intel XEON

Case Study

AI-Based Recommendation Systems
Intel® Xeon® Scalable Processors

MetaApp Delivers a New AI-Based Recommendation System with Alibaba Cloud and Intel

MetaApp uses the Alibaba Cloud Elastic Compute Service (ECS) compute-optimized c8i instance family with 4th Gen Intel® Xeon® Scalable processors to lower the cost of its AI-based recommendation system by 22 percent while completing the same number of queries per second (QPS).[1]

By using Alibaba Cloud instances running on 4th Gen Intel Xeon Scalable processors, MetaApp needed 25 percent fewer cores to run AI inference on its recommendation system than using instances with the previous-generation processor.[1]

MetaApp offers China's leading game platform for interactive entertainment, with more than 1,000 small- and medium-sized game-development teams using its game-creation and distribution services.[2] The company's services run on Alibaba Cloud, China's top cloud service provider (CSP), with a 34 percent market share.[3]

MetaApp provides game development tools and an AI-based recommendation system. The recommendation system helps game developers increase end-user traffic, boosting the monetization capabilities of their games. The company collaborated with Alibaba Cloud and Intel to build an enhanced recommendation system critical to its growth strategy.

MetaApp built the system on the Alibaba Cloud Elastic Compute Service (ECS) c8i instance. The company used DeepRec, an open source deep learning (DL) framework enhanced by Intel® oneAPI Deep Neural Network Library (Intel® oneDNN) to harness the power of the underlying CPU, the 4th Gen Intel Xeon Scalable processors, including the built-in Intel® Advanced Vector Extensions 512 (Intel® AVX-512) accelerator.

Challenges

AI recommendation systems consume enormous computational power for DL networks, in addition to using significant amounts of memory to support embedding subsystems. The recommendation system starts by narrowing the list of items to recommend from millions to hundreds. The list is referred to as "the candidates." Each item is then scored and ranked using DL-based scoring mechanisms and insights about entity relationships stored in embedded subsystems. Finally, the system fine-tunes the ranking using DL network insights, allowing it to consider complex constraints.

The end-to-end latency of a recommendation system depends on the combined time for candidate generation, ranking, and any re-ranking. In practice, most recommendation systems aim for sub-second to low-second latencies for the entire recommendation process.[4]

The dual requirements for high computational power and substantial amounts of memory are expensive and can challenge profitability. Although GPUs can address the computational power issue to some extent, most of the data still needs to be processed on the CPU side due to memory requirements.

MetaApp identified several opportunities to enhance its recommendation system with a new design. They planned to improve resource usage and resource elasticity. Additionally, they envisioned simplifying the design. Finally, they sought ways to reduce or eliminate reliance on GPUs, which are typically priced at a premium in cloud instances, to reduce system TCO.

Solution

With support from Alibaba Cloud and Intel, MetaApp deployed a new recommendation system on the Alibaba Cloud ECS c8i instance family with 4th Gen Intel Xeon Scalable processors. MetaApp moved its AI training (fine-tuning) workload to the c8i instance from an instance powered by 2nd Gen Intel Xeon Scalable processors. It moved its AI inference workload to the c8i instance from an instance powered by 3rd Gen Intel Xeon Scalable processors.

4th Gen Intel Xeon Scalable processors feature built-in accelerators to help improve performance workload efficiency, especially for workloads powered by AI.[5] They also offer performance-enhancing features like PCIe Gen 5 to unlock input/output (I/O) speeds and DDR5 memory for 1.5x the memory bandwidth of DDR4 memory found in previous processor generations.[6]

MetaApp adopted DeepRec, a high-performance recommendation DL framework built on TensorFlow. The company also used oneDNN, an open source cross-platform performance acceleration library integrated into DeepRec. With oneDNN, MetaApp harnessed the capabilities of Intel AVX-512, a built-in accelerator within 4th Gen Intel Xeon Scalable processors, to accelerate the performance of computational tasks.

Results

MetaApp's new recommendation system helped the company meet its flexibility, cost, and speed goals.

DeepRec delivers dynamic resourcing

MetaApp saved time and development costs using the DeepRec framework and oneDNN, which has helped its developers produce fast, platform-independent AI applications through optimized building blocks.[7]

MetaApp worked with Alibaba Cloud and Intel to optimize DeepRec for 4th Gen Intel Xeon Scalable processors at various levels of the recommendation system. These optimizations have allowed MetaApp to address its end-to-end needs, from offline training to online inference.

DeepRec with Intel optimizations has allowed MetaApp to break free of existing dependencies on GPUs. The company has gained the ability to dynamically adjust resources and achieve flexible scalability in Alibaba Cloud.

Fewer cores, lower cost

By building its recommendation system on Alibaba Cloud ECS c8i instances with 4th Gen Intel Xeon Scalable processors, MetaApp was able to use 25 percent fewer cores to run AI inference at the same rate of queries per second (QPS) as on its original system, but at a 22 percent lower cost.[1]

Better AI training performance

By using Alibaba Cloud ECS c8i instances with 4th Gen Intel Xeon Scalable processors, MetaApp realized 64 percent higher training (fine-tuning) performance than those with 2nd Gen Intel Xeon Scalable processors.[8] Higher training performance allows the recommendation system to understand user preferences faster.

MetaApp also benefitted from 160 percent higher training (fine-tuning) performance per price than instances running on 2nd Gen Intel Xeon Scalable processors.[8] Higher performance per price contributes to lower TCO.

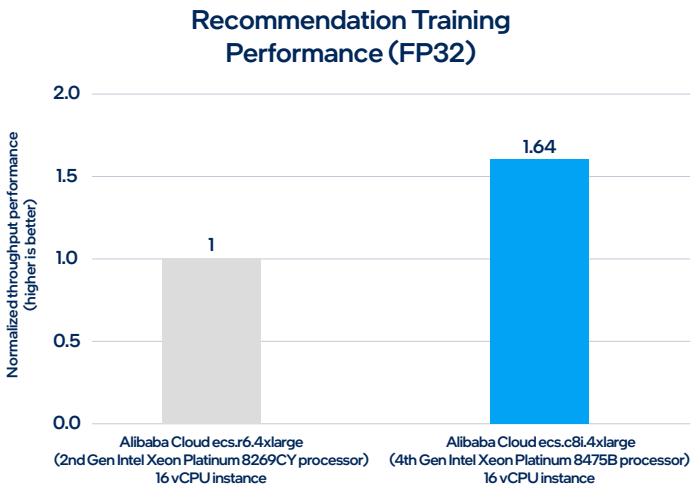

Figure 1. MetaApp's training performance increased by 64 percent[8]

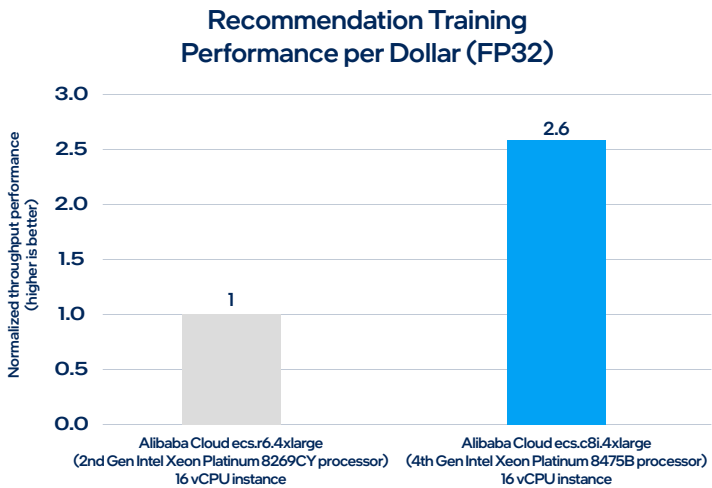

Figure 2. MetaApp's training performance per price increased by 2.6x[8]

> "Utilizing Alibaba Cloud CPU-based instances empowers us to conduct training and inference cost-effectively, enabling innovation with our AI recommendation engine. This approach ensures a fast, responsive user experience required for success and facilitating future scalability."
>
> — Ruozhou Zang, Head of AI Research and Development, MetaApp[9]

Energy efficiency

In addition to helping customers like MetaApp reduce cloud TCO, Alibaba Cloud reaps other benefits from its c8i instances. Intel technologies allow MetaApp to achieve a 2.9x average performance-per-watt efficiency improvement compared to previous-generation processors.[10] This enhanced energy efficiency helps Alibaba Cloud progress toward sustainability goals while reducing costs.

AI acceleration

Intel has a history of providing hardware-based AI acceleration. 2nd Gen Intel Xeon Scalable processors introduced Intel® Deep Learning Boost (Intel® DL Boost), a set of acceleration features to accelerate AI training and inference, as shown in Figure 4.

These features include Vector Neural Network Instructions (VNNI), which uses Intel AVX-512 to reduce the number of operations per clock cycle. 2nd Gen Intel Xeon Scalable processors also natively support the INT8 data type, which can improve inference speed and efficiency.

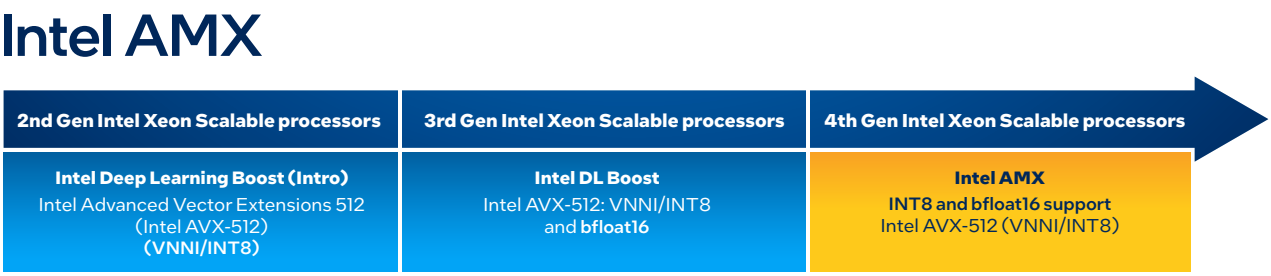

Figure 3. Innovations like Intel AMX build on the accelerators built into previous-generation Intel Xeon Scalable processors, enabling even greater performance

3rd Gen Intel Xeon Scalable processors added native support for the bloat16 (BF16) format, which can improve AI training efficiency. 4th Gen Intel Xeon Scalable processors further improve upon the previous generations by offering built-in accelerators, including Intel® Advanced Matrix Extensions (Intel® AMX). Intel AMX improves the performance of DL training and inference. This accelerator enables you to run AI inference on the CPU instead of offloading the workload to discrete accelerators, which can provide a significant performance boost.[11] The Intel AMX architecture also supports the BF16 and INT8 data types.

Success with AI

MetaApp worked with Alibaba Cloud and Intel to develop a new AI-based recommendation system to help its game-developer customers better monetize their games. The system is faster and costs less. For AI training (fine-tuning), it delivers 64 percent higher performance and 160 percent higher training performance per price than the previous system.[8] AI inferencing uses 25 percent fewer cores at 22 percent less cost than the previous system.[1] Additionally, software optimizations for Intel hardware make dynamic scheduling and flexible scaling possible while bypassing the need for GPUs.

MetaApp's journey is not unique. To grow and succeed, today's digitally native businesses, from e-commerce to social media platforms, must enhance customer experiences through personalized services. AI-based recommendation systems play a pivotal role. Alibaba Cloud and Intel offer services and technologies to help customers like MetaApp build recommendation systems that are fast, efficient, cost-effective, and flexible.

Get started today

Bring AI everywhere with Intel® artificial intelligence (AI) solutions.

Learn how to use AI to enhance the relevance and value of recommendation systems.

[1] See [P12] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.
[2] Source: Interview with Intel account manager. January 2024.
[3] Canalys. "Mainland China cloud service spending grew 6% in Q1 2023." June 2023. canalys.com/newsroom/china-cloud-market-Q1-2023.
[4] Rocketset. "A Blueprint for a Real-World Recommendation System." December 2023. rockset.com/blog/a-blueprint-for-a-real-world-recommendation-system/.
[5] Intel. Intel® Accelerator Engines webpage. intel.com/content/www/us/en/products/docs/accelerator-engines/overview.html.
[6] 4th Gen Intel Xeon Scalable processor: 8 channels DDR5, up to 4,800 MT/s (1 DPC) vs. 3rd Gen Intel Xeon Scalable processor: 8 channels DDR4, 3,200 MT/s (2 DPC).
[7] Intel. "Intel® oneAPI Deep Neural Network Library." Accessed February 2024. intel.com/content/www/us/en/developer/tools/oneapi/onednn.html#gs.4bydv5.
[8] See [P13] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.
[9] Intel. Intel interview with MetaApp. February 2024.
[10] See [E1] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.
[11] See [A16] and [A17] at intel.com/processorclaims: 4th Gen Intel Xeon Scalable processors. Results may vary.

Printed in USA 0224/TB/PRW/PDF Please recycle 358951-001US