

## 記事

[Toshihiko Minamoto](#) · 2021年7月15日 16m read

## InterSystems IRIS および Intel Optane DC パーシステントメモリ

InterSystems および Intel は先日、InterSystems IRIS を「Cascade Lake」としても知られる第 2 世代 Intel® Xeon® スケーラブルプロセッサおよび Intel® Optane™ DC

パーシステントメモリ (DCPMM) と組み合わせて一連のベンチマークを実施しました。

さまざまなワークロード設定とサーバー構成で、Intel の最新のサーバーテクノロジーを使用した InterSystems IRIS のパフォーマンスとスケーラビリティ機能を実証するのがこのベンチマークの目的です。

このレポートには、さまざまなベンチマークの結果とともに、Intel DCPMM と InterSystems IRIS のユースケースが 3 つ示されています。

### 概要

パフォーマンスとスケーリングを実証するために 2

種類のワークロードが使用されています。読み取り集中型のワークロードと書き込み集中型のワークロードです。このように分けて実証するのは、読み取り集中型ワークロードにおけるデータベースキャッシュ効率の増加と、書き込み集中型ワークロードにおけるトランザクションジャーナルの書き込みスループットの増加のそれぞれに特化したユースケースで、Intel DCPMM の影響を示すためです。

両方のユースケースシナリオにおいて、InterSystems IRIS

のスループット、スケーラビリティ、およびパフォーマンスの大幅なゲインが達成されています。

- **読み取り集中型ワークロード**では、4 ソケットサーバーと、合計約 1.2 TB のデータを持つデータセットを使用する大量の長期実行分析クエリが使用されました。DCPMM を「Memory Mode」で使用した場合のベンチマーク比較では、メモリの少ない前世代の Intel E7v4 シリーズプロセッサと比べた場合、経過実行時間が大幅に短縮され、およそ 6 倍高速になりました。E7v4 と、DCPMM を使った最新のサーバーを同じメモリサイズで比較した場合は、20% の改善が見られました。これは、DCPMM による InterSystems IRIS データベースキャッシュ機能の向上と最新の Intel プロセッサアーキテクチャによるものです。
- **書き込み集中型ワークロード**では、2 ソケットサーバーと InterSystems HL7 メッセージングのベンチマークが使用されました。多数のインバウンドインターフェースで構成されており、各メッセージには複数の変換が伴い、インバウンドメッセージごとに 4 つのアウトバウンドメッセージが使用されています。高スループットを維持する上で重要なコンポーネントの 1 つは、IRIS for Health のメッセージ耐久性保証で、その操作においては、トランザクションのジャーナル書き込みのパフォーマンスが重要となります。「APP DIRECT」モードで DCPMM を使用して、DAX XFS でトランザクションのジャーナル用の XFS ファイルシステムを提供した場合、このベンチマークのメッセージスループットには 60% の向上が示されました。

テスト結果と構成を要約すると、DCPMM は適切に設定された InterSystems IRIS

とワークロードで使用された場合にスループットを大幅に向上させることができます。

高レベルのメリットとしては、読み取り集中型ワークロードではデータベースのキャッシュ効率の向上とディスク I/O ブロック読み取りの抑制、書き込み集中型ワークロードではジャーナルの書き込みスループットの向上を得られます。

さらに、古いハードウェアを更新し、パフォーマンスとスケーリングの改善を検討しているユーザーにとって、DCPMM を備えた Cascade Lake に基づくサーバーは優れた更新パスとなります。InterSystems のテクノロジーアーキテクトと相談しながら、既存のワークロードに推奨される構成についてのアドバイスを得ることができます。

## 読み取り集中型ワークロードベンチマーク

読み取り集中型ワークロードでは、512 GiB と 2 TiB のデータベースキャッシュサイズの E7v4 (Broadwell) と Intel® Optane™ DC パーシステントメモリ (DCPMM) を使用した 1 TB と 2 TB のデータベースキャッシュサイズの最新の第 2 世代 Intel® Xeon® スケーラブルプロセッサ (Cascade Lake) と比較する分析クエリベンチマークを使用しました。

より大規模なキャッシュの影響とパフォーマンスの向上を示すために、さまざまなグローバルバッファサイズで複数のワークロードを実行しました。構成を反復するごとに、COLD と WARM で実行しています。COLD は、データベースキャッシュにデータが事前に入力されていない場合で、WARM は、データベースキャッシュがすでにアクティブになっており、ディスクからの物理的な読み取りを減らすために、データが入力済みである (または少なくとも可能な限り入力されている) ことを示します。

### ハードウェア構成

古い 4 ソケット E7v4 (Broadwell) ホストを DCPMM を使った 4 ソケット Cascade Lake サーバーと比較しました。この比較が選択されたのは、InterSystems IRIS を使ってハードウェアの更新を検討している既存のお客様がパフォーマンスの向上を得られることを実証するためです。バージョン間のソフトウェアの最適化が要因とならないように、すべてのテストには同じバージョンの InterSystems IRIS が使用されました。

ディスクパフォーマンスが比較の要因とならないように、すべてのサーバーには同一のストレージレイにある同一のストレージが使用されています。ワーキングセットは 1.2 TB のデータベースです。図 1 にはこのハードウェア構成と、それぞれの 4 ソケット構成の比較が示されています。

#### 図 1: ハードウェア構成

サーバー #1 の構成

プロセッサ: 4 x E7-8890 v4 @ 2.5Ghz

メモリ: 2TiB DRAM

ストレージ: 16Gbps FC all-flash SAN @ 2TiB

サーバー #2 の構成

プロセッサ: 4 x Platinum 8280L @ 2.6Ghz

メモリ: 3TiB DCPMM + 768GiB DRAM

ストレージ: 16Gbps FC all-flash SAN @ TiB

DCPMM: Memory Mode のみ

### ベンチマークの結果と結論

512 GiB を 1 TiB か 2 TiB のいずれかの DCPMM

バッファプールサイズと比較した場合、経過実行時間に大幅な短縮が見られます (約 6 分の 1)。さらに、2 TiB E7v4 DRAM と 2 TiB Cascade Lake DCPMM の構成を比較した場合には約 20% の改善も見られました。

バッファプールのサイズが同じであるとした場合、この 20%

の増加は、ほぼ新しいプロセッサのアーキテクチャとプロセッサのコア数の増加によるものだと考えられますが、

それでも、テストされた 4 ソケット Cascade Lake にインストールされていたが 24 x 128 GiB DCPMM

のみであることを考慮すると深い意義があります。DCPMM は 12 TiB

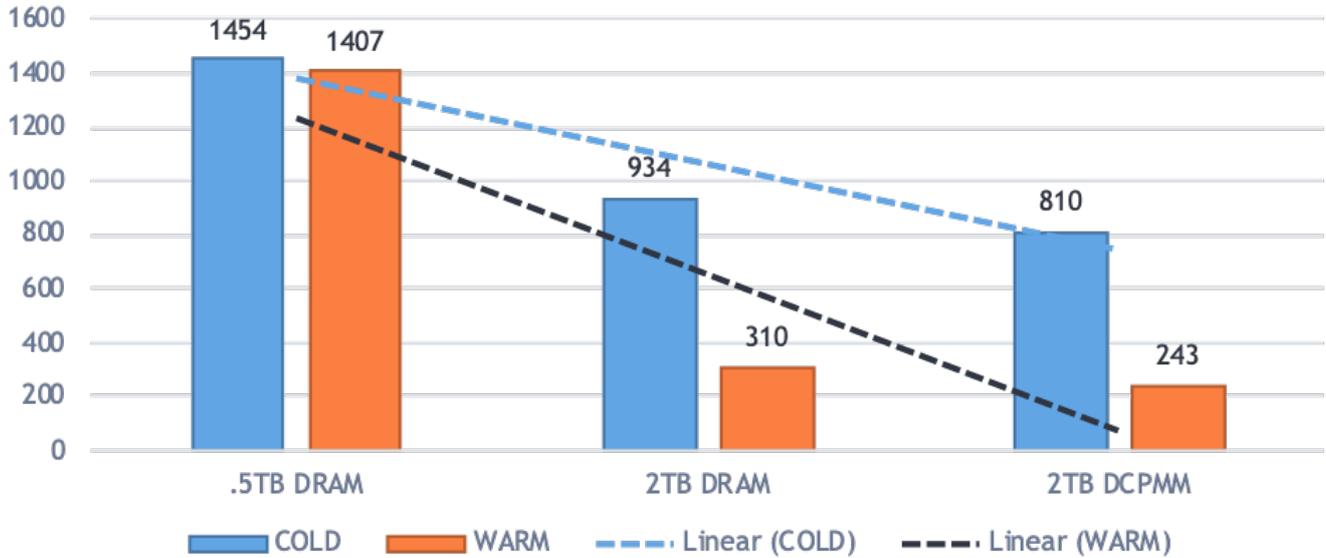
までスケールアップすることが可能であり、これは E7v4 が同じ 4

ソケットサーバーのフットプリントでサポートできるメモリの約 4 倍のメモリです。

以下の図 2 に示されるグラフは、この比較の結果を示しています。両方のグラフの y 軸は経過時間 (値が小さくなるほど良) で、さまざまな構成で得た結果が比較されています。

#### 図 2: 各種構成の経過時間の比較

## Comparing E7v4 .5TB & 2TB DRAM to 2TB Cascade Lake DCPMM (Memory Mode)



### 書き込み集中型ワークロードベンチマーク

このベンチマークのワークロードは、すべての T4 タイプのワークロードを使用した HL7v2 メッセージングワークロードです。

- T4 ワークロード**は、ルーティングエンジンを使って、個別に変更されたメッセージを 4 つの各アウトバウンドインターフェースにルーティングしました。平均して、インバウンドメッセージの 4 つのセグメントが変換ごとに変更されました (4 つの変換で 1 件につき 4 件)。各インバウンドメッセージでは、4 つのデータ変換の実行により 4 つのメッセージがアウトバウンドに送信され、5 つの HL7 メッセージオブジェクトがデータベースに作成されました。

各システムは 128 個のインバウンドビジネスサービスと各インバウンドインターフェースに送信される 4800 件のメッセージ (インバウンドメッセージ合計 614,400 件、アウトバウンドメッセージ合計 2,457,600 件) で構成されています。

このベンチマークワークロードのスループットの単位は「1秒あたりのメッセージ数」です。トランザクションジャーナルのスループットとレイテンシは高スループットを維持する上で重要なコンポーネントであるため、ベンチマーク実行中のジャーナルの書き込みにも関心があります (記録されています)。IRIS for Health のメッセージ耐久性保証のパフォーマンスに直接影響を与えるため、その操作において、トランザクションジャーナルの書き込みパフォーマンスが重要となります。ジャーナルのスループットが低下すると、アプリケーションプロセスによってジャーナルバッファの可用性が阻止されてしまいます。

### ハードウェア構成

書き込み集中型ワークロードでは、2 ソケットサーバーを使用することにしました。192 GB の DRAM と 1.5 TiB の DCPMM しかないため、この構成は前述の 4 ソケット構成よりも小さくなります。DCPMM を使用した Cascade Lake のワークロードを以前の初代 Intel® Xeon® スケーラブルプロセッサ (Skylake) サーバーに比較しました。両サーバーには 750GiB Intel® Optane™ SSD DC P4800X ドライブがローカル接続されています。

図 3 にはこのハードウェア構成と、それぞれの 2 ソケット構成の比較が示されています。

図 3: 書き込み集中型ワークロードのハードウェア構成

サーバー #1 の構成

プロセッサ: 2 x Gold 6152 @ 2.1Ghz

サーバー #2 の構成

プロセッサ: 2 x Gold 6252 @ 2.1Ghz

サーバー #1 の構成  
メモリ: 192GiB DRAM  
ストレージ: 2 x 750GiB P4800X Optane SSD

サーバー #2 の構成  
メモリ: 1.5TiB DCPMM + 192GiB DRAM  
ストレージ: 2 x 750GiB P4800X Optane SSD  
DCPMM: Memory Mode & App Direct モード

## ベンチマークの結果と結論

**テスト 1:** このテストでは、図 3 のサーバー #1 構成に示される Skylake サーバーにおいて前述の T4 **ワークロード** を実行しました。Skylake サーバーは、2010 ジャーナル書き込み/秒のジャーナルファイル書き込み速度で約 3355 件のインバウンドメッセージの持続的なスループットを示しました。

**テスト 2:** このテストでは、図 3 のサーバー #2 構成に示される Cascade Lake サーバーにおいて、DCPMM の Memory Mode を指定して同じワークロードを実行しました。これは、2400 ジャーナル書き込み/秒のジャーナルファイル書き込み速度で約 4684 件のインバウンドメッセージの持続的なスループットという大幅な向上を示しました。これは、**テスト 1 に比較すると 39% の増加です。**

**テスト 3:** このテストでは、図 3 のサーバー #2 構成に示される Cascade Lake サーバーにおいて同じワークロードを実行しましたが、今度は DCPMM を App Direct Mode に指定し、DCPMM による何らかの実行を構成せずに実行しました。このテストの目的は、DRAM のみを使用した Cascade Lake のパフォーマンスとスループットを DCPMM と DRAM を使用した Cascade Lake に比較して測定することです。DCPMM が使用されていない場合でもスループットは（比較的小さいとは言え）向上したという、特に驚くことでもない結果が出ました。これは、2540 ジャーナル書き込み/秒のジャーナルファイル書き込み速度で約 4845 件のインバウンドメッセージの持続的なスループットという向上を示しました。DCPMM は DRAM に比べてより高いレイテンシがあり、更新が大量に流入すればパフォーマンスが低下するため、予想された動作と言えます。別の見方をすると、まったく同じサーバーで DCPMM を Memory Mode で使用する場合、書き込みの取り込みワークロードに 5% 未満の低下があることとなります。また、Skylake を Cascade Lake (DRAM のみ) に比較した場合、**テスト 1 の Skylake サーバーに比べて 44% の増加が得られています。**

**テスト 4:** このテストでは、図 3 のサーバー #2 構成に示される Cascade Lake サーバーにおいて同じワークロードを実行しましたが、今度は DCPMM を App Direct Mode に指定し、App Direct Mode をジャーナルファイルシステム用にマウントされた DAX XFS として使用して実行しました。これは 2630/秒のジャーナルファイル書き込み速度で 1 秒あたり 5399 件のインバウンドメッセージというさらに高いスループットを示しました。この種のワークロードでは App Direct Mode の DCPMM がより適した DCPMM の使用方法であることが示されています。これらの結果を最初の Skylake 構成と比較すると、**テスト 1 の Skylake サーバーに比べ、スループットが 60% 増加しています。**

## InterSystems IRIS の推奨される Intel DCPMM ユースケース

Intel® Optane™ DC パーシステントメモリを使用することで InterSystems IRIS にメリットが与えられるユースケースと構成にはいくつかあります。

### Memory Mode

このモードは、単一の InterSystems IRIS デプロイや大規模な InterSystems IRIS シャードクラスタでの膨大なデータベースキャッシュに最適です。後者の環境ではより多く（またはすべて）のデータベースをメモリにキャッシュできます。DCPMM と DRAM の比率は最大 8:1 にすることをお勧めします。「ホットメモリ」を L4 キャッシュレイヤーとして機能する DRAM に保持する際に重要です。これは、リソース占有やその他のメモリキャッシュラインなど、共有内部 IRIS メモリ構造において特に重要となります。

### App Direct Mode (DAX XFS) – ジャーナルディスクデバイス

このモードは、DCPMM をトランザクションジャーナルファイルのディスクデバイスとして使用する場合に最適です。DCPMM は Linux

にマウントされた XFS ファイルシステムとしてオペレーティングシステムに表示されます。DAX XFS を使用するメリットは、これによって PCIe バスのオーバーヘッドとファイルシステムからのダイレクトメモリアクセスが緩和されることにあります。HL7v2 ベンチマークの結果に示されるように、書き込みレイテンシによって HL7 メッセージングのスループットは大幅に向上します。また、ストレージには従来のディスクデバイスと同様に、再起動や電源サイクル時における永続性と耐久性が備わっています。

## App Direct Mode (DAX XFS) – ジャーナル + 書き込みイメージジャーナル (WIJ) ディスクデバイス

このユースケースでは、App Direct モードの用法がトランザクションジャーナルと書き込みイメージジャーナル (WIJ) の両方に拡張されます。両ファイルは書き込み集中型であるため、超低レイテンシと永続性のメリットを確実に得られます。

### Dual Mode: Memory + App Direct Modes

DCPMM を Dual Mode で使用すると DCPMM のメリットが拡大し、トランザクションジャーナルや書き込みイメージジャーナルデバイスで大規模なデータベースキャッシュと超低レイテンシを実現できるようになります。このユースケースでは、DCPMM は OS にマウントされた XFS ファイルシステムとオペレーティングシステムの RAM として表示されます。これは、DCPMM の一定の割合を DAX XFS に割り当て、残りを Memory Mode に割り当てることで可能です。前述のように、インストールされている DRAM はプロセッサの L4 のようなキャッシュとして機能します。

### 「疑似」Dual Mode

疑似 Dual Mode 寄りにユースケースモデルを拡張するために、OLTPタイプのワークロード用と分析または大規模なクエリーニーズ用に高速のインバウンドトランザクションと更新が伴うトランザクションと分析の並行ワークロード (HTAP ワークロードとしても知られています) タイプのワークロードがあり、さらに [InterSystems IRIS 共有クラスタ](#)内ではそれぞれの InterSystems IRIS ノードタイプが DCPMM のさまざまなモードで稼働しています。

この例では、グローバルバッファの大規模データベースキャッシュと、トランザクションワークロード用に DAX XFS として Dual Mode または App Direct のいずれかで実行する [データノード](#)のメリットを得られるように、DCPMM Memory Mode で実行する大規模なクエリ/分析ワークロードを処理する InterSystems IRIS [計算ノード](#)が追加されています。

## まとめ

インフラストラクチャの選択に関して言えば、InterSystems IRIS には多数のオプションが提供されています。インフラストラクチャの要件はアプリケーション、ワークロードプロファイル、およびビジネスニーズによって決まり、これらのテクノロジーとインフラストラクチャの選択が、ビジネスにおけるアプリケーションの成功、採用、および重要性を左右します。第 2 世代 Intel® Xeon® スケーラブルプロセッサと Intel® Optane™ DC パーシステントメモリを使用した InterSystems IRIS は、ビジネスに大きな影響を与える InterSystems IRIS ベースアプリケーションに画期的なスケーリング性能とスループット性能を与えることができます。

InterSystems IRIS と Intel DCPMM 対応サーバーには、次のようなメリットがあります。

- Memory Mode の DCPMM を使用した InterSystems IRIS または InterSystems IRIS for Health データベースキャッシュにマルチテラバイトのデータベースが完全に収まるようにメモリ容量を増加します。ストレージ (ディスク) からの読み取りと比較した場合、サイズが増加するにつれてシステムメモリを利用する InterSystems IRIS の実証されたメモリキャッシュ機能によって、コードを変更することなくクエリへの応答パフォーマンスを 6 倍向上させることができます。
- 同一のプロセッサを使って、利用可能な最速の NVMe ドライブから、App Direct モードによって DAX XFS ファイルシステムとして DCPMM を利用するようにトランザクションジャーナルディスクを変更するだけで、HL7 変換など、InterSystems

IRIS と InterSystems IRIS for Health

に基づく高速データ相互運用性スループットアプリケーションのパフォーマンスを最大 60% 増のスループットに改善します。

メモリ速度のデータ転送とデータの永続性を活用することで、InterSystems IRIS と InterSystems IRIS for Health に大きなメリットが与えられます。

- 読み取り集中型ワークロードが書き込み集中型ワークロードか、またはその両方のワークロードかに関係なく、Mixed Mode の DCPMM を使った 1 つのリソースコンポーネントの為だけにサーバー全体を過剰に割り当てることなく、必要に応じて計算リソースを拡大します。

お客様の InterSystems IRIS

ベースのアプリケーションに最適なハードウェア構成についてのご相談は、InterSystems テクノロジーアーキテクトにお問い合わせください。

[#HL7 #インターシステムズビジネスソリューションとアーキテクチャ #シャーディング #テスト #パフォーマンス #ビッグデータ #HealthShare #InterSystems IRIS #InterSystems IRIS for Health #TrakCare](#)

---

ソースURL:<https://jp.community.intersystems.com/post/intersystems-iris-%E3%81%8A%E3%82%88%E3%81%B3-intel-optane-dc-%E3%83%91%E3%83%BC%E3%82%B7%E3%82%B9%E3%83%86%E3%83%B3%E3%83%88%E3%83%A1%E3%83%A2%E3%83%AA>