UP TO 4.13X improvement

in inference performance of conventional vision models from converting models from FP32 to BF16.[1]

3X increase

in overall efficiency of online resources and saved 70% on service costs.[2]

Meituan Accelerates Vision AI Inference Services and Optimizes Costs

For Meituan, vision AI has become the key to driving business model innovation, delivering more accurate and personalized services to users, and enhancing competitive advantages. However, Meituan's vision of AI inference also faces various challenges in computing power and costs. Meituan needs to improve the throughput of its vision AI inference without compromising accuracy to support more intelligent operations. 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. To accelerate AI inference, Meituan utilizes advanced hardware capabilities such as 4th Gen Intel® Xeon® Scalable processors and the built-in Intel® Advanced Matrix Extensions (Intel® AMX). By combining these technologies with header service optimization strategies such as dynamic scaling, Meituan has increased the overall efficiency of its online resources.

Products and Solutions	Industry	Organization Size	Country	Learn more
4th Gen Intel® Xeon® Scalable Processors Intel® Advanced Matrix Extensions Intel® Integrated Performance Primitives	Internet, E-commerce	10,001+	China	Case Study

1, 2 For more complete information about performance and benchmark results, visit https://www.intel.com/content/www/us/en/customer-spotlight/stories/meituan-vision-ai-customer-story.html