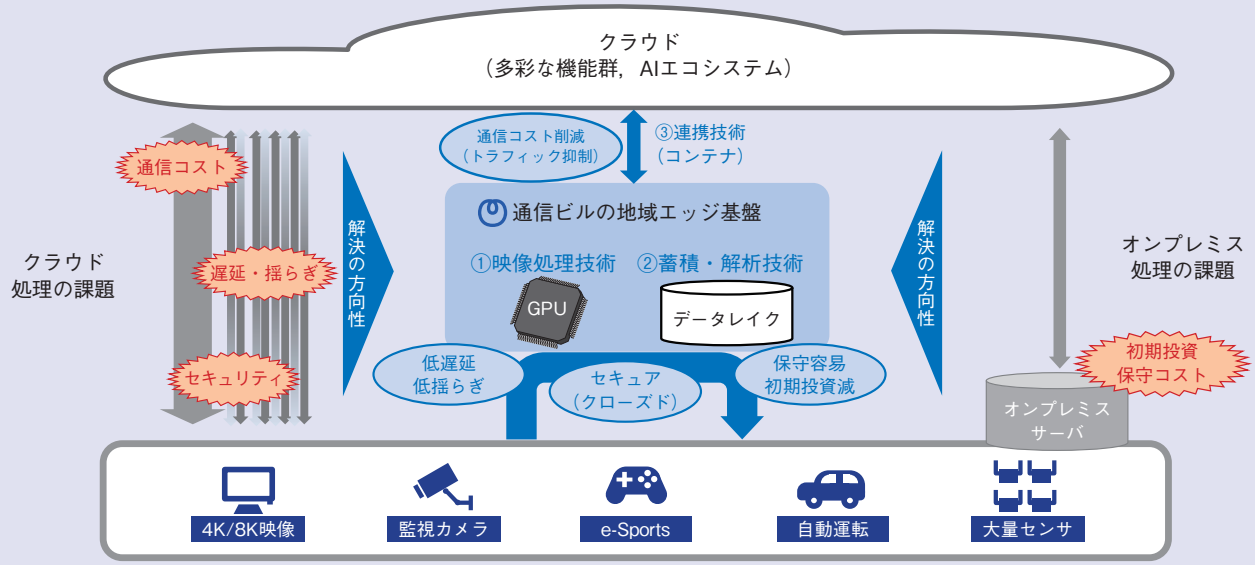
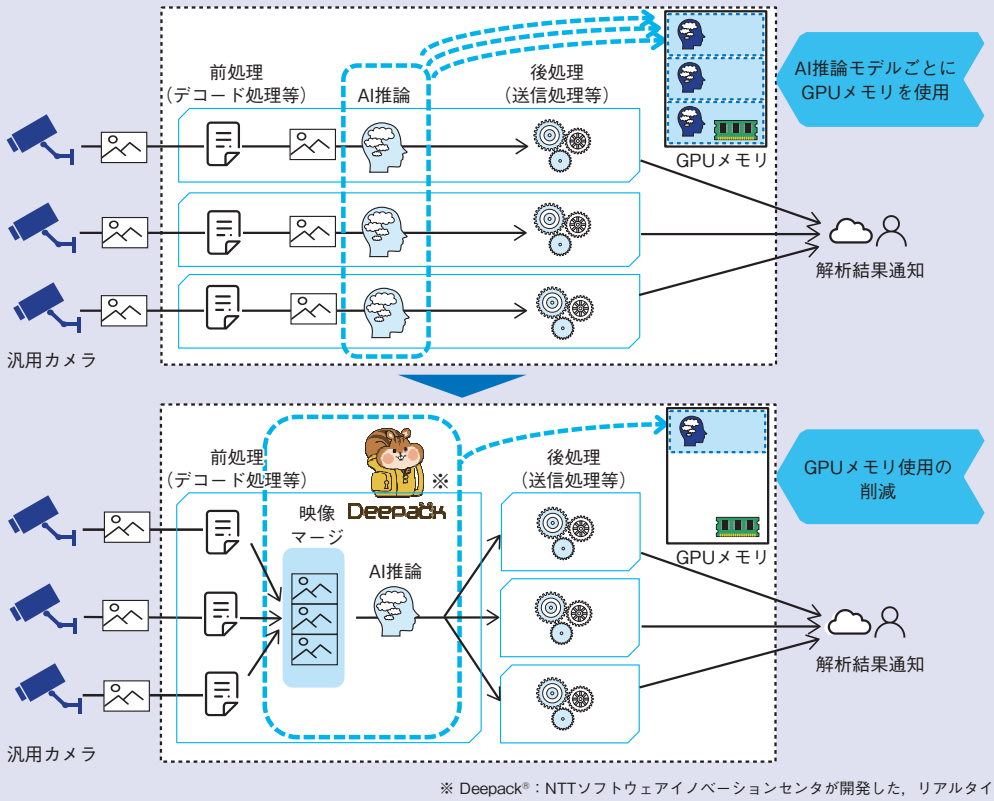


図1　通信ビルエッジへの機能配置による解決



※ Deepack®：NTTソフトウェアイノベーションセンタが開発した，リアルタイム映像解析技術

図 2　GPU高収容化

を進めています．具体的には，GPU（Graphics Processing Unit）*2を用いた映像のAI解析技術，大容量の映像や地域に散在するデータをセキュアに蓄積し使いたいときに利用可能なデータレイク技術を確立し，その有効性の検証を進めています．

映像AI解析技術

（1） GPU高収容化技術

高度な映像AI解析を実現するためには，AI推論*3を行うGPUが必要となります．しかし，これまでの映像AI解析サービスでは，GPUを搭載するコンピュータをカメラと一体で提供する形態が主流でした．この形態では，カメラごとに高価なGPUを占有して使用するため，高コストとなることが課題でした．そこで，通信ビルエッジに設置したGPUサーバをネットワーク経由で複数カメラ，複数ユーザで共有利用することで，GPUを低コストに利用可能とすることをめざし，GPU高収容化技術の確立に取り組んできました（**図2**）．

GPU高収容化技術は，NTTソフトウェアイノベーションセンタが開発したリアルタイム映像解析技術「Deepack®」を導入し，複数の映像ストリームを束ねてリアルタイムにAI推論を行う方式により実現しました．これはAI推論を一括で行うことでGPUメモリの使用を低

＊1　REIWAプロジェクト：Regional Edge Interconnected Wide-Area Networkの略．NTTのアセットを活用し，地域課題を解決，地域貢献する弊社のプロジェクトの総称．
＊2　GPU：3Dグラフィックスなどの画像描写処理を行う半導体チップ．ディープラーニングを使ったAI学習やAI推論においても活用されています．
＊3　AI推論：AI学習で作成されたＡＩ推論モデルに，画像等のデータを入力し，解析結果を得ること．

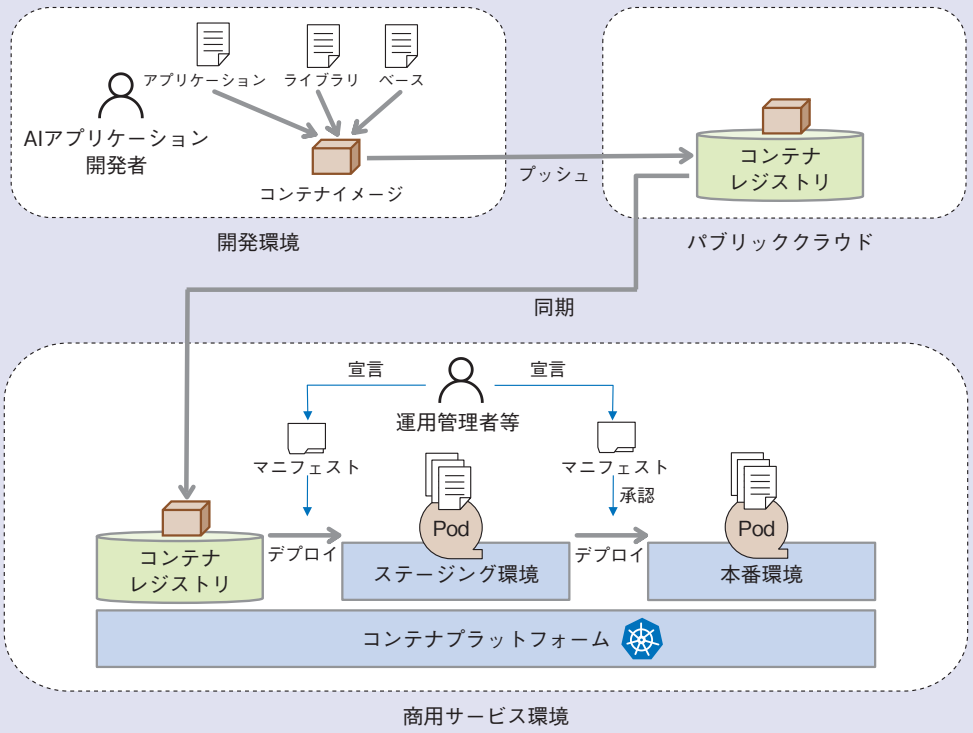


図 3　コンテナベースの運用

減し，同時にAI推論を可能とする映像ストリーム数を増加させるものです．さらに，このAI推論処理，およびAI推論の前後処理をハードウェア（GPU，CPU），ミドルウェア，OS，アプリケーションをレイヤ横断的に最適化することにより，GPUへの映像ストリームの収容数を世界的に類を見ないレベルに高めています．

（2）　コンテナベースの運用

開発したアプリケーションを商用サービスの環境に適用するには，開発環境で手順書を作成し，その後，商用サービスの環境で手順書に従って構築する手法が一般的でした．しかし，この手法では人的ミスやソフトウェアバージョンの不一致などによる構築時のトラブルのリスクがあります．加えて，AIアプリケーションではGPUやさまざまな推論モデルを利用するため構成が複雑化しており，手順書による構築に代わる構築手法の確立が必要となります．これらの課題を解決するため，開発環境と商用サービス環境の間のポータビリティを実現するコンテナ*4技術を採用しました．また，コンテナの管理には宣言型のコンテナオーケストレーションを採用することにより運用の自動化を実現し，さらにアプリケーションデプロイの自動化にも取り組んできました（**図3**）．

まずAIアプリケーションの開発者はクラウドなどの開発環境で作成したコンテナイメージをコンテナレジストリに置きます．次にコンテナオーケストレーションによるアプリケーションのマニフェスト（宣言）の変更をトリガとして自動的にコンテナイメージをステージング環境にデプロイを行います．検証後，運用管理者の承認により商用サービスの本番環境へのデプロイを行います．これによりAIアプリケーションのような更新頻度の高いアプリケーションをスピーディに商用サービス環境へ配備できる仕組みを

*4　コンテナ：アプリケーションが動作するのに必要な実行ファイル・ライブラリ一式をパッケージ化することによりアプリケーションをさまざまな環境上で動作可能とする仮想化技術．近年急速に普及し世界の主流になりつつあります．

実現します．

（3） AIサービサー向けのAPI

映像AI解析の商用サービスでは，AIアプリケーションを開発する多くのサービサーにNTT東日本がAI推論の基盤を提供することにより，エンドユーザに多種多様な映像AI解析サービスを提供することをめざしています．そのためには，AI推論の基盤が安価に利用できるだけでなく，AIサービサーにとって便利に利用できることが重要となります．そこで，AI推論の基盤をAIサービサーに提供するAPI（Application Programming Interface）の開発に取り組んでいます（**図4**）．このAPIの開発では，国内外のAI関連イベント，コミュニティなどのオープンな場で情報発信を行い，AIサービサーの方々のご意見，ご要望をいただきながら技術開発を進めていく予定です．

データレイク技術

（1） リアルタイム処理と蓄積処理

AI解析サービスでは，データをリアルタイムに解析するだけでなく，蓄積されたデータを利用してAI解析を行うサービスが考えられます．例えば，店舗に設置されたカメラの映像解析では，不審者の検出などの防犯目的の利用はリアルタイムでのAI解析が求められますが，顧客の動線分析などのマーケティング目的の利用では，蓄積された映像を後日引き出してAI解析を行うことが考えられます．このような利用では，大量のデータを低コストかつセキュアに蓄積し，必要に応じて速やかに利用できる仕組みが必要となります．そこで通信ビルエッジを活用したデータレイク技術の確立に取り組んでいます．

（2） データレイクのアーキテクチャと要素技術

データレイクは，カメラ映像や地域の各種データベースから収集したデータを構造化・非構造化などのデータ種別やセキュリティなどの条件に基づいて，コストや性能が最適なストレージに蓄積するストレージサービス（**図5**①），蓄積されたデータを利用するときに目的に応じてデータを変換，抽出するデータレイクコンポーネントサービス（図5②），およびユーザ管理やオペレーションなどの非サービス系機能（図5③）を具備します．さらにサービスAPIを介してクラウド上などの各種データ解析サービス（図5④）や映像AI解析のプラットフォーム（GPU高収容化）と連携します．

OSSコミュニティ，ベンダ製品，NTTグループサービス，NTT研究所から優れた技術を取り入れ，REIWAプロジェクトのユースケースへの技術展開，フィードバックを通じて技術を磨き，データレイクサービスの実用化に向けて取り組みます．

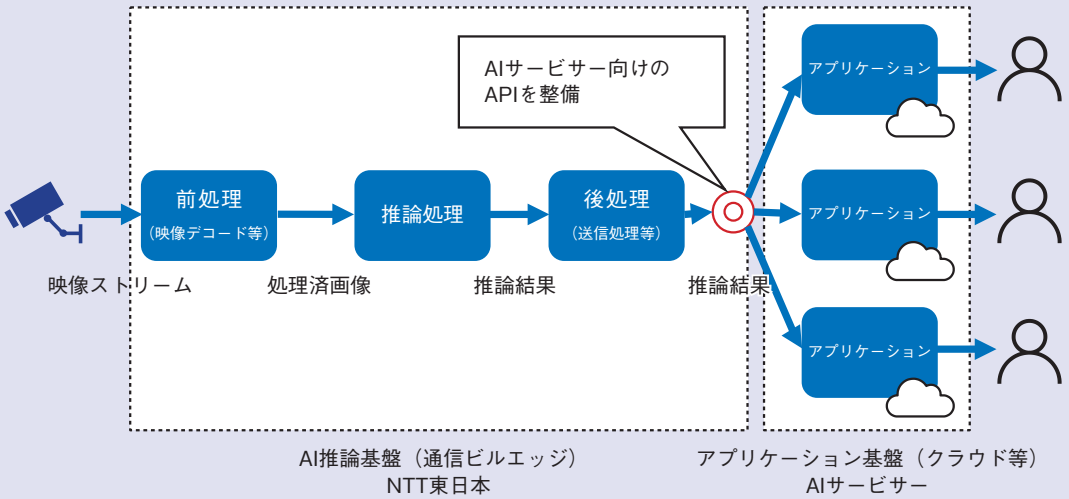


図4　AIサービサー向けAPIのイメージ

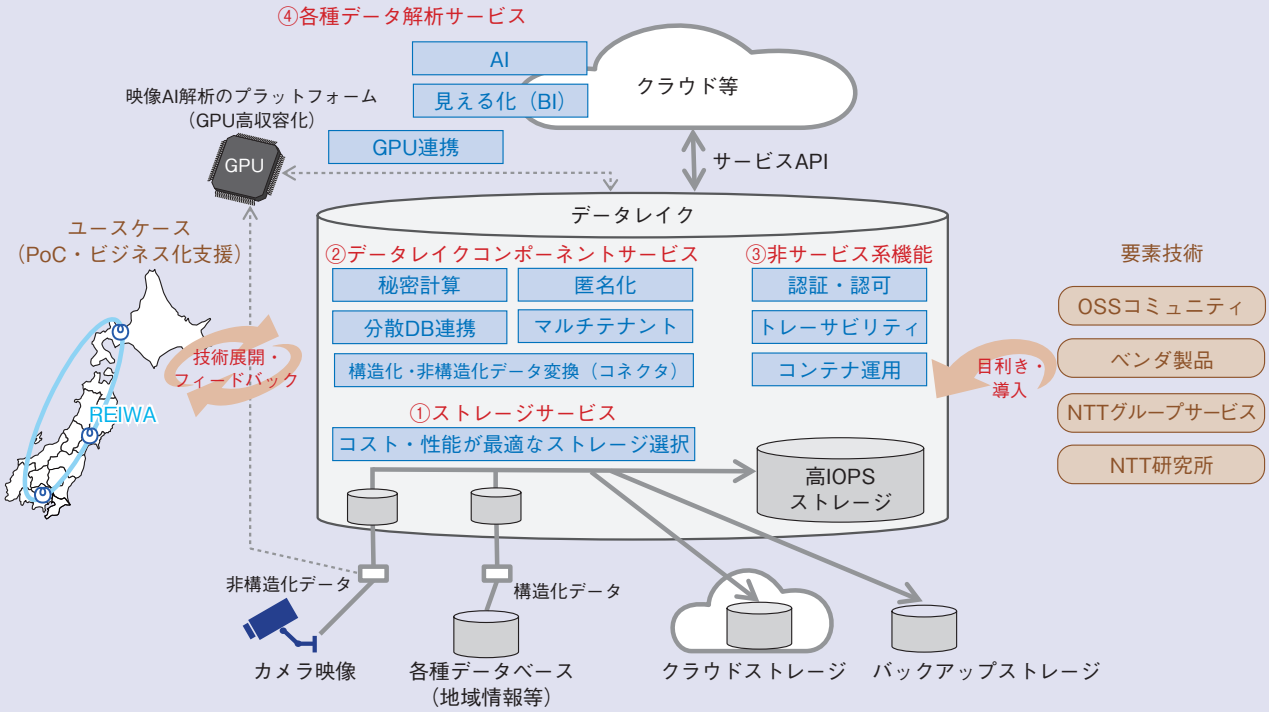


図5　データレイクの構成

まとめ

映像AI解析技術，データレイク技術の開発では，ハードウェア，ソフトウェアのレイヤ横断的な知見，ノウハウを必要とし，方式策定，PoCから商用展開までをスピーディに進めることが求められます．そこでビジネス開発本部，ネットワーク事業推進本部，デジタル革新本部による組織横断的な“One Team”により開発を進めています．さらにNTTグループ企業やNTT研究所，先端技術を有するグローバル企業やベンチャー企業，地域の産官学のパートナーとオープンに連携し，社会課題の解決に資するプラットフォームの実現に取り組んでいきます．

◆問い合わせ先
NTT東日本
デジタル革新本部　デジタルデザイン部 プラットフォーム開発部門
TEL　03-5359-4185
E-mail　dd-pfd-arc@east.ntt.co.jp