



エッジで低消費電力の人工知能(AI)実装を加速

ラティスセミコンダクターホワイトペーパー

2018年5月

スマートファクトリー、シティ、ホームそして、モバイルの導入によって、エッジでより高いインテリジェンスが必要なシステムアーキテクチャ・新しいアプリケーションに変化が促されています。AIに基づく新世代のコンピューティングアプリケーションの要件を満たすため、人工知能(AI)/機械学習シリコンソリューションは、必要不可欠です。

エッジでコンピューティングソリューションを構築する設計者は、激しい競争市場の中で、性能を下げることなく柔軟で低消費電力、小型ファクタ、低コストという新しい要件を満たすため、絶えず挑戦を続けています。低消費電力オペレーション向けに最適化された低密度FPGAを使用したIoTデータソースに近い低消費電力推論を統合したシステムは、市場投入までの時間を短縮しながら、ネットワークエッジに必要な厳しい性能と電力制約を満たすことができます。

包括的な開発エコシステムである新しいLattice sensAI™スタックは、エッジに最適化された柔軟性のある推論ソリューションの構築を可能にします。様々なIP、ツール、リファレンスデザイン、設計専門の開発者はエコシステムを利用し、革新的なソリューションで市場投入を迅速化します。

ラティスセミコンダクター
111 5th Ave., Suite 700
Portland, Oregon 97204 USA
Telephone: (503) 268-8000
www.latticesemi.com

目次

目次	2
増加するエッジインテリジェンスにおけるアーキテクトの変化と需要	3
エッジコンピューティングの要件	5
FPGAを基にしたエッジにおける機械学習推論	6
Lattice sensAI™ スタックの紹介	7
カスタム設計サービス	10
まとめ	11

増加するエッジのインテリジェンスにおけるアーキテクトの変化・需要

初代のコンピューターが発明されて以来、理想的なシステムアーキテクチャを確定するのは簡単ではありません。コンピューティングの歴史を探ると、システム設計者は、計算リソースをユーザーから遠くに配置した集中構成と、処理リソースをユーザーの近くに配置する分散アーキテクチャの間を行ったり来たりしていることがわかります。例を挙げると、1970～80年台に普及したサーバーベースの方法は、計算能力とストレージ機能をプールするため、高度な集中手法を取り入れました。しかし、これも低コストなPCの登場とインターネットの普及により、1980年代と1990年代の間で急速に衰退しました。新しいアーキテクチャモデルでは、計算タスクがますます個々のPCに委ねられます。

PCを中心に構築された高度な分散型アプローチはすっかり普及したように見えてきましたが、今度はスマートフォン、タブレット、そしてラップトップなどの持ち運び可能な機器の増加によって、計算ハードウェアとストレージリソースが重荷となりました。システム設計者は次第に、タスクを実質上無制限である計算・ストレージリソースと高い信頼性を保ち、かつ低コストで利用できるクラウドに移行し始めました。

企業は、クラウドを使用して資本コストを削減し、ITインフラストラクチャ関連の運用、保守コストをより効率的に管理するようになっていきます。企業が機械学習技術を採用し、より高度な人工知能を導入するにつれて、クラウドはますます中心的な役割を果たすようになります。次世代のスマートファクトリー、スマートシティ、スマートホームは、マシンビジョンシステムを効率的に管理し、交通パターンを調整し、消費電力を最小限に抑えるためにクラウドを必要とします。

全てのアプリケーションがクラウドから実行するわけではありません。専門家は、集中管理システムから分散システムのアーキテクチャに移行したことを表す最初の兆候だと述べています。次の移行が来るかどうかに関わらず、ある1つのことは明らかです。より低いレイテンシ、プライバシーへの配慮、通信帯域幅の制限はエッジインテリジェンスの需要を駆り立てています。設計者はエッジアプリケーションに高度なインテリジェンスを追加するため、変化する環境条件にすばやく対応するシステムが必要です。例えば、ロボットカーがスマートシティに入ったとき、衝突を避ける方法をクラウドに尋ねる時間はありません。すぐに反応し、自分で判断をしないといけません。また、自宅でAIに基づいたセキュリティカメラが動きを感知するときは、デバイス上のリソースを使って侵入者がいるかどうかを判断し、警察を呼ぶ必要があります。

新しいアプリケーションはクラウドよりもIoTセンサデータのソース近くに配置されたAI/機械学習ベースのコンピューティングソリューション必要とします。どれくらいの長さが必要でしょうか？巨大なものが必要だと思う人もいます。Garner社のアナリストは2022年までに企業で生成されたデータの最大50%が従来の集中データセンターの外もしくはクラウドで処理されると予想しています。（図1参照）

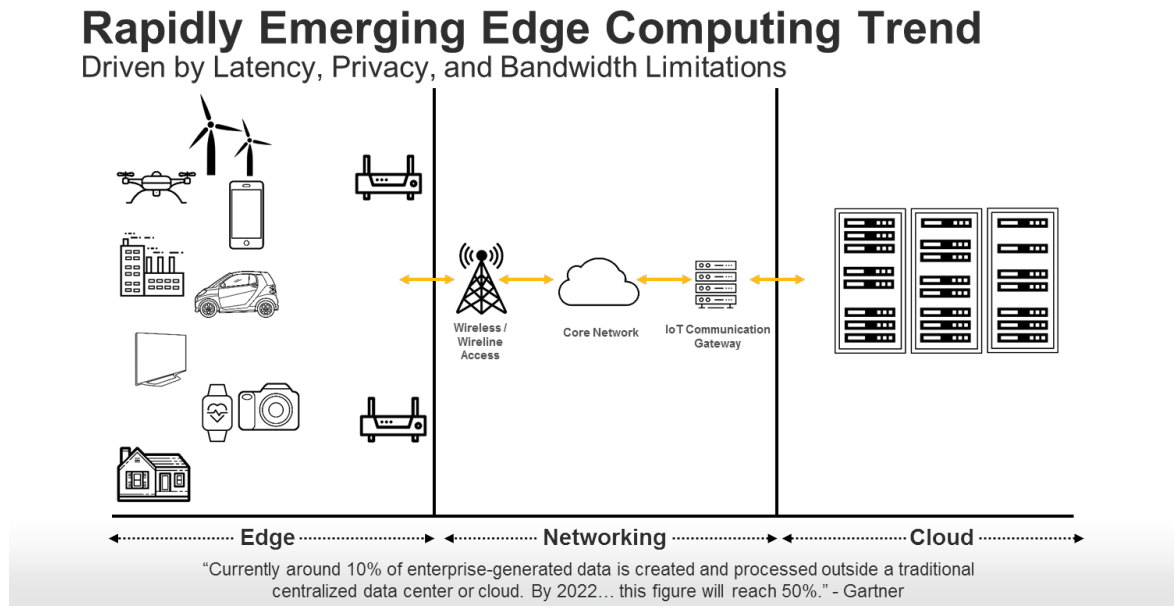


図1. レイテンシ・プライバシーそしてネットワーク帯域幅制限によって急速に変化するエッジコンピューティングのトレンド

エッジコンピューティングの要件

エッジのコンピューティングソリューションを構築する設計者が直面する最も困難な問題の1つは、柔軟性と低消費電力、小型ファクタ、低コスト要件などの固有な組み合わせに対応することです。（図2参照）

Edge Computing Requirements

Flexible, Low Power, Small Form Factor, Production Priced Silicon

	MOBILE	SMART HOME	SMART CITY	SMART FACTORY	SMART CAR
Applications	Face detection	Speech recognition	License plate recognition	Fault detection	Collision avoidance
End Devices	Smartphones Wearables	Smart speakers AR/VR	Surveillance cameras Drones	Machine vision Robotics	Automation Levels 3,4,5
Power	<10 mWs	<500 mWs	<1W	<3W	>10W
Price	<\$1	<\$3	<\$5	<\$10	>\$20
Performance	~1 BOPS	~10 BOPS	~100 BOPS	~1,000 BOPS	>10,000 BOPS

図2. AIを基にした新世代のエッジコンピューティングアプリケーションは
それぞれの独自の要件が必要とされる

性能を下げることなく、低消費電力で小型、低コストのエッジソリューションをどのように構築できるでしょうか？何よりもまず、最大限の設計柔軟性をもたらすシリコンが必要となります。急速に進歩するニューラルネットワークのアーキテクチャとアルゴリズムを利用でき、かつ、幅広いI/Oインターフェースに対応したシリコンが必要になります。そして、電力を節約するため、カスタム可能な量子化によって精度をトレードオフできるソリューションも必須です。

エッジ製品ではフットプリントの制約があるので、設計者はサイズや温度管理の制限を守りながら、小型で高性能なAIデバイスが構築できるシリコンが必要です。また、コストも重要な要素で、どのソリューションも他の大容量エッジソリューションに匹敵しなければなりません。最後に、エッジでさえ市場投入までの時間も重要です。一番最初にソリューションを市場にもたらした人がエッジの恩恵を享受することができます。したがって、あらゆる視点ソリューションは、デモ、リファレンスデザイン、設計サービスのいずれであっても、ソリューションをカスタマイズし、開発サイクルを短縮するために必要なリソースにアクセスできる必要があります。

FPGAを基にしたエッジにおける機械学習推論

FPGAはエッジでどのような役割を果たしているでしょうか？機械学習は通常、2種類の計算作業を求められます。学習中のシステムは既存のデータから新しい機能を取り入れます。例えば、顔検出機能は何万枚もの画像を収集し、分析することで人間の顔を認識することを学習します。この初期の学習段階は高度な数値計算です。開発者は通常、そのような大容量のデータを処理するため、データセンターで高度なハードウェアを使用します。

機械学習、推論における次の段階はパターンを確定し、タスクを実行することで新しいデータをシステムの機能に応用することです。例えば、顔検出機能は、現地で働くにつれて人間の顔を正しく認識するよう改良され続けると以前話しました。この段階では、システムが動作するにつれて学習し、時間の経過と共にインテリジェンスを高めていきます。エッジ機器の起動には多くの制約があるため、設計者はクラウドで推論を実行する余裕がありません。その代わりにエッジ上でデータソースの近くで計算タスクを実行することでシステムのインテリジェンスを拡充します。

しかし、クラウドで利用できる膨大な計算リソースをどのように置き換え、エッジの推論を実行できるのでしょうか？唯一の方法は、FPGAの並列処理能力を利用してニューラルネットワークの性能を向上させることです。低消費電力で起動するために最適化された低密度FPGAを使用することで、ネットワークエッジに必要な厳しい性能と電力制限を満たすことができます。ラティスのECP5とiCE40 UltraPlus FPGAはまさにこのニーズを満たしています。ECP5 FPGAは1W以下でニューラルネットワークを加速し、iCE40 UltraPlus FPGAはmW範囲のニューラルネットワークを加速し、効率的にAIベースのエッジコンピューティングアプリケーションを構築することができます。（図3参照）

Low Power Machine Learning Inferencing

From Under 1mW to 1W

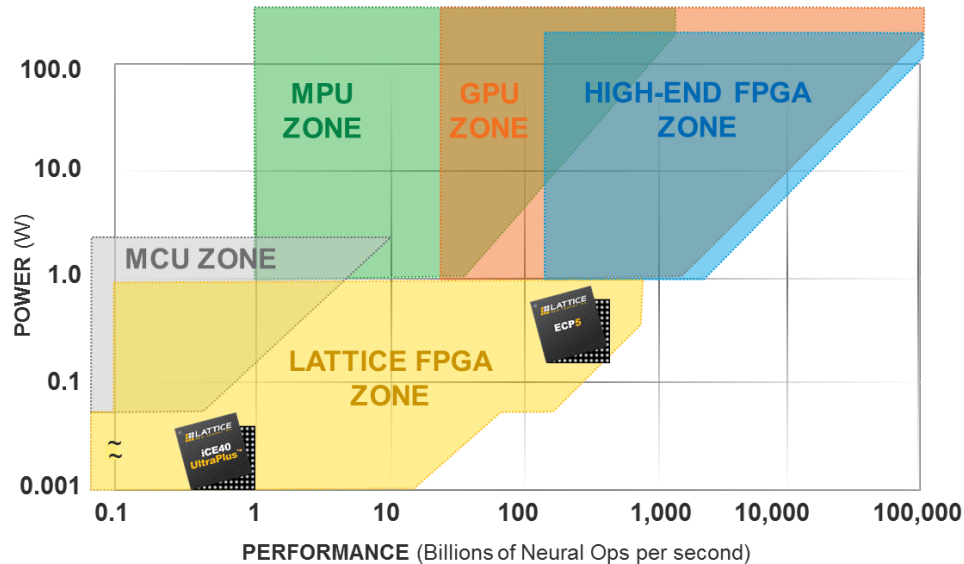


図3. 1mWから1W以下から、ラティスのFPGAを基にした低消費電力の機械学習推論を実現

Lattice sensAI™ スタックの紹介

効率的なソリューションを構築し迅速に市場投入するためには、計算ハードウェアへのアクセスに加えて様々なIP、ツール、リファレンスデザイン、そして設計の専門知識が必要です。

増加し続ける設計者の課題に取り組むため、ラティスはiCE40 UltraPlusとECP5 FPGAファミリを基にした新しい包括的な開発エコシステムを提供しています。スマートホーム、スマートシティ、スマートファクトリー、スマートカー、モバイルアプリケーション向けのAIエッジソリューションを迅速に構築できるように設計されたLattice sensAIスタックは、最適な柔軟な推論ソリューションです。

図4のように、モジュラーハードウェアプラットフォーム、ニューラルネットワークIPコア、ソフトウェアツール、リファレンスデザイン、そしてエコシステムの提携企業によるカスタム設計サービスを組み合わせたLattice sensAIは、1mW～1W以下の低消費電力、パッケージサイズが5.5mm²、そして大量販売時価格が米1から10ドルという安価で柔軟な推論ソリューションの構築を簡素化します。

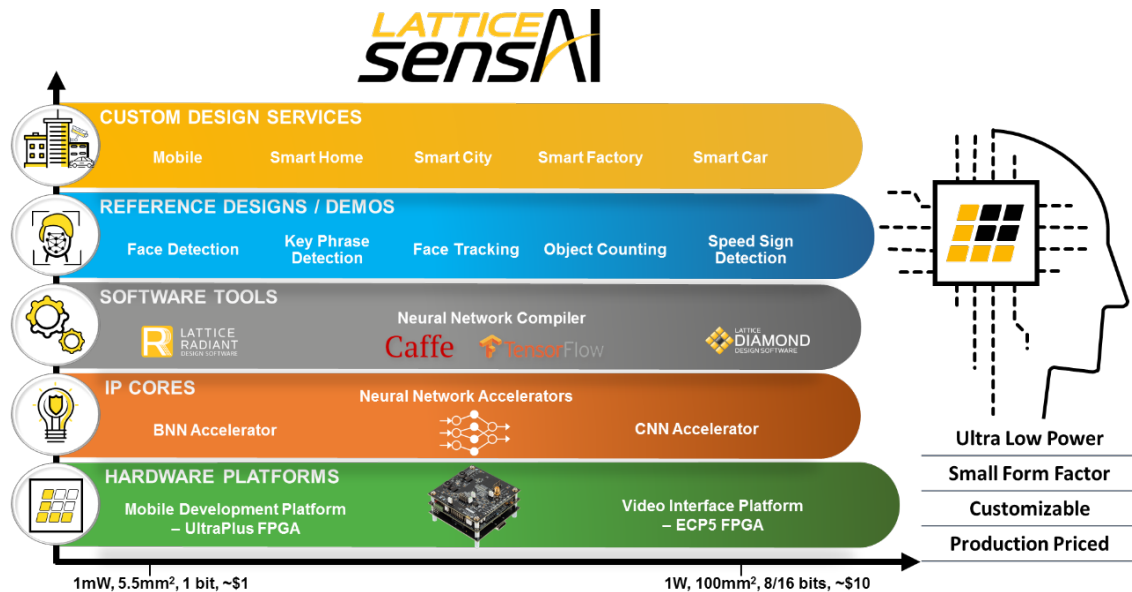


図4: ラティスのsensAIスタックはエッジのコンピューティングソリューション開発に向けた強固な土台を提供します

図4で示すように、Lattice sensAIスタックはラティスのハードウェアキットで始まります。この機能を実行するラティスは新しいモジュラーハードウェアプラットフォームは、幅広い性能と電力要件を満たす機械学習設計の試作を加速化します。わずか数mWの電力消費でのAI設計によってラティスは低消費電力のiCE40 UltraPlus FPGAに基づいたモバイル開発プラットフォーム(MDP)を提供しています。このMDPは画像センサ、マイクロフォン、コンパス/圧力/ジャイロ、その他を含む様々なオンボードセンサーを搭載しています。多くの電力を必要とし、通常1W以下で起動するアプリケーション向けにラティスにはECP5 FPGAファミリをベースとしたモジュラービデオインターフェースプラットフォーム(VIP)があります。VIPはMIPI CSI-2、組込みDisplayPort (eDP)、HDMI、GigE VisionおよびUSB3などの幅広いインターフェースとの接続性をもたらします。ラティスが提供している最初のハードウェアプラットフォームの1つは、受賞歴を持つ組込みビジョン開発キットです。このモジュラープラットフォームは、ESP5プロセッサボードとHDMI出力ボードを組み合わせたCrossLink入力ボードです。最新のeDPとUSB3 GigE I/Oボードの導入で、他のアプリケーション対応の出力ボードへと簡単にスワップアウトすることができます。

ハードウェア層に加えて、FPGAで簡単にインスタンス生成が可能な新しいニューラルネットワークアクセラレータIPコアも提供します。このソフトIPラインナップには、iCE40 UltraPlus FPGA 向けに最適化された2値化ニューラルネットワーク(BNN)アクセラレータも含まれます。このIPで、iCE40 UltraPlus FPGAを使用して BNNにディープラーニングアプリケーションを実装できます。ラティスはまた、畳み込みニューラルネットワーク(CNN)アクセラレータコアも提供します。完全パラメータ設定可能なコアはラティスのECP5 FPGAに実装でき、ラティスのECP5 FPGAに実装することができ、可変量子化対応なので、消費電力を節約するために精度のトレードオフができます。

図5で示すように、Lattice sensAIスタックは使い易いツールで迅速な設計の空間探索とトレードオフを可能にします。ネットワークはCaffeやTensorFlowなどの業界規格のフレームワークで訓練されます。ニューラルネットワークコンパイラツールは、訓練されたネットワークモデルを固定小数点表示へマッピングし、重みとアクティベーションの可変量子化に対応します。さらに、ニューラルネットワークコンパイラは以前にRTLの経験がない場合でも、ラティスのCNN / BNN アクセラレータIPコアに実装するために、様々な種類のネットワークの解析、シミュレーション、およびコンパイルを手助けします。RadiantやDiamondなどの従来のFPGA設計ソフトウェアツールは、前後処理ブロックの残りを含めたFPGA全体の設計を実装するために使用されます。

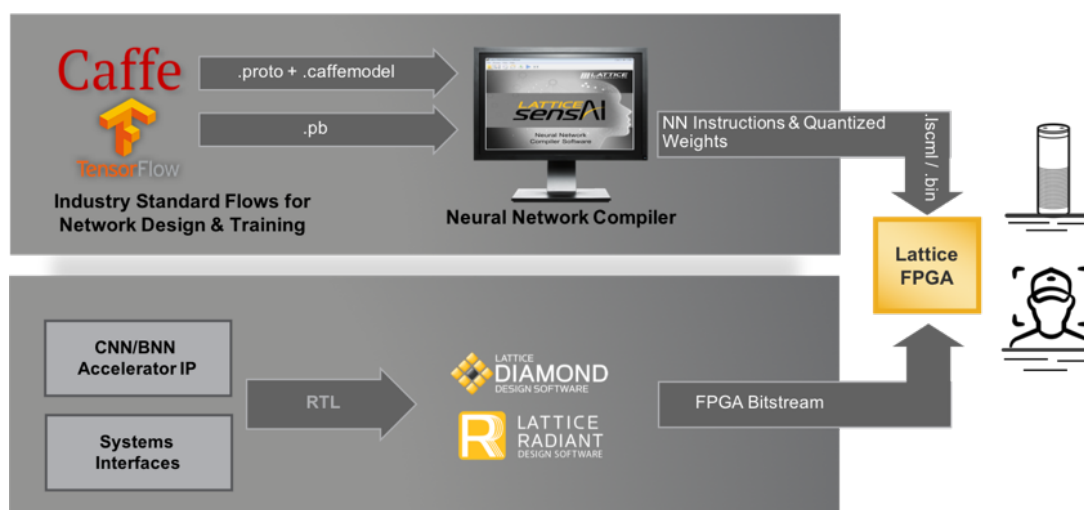


図5: 使い易いLattice sensAIツールで迅速な設計空間探索とトレードオフが可能に

一般的なAI機能の実装を簡素化するため、Lattice sensAIスタックは、スタックのハードウェアプラットフォーム、IPコア、そしてソフトウェアツールを使用する幅広いリファレンスデザインとデモを含んでいます:

低消費電力の人間顔検出 – このデモには、ニューラルネットワークモデルを使用してネットワークエッジで動作するように設計された、高速で低消費電力の、人間の顔検出機能が含まれています。2値化された重さとアクティベーションを利用して、このiCE40 UltraPlus FPGA に基づいたデモは1mW以下の電力消費で顔検出の実装を可能にします。

車載機器のアフターマーケットカメラ – このデモのターゲットは新興の車載機器アフターマーケットカメラのマーケットです。FPGA本来の並列化を利用して速度標識検出機能を実装する方法を示しています。このデモではECP5 FPGAに実装された畳み込みニューラルネットワークが交通標識を読み取って訓練されています。訓練が完了すると、カメラが標識を通過したときに速度制限を検出し、表示することが可能です。

音声コマンドをシステム実行へ変換 – このデモは音声コマンドをシステムの実行へと変換するための

設計図を提供します。サブ5mWのキープレーズ検出機能は2値化されたニューラルネットワークを使用します。このデモはデジタルマイクロフォンをラティスの推論エンジンに直接接続し、キープレーズ検出によって常時オンの聞き取りを可能にする方法を説明します。

顔追跡向けオブジェクト検出ソリューション – このデモでは、顔追跡アプリケーション向けにAIベースのオブジェクト検出ソリューションの設計について解説します。また、8つのニューラルネットワークエンジンが実装された畳み込み層を持つ畳み込みニューラルネットワーク(CNN)アクセラレーション向けのラティス ECP5-85 FPGAの使い方も説明します。このソリューションはラティスの組込みビジョン開発キットに基づいたスタンドアロン動作が可能で、電源投入後は、90 x 90 RGB入力で14fpsで起動します。ECP5の総電力消費はわずか0.85Wです。

カスタム設計サービス

開発チームはしばしばカスタムソリューションを開発するため、設計サービス提携企業の専門知識が必要となります。AI市場も例外ではありません。ニーズを満たすため、ラティスはスマートファクトリー、スマートシティ、スマートカー、スマートホーム、そしてモバイルアプリケーションなどの幅広い領域で、多くの設計サービス提携企業と関係性を築いています。一例として、ラティスの認定パートナーは VectorBlox社は、ニューラルネットワークに基づいた推論ソリューションのディベロッパーです。パートナーシップの一環として、VectorBox社とラティスは最近、顔検出アプリケーションに向けてiCE40 UltraPlus FPGA に5000 LUT以下のニューラルネットワークを実装しました。独自のアクセラレータを備えたオープンソースのRISC Vソフトプロセッサを使用することで、消費電力を大幅に削減し、応答時間を短縮できます。

ラティスのFPGAで更に迅速な推論ソリューションを実装するため、ニューラルネットワークの設計と訓練の知識を持つ設計サービスの専門知識が必要な場合もあります。そしてこの知識は、従来のRTL設計に加え、CaffeやTensorなどのフレームワーク経験もセットで必要であることが多くあります。そのような専門知識の必要性に応えるため、ラティスのLattice sensAIには設計サービスプログラムがあります。このプログラムは、ラティスFPGAでディープラーニング実装の設計を加速するために必要な設計サービス企業へのアクセスを提供し、これらの企業はアプリケーション向けRTLを開発する機能に加えて、ネットワークの開発と訓練の専門知識をもたらします。

このプログラムに協賛しているサードパーティ企業は、ラティスと共にニューラルネットワークの開発やハードウェアのトレーニングと実装スキルのデモを行っています。

まとめ

エッジコンピューティングの革命は、AIに基づいたシステムの専門知識を持つ開発者が必要です。ユーザーがより高度なインテリジェンスを求めるにつれて、IoT DATのソースに近くで必要な低消費電力の推論を統合するシステムの需要が増加します。Lattice sensAIスタックは柔軟で超低消費電力、小型、そしてエッジ向けに最適化された生産価格のソリューションを提供します。幅広いハードウェアプラットフォーム、ソフトIP、設計ツール、リファレンスデザイン、サードパーティの専門知識によるサポートが可能なラティスの超低消費電力FPGAは、成功まで最速の方法を提供します。