Case Study

Intel® Xeon® Processors

Netflix Optimizes Amazon Instance Performance, and Reduces Costs, Using Intel® Xeon® Processors and Intel® Analysis Tools

The latest Intel tools help identify bottlenecks down to the micro-architecture level.

Solution Summary

- Intel® Xeon® Processors
- Intel® PerfSpect
- Intel® VTune™ Profiler
- Amazon EC2 Instances

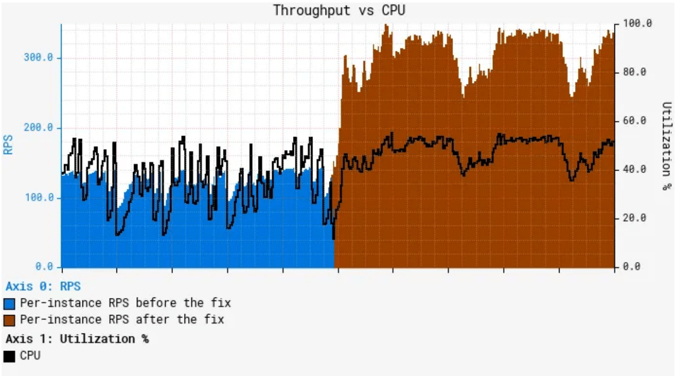
Executive Summary

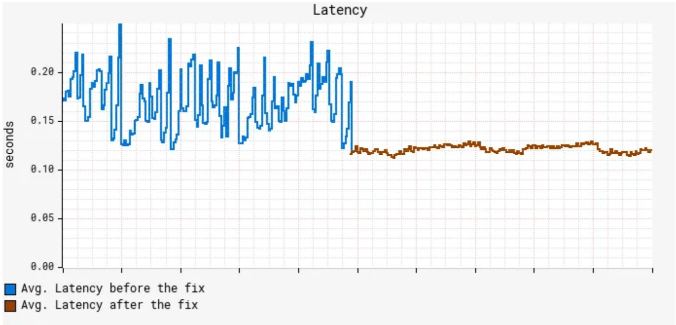
As a global entertainment streaming provider, Netflix is among the largest AWS customers. While evaluating its Amazon EC2 instances supported by Intel® Xeon® processors, the Netflix team realized their solution did not scale linearly, as expected. The team identified the source of lag by partnering with Intel and using analysis tools like Intel® PerfSpect, and Intel® VTune™ Profiler to evaluate the system at a micro-architecture level. After implementing a fix, Netflix increased its throughput per CPU by 3.5x, allowing the company to consolidate its Amazon EC2 instances and realize significant cost savings.

Challenge

Netflix regularly evaluates its Amazon EC2 instances to ensure it obtains the greatest performance for its workloads. Inspection by their engineering team revealed that supporting Amazon EC2 instances could not maintain linear scale, so they looked for the source of lag starting at the application layer. However, performance hindrances can also reside at the multi-core processor micro-architecture level. For example, unoptimized multi-threading, hierarchical cache subsystems, non-uniform memory, or out-of-order execution can also significantly impact performance. Netflix sought help from Intel to explore these other possible bottlenecks.

Using Intel technologies to identify bottlenecks, Netflix nearly tripled the performance of their Amazon EC2 instances while minimizing cloud spend. *Photo courtesy of Netflix.*





The graphs show the results of addressing the true sharing problem reached on the Amazon EC2 m5.12xl instance.[1]

Solution

Obtaining performance insights requires a top-down analysis of system elements, including interaction within a processor. The team executed a detailed evaluation of applications and the software's interaction with available hardware resources. Intel VTune Profiler, an analysis tool that evaluates multi-threaded and serial applications, helped find code segments that did not optimally use processor time. In addition, Intel PerfSpect explored the interaction among programmed sequences and microarchitectural subsystems to identify the source of lag. Ultimately, the team pinpointed the bottleneck in a set of instructions within the Java Virtual Machine.

> "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*

Results

After identifying the issue and fixing it, the throughput per CPU resulted in **a 3.5x improvement** over the throughput initially reached on the Amazon EC2 m5.12xl instance[1], along with a reduction in both average and tail latency.[2] The dramatic performance boost allowed Netflix to consolidate its Amazon EC2 instances for significant cost savings. Moving forward, Netfix will continue its periodic evaluation of instance performance and optimize it as needed with the aid of Intel hardware and tools.

Key Takeaways

- Maximizing cloud instance performance requires workload analysis from the application layer down to the CPU/GPU micro-architecture level.
- A systematic, top-down evaluation approach proves most effective.
- The findings from Netflix's optimization efforts extend beyond the company. Many companies depend on Java workloads, so if a problem is fixed in the Java Development Kit by Intel's team, many other enterprises benefit.

Where to Get More Information

Explore Intel Xeon processors.

Learn about Intel VTune Profiler.

Find out more about Intel PerfSpect on GitHub.

Learn more about Netflix's performance optimization process.

Read about best practices for Amazon EC2 instances.

[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.

092023/RJMJ/JS/PDF ♻ Please Recycle 357108-001US