DPU を基盤としたハードウェア アクセラレーション: ソフトウェアの観点

Bob Wheeler
主席アナリスト

2021 年 6 月

www.linleygroup.com

DPU を基盤としたハードウェア アクセラレーション: ソフトウェアの観点

The Linley Group 主席アナリスト Bob Wheeler

データ プロセッシング ユニット (DPU) はデータセンターの効率を確実に高めますが、低レベル プログラミングが必要であることから、幅広い導入には至っていません。NVIDIA は、BlueField DPU のプログラミングを抽象化する DOCA フレームワークを使用して、この課題を解決することを目指しています。さらに、DOCA と CUDA の組み合わせにより、お客様が将来の DPU+GPU コンバージド ハードウェアをプログラミングできるようにします。 NVIDIA はこのホワイトペーパーの作成に出資していますが、意見や分析は著者によるものです。

DPU+GPU コンバージェンスへの道筋

データ プロセッシング ユニット (DPU) は、データセンターの効率を確実に高めるもので、異種プロセッシング ミックスの新たな要素です。DPU は、データセンターのディスアグリゲーションに重要です。DPU がネットワーク コンピューティングとストレージ間のデータ移動を処理するので、サーバー プロセッサはコンピューティング タスクのみを実行できるようになります。クラウド サービス プロバイダーは、DPU ベースのスマート ネットワーク インターフェイス カード (NIC) を使用して、サーバー プロセッサのコンピューティング サイクルを節約し、サービスから収益を創出することができます。DPU はまた、サーバー プロセッサより効率的にネットワーク トラフィックを処理するので、データセンターの消費電力が削減されます。ストレージ システムで標準のプロセッサの代わりに DPU を使用すると、消費電力を抑えつつ、SSD アレイの大量のスループットを処理できます。

現在、いくつかのベンダーが DPU という位置付けでプロセッサを提供しています。The Linley Group はこの製品分野について研究し、DPU を、「ネットワーク ポートから PCI Express (PCIe) インターフェイスまでの主要機能をすべて統合する、プログラム可能なネットワーク SoC」と定義しました。高帯域幅 PCIe インターフェイスは、従来の組み込みプロセッサだけでなく、プログラム可能なイーサネット スイッチ チップからも DPU を分離します。PCIe インターフェイスを高速パケット処理用の統合データ プレーンと組み合わせると、DPU がスマート NIC 内でネットワーク トラフィックを終端したり、ストレージ コントローラー カード内で SSD を接続したりするのに適したものになります。

NVIDIA は、2020 年の Mellanox 買収と、それに伴って獲得した BlueField ファミリによって、DPU 分野に参入しました。現在正式リリースされている BlueField-2 は、スマート NIC やストレージ コントローラーに使用され、最大 200Gbps のイーサネット ポートと高帯域幅 PCIe インターフェイスを統合します。このチップは、ハイパフォーマンスの 8 コア Arm コンプレックスや、IPSec と TLS のインラインの暗号化機能を備えたラインレートのプログラマブル データ プレーンを統合します。

BlueField-2 には、正規表現 (RegEx) アクセラレータが含まれています。これは、侵入検知、アンチウイルス、スパム フィルタリングなどのアプリケーションの文字列検索をオフロードするものです。BlueField-2 は、公開鍵暗号エンジン、真性乱数ジェネレーター (TRNG)、セキュア ブートも提供します。BlueField-2 の PCIe Gen4 x16 ホスト インターフェイスは、200Gbps のネットワーク スループットを処理できます。

GTC 2021 において、NVIDIA は BlueField-3 を 2022 年にリリースすることを発表し、図 1 に示す BlueField-4 のロードマップを公開しました。16 コアの Cortex-A78 を使用する BlueField-3 は、BlueField-2 に比べてコンピューティングのパフォーマンスが約 4 倍、ネットワーク帯域幅が 2 倍になります。最大 400Gbps のポート速度のイーサネットおよび InfiniBand を処理し、PCIe Gen5 ホスト インターフェイスで x16 スロットから 2 倍の帯域幅を利用できるようになります。

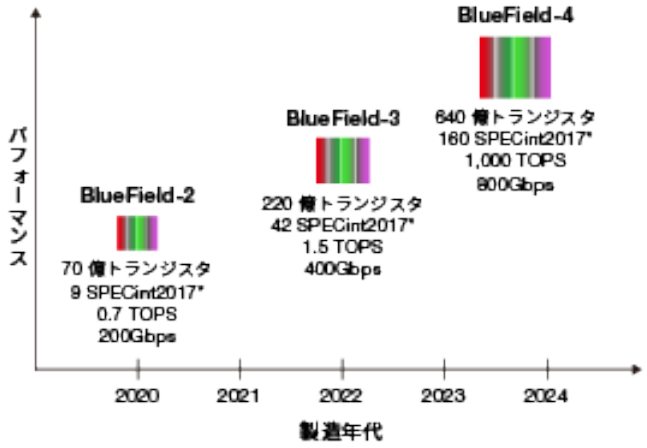

図 1.BlueField DPU のロードマップ。BlueField-3 では BlueField-2 に比べて汎用コンピューティングとネットワークのパフォーマンスが拡張されるのに対し、BlueField-4 では AI を高速化するための GPU も追加されます。*SPECrate 2017 Integer。

BlueField-2 と BlueField-3 は AI 処理に Arm コアを使用しますが、BlueField-4 には AI を高速化するための GPU が組み込まれます。これによって、チップの AI パフォーマンスは、NVIDIA の A100 などの最先端アクセラレータと同等クラスになります。その一方で NVIDIA は、2 つのチップを使用して 75 TOPS のアクセラレータを追加し、DPU+GPU ソリューションを 1 つの PCIe スロットに実装する BlueField-3X カードの提供を計画しています。サイバー セキュリティ、ソフトウェア デファインド ネットワーク、クラウド オーケストレーションなどのアプリケーションに AI 機能が追加されているので、NVIDIA の DPU+GPU ハードウェアおよびソフトウェアを採用すれば、AI 処理とネットワーク処理をどちらも高速化できます。

ハードウェア アクセラレーションを実現する BlueField

サーバー プロセッサと異なり、DPU はネットワーク パケットの処理専用です。アーキテクチャはさまざまですが、ほとんどの DPU にはコントロール プレーンとアプリケーション コード用の CPU コアに加え、プログラマブル データ プレーンが含まれています。DPU の専用データ パスは、CPU コアを使用するより効率的であるだけでなく、パフォーマンスもはるかに優れています。

図 2 に示すように、基本的に BlueField のアーキテクチャは、プログラマブル データ パスを備えた NIC サブシステム (ベースは ConnectX) と、暗号、圧縮、RegEx 用のハードウェア アクセラレータ、コントロール プレーン用の Arm コンプレックスを融合したものです。BlueField-3 では、プログラム可能なパケット プロセッサが 16 コアを構成して 256 スレッドを処理し、Arm コアに負荷をかけることなくデータ パスを処理できます。多くのアプリケーションでは、データ パスが既知のネットワーク フローを自律的に処理し、新しいフローなどの例外やコントロール プレーンの機能を Arm コアが処理します。

NVIDIA は BlueField-3 の仕様を開示していませんが、BlueField-2 のインライン処理には、100Gbps の IPSec 暗号化と 200Gbps の TLS 暗号化が含まれています。プログラマブル データ パスは、2 億 1,500 万パケット/秒 (215Mpps) の速度でステートフル フロー テーブル (SFT) や NVMe-over-Fabrics (NVMe-oF) プロトコルを実装できます。BlueField-3 が 400Gbps のイーサネットと InfiniBand を処理するようになると考えると、アクセラレータのスループットも 2 倍になると予測されます。

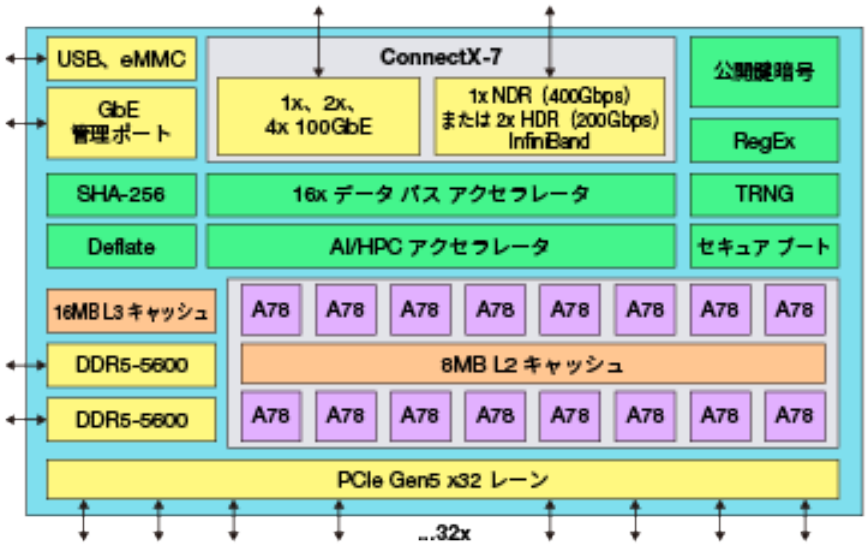

図 2.BlueField-3 DPU。プログラマブル データ パスとハードウェア アクセラレーション ブロックを組み合わせると、Arm コンプレックスにアクセスすることなく、ラインレートでの処理が可能になります。

ネットワークについては、Open vSwitch、オーバーレイ プロトコル (VXLAN など)、ネットワーク アドレス変換 (NAT)、負荷分散、きめ細かいトラフィック管理など、高度なデータセンター SDN とネットワーク機能仮想化 (NFV) が DPU によって高速化されます。 ストレージについては、DPU によって RoCE (RDMA)、NVMe-oF、保存データの暗号化、データ重複削除、分散エラー訂正、データ圧縮が高速化されます。

オフロードによってサーバー CPU コアが解放され、アプリケーション処理を行えるようになる一方、DPU はアプリケーション ドメインをインフラストラクチャ サービス ドメインから分離して、システム セキュリティを向上させることもできます。クラウド サービス プロバイダーは、この分離を利用してベアメタル コンピューティング インスタンスを提供し、仮想化されたネットワークやストレージが仮想サーバーから見えないようにします。ホストのオペレーティング システムと DPU ベースのサービス間のこのような「エア ギャップ」も、データセンター内の仮想サーバー上で開始されるマルウェア攻撃からインフラストラクチャを保護します。

DPU を導入しやすくする DOCA

DPU には確かなメリットがありますが、低レベル コードを作成する必要があるため、早期導入するお客様は限られていました。ISV、サービス プロバイダー、教育機関が DPU を導入できるようにするため、NVIDIA は DOCA (オンチップ型データセンター アーキテクチャ) を開発しました。DOCA は、実績あるドライバーを基礎とするライブラリ、ランタイム、サービスからなるフレームワークです。オープンソース プロジェクトに関連するライブラリもあれば、NVIDIA 固有のライブラリもあります。CUDA が GPU プログラミングを抽象化するように、DOCA は DPU プログラミングを高レベルに抽象化します。

図 3 は DOCA 1.1 のスタックを示したもので、ドライバー、ライブラリ、サービス エージェント、リファレンス アプリケーションが含まれています。NVIDIA は、開発者向けの DOCA SDK と、そのまま展開できる DOCA ランタイムを組み合わせて、このスタックを実現しています。たとえば、ASAP^2 はネットワーク データ パス ドライバーで、バイナリとして提供されます。フロー トラッキングおよび RegEx アクセラレータを設定する低レベル API だけでなく、VirtIO を使用して、ネットワーク デバイスのエミュレーションを行うことができます。セキュリティ ドライバーは、TLS 用にインライン カーネルをオフロードします。ストレージに関しては、SNAP ドライバーが NVMe 仮想化を提供します。これは、NVMe-oF をローカル デバイスのように使用して、リモート ブロック ストレージを接続するものです。

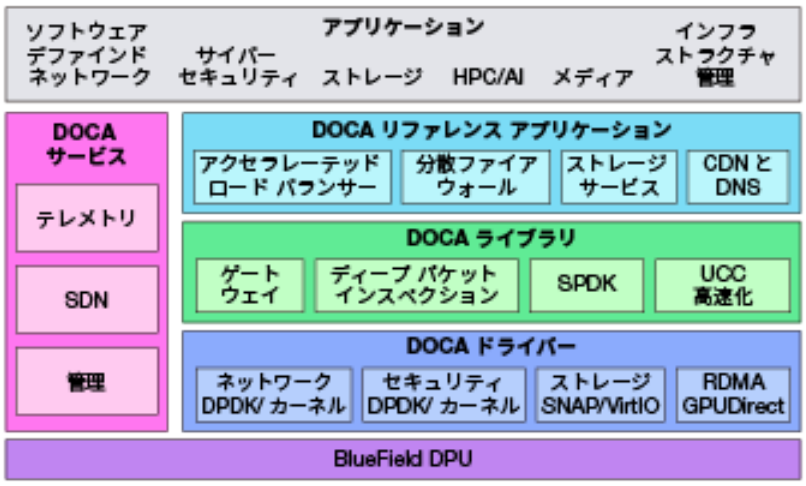

図 3.DOCA 1.1 のスタック。DOCA ライブラリはアプリケーション開発者に高レベル API を提供するので、低レベル プログラミングが不要になります。

スタックの上位にあるフロー ゲートウェイ ライブラリは、データ パスの SFT を利用して構築されたハードウェア アクセラレーテッド ゲートウェイを実装します。このライブラリは、ネットワーク トラフィックのフィルタリングや分散を行うゲートウェイ アプリケーションに、DPDK の汎用フロー API (rte_flow) より高レベルの抽象化を提供します。同様に、ディープ パケット インスペクション (DPI) ライブラリは、SFT と RegEx の高速化を組み合わせて、高レベル API をアプリケーション レイヤーに公開します。このライブラリは、コンパイル済みシグネチャ データベースでパケット ペイロードの非固定検索を行うことができます。

ストレージに関して言えば、DOCA は、ユーザー空間ライブラリを提供するオープンソースのストレージ パフォーマンス開発キット (SPDK) をサポートしています。HPC と AI に関しては、DOCA にはもともとランタイム コンポーネントとして Unified Collective Communication (UCC) ライブラリが含まれていますが、将来のリリースでは SDK のサポートが予定されています。

DOCA のサービスには、CollectX、NetQ、WJH (What Just Happened) など複数のテレメトリ エージェントと、プロビジョニング、管理、オーケストレーション用の DPU 運用ツールが含まれています。NVIDIA は、ネットワーク仮想化のパフォーマンス向上のため、標準的な OVS および OVS-over-DPDK アプリケーションを提供しています。負荷分散やアプリケーション認識を実装するリファレンス アプリケーションのソース コードも提供しています。

NVIDIA は 1.1 リリース以降、ドライバーとライブラリを追加して、DOCA のアプリケーション範囲を拡張する予定です。また、お客様が DPU データ パスをプログラミングできるよう、P4 言語と P4Runtime API のサポートを計画しています。計画中のライブラリには、Telco およびメディア アプリケーション向けの Time as a Service (TSDC)、アウトオブバンドのマルウェアを検知するホスト イントロスペクション、SPDK 向けの圧縮の高速化などがあります。

ライブラリによって低レベル プログラミングは不要になりますが、多くのアプリケーションはだ Arm に対応していません。この移行を容易にするため、DOCA は x86 ホストでの開発と展開の両方を行えるようにします。エミュレーテッド Arm コンテナーは、DPU ターゲットに x86 ベースの開発環境を提供するものです。一方で NVIDIA は、アプリケーションを Arm アーキテクチャに移植する準備ができていない、または移植できない開発者に、x86 向けの DOCA ランタイムを提供します。この場合、gRPC クライアントが DPU 上で動作し、x86 ランタイムとの通信チャネルを確立します。アプリケーションは DPU ランタイム コンポーネントにアクセスできるので、開発者が Arm コードをコンパイルする必要はありません。

DOCA のユースケース

ネットワークに DPU を使用する初期のユースケースは、サーバー CPU から仮想スイッチングをオフロードすることです。Red Hat は、BlueField DPU を使用して OVS を高速化することにより、同じ機能に 8 CPU コアを使用するのに比べ、パフォーマンスを 53 倍に向上させました。CPU コアの代わりに DPU を使用することで節約されるコストを示すため、Red Hat は投資収益率 (ROI) 分析も実施しました。100 万台の仮想マシンをホスティングするパブリック クラウドのデータセンターの場合、推定される合計 OVS スイッチング要件は 100 億パケット/秒 (VM あたり 10,000pps) でした。次に、コアあたりわずか 43,750pps のサーバー CPU を使用して、標準的な OVS のパフォーマンスを評価しました。サーバーあたり 24 コアと仮定すると、ネットワークを処理するには、9,524 台のサーバーと同等のものが必要です。

ROI を見積もるため、Red Hat はサーバー価格として 8,500 ドル、標準的なイーサネット NIC と比べた DPU の割増価格として 300 ドルを使用しました。100 万コアの場合、より少ない数の DPU 対応サーバーを使用することによる設備投資の節約額合計は 6,850 万ドルでした。また、DPU アクセラレーテッド サーバーの消費電力は合計で 4 メガワット少なく、運用経費も節約できました。この分析を考慮すると、Red Hat や VMware などのエコシステム パートナーが NVIDIA と提携して、DPU アクセラレーテッド ネットワークのインボックス サポートを提供している理由が簡単に理解できます。

DPU を使用して OVS をオフロードすることは簡単です。DPU データ パスは既知のフローのステートフル トラッキングを処理でき、Arm コアやホストが新しいフローを処理するからです。データ パスのインライン IPSec 機能は、OVS トンネルの暗号化もオフロードできます。OVS コントロール プレーンを Arm コンプレックス上で動作させることにより、DPU は vSwitch の処理をホストから完全にオフロードできます。しかし、他のアプリケーションは、BlueField の公開鍵暗号化、RegEx、圧縮、ハッシュ化用のハードウェア アクセラレータからも恩恵を受けることができます。

図 4 は、アプリケーションが DPU をデータ パスのオフロードだけでなく、ルックアサイド アクセラレータとしても使用する例を示しています。たとえば、次世代ファイアウォール (NGFW) は、SFT とインライン IPSec をデータ パスで使用しつつ、ルックアサイド API を使用して DPI と公開鍵暗号を高速化できます。DPI のその他のユースケースには、URL フィルタリング、侵入検知、アプリケーション認識などがあります。最高レベルのオフロードと分離を実現するため、ISV はアプリケーションを DPU の Arm コンプレックスに移植できますが、x86 アプリケーションであっても DPU のすべてのアクセラレータにアクセスできます。Palo Alto Networks は、DPU ベースのオフロードを VM シリーズの NGFW に追加する一方、アプリケーションを引き続き x86 ホスト上で動作させてスループットを 4 ～ 6 倍向上させ、100G イーサネット サポートを実現しました。

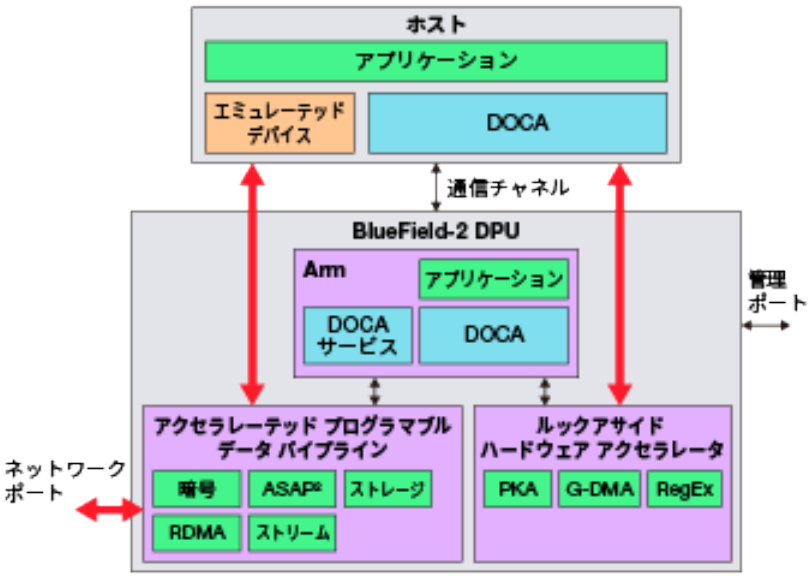

図 4.DPU オフロード アプリケーションの例。DOCA を使用すると、ホスト上で動作する x86 ベースのアプリケーションと、DPU 上で動作する Arm ベースのアプリケーションのどちらも、アクセラレーテッド データ パスとルックアサイド アクセラレータにアクセスできます。

サイバー セキュリティのもう 1 つの例が、自然言語処理モデルを使用してデータ リークを特定する、NVIDIA の Morpheus AI フレームワークです。このフレームワークは、GPU アクセラレーテッド サーバーで動作しますが、分散 BlueField DPU をネットワーク全体でセンサーとして使用します。DPU がリアルタイム テレメトリ データを Morpheus サーバーに送信すると、このフレームワークがセキュリティ ポリシーを DPU にプッシュして、脅威に対応することができます。DPU を各サーバーに配置することで、サーバーのオペレーティング システムから切り離され、サーバー CPU に負荷をかけない、マイクロセグメント化されたセキュリティを実装できます。

NVIDIA の DPU に固有のユースケースは、ストレージの仮想化とディスアグリゲーションです。SNAP は、NVMe ストレージ向けの PCIe 物理機能 (PF) を使用して BlueField を設定します。サーバーに対し、この PF は直接接続された SSD のように機能しますが、実際にはネットワーク ストレージに

アクセスします。DPU がネイティブな NVMe プロトコルを NVMe-oF パケットに変換し、オールフラッシュ アレイなどのストレージ ターゲットに送信します。DPU のデータ パスが NVMe-oF コマンド カプセルを処理し、Arm コアの利用を最小限に抑えます。NVMe-oF は RDMA を使用するので、SSD のアクセス時間に比べ、ネットワークのレイテンシの増加は無視できます。DOCA の SPDK ライブラリが、ユーザー空間 API を SNAP の透過的なストレージ仮想化の上に追加します。

まとめ

ネットワーク専用 SoC である DPU がサーバー プロセッサより効率的にネットワーク、セキュリティ、ストレージのタスクを処理できる理由は、簡単に理解できます。その一方、x86 アーキテクチャは広く普及しているため、アプリケーション開発者にとって望ましいターゲットになっています。たとえば、従来ハードウェアを販売していたネットワーク セキュリティ ベンダーは、仮想マシン上で動作する仮想アプライアンスとして x86 アプリケーションのライセンスを供与できます。移植性を確保するため、開発者は高レベル API を使用して、基礎となるハードウェアに依存しないようにする必要があります。逆に言うと、DPU の導入にはカスタムの低レベル コードが必要で、これがアプリケーション開発者にとって障壁となっています。

NVIDIA は、DOCA を使用して DPU プログラミングのための高レベルの抽象化を提供し、この課題を解決することを目指しています。開発者は、このフレームワークが提供するランタイムのバイナリと高レベル API を利用して、アプリケーション コードの作成に集中できます。複雑な DPU ハードウェアについて学ぶ必要はありません。Arm サーバーはパブリック クラウドに早期導入されつつありますが、多くのアプリケーション開発者には大量の x86 コード ベースがあり、Arm 移植の準備ができていません。このようなお客様は、NVIDIA の x86 向け DOCA ランタイムを使用すれば、Arm 移植のハードルがなくなり、今すぐに DPU を導入して、後で最適化することができます。

AI についても同様の葛藤があります。x86 サーバー プロセッサでコードを実行するか、それとも、GPU のように最適化されたハードウェアを使用してコード実行を高速化するか、という点です。競争は激しくなっていますが、NVIDIA は AI 高速化のリーダーであり続けています。その一因は、CUDA ソフトウェアの成熟度と幅広さにあります。オープンソースのニューラル ネットワーク フレームワークは、基本的に CUDA を高速化のためのデフォルト ソリューションとして利用します。NVIDIA は AI 分野でこのようなリーダーシップをとっているので、ハードウェアおよび DOCA と CUDA を組み合わせた統合開発環境からなる DPU+GPU コンバージド ソリューションを提供するという、独自のポジションを維持できているのです。

CUDA ではさまざまな世代の GPU の後方および前方互換性がサポートされています。同じように、DOCA を利用すれば、今後リリースされる BlueField-3 でもコードがシームレスに動作するとわかった上で、当面は BlueField-2 を使用し、DPU を扱い始めることができます。同様に、将来的には

CUDA コードが BlueField-4 で動作するとわかった上で、今は A100 PCIe カードなどの NVIDIA GPU を導入するといったことも可能です。NVIDIA のビジョンは、DPU が異種コンピューティングの第 3 の柱となり、CPU と GPU を補完するというものです。幅広いアプリケーションでこのビジョンを実現するには、DOCA が不可欠です。

*Bob Wheeler は The Linley Group の主席アナリストであり、*Microprocessor Report *のシニア エディターです。The Linley Group は、マイクロプロセッサと SoC 設計に関して非常に包括的な分析を提供しています。ビジネス戦略だけでなく、社内テクノロジも分析しています。詳細記事では、組み込みプロセッサ、モバイル プロセッサ、サーバー プロセッサ、AI アクセラレータ、IoT プロセッサ、プロセッサ IP コア、イーサネット チップなどのトピックを扱っています。詳細については、当社 Web サイト www.linleygroup.com をご覧ください。*