Solution Brief

Deep Learning Models
4th Gen Intel Xeon Scalable processors

Intel® Advanced Matrix Extensions (Intel® AMX) Enhances AI Inference Performance for Alibaba Cloud Address Purification

4th Gen Intel Xeon Scalable processors with Intel AMX boost end-to-end inferencing performance 2.48x as compared to the previous generation.[1]

As an important technique of artificial intelligence (AI), deep learning (DL) has been widely implemented in many areas, such as computer vision (CV), natural language processing (NLP) and recommendation systems. However, with the explosive growth of data and the increasing complexity of DL models, using inferencing in production can be challenging. Users expect to optimize hardware, software, and algorithms to improve performance and reduce overall cost. Optimizing DL inferencing helps users adopt more complex DL models to improve accuracy while maintaining the same latency.

To improve the performance of address-purification services, Alibaba Cloud's machine learning platform (PAI) and the Alibaba Academy for Discovery, Adventure, Momentum and Outlook (DAMO Academy) NLP team collaborated with Intel. 4th Gen Intel® Xeon® Scalable processors, with Intel AMX, along with optimization tools, improved end-to-end inferencing by up to 2.48 times, compared to using a previous-generation platform.[1]

Alibaba Cloud Address Purification

Address purification is the automated process of standardizing, correcting, and validating postal address. It is used in many industries including logistics, e-commerce, retail, and finance. Alibaba Cloud Address Purification is an efficient standard address algorithm as a service (AaaS) developed by the NLP team of Alibaba DAMO Academy based on Alibaba Cloud's enormous address collection.[2] Faster end-to-end performance translates to better business results for Alibaba Cloud's customers. This AaaS is a one-stop, closed-loop service platform for address data processing. It uses the NLP algorithm to correct, complete, normalize, structurize, and label the address data registered in business systems. It supplies more than 20 types of address services[3] and can be deployed on public, private, or hybrid clouds. Alibaba Cloud objectives are:

- Accelerate one-stop performance of the platform with an overall consideration in multiple workloads such as data cleaning and model inference
- Use existing hardware resources efficiently and make full use of customers' server resources in public, private, and hybrid clouds to reduce hardware costs

Optimizing Alibaba service with Intel® technology

The Bidirectional Encoder Representations from Transformers (BERT) model is a deep learning technique for natural language processing (NLP) that helps artificial intelligence (AI) programs understand the context of ambiguous words in text. Alibaba uses BERT as the search module for its Address Purification service.[4] It is used for multitask vector recall and fine sorting. Intel provides a variety of solutions that can help significantly accelerate the solution's performance.

Intel AMX

4th Gen Intel Xeon Scalable processors offer a built-in accelerator called Intel AMX that helps Alibaba Cloud Address Purification to achieve outstanding performance, cost effectiveness, and scalability. The 4th Gen Intel Xeon Scalable processors with Intel AMX can be deployed in a wide range of DL use cases including recommender systems, natural language processing, and retail e-commerce software solutions.

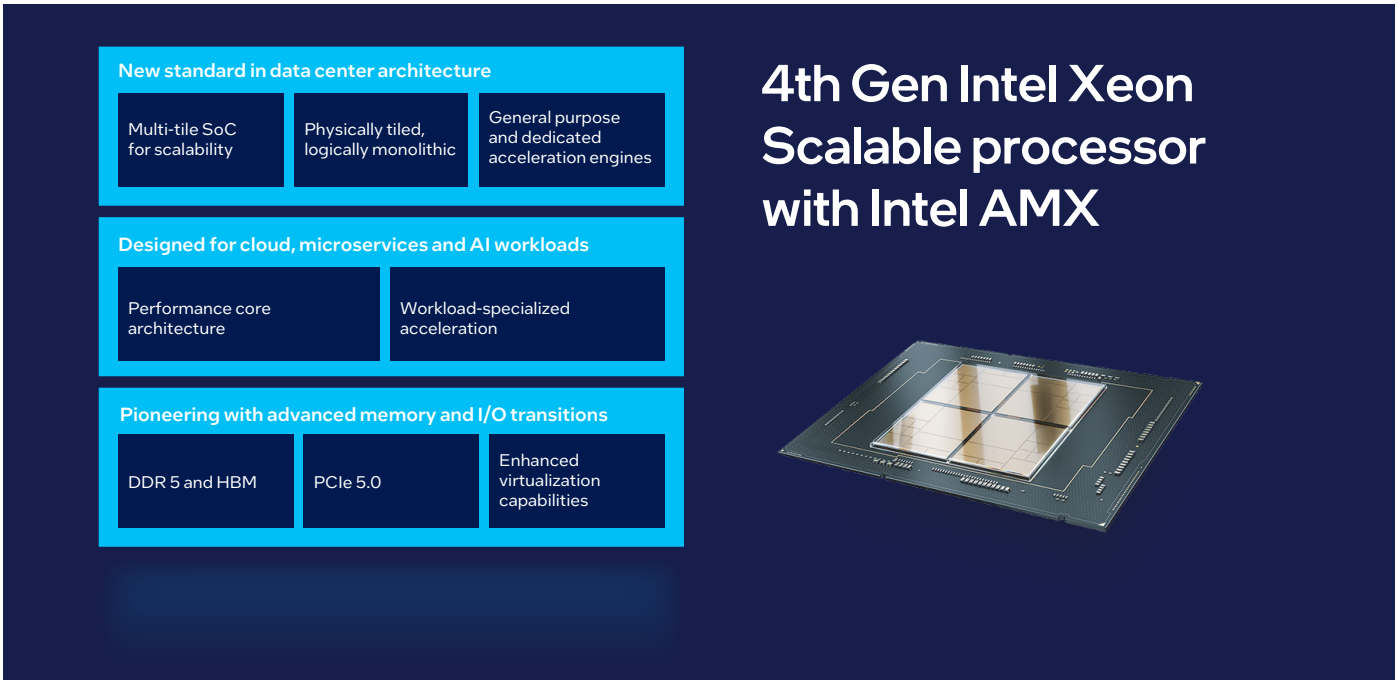

Figure 1. 4th Gen Intel Xeon Scalable processor with Intel AMX

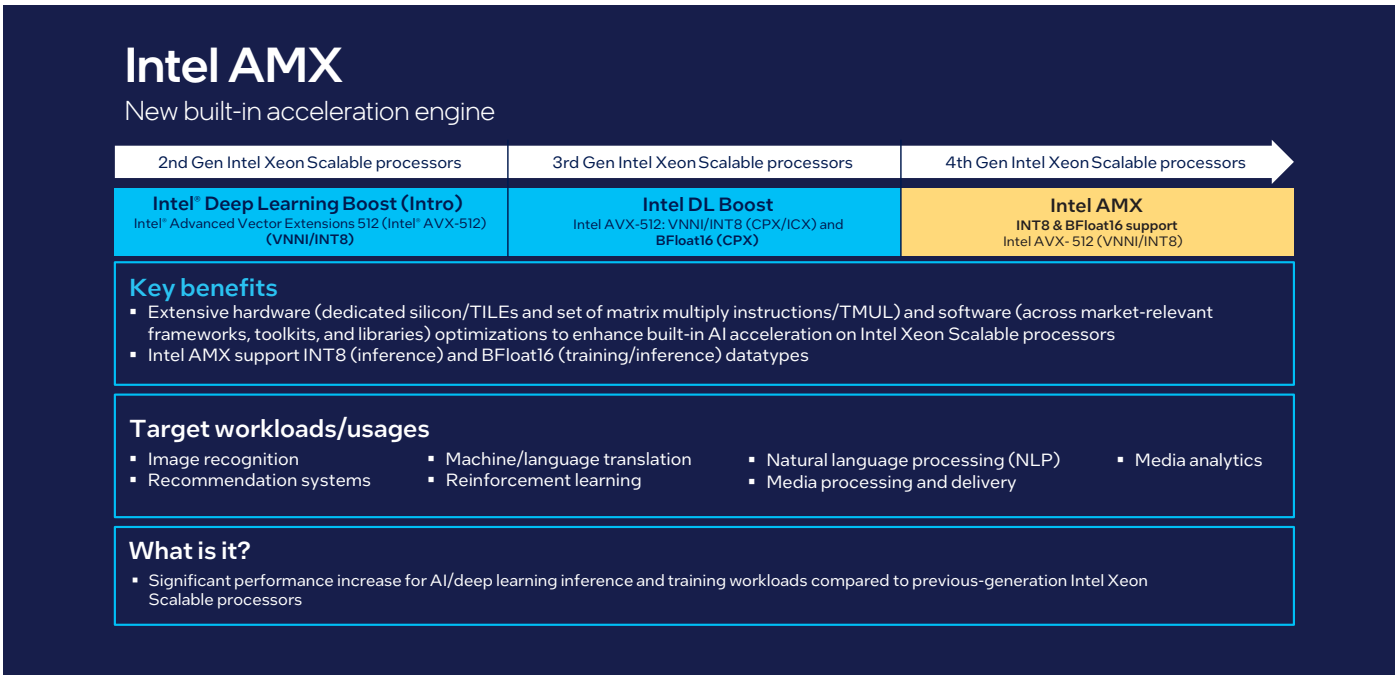

Figure 2. Intel AMX overview

Blade: A General Tool for Inference Optimization

The Alibaba Cloud Address Purification solution adopts Blade, a general inference optimization tool introduced by Alibaba Cloud Machine Learning PAI team, to optimize the inference performance of address purification. Blade integrates a number of optimization methods, including computational graph optimization, optimization libraries like Intel® oneAPI Deep Neural Network Library (Intel® oneDNN), the BladeDISC compiler, the Blade high-performance operator library, the Intel custom backend, and Blade mixed precision.

Integrating Intel Custom Backend into Blade

Intel custom backend,[5] as the software backend of Blade, boosts the model performance in terms of quantization and sparsification inference. Intel custom backend mainly contains three levels of optimization: first, use the primitive cache strategy to optimize memory; second, optimize the graph fusion; finally, at the operator level, it builds an efficient operator library including custom and sparse kernels.

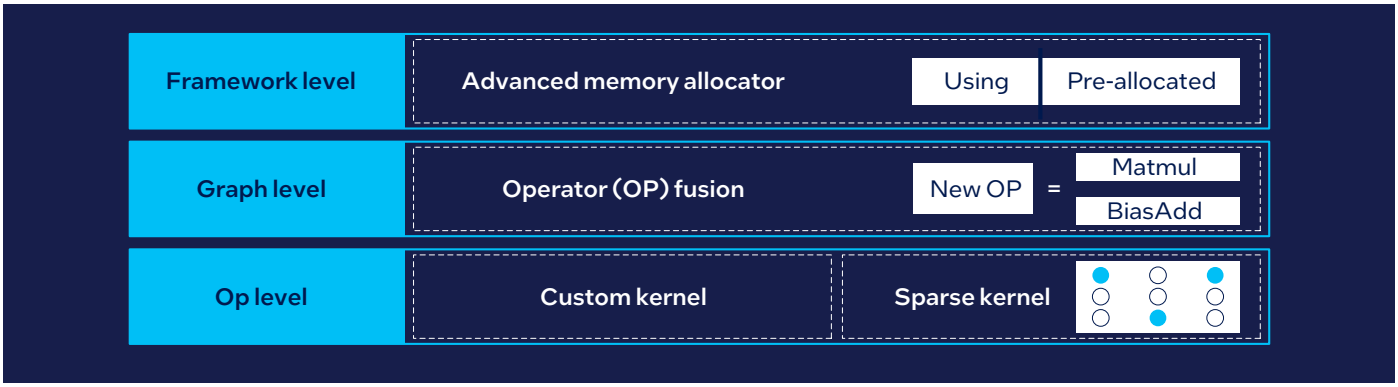

Figure 3. Structure of Intel custom backend

Starting from the second generation, Intel Xeon Scalable processors provided VNNI specifically for INT8 quantization, which optimizes AI performance based on the INT8 data type and is widely used in model quantization solutions.

Intel AMX greatly improves the capability of INT8, and is supported by making use of Intel oneDNN. INT8 quantization based on Intel AMX can significantly improve the model performance compared to VNNI. Figure 4 describes how Intel AMX works.

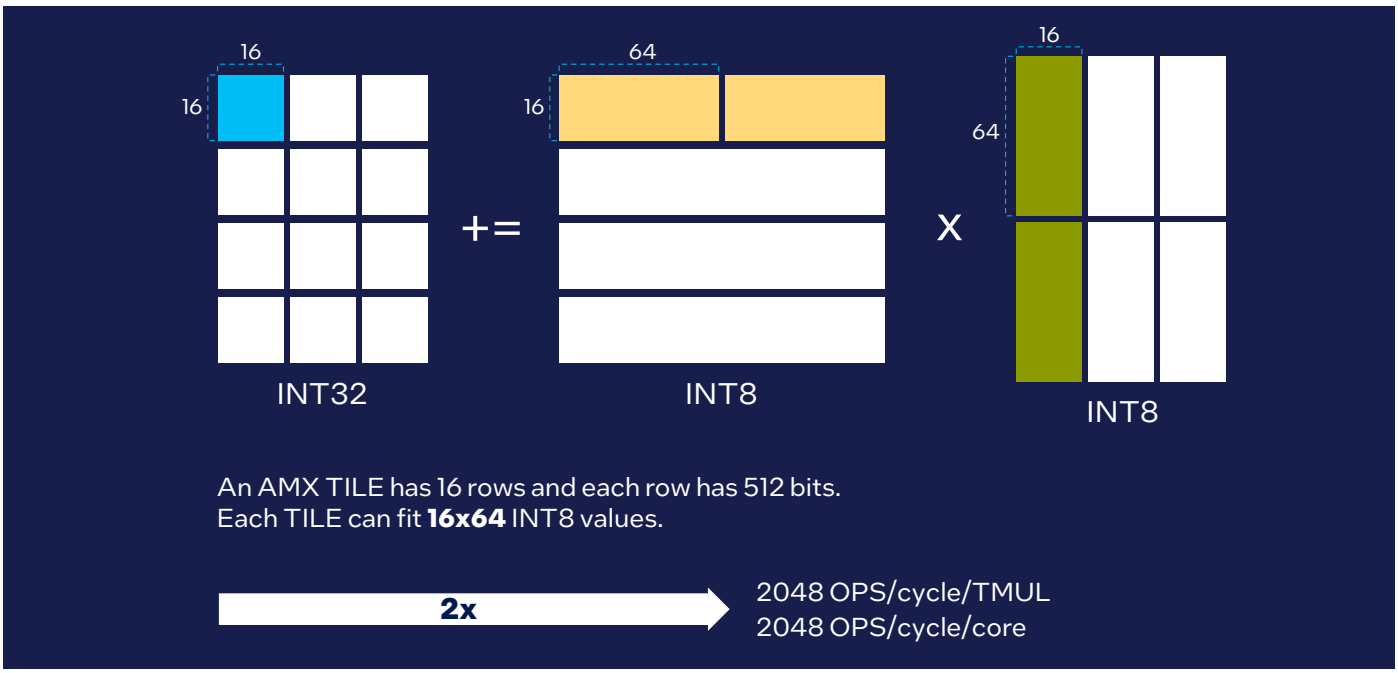

Figure 4. Intel Advanced Matrix Extensions (Intel AMX) capability

Performance and Optimization Gains

Alibaba Cloud and Intel have also tuned the address purification models to improve their inference performance, achieving with PAI a 2.48x performance gain when using 4th Gen Intel Xeon Scalable processors with Intel AMX, compared to the previous-generation platform.[1] The Intel AMX-based Intel custom backend optimizes the 4-layer BERT model of a fixed shape size (10 x 53) to achieve the gain. See Figure 5.

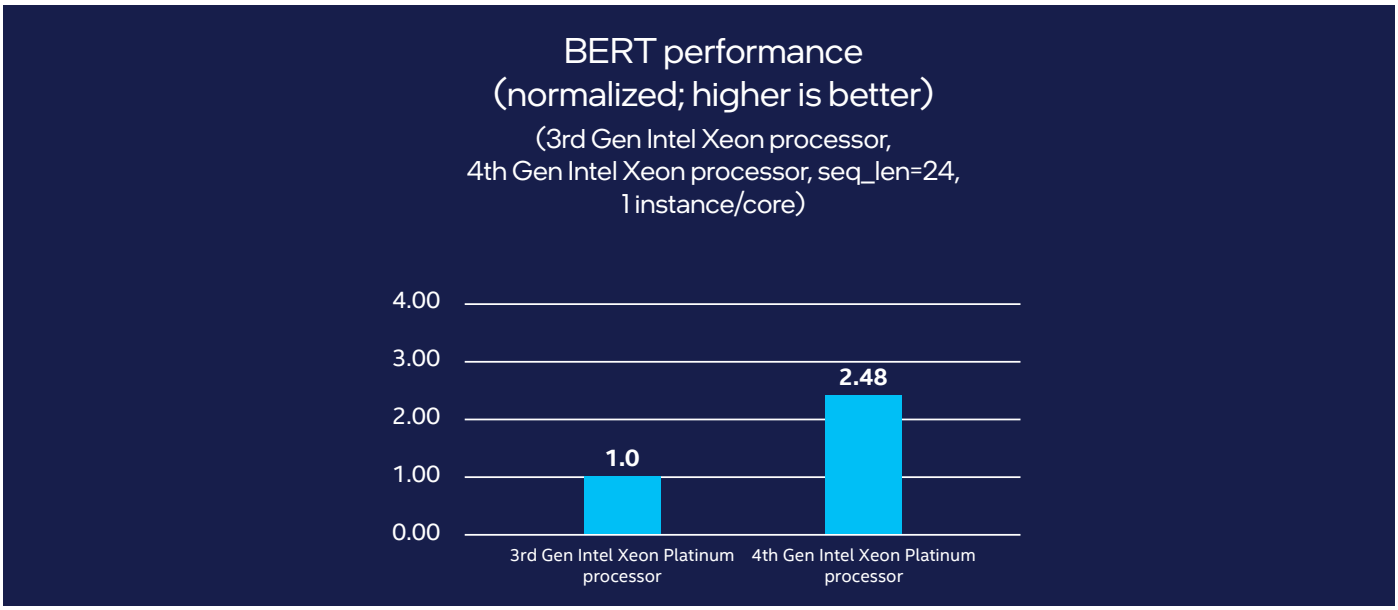

Figure 5. Inference performance of BERT model[1]

The CCKS2021 Chinese NLP address-correlation task was used to verify the models. The floating-point 32 (FP32)-based optimization scored 78.72, while the INT8-based optimization scored 78.85.[6] The higher score is better.

Summary

Alibaba Cloud used the 4th Gen Intel Xeon Scalable processor with Intel AMX to optimize the AI inference of its address purification service. Faster end-to-end performance translates to better business results for Alibaba's customers in logistics, e-commerce, energy, retail, and finance. Intel AMX also reduces the overhead Alibaba might have if the company had deployed dedicated accelerators such as independent GPUs. Using a built-in accelerator helps Alibaba control the total cost of ownership (TCO) of its address purification service.

To boost the end-to-end performance of additional DL models, Intel and Alibaba are expanding cooperation with their customers to optimize software and hardware integration. The goal is to accelerate the performance of DL models and tap the value of Intel technologies to the greatest extent. Intel also expects to have more in-depth cooperation with industry partners and to contribute to the deployment and implementation of AI technology.

[1] Configuration: BASELINE: Test by Intel as of 10/19/2022. 1-node, 2x 3rd Gen Intel Xeon Platinum processor, Intel® Hyper-Threading Technology (Intel® HT Technology) on, Intel® Turbo Boost Technology on, total memory 256 GB (16 slots/16 GB/3,200 MT/s [run @ 3,200 MHz]), WLYDCRB1.SYS.0029.P30.2209011945, 0xd00037b, CentOS Linux 8, 4.18.0-305.12.1.el8_4.x86_64, GCC 8.5.0, NLP Toolkit v0.3, Pytorch 1.11, BERT-mini, INC 1.13, transformer 4.18.0, 1 instance/core,BS=32, seq_len=24, Datatype: INT8

NEW-1: Test by Intel as of 10/19/2022. 1-node, 2x 4th Gen Intel Xeon Platinum processor, Intel HT Technology on, Intel Turbo Boost Technology on, total memory 256 GB (16 slots/16 GB/4,800 MHz [run @ 4,800 MHz]), EGSDCRB1.SYS.0090.D03.2210040200, 0x2b0000c0, CentOS Stream 8, 5.15.0-spr.bkc.pc.8.8.5.x86_64, GCC 8.5.0, NLP Toolkit v0.3, Pytorch 1.11, BERT-mini, INC 1.13, transformer 4.18.0, 1 instance/core, BS=32, seq_len=24, Datatype:INT8.

[2] Alibaba Cloud. "Address Normalization." aliyun.com/product/addresspurification/addrp.

[3] Alibaba Cloud. "What is Address Normalization?" https://help.aliyun.com/document_detail/169746.html.

[4] Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." ACL Anthology. June 2019. https://aclanthology.org/N19-1423/.

[5] GitHub. "Intel Neural Compressor." https://github.com/intel/neural-compressor/commits/inc_with_engine.

[6] Alibaba Cloud. "'Intel Innovation Master Cup' Deep Learning Challenge Track 3: CCKS2021 Chinese NLP Address Correlation Task." November 2021. https://tianchi.aliyun.com/competition/entrance/531901/introduction.