エッジ AI の新しいカタチ：インテル® FPGA と LeapMind による極小量子化 AI 向け MLOps

インテル プログラマブル・ソリューションズ事業本部、アジア地域事業開発ディレクター

生嶋 孝之

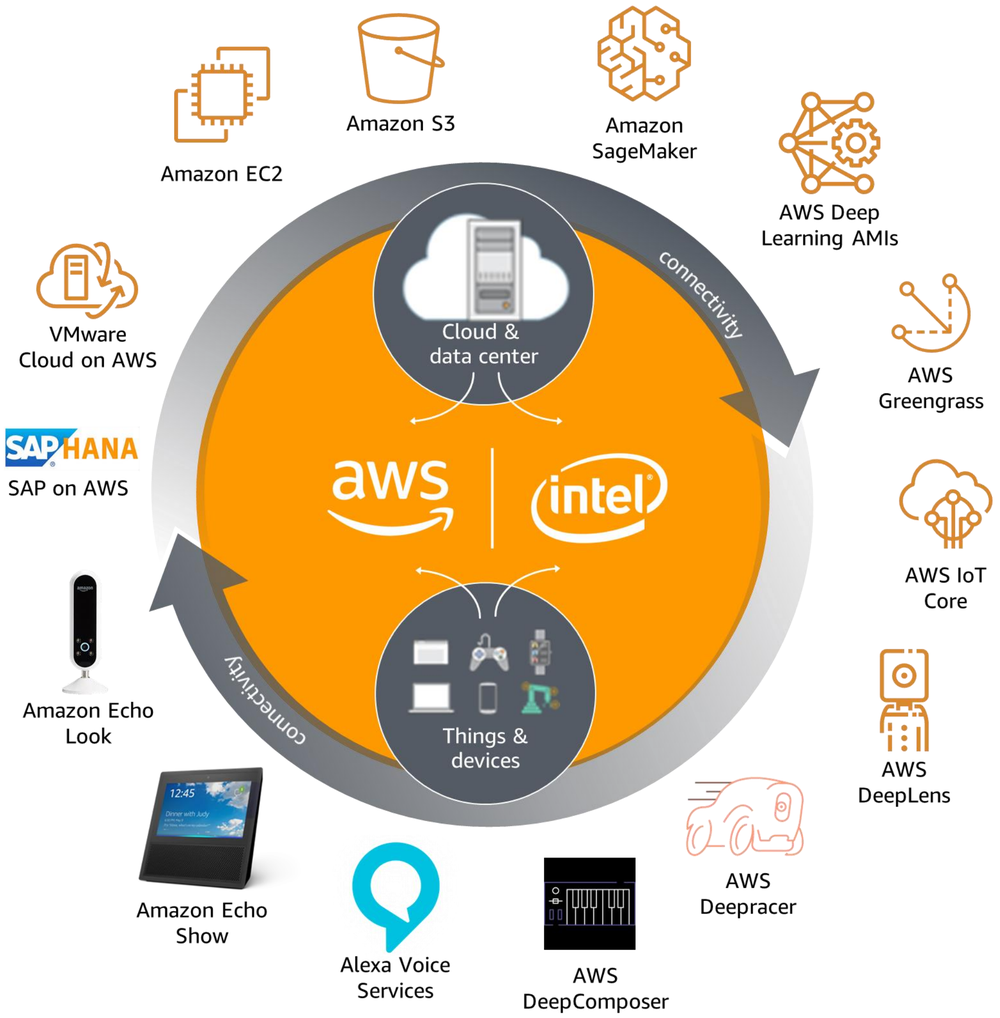
2020年10月21日

intel®

免責条項

- インテル® テクノロジーの機能と利点はシステム構成によって異なり、対応するハードウェアやソフトウェア、またはサービスの有効化が必要となる場合があります。実際の性能はシステム構成によって異なります。詳細については、ハードウェア・メーカーまたは販売店にお問い合わせいただくか、http://www.intel.co.jp/ を参照してください。
- 絶対的なセキュリティーを提供できる製品またはコンポーネントはありません。
- インテルは、本資料で参照しているサードパーティーのベンチマーク・データまたはウェブサイトについて管理や監査を行っていません。本資料で参照しているウェブサイトにアクセスし、本資料で参照しているデータが正確かどうかを確認してください。
- 本資料に掲載されている情報は、現状のまま提供されます。本資料は、明示されているか否かにかかわらず、また禁反言によるとよらずにかかわらず、いかなる知的財産権のライセンスも許諾するものではありません。インテルはいかなる責任も負うものではなく、また本資料に関する明示または黙示の保証 (特定目的への適合性、商品適格性、あらゆる特許権、著作権、その他知的財産権の非侵害性への保証を含む) に関してもいかなる責任も負いません。
- Intel、インテル、Intel ロゴ、OpenVINO、Xeon は、アメリカ合衆国および / またはその他の国における Intel Corporation またはその子会社の商標です。
- *その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。
-

インテルと AWS の連携



共通の歴史と価値

10 年以上にわたるエンジニアリングの連携

デジタル・トランスフォーメーション

顧客優先の視点の共有

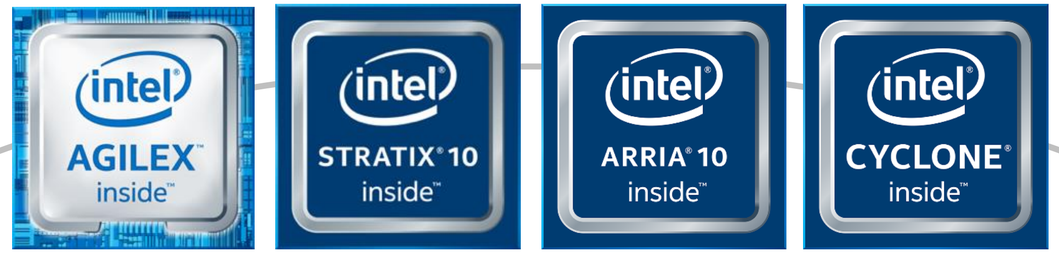
高パフォーマンス + 低コスト

世界規模のサプライチェーン

あらゆる分野での導入が進む分析と AI

2020年の戦略的テクノロジーのトレンド上位 10 すべてに関与

AI に FPGA が採用される理由



組込み、エッジ、データセンターの
幅広いアプリケーションで
リアルタイムの AI を実現

進化する AI ワークロードを高速化するための開発期間

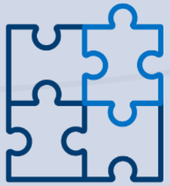
- 精度
- レイテンシー
- スパース化、量子化
- 敵対的生成ネットワーク
- 強化学習
- ニューロモーフィック・コンピューティング
- …

メモリー要件の厳しい低レイテンシーのワークロード

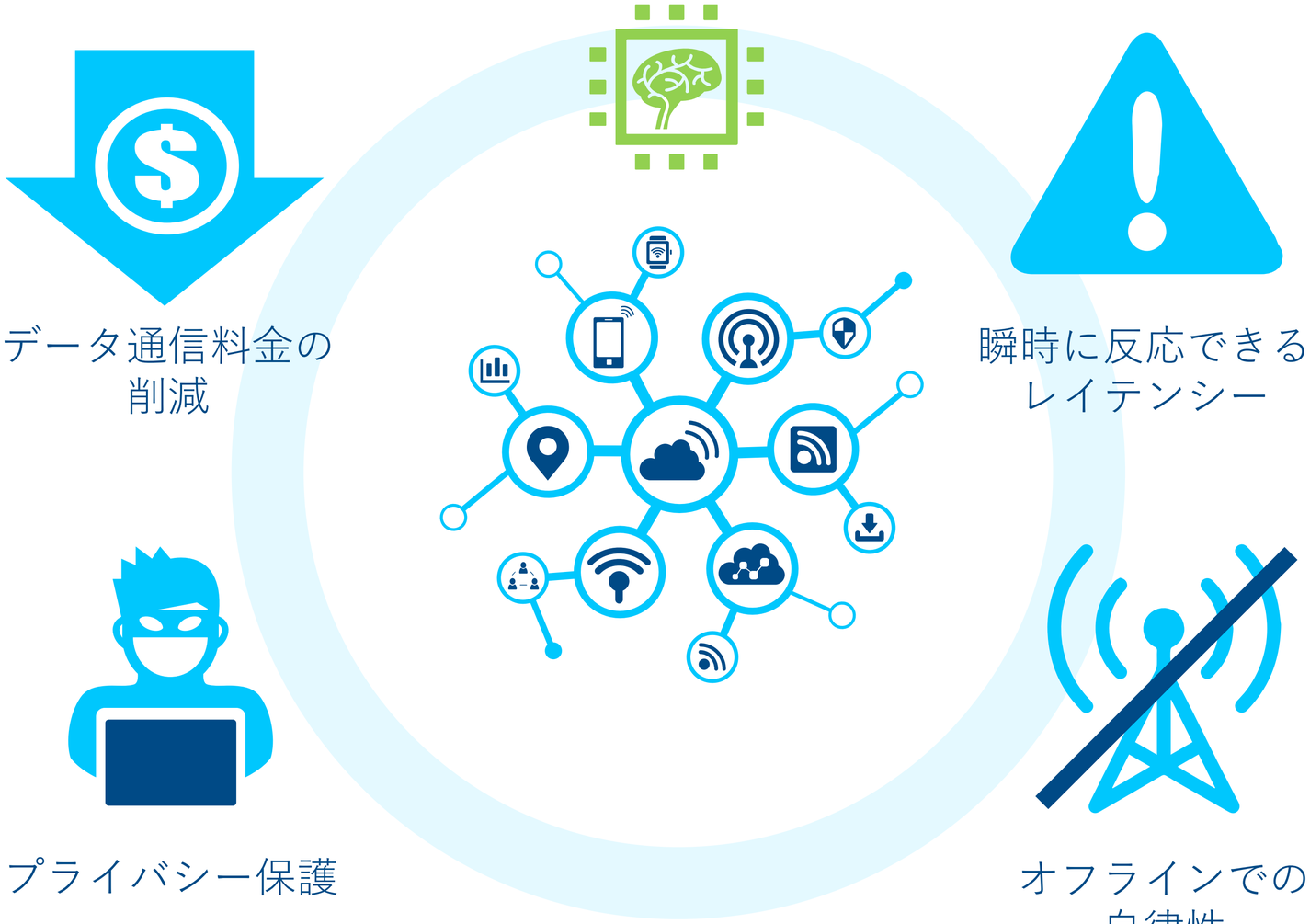
- RNN
- LSTM
- 音声
- …

AI + α の柔軟な機能をシステムレベルで実現

- AI + I/O
- AI + ネットワーキング
- AI + セキュリティー
- AI + 前後の信号処理
- …



Edge Deep Learning の必要性



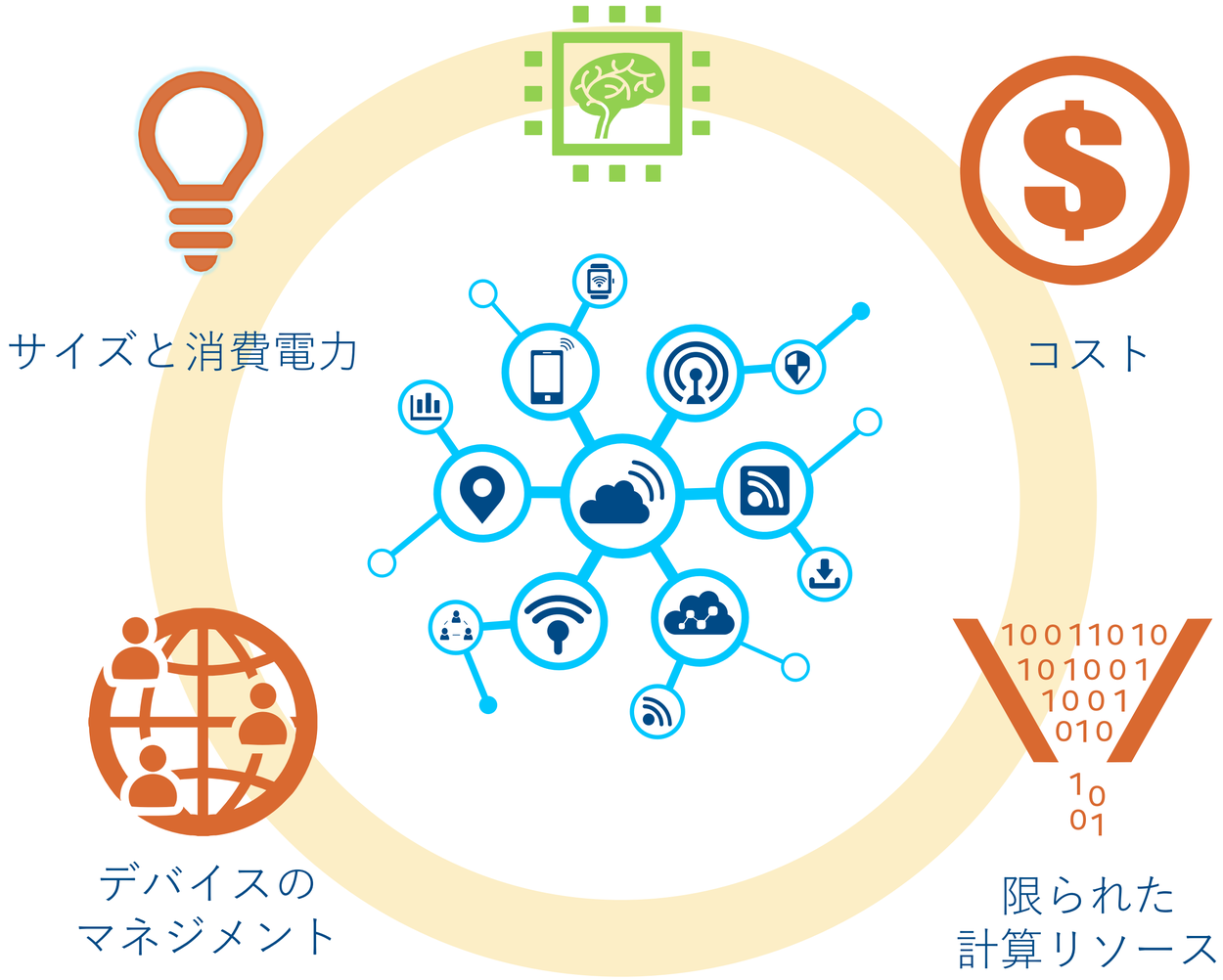
データ通信料金の
削減

瞬時に反応できる
レイテンシー

プライバシー保護

オフラインでの
自律性

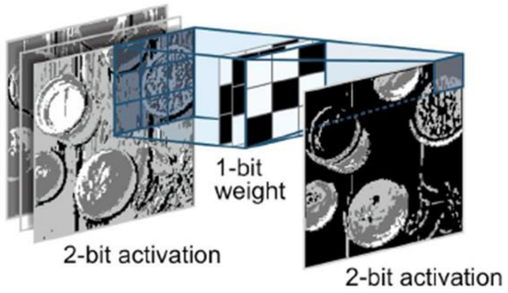
Edge Deep Learning の課題

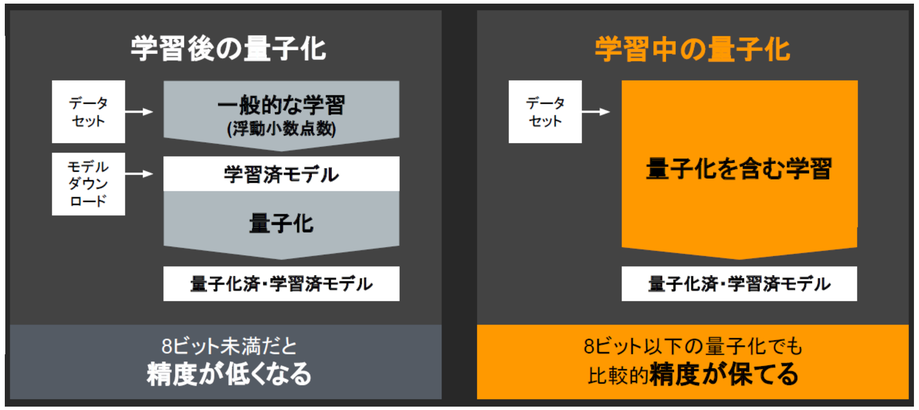


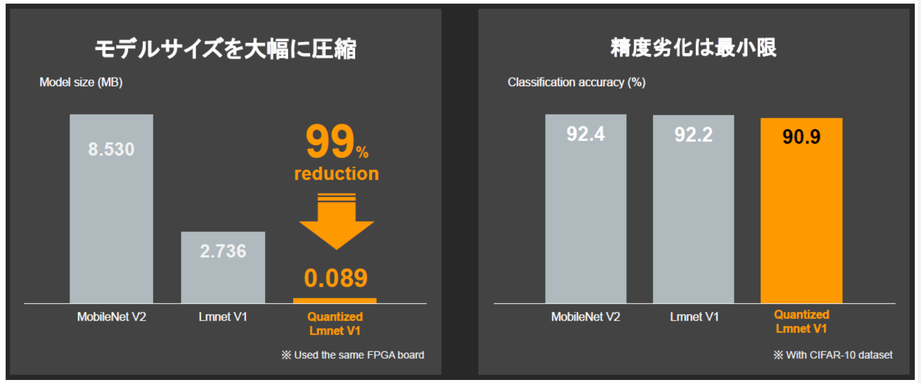
LeapMind の「極小量子化技術」

- 1 ビットまたは 2 ビットまでモデルを量子化
- 量子化を含む学習で推論精度の劣化を抑止
- 超低消費電力 AI 推論アクセラレータ IP 「Efficiera*」を開発

EFFICIERA®



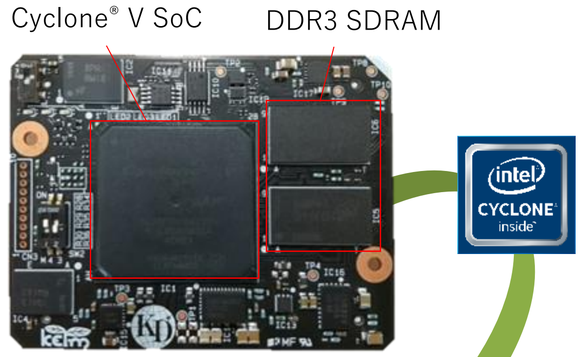




*その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。

商用可能な新しいエッジ AI デバイス

- **特長**
  - **超小型プラットフォーム**
  - **SoC FPGA によるソフトとハードの柔軟性**
  - **設計・評価済みの SoC 周辺回路**
  - **エッジ AI を容易にする開発環境**
  - **試作から中小ロット量産まで対応可**



製造：株式会社近藤電子工業

販売：株式会社マクニカ

KEIm-CVSoC SoM 基本仕様

デバイス	Cyclone® V SX SoC FPGA • CPU: Arm* Cortex-A9 x2 • FPGA: 85KLE
メモリー	2GB DDR3, QSPI Flash, EEPROM
高速 I/O	3.125Gbps x6, PCIe
外形寸法	55x43mm

KEIm-CVSoC 開発キット基本仕様

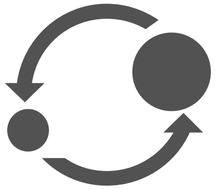
CMOS カメラ	OmniVision OV5642(5MPixel)
映像出力	DVI、ミニ HDMI
外部 I/O	Ethernet (10/100/1000BASE-T), USB2.0, M.2, UART, Micro SD, JTAG
外形寸法	110 x 80 x 50mm

*その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。

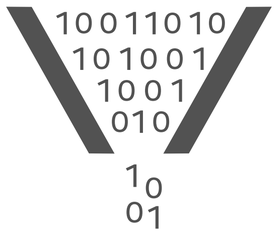
これまでのエッジ AI のカタチ

開発フローが煩雑で工数が多く、PoC と運用の「壁」により事業化が難しい

環境構築
- 学習サーバーの購入・設置
- ストレージの確保
- CPU、メモリーの増設
- バージョン管理



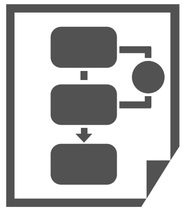
学習
- モデリング
- 学習
- 評価・検証
- 再学習



モデルの最適化、圧縮
- モデルの最適化
- モデルの圧縮
- FP32 → INT

開発キットへの実装、評価
- 推論速度
- モデルの精度
- 消費電力

PoC の壁



アプリケーション開発
- 要件定義
- 仕様作成
- コーディング
- 実装・テスト

商用デバイスの設計・選定
- 環境要件
- サイズ、電力、熱
- コスト
- 基板設計、外装設計

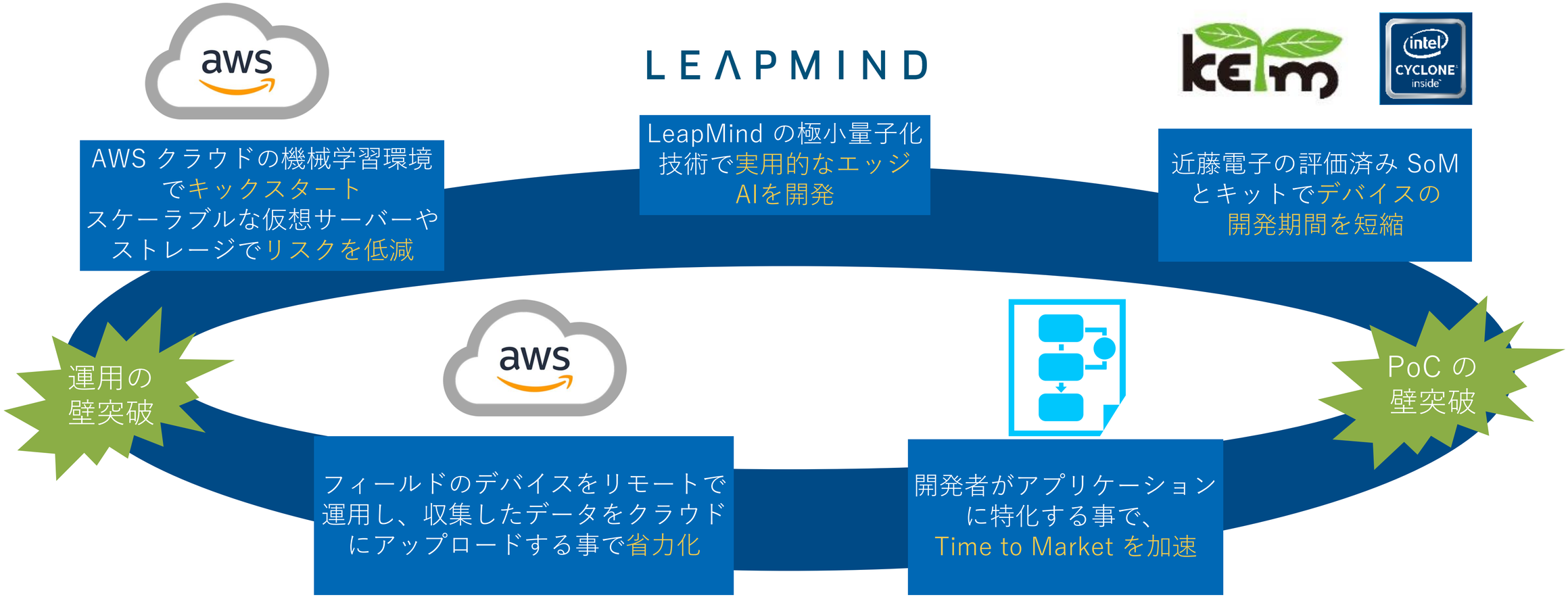
商用デバイスの評価・認定
- 機能評価
- 性能評価
- 環境試験
- ノイズ試験
- 安全認証

運用
- デバイスの調達・製造
- 販売デバイスの管理
- サービスのシステム運用
- メンテ、アップグレード
- 再学習のデータ収集

運用の壁

新しいエッジ AI のカタチ

シンプルな開発フローでスモールスタートし、リスクとコストを低減



*その他の社名、製品名などは、一般に各社の表示、商標または登録商標です。