Case Study

Intel® Xeon® Processors

Netflix Chooses Amazon EC2 Instances with Intel® Xeon® Processors to Provide Fast and Seamless Streaming Experiences

Intel technologies supporting AI help deliver performant, high-quality video content and micro-services to subscribers while significantly reducing Netflix's cloud spend.

Solution Summary

- Intel® Xeon® Processors
- Intel® oneAPI Deep Neural Network Library
- Intel® Deep Learning Boost with Vector Neural Network Instructions (VNNI)
- Intel® Advanced Vector Extensions 512 (Intel® AVX 512)
- Intel® VTune™ Profiler
- Intel® PerfSpect
- Amazon EC2 Instances

Executive Summary

Netflix endeavors to transform home entertainment and provide 260 million subscribers with reliable, customized experiences on any device. To accomplish that, Netflix must accelerate data movement and AI workloads using advanced technologies, including Amazon EC2 instances supported by Intel® Xeon® processors. Working with Intel, Netflix:

- Optimized Amazon instances at the micro-architecture level to increase performance and reduce cloud spend. After upgrading its EC2 instances, Netflix achieved a 3.5x performance improvement per CPU, exceeding anticipated linear scaling.[1]
- Used Intel® oneAPI Deep Neural Network Library (oneDNN) and the Intel AVX instruction set to optimize video encoding speed during hours of lower user demand. Intel solutions provided a significant improvement in frame-per-second encoding.

Challenge

Netflix seeks to deliver seamless, on-demand content to its global customer base regardless of the device they use to view it. The process requires multiple micro-services optimized for the workloads supporting subscriber experiences. Some backend microservices must handle content development, rendering, and encoding tasks. On the user-facing end, Netflix subscribers need a tailored home page view that identifies and recommends the most relevant content from thousands of

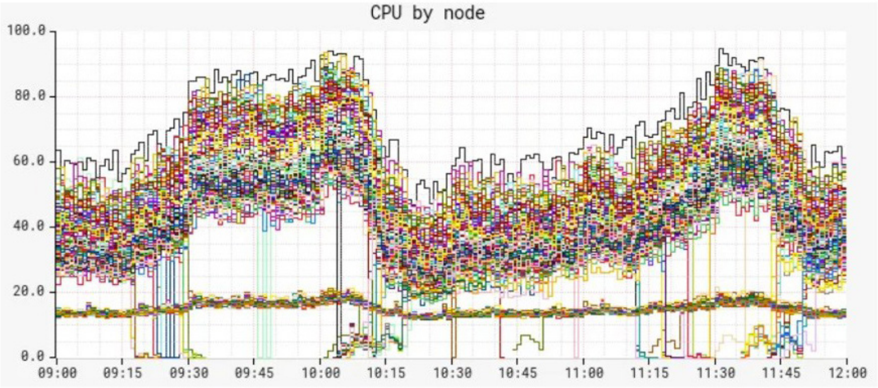


Figure 1. The graph illustrates the breakdown of CPU by node. While the analysis found roughly equal traffic distribution between nodes, CPU metrics demonstrated different bimodal distribution patterns.[1]

titles. Plus, Netflix continuously strives to serve its members with excellent entertainment and exceptional streaming quality, available anytime.

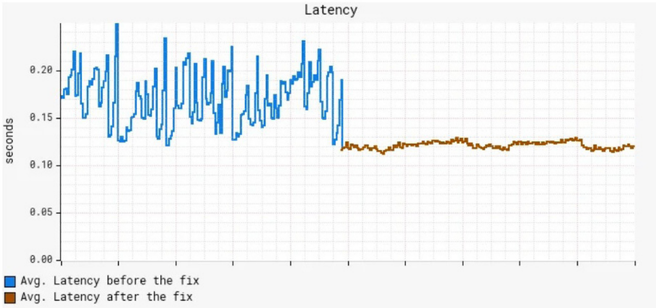
To accomplish all these tasks and more, Netflix requires a reliable, highly scalable, and AI-ready cloud solution with advanced tools to simplify troubleshooting should problems arise. For example, the Netflix team discovered an unexpected latency challenge while evaluating its Amazon EC2 instances for performance. They needed an effective way to assess instances down to the CPU's micro-architecture level to accelerate workloads while minimizing cloud spend.

Solution

To provide subscribers with fast and customized streaming experiences, Netflix tapped the power of Amazon EC2 instances supported by Intel® Xeon® processors. Netflix's performance team worked closely with Intel to scrutinize the software's interaction with available hardware resources and identify bottlenecks. Intel® VTune™ Profiler found code segments that did not use processor time optimally. Intel® PerfSpect provided additional insights by evaluating micro-architectural subsystems and programmed sequences. The tools ultimately helped pinpoint the instance bottleneck in a set of instructions within the Java Virtual Machine.

By using Amazon instances with underlying Intel Xeon processors rather than GPUs, Netflix found ways to save money by using each instance to accomplish multiple tasks. During peak viewing hours, Amazon instances can focus resources on streaming. When user demand lessens, instances can divert compute power to accelerate video encoding.

> "To ensure our customers have the best experiences with our streaming service, speed counts. Using Intel technologies to identify bottlenecks, we nearly tripled the performance of our Amazon EC2 instances while minimizing our cloud spend."
>
> *– Vadim Filanovsky, Performance Engineer at Netflix*



The significant latency decrease achieved after identifying and addressing the true sharing problem is dramatically illustrated in this graph.[1]

Results

With Intel's support to identify the instance bottleneck, Netflix realized a performance improvement per CPU by 3.5x compared with the initial throughput on Amazon EC2 instances.[1] They also benefitted from a significant reduction in average and tail latency.[2] Other companies utilizing Java workloads can also benefit from Netflix's CPU optimization approach since Intel addressed the source of latency in the open Java Development Kit.

By using oneDNN which taps the Intel AVX instruction set, Netflix claimed a substantial percent improvement in frame-per-second encodings, delivering excellent video quality on all devices.

Netflix's Amazon instances with Intel Xeon processors can also serve multiple purposes effectively using auto-scaling. Efficiencies gained through the CPUs allow Netflix to reduce the number of instances required for mission-critical workloads and significantly cut their overall cloud infrastructure spending.

intel. + aws

[1] To achieve 3.5x improvement over the initial results, there were two distinct steps: a) eliminating false sharing; and b) avoiding true sharing. The graph only represents the results of step (b). See charts representing the "before" and "after" false sharing fix here.

[2] https://netflixtechblog.com/seeing-through-hardware-counters-a-journey-to-threefold-performance-increase-2721924a2822

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.

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

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

No product or component can be absolutely secure.

RJMJ/SB/022024 ♻ Please Recycle