ソリューション概要

FPGA
ストリーミング・データのリアルタイム分析

Megh Computing プラットフォームがインテル® Arria® 10 FPGA でリアルタイム・ストリーミング分析を加速

マシンラーニングとディープラーニングにより、ストリーミング・データからより多くの価値をシームレスに引き出す

「データセンターにおけるコンピューティングに第 3 の波が押し寄せています。FPGA をベースとした専用ハードウェア・アクセラレーターが、オンプレミス、さらにクラウドで利用可能になり、分析や他のアプリケーションに向けたリアルタイム・ストリーミング、マシンラーニング、ディープラーニングのニーズをサポートします」

— **Megh Computing、CEO Prabhat K. Gupta 氏**

概 要

リアルタイム・ストリーム処理の需要は、センサーやウェブ、その他のソースからのストリーミング・データの急増とともに急速に高まっています。企業はデータの移動に合わせてデータを処理することでビジネス価値を生み出したいと考えており、ストリーミング・データから情報を抽出するために、マシンラーニング / ディープラーニング・ライブラリーを備えた Spark* Streaming フレームワークの採用を急いでいます。今日使用されているソフトウェアのみのソリューションでは、低レイテンシーかつ高スループットというパフォーマンス要件を満たすことはできません。低いレイテンシーと高い計算能力を実現するためには、ハードウェアとソフトウェアが統合された新たなプラットフォームとツールが必要です。

Megh Computing は、Spark* Streaming とその他のフレームワークを使用して、リアルタイム分析を加速する革新的なプラットフォームを提供します。このソリューションは、インテル® Arria® 10 FPGA ベースのハードウェア / ソフトウェア・プラットフォーム、インテル® FPGA プログラマブル・アクセラレーション・カード (インテル® FPGA PAC)、インテル® アクセラレーション・スタック (インテル® Xeon® CPU & FPGA 対応) 上で動作し、データ・ストリーミング処理によって価値を引き出すためにマシンラーニングとディープラーニングのアルゴリズムを使用するアプリケーションのシームレスなアクセラレーションを実現します。Megh Computing とインテルのソリューションにより、低レイテンシー、低 TCO でパフォーマンスを向上させて、新しいユースケースとワークロードをサポートできるようになります。

課 題

ストリーミング分析市場は、2022 年までに 159 億米ドルの市場規模に達成することが予想され、33.1% という年間成長率で拡大しています。[1] こうした成長の原動力は、IoT センサーとイベント、ソーシャルメディア、使用状況や行動に関するデータを記録するウェブアプリから、幅広い業種にわたるトランザクションおよび運用データまで、多彩なソースから流れてくる構造化および非構造化データの急増によるものです。従来型のビジネス・インテリジェンス (BI) に依存していたビジネスが、マシンラーニングやディープラーニングによる高度な分析へと移行するにつれて、新たなレベルのパフォーマンス、低レイテンシー、高速処理を必要とする大量のストリーミング・データにより、コンピューティングとインフラストラクチャーに対する要求の高まりが生じています。

今日のリアルタイム分析ソリューションは、データをストリーミングするための Apache Kafka*、Apache Flink*、または同様のフレームワークを使用するソフトウェア・プラットフォームと、データを処理するための分散フレームワークとしての Spark* Streaming フレームワークに基づいています。典型的な動作ステージには、(1) データを取り込んで処理のための準備をする抽出、変換、ロード (ETL) と、(2) データから情報を推論するためのマシンラーニング / ディープラーニング・ライブラリーが含まれます。

これらのオープンソースまたは独自のソフトウェアのみのソリューションでは、リアルタイム分析が要求する計算処理の増加に応えたり、リアルタイム分析のユースケースをサポートするために必要な低レイテンシーを実現することはできません。さらに、増加し続けるデータを処理するためにノード数を増やしても、パフォーマンスが直線的に拡大しないことに多くの組織は気づいていました。

高速処理におけるレイテンシーの短縮は、多くの場合、CPU だけでは実現できません。FPGA をヘテロジニアスな CPU および FPGA プラットフォームの一部として使用して、再構成可能なアプリケーション固有のアクセラレーターを実装することで、低レイテンシーでより高いパフォーマン

スを実現できます。高速化されたパフォーマンス、柔軟性、低消費電力、プログラマビリティーなどの利点を提供する FPGA は、GPU や他のアクセラレーターと比較して、こうした種類のアプリケーションに適しています。ただし、その代償として、FPGA のプログラミングと管理が複雑化し、データセンター・オーケストレーション・ソフトウェアへの統合が必要になることがあります。

マシンラーニングやディープラーニングで Spark* Streaming を使用する現在のソリューションは、パフォーマンスの要件を満たせないことがある

ソリューション

Megh Computing のリアルタイム分析プラットフォームは、音声、ビデオ、テキストなどさまざまな形式でのデータ・ストリーミングを可能にするために、リアルタイム分析の主要課題に対処することを目的として設計されています。

Megh Computing のソリューションは、標準的な大容量インテル® Xeon® プロセッサー搭載サーバーに、インテル® プログラマブル・アクセラレーション・カード（インテル® Arria® 10 GX FPGA 搭載版）とインテル® アクセラレーション・スタック（インテル® Xeon® CPU & FPGA 対応）を組み合わせて使用するヘテロジニアスな CPU + FPGA プラットフォームに基づいています。Megh Computing のソリューションには、主に 3 つの属性があります。

- FPGA への完全なリアルタイム分析パイプラインの実装。

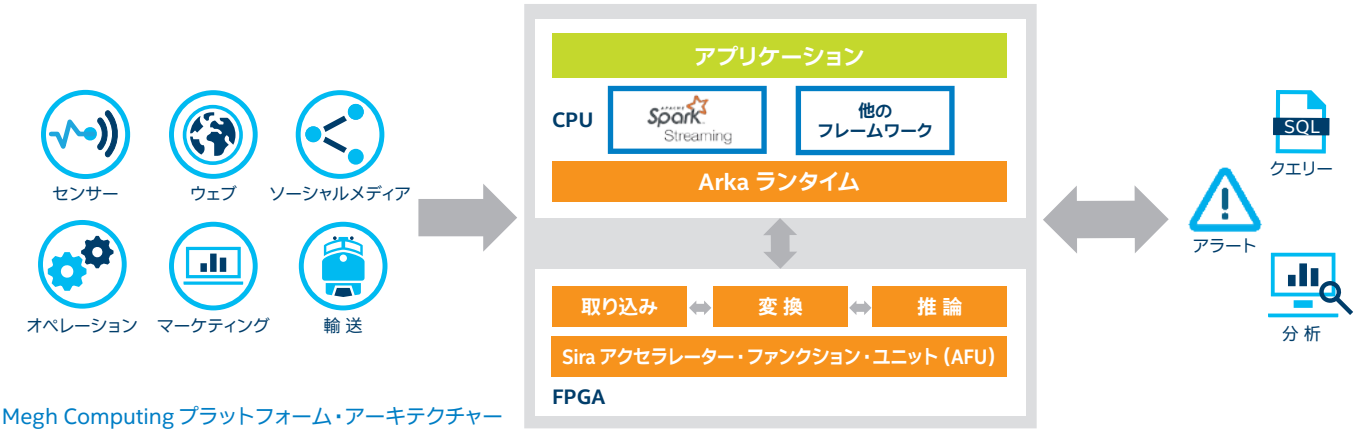
  取り込み：さまざまなフォーマットのストリーミング・データ用

  変換：データ処理のための変換用

  推論：マシンラーニングとディープラーニング推論用

- コード変更が不要な、マシンラーニング・ライブラリーとディープラーニング・ライブラリーを使用した標準およびカスタムのデータ分析フレームワークのサポート。例えば、推論には Spark* Streaming および BigDL ライブラリーを使用します。
- Megh Computing の Arka ランタイム・ソフトウェアおよび Sira FPGA ハードウェア・ライブラリー用の SDK を使用した、さまざまなユースケースに合わせたカスタマイズ。

Megh Computing とインテルにより、ASIC と同等の使いやすさを得られると同時に、FPGA のプログラマビリティーと再構成（リコンフィグレーション）の利点も得られます。インテル® プラットフォーム上の Megh Computing の統合ソフトウェアおよびハードウェア・ソリューションは、あらかじめ最適化されたライブラリーを高位 API（実際のライブラリー呼び出しを行い、それらを FPGA に再マッピングする）を介して公開することによって、FPGA の複雑さを抽象化しています。インテル® Arria® 10 FPGA はインライン・ストリーミング・データを直接処理し、他のデータセンターの活動に CPU を利用できるようにします。



Megh Computing プラットフォーム・アーキテクチャー

Megh Computing のリアルタイム分析プラットフォームの利点	
パフォーマンスと効率性を向上	• 総運用コストの削減額を **2 倍以上に拡大**[2] • 性能効率を **5 倍から 10 倍に向上**[2] • レイテンシーを **10 分の 1 以下に低減**[2]
価値実現までの時間を短縮	• 大容量ハードウェア・プラットフォームを介した新しいリアルタイム分析ユースケースの展開をサポート • オープンソース・ソフトウェア・フレームワークの柔軟性と Megh SDK を使用したカスタマイズ機能の獲得
展開を簡素化	• パブリック・クラウド、プライベート・クラウド、またはエッジクラウドにコンテナまたは仮想マシン（VM）を使用して FPGA Accelerator as a Service を展開し、クラウド・オーケストレーション・レイヤーと統合

データ主導の経済社会のための FPGA

柔軟でプログラマブルな FPGA により、ビジネスおよび産業分野において、マシンラーニングおよびディープラーニング・アルゴリズムを使用したリアルタイム・ストリーミング分析用のスケールアウト・アクセラレーターに対するニーズの高まりに応えることが可能になります。

FPGA は、以下を実現する再構成可能なハードウェア・ゲートを提供します。

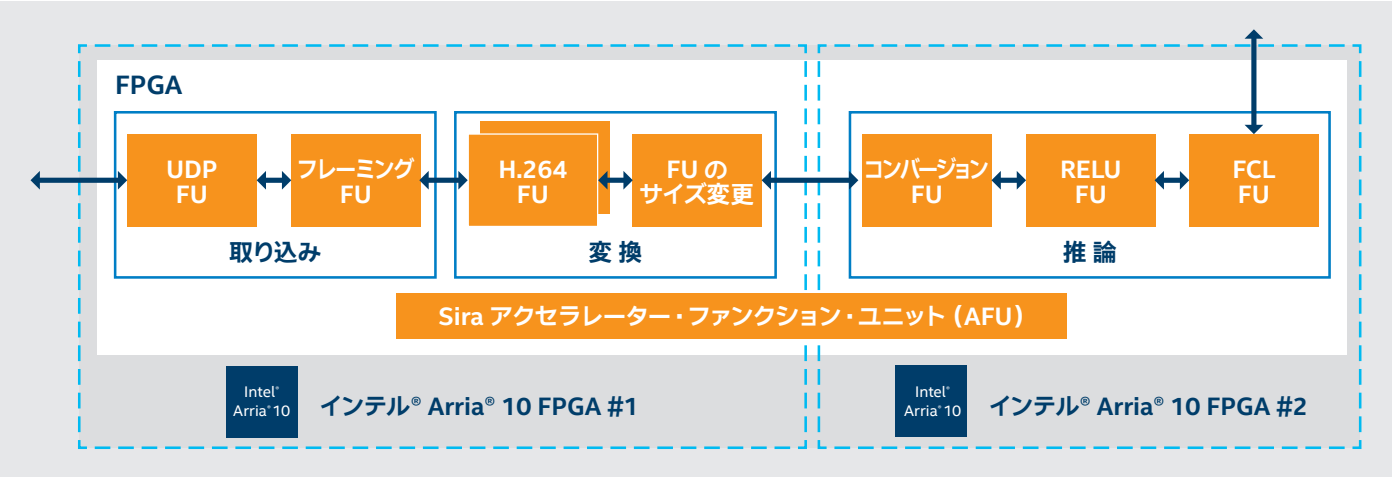
- 直接の I/O 接続と低遅延のカスタム・ハードウェア・アクセラレーターを設計する
- 単一のアプリケーションで性能効率を向上させる
- デバイスを別のアプリケーション用の新しいアクセラレーターとして即座に再設定する

Megh Computing ソリューションはこれらの機能を利用して、完全なリアルタイム分析パイプラインを FPGA に実装します。例えば、Megh Computing は、小売業での不正防止やその他のユースケースをターゲットにしたビデオ分析パイプラインを作成しました。その完全なパイプラインは、取り込み、変換、推論段階を実装するためのさまざまな機能ユニット (FU) で構成されています。これらの機能は、2 つのインテル® Arria® 10 FPGA にまたがってマッピングされています。

- 取り込み段階と変換段階は、最初の FPGA に実装
- 推論段階は、2 番目の FPGA に実装

データは最初の FPGA から 2 番目の FPGA に直接転送されます。これらすべての機能は、単一のインテル® Stratix® 10 FPGA にもマッピングできます。あるいは、このソリューションは、2 つのインテル® Stratix® 10 FPGA を使用して、より大容量のソリューションを実現できます。

近い将来、Megh Computing はテキスト分析のための同様のパイプラインを構築し、その後、音声やその他のパイプラインも構築する予定です。これらすべてのパイプラインは Megh Computing が提供するアクセラレーター・ファンクション・ユニット (AFU) のライブラリーを利用するか、サードパーティーが開発した FU を統合します。



インテル® Arria® 10 FPGA への Megh Computing ビデオ分析パイプライン・マッピング

ユースケースの例

トランザクション処理による大容量のストリーミング・データは、最先端の企業にとってはいうまでもなく、金融、小売、電気通信、製造などの分野の企業にとっても、その日々のオペレーションや意思決定をするための中核となる存在です。Megh Computing とインテルのソリューションは、エッジであろうとクラウドであろうと、ストリーミング・データの送信、処理、分析に低レイテンシーと高速パフォーマンスを必要とするユースケースに最適です。

例えば、金融取引分析では 100 ミリ秒未満のリアルタイム要件を満たす必要があります。株式市場の変動はリアルタイムで可視化されなければならず、ポートフォリオは正確かつ最新のリスク評価に基づいてバランスの最調整が行われます。

金融と小売、いずれの業種にとっても不正防止は非常に重要であり、ミリ秒単位での取引検証が必要です。e コマースや金融分野の各社は、マシン駆動型アルゴリズムによる監視を利用することで、データのバリエーションを発見し、それが発生した瞬間に不正を検出しようとしています。小売業では、Point of Sale (POS) 端末は SKU の不正確なスキャンまたはラベリングをリアルタイムで検出する必要もあります。

IoT は、リアルタイムの物体 / 画像 / 顔認識、拡張現実や仮想現実、コンピューター・ビジョン・アプリケーションなど、質の高いエンドユーザー体験を提供するためにミリ秒単位の応答時間を必要とする、最先端の新しいユースケースを数多く生み出し続けています。

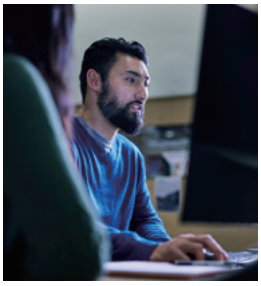
従来の多くの運用プロセスでは単に素早い応答時間が期待される一方で、リアルタイム・データに基づいてダッシュボードを更新する幅広いアプリケーションでは 1 秒未満という厳密な応答時間が要求されます。

小 売

- SKU のスキャンミスによる不正の防止
- 品質管理のための生鮮食品の画像分類
- 商品の置き忘れ回避による在庫管理の改善

金 融

- クレジットカード取引における不正防止
- 取引業務のリアルタイム・モニタリング



電気通信

- ネットワーク・プローブに基づくネットワーク・パラメーターの分析および調整
- セキュリティー・チェックのためのネットワーク・データのリアルタイム・モニタリング

工業生産

- 運用効率のための生産ラインのリアルタイム・モニタリング
- 予測型の設備資産管理

企 業

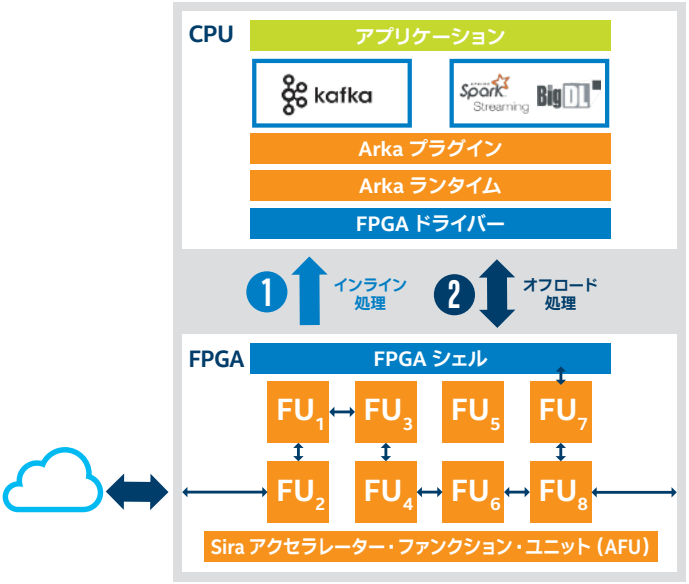
- セキュリティー・イベント管理
- 意思決定の改善のためのリアルタイムおよび履歴データの分析

ソリューション・アーキテクチャー

本質的に、Megh Computing ソリューションはオープンソースとカスタム・フレームワークで動作するプラグイン・ライブラリーを介して FPGA の複雑さを抽象化します。

Megh Computing はソフトウェアとハードウェアを組み合わせることで、ストリーミング・データのリアルタイム分析処理を可能にするために必要なパフォーマンスと柔軟性を提供します。ハードウェア・プラットフォームは、インテル® CPU ベースのサーバー上で動作するインテル® Arria® 10 FPGA またはインテル® Stratix® 10 FPGA を搭載したインテル® FPGA プログラマブル・アクセラレーション・カード (インテル® FPGA PAC) で構成されます。ソフトウェアに関しては、BigDL ライブラリーを含む Spark* Streaming フレームワークが、Arka ランタイムおよび Sira アクセラレーター・ファンクション・ユニット (AFU) 用の Megh Computing のライブラリーと組み合わされています。

- Accelerator as a Service 機能を公開する Arka ランタイムは、インテル® FPGA を管理し、AFU のソフトウェア・フォールバックをサポートします。また、フレームワーク用のカスタム・データフロー・アクセラレーター・パイプラインとプラグインの構築にも役立ちます。
- 実際のアクセラレーションを実現する Sira AFU は、インテル® FPGA にダウンロードされるライブラリーとして実装されています。アプリケーションは変更されることなく、標準またはカスタム API を使用して実行されます。



Megh Computing プラットフォームは、アプリケーションとインテル® Arria® 10 FPGA の間の中間レイヤーとして機能し、ストリーミング・データのリアルタイム分析処理を簡素化および高速化

1. 高度に最適化された Sira アクセラレーター・ファンクション・ユニットは、高レベルの API を介して Arka ランタイム・フレームワークによってアプリケーションに公開されます。
2. アプリケーションは変更されることなく実行され、API を呼び出すことができます。
3. 機能がインテル® Arria® 10 FPGA にロードされ、そこで実行されます。

Megh Computing ソリューションの利点 :

- FPGA に実装された完全なリアルタイム分析パイプライン
  - リアルタイムのユースケースを実行するための固定された低レイテンシー
  - 高いスループットで TCO を削減
- 異なるフレームワークを使用した実装を容易にするための FPGA のソフトウェア抽象化レイヤー
  - FPGA の管理と設定
  - FPGA Accelerator as a Service を公開
- 市場投入までの時間を短縮するためのカスタマイズ・ツール
  - カスタム・アクセラレーターとパイプラインを開発するための Arka/Sira SDK
  - Megh Computing が提供するサポートサービス

インテル® Arria® 10 FPGA およびインテル® Stratix® 10 FPGA

ビッグデータと AI に対する要求が高まるにつれて、FPGA の再プログラム可能なテクノロジーは、データセンター・アプリケーションの処理要件や変化するワークロードへの対応に威力を発揮します。再構成可能なロジック、メモリー、デジタル信号処理ブロックを使用することで、FPGA はあらゆる種類の機能を高いスループットとリアルタイム性能で実行するようにプログラムできるため、多くの重要エンタープライズおよびクラウド・アプリケーションにとって最適な選択肢となります。

インテル® Xeon® プロセッサーとインテル® FPGA のアクセラレーション・スタックは、業界をリードするオペレーティング・システム、仮想化およびオーケストレーション・ソフトウェアに対応しています。ソフトウェア開発者は収益化までの期間短縮、管理の簡素化、さらにアクセラレーション・ワークロード関連で活発な活動が行われているエコシステムへのアクセスを実現することができます。

インテル® PAC インテル® Arria® 10 GX FPGA 搭載版と同様に、インテル® Stratix® 10 FPGA を搭載した最新のインテル® FPGA PAC D5005（インテル® FPGA プログラマブル・アクセラレーション・カード D5005）は IP をサポートし、幅広いアプリケーション・ワークロードを高速化します。インテル® Stratix® 10 SX FPGA を搭載したインテル® FPGA PAC は、ストリーミング分析やビデオ・トランスコーディングなど、インライン処理とメモリーを大量に使用するワークロード用に構築された、より大型のフォーム・ファクター・カードです。インテル® Arria® 10 FPGA を搭載した小型フォームファクターのインテル® FPGA PAC は、バックテスト、データベースの高速化、および画像処理のワークロードに最適です。

将来に向けて

Megh Computing のリアルタイム分析プラットフォームは、インテル® FPGA アクセラレーションが実現する待ち時間の短縮やスループットの向上によって恩恵を受けるユースケースをターゲットにしています。標準化されたハードウェアおよびソフトウェアで動作するように設計されたこのソリューションは、単一ノードから大規模クラスターへの拡張、およびパブリック / プライベート / エッジクラウドへの展開を簡素化します。

Megh Computing とインテルにより、組織は現在のインフラストラクチャーを有効に活用しながら、マシンラーニングとディープラーニングの利点を実現し、高速化することが可能になります。

詳細情報

Megh Computing のリアルタイム分析プラットフォームについては、http://megh.com/（英語）を参照してください。

インテル® プログラマブル・アクセラレーション・カードとアクセラレーション・スタックの詳細については、こちらを参照してください。

ソリューション・コンポーネント

Megh Computing リアルタイム分析プラットフォームとライブラリー

Arka ランタイム

- Accelerator as a Service の公開と管理
- 連続ストリームのサポート
- 複数の FPGA のサポート

Sira アクセラレーター・ファンクション・ユニット（AFU）

- Packet Processing Engine（PPE：パケット処理エンジン）：UDP パケット処理と TCP バイパス
- Stream Processing Engine（SPE）：データ変換
- ディープラーニング・エンジン（DLE）：さまざまなトポロジーに対するインライン推論

インテル® Arria® 10 FPGA を搭載したインテル® FPGA プログラマブル・アクセラレーション・カード（インテル® FPGA PAC）

インテル® プロセッサー搭載の認定サーバー

Megh Computing について

Megh Computing の使命は、スケールアウト FPGA ベースのアクセラレーターを使用して、データセンターでコンピューティングの第 3 の波を実現することです。Megh Computing は 2017 年、データセンターでの FPGA の使用を開拓したインテルのチームによって設立されました。過去 10 年間にわたり、このチームは、コンセプトから製造まで、CPU および FPGA プラットフォームを使用したヘテロジニアス・コンピューティングを約束してきました。私たちが現在目指しているのは、フルスタック開発における当社の専門知識を活用して、パブリック / プライベート / エッジクラウドで FPGA アクセラレーターを使用してリアルタイム分析を加速するプラットフォームを提供することです。

http://megh.com/（英語）

[1] http://www.marketresearchandstatistics.com/ad/global-streaming-analytics-market/（英語）

[2] CPUのみの構成は、インテル® Xeon® プロセッサー8000番台、メモリー、および1つのエグゼキューターに基づいています（46コア）。BigDLライブラリーを使用したResNet-50モデルによる推論では、消費電力の見積もりは400 Wとなり、パフォーマンスは182フレーム / 秒です。FPGAシステムは、2枚のインテル® Arria® 10 FPGAカードを搭載したインテル® Xeon® プロセッサー3000番台で構成されています。消費電力の見積もりは、CPUと50WのFPGA x 3の構成により1ワット当たり150ドルとなり、ResNet-50モデルのFPGA上での実行によるパフォーマンスは939フレーム / 秒です。

性能の測定結果は2018年12月時点のテストに基づいています。また、現在公開中のすべてのセキュリティー・アップデートが適用されているとは限りません。詳細については、公開されている構成情報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はありません。

性能に関するテストに使用されるソフトウェアとワークロードは、性能がインテル® マイクロプロセッサー用に最適化されていることがあります。SYSmark* やMobileMark*などの性能テストは、特定のコンピューター・システム、コンポーネント、ソフトウェア、操作、機能に基づいて行ったものです。結果はこれらの要因によって異なります。製品の購入を検討される場合は、他の製品と組み合わせた場合の本製品の性能など、ほかの情報や性能テストも参考にして、パフォーマンスを総合的に評価することをお勧めします。詳細については、http://www.intel.com/benchmarks/（英語）を参照してください。

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。絶対的なセキュリティーを提供できるコンピューター・システムはありません。詳細については、各システム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。

338557-001JA