Case Study

4th Gen Intel® Xeon® Scalable Processors Help Meituan Accelerate Vision AI Inference Services and Optimize Costs

Challenges

For Meituan, vision AI has become the key to driving business model innovation, delivering more accurate and personalized Internet services to users, and enhancing competitive advantages. However, Meituan's vision of AI inference also faces various challenges in computing power and costs.

Performance

As Meituan's business and user base continue to snowball, more applications require the development of intelligent processes through vision AI. Meituan needs to improve the throughput of its vision AI inference without compromising accuracy to support more intelligent operations.

Costs

Huge infrastructure investments are required to perform vision AI inference on massive data. While discrete GPUs can meet performance requirements, their price is relatively high. For low-traffic long-tail model inference services, CPUs are often more cost-effective.

Flexibility

Meituan hopes to improve the agility of its vision AI services and meet the AI inference requirements of long-tail scenarios through flexible resource scheduling across multiple architectures.

Solution Overview

At present, computer vision powered by artificial intelligence (AI) has become a critical method for companies to obtain data insights and drive the intelligent transformation of their business. By leveraging enhanced deep learning neural networks, vision AI captures data in a more sophisticated manner and takes analytics to a new level, optimizing work efficiency, reducing costs, improving revenues, and enhancing customer satisfaction. With the explosive growth of visual data and continued business development, companies hope to obtain higher returns on investment while accelerating the training of computer vision AI models and improving inference performance.

As a leading retail technology company, Meituan is committed to the mission to help people eat better and live better through a "retail + technology" strategy. Meituan attaches great importance to using innovative vision AI technology to empower businesses such as catering, travel, tourism, shopping, and entertainment. To accelerate AI inference, Meituan utilizes advanced hardware capabilities such as 4th Gen Intel® Xeon® Scalable processors and their built-in Intel® Advanced Matrix Extensions (Intel® AMX). Through methods such as converting models from FP32 to BF16, the inference performance of conventional vision models can be improved by around 3.38-4.13x[1]. By combining these technologies with header service optimization strategies such as dynamic scaling, Meituan has increased the overall efficiency of its online resources by over 3x and saved 70% on service costs[2].

Meituan Utilizes Vision AI Applications to Empower Intelligent Transformation

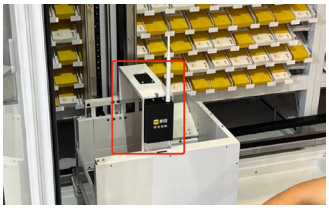
Vision AI has penetrated content creation, content review, distribution, user interaction, value realization, and other links of Meituan's business. Meituan vision AI empowers the industry with scenario-based products and innovative tools. It has been successfully applied in scenarios such as intelligent image processing, merchant registration certificate identification, QR code bike lock, pharmaceutical package scanning, identity verification, and more to enable the intellectual transformation of businesses.

Open catering platform vision services

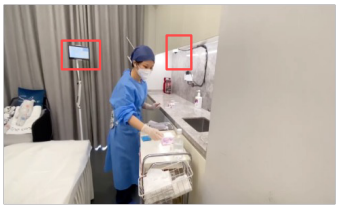
Merchant registration certificate identification

QR code bike lock

Pharmaceutical package scanning



Unmanned medicine warehouse dispensing monitoring



Cosmetology "beauty assured" performance monitoring

Figure 1. Typical applications of vision AI technology in Meituan businesses

To further optimize vision AI inference services, Meituan transformed the original algorithm service process into a microservices-based vision AI algorithm process, separating CPU services from other accelerator services to ensure that different workloads can be run on different devices. Scheduling is performed through the scheduling service in the middle layer, thereby improving hardware utilization.

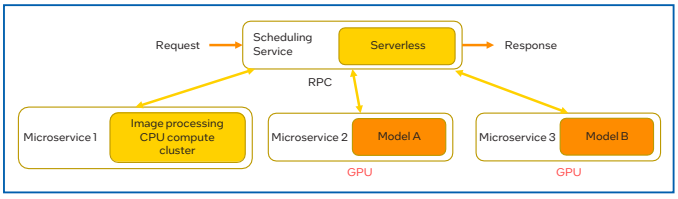


Figure 2. Meituan microservices-based vision AI algorithm process flow

Facing the cost challenge of vision AI inference services, Meituan adopted a CPU-based strategy for its low-traffic long-tail model inference services. Intel® Xeon® Scalable processors were used to handle model inference services with relatively low workload pressure and latency requirements. This approach ensures that Meituan can efficiently utilize existing CPU resources and reduce GPU deployment and O&M costs.

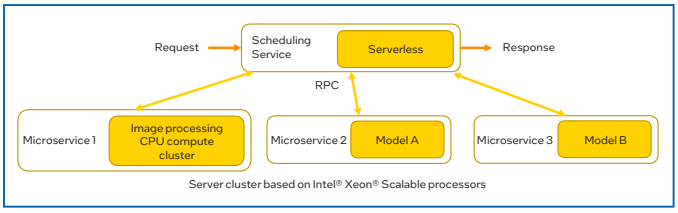


Figure 3. Meituan CPU-based low-traffic long-tail model inference service

Meituan also adopts a traffic-sensitive header service optimization strategy that separates online operations that are sensitive to delay, require high stability, and have fluctuating traffic from offline batch processing operations that are not sensitive to delay, require low stability, and have uniform traffic. Through dynamic scaling, many resources are freed up during off-peak periods for offline batch processing, thereby saving on resources and improving the overall AI inference performance.

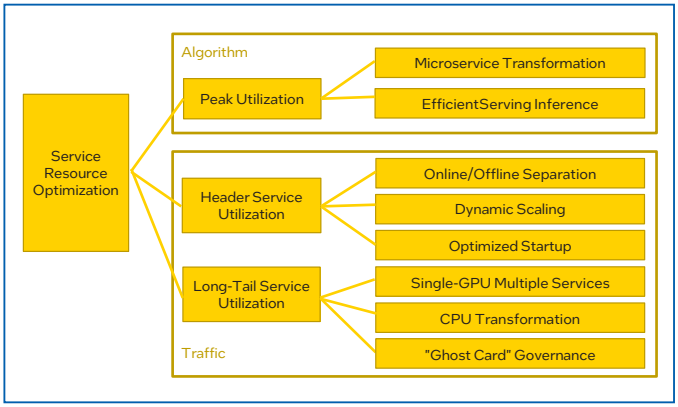


Figure 4. Meituan's vision AI inference optimization strategy

4th Gen Intel® Xeon® Scalable Processors Empower the Inference Performance of Vision AI

To further accelerate the performance of vision AI inference services, Meituan utilized 4th Gen Intel® Xeon® Scalable processors, the integrated Intel® AMX accelerator, and software suites such as Intel® Integrated Performance Primitives (Intel® IPP) for optimization.

4th Gen Intel® Xeon® Scalable processors increase the instructions per cycle (IPC) through innovative architecture. With up to 60 cores per socket and support for 8-channel DDR5 memory, the processors improve memory bandwidth and speed while achieving higher memory bandwidth per PCIe 5.0 (80 channels). 4th Gen Intel® Xeon® Scalable processors deliver modern performance and security, along with the ability to scale with business demand. With built-in accelerators, the processors provide users with optimized performance across AI, analytics, cloud and microservices, networking, databases, storage, and other workloads. When combined with a robust ecosystem, 4th Gen Intel® Xeon® Scalable processors help users build more efficient and secure infrastructure.

4th Gen Intel® Xeon® Scalable processors take AI performance to the next level and come equipped with the innovative Intel® AMX accelerator. Unlike the Intel® AVX-512 provided in previous Intel® Xeon® Scalable processors, Intel® AMX adopts a new instruction set and circuit design. Providing matrix operations significantly increases the instructions per cycle of AI applications and empowers the performance of the training and inference of AI workloads.

In real-world workloads, Intel® AMX can support BF16 and INT8 data types. BF16 has the same dynamic range as standard IEEE-FP32 but with lower precision than FP32. In most cases, BF16 has the same model inference precision as FP32, but as BF16 only needs to process data half the size of FP32, its throughput is much higher, and memory resource requirements are much lower. Intel® AMX can achieve 2048 INT8 operations and 1024 BF16 operations per cycle per physical core[3], significantly improving the efficiency of AI workloads compared to Intel® AVX-512 acceleration technology.

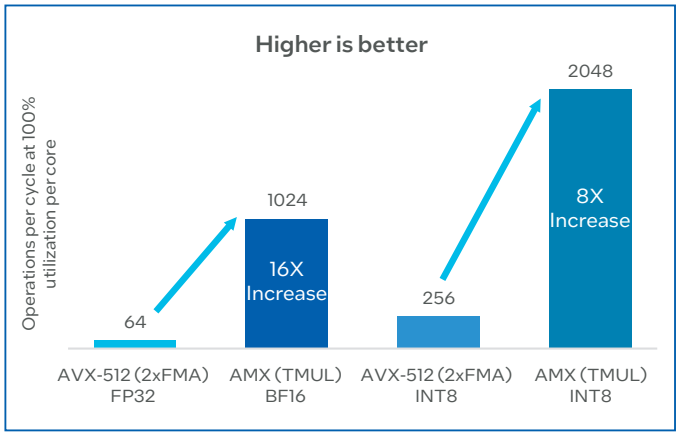


Figure 5. Intel® AMX offers significant matrix operation efficiency improvements over Intel® AVX-512[3]

Meituan also integrates Intel® Extension for PyTorch (Intel® IPEX) to accelerate PyTorch. Intel® IPEX is an open-source extension project launched by Intel, implemented based on the PyTorch extension mechanism. By providing additional software optimization to utilize hardware features fully, Intel® IPEX improves the computing performance of deep learning inference and training on Intel® processors through native PyTorch.

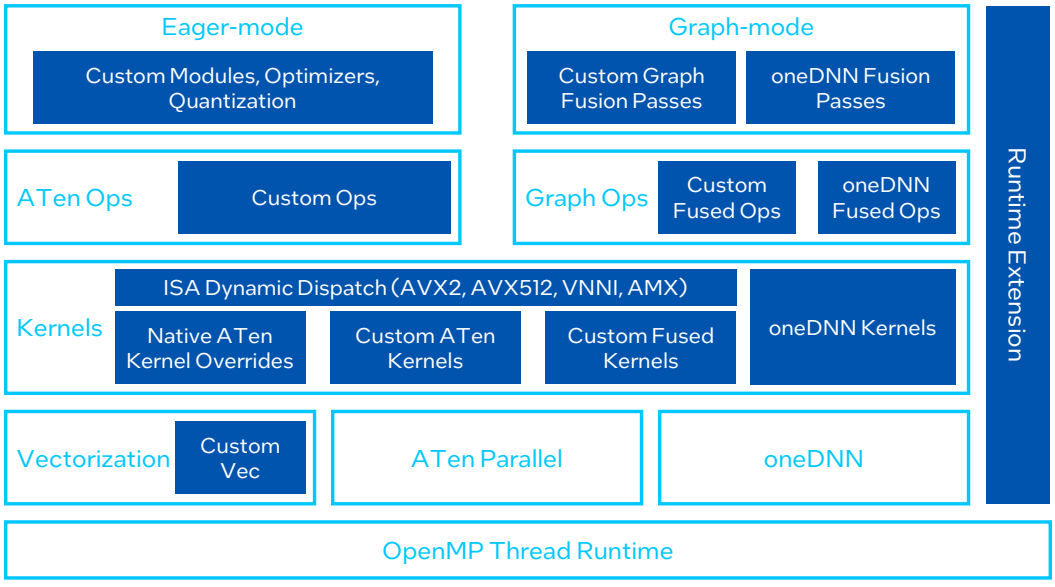


Figure 6. Intel® IPEX architecture

Meituan dynamically converts the data type of a variety of vision AI models from FP32 to BF16 with Intel® AMX acceleration technology to increase throughput and accelerate inference with an acceptable loss of precision. Meituan compared the inference performance of the BF16 model converted using Intel® AMX acceleration technology to verify post-optimization performance with the baseline FP32 model. As shown in the test data in Fig. 7, the inference performance of the model can be improved by 3.38-4.13x after conversion to BF16. Most of the precision loss of the "Top1" and "Top5" sections can be controlled within the 0.01%-0.03% range[1].

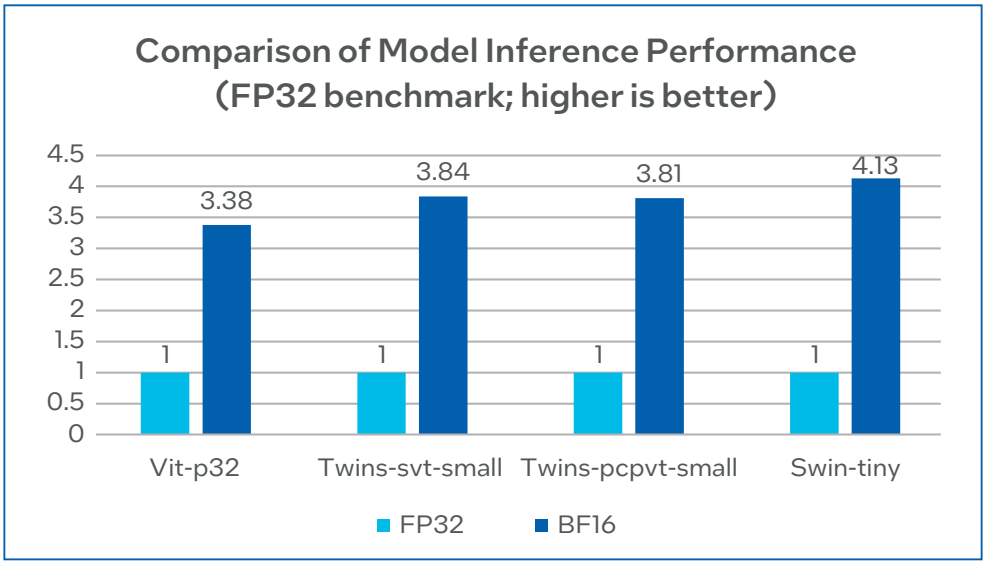


Figure 7. FP32/BF16 model inference performance comparison[1]

Benefits

Through adopting 4th Gen Intel® Xeon® Scalable processors and integrating optimization strategies such as microservices transformation, online/offline separation, dynamic scaling, and optimized startup, Meituan has obtained various benefits.

- The post-optimization performance of Vit-p32, Twins-svt-small, Twins-pcpvt-small, Swin-tiny, and other models has been increased by 3.38-4.13x, with the majority of the precision loss of the "Top1" and "Top5" sections can be controlled within the 0.01%-0.03% range[1]. The overall efficiency of online resources has been improved by over 3x[2].
- Empowered by improved performance, Meituan can take full advantage of the potential of existing infrastructure, reduce the investment requirement of vision AI services, and reduce service costs by 70%[2].
- Agile resource scheduling supports the efficient innovation of vision AI services.

Outlook

Meituan's vision AI inference optimization shows that 4th Gen Intel® Xeon® Scalable processors integrated with the Intel® AMX acceleration engine can enhance AI inference performance and reduce the total cost of ownership (TCO) of vision AI inference services. Meituan and Intel are also committed to using hardware innovation and software optimization to continuously improve inference performance and fully unlock the value of vision AI services.

As the trend of intelligent and digital transformation continues, Intel will further cooperate with Meituan and other partners to empower business innovation with computing, storage, network, and other capabilities, accelerate the development of the AI sector, and drive the implementation of AI technology and practice. Intel and its partners aim to provide more accurate and personalized services for end users, all the while reducing the performance, cost, and technical threshold of AI deployment and driving the intelligent transformation of the industry.

About Meituan

As a tech-driven retail company, Meituan has a strategic focus on "Retail + Technology" and adheres to our mission of "We help people eat better, live better." Since its establishment in March 2010, Meituan has advanced the digital upgrading of services and goods retail on both supply and demand sides. Together with our partners, we provide quality services for consumers. On 20 September 2018, Meituan was listed on the Main Board of the Stock Exchange of Hong Kong Limited. Meituan has always been a customer-obsessed company, and we will continue to increase our R&D investment in new technologies. Meituan will collaborate with all partners to fulfill our social responsibility and create more value for society.

About Intel

Intel (Nasdaq: INTC) is an industry leader, creating world-changing technology that enables global progress and enriches lives. Inspired by Moore's Law, we continuously work to advance the design and manufacturing of semiconductors to help address our customers' greatest challenges. By embedding intelligence in the cloud, network, edge and every kind of computing device, we unleash the potential of data to transform business and society for the better. To learn more about Intel's innovations, go to newsroom.intel.com and intel.com.

intel.

---

[1] Data from internal test results conducted by Meituan in October 2022. Test Configuration: BASELINE-1 - 2-node, 2x Intel® Xeon® Platinum 8468V processors, 48 cores, HT On, Turbo On, Total Memory 256 GB (16 slots/ 16 GB/ 4800 MHz), <EGSDCRB1.SYS.8901.P01.2209200243>, <0x2b0000a1>, <CentOS Stream 8>, <5.16.0-intel-next-01783-g51456e>, <gcc 11.2.1>, <ViT-p32>, <Stock PyTorch v1.12.0>, <ViT-p32>, score=544.40<fps>; NEW-1 - 2-node, 2x Intel® Xeon® Platinum 8468V processors, 48 cores, HT On, Turbo On, Total Memory 256 GB (16 slots/ 16 GB/ 4800 MHz), <EGSDCRB1.SYS.8901.P01.2209200243>, <0x2b0000a1>, <CentOS Stream 8>, <5.16.0-intel-next-01783-g51456e>, <gcc 11.2.1>, <ViT-p32>, <Intel-Extension-for-PyTorch v1.12.300>, <ViT-p32>, <OneDNN 2.6>, score=1839.85<fps>; BASELINE-2 - 2-node, 2x Intel® Xeon® Platinum 8468V processors, 48 cores, HT On, Turbo On, Total Memory 256 GB (16 slots/ 16 GB/ 4800 MHz), <EGSDCRB1.SYS.8901.P01.2209200243>, <0x2b0000a1>, <CentOS Stream 8>, <5.16.0-intel-next-01783-g51456e>, <gcc 11.2.1>, <Twins-svt-small>, <Stock PyTorch v1.12.0>, <Twins-svt-small>, score=626.91<fps>; NEW-2 - 2-node, 2x Intel® Xeon® Platinum 8468V processors, 48 cores, HT On, Turbo On, Total Memory 256 GB (16 slots/ 16 GB/ 4800 MHz), <EGSDCRB1.SYS.8901.P01.2209200243>, <0x2b0000a1>, <CentOS Stream 8>, <5.16.0-intel-next-01783-g51456e>, <gcc 11.2.1>, <Twins-svt-small>, <Intel-Extension-for-PyTorch v1.12.300>, <Twins-svt-small>, <OneDNN 2.6>, score=2409.37<fps>; BASELINE-3 - 2-node, 2x Intel® Xeon® Platinum 8468V processors, 48 cores, HT On, Turbo On, Total Memory 256 GB (16 slots/ 16 GB/ 4800 MHz), <EGSDCRB1.SYS.8901.P01.2209200243>, <0x2b0000a1>, <CentOS Stream 8>, <5.16.0-intel-next-01783-g51456e>, <gcc 11.2.1>, <Twins-pcpvt-small>, <Stock PyTorch v1.12.0>, <Twins-pcpvt-small>, score=550.10<fps>; NEW-3 - 2-node, 2x Intel® Xeon® Platinum 8468V processors, 48 cores, HT On, Turbo On, Total Memory 256 GB (16 slots/ 16 GB/ 4800 MHz), <EGSDCRB1.SYS.8901.P01.2209200243>, <0x2b0000a1>, <CentOS Stream 8>, <5.16.0-intel-next-01783-g51456e>, <gcc 11.2.1>, <Twins-pcpvt-small>, <Intel-Extension-for-PyTorch v1.12.300>, <Twins-pcpvt-small>, <OneDNN 2.6>, score=2094.38<fps>; BASELINE-4 - 2-node, 2x Intel® Xeon® Platinum 8468V processors, 48 cores, HT On, Turbo On, Total Memory 256 GB (16 slots/ 16 GB/ 4800 MHz), <EGSDCRB1.SYS.8901.P01.2209200243>, <0x2b0000a1>, <CentOS Stream 8>, <5.16.0-intel-next-01783-g51456e>, <gcc 11.2.1>, <Swin-tiny>, <Stock PyTorch v1.12.0>, <Swin-tiny>, score=382.82<fps>; NEW-4 - 2-node, 2x Intel® Xeon® Platinum 8468V processors, 48 cores, HT On, Turbo On, Total Memory 256 GB (16 slots/ 16 GB/ 4800 MHz), <EGSDCRB1.SYS.8901.P01.2209200243>, <0x2b0000a1>, <CentOS Stream 8>, <5.16.0-intel-next-01783-g51456e>, <gcc 11.2.1>, <Swin-tiny>, <Intel-Extension-for-PyTorch v1.12.300>, <Swin-tiny>, <OneDNN 2.6>, score=1579.61<fps>. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

[2] Data from internal test results conducted by Meituan. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.

[3] Data from the Intel official website. For more details, visit https://edc.intel.com/content/www/us/en/products/performance/benchmarks/architecture-day-2021/

Performance varies by use, configuration and other factors. Learn more at www.Intel.com/PerformanceIndex

Performance results are based on testing as of dates shown in configurations and may not reflect all publicly available updates. See backup for configuration details. No product or component can be absolutely secure.

Your costs and results may vary.

Intel technologies may require enabled hardware, software or service activation.

Intel disclaims all express and implied warranties, including without limitation, the implied warranties of merchantability, fitness for a particular purpose, and non-infringement, as well as any warranty arising from course of performance, course of dealing, or usage in trade.

Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.