64 コアと 96 コアのインテル® Xeon® 6 プロセッサー上でのインテル® AI for Enterprise RAG パフォーマンスのスケーリング

この記事は、インテルのブログで公開されている「Scaling Intel® AI for Enterprise RAG Performance: 64-Core vs 96-Core Intel® Xeon®」の日本語参考訳です。原文は更新される可能性があります。原文と翻訳文の内容が異なる場合は原文を優先してください。

エグゼクティブ・サマリー

この評価により、インテル® AI for Enterprise RAG 推論において、64 コアから 96 コアのインテル® Xeon® 6 プロセッサー・ベースの構成に移行することで、同時実行性が大幅に向上し、レイテンシーのスケーリングが改善されることが示されました。96 コア SKU は、Llama-AWQ および Mistral-AWQ において、SLA に準拠した同時実行性をすべてのワークロードで倍増 (32 ユーザーから 64 ユーザー) させ、Qwen-AWQ の SLA 準拠の同時実行性も 64 コアシステムと比較して 33～50% 向上 (ワークロードにより異なる) させます。

はじめに

エンタープライズ RAG

検索拡張生成 (RAG) は、大規模言語モデルの能力とリアルタイムの情報検索機能を組み合わせた、人工知能 (AI) への革新的なアプローチです。エンタープライズ RAG システムは、この概念を拡張し、ビジネス・クリティカルなアプリケーションの厳しい要件に対応し、組織が求める信頼性、セキュリティー、パフォーマンス基準を備え、エンタープライズ規模のワークロードを処理できる、実働環境に対応したソリューションを提供します。

インテル® AI for Enterprise RAG (英語) ソリューションは、包括的な RAG パイプラインを提供します。このエンドツーエンドのソリューションは、埋め込みモデル、ベクトル・データベース、リランキング機能、大規模言語モデルを統一されたスケーラブルなアーキテクチャーに統合し、主要なエンタープライズ課題 (高同時ユーザー負荷での低レイテンシーの維持、一貫した応答品質の確保、ビジネス要件を満たす測定可能な SLA の提供) に対応できるように設計されています。

評価の目的

前回の調査 (英語) では、64 コアのインテル® Xeon® 6767P プラットフォームを分析しました。今回のフォローアップでは、その作業を要約し、「AWQ 量子化 LLM 推論をデュアルソケットの開発コード名 Granite Rapids 64 コア構成から、さらに高密度な 96 コア構成 (インテル® Xeon® 6972P プロセッサー) に移行することで、実際にどの程度のキャパシティーとレイテンシーの向上が実現されるのか?」という点に注目します。

テスト方法: 包括的なパフォーマンス評価

テスト方法、データ準備、認証モデル、同時実行ハーネス、および検索/リランク・パイプラインは、前回の調査から変更ありません。2 つのテストシナリオ間で異なるのは、ハードウェア構成と vLLM レプリカの数のみです。

テスト対象の AWQ モデル

前回と同様に、このベンチマークでは 3 つの異なる大規模言語モデル構成を評価します。今回は AWQ (Activation-aware Weight Quantization) バージョンのみを評価します。それぞれ、異なるエンタープライズ展開シナリオと地域市場の要件に最適化されています。

- **Llama** hugging-quants/Meta-Llama-3.1-8B-Instruct-AWQ-INT4 (英語)
- **Qwen** Qwen/Qwen3-14B-AWQ (英語)
- **Mistral** solidrust/Mistral-7B-Instruct-v0.3-AWQ (英語)

ハードウェアとソフトウェアの構成

インテル® Xeon® 6767P プロセッサー (64 コア) - ベースライン構成

- **プロセッサー:** 2x インテル® Xeon® 6767P プロセッサー (ソケットごとの物理コア数 64、350W TDP、SNC 無効)
- **メモリー:** 512GB DDR5-6400 (16x 32GB モジュール)
- **vLLM レプリカ:** 2 ポッド
- **ネットワーク:** 4x BCM57412 NetXtreme-E 10Gb RDMA イーサネット・コントローラー
- **ストレージ:** 447.1GB HFS480G3H2X069N + 447.1GB Dell BOSS-N1
- **BIOS:** バージョン 1.2.6、マイクロコード 0x10003a2

インテル® Xeon® 6972P プロセッサー (96 コア) - 拡張構成

- **プロセッサー:** 2x インテル® Xeon® 6972P プロセッサー (ソケットごとの物理コア数 96、500W TDP、SNC 無効)
- **メモリー:** 1536GB DDR5-6400 (24x 64GB モジュール)
- **vLLM レプリカ:** 4 ポッド (スループット向上ため 2 倍に増設)
- **ネットワーク:** 2x インテル® イーサネット・コントローラー X710 10GBASE-T + 1x I210 Gigabit
- **ストレージ:** 894.3GB SAMSUNG MZ1L2960HCJR SSD
- **BIOS:** BHSDCRB1.IPC.3544.P60.2504160256、マイクロコード 0x10003c1

共通のソフトウェア・スタック構成

- **オペレーティング・システム:** Ubuntu 24.04.2 LTS
- **組込みサービス:** TorchServe 0.12.0、4 ポッドレプリカ、モデル: BAAI/bge-base-en-v1.5
- **ベクトル・データベース:** Redis 7.4.0-v2、1M ベクトル
- **リトリーバー:** 1 ポッドレプリカ、k=5
- **リランカー:** TorchServe 0.12.0、2 ポッドレプリカ、top_n=1、モデル: BAAI/bge-reranker-base
- **LLM サービス:** vLLM 0.9.2、BF16 精度
- **アプリケーション:** インテル® AI for Enterprise RAG 1.4.0

パフォーマンス分析: 64 コアと 96 コアの比較

以下のパフォーマンス・データは、複数の LLM モデルにおいて、ワークロード・パターンと同時実行レベルが異なる場合の、インテル® AI for Enterprise RAG パイプラインのエンドツーエンド・パフォーマンスを示しています。TTFT (Time to First Token) は、埋め込み、検索、リランキング、LLM といったすべての RAG コンポーネントを含むエンドツーエンドのレイテンシーで、秒単位で報告され、TPOT (Time Per Output Token) はミリ秒単位で報告されます。TTFT が低いほど応答性が向上し、TPOT が低いほど持続的なトークン・スループットが向上します。

ワークロード: 128 入力/128 出力
簡潔な Q&A と迅速な応答に最適化

concurrent users	TTFT (seconds)						TPOT (milliseconds)					
	Llama		Qwen		Mistral		Llama		Qwen		Mistral	
	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores
12	0.99	1.06	1.14	1.30	1.03	1.07	53.4	56.7	97.2	92.2	46.5	49.1
16	1.08	1.10	1.21	1.34	1.25	1.18	58.6	59.3	109.3	96.4	57.8	51.8
24	1.15	1.26	1.21	1.50	1.27	1.30	68.4	64.1	137.4	106.0	60.8	56.4
32	1.27	1.40	1.26	1.57	1.36	1.51	78.2	69.7	154.1	113.7	70.5	62.1
64	1.61	1.98	1.57	1.93	1.78	2.16	122.1	92.7	223.1	156.2	109.4	97.9
128	2.22	2.73	2.29	2.95	2.16	3.13	224.6	153.8	386.1	261.0	206.5	139.5

図 1. ワークロード: 128 入力/128 出力のパフォーマンス値

ワークロード: 256 入力/256 出力
バランスの取れた応答要件を持つ中程度の長さのクエリー

concurrent users	TTFT (seconds)						TPOT (milliseconds)					
	Llama		Qwen		Mistral		Llama		Qwen		Mistral	
	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores
12	1.23	1.24	1.35	1.63	1.18	1.25	53.8	57.6	92.5	93.3	47.3	50.0
16	1.25	1.26	1.43	1.68	1.34	1.29	59.7	59.9	104.5	96.8	53.1	52.2
24	1.37	1.36	1.94	1.77	1.44	1.38	69.5	64.5	126.1	104.9	63.7	56.8
32	1.47	1.46	1.74	1.83	1.62	1.53	79.9	69.8	153.8	111.7	74.0	62.2
64	1.83	1.70	2.06	2.08	1.87	1.82	126.4	92.9	218.2	150.9	117.5	85.4
128	2.60	2.12	4.15	2.51	2.58	2.31	227.0	147.5	372.1	240.5	213.9	140.3

図 2. ワークロード: 256 入力/256 出力のパフォーマンス値

ワークロード: 256 入力/512 出力
包括的な応答を必要とする中程度の長さのクエリー

concurrent users	TTFT (seconds)						TPOT (milliseconds)					
	Llama		Qwen		Mistral		Llama		Qwen		Mistral	
	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores
12	1.11	1.21	1.36	1.61	1.12	1.23	52.5	57.3	88.1	92.0	45.6	49.7
16	1.17	1.22	1.48	1.71	1.17	1.23	57.2	59.0	99.0	95.4	50.6	51.6
24	1.28	1.33	1.77	1.76	1.21	1.29	65.5	62.6	119.2	99.3	58.9	54.8
32	1.34	1.37	1.62	1.83	1.28	1.36	73.7	67.9	143.3	106.4	67.4	59.5
64	1.57	1.48	2.67	2.06	1.53	1.48	112.0	87.7	185.1	141.6	102.9	78.7
128	2.06	1.67	3.88	2.27	1.86	1.64	198.6	131.6	314.0	213.7	182.1	119.4

図 3. ワークロード: 256 入力/512 出力のパフォーマンス値

ワークロード: 256 入力/1024 出力
詳細な分析と説明のための拡張応答生成

concurrent users	TTFT (seconds)						TPOT (milliseconds)					
	Llama		Qwen		Mistral		Llama		Qwen		Mistral	
	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores	64 cores	96 cores
12	1.13	1.21	1.50	1.65	1.12	1.23	53.4	58.5	87.8	93.8	45.7	50.6
16	1.18	1.25	1.48	1.68	1.15	1.25	57.1	60.1	98.3	96.6	51.1	52.0
24	1.29	1.30	1.64	1.81	1.17	1.26	64.9	62.8	118.7	99.5	58.9	55.2
32	1.33	1.32	1.95	1.83	1.23	1.28	73.0	67.4	136.3	107.5	67.2	59.4
64	1.41	1.41	2.80	2.06	1.36	1.41	111.0	86.8	175.9	136.1	101.3	79.4
128	2.07	1.61	4.82	2.56	1.61	1.49	189.9	131.1	296.6	201.1	176.5	119.2

図 4. ワークロード: 256 入力/1024 出力のパフォーマンス値

サービスレベル合意 (SLA) キャパシティー分析

広範なユーザー・エクスペリエンス調査に基づいた、以下の SLA しきい値を採用しています。

- **Time to First Token (TTFT):** < 3 秒 (クエリー送信から最初の応答トークンが出現するまでのレイテンシーを測定)
- **Time Per Output Token (TPOT):** < 100 ミリ秒 (最初の応答が開始された後、後続のトークンが生成されるレートを定量化)

SLA に基づく最大同時ユーザー数

以下の表は、両方の SLA しきい値を同時に満たす最大持続同時実行性を報告しています。

input tokens	max output tokens	Llama		Qwen		Mistral	
		64 cores	96 cores	64 cores	96 cores	64 cores	96 cores
128	**128**	32	64	12	16	32	64
256	**256**	32	64	12	16	32	64
256	**512**	32	64	16	24	32	64
256	**1024**	32	64	16	24	32	64

図 5. SLA に基づく最大同時ユーザー数

主要なパフォーマンス分析

96 コア構成は、高同時ユーザー負荷下で最も大きな利点を示しています。

- **同時ユーザー数 64 人以上:** 96 コアシステムは、高負荷環境下でも 64 コアシステムよりも大幅に優れたパフォーマンスを維持
- **TPOT の改善:** 高同時実行レベルでのトークン生成パフォーマンスが向上
- **負荷時の安定性:** ユーザー数の増加に伴うパフォーマンス特性の一貫性が向上

モデル別のパフォーマンス向上

- **Llama-AWQ:** 全ワークロードでキャパシティーが 100% 増加 (32 → 64 同時ユーザー)
- **Llama-AWQ:** ワークロードの複雑さに応じてキャパシティーが 33%～50% 増加 (12～16 → 16～24 同時ユーザー数)
- **Mistral-AWQ:** 全ワークロードでキャパシティーが 100% 増加 (32 → 64 同時ユーザー)

費用対効果分析

ハードウェア投資比較:

- **コア数:** 50% 増加 (64 → 96 コア)
- **キャパシティーの増加:** Llama-AWQ および Mistral-AWQ モデルのユーザー・キャパシティーが 100% 増加

エンタープライズ向け価値提案:

- **ユーザー密度:** 中程度のハードウェア追加投資で同時ユーザー・キャパシティーを倍増
- **コスト・パフォーマンス:** リソース利用率の向上による効率性の向上
- **スケーリングの経済性:** 大きなキャパシティー要件に対する優れた費用対効果

導入に関する推奨事項

96 コア構成の最適なユースケース:

- **大規模導入:** 50 人以上の同時ユーザーをサポートする予定の組織
- **マルチモデル環境:** 複数の AWQ モデルを同時に必要とする組織
- **ピーク負荷管理:** バースト・キャパシティーを必要とする、大幅な使用量の増加があるアプリケーション
- **統合戦略:** 複数の 64 コア導入環境の統合を目指すデータセンター

64 コア構成で十分なシナリオ:

- **中程度のキャパシティー要件:** 同時ユーザー数が 32 人未満の導入環境
- **予算重視の導入:** 初期投資を最小限に抑えることを優先する組織
- **パイロット導入:** 将来的に拡張を計画している初期の RAG 実装

まとめ

インテル® Xeon® 6972P プロセッサー (96 コア) 構成は、64 コアのベースラインと比較して実質的なパフォーマンス向上をもたらし、Llama-AWQ および Mistral-AWQ の SLA に準拠した同時実行性を倍増させるとともに、高負荷での Qwen-AWQ のスケーラビリティーを大幅に強化します。コア、メモリー、および計算リソースへの増分投資は、持続的な高キャパシティー RAG 展開を必要とする組織にとって、強力な費用対効果 (ROI) をもたらします。

主要な推奨事項:

- 40 人以上の同時ユーザーを対象とするエンタープライズ実装には、**96 コアシステム**を導入してください。
- 統合戦略とマルチモデル環境には、**強化されたキャパシティー**を活用してください。
- ユーザー数の増加予測とピーク負荷要件に基づいて、**インフラストラクチャーの拡張をプロアクティブに計画**してください。
- 64 コアと 96 コアの構成を選択する際には、**リージョンモデルの最適化**を検討してください。

96 コア・プラットフォームは、即座にキャパシティーを増強し、将来のマルチモデル拡張やワークロードの複雑性の変化に対応できるスケーラブルな基盤を構築します。

法務上の注意書き

性能は、使用状況、構成、その他の要因によって異なります。詳細については、http://www.intel.com/PerformanceIndex/ (英語) を参照してください。

インテルのテクノロジーを使用するには、対応したハードウェア、ソフトウェア、またはサービスの有効化が必要となる場合があります。

絶対的なセキュリティーを提供できる製品またはコンポーネントはありません。

実際の費用と結果は異なる場合があります。

インテルは、サードパーティーのデータについて管理や監査を行っていません。ほかの情報も参考にして、正確かどうかを評価してください。