ケーススタディー

インテル® Xeon® プロセッサー

Netflix は高速でシームレスなストリーミング体験を提供するため、インテル® Xeon® プロセッサー搭載の Amazon EC2* インスタンスを選択

AI をサポートするインテルのテクノロジーにより、Netflix はクラウド費用を大幅に削減しながら、パフォーマンスと品質の高いビデオコンテンツとマイクロサービスを加入者に提供

ソリューション

- インテル® Xeon® プロセッサー
- インテル® oneAPI ディープ・ニューラル・ネットワーク・ライブラリー (インテル® oneDNN)
- インテル® ディープラーニング・ブースト (インテル® DL ブースト) のベクトル・ニューラル・ネットワーク命令 (VNNI)
- インテル® アドバンスト・ベクトル・エクステンション 512 (インテル® AVX-512)
- インテル® VTune™ プロファイラー
- インテル® PerfSpect
- Amazon EC2* インスタンス

概要

Netflix は、ホーム・エンターテイメントを変革し、2 億 6,000 万人の加入者にあらゆるデバイスで信頼性の高いカスタマイズされた体験を提供することを目指しています。これを達成するには、インテル® Xeon® プロセッサー搭載の Amazon EC2* インスタンスなどの高度なテクノロジーを使用して、データの移動と AI ワークロードを高速化する必要があります。インテルと協力して Netflix は以下を実現しました。

- Amazon インスタンスをマイクロアーキテクチャー・レベルで最適化して、パフォーマンスを向上し、クラウド費用を削減しました。EC2* インスタンスのアップグレード後、CPU あたりのパフォーマンスが 3.5 倍向上し、予想された線形スケーリングを上回りました。[1]
- インテル® oneAPI ディープ・ニューラル・ネットワーク・ライブラリー (インテル® oneDNN) とインテル® アドバンスト・ベクトル・エクステンション (インテル® AVX) 命令セットを使用して、ユーザー需要が少ない時間帯のビデオエンコード速度を最適化しました。インテルのソリューションにより、エンコード FPS が大幅に改善されました。

課題

Netflix は、使用するデバイスに関係なく、シームレスなオンデマンド・コンテンツを世界中の加入者に配信することを目指しています。このプロセスには、加入者の体験をサポートするワークロード向けに最適化された複数のマイクロサービスが必要です。バックエンド・マイクロサービスは、コンテンツの開発、レンダリング、エンコードを処理しなければなりません。加入者側では、数千のタイトルから最も関連性の高いコンテンツを識別して推薦する、カスタマイズされたホームページ・ビューが必要です。さらに、Netflix は、いつでも利用できる優れたエンターテイメントと卓越したストリーミング品質を加入者に提供できるよう継続的に取り組んでいます。

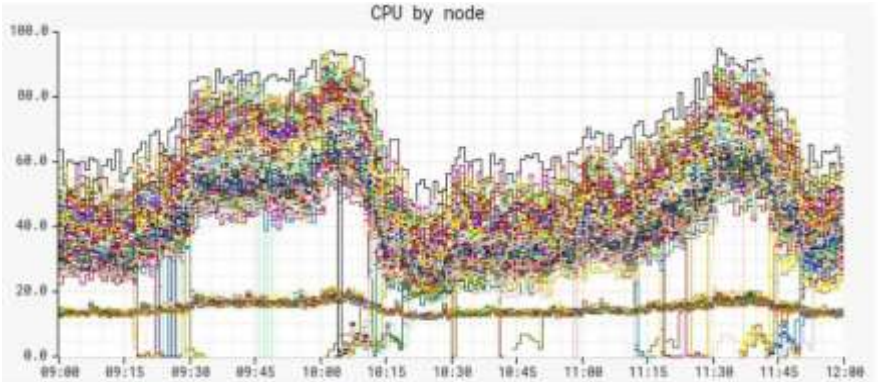

図 1. ノードごとの CPU 利用率を示すグラフ。解析の結果では、ノード間のトラフィック分布はほぼ均等ですが、CPU メトリックは異なる二峰性分布パターンを示しています。[1]

すべてのタスクを達成するため、Netflix は、問題が発生した場合のトラブルシューティングを簡素化する高度なツールを備えた、信頼性と拡張性の高い、AI 対応のクラウド・ソリューションを必要としていました。例えば、Netflix チームは、Amazon EC2* インスタンスのパフォーマンスの評価中に、予期しないレイテンシーの問題を発見しました。クラウド費用を最小限に抑えながらワークロードを高速化するには、インスタンスを CPU のマイクロアーキテクチャー・レベルまで評価する効果的な方法が必要でした。

ソリューション

加入者に高速でカスタマイズされたストリーミング体験を提供するため、Netflix はインテル® Xeon® プロセッサー搭載の Amazon EC2* インスタンスの能力を活用しました。Netflix のパフォーマンス・チームはインテルと緊密に協力して、ソフトウェアと利用可能なハードウェア・リソースの相互作用を精査し、ボトルネックを特定しました。インテル® VTune™ プロファイラーでプロセッサー時間を最適に使用していないコード領域を検出し、インテル® PerfSpect でマイクロアーキテクチャーのサブシステムとプログラムシーケンスを評価してさらなる情報を得ることができました。これらのツールは、最終的に、Java 仮想マシン内の一連の命令におけるインスタンスのボトルネックをピンポイントで特定するのに役立ちました。

Netflix は GPU ではなく、インテル® Xeon® プロセッサー搭載の Amazon インスタンスを使用することで、各インスタンスで複数のタスクを実行してコストを削減する方法を見つけました。視聴のピーク時間帯には、Amazon インスタンスはリソースをストリーミングに集中させ、ユーザー需要が減少したら、コンピューティング能力をビデオエンコードの高速化に転用できます。

「当社のストリーミング・サービスで加入者に最高の体験を提供するには、速度が重要です。インテルのテクノロジーを利用してボトルネックを特定することで、クラウド費用を最小限に抑えながら、Amazon EC2* インスタンスのパフォーマンスをほぼ 3 倍にすることができました。」

– Netflix 社パフォーマンス・エンジニア Vadim Filanovsky 氏

図 2. 真の共有問題を特定して対処した後、レイテンシーが大幅に減少しているのが分かります。[1]

結果

インスタンスのボトルネックを特定するインテルのサポートにより、Netflix は、Amazon EC2* インスタンスの初期スループットと比較して、CPU あたりのパフォーマンスが 3.5 倍向上しました。[1] また、平均レイテンシーとテール・レイテンシーの大幅な削減も達成しました。[2] インテルは OpenJDK* (オープンソースの Java* 開発キット) でレイテンシーの原因に対処したため、Java* ワークロードを利用しているほかの企業も Netflix の CPU 最適化アプローチから恩恵を受けることができます。

Netflix は、インテル® AVX 命令セットを利用するインテル® oneDNN により、エンコード FPS を大幅に向上し、すべてのデバイスで優れたビデオ品質を提供できると主張しています。

インテル® Xeon® プロセッサー搭載の Netflix の Amazon インスタンスは、自動スケーリングにより、さまざまな目的に効果的に対応することもできます。CPU によって効率が向上したことで、Netflix はミッションクリティカルなワークロードに必要なインスタンスの数を減らし、クラウド・インフラストラクチャー全体の大幅な費用削減を実現しました。

intel + aws

[1] 初期の結果と比較して 3.5 倍の向上を達成するため、a) 偽の共有を排除する、b) 真の共有を回避する、という 2 つの明確なステップが実施されました。図 2 のグラフはステップ (b) の結果のみを表しています。偽の共有を排除する前と後のグラフはこちら (英語) を参照してください。

[2] https://netflixtechblog.com/seeing-through-hardware-counters-a-journey-to-threefold-performance-increase-2721924a2822 (英語)