第 2 世代インテル® Xeon® スケーラブル・プロセッサーの技術概要

この記事は、インテル® デベロッパー・ゾーンに掲載されている「Second Generation Intel® Xeon® Processor Scalable Family Technical Overview」の日本語参考訳です。

はじめに

この記事では、第 2 世代インテル® Xeon® スケーラブル・プロセッサーの新機能と拡張機能、および開発者がそこから利点を得る方法について説明します。この新しいプロセッサーは、前世代のインテル® Xeon® スケーラブル・プロセッサーと同じ機能をすべて備えている上、AI ワークロード向けのインテル® AVX-512 Deep Learning Boost (インテル® AVX-512 DL Boost)、インテル® Optane™ DC パーシステント・メモリー、およびインテル® Speed Select Technology などの新たな機能を提供します。

表 1. 次世代のインテル® Xeon® スケーラブル・プロセッサー・マイクロアーキテクチャーの概要
注: 新しい機能と前世代に対する変更は**太字**で示します。

CPU	**第 2 世代インテル® Xeon® スケーラブル・プロセッサー**: 最大 28 コア、 インテル® ハイパースレッディング・テクノロジー対応、消費電力 70W から 205W
新機能	**プロセッサー周波数の向上**
	インテル® AVX512-Deep Learning Boost (インテル® AVX-512 DL Boost)
	インテル® Speed Select Technology (一部の SKU)
ソケット	ソケット P
スケーラビリティー	2S、4S、およびグルーレス 8S (xNC サポートにより >8S)
メモリー	CPU ごとに 6 チャネル DDR4 R/LDIMM、ソケットごとに 12 DIMM、 **最大 2666MT/秒 2DPC、最大 2933MT/秒 1DPC、16GB DDR4 ベースの DIMM (一部の SKU)**
	インテル® Optane™ DC パーシステント・メモリー (モジュールごとに最大 512GB) (一部の SKU)
UPI	CPU ごとに最大 3 リンク
インテル® ウルトラ・パス・インターコネクト (インテル® UPI)	x20、速度: 9.6 と 10.4GTS
PCIe*	PCIe* Gen 3: CPU ごとに 48 レーン (分岐サポート: x16、x8、x4)
ホスト・ファブリック	ディスクリート・インテル® Omni-Path アーキテクチャー・アダプター (100GB/秒)
	[統合ファブリック SKU は、前世代のインテル® Xeon® スケーラブル・プロセッサーでのみ利用可能]
インテル® C620 シリーズ・チップセット	インテル® QuickAssist テクノロジー (インテル® QAT) 拡張シリアル・ペリフェラル・インターフェイス (eSPI) 統合インテル® イーサーネット接続
	最大 4x10GB/1GB ポート、PCIe* 3.0 最大 20 ポート (8GT/秒)
	SATA 3 最大 14、USB 2.0 最大 14、USB 3.0 最大 10

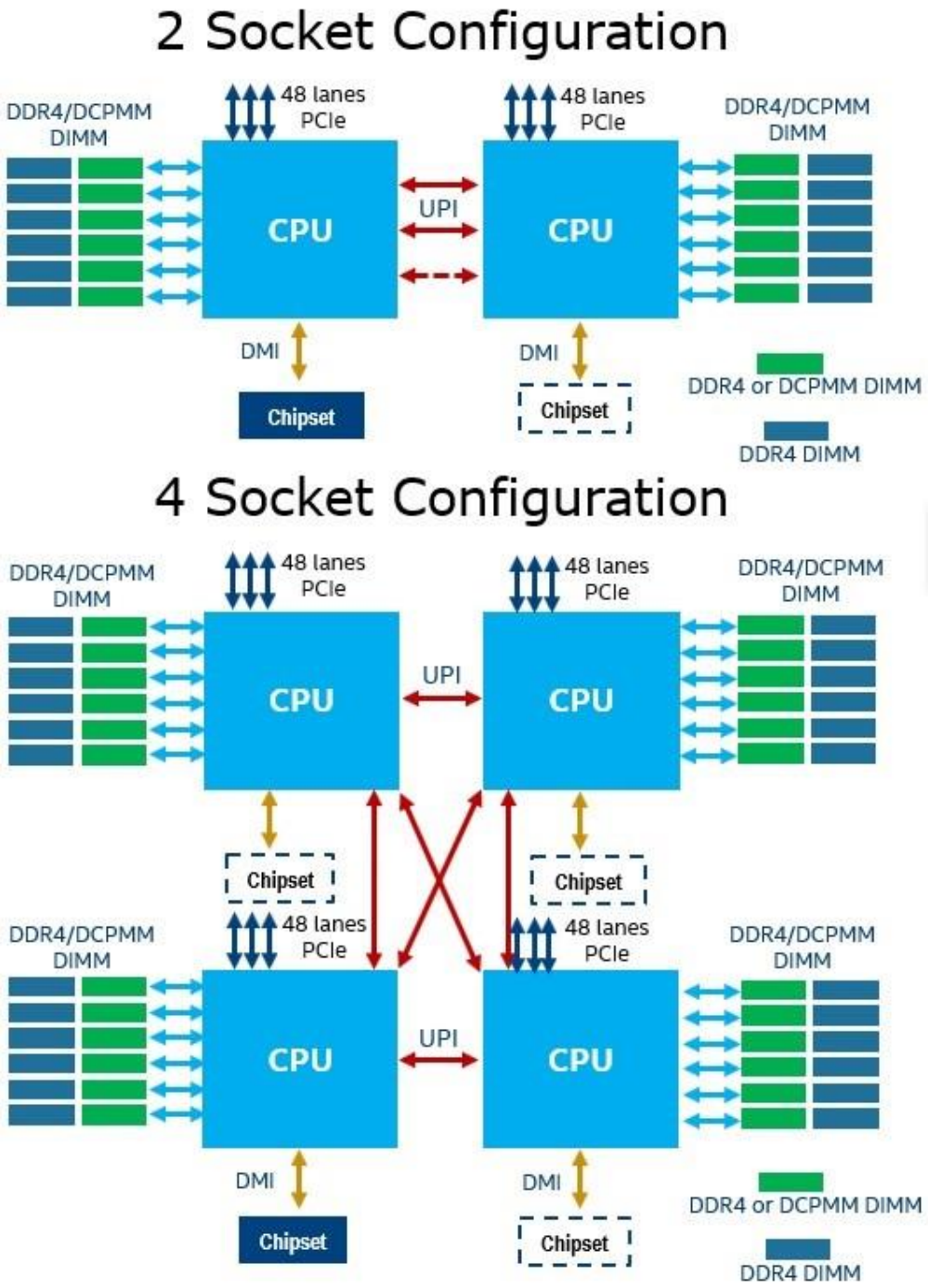

図 1. プラットフォーム構成ダイアグラム

インテル® Optane™ DC パーシステント・メモリー・テクノロジー

インテル® Optane™ DC パーシステント・メモリー・モジュールは、揮発性または不揮発性のどちらの状態でも使用することができ、DRAM とストレージ間に新しい階層を提供する新しい形態の大容量メモリーです。コンピューター・アーキテクチャーにおける伝統的なメモリーは揮発性です。揮発性メモリーの内容は、システムの電力が供給されている間だけ保持されます。電力が停止するとデータは即座に失われます。パーシステント・メモリーは、システム電源が切られたり、再起動されても、メモリーに保存されている情報の整合性を保持します。パーシステント・メモリーはバイト・アドレス指定が可能で、キャッシュ・コヒーレントで、ページングなしでソフトウェアに永続的な直接アクセスを提供します。

ハードウェア設計

インテル® Optane™ DC パーシステント・メモリー・モジュール・ベースの DIMM は、インテル® アーキテクチャー (IA) ベースのサーバー・プラットフォームの標準 DDR4 unbuffered/registered/load reduced (LR) コネクターに装着されます。DDR4 ハードウェア仕様に準拠していますが、ネイティブの DDR4 インターフェイス・プロトコルとは互換性がありません。代わりに、DDR-T と呼ばれる固有のエンコード・トランザクション・プロトコルを使用して通信が行われます。

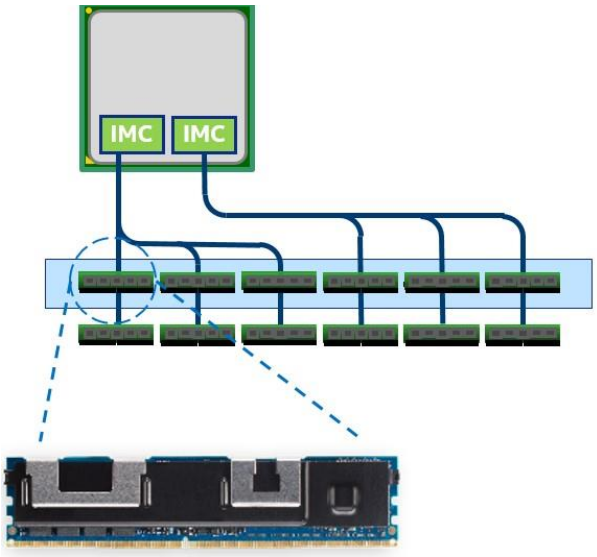

図 2. インテル® Optane™ DC パーシステント・メモリー・モジュール・ベースの DIMM、メモリー・コントローラーおよびメモリーチャネル

プロセッサーのパッケージには 2 つのメモリー・コントローラーがあり、それぞれ 3 つのメモリーチャネルとチャネルあたり 2 つの DIMM を持ちます。すべてのメモリースロットに標準の DDR4 メモリーを装着するか、スロットの半分にパーシステント・メモリー・モジュール (図 2 の青色枠) を装着できます。システムにパーシステント・メモリー・モジュールと DDR4 メモリーの両方を装着すると、追加の利点が得られます。パーシステント・メモリー・テクノロジーで利用可能な DIMM 容量は、128GB、256GB、および 512GB でプロセッサー・ソケットあたり最大 3TB です。3TB 方式をサポートするプロセッサー・モデルには、モデル番号の最後に「L」が付きます。最大 3TB のパーシステント・メモリーに加え標準のメモリーを含み、メモリーソケットあたり最大 4.5TB がサポートされます。インテル® Optane™ DC パーシステント・メモリー・ベースの DIMM は、最大 2666MT/秒の速度で動作します。

ハードウェアとソフトウェアの要件

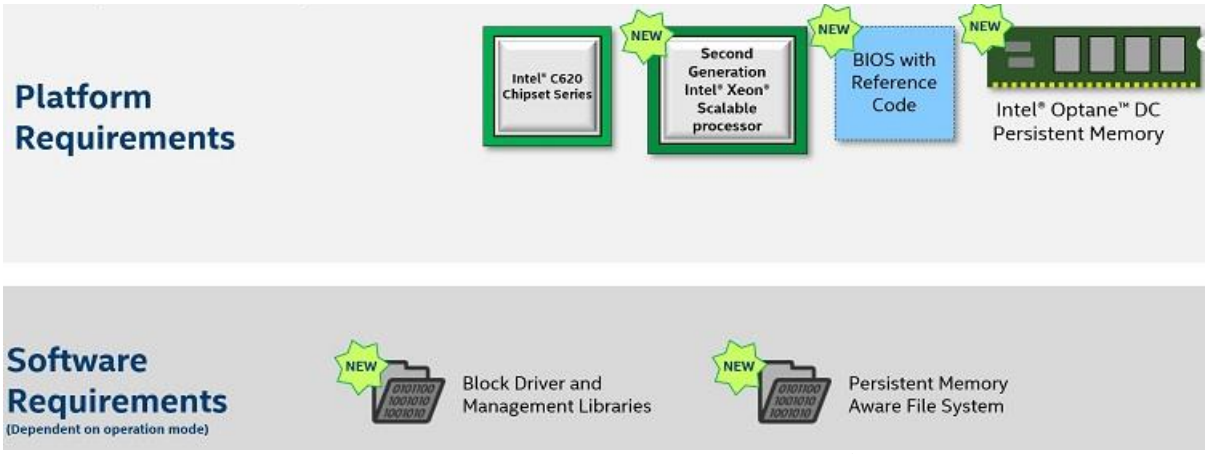

図 3. インテル® Optane™ DC パーシステント・メモリー・モジュールの技術要件

インテル® Optane™ DC パーシステント・メモリー・モジュールを利用するには、この新しいテクノロジーをサポートする BIOS、CPU、およびプラットフォームが必要です。また、サポートするブロックドライバーと管理ライブラリーを提供するオペレーティング・システム、およびインテル® Optane™ DC パーシステント・メモリー・モジュールの利点を最大限に活用するパーシステント・メモリー対応のファイルシステム (PMFS) も必要です。

インテル® Optane™ DC パーシステント・メモリー・モジュールを揮発性、不揮発性、または両方の組み合わせとして設定するには、BIOS やオペレーティング・システムを介して準備します。システムがオペレーティング・システム・レベルで起動すると、OS は適切なドライバーをロードして選択された設定を利用します。データセンターでは、インテル® Optane™ DC パーシステント・メモリー・モジュールの準備は、必要に応じて BMC (Base Management Controller) を介してリモートで行うこともできます。

インテル® Optane™ DC メモリーの操作モード

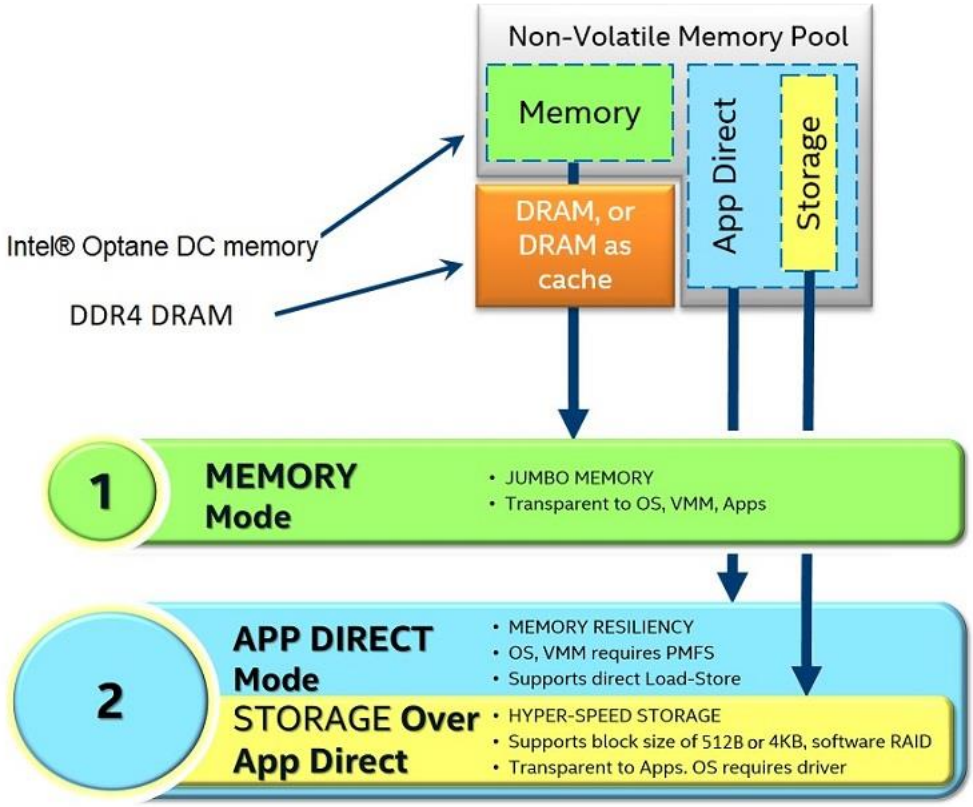

図 4. インテル® Optane™ DC メモリーの操作モード

インテル® Optane™ DC パーシステント・メモリー・モジュールは、Memory モード、App Direct モード、App Direct によるストレージモードの 3 つのモードで利用できます。Memory モードでは、インテル® Optane™ DC パーシステント・メモリー・モジュールは従来の揮発性メモリーとして、オペレーティング・システム、アプリケーション、または仮想マシンモニター (VMM) から透過的に見えます。システム上の DRAM メモリーは、自動的にメモリーシステムのキャッシュとなります。DRAM メモリーがメモリーキャッシュとして動作する場合、その領域はオペレーティング・システムが利用可能な合計揮発性メモリープールには含まれません。

App Direct モードでは、インテル® Optane™ DC パーシステント・メモリー・モジュールをパーシステント・メモリーとして動作させ、アプリケーションが明示的に利用することができます。この動作モードでは、オペレーティング・システムや VMM はパーシステント・メモリー・ファイル・システムを必要とします。App Direct モードを使用すると、アプリケーションは 64 バイトのキャッシュライン・サイズでパーシステント・メモリーを直接ロードおよびストアできます。DRAM メモリーは、システム上の揮発性メモリーの独立したプールとして見えます。

アプリケーションは、揮発性の DRAM に書き込むデータとインテル® Optane™ DC パーシステント・メモリー・モジュールに書き込むデータを明示的に決定する必要があり、ソフトウェア開発者は最適化を考慮する必要があります。Memory モードで DRAM キャッシュを介するデータの循環が早すぎるワークロードでは、キャッシュミスが多発する可能性があります。Memory モードでは DRAM キャッシュはアドレス指定できず、App Direct モードでは開発者は DRAM を見ることができます。これにより、開発者は実行頻度の高いコードを高速な DRAM メモリーに配置し、ストレージよりも高速にアクセスする必要があるデータを DRAM よりも高レイテンシーのパーシステント・メモリーに格納して、アプリケーションを最適化できます。

最後のモードは App Direct によるストレージモードです。これは、OS ネイティブの不揮発性デュアル・インライン・メモリー・モジュール (NVDIMM) のドライバーを使用して、ブロックデバイスを使用する既存のアプリケーションをサポートする機能を提供します。ドライバーは、512 バイトと 4K バイトのブロックサイズで通信でき、アプリケーションがパーシステント・メモリーをストレージデバイスとして使用することを可能にします。

プログラミングの観点から見ると、App Direct モードは 2 つの方法でパーシステント・メモリー領域へのアクセスを可能にします。これには、標準ファイル API を使用する従来の方法と、メモリー・マッピング・ファイルを使用する方法があります。これによりロードとストア命令による 64B キャッシュライン・サイズでの直接アクセスが可能になります。

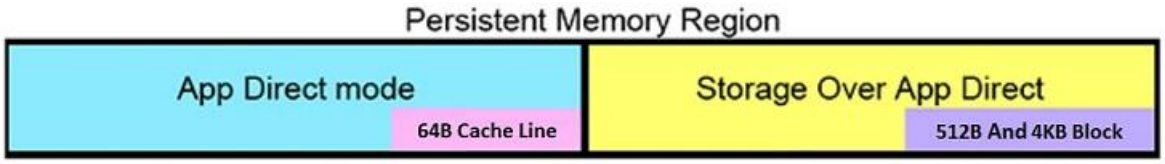

図 5. パーシステント・メモリー領域のプログラミング

ソフトウェア・アーキテクチャー

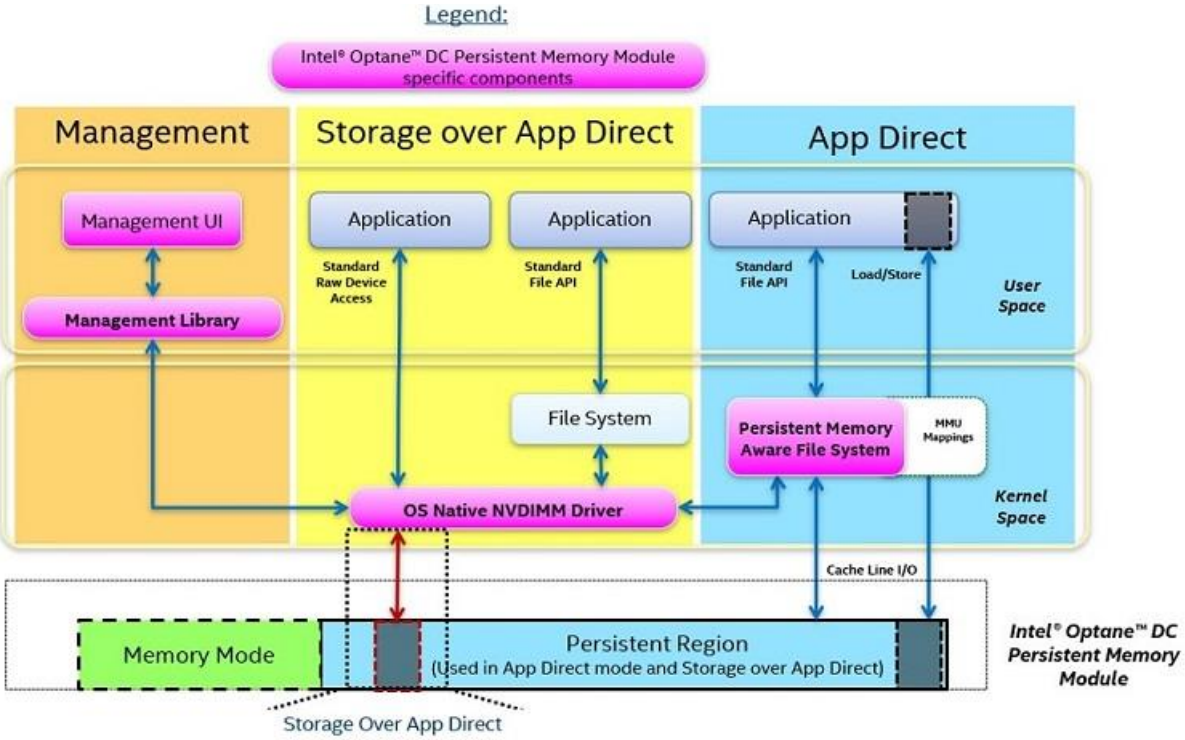

図 6. ソフトウェア・アーキテクチャー概略図

図 6 は、インテル® Optane™ DC パーシステント・メモリー・モジュールのプールを持つシステムを示しています。このモジュールには、揮発性メモリーとして割り当てられているメモリーと不揮発性 (パーシステント) メモリーとして割り当てられているメモリー領域があります。この図は、パーシステント・メモリーを使用するソフトウェアの観点から、App Direct モードと App Direct によるストレージモードの動作に焦点を当てています。

オペレーティング・システムは、OS ネイティブの NVDIMM ドライバー、または異なるモードとインテル® Optane™ DC パーシステント・メモリー・モジュール間のインターフェイスとなるドライバーのセットを利用します。図 6 の左にあるオレンジ色枠に示すように、ドライバーは管理ソフトウェアとメモリーのインターフェイスを可能にし、メモリーの分割、メモリーの健全性チェック、メモリーに関連するさまざまなセンサーデータのチェックを可能にします。これは、コマンドライン・インターフェイスまたはソフトウェア管理 API によって実行できます。

App Direct モードには、図 5 の青色枠に示されるパーシステント・メモリー領域との 2 つのインターフェイスがあります。最初のインターフェイスは、パーシステント・メモリーに対応して変更された、パーシステント・メモリーをサポートするファイルシステムを介するものです。パーシステント・メモリー対応のファイルシステムは、OS ネイティブの NVDIMM ドライバーと連携してパーシステント・メモリーのアドレス範囲を取得します。これにより、パーシステント・メモリー対応のファイルシステムは、オープン、リード、およびライトなどの通常のファイルシステム操作に加え、ブロックスタックを介さない直接ロードとストア操作を使用できます。パーシステント・メモリーをサポートするファイルシステムには、Windows* の NTFS、Linux* の EXT4 と XFS があります。

App Direct モードでパーシステント・メモリーを操作する 2 つ目の方法は、メモリー・マップ・ファイルを使用するものです。メモリーマッピングは以前からある手法で、今日ではすべてのオペレーティング・システムでサポートされています。通常、標準 DRAM を使用するメモリーマッピングはパフォーマンスに影響します。これは、バイト・アドレス指定可能な方法でメモリーにアクセスするには、ファイルをメモリー領域にページングする必要があり、コンテキスト・スイッチが発生するためです。しかし、パーシステント・メモリーでは、メモリーマッピングされたファイルでページングが発生せず、アプリケーションが直接アクセスできるため、これは当てはまりません。メモリーマッピングされたファイルを扱うアプリケーションは、メモリーマッピングによるコンテキスト・スイッチおよび割り込みの発生を回避し、ロードとストア操作を最短のコードパスで行うことができます。この直接アクセスする方法は DAX とも呼ばれます。DRAM のメモリーマッピングと同様に、ストアは Linux* の **msync()** や Windows* の **FlushFileBuffers()** などの API を使用してフラッシュされるまで永続的ではありません。CPU キャッシュのフラッシュは、**CFLUSHOPT** と **CLWB** 命令を使用して直接行うこともできます。これについては後述します。

図 6 の黄色枠に示される App Direct によるストレージモードでは、OS ネイティブの NVDIMM ドライバーは、パーシステント・メモリー領域をブロックサイズで読み書きします。これにより、既存のブロック・ストレージ・アプリケーションやファイルシステムは、ストレージデバイスのように見える透過的な方法でパーシステント・メモリー領域を直接アクセスすることが容易になります。

パーシステント・ドメイン

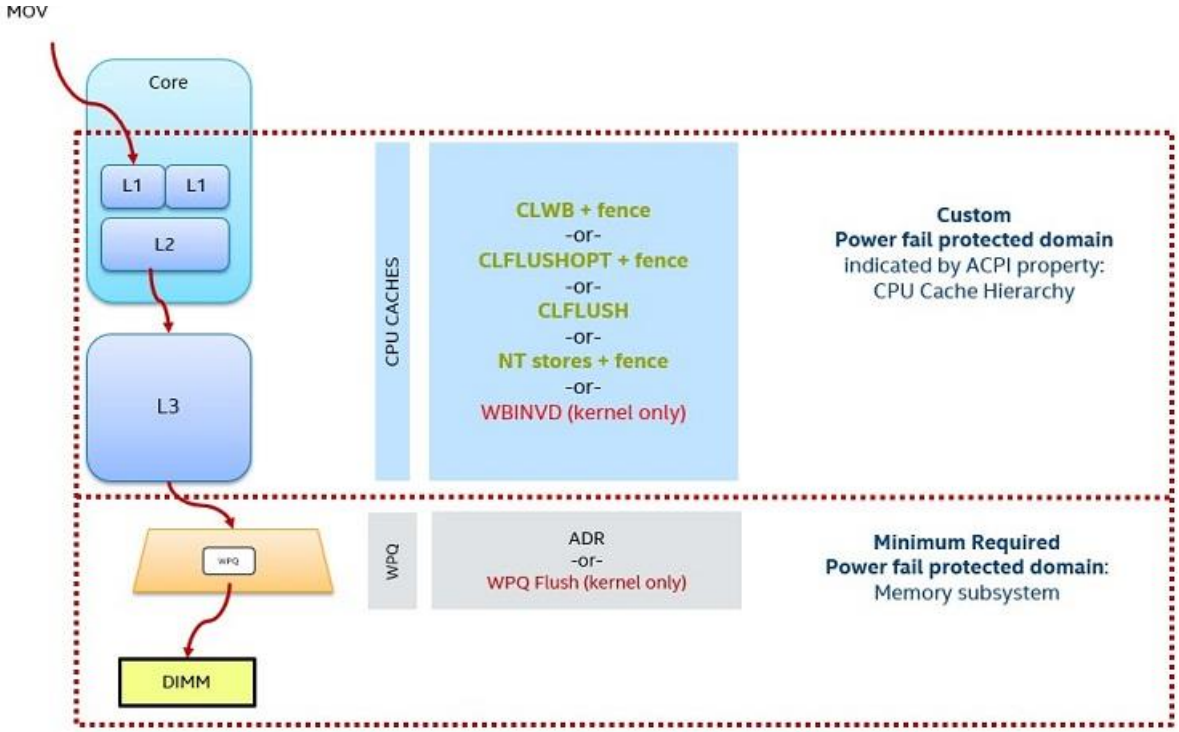

図 7. パーシステント・ドメインの概要

パーシステント・ドメインは、ストア命令が永続的であると見なされる条件を表します。プロセッサー・キャッシュは、パフォーマンスの利点をもたらすため、すべてのストアを直ちにパーシステントにできるわけではありません。キャッシュをライトスルーにするとパフォーマンスに影響します。プロセッサー・キャッシュの容量が大きいと、システムが停止したときに自動フラッシュを行うのが困難になります。これは、容量の大きいプロセッサー・キャッシュをクリアするために追加の電力が必要になるためです。

つまり、アプリケーションはプロセッサー・キャッシュをフラッシュする必要があります。これは、図 7 の中央の緑色のコマンドリストを使用して行います。ストアがプロセッサー・キャッシュからチップセットのライト・ペンディング・キュー (WPQ) に入ったときにシステムがダウンする状況を考えてみてください。プラットフォームには、WPQ が失敗した際に確実にフラッシュし、ストアをパーシステント・メモリーに保持するのに十分な電力が残っています。パーシステント・ドメインとして識別される図 7 の下の赤色枠に分類されるストアは、パーシステント・メモリーに保存されることが保証されます。パーシステント・メモリー開発キット (以下を参照) は、これらの規則に従う API を提供します。

インテル® Optane™ DC パーシステント・メモリー領域と名前空間の使い方

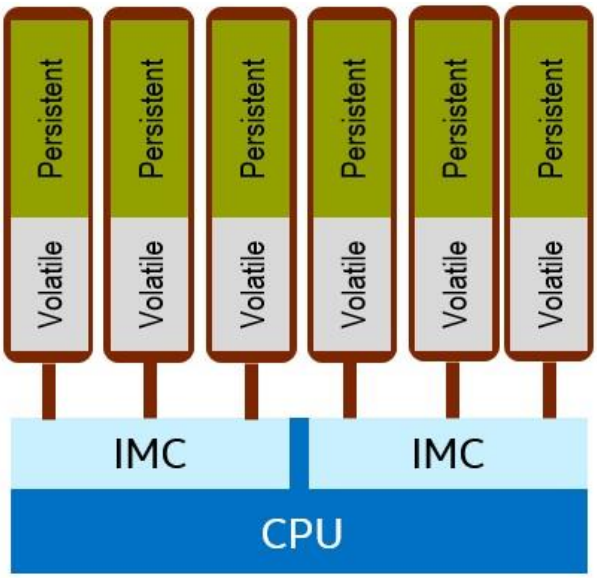

図 8. メモリー領域

インテル® Optane™ DC パーシステント・メモリー・モジュールは、揮発性、不揮発性、またはその 2 つを組み合わせたメモリー領域に設定できます。パーシステント領域は、さらに App Direct (ロード/ストアサイズのデータ) と App Direct によるストレージ (ブロックサイズのデータ) に分割することができます。図 8 に示すように、ソケット上に作成される領域は DIMM 全体で同種である必要があります。この図では、インテル® Optane™ DC パーシステント・メモリー・モジュールに揮発性領域 (Volatile) と不揮発性領域 (Persistent) の 2 つがあります。Linux* または UEFI* で DIMM 上の領域を設定または管理するには Ipmctl (英語) を使用します。

パーシステント領域の分割方法を検討する場合、それらを名前空間と呼ばれる論理デバイスに分割できます。名前空間は、オペレーティング・システムからは永続的な非連続なまたは連続した領域として見えます。名前空間には、ブロックまたは DAX 対応の 2 種類のパーシステント・メモリーがあります。Linux* で名前空間を設定および管理するには ndctl (英語) を使用します。

パーシステント・メモリー・ライブラリー

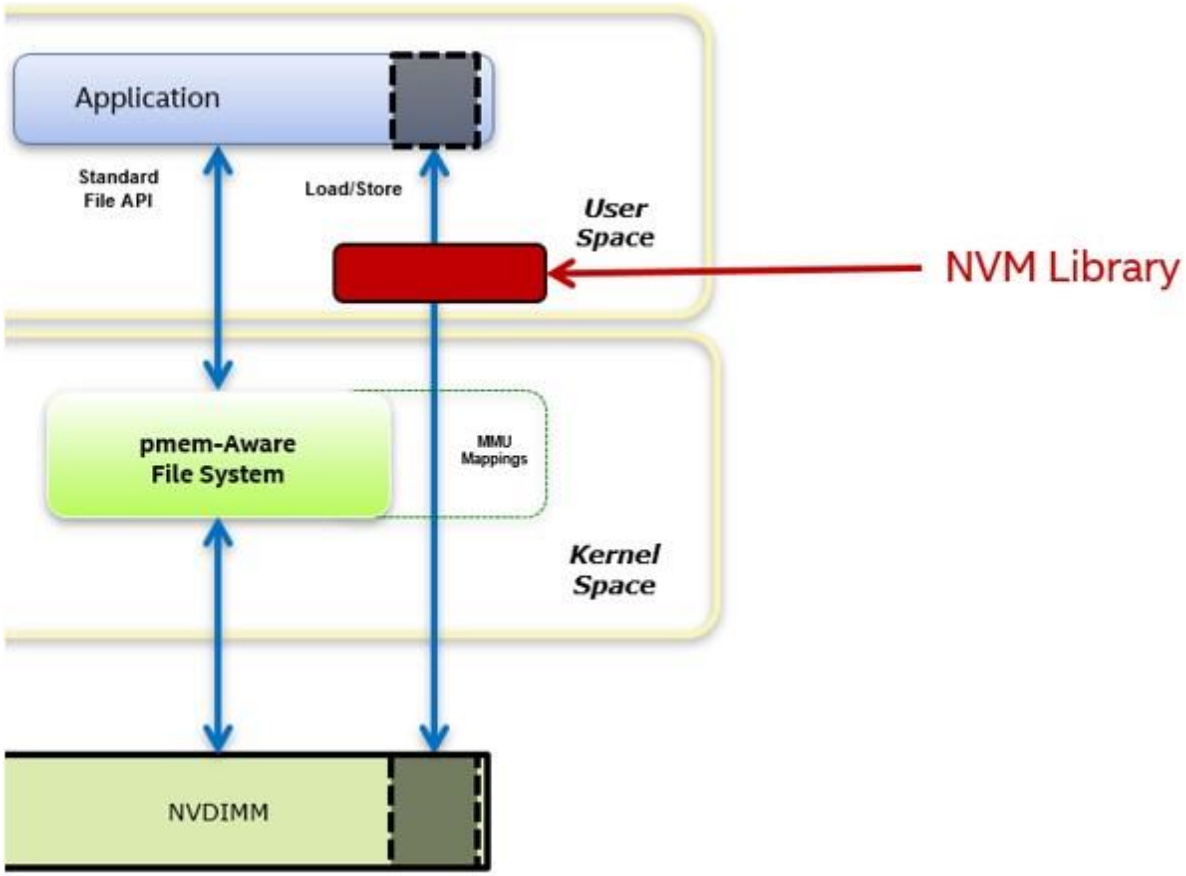

図 9. 不揮発性メモリー (NVM) ライブラリー

パーシステント・メモリー開発キット (PMDK) は、オープン・パーシステント・メモリー・プログラミング・モデルに基づいています。ライブラリーはオープンソースであり、malloc/free 形式のプログラミング、トランザクション・サポート、ハードウェアの独立性、および高可用性向けの C/C++ 言語ライブラリーを含みます。この開発キットの使用は必須ではありませんが、パーシステント・メモリー・アプリケーションの開発に役立ちます。

これらのライブラリーは、トランザクション・アルゴリズムを処理する複雑性を排除し、システムやアプリケーションのクラッシュ時にメモリーのデータを失うことなく一貫性を保つことができるため、開発者にとって重要です。パーシステント・メモリーを揮発性のように見せる利用ケースをサポートするライブラリーもあります。

CLWB 命令

メモリーマップの変更をディスクに保存する場合、変更は Linux* の **msync** や Windows* の **FlushFileBuffers** などのフラッシュ操作を行うまで保証されません。メモリーマップがパーシステント・メモリーに格納されている場合、CPU キャッシュのフラッシュ操作が必要になります。インテル® Optane™ DC パーシステント・メモリー・モジュールでは、CLWB 命令を使用して行います。

コード例

MOV X1, 10 - 10 を X1 にストア
MOV X2, 20 - 20 を X2 にストア
MOV R1, X1 - X1 と X2 へのストアはグローバルに見えますが、それでも潜在的には揮発性です
CLWB X1 - X1 と X2 をキャッシュからフラッシュ
CLWB X2
SFENCE - ストアはこの前に配置します

インテル® Optane™ DC パーシステント・メモリーの利用例

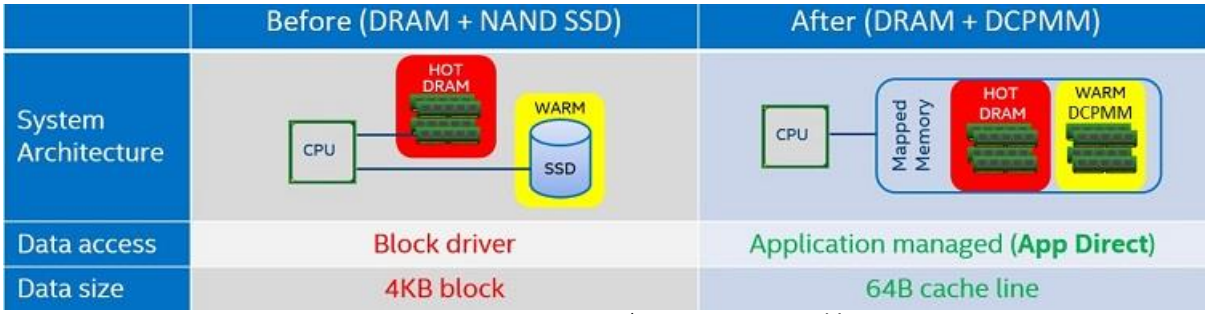

図 10. インメモリー・データベースの比較

メモリーに常駐するデータベースは、メモリー容量の増加による恩恵を受けます。これは、このような使用例におけるインテル® Optane™ DC パーシステント・メモリー・モジュールの明らかな利点の 1 つです。別の利点として、例えば、オペレーティング・システムにセキュリティー・パッチを適用するため、インメモリー・データベースをシャットダウンする場合が考えられます。パッチを適用してシステムを再起動したら、データベースをメモリーにロードし直す必要があります。しかし、テラバイト規模のデータベースではしばらく時間を要します。このシナリオでインテル® Optane™ DC パーシステント・メモリー・モジュールを使用すると、メモリーの永続性により DRAM と SSD 間でデータをページングする必要がなくなります。

さらに、小さなデータサイズでデータベースをアクセスすることも可能です。ブロックドライバーを使用する場合、4K バイトのブロック単位でデータにアクセスする必要があります。App Direct モードではバイト・アドレス指定が可能であり、ページ 4K バイト全体ではなく 64 バイトのキャッシュラインでデータベースを読み書きしたり、ブロックの一部をディスクにフラッシュする新しい選択肢を提供します。データベース・ソフトウェアは、サービスレベルの管理要件に従ってデータの優先順位に応じて、データを DRAM に保存するかインテル® Optane™ DC パーシステント・メモリー・モジュールに保存するかを選択できます。

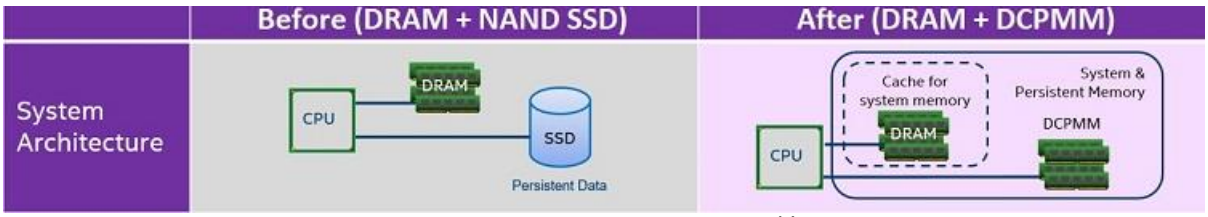

図 11. メモリーキャッシュの比較

前述のインメモリー・データベースの使用例と同様に、メモリーキャッシュを使用した場合も同じ利点があります。大容量のインテル® Optane™ DC パーシステント・メモリー・モジュールにより、SSD よりもプロセッサーに近いメモリーが利用可能となり、DRAM はシステムメモリー用のキャッシュとして動作します。追加の考慮事

項として、導入コストの観点からインテル® Optane™ DC パーシステント・メモリー・モジュールは、ソケットあたりの高いメモリー密度により、少ないノードで同数のクライアントをサポートできます。

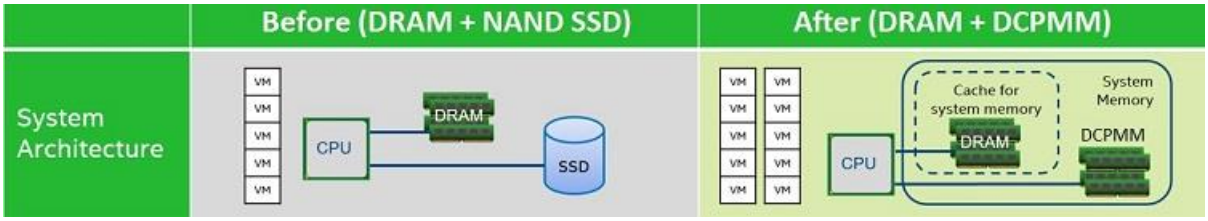

図 12. 複数使用とメモリー・スケーリングの比較

複数使用の VM 利用ケースでは、インテル® Optane™ DC パーシステント・メモリー・モジュールは、標準の DRAM よりもソケットあたりのメモリー容量が大きく、さらに多くの仮想マシン (VM) インスタンスを実行したり、VM あたりより大きなメモリー容量をもたらし、DRAM と SSD 間のページングも排除します。

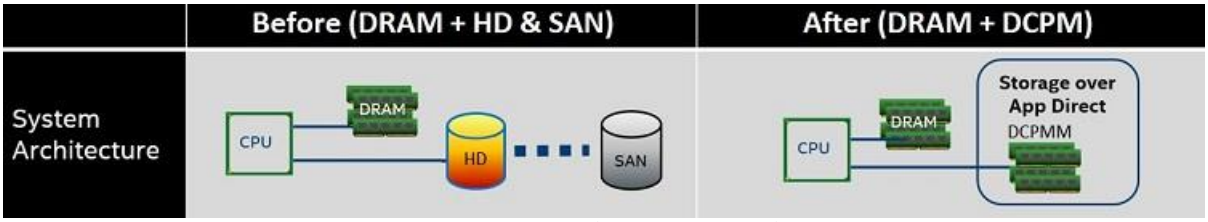

図 13. 直接接続ストレージまたは HPC ジャーナリング

App Direct によるストレージでは、多くの利用ケースで利点が得られます。典型的なストレージの利用例としては、ストレージアレイと比較して、メモリーと直接接続できることから高速化が期待できます。単一障害点を考慮してストレージデータを複製する必要がある場合、1 つの選択肢としてリモート・ダイレクト・メモリー・アクセス (RDMA) を使用する方法があります。これは、リモートノード上のパーシステント・メモリーに直接アクセスするために使用できます。これには、PMDK のライブラリーが有用です。PMDK ライブラリーは、リモートでパーシステント・メモリーに書き込みを行うメッセージ・パッシング・インターフェイス (MPI) アプリケーションなど、RDMA 上に構築されたものに対して機能します。

パーシステント・メモリーは、ハイパフォーマンス・コンピューティング (HPC) データの可用性を向上させ、大規模な HPC ノード上のブロックまたはキャッシュラインでジャーナリングを行う信頼度の高い書き込みキャッシュとして機能します。長時間計算を行う HPC の利用ケースでは、マシンがクラッシュした際に計算データが失われないようにデータのスナップショットの作成が必要になることがあります。これは、ジャーナリングやチェックポイントとして知られています。スナップショットが作成される場合、計算を続行する前に一時停止する必要があります。スナップショットがネットワークを介してストレージに作成される場合、データサイズにより停止時間が非常に長くなることがあります。しかし、スナップショットをローカルのパーシステント・メモリーに作成すると、はるかに高速になります。作成するチェックポイントは、計算実行中にインテル® Optane™ DC パーシステント・メモリー・モジュールからリモートのストレージに移行することもできます。

インテル® Optane™ DC パーシステント・メモリーの信頼性、可用性、保守性 (RAS) 機能

第 2 世代インテル® Xeon® スケーラブル・プロセッサーは、インテル® Optane™ DC パーシステント・メモリー・モジュール向けに特有の信頼性、可用性、および保守性 (RAS) 機能を備えています。RAS 機能は、インテル® Optane™ DC パーシステント・メモリー・モジュール内と DDR-T プロトコルレベルで、データの完全

性を維持するように設計されています。これにより、エラーが識別され、メモリー空間内の破損に対する保護が試行されます。RAS 機能によって識別されるインテル® Optane™ DC パーシステント・メモリー・モジュールで検出されるエラーは、標準メモリーで発生するエラーと同じ方法でアプリケーションに通知される必要があります。DRAM に関連する RAS 機能については、インテル® Xeon® スケーラブル・プロセッサーの技術概要 (英語) を参照してください。

注目すべき利用可能な RAS 機能:

- パトロール・スクラビング - 訂正可能および訂正不可能なエラー訂正コード (ECC) エラーを自律的に探知します。DDR4 メモリーと不揮発性 DDR DIMM メモリーの両方に独立したスクライバーがあり、これらは相互に非同期で動作します。
- エラー・インジェクション - OS カーネルやアプリケーション・レベルでソフトウェア・エラー処理をテストするために使用されるエラー・インジェクション・メカニズム。
- ウイルスエラーの隔離 - データ汚染メカニズムでは隔離できないエラーに対するプラットフォーム全体の抑制メカニズムです。
- データの汚染 - NVDIMM で訂正不可能なエラーが発生すると、POISON ステータスが返され、破損したデータを隔離します。

インテル® Optane™ DC パーシステント・メモリーのセキュリティー機能

インテル® Optane™ DC パーシステント・メモリー・モジュールの重要な機能は、メモリー内のデータをメモリー・スクレイピングなどさまざまな攻撃から保護するのに役立つことです。これを達成する方法の 1 つとして、安全なブートを実現するため、メモリー・コントローラーのファームウェアは、ブート時に読み取り専用のメモリーベースの信頼された経路で認証されます。もう 1 つの方法は、XTS-AES 256 暗号化レベルによってパーシステント・メモリー領域と揮発性メモリー領域の両方でメモリー内のデータを保護します。

パーシステント・メモリーの暗号化は、BIOS レベルで利用者から提供されるパスフレーズと相互作用します。パスフレーズは、各ブートサイクルでメモリー内のパーシステント・データの整合性を維持するように管理する必要があります。パーシステント・メモリーの暗号消去には、インバウンドとアウトバウンドの両方のオプションが用意されています。

揮発性メモリー領域の暗号化は、システムによって自動的に管理されるため利用者は操作する必要がありません。揮発性メモリーの暗号化キーは、システムのシャットダウン、または電源切断時に消去され、起動サイクルごとに再生成されます。これは、停電後もデータの一部を保持することがある標準 DRAM よりも安全です。

インテル® Optane™ DC パーシステント・メモリーと DIMM の障害

DIMM に障害が発生した場合、保守方法は標準 DRAM と同じで、システムを停止してユニットを交換します。RAID 5/1 などのソフトウェア RAID で App Direct によるストレージモードを使用する場合、データは新しい DIMM で再構築されます。App Direct によるストレージモードが RAID なしで使用される場合、DIMM を交換する前にデータをバックアップする必要があります。システムが新しい DIMM で起動したら、データをバックアップ・コピーから復元します。Memory モードではデータは失われます。

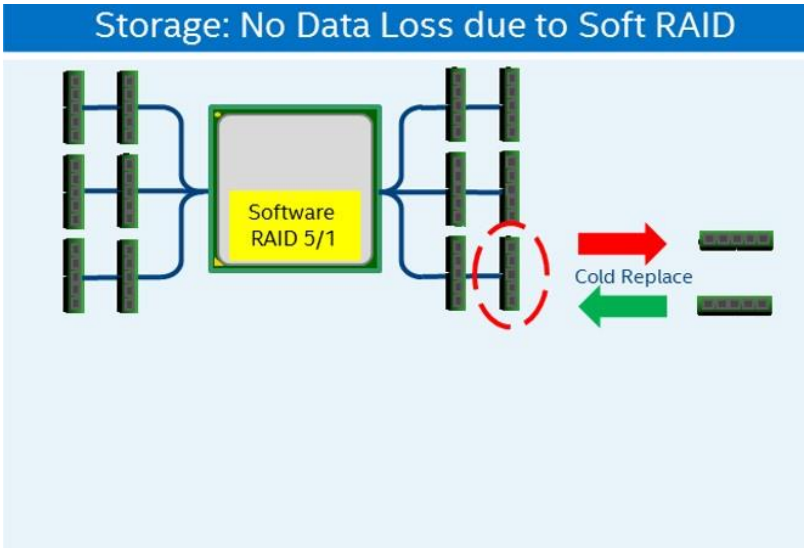

図 14. ソフトウェア RAID とインテル® Optane™ DC パーシステント・メモリー・モジュール

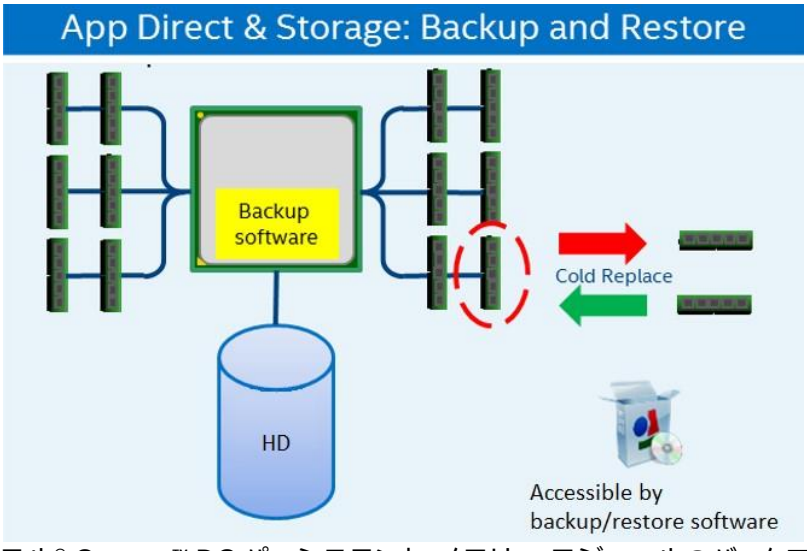

図 15. インテル® Optane™ DC パーシステント・メモリー・モジュールのバックアップと復元

インテル® Optane DC パーシステント・メモリーとプラットフォーム障害

プラットフォームで障害が発生しても、データはパーシステント・メモリー領域に保持され、手動でデータを保存し復元する必要はありません。ただし、揮発性メモリー内のデータは失われます。DIMM は同じ交換用のマザーボードに移動し、元のマザーボードの DIMM スロットと同じ順番で装着する必要があります。パーシステント・メモリー領域の暗号化セキュリティー機能により、利用者は新しいシステムのパーシステント・メモリー領域にアクセスするため、BIOS で正しいパスフレーズを指定する必要があります。

インテル® Optane™ DC パーシステント・メモリーのポリシー・プロビジョニング

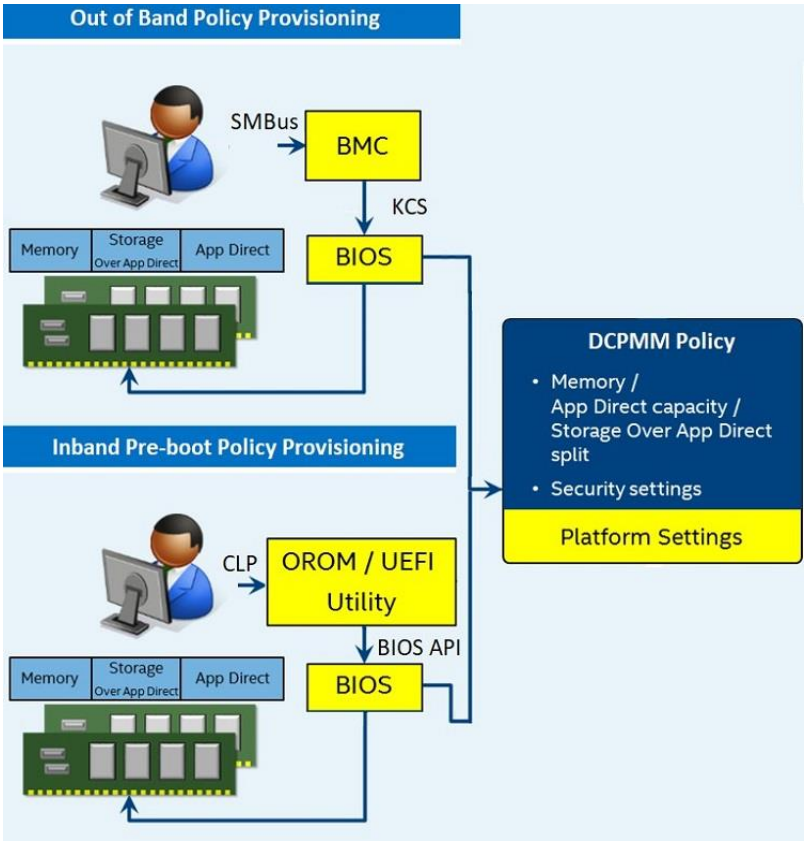

図 16. メモリー・プロビジョニングの概要

ポリシーを介したインテル® Optane™ DC パーシステント・メモリー・モジュールのプロビジョニングが有益な、クラウド・ワークロードおよびデータセンター環境のシナリオがあります。インテル® Optane™ DC パーシステント・メモリー・モジュールは、起動時のパーティション分割を含む、メモリー・パーティションと App Direct パーティションを同時にサポートします。これにより、管理者はパーティション・ポリシーを通してプラットフォーム・エージェントによってデータセンター全体に変更を適用できます。クラウド環境では、ポリシーは VMM または OS の制御下でパーティションの一部を割り当てできます。この手順は、ODROM/UEFI* と通信するコマンドライン・プロトコルを介したインバウンドで、または BMC との通信を介したアウトバウンドで実行できます。プロビジョニングに加え、インバウンドおよびアウトバウンドのオプションにより、パーシステント・メモリー領域の安全な暗号化消去やセキュリティー・パスフレーズの変更など、メモリー空間にその他の変更を加えることができます。

インテル® Optane™ DC パーシステント・メモリーの監視

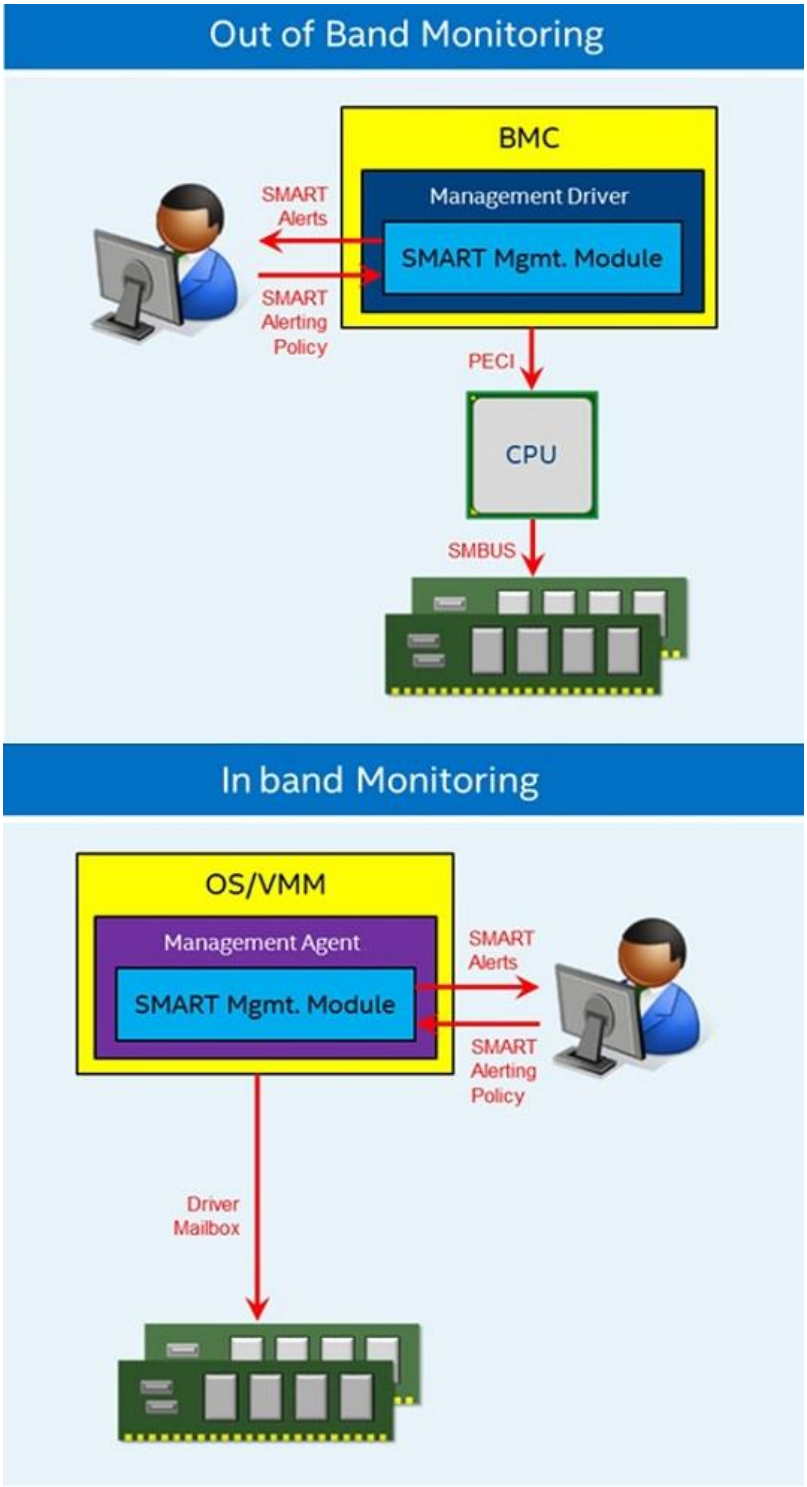

図 17. メモリー監視の概要

インテル® Optane™ DC パーシステント・メモリー・モジュールの状態を監視するインバウンドとアウトバウンドのオプションがあります。選択されているメモリーの監視方法に応じて、警告の設定に加えて、メモリーの電力、温度、および健康状態などを含む各種センサー情報を利用することができます。インテル® ノード・マネージャー (インテル® NM) は、メモリーおよびその他のプラットフォームの状況を監視するのに役立ちます。

パーシステント・メモリーのリソース

パーシステント・メモリーに関連するこれらの記事、ビデオ、およびサンプルコードは、インテル® Optane™ DC パーシステント・メモリー・モジュールの採用を促進する目的で提供されています。

- インテル® デベロッパー・ゾーン (インテル® DZ) のパーシステント・メモリー関連ページ (英語)
- Java* 向けの低レベル・パーシステント・ライブラリー (LLPL) の紹介ビデオ (英語)
- Pmemcheck でパーシステント・メモリー・プログラミングのエラーを検出 (英語)
- PMDK を使用してパーシステント・メモリーのリークを検出 (英語)
- NoSQL データベースでパーシステント・メモリーに対応 - Apache Cassandra* の例 (英語)
- パーシステント・メモリー・プログラミング向け Java* API の紹介 (英語)
- パーシステント・メモリーで C++ アプリケーションをブーストできるか? 簡単な grep サンプル (英語)
- Java* 向けの低レベル・パーシステント・ライブラリー (LLPL) の紹介 (英語)
- コード例: Panaconda - ゲーム Snake のパーシステント・メモリー・バージョン (英語)
- パーシステント・メモリー向けの Apache Cassandra* トランスフォーメーション (英語)
- パーシステント・メモリー開発キット (PMDK) (英語)
- オープン・パーシステント・メモリー・プログラミング・モデル (英語)
- アーキテクチャー全体への導入 (英語)
- パーシステント・メモリーのエミュレート (英語)

インテル® AVX-512 Deep Learning Boost (インテル® AVX-512 DL Boost)

8 ビット整数データ型の計算スループットを高めるインテル® AVX-512 DL Boost は、ニューラル・ネットワークにおけるディープラーニング推論に役立つように設計された新しいインテル® AVX-512 命令です。ディープラーニングは、訓練と推論の 2 つで構成されます。

ディープラーニングではさまざまなアルゴリズムを利用しますが、その主なものは画像認識、音声認識、および言語翻訳です。これらすべてのケースには、多数の既知の入力による訓練フェーズを持つ計算モデルが含まれます。例えば、写真に写る猫や犬を識別するように設計された画像認識アルゴリズムでは、猫や犬に対する多くの既知の画像をベースにモデルを訓練します。アルゴリズムは写真を処理して、イメージが猫もしくは犬であるかを予測します。ある猫のイメージに対し 80% が猫で 20% が犬という予測が返された場合、アルゴリズムは精度を向上するため何らかの補正を必要とします。

アルゴリズムは自動的に特定の重みを調整して、既知の入力に対する以降のパスの精度を高め、猫または犬の既知の写真に対する推測が 100% に近くなるように試みます。訓練フェーズではこのような処理が行われ、これらは計算集約的です。計算の正確さが重要であり、そのためにはさらに広範囲の数を必要とします。これらの数値の精度を維持するため、訓練フェーズでは浮動小数点データ型が有用です。

インテル® AVX-512 DL Boost は、より小さなデータ型を使用してワークロードの推論を支援するように設計されています。推論プロセスの解析と比較は、訓練フェーズと比較して小さなバッチデータを使用します。ディープラーニングの推論は、アルゴリズムが既知のイメージで訓練済みであり、未知のイメージが解析のため提供された場合に行われます。訓練フェーズに基づいて、アルゴリズムは未知のイメージが何であるかを推測します。この例では、イメージが猫であるか犬であるかを判断します。インテル® AVX-512 DL Boost 向けのソフトウェア開発のサポートは、ディープ・ニューラル・ネットワーク向けインテル® マス・カーネル・ライブラリー (インテル® MKL-DNN) では有効になっています。アプリケーションがすでにインテル® AVX-512 基本命

令を活用している場合、最小限の作業でオペレーティング・システムや VMM のインテル® AVX-512 DL Boost サポートを有効できます。

図 18. 新しい VPDPBUSD 命令で古い 3 つの命令を置き換える

インテル® AVX-512 DL Boost には次のような特徴があります。

- VPDPBUSD の導入。内部ループにある 3 つの命令シーケンス (VPMADDUBSW、VPMADDWD、および VPADDD) を新しい単一の 8 ビット融合命令に置き換えます。
- 128 ビットと 256 ビットのベクトルを処理する機能

インテル® Speed Select

インテル® Speed Select Technology は、CPU のパフォーマンスを細かく制御できる機能の集まりです。これまでプロセッサーには、固定基本周波数、発熱制限、電力エンベロープなど、パッケージ上のすべてのコアで共有される特性が備わっていました。インテル® Speed Select Technology - Performance Profile (インテル® SST-PP) は、プロセッサー・コアのグループに固有の特性を割り当てることができます。

インテル® SST-PP は、プロセッサーが 3 つの異なる動作ポイントで動作可能な新機能を提供します。それぞれの動作ポイントは、コア数、基本コア周波数、熱設計電力、および最大温度で構成される固有のプロファイルを持っています。これらの動作ポイントは、システム起動時に BIOS によって検出および実装されます。

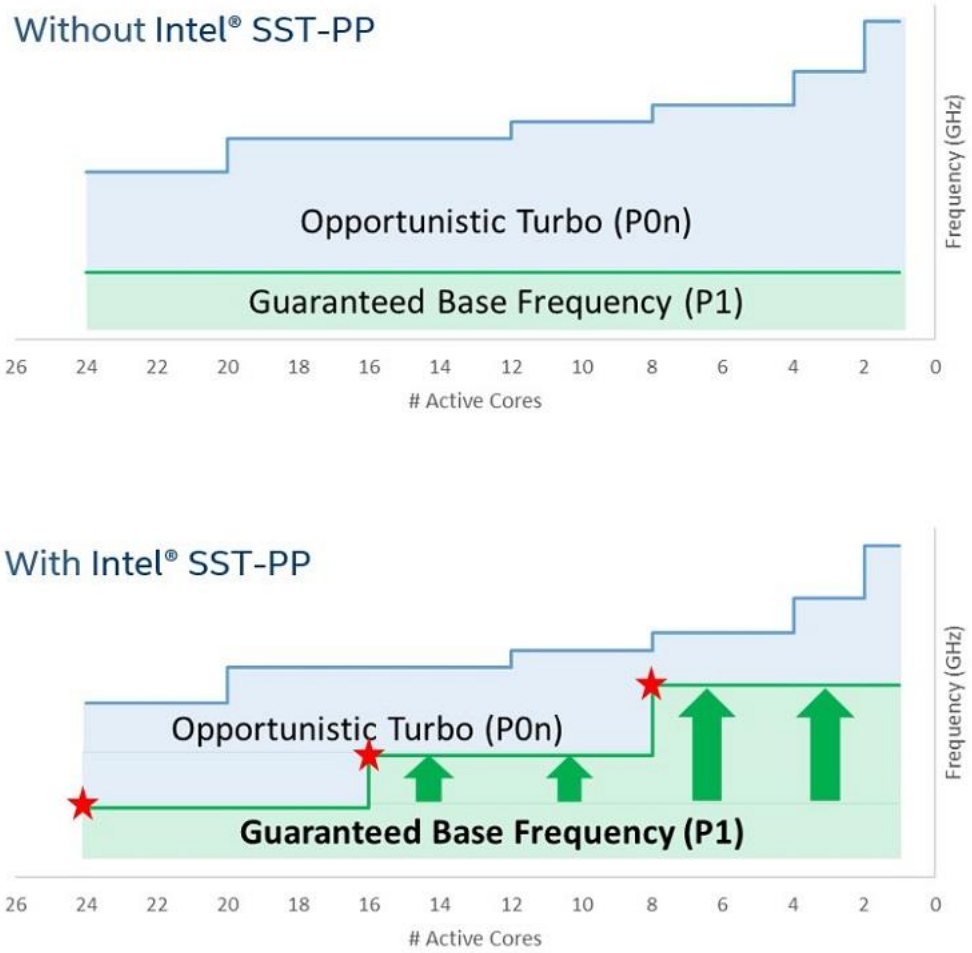

図 19. インテル® SST-PP あり/なしのプロセッサーの比較
(周波数とコア数は説明を目的としたものです)

インテル® Speed Select は、サービス・レベル・アグリーメント (SLA) のコアあたりで保証されるパフォーマンスを向上させたり、特定のワークロードや仮想マシンの要件に基づいて複数プロセッサーのプロファイルを作成することができます。データセンターのワークロードは時間とともに変化します。リソースの変更や物理的な移動を管理するのに人手を費やすのではなく、インテル® Speed Select は、変化するワークロード要件を満たすためシステムを自動的に再構成するリモートアクセスによる解決策を提供します。

これら動作ポイントの設定は、BIOS または RESTful* Redfish* 管理フレームワークの API を介して行うことができます。インテル® ラック・スケール・デザイン (インテル® RSD) を使用するサードパーティーのオーケストレーション・ソフトウェアを介して行うこともできます。

モデル番号の末尾が「Y」である第 2 世代インテル® Xeon® スケーラブル・プロセッサーのみが、インテル® SST-PP をサポートします。

図 20. インテル® Speed Select Technology - Base Frequency (インテル® SST-BF) あり/なしのプロセッサーの比較 (コア数は説明を目的としたものです)

インテル® SST-BF は、ネットワークと仮想化のワークロードに重点を置いた、プロセッサー・モデルに関連するもう 1 つの新機能です。インテル® SST-BF は、優先順位の高いコアの基本周波数を上げてボトルネックに対処し、優先順位の低いワークロードを実行するほかのコアの基本周波数を下げる機能を提供します。

インテル® SST-BF をサポートするのは、プロセッサー・モデル番号の最後に「N」が付くもののみです。

インテル® リソース・ディレクター・テクノロジー (インテル® RDT)

インテル® リソース・ディレクター・テクノロジー (インテル® RDT) は、共有リソースの監視と管理に役立つ機能を提供します。キャッシュ監視テクノロジー (CMT)、キャッシュ割り当てテクノロジー (CAT)、メモリー帯域幅監視 (MBM)、コードデータ優先付け (CDP) などの機能は前世代のプロセッサーでもサポートされています。

インテル® リソース・ディレクター・テクノロジー (インテル® RDT)

インテル® RDT の背景にある主な原則を説明するアニメーションについては、インテル® リソース・ディレクター・テクノロジーによるリソース利用率の最適化 (英語) を参照してください。

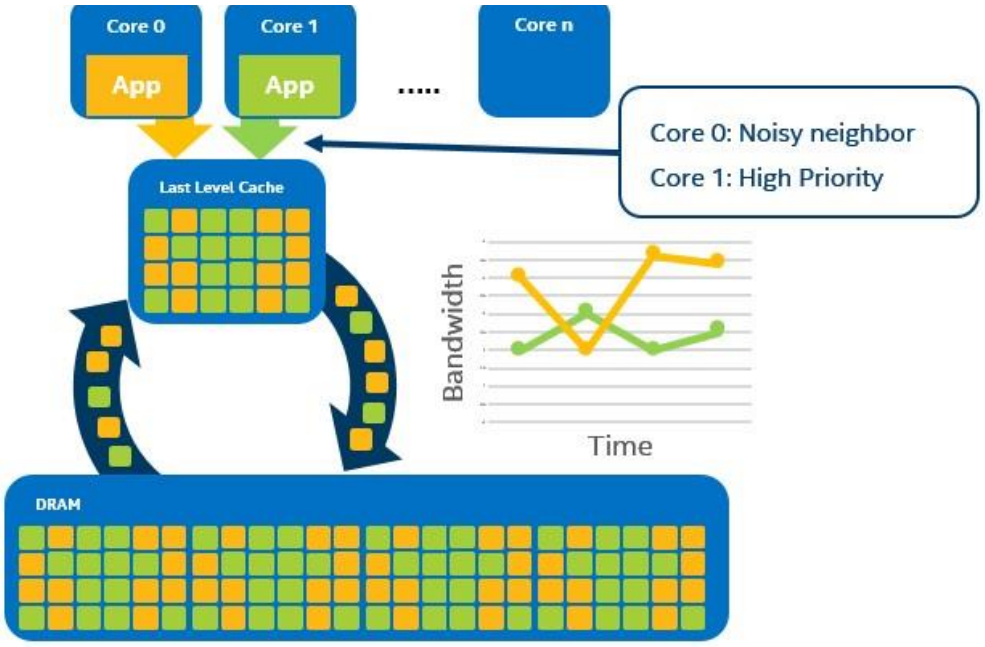

図 21. メモリー帯域幅モニタリング (MBM) を使用して うるさい隣人 (Noisy neighbor - core 0) を特定

第 2 世代インテル® Xeon® スケーラブル・プロセッサーでは、メモリー帯域幅割り当て (MBA) と呼ばれる新しい機能が導入されています。これは、スレッドごとのメモリー帯域幅を制御するために追加されました。この機能は、うるさい隣人 (noisy neighbor) を分離するため、MBM と組み合わせて利用されます。図 21 は、メモリー帯域幅を占有している VM またはアプリケーション (コア 0 上のオレンジ色のアプリ) と、リソースを使い果たしている 2 番目の VM またはアプリケーション (コア 1 の緑色のアプリ) を示しています。メモリー帯域幅がどのように変化しているかをこのレベルで把握するには、インテル® RDT で MBM を使用します。MBM は、ソフトウェア・スレッドに割り当てられているリソース監視 ID (RMID) を利用します。オペレーティング・システムや VMM は、いつでも任意の RMID に関連するメモリー帯域幅利用率を読み取ることができます。

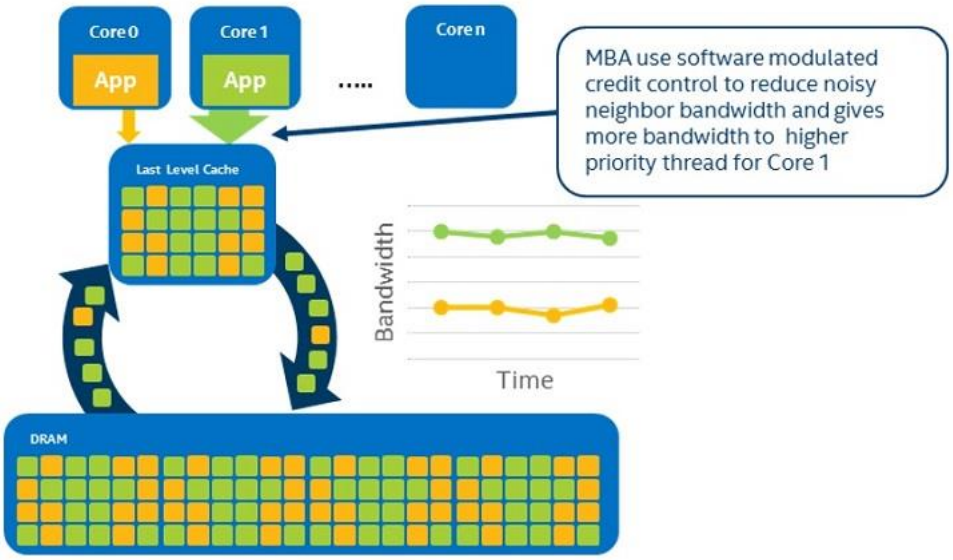

図 22. MBA はメモリー帯域幅を優先付け

MBM でうるさい隣人を特定したら、RMID にいくつかのクレジットを割り当てるため MBA の機能を使用します。サービス・レベル・アグリーメント (SLA) の範囲内で動作できるように、うるさい隣人に特定数のクレジットを付与し、クレジットがなくなったら、メモリー帯域幅へアクセスできなくします。これにより、2 番目の VM またはアプリケーションが同じリソースを使用できるようになります。

『インテル® 64 および IA-32 アーキテクチャー・ソフトウェア開発者マニュアル (Vol. 3)』(英語) の 17.16 節でインテル® RDT 機能を使用したプログラミングの詳細が説明されています。

この機能を利用するには、OS または VMM レベルで有効にする必要があります。また、BIOS レベルで IA-32、インテル® 64 およびインテル® アーキテクチャー向けのインテル® バーチャライゼーション・テクノロジー (インテル® VT-x) 機能を有効にする必要があります。インテル® VT-x の設定方法については、OEM BIOS ガイドを参照してください。

インテル® RDT リソース

- リファレンス・アプリケーション (英語)
- カーネルがサポートするリソース・コントロール・ファイル・システム (英語)
- c-group でサポートされるキャッシュ割り当て技術 (英語)
- Perf でサポートされるキャッシュ・モニタリング技術 (英語)
- インテル® Performance Counter Monitor 向けのキャッシュ・モニタリング技術サポート
- インテル® RDT のエラッタ (英語)

ハードウェアによるサイドチャネル攻撃の軽減

サイドチャネル方式に関連するセキュリティー上の脆弱性を改善するため、第 2 世代インテル® Xeon® スケーラブル・プロセッサーではハードウェアに変更が加えられました。これらの変更により、Branch Target Injection、Rogue Data Cache Load、および L1 Terminal Fault などのサイドチャネル攻撃が軽減されます。ハードウェアの修正は、一般に入手可能なすべてのプロセッサー SKU で実装されています。これらのハードウェアの修正の中には、OS カーネルと VMM のアップデートが必要なものもあります。サポートについてはソフトウェア・ベンダーにお問い合わせください。

サイドチャネル解析とインテル製品について (英語)
ソフトウェア開発者向けのリソースとガイド (英語)

ソフトウェア・ツール

以下は、第 2 世代インテル® Xeon® スケーラブル・プロセッサーを使用するため最適化されたソフトウェア・ツールの一覧表です。

表 2. ソフトウェア・ツール

ツール	バージョン	説明
インテル® Performance Counter Monitor (インテル® PCM) (英語)	2.11.1	プロセッサー内部のリソースの使用状況を推定するサンプル C++ ルーチンとユーティリティーを提供します。
インテル® Memory Latency Checker (インテル® MLC) (英語)	3.5	システムのメモリー帯域幅をテストします。
インテル® VTune™ Amplifier		ワークロードのパフォーマンスを解析およびプロファイルして最適化に役立てます。
インテル® C++ コンパイラー	Windows* 版	最適化コンパイラーと MP LINPACK ライブラリー。
	Linux* 版	
インテル® Parallel Studio XE		最適化された MP LINPACK* ライブラリー。
ストレージ、ネットワーク、およびパーシステント・メモリー向けに最適化された開発キットとライブラリー	SPDK	spdk.io
	DPDK	dpdk.org
	PMDK	pmem.io
インテル® Xeon® スケーラブル・プロセッサーのアンコア・パフォーマンス・モニター (zip) (英語)		アンコアとアンコア・サブコンポーネントの監視機能について詳しく説明しています。

著者

David Mulnix は、インテル コーポレーションに 20 年以上勤務しているソフトウェア・エンジニアです。David が注目する分野は、ソフトウェアの自動化、サーバー電力、およびパフォーマンス解析であり、Server Efficiency Rating Tool (SERT*) の開発に貢献してきました。

寄稿者: Andy Rudoff

関連情報

第 1 世代インテル® Xeon® スケーラブル・プロセッサーの技術概要 (英語)
インテル® 64 および IA-32 アーキテクチャー・ソフトウェア開発者マニュアル (SDM) (英語)
インテル® アーキテクチャー命令セット拡張プログラミング・リファレンス (英語)
インテル® リソース・ディレクター・テクノロジー (インテル® RDT)
インテル® Xeon® スケーラブル・プロセッサーのアンコア・パフォーマンス・モニター (zip) (英語)
インテル® C620 シリーズ・チップセット・プラットフォーム・データシート (英語)
インテル® AI 開発者プログラム (英語)
インテル® リソース・ディレクター・テクノロジーによるリソース利用率の最適化 (英語)
インテル® メモリー・プロテクション・エクステンション (インテル(R) MPX) イネーブリング・ガイド (英語)
インテル® Run Sure テクノロジー (英語)
インテリジェント・リテール・デバイス向けのインテル® ハードウェア・ベースのセキュリティー技術 (英語)
ジェームス・レインダース氏によるプロセッサー・トレース (英語)
インテル® ノード・マネージャー Web サイト (英語)
インテル® ノード・マネージャー・プログラマーズ・リファレンス・キット (英語)
インテル® ノード・マネージャー向けオープンソース・リファレンス・キット (英語)
インテル® ノード・マネージャーの設定方法 (英語)
インテル® キャッシュ・アクセラレーション・ソフトウェア (インテル® CAS) (英語)
インテル® データセンター・マネージャー (インテル® DCM)
インテル® Performance Counter Monitor - より優れた CPU 使用率の測定方法
インテル® Memory Latency Checker (インテル® MLC) (英語)
インテル® VTune™ Amplifier
インテル® Data Center Modernization Estimator (英語)

コンパイラーの最適化に関する詳細は、最適化に関する注意事項を参照してください。