NeurIPS 億単位の近似最近傍探索コンテストの優勝作品

インテル® Xeon® プロセッサーとインテル® Optane™ テクノロジーで検索パフォーマンスを大幅に向上

Mariano Tepper、Cecilia Aguerrebere、および Ted Willke インテル ラボ
Sourabh Dongaonkar および Jawad B Khan インテル ファウンドリー・サービス
Mark Hildebrand カリフォルニア大学デイビス校

近似最近傍探索（ANN）（英語）とも呼ばれる類似性探索は、ウェブスケールのデータベース上で検索、推薦、ランキング操作を必要とする多くの AI アプリケーションのバックボーンであり、精度、スピード、スケール、コスト、そしてサービス品質制約が重要になります。この記事では、インテル® Xeon® プロセッサーとインテル® Optane™ メモリーの能力を活用することで、ANN を進化させるソリューションについて説明します。これらの進化を示すため、我々は NeurIPS'21 億単位の近似最近傍探索コンテスト（英語）に参加し、カスタム・ハードウェア・トラックで優勝しました。我々のソリューションは、CAPEX（資本的支出）を 1/8 ~ 1/19 に削減し、5 年間の OPEX（営業費用）を第 2 位のソリューション（英語）と同等に抑えました。これにより、現代の大規模で高精度かつハイパフォーマンスなシナリオにおける類似性検索の参入障壁を大幅に下げ、普及を促進します。

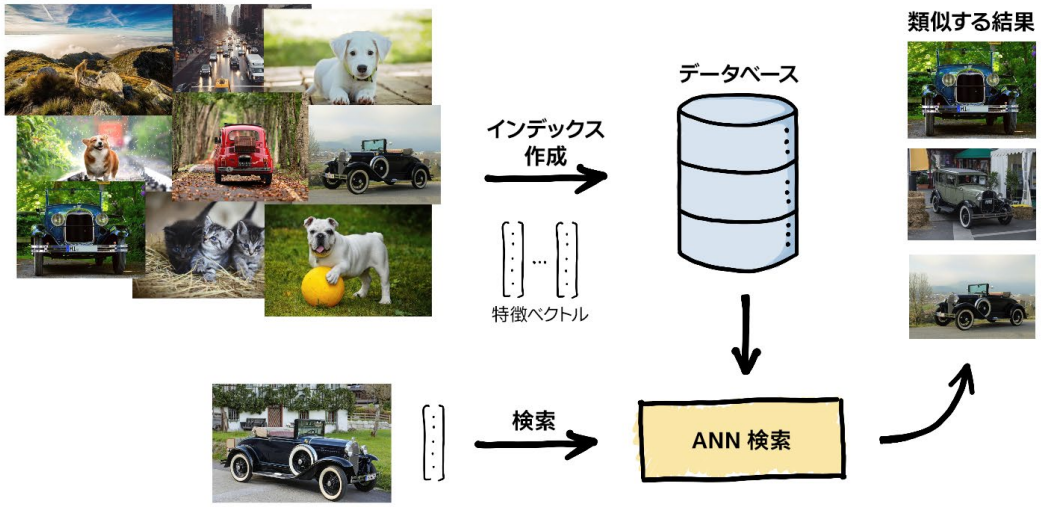


図 1. ANN 検索パイプラインの概略図

近似最近傍探索

類似性検索の目的は、与えられた高次元特徴ベクトルのデータベースと同次元のクエリーベクトルに対し、何らかの類似性関数に基づいて、クエリーと最も類似するデータベース・ベクトルを検索することです（**図 1**）。現代のアプリケーションでは、これらのベクトルはデータ（画像、音声、テキストなど）の内容を表し、類似したベクトルが意味的に関連する項目に対応するように、ディープラーニング・システムを使用して抽出および要約されます。

実際に有用であるには、類似性検索ソリューションはさまざまな価値を提供する必要があります。

- **精度：**検索結果が実用に耐えられる品質でなければなりません（つまり、検索された項目がクエリーに類似している必要があります）。
- **パフォーマンス：**検索は高速で、多くの場合、厳格なサービス品質制約を満たす必要があります。
- **スケーラビリティー：**データベースは、含まれる項目の数や次元数など、ますます大きくなっています。
- **コスト：**本番環境やデータセンターで使用されるため、総所有コスト（TCO）を最小化する必要があります。TCO は通常、資本的支出（CAPEX）と営業費用（OPEX）の組み合わせとして測定されます。

自然なソリューションとして、データベース内の各ベクトルを線形にスキャンし、クエリーと比較し、結果を類似度の降順でランク付けし、最も類似したベクトルを返すことが考えられます。しかし、データの膨大さと豊富さから、このアプローチは不可能であり、計算とメモリーの両方を多用する大規模な類似性検索は、非常に困難な問題です。一般に、次の 2 つのフェーズを含む優れたソリューションを必要とします。

1. インデックス作成時には、データベースの各要素を高次元ベクトルに変換し、検索時にデータベースの一部にのみアクセスするようにインデックスを作成します。
2. 検索時には、与えられたクエリーベクトルに対し、アルゴリズムはインデックスを使用してデータベースをふるいにかけます。その結果は、最終的な用途に応じて、またこれらの意味的に関連した結果に基づいて、さまざまな情報に基づいた行動を取るために使用されます。

NeurIPS'21 億単位の近似最近傍探索コンテスト

2021 年 12 月、NeurIPS カンファレンスの一環として、初の億単位の類似性検索コンテストが開催されました。コンテストの目的は、精選された実際のデータセットで最先端の類似性検索を比較検討し、新しいソリューションの開発を促進することでした。我々は、インテルのハードウェアをフル活用できるカスタム・ハードウェア・トラックに参加しました。インテル® Xeon® プロセッサーとインテル® Optane™ パーシステント・メモリー（PMem）の機能をフル活用したソリューションを開発し、ワンツーアプローチを実現した結果、最終的にコンテストで優勝を勝ち取りました。

データセット間で比較される基本的なメトリックは TCO で、90% の再現率と 10 万クエリー / 秒 (QPS) のスループットを持つソリューションの CAPEX + 5 年間の OPEX として定義されました。CAPEX と OPEX（英語）は、コンテスト主催者によって次のように定義されています。

- CAPEX = （すべてのハードウェア・コンポーネントの希望小売価格）×（100,000QPS をサポートするのに必要な最小システム数)
- OPEX = （ベースライン再現率 @10 しきい値以上の最大 QPS）×（キロワット時 / クエリー）×（秒 / 時間）×（時間 / 年）×（5 年間）×（ドル / キロワット時）×（100,000QPS に対応するのに必要な最小システム数)

これらのメトリックは、各ソリューションの電力効率 (OPEX) と生のパフォーマンス (CAPEX) のバランスをとります。

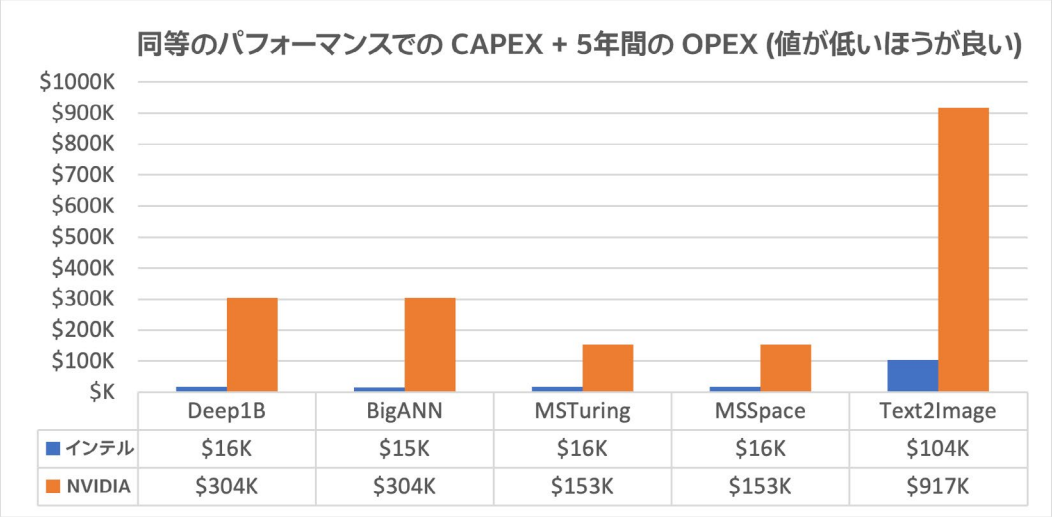


図 2. 優勝したインテルベースのソリューションと第 2 位の NVIDIA* ベースのソリューションの TCO の相違点（5 つの異なるデータセット（X 軸）で最大 20 倍の向上を達成)

我々のソリューションは TCO を画期的に改善し、第 2 位の複数のデータセットで NVIDIA DGX* A100 GPU システムを使用する `cuanns_multigpu` と比較して、8.85 ~ 19.7 倍の効率を実現しています（**図 2**）。この圧倒的な差は、我々と第 2 位の NVIDIA* のハードウェア構成を比較すると明白です（**表 1**）。インテル® Optane™ パーシステント・メモリーを搭載した安価な 1U 2S インテル® Xeon® プロセッサー・ベースのサーバー 1 台で、A100 GPU を 8 個、ANN 検索ワークロード用の 64 コア CPU を 2 個搭載した NVIDIA* のフラッグシップである NVIDIA DGX* A100 サーバーと同等のパフォーマンスを達成することが可能です。

	インテル® Xeon® プロセッサー + インテル® Optane™ メモリー	**NVIDIA DGX* A100**
CPU	デュアル・インテル® Xeon® Gold 6330N プロセッサー 合計 56 コア	デュアル AMD EPYC* 7742 (開発コード名 Rome) 合計 128 コア
システムメモリー	512GB DDR4 2TB インテル® Optane™ パーシステント・メモリー 200 シリーズ	2TB DDR4
GPU	なし	8x NVIDIA* A100 80GB GPU
GPU メモリー	なし	640GB
電力	最大 1.2KW	最大 6.5KW
合計コスト	14,664 ドル（英語）	150,000+ ドル（英語）

表 1. BigANN コンテストのインテルと NVIDIA* のハードウェア構成の比較—2 つの構成は同様のパフォーマンスを達成

インテルベースのソリューションでは同等のパフォーマンスにおける CAPEX が非常に低いことに加え、電力効率も大幅に向上しています。これは、コンテストの全マシンで標準 IPMI インターフェイスにより測定されたクエリー当たりの電力（ジュール単位）でも示されています。インテルベースのソリューションのクエリー当たりの電力は、NVIDIA* ベースのソリューションと比較して、最大 5 倍も優れています（**図 3**）。これは、ANN 検索問題に対するより持続可能なソリューションであるとともに、長期的な OPEX の改善にもつながります。

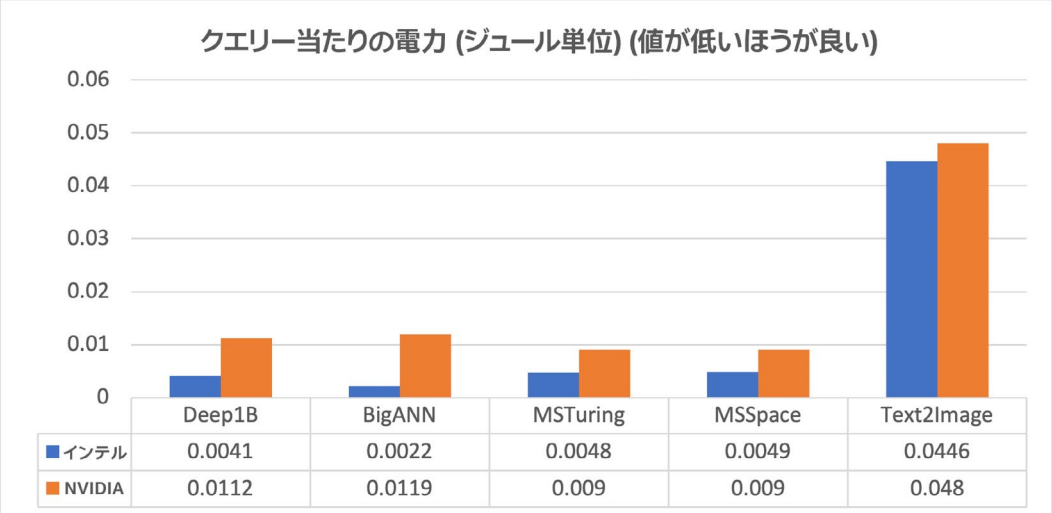


図 3. インテルベースのソリューションのクエリー当たりの電力は NVIDIA* ベースのソリューションと比較して最大 5 倍優れている

このような TCO の大幅な改善は、インテル® Xeon® プロセッサーの利点とインテル® Optane™ パーシステント・メモリーの優れた容量とスループットに加え、ハードウェア・リソースの最適な利用を可能にするアルゴリズムの革新によって実現されました。以下では、このハードウェアとソフトウェアの複合的なアプローチにより、いかにして圧倒的な差をつけてこのコンテストを勝ち抜いたのか、その詳細を紹介します。

アルゴリズムによるアプローチ

ここでは、インテル® Xeon® プロセッサーとインテル® Optane™ パーシステント・メモリーによる ANN 検索アルゴリズム（GraphANN）のパフォーマンスを示します。GraphANN は、インテル® Optane™ パーシステント・メモリー向けに高度に最適化されたグラフベースの Vamana アルゴリズムの拡張版です。データポイントのインデックスを作成するため有向グラフを構築し、グラフをナビゲートして新しいクエリーの最近傍を見つけるため貪欲法による検索を行います。検索中は、2 つの主要なデータ構造（グラフと特徴ベクトル）が使用されます。このソリューションでは、グラフをインテル® Optane™ パーシステント・メモリーに格納し、可能な限り特徴ベクトルを DRAM に保持します。この組み合わせにより、1 ドル当たりのスループットとパフォーマンスが飛躍的に向上します。さらに、インテル® Optane™ パーシステント・メモリーは従来の DRAM よりもはるかに大容量であるため、大規模なデータセットに必要なスケーリングが可能です。最後に、パーシステント・メモリーを使用することで、億単位のデータセットでは非常に時間のかかるメモリーへのインデックスのロードが不要になるという利点もあります。

インテル® Optane™ パーシステント・メモリー（PMem）

インテル® Optane™ パーシステント・メモリーは、SSD やパーシステント・メモリー・アプリケーションに使用できるストレージクラスのメモリーです。歴史的に、メモリーとストレージのパフォーマンスの間には常にギャップがありました。インテル® Optane™ メモリー・テクノロジーはこのギャップを埋めるために設計されています（**図 4**）。高密度でトランジスターを必要としない 3 次元積層可能な設計で、メモリーセルを個別にアドレス指定できます。これらの機能は、手頃な価格で大容量とデータ永続性のサポートというユニークな組み合わせを提供します。特徴的な動作モードを提供する革新的なテクノロジーにより、さまざまなワークロードのニーズに適応できます。例えば、インテル® Optane™ テクノロジーは、ストレージがボトルネックとなっている大規模アプリケーションのログやキャッシュ層のストレージを高速化するために使用されています。

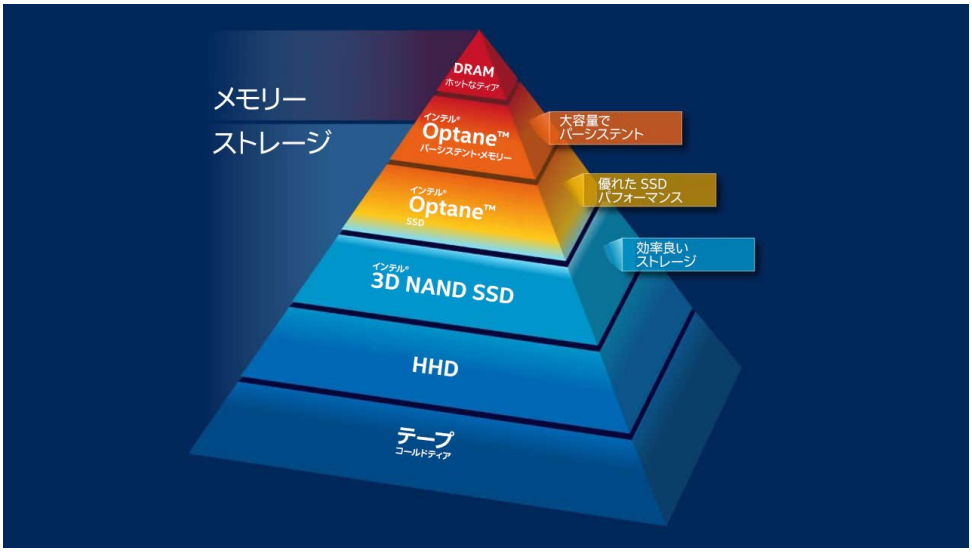


図 4. メモリーとストレージの階層とインテル® Optane™ テクノロジーの位置付け

インテル® Optane™ パーシステント・メモリーと DRAM の類似点は、DIMM 規格パッケージであり、DRAM と同じバス / チャネル上に存在し、DRAM と同様に揮発性データを格納できます。DRAM との相違点は以下の通りです。

- インテル® Optane™ パーシステント・メモリーは従来の DRAM よりもはるかに大きな容量を実現します。モジュールの容量には 128GB、256GB、512GB があり、一般に 16GB から 64GB の範囲である DRAM モジュールよりもはるかに大容量です。
- インテル® Optane™ パーシステント・メモリーは、モジュールへの電源供給がない状態でデータを保存できるパーシステント・モードで動作可能であり、組込みのハードウェア暗号化機能によりデータを安全に維持します。TCO は、GB 当たりのコスト比較において DRAM よりも大幅に改善されており、また DRAM を超える容量への拡張が可能です。

柔軟性を高めるため、インテル® Optane™ パーシステント・メモリーにはメモリーモードとアプリ・ダイレクト・モードの 2 つの動作モードがあります。メモリーモードは、永続性なしでメインメモリー容量を拡張します。インテル® Optane™ パーシステント・メモリーと、そのダイレクト・マップ・キャッシュとして機能する従来の DRAM の組み合わせです。アプリ・ダイレクト・モードでは、インテル® Optane™ パーシステント・メモリーは、DRAM とは別にアドレス指定可能なパーシステント・メモリー・デバイスのように見えます。

インテル® Optane™ パーシステント・メモリー（PMem）を使用した ANN

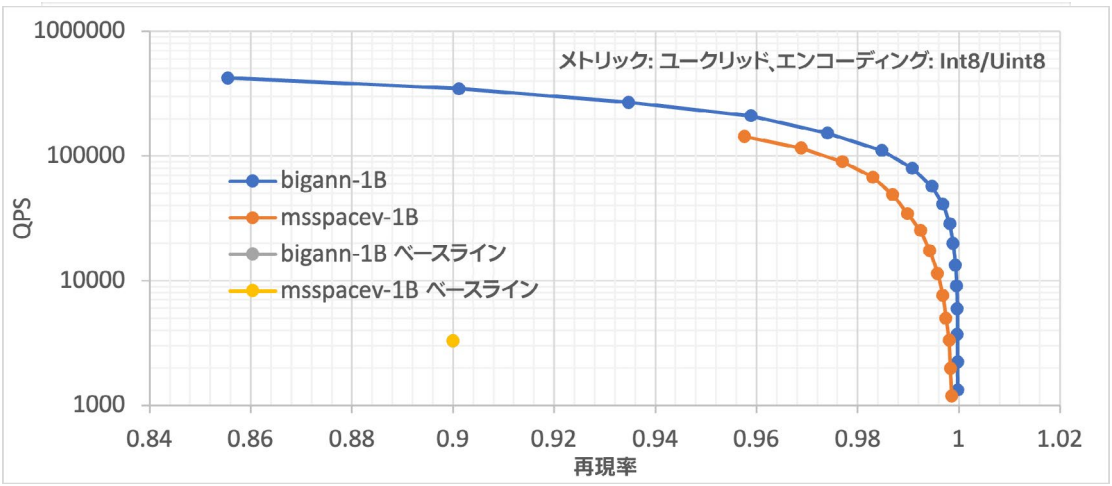
Vamana アルゴリズムのアクセスパターンを調査した結果、クエリー間のデータ再利用は限定的であることが分かりました。インテル® Optane™ パーシステント・メモリーをメモリーモードで使用しても、キャッシュとしての価値があまりないため、ここではアプリ・ダイレクト・モードを使用しました。

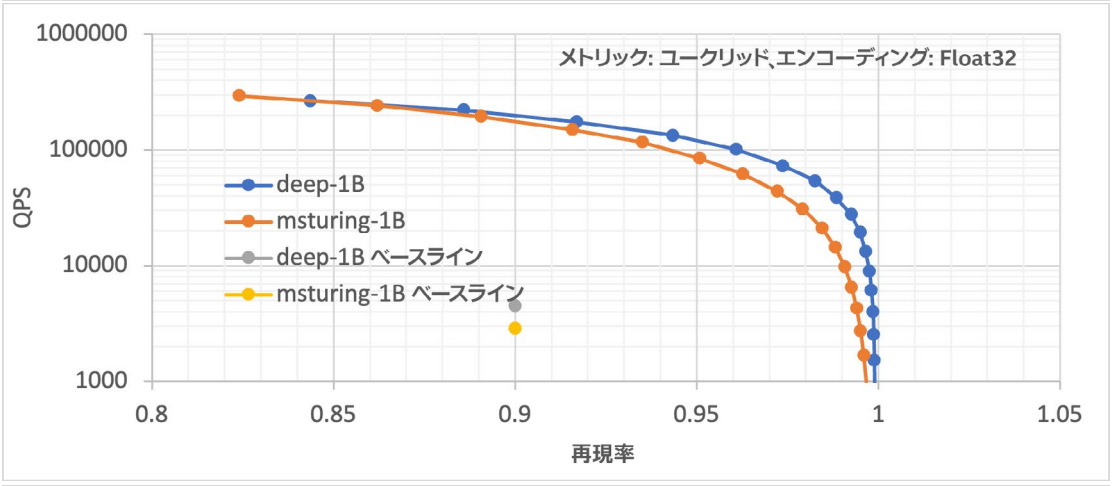
パフォーマンスを最大限に引き出すため、グラフをインテル® Optane™ パーシステント・メモリーに格納し、特徴ベクトルは DRAM に格納するようデータを整理しました。グラフのアクセスパターンは、非常にランダムで、非局所的です。インテル® Optane™ パーシステント・メモリーは 64 バイトのブロックサイズを持つため、パフォーマンスを維持したままグラフ要素にアクセスできます。ちょうど 4 回のアクセスで任意のノードの近傍を取得できるように、各ノードの最大出次数は 127 に制限します（近傍ごとに 4B と近傍数に 4B を使用します）。データを連続して保存することで、この 4 回のアクセスをパイプライン化できます。

Vamana アルゴリズムの最適化バージョンは、Julia（英語）プログラミング言語で記述されています。最適化は、一般的なもの（ソフトウェア・ベース）とインテル® Optane™ テクノロジー固有のものに分けることができます。一般的な最適化は、グラフとデータの表現の最適化、距離計算での VNNI 命令の使用、データベクトルの静的サイズ変更、メモリー・アライメントなどに関連しています。重要な最適化の 1 つは、距離計算ループをベクトルフェッチ・ループから切り離す「プリフェッチ・ホイスティング」です。このアプローチでは、距離計算のステップを開始する前に、x86 組込み命令を使用してメモリーから可能な限り多くのデータをプリフェッチします。これにより、メモリー・レイテンシーがクエリーに与える影響を最小限に抑えられます。もう 1 つの重要な最適化は、DRAM と PMem の間でベクトルを分割することで得られます。距離計算のベクトルのフェッチは検索で最も時間のかかるステップであり、DRAM に可能な限り多くのベクトルを保持することで、PMem トラフィックが軽減され、パフォーマンスが大幅に向上します。

マルチスレッド・アーキテクチャーは、ワーカースレッドに動的に負荷分散されるクエリーの小さなバッチを作成します。各スレッドは、バッチ内で一度に 1 つのクエリーを処理します。さらに、クエリー処理に必要なすべての中間スクラッチ空間のデータ構造は事前に割り当て、各スレッドは専用のプライベート・スクラッチ空間を所有します。これにより、クエリー処理中の動的なメモリー割り当てが不要になり、スレッド間の同期を最小限に抑えられます。

これらの最適化により、従来の最適解（GPU 上で動作する FAISS アルゴリズム）と比較して、ANN 検索パフォーマンスが 10 ~ 100 倍以上向上しました。**図 5** は、5 つの異なるデータセットで最適化したアプローチによって達成された改善を示しています。これらのデータセットには、異なるエンコーディング（Int8、UInt8、Float32）と、異なる距離メトリック（ユークリッドと内積）が含まれます。これらのデータセットにおいて、インテル® Xeon® プロセッサーとインテル® Optane™ メモリー上で GraphANN を実行すると、ベースライン・パフォーマンスが劇的に向上することが分かります。





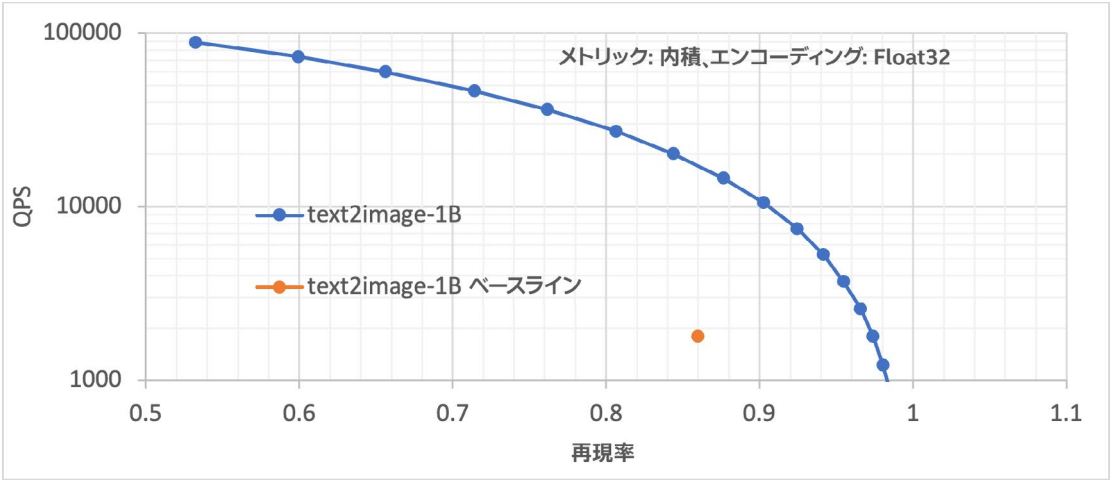


図 5. 5 つの異なるデータセットでの 1 秒当たりのクエリー数（スループット）と再現率—これまで最高のソフトウェア（FAISS）とハードウェア（GPU）に対して、インテル® Optane™ パーシステント・メモリーを使用したソリューションによる改善の程度を示している

まとめ

この記事では、NeurIPS 2021 の億単位の近似最近傍探索コンテストで優勝したアルゴリズム、設計上の選択、および関連ハードウェア設定について説明しました（コンテストのリーダーボードはこちら（英語）を参照してください）。また、関連するコードの変更を伴わないハードウェアのアップグレードから、コードの本格的なカスタム書き換えまで、さまざまな設計シナリオにおいて、インテル® Optane™ パーシステント・メモリーが類似性検索アルゴリズムのパフォーマンスを大幅に向上させることを示しました。