2025年度未踏アドバンスト事業

オンプレと連携可能な Wasm に特化したクラウドサービス
－ Wasmによる更新容易で軽量なエッジAI実行基盤 －

野崎 愛（東京大学大学院）・上田 蒼一朗（京都大学大学院）

エッジAIシステムの運用における課題：AIモデルの頻繁な更新

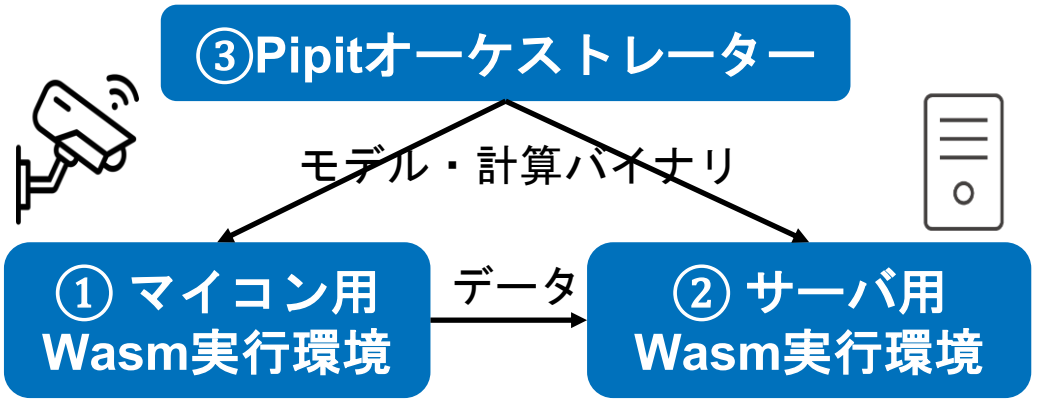
一般の推論基盤ではコンテナやPythonのパッケージングマネージャーが利用されている
しかしエッジAI環境では
❌ **リソース制約**：大きなバイナリサイズは扱えない（Ex. Pytorch入りDockerは2GB超）
❌ **多様な実行環境**：多種多様なCPU・OS・NPUの組み合わせ
→ **軽量かつポータブルな推論実行環境がエッジAIに求められている**

Pipit: WebAssemblyによるエッジAI実行環境

モデル・計算の配布フォーマットとしてWebAssembly (Wasm) を採用したプラットフォーム

⭕ **軽量**：推論モデルと前後処理のみを含むバイナリ（推論エンジンを含まないためMBオーダー）
⭕ **ポータブル**：WasmはCPU/OS非依存

これにより
- リソース制限の厳しいエッジ環境で俊敏な推論更新を可能に
- エッジサーバ基盤（ex. 基地局・CDN）での推論ワークロードの凝縮性を高める



2025年度未踏アドバンスト事業

オンプレと連携可能な Wasm に特化したクラウドサービス
－ Wasmによる更新容易で軽量なエッジAI実行基盤 －

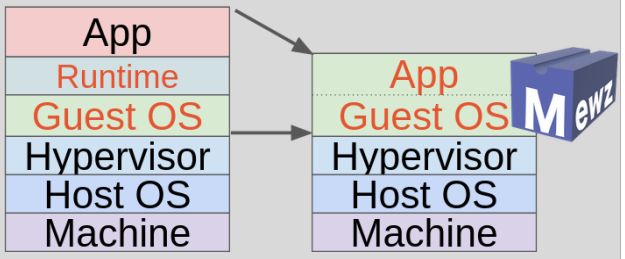
野崎 愛（東京大学大学院）・上田 蒼一朗（京都大学大学院）

waiot
－ マイコン用Wasm実行環境 －

ESP32等のマイコン上でアプリケーションを再起動なしで動的に実行可能にするランタイムである。従来の組込み開発ではファームウェアとアプリケーションが一体化し、更新には分単位の再起動が必要であった。waiotではこれらを分離し、waiotがファームウェアとして動作し、その上でWasmが実行される。これによりネットワーク越しにWasmのみを秒単位で更新可能となる。

Mewz
－ サーバ用Wasm実行環境 －

Wasmを実行するのに特化したOSカーネルである。仮想マシン上で高い隔離性を保って実行できる。Wasmを実行するのに最低限の機能のみを持ったOSであり、Linux等の汎用的なOSと比較してより軽量かつ高凝集にWasmを実行可能である。ResNetを実行するPythonコードをコンテナ化するとPytorch等の推論ランタイムも含みイメージサイズは25GBほどになる。一方でPipitで配布するWasmバイナリは100MB程度で同じ機能を実現できる。



Pipit
オーケストレーター

Kubernetesをベースにしたシステムの統合管理ツールである。クラウドのインスタンスやエッジサーバーからLinuxが搭載されていないマイコンまでKubernetesで管理する。クラスタにWasmをデプロイすることで統一的なインターフェースでwaiotやMewzでWasmを実行可能である。