インテル® Optane™ DC パーシステント・メモリーを使用するアプリケーションの最適化

IA Software User Society (iSUS)
編集長　すがわら　きよふみ

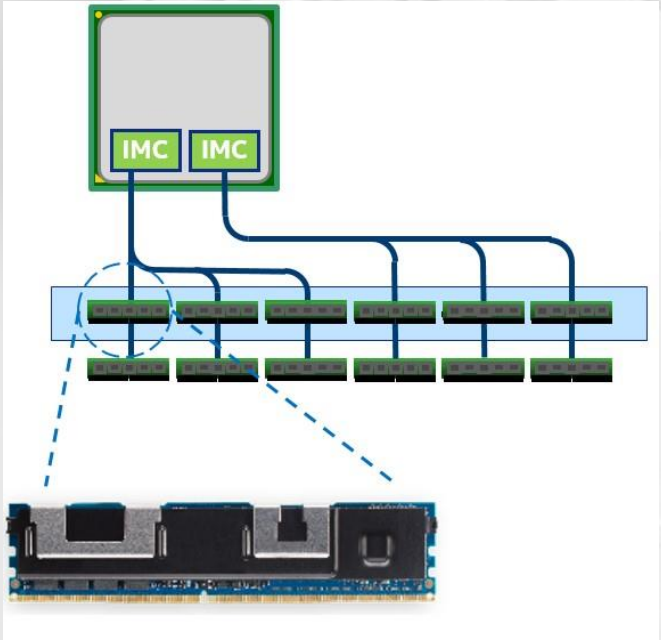
内容

インテル® Optane™ DC パーシステント・メモリーとは

- ハードウェア設計と要件
- 動作モード
- ソフトウェア・アーキテクチャーとパーシステント・ドメイン
- パーシステント・メモリー・ライブラリー
- 利用例
- 計画から導入まで
- BIOS 設定の例
- App Direct によるストレージモードの例 (Windows* 環境)
- App Direct モードの例

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

- 第 2 世代インテル® Xeon® スケーラブル Gold プロセッサーでサポート
- 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) コネクターに装着できます
- 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
- 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

- 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
- 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

プラットフォーム要件

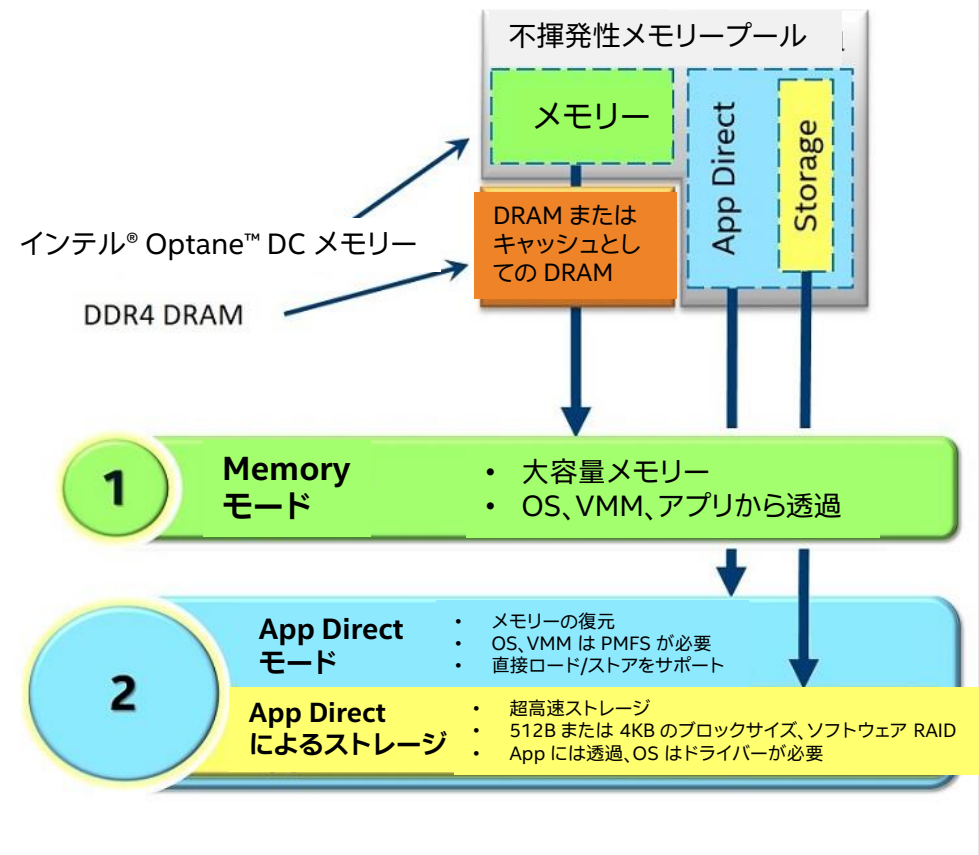
ソフトウェア要件

ブロックドライバーと管理ライブラリー

パーシステント・メモリー対応のファイルシステム

DCPMM の動作モード

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



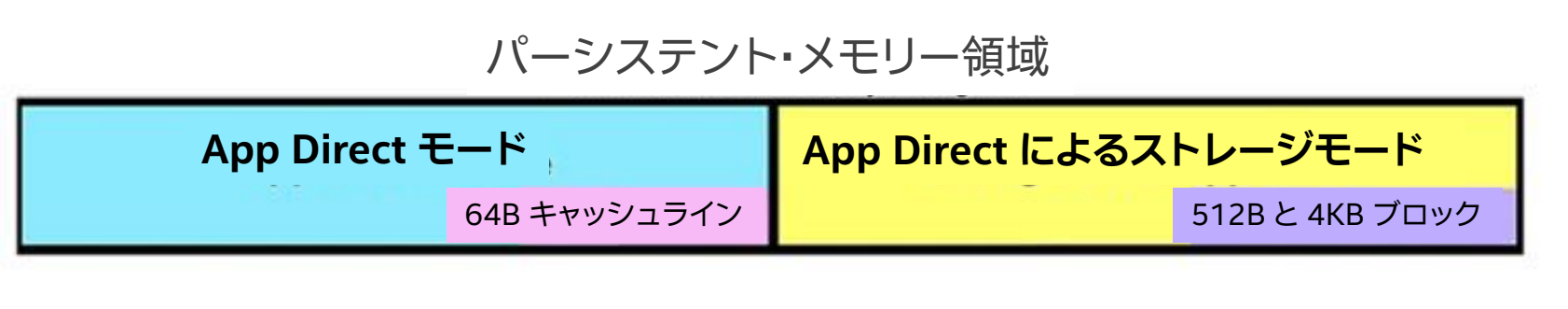
- Memory モードでは、インテル® Optane™ DC パーシステント・メモリー・モジュールは従来の揮発性メモリーとして、オペレーティング・システム、アプリケーション、または仮想マシンモニター (VMM) から透過的に見えます
- App Direct モードでは、インテル® Optane™ DC パーシステント・メモリー・モジュールをパーシステント・メモリーとして動作させ、アプリケーションが明示的に利用することができます
- App Direct によるストレージモードでは、OS ネイティブの不揮発性デュアル・インライン・メモリー・モジュール (NVDIMM) のドライバーを使用して、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

DCPMM の動作モード (2)

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:

- 標準ファイル API を使用する従来の方法
- メモリー・マッピング・ファイルを使用する方法

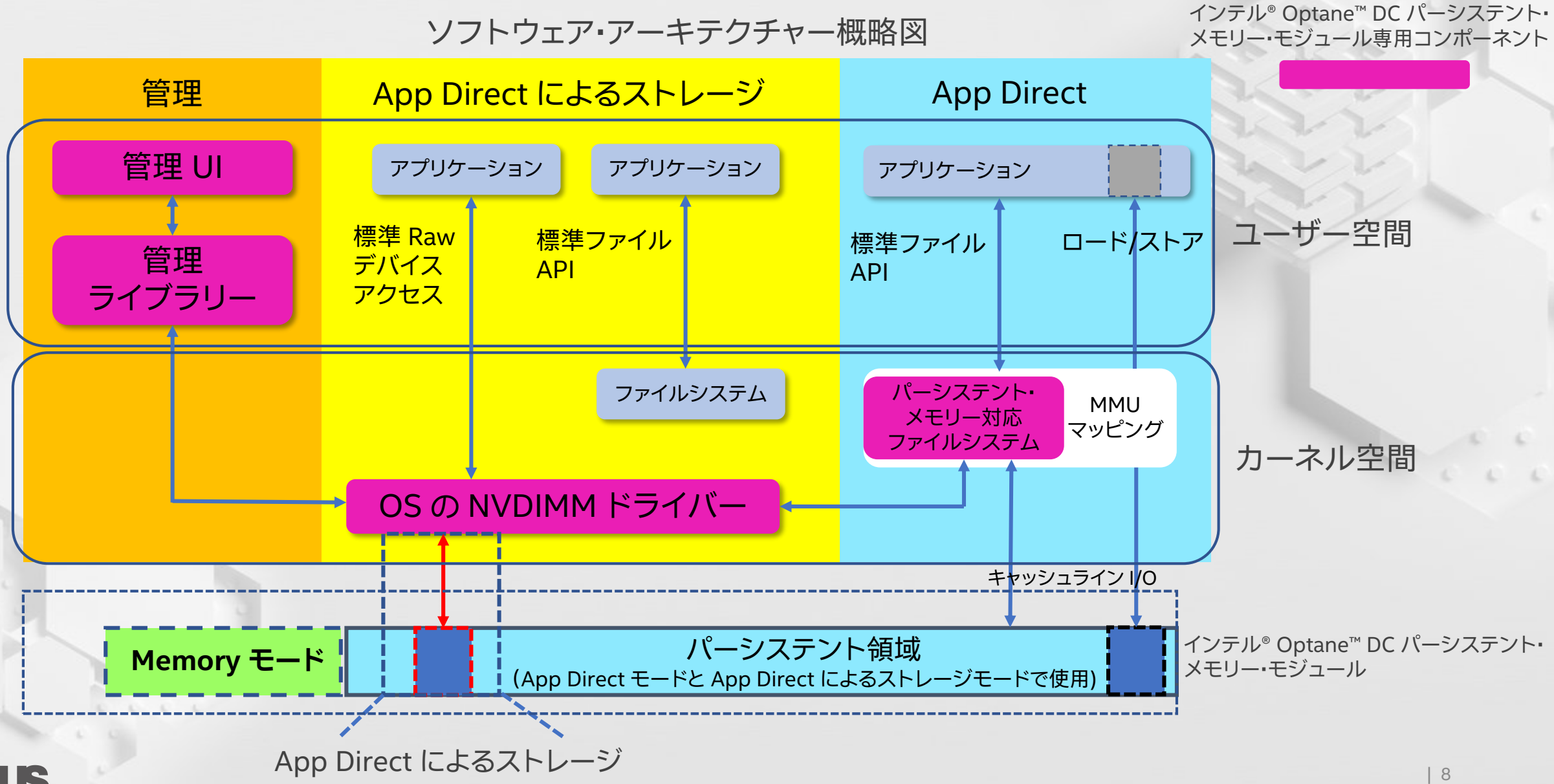
これによりロードとストア命令による 64B キャッシュライン・サイズでの直接アクセスが可能になります



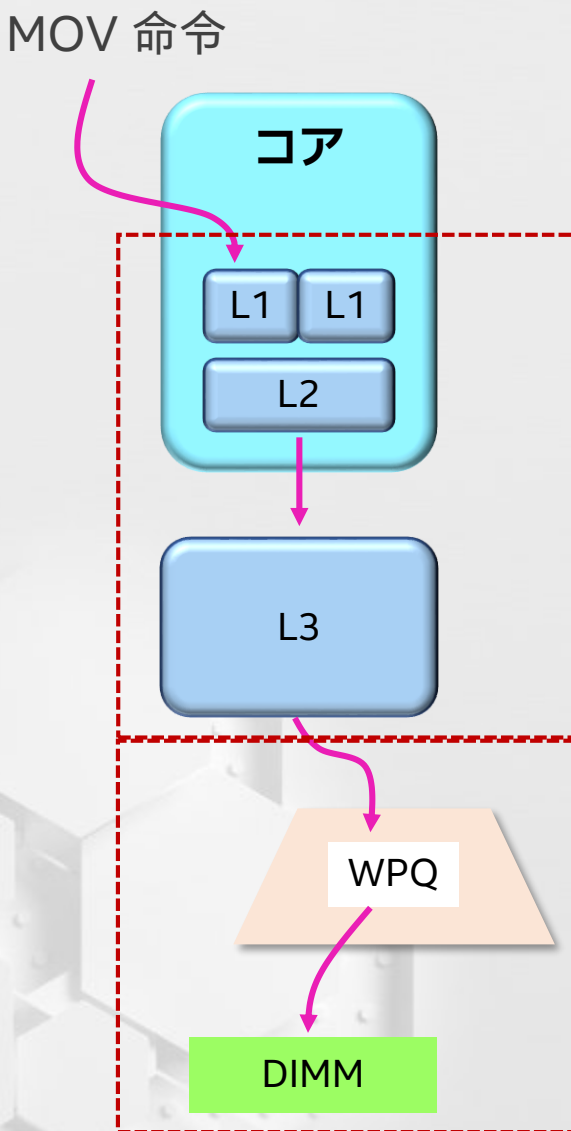
パーシステント・メモリー領域のプログラミング

ソフトウェア・アーキテクチャー

ソフトウェア・アーキテクチャー概略図



パーシステント・ドメイン



CPU キャッシュ

CLWB + フェンス
or
CLFLUSHOPT + フェンス
or
CLFLUSH
or
NT ストア + フェンス
or
WBINVD (カーネルのみ)

APCI プロパティーで示されるカスタム電力障害保護ドメイン:
CPU キャッシュ階層

WPQ

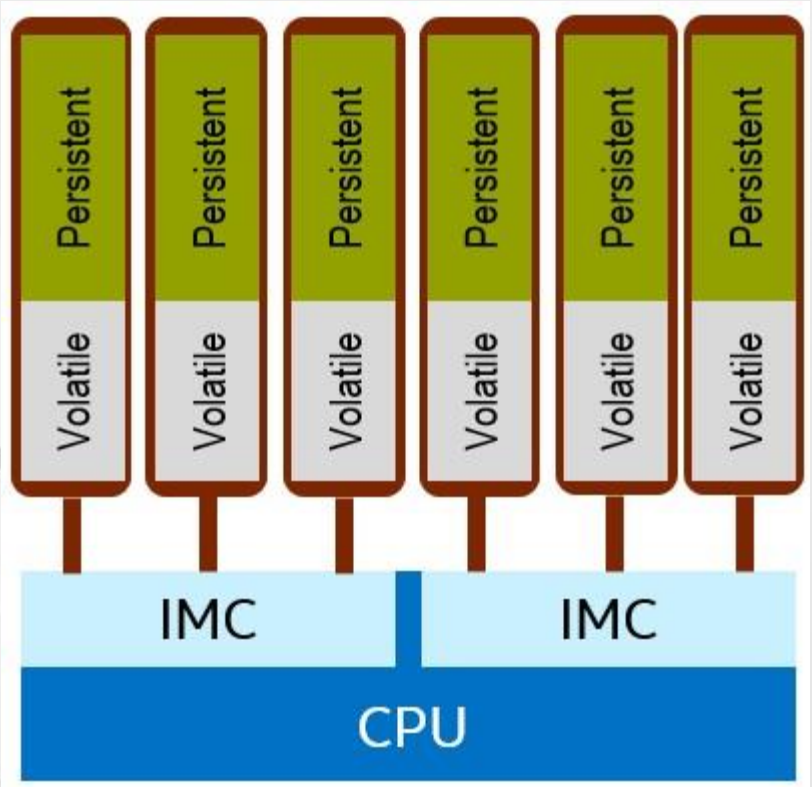
ADR
or
WPQ Flush (カーネルのみ)

最低限必要な電力障害保護ドメイン:
メモリー・サブシステム

WPQ: ライト・ペンディング・キュー

インテル® Optane™ DC パーシステント・メモリー領域と名前空間

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



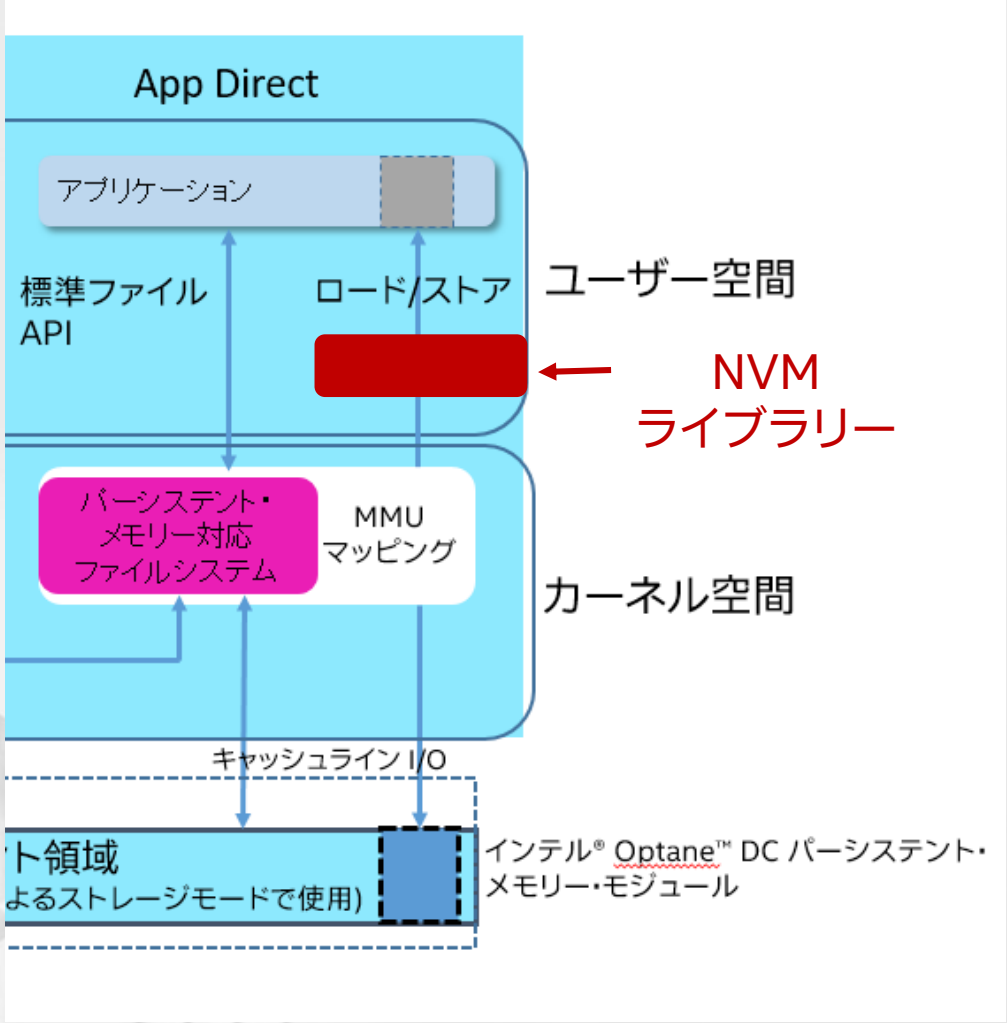
ソケット上に作成される領域は DIMM 全体で同種である必要があります。この図では、DPCMMに揮発性領域 (Volatile) と不揮発性領域 (Persistent) の 2 つがあります

- 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



- 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

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

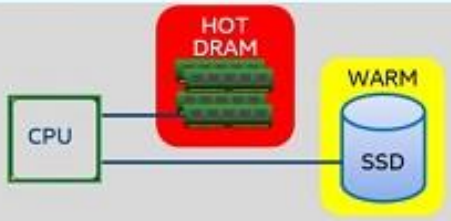
例: CLWB 命令

- 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 SFENCE の前に配置します
```

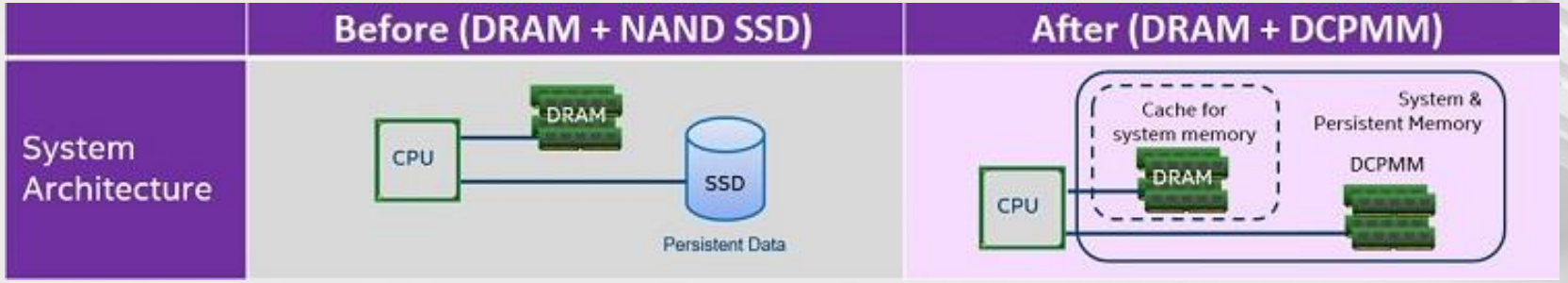
インテル® Optane™ DC パーシステント・メモリーの利用例

	Before (DRAM + NAND SSD)	After (DRAM + DCPMM)
System Architecture		
Data access	Block driver	Application managed (**App Direct**)
Data size	4KB block	64B cache line

- 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

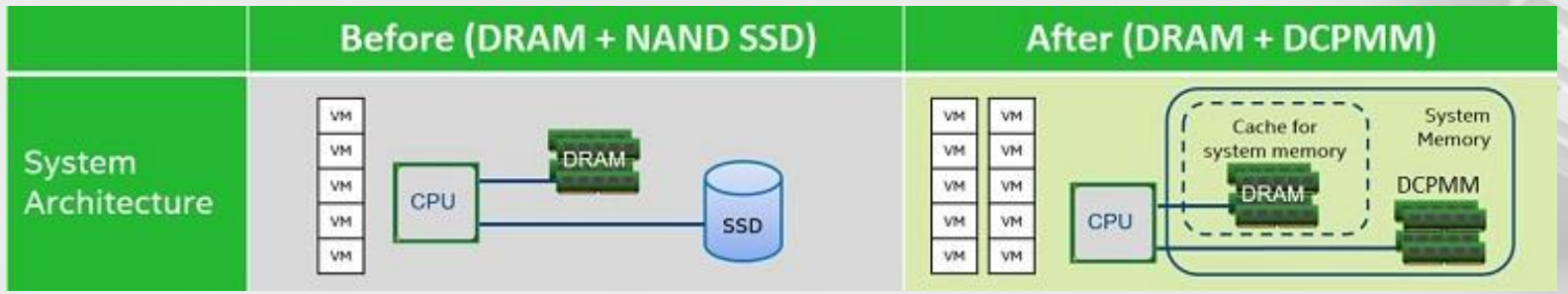
インテル® Optane™ DC パーシステント・メモリーの利用例 (2)



- 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

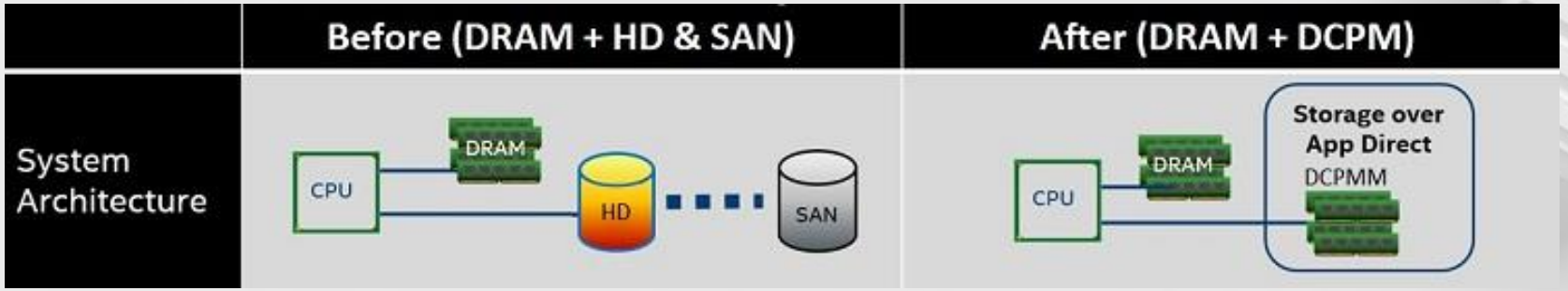
- データベース・ソフトウェアは、サービスレベルの管理要件に従ってデータの優先順位に応じて、データを DRAM に保存するかインテル® Optane™ DC パーシステント・メモリー・モジュールに保存するかを選択できます

インテル® Optane™ DC パーシステント・メモリーの利用例 (3)



- 複数の VM を利用するケースでは、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

インテル® Optane™ DC パーシステント・メモリーの利用例 (4)



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
- 障害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

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

計画から導入まで

インテル® Optane™ DC パーシステント・メモリー (DCPMM) の利点

A) DRAM より価格が安いと言われる

B) DIMM ソケットにそのまま装着できる

C) 電源を切っても内容が保持される (不揮発性) 機能を持つ

D) DRAM よりは低速であるが、NVMe SSD よりはおよそ 3 倍高速

インテル® Optane™ DC パーシステント・メモリー (DCPMM) 利用上の注意点

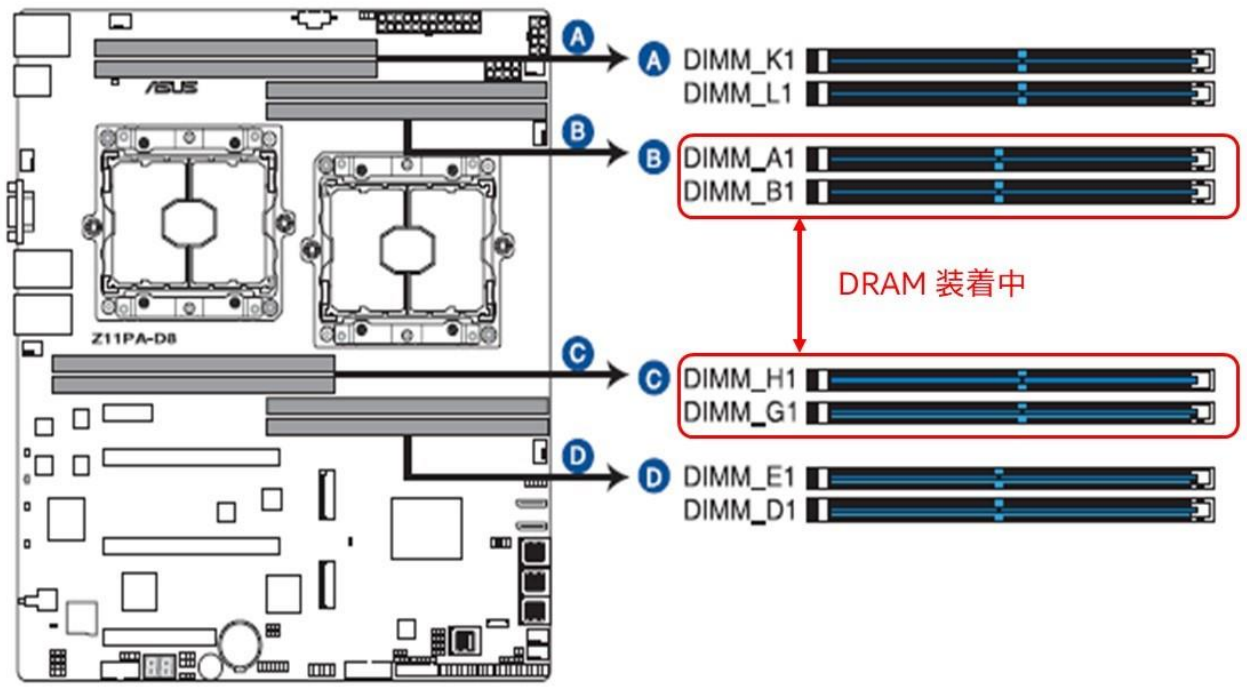
A) 第 2 世代インテル® Xeon® Gold プロセッサー以上が必要 (Silver やBronze ではサポートされない)

B) チャネルあたり 1 つの DCPMM しか装着できない (同じチャネルに DRAM が搭載されている必要がある)

C) システム BIOS とオペレーティング・システムが DCPMM に対応している必要がある

D) ユーザーは Memory モード (MM)、App Direct モード (AP) または MM + AP モードを選択する必要がある

DCPMM の装着例



K1 & L1 および E1 & D1 ソケットが空いているわけですが、ここに 4 本の DCPMM を装着してもシステムは、不明なエラーを起こし起動できません

2 CPU Configuration				
Modes	DIMM_A1	DIMM_B1	DIMM_D1	DIMM_E1
AD	DRAM	DCPMM	DRAM	DCPMM
MM	DRAM	DCPMM	DRAM	DCPMM
AD+MM	DRAM	DCPMM	DRAM	DCPMM

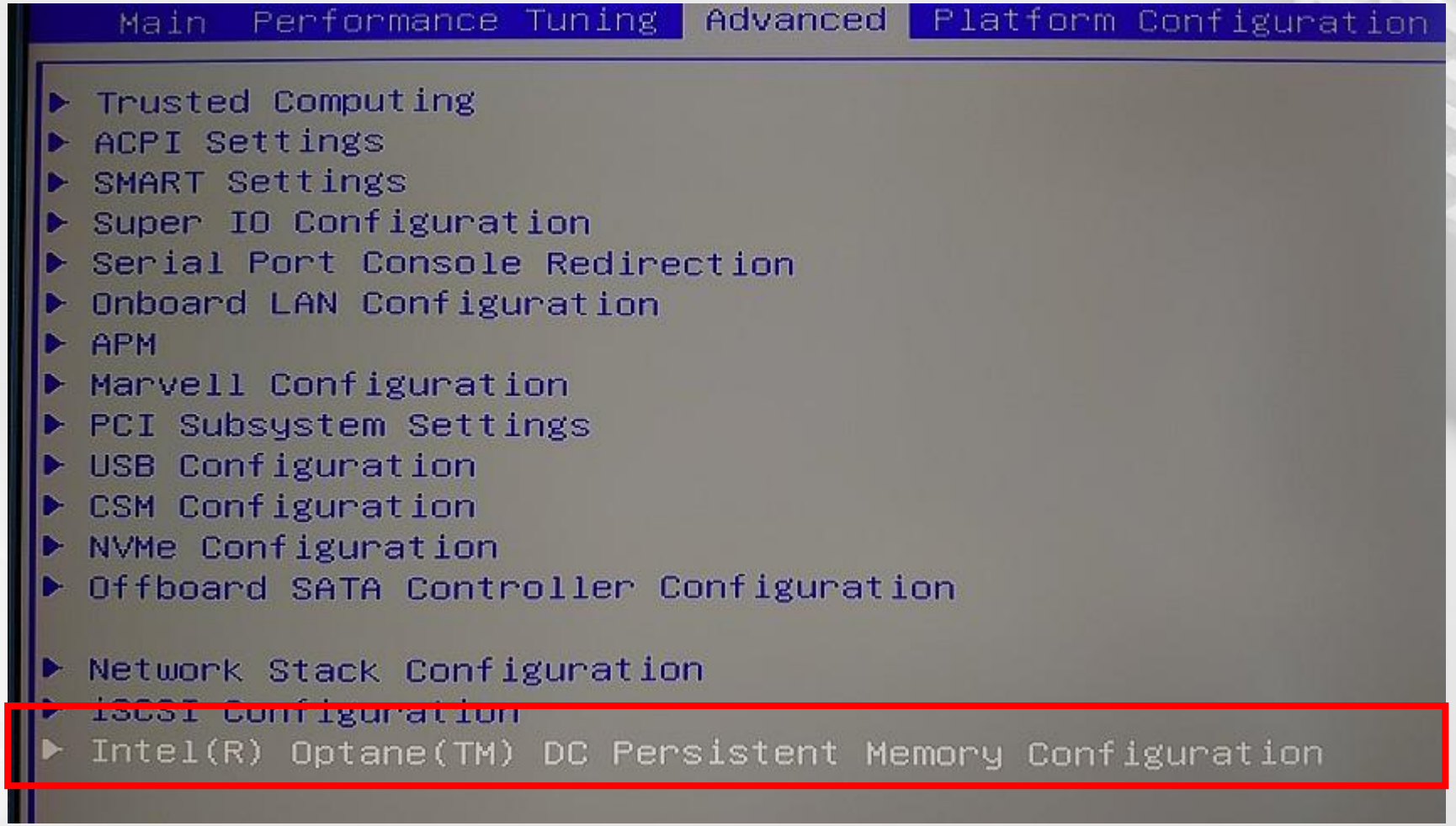
2 CPU Configuration				
Modes	DIMM_G1	DIMM_H1	DIMM_K1	DIMM_L1
AD	DRAM	DCPMM	DRAM	DCPMM
MM	DRAM	DCPMM	DRAM	DCPMM
AD+MM	DRAM	DCPMM	DRAM	DCPMM

B1 と H1 に装着されている DRAM をそれぞれ D1 と K1 に移動し、空きスロットに DCPMM を 4 本装着します。

出典: ASUS Z11PA-D8 V6 ユーザーズマニュアル

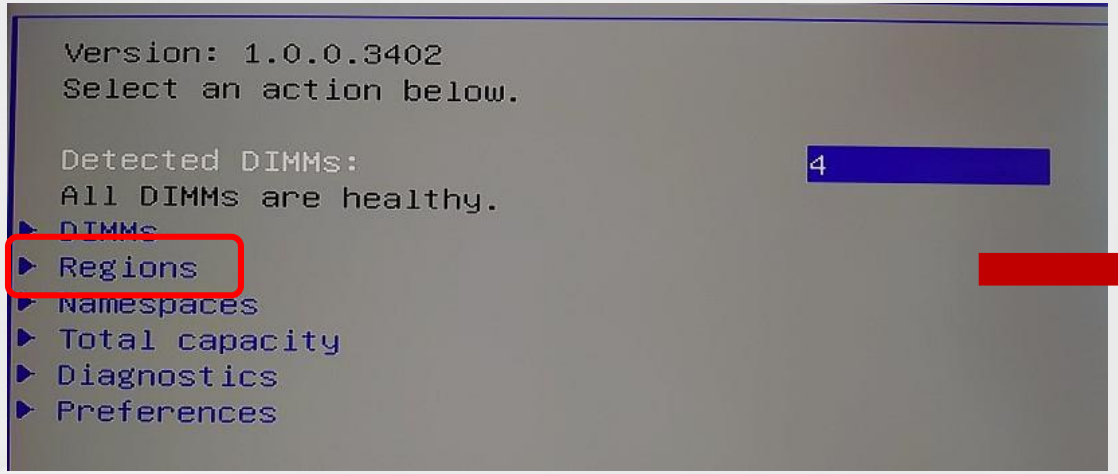
BIOS 設定

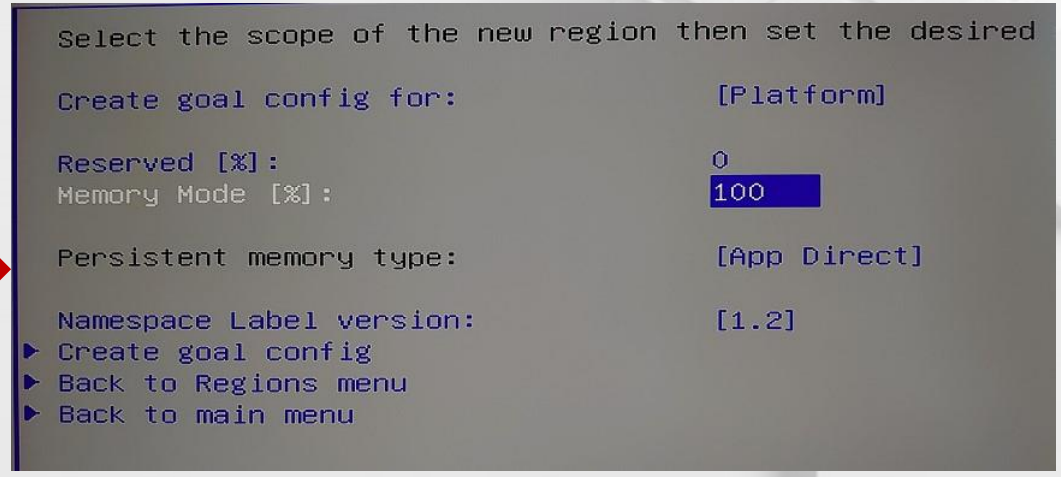
インテル® Optane™ DC パーシステント・メモリー BIOS 設定



ASUS 社のマザーボード Z11PA-D8 の例

BIOS 設定 (2)





- **[Regions]** を選択して、メモリー領域を作成
- DCPMM をすべて Memory モードで使用するには、**[Regions]** から **[Create goal config]** を選択して、**[Memory Mode [%]]** に **100** を設定します。その後 BIOS 設定を保存してシステムを再起動します

OS を起動して確認

Windows® 10 Pro と 128GB の DRAM が搭載されているシステムに、DCPMM (128GB x 4) を装着し BIOS で Memory モード 100% で設定したシステムで、リソースモニターで確認

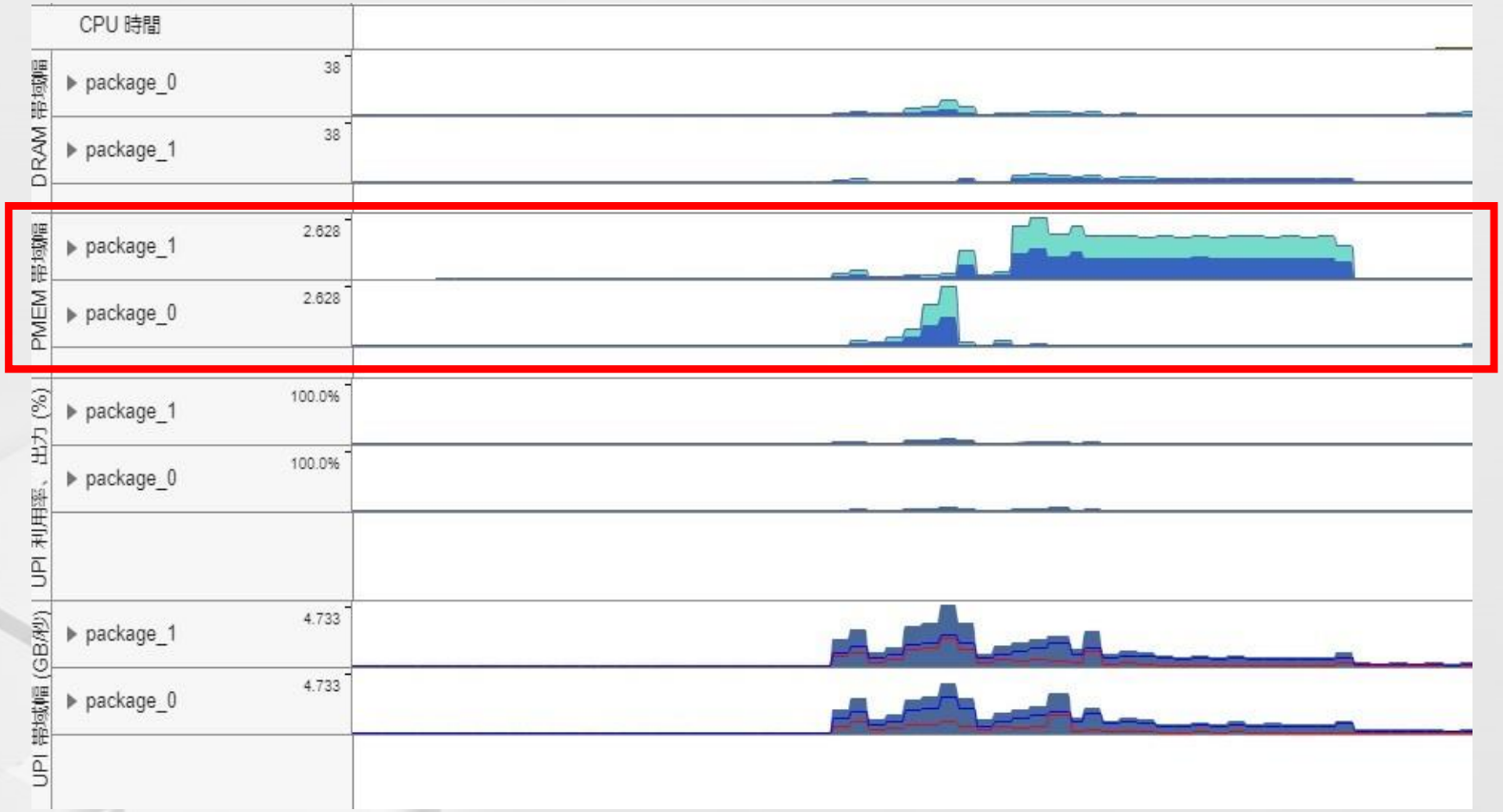
Memory モード

App Direct モード

インストール済み 648704MB で合計 515709MB と DRAM 容量が DCPMM の分だけ増加していることが分かります。DRAM 128GB はメモリープールには割り当てられず、DCPMM に対するキャッシュとして使用されます

インストール済み DIMM は 648704MB ですが、合計は 130685MB となっています。これは、DCPMM がDRAM として割り当てられない App Direct モードとして設定されており この領域を使用するには DCPMM をサポートする OS とアプリケーションが必要になります

Memory モードの動作



- Windows® 10 Pro で DCPMM を Memory モードに設定して起動後、大量の malloc を行う簡単なプログラムで 400GB 以上のメモリーを確保し、スワップが発生することなくプログラムを実行できることが確認できました
- インテル® VTune™ Amplifier 2019 を使用してプロファイルを行ったところ、PMEM (DCPMM のこと) がアクセスされていることを確認

App Direct によるストレージモード

Windows* 環境での例

不揮発性ディスクを設定

Windows* PowerShell* を管理者モードで起動して、Get-PmemUnusedRegion コマンドで DCPMM のリージョンを確認します

```
管理者: Windows PowerShell
PS C:¥users¥kiyo¥desktop¥optane> Get-PmemUnusedRegion

RegionId TotalSizeInBytes DeviceId
-------- ---------------- --------
       1     270582939648 {10, 110}
       3     270582939648 {1010, 1110}
```

ソケット 1 と 2 に装着されている 2 枚ずつの DCPMM は、それぞれリージョン ID 1 と 3 で認識されています。Get-PmemUnusedRegion 1 | New-PmemDisk コマンドで永続メモリーをディスクとして作成します。正常に作成されるとプロンプトに返ります

```
管理者: Windows PowerShell
PS C:¥users¥kiyo¥desktop¥optane> Get-PmemUnusedRegion 1 | New-PmemDisk
新しい永続メモリ ディスクを作成しています。これには数分かかる場合があります。
PS C:¥users¥kiyo¥desktop¥optane>
```

Windows* 環境での例 (2)

Get-PmemUnusedRegion コマンドで DCPMM のリージョンを再度確認すると、リージョン 1 がすでに未使用ではないことが分かります

```
管理者: Windows PowerShell
PS C:¥users¥kiyo¥desktop¥optane> Get-PmemUnusedRegion

RegionId TotalSizeInBytes DeviceId
-------- ---------------- --------
       3     270582939648 {1010, 1110}
```

このままではまだディスクとして利用することはできません。Get-PmemDisk コマンドでディスク番号を確認します。ここではディスク番号は 4 となっています

```
管理者: Windows PowerShell
PS C:¥users¥kiyo¥desktop¥optane> Get-PmemDisk

DiskNumber Size   HealthStatus AtomicityType CanBeRemoved PhysicalDeviceIds UnsafeShutdownCount
---------- ----   ------------ ------------- ------------ ----------------- -------------------
4          252 GB Healthy      None          True         {10, 110}         0
```

Windows* 環境での例 (3)

Initialize-Disk コマンドでディスク番号 4 を初期化します。これでディスクとして利用できる状態になりましたが、パーティションの作成とドライブレターの割り当ては行われていません

```
管理者: Windows PowerShell
PS C:¥users¥kiyo¥desktop¥optane> Initialize-Disk -Number 4
PS C:¥users¥kiyo¥desktop¥optane> _
```

NewPartiton -Disknumber 4 -UseMaximumSaize -DriveLetter V コマンドで、パーティションを作成して V ドライブに割り当てます

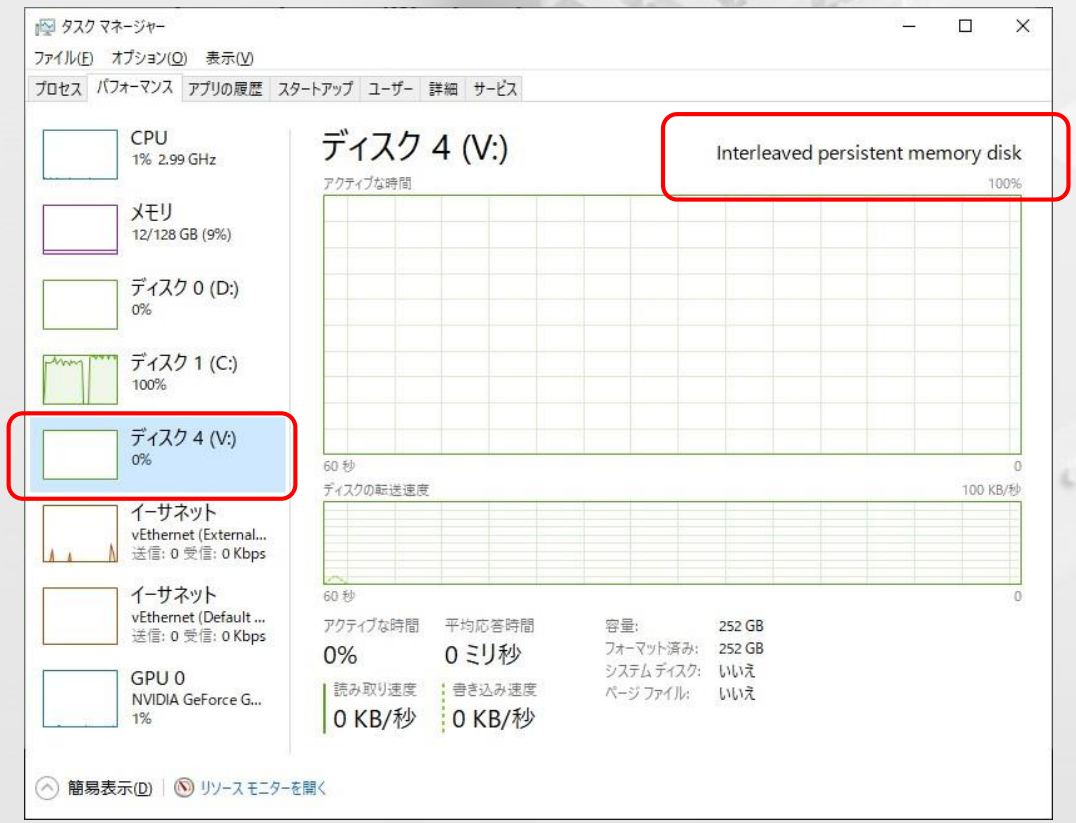
```
管理者: Windows PowerShell
PS C:¥users¥kiyo¥desktop¥optane> New-Partition -Disknumber 4 -UseMaximumSize -DriveLetter V


   DiskPath: ¥¥?¥scmld#ven_8980&dev_097a&subsys_89804151&rev_0018#3&1b1819f6&0&0301808916d7c86625dbc9478178da74517d97a5
#{53f56307-b6bf-11d0-94f2-00a0c91efb8b}

PartitionNumber  DriveLetter Offset                                        Size Type
---------------  ----------- ------                                        ---- ----
2                V           16777216                                 251.98 GB Basic
```

Windows* 環境での例 (4)

最後に Format コマンドでパーティションをフォーマットします

管理者: Windows PowerShell

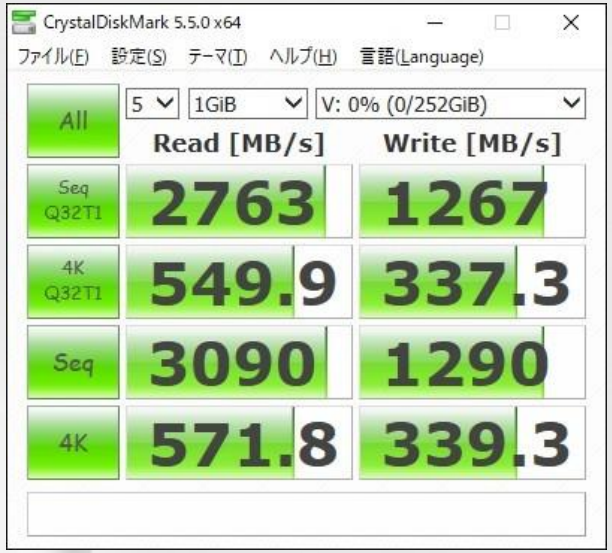
```
PS C:¥users¥kiyo¥desktop¥optane> format /q /y /dax V:
ファイル システムの種類は RAW です。
新しいファイル システムは NTFS です。
クイック フォーマットしています  252.0 GB
ファイル システム構造を作成します。
フォーマットは完了しました。
    252.0 GB: 全ディスク領域
    251.9 GB: 使用可能領域
```

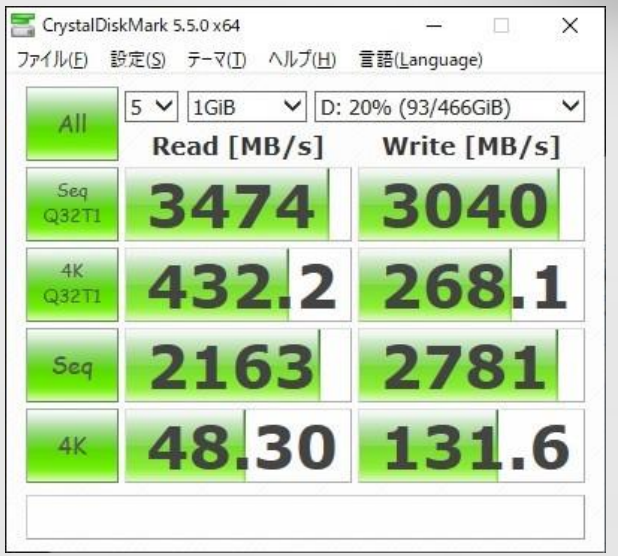


これで、V ドライブとして DCPMM は不揮発性ディスクとして利用できるようになりました。タスクマネージャーでは、Interleaved persistent memory disk として見えています

Windows* 環境での例 – パフォーマンス測定

CristalDiskMark を使用して NVMe SSD と DCPMM のパフォーマンスを比較してみます。比較する SSD は、シリコンパワー NVMe 1.3 NAND M.2 SSD (P34A80 シリーズ) 1TB です。コントローラは Phison PS5012-E12 を搭載し、シーケンシャル読み込み速度は3200MB/秒、シーケンシャル書き込み速度は3000MB/秒のカタログ値を持ちます





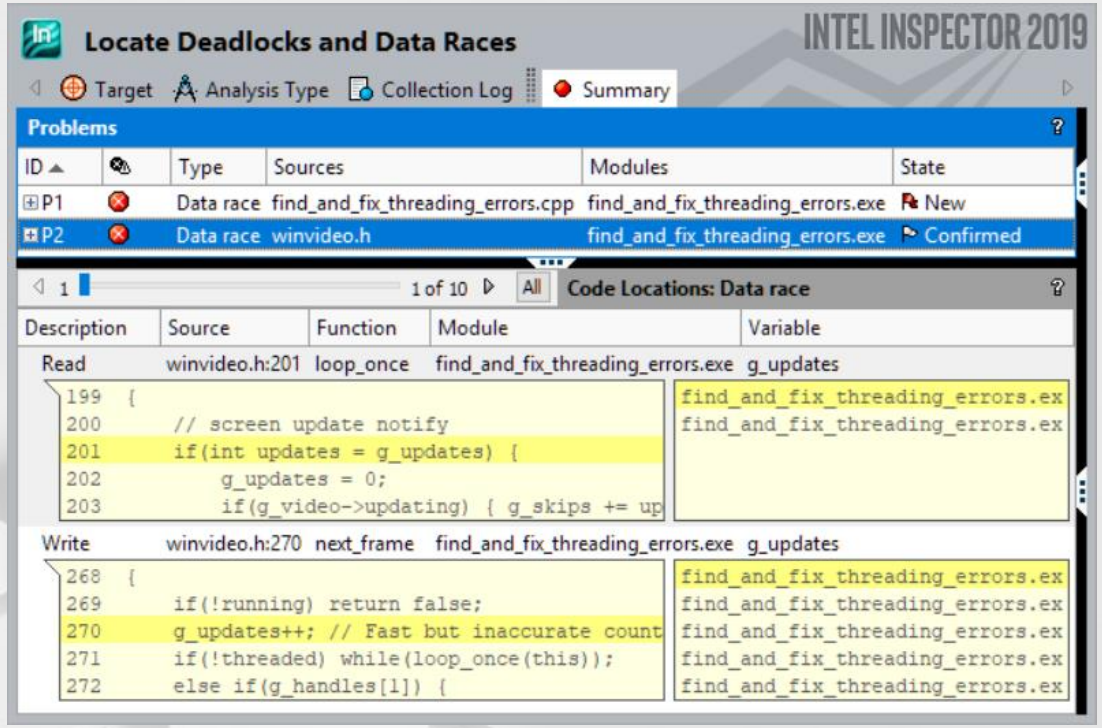
左が DCPMM で右がシリコンパワー P34A80 です。Seq Q32T1 では、P34A80 が上回っていますが、ブロックサイズが小さくなると DCPMM のパフォーマンスが高い傾向にあることが分かります

App Direct モードの例

- インテル® Optane™ DC パーシステント・メモリー導入への道: その 3 – 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 (簡単な grep の例)
- Java* によるインテル® Optane™ DC パーシステント・メモリーのサポート
- サンプルコード: 不揮発性メモリーキャッシュの実装 – 簡単な検索の例
- 不揮発性メモリーのアウトオブオーダー解析にインテル® Inspector を使用する

インテル® Inspector によるメモリーとスレッドのデバッグ

メモリーリーク、異常、データ競合、デッドロックの特定とデバッグ



正当性検証ツールにより ROI が 12%-21%[1] 向上

- 早期に問題を発見したほうが修正コストが少なくて済む
- 競合やデッドロックは簡単に再現できない
- メモリーエラーをツールなしで発見するのは困難

デバッガー統合により迅速な診断が可能

- 問題の直前にブレークポイントを設定
- デバッガーで変数とスレッドを確認

バージョン 2019 の新機能

不揮発性メモリーエラーを発見

- 不足している/冗長なキャッシュフラッシュ
- ストアフェンスの不足
- アウトオブオーダーの不揮発性メモリーストア
- PMDK トランザクションの Redo (やり直し) ログエラー

詳細: isus.jp/intel-inspector-xe/

[1]コスト要因 - Square Project による分析
CERT: U.S. Computer Emergency Readiness Team および Carnegie Mellon CyLab
NIST: National Institute of Standards & Technology : Square Project の結果

不揮発性メモリーの利用例

- DCPMM にあるデータは、バイトアドレス指定可能で、システムやプログラムがクラッシュしても保持されます。DCPMM のアクセス・レイテンシーは、DRAM に匹敵します
- プログラムは、通常の CPU ロード/ストア命令を使用して DCPMM を読み書きします
  次のコード例について考えてみます

```
struct address {
    char name[64];
    char address[64];
    int valid;
}
```

```
int main()
{
    struct address *head = NULL;
    int fd;
    fd = open("addressbook.pmem", O_CREAT|O_RDWR, 0666);
    posix_fallocate(fd, 0, sizeof(struct address));

    head = (struct address *)mmap(NULL, sizeof(struct address), PROT_READ|PROT_WRITE,
      MAP_SHARED, fd, 0);
    close(fd);

    strcpy(head->name, "Clark Kent");
    strcpy(head->address, "344 Clinton St, Metropolis, DC 95308");
    head->valid = 1;

    munmap(head, sizeof(struct address));

    return 0;
}
```

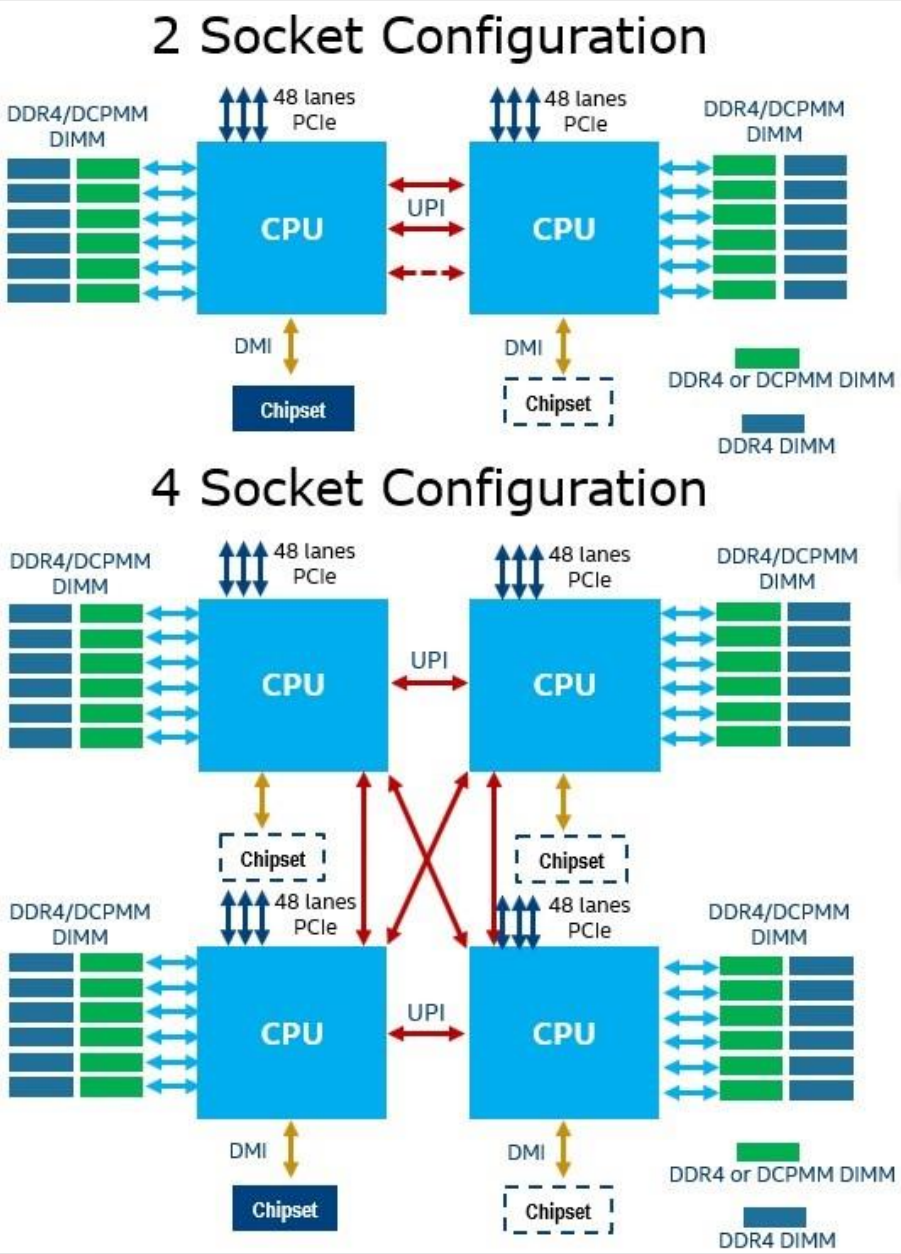
munmap を呼び出す前に電力が失われた場合、

- head->name、head->address、および head->valid のいずれも、メモリーシステムに到達せず、永続化されない
- 3 つすべてが永続化される
- 1 つまたは 2 つのみが永続化される

まとめ

- インテル® Optane™ DC パーシステント・メモリーは、データセンターでの利用に特化して設計された新しい種類のメモリーおよびストレージ・テクノロジーです
- この製品は、大容量、低価格、および永続性というこれまでにない組み合わせをもたらします
- 手ごろな価格でシステムメモリーの容量を拡張できるため、利用者はこの新しいメモリーを搭載するシステムを使用して、プロセッサーに近いメモリー階層に大量のデータを移動または保持し、システムストレージからデータを取り込むレイテンシーを最小限に抑えることで、ワークロードを最適化できます

参考資料


2 Socket Configuration
DDR4/DCPMM DIMM
48 lanes PCIe
CPU
UPI
CPU
48 lanes PCIe
DDR4/DCPMM DIMM
DMI
Chipset
DMI
Chipset
DDR4 or DCPMM DIMM
DDR4 DIMM
4 Socket Configuration
DDR4/DCPMM DIMM
48 lanes PCIe
CPU
UPI
CPU
48 lanes PCIe
DDR4/DCPMM DIMM
Chipset
Chipset
DDR4/DCPMM DIMM
48 lanes PCIe
CPU
UPI
CPU
48 lanes PCIe
DDR4/DCPMM DIMM
DMI
Chipset
DMI
Chipset
DDR4 or DCPMM DIMM
DDR4 DIMM

iSUS

ソフトウェア・セミナー