CASE STUDY

Intel® Xeon® Platinum 8168 processor

Taboola Optimizes Artificial Intelligence for Smarter Content Recommendations

Hundreds of billions of recommendations to a billion users a month powered by an optimized combination of hardware from Intel and TensorFlow* software

"The use of the new CPU will allow us to serve more clients with less footprint, providing for better content recommendations faster. To do all that with the same investment in hardware is a great win for both Taboola and Intel."

Ariel Pisetzky, Taboola VP of IT.

In the world of online, always-on and on-demand media, being slow is the same as not being there at all.

Taboola's software-as-a-service (SaaS) solution delivers 360 billion content and article recommendations to over a billion unique users every month on thousands of publishers' sites. Each one is served in a tenth of a second by an artificial intelligence-based (AI) recommendation engine that analyzes contextual information about web visitors and their preferences. Getting this right drives the clicks, views and shares that are the foundation of modern publishing; getting it wrong, or not delivering at all, undermines this key source of revenue and disrupts user experience.

Taboola believes that answering more recommendation requests in a lower latency environment provides its customers, including some of the largest publishers in the world, with a considerable competitive advantage. Its search for an optimized hardware and software combination to help it service more requests per server led it straight to the long-standing technology partnership Intel.

Deep learning, deep personalization

Clicks and views contribute directly to the bottom line for online publishers so Taboola's global recommendation engine is a strategic solution for its customers.

Keeping web visitors engaged requires deeply personalized content recommendations, which depends on the analysis of contextual information that ranges from the simple – like time of day and recently-viewed content – to the complex, such as modelling inferences based on the language of specific content pieces.

With no opportunity to prepare in advance of web users' visits, Taboola's algorithms must work in real time to infer the right content recommendations within 100 milliseconds. This recommendation engine works in two parts, the first is machine learning using deep neural networks to infer user preferences, and the second is the processing and delivery of real-time content recommendations.

Instead of investing in extra hardware capacity to meet growing demand, Taboola wanted to understand how an optimized combination of hardware and its own recommendation engine, based on TensorFlow* open source machine intelligence software, could achieve the same results.

Hardware and software – the perfect team

Taboola decided to test its software on the newest Intel® Xeon® Platinum 8168 Processor in a Docker* container environment, while using Intel® VTune™ software to analyze system performance and flag improvement opportunities[1].

The Intel® Xeon® Platinum 8168 processors' higher CPU frequency and memory bandwidth, compared to Intel's previous generation processors the company was using, immediately boosted the software's performance. The highly parallel processor, running the Intel® Math Kernel Library (Intel® MKL) underneath the TensorFlow application, helped Taboola accelerate math processing routines, increase application performance, and reduce development time.

Intel VTune, an Intel software solution, also allowed Taboola to analyze its algorithm's choices and quickly find serial and parallel code bottlenecks, so that it could speed up execution by better understanding where and how the application could benefit from available hardware resources.

Testing required several iterations, so working in a Docker container environment made the process streamlined and efficient by removing the need to configure and reconfigure servers for every test.

A great win for Taboola and Intel

Within a matter of weeks, Taboola achieved an incremental gain up to 149 percent in throughput of its key deep learning model, by taking advantage of Intel® architecture and Intel® MKL and Intel® Advanced Vector Extensions 2 (Intel® AVX2) instructions at the processor level[1].

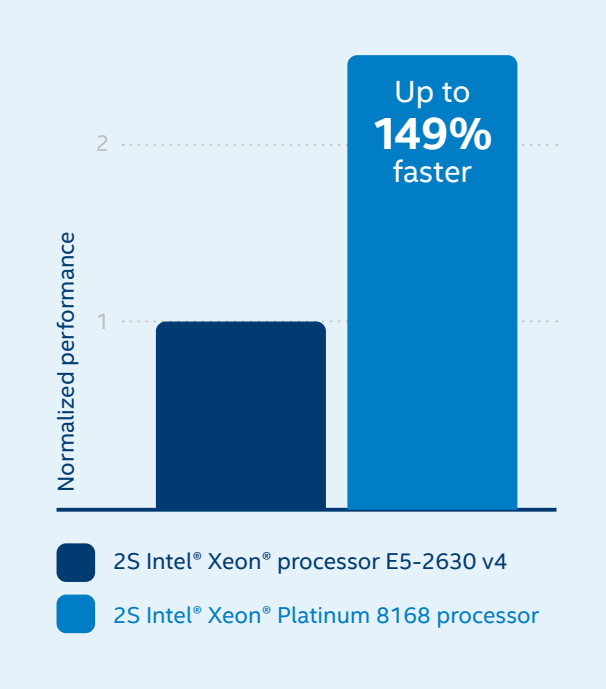


Figure 1 Taboola's TensorFlow* inference engine increased performance with the Intel® Xeon® Platinum 8168 processor

Taboola's vice president of IT, Ariel Pisetzky, said: "The use of the new CPU will allow us to serve more clients with less footprint, providing for better content recommendations faster. To do all that with the same investment in hardware is a great win for both Taboola and Intel."

Intel VTune provides system wide profiling of applications and visualizes the time each function is taking in easy-to-read Microsoft Excel* spreadsheets. By first identifying the exact contribution of each software module to the overall run-time of the application, the solution then flags the precise lines of source code within those modules that could be optimized.

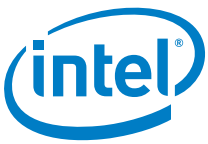
Before using Intel VTune, Taboola only had insights into performance at the block level within TensorFlow, which only showed serious mistakes and bad configurations. This meant it could only see how much time each block takes under one prediction but couldn't get under the hood to see what actual code was being called.

With Intel VTune, Taboola could profile and understand the application in a deeper way than ever before. This helped to reveal tuning opportunities in both how TensorFlow performed with the Intel MKL and how its AI models that run on top of TensorFlow interact with each other. The company says that in the past it would not have had the visibility to be able to optimize its recommendation engine at this level. Instead, it would have needed to wait for further developments in TensorFlow's capabilities or to invest in further hardware.

Finally, Taboola was one of the first companies in the world to use Intel VTune in a Docker container environment, without which the company believes it would have been many times more difficult to achieve the improvement gains it did in such a short timeframe.

A meeting of minds

The close collaboration between Intel and Taboola was critical to the project's success, and the expertise exchanged between the companies was just as important as the technology used. With Intel's support, Taboola's benefited from gaining a deeper understanding of the impact of hardware on application performance, and the two are already exploring how the new Intel® Xeon® Scalable Processor family can contribute to continue uplifts in performance.



[1] Testing conducted on ISV* software comparing Intel® Xeon® Platinum 8168 processor to 4S Intel® Xeon® Processor E7-8890 v4 Testing done by Intel. Taboola code written in TenserFlow: OS: CentOS 7 kernel 3.10. Testing by Intel and Taboola June 2017. BASELINE: 2S Intel® Xeon® processor E5-2630 v4, 2.2GHz, 20 cores, turbo on and HT off, 128GB total memory, 8 slots / 16GB / 2133 MT/s / DDR4, 400Gb, Dell SSD DC S3610, CentOS 7 kernel 3.10.0-514. NEW: Intel® Xeon® Platinum 8168 processor, 2.7GHz, 24 cores, turbo on, HT off, 192GB total memory, 12 slots / 16GB / 2666 MT/s / DDR4, Intel SSD, CentOS 7 kernel 3.10.0-514

Software and workloads used in performance tests may have been optimized for performance only on Intel microprocessors. Performance tests, such as SYSmark and MobileMark, are measured using specific computer systems, components, software, operations and functions. Any change to any of those factors may cause the results to vary. You should consult other information and performance tests to assist you in fully evaluating your contemplated purchases, including the performance of that product when combined with other products.
For more information go to www.intel.com/benchmarks.

Intel does not control or audit third-party benchmark data or the web sites referenced in this document. You should visit the referenced web site and confirm whether referenced data are accurate. Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Performance varies depending on system configuration. No computer system can be absolutely secure. Cost optimization scenarios described are intended as examples of how a given Intel- based product, in the specified circumstances and configurations, may affect future costs and provide cost savings. Circumstances will vary. Intel does not guarantee any costs or cost reduction. Check with your system manufacturer or retailer or learn more at intel.com.

Optimization Notice: Intel's compilers may or may not optimize to the same degree for non-Intel microprocessors for optimizations that are not unique to Intel microprocessors. These optimizations include SSE2, SSE3, and SSSE3 instruction sets and other optimizations. Intel does not guarantee the availability, functionality, or effectiveness of any optimization on microprocessors not manufactured by Intel. Microprocessor-dependent optimizations in this product are intended for use with Intel microprocessors. Certain optimizations not specific to Intel microarchitecture are reserved for Intel microprocessors. Please refer to the applicable product User and Reference Guides for more information regarding the specific instruction sets covered by this notice.

Intel, the Intel logo, VTune and Xeon are trademarks of Intel Corporation in the U.S. and/or other countries.

*Other names and brands may be claimed as the property of others.

© 2017 Intel Corporation.

0919/CAT/MS/PDF

♻ Please Recycle

336532-001EN