FUJISOFT

次世代IoTに向けたAIの組み込み実装への取り組み

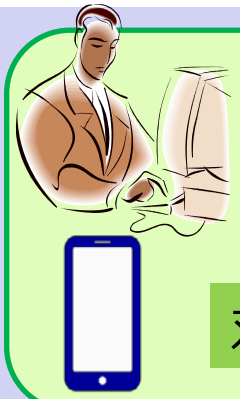
AIの推論機能をFPGAに実装するための技術とソリューション提案

Embedded Product Business
Development Department
FUJISOFT INCORPORATED

Agenda

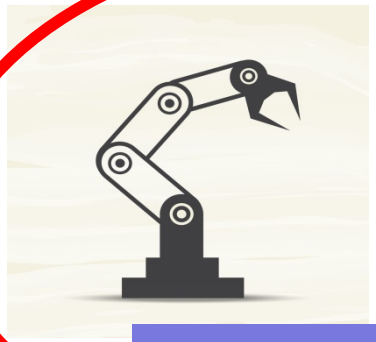
1. エッジAIの現状
2. 組み込みAIのニーズ
3. FPGAとエッジAI
4. 組み込み向けエッジAI実装の特性(GPUとFPGA)
5. エッジAI導入に向けた計画と検証の重要性
6. エッジAI設計とFPGA実装の提案
7. 富士ソフトのエッジAI実装サービス

エッジAIの現状　-1



- 推論アクセラレーション
- 機能特化
- SW-AIアプリ
- WebDNN

対人向けエッジAI



- 高速推論
- 常時推論
- 低遅延・定常性
- 機能特化
- 小クラス分類

マシン向けエッジAI

エッジAI

デバイス組み込み型モデル
(ローカルで用途特化・限定カテゴリ分類)

クラウド型
AIサービス
多種多様で従量課金

サービス型ビジネスモデル
(Cloudで多用途・多数カテゴリ分類)

クラウドAI

エッジAI：特定機能特化型・小型軽量・最適化・スタンドアロン動作・省電力

エッジAIの現状　-2

今、どういう課題があり、なぜエッジAIが必要とされているのか？

- **第１位：　コスト**
  - リアルタイム・センサーデータを使ったクラウドAIベースの傾向予測：
    運用費：月額５～３万円＋通信費/件　→　～５０万円/件の年間運用コスト
- **第２位：　レイテンシー(遅延)**
  - 自動作業工程でのAI機能導入を検討したが、クラウドAIではネットワーク遅延等で成立不可
- **第３位：　高速処理と最小化**
  - AIによる画像の数種類分類（軽いAI）を安定的に高速処理（マシンスピード）で推論と十分な処理帯域が必要

エッジAIの現状　-3

AI技術進化の追従性がポイント

- 日々進化途中の技術　→6カ月後には新しいブレークスルーの可能性
- 多様性と最適化　→目的や規模ごとにAI最適化が進み進化が細分化

最大の資産は学習データ

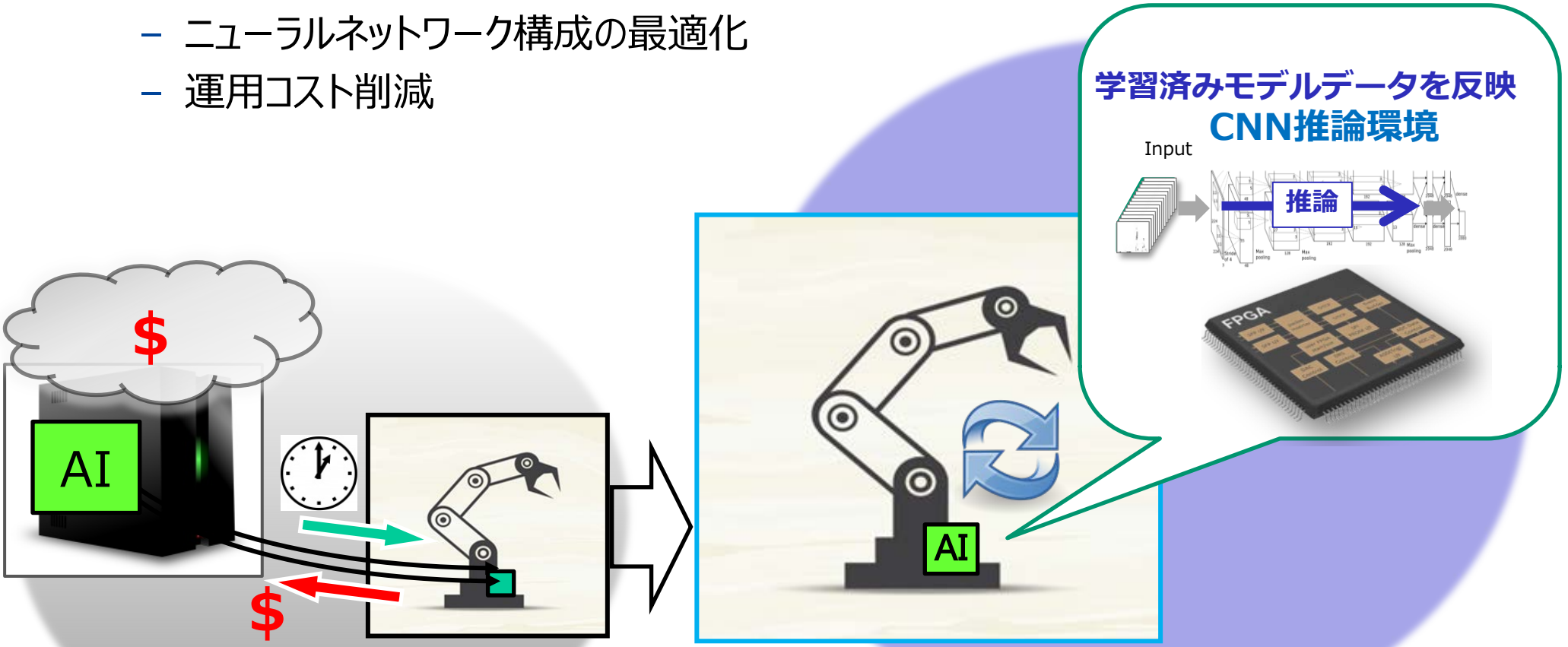
- 学習データがあれば、随時、新しいAI技術へ乗り換え可能

既存サービスでの落とし穴

- 囲い込まれていませんか？
- 標準開発環境・オープンスタンダードは？
- ノウハウの流出の懸念

組み込みのAIのニーズ　-1

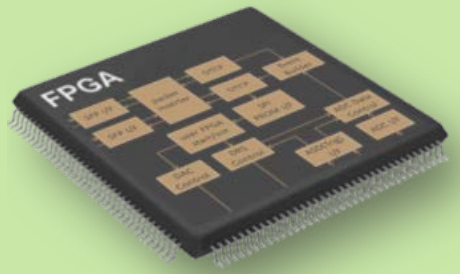
- **クラウド・サーバー非依存型　→AI推論機能を機器に組み込む**
  - レイテンシー(遅延）最小化
  - ニューラルネットワーク構成の最適化
  - 運用コスト削減



用途に合った適正な精度と最小構成のバランス重視型

組み込みAIのニーズ　-2

- 低遅延で確立された安定したレスポンスが必要
- セキュリティ重視のインターネットを使わない環境
- 既存システムの流用・連携、省電力、省スペース
- 特定の用途に特化した最適化AI設計
- マシン制御系連携の処理速度に準じた推論速度と帯域
- 使用環境の自由度（ファンレス・温度拡張対応・連続運用・長期供給）

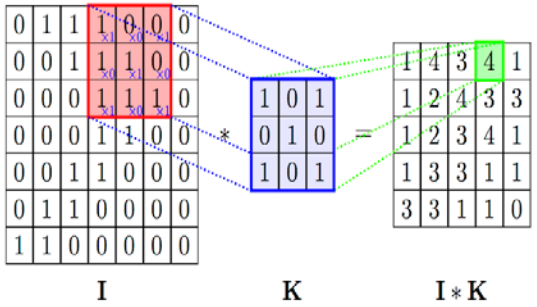


組み込みAIのニーズにはエッジAIの形態が有効　→FPGAの有効性

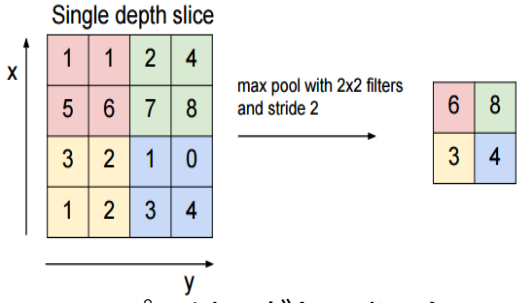
FPGAとエッジAI -1

- CNNの特徴は、畳み込みとプーリング
  - 畳み込み：特徴マップ、カーネルなどと呼ばれる2次元フィルタを1ピクセルづつずらしながら元画像に重ね合わせ、特徴を検出する
  - プーリング：近隣の特徴量を一つの特徴量としてまとめて、ニューロンの数を減らす（低解像度化する）

いずれも2次元のデータ配列を何十万回もスキャンして積和演算する
⇒ **極めて大量の並列演算**



畳み込み(Convolution)



プーリング(Pooling)

FPGAによるCNNの高速推論処理が有効

膨大な行列積和演算をFPGAで高速並列処理
⇒ CNNの高速演算を実現

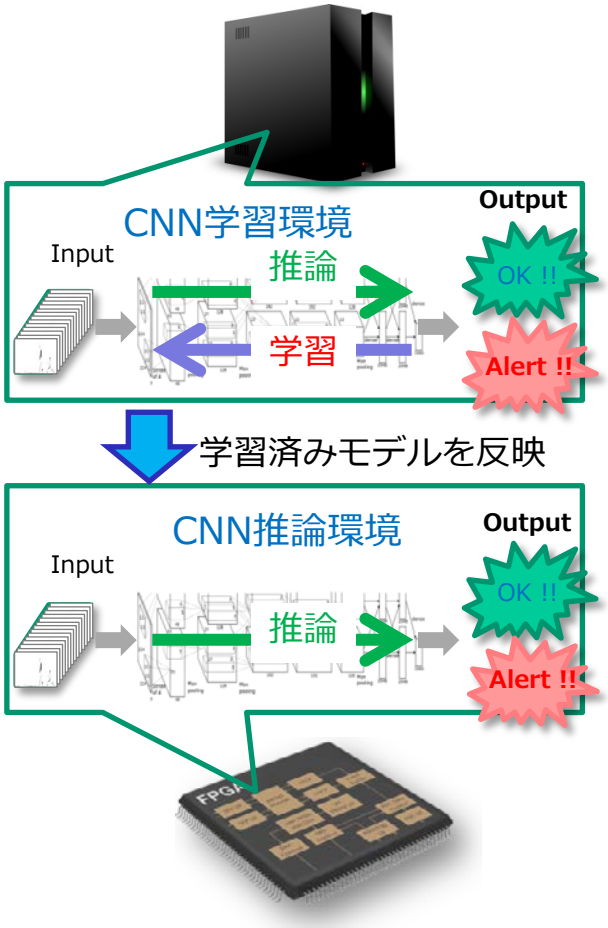
FPGAとエッジAI　-2

❖ **CNNでの並列演算は、1つ1つの演算は単純だが回数が膨大**

- ◆ コア数、規模、メモリーアクセスを自由にデザインできるFPGAは並列演算を最適できる
- ◆ 組み込まれたコアを常に動かし、外部メモリーへの常時アクセスが必要なGPUより、最適な数のコアを生成するFPGAの方が消費電力を最適化できる

❖ **学習時の除算はロジック消費が大きい　→FPGAでは非効率**

- ◆ 学習はサーバ側で時間をかけて行う必要があり、エッジ側ではリアルタイムに学習処理するニーズは薄い　→サーバー処理が有効
- ◆ エッジ側は性能保証・低コスト・省電力・省スペースが求められる



学習はパフォーマンス重視のサーバ側で実施　・　エッジ側ではFPGAが推論だけを実施
電力効率が求められ、長期的な利用を想定したIoTエッジ端末にはFPGAが適している

組み込み向けエッジAI実装の特性(GPUとFPGA)
- AIの推論機能を組み込むには？

① AI推論機能をプログラム実装し機器内のプロセッサ(CPU)でSW処理

- ◎ 追加のHW不要で容易に実装
- × 推論機能の制限と処理速度の問題
- × 既存処理とプロセッシング・リソースの食い合い

② GPU＆メモリーをモジュール化して機器に組み込む

- ◎ 一般的に学習と近い環境で学習済みネットワークモデルを利用できる
- ◎ 高速推論性能
- × 発熱対策の考慮が必要　→サーマルスロットリングによる処理速度の低下
- △ 製品のライフサイクルを考慮した運用

③ FPGAで実装する

- ◎ 省電力・省スペースでの実装が可能
- ◎ 安定・高速推論性能
- ◎ 安心感：　組み込み市場での実績、長期供給、拡張温度対応
- △ FPGAの規模に応じたDNNの設計や調整が必要

組み込み特有のニーズと制約　→FPGAが優位

組み込み向けエッジAI実装の特性(GPUとFPGA)

- 組み込みでの有効性比較

FUJISOFT INCORPORATED

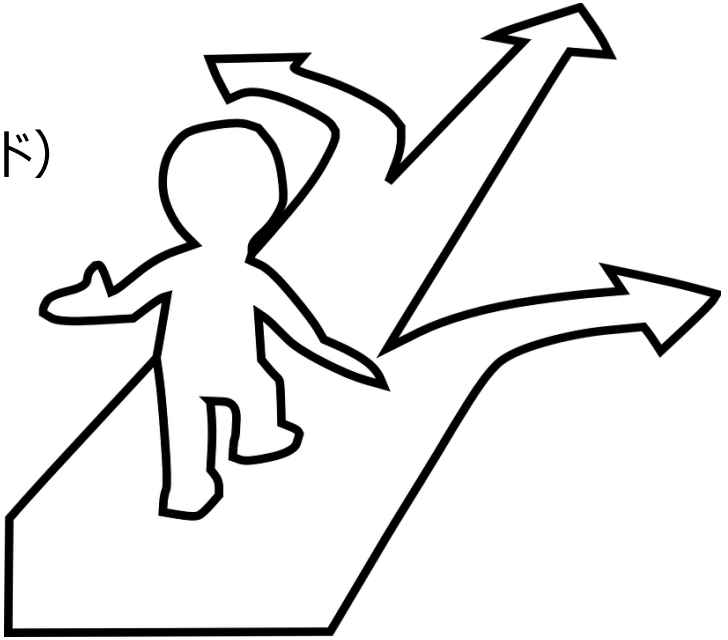
	FPGA	GPU	CPU
■ 安定した性能	◎	△	×
■ 電力効率(発熱)	◎	×	×
■ 省スペース	◎	△	◎
■ 推論速度と帯域	○	◎	×
■ 温度拡張品	◎	○	○
■ 工業用使用実績	◎	△	◎
■ 長期供給	◎	×	△
■ 価格	△	×	○

組み込み向けエッジAIではFPGAの有効性が高い

エッジAI導入に向けた計画と検証の重要性　-1

AI開発時の選択

- 学習データ
- AIアルゴリズム・フレームワーク
- 学習環境（オンプレミス　or　クラウド）



FPGAの選択

- 回路規模や動作速度
- パッケージや動作温度範囲
- メーカー

高位合成環境の選択

- FPGAデバイスメーカにより開発環境が異なる
- オンプレミス環境　or　クラウド環境

最適な実装には多くの選択肢　→　多種多様な専門知識が必要

エッジAI導入に向けた計画と検証の重要性 -2

- AI導入には大きな費用が発生 → 事前の費用対効果検証が重要
- 学習データの収集と作成が精度に大きく影響
- AI技術は日々革新の連続 → 最新AI技術へ短期間で移行が前提

製品開発の前に、AI導入の効果検証を
最小限のリソースで短期間に実現することが重要

→ 適切なAI性能・機能の見極め
→ 短期間での実装、改良、アップデート
→ 適切な回路規模のFPGAデバイスの選択

FPGAで費用対効果検証・ニーズに合ったエッジAI設計が有効

Agenda

1. エッジAIの現状
2. 組み込みAIのニーズ
3. FPGAとエッジAI
4. 組み込み向けエッジAI実装の特性(GPUとFPGA)
5. エッジAI導入に向けた計画と検証の重要性
6. エッジAI設計とFPGA実装の提案
7. 富士ソフトのエッジAIソリューション

エッジAI設計とFPGA実装方法

学習

サーバー
+
一般的な
Deep Learning
フレームワーク

学習データ
+
学習済みネットワークモデル
精度評価

- FPGAやAI規模に応じたDNNのデザイン
- 学習データの再利用
- 標準ツール (インテル® FPGA SDK for OpenCL™) で高位合成
- 高位合成後の実行ファイルをFCGWで検証

ラベルデータ

学習データ

FPGA毎に最適化したDNNをデザイン

C++のヘッダファイルに変換

FPGA用再学習で重み生成

標準のインテル® FPGA SDK for OpenCL™でコンパイル

SoC FPGA実行ファイル

Binary

aocx

推論

インテル® Arria® 10 SoC FPGA
内蔵IoTフォグコンピューティングゲートウェイ

開発　機能/性能検証　POC製作を　同時進行

富士ソフトのエッジAIソリューション：IoTフォグコンピューティングゲートウェイ

- フォグコンピューティングを想定した高性能ゲートウェイを開発
- エッジAI向けPOCプラットフォーム
- FPGA評価キット

インテル® Arria® 10 SoC FPGA搭載
ARM：800MHz Dual Core
FPGA：160K LE
(320にマイグレーション可)

組み込みOS
Ubuntu 16.04 LTS , Kernel 4.1.33

省スペース・省電力
筐体サイズ 220×156×90
組み込み機器ならではの省電力

センサーI/F
USB 2.0 Host x 4ポート

ネットワークI/F
GbEthernet x 2ポート
Wi-Fi (Option)
LTE/3G (Option)

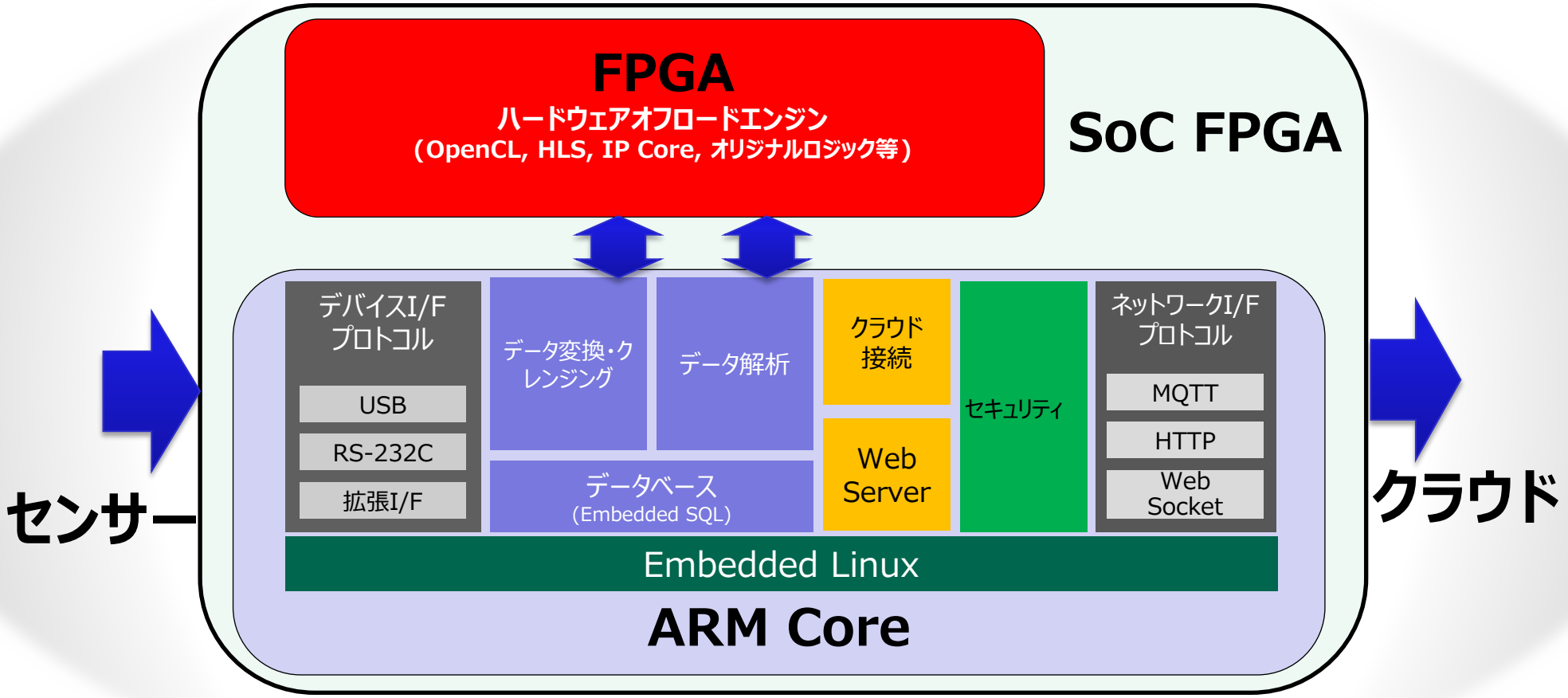
高い拡張性
mini PCIeオプションボードによってインターフェースの追加が可能

日々進化する脅威に対応し、異変を検知・防御するセキュリティ
TrendMicro社の IoT Security (Option)

現在、限定サンプル（Beta 版）.　製品版：量産Q1'18予定

IoT フォグコンピューティングゲートウエイ
- SW実装例

- ゲートウェイとしての通信機能は SoC FPGA の ARM部分で処理
- フォグコンピューティングのデータ制御・AI推論の演算処理をFPGAにオフロード



エッジAI設計とFPGA実装の提案

- 内容：推論エンジンをFPGAに実装するためのAI開発者向けセット
- 目的：POC(概念実証)の組み込みAIシステム実装開発
  ⇒エッジAI開発・評価環境・POC開発を集約

FPGA向け推論CNN開発・実装ツール

FPGA内蔵
IoTフォグコンピューティング
ゲートウェイ
開発キット

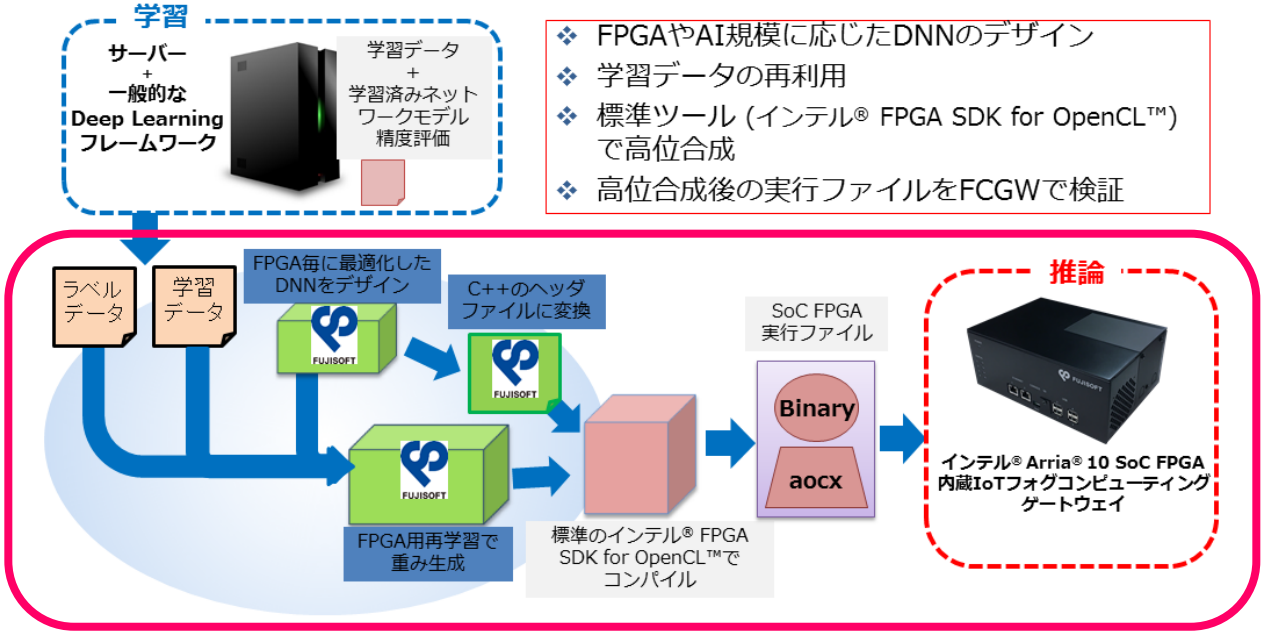
FUJISOFT

Deep Learning学習環境一式

エッジAIの実装開発環境をパッケージで提供 （2018年１H予定）

“AI-ON-FPGA”エッジAI実装サポートサービス

エッジAI実装コンサルテーション(12月予定)



AI-ON-FPGA 実装技術サポート

- **AI技術者による、メニューに沿った技術サポート提供**

商用ライセンス販売

- **各種NNモデルのRTLライブラリー**
- **推論用DNNのIP**

推論FCGW提供

FPGA内蔵 フォグコンピューティング ゲートウェイ（FCGW）

FUJISOFT <https://www.fsi-embedded.jp>

FUJISOFT