White Paper

Internet Discovery
Artificial Intelligence

Taboola Partners with Intel Technologies to Optimize and Future-Proof Their Solution Data Centers with Innovative IT Infrastructure

To gain the steady increases in power needed to deliver thousands of content recommendations per second, Taboola leverages Intel Technologies to optimize and future-proof their solution infrastructure, with results that rival and surpass cloud computing.[1]

About Taboola

Taboola powers recommendations for the open web, helping people discover things they may like. The company's platform, powered by artificial intelligence, is used by digital properties, including websites, devices, and mobile apps, to drive monetization and user engagement. Taboola has long-term partnerships with some of the top digital properties in the world, including CNBC, BBC, NBC News, Business Insider, The Independent, and El Mundo.[2] More than 15,000 advertisers use Taboola to reach nearly 600 million daily active users in a brand-safe environment.[1] Following the acquisition of Connexity in 2021, Taboola is a leader in powering e-commerce recommendations, driving more than 1 million monthly transactions.[1] Leading brands including Walmart, Macy's, Wayfair, Skechers, and eBay are among key customers.[2]

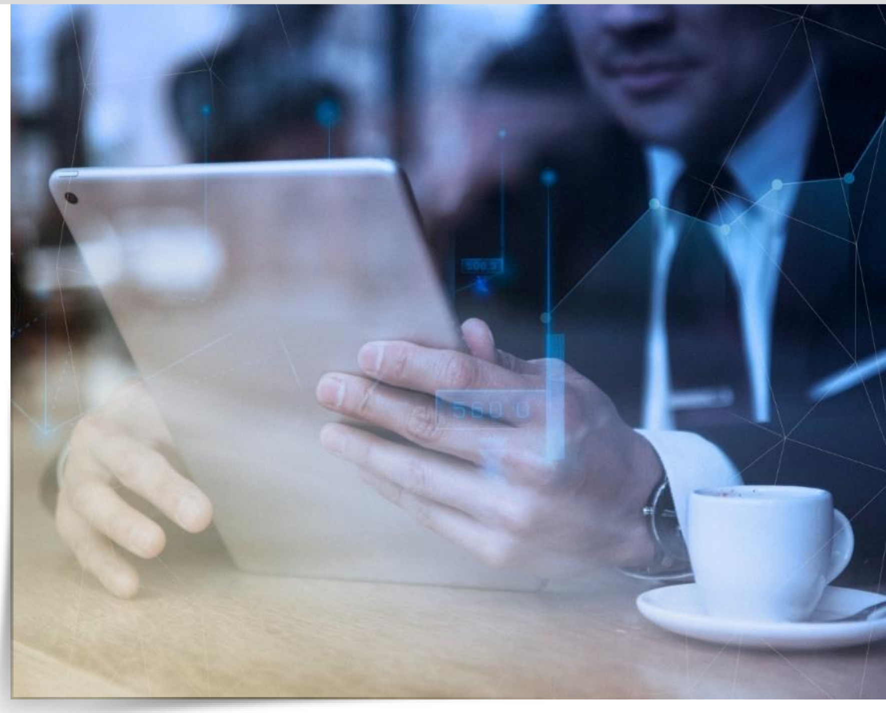
Executive Summary

For AI solution providers, the choice between cloud computing, data centers, or a hybrid approach is pivotal to optimizing their IT infrastructure and services. Although cloud services are highly popular for enabling organizations to access and scale computing resources on-demand without significant upfront infrastructure investment, many solution providers are surprised to learn that data centers may be the more cost-effective and rewarding option in the long run depending on a number of factors.

Specifically, processing and storing mass amounts of data, which is typically generated by AI, computer vision, and machine leanings solutions, in the cloud can become very expensive for solution providers. As computing requirements inevitably rise due to increasing user bases, data generation, and more, solution providers can become inundated with exorbitant cloud infrastructure costs that will only continue to rise, as they don't have direct ownership of their infrastructure.

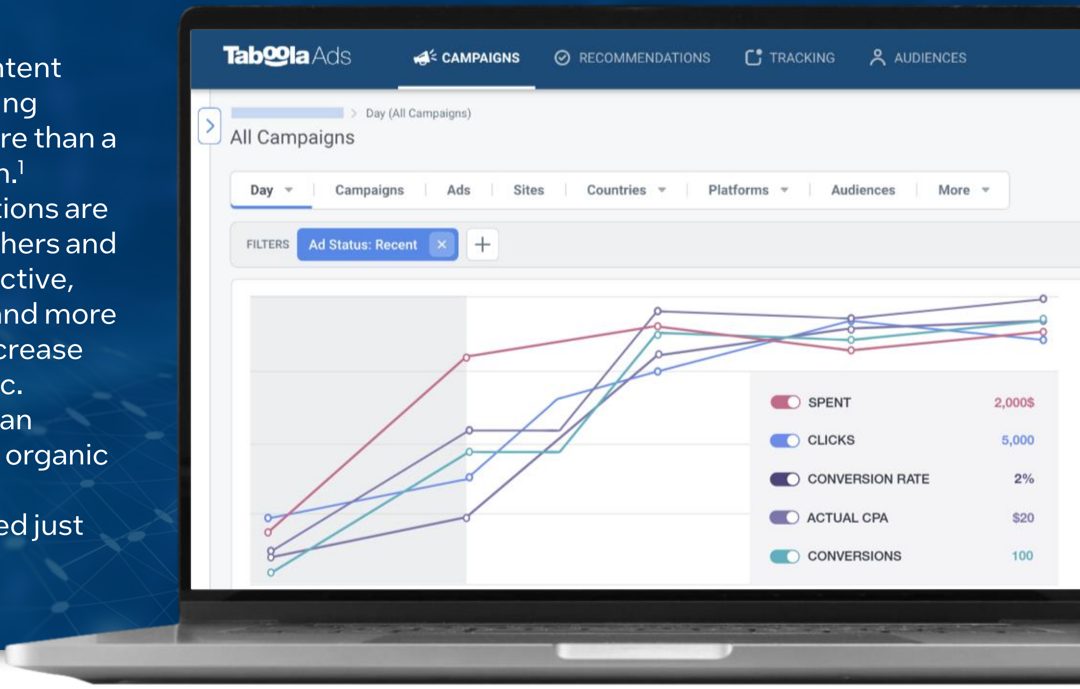
For companies that offer technology solutions with high computing requirements and stabilized workloads, data centers offer full ownership and control of their infrastructure. Data centers enable organizations to do extremely high levels of computing without paying a cloud service provider to host the infrastructure required for their solutions. Despite the upfront expenditure required to build a data center, having direct ownership over one's solution infrastructure enables easier and more cost-effective in-house customization, optimization, and workload fluxing than hiring a third-party cloud service provider or hyper scaler. Even by partially transitioning to data center computing, solution providers can mitigate much of the costs of cloud computing and open their product up to the latest generations of Intel technology.

By leveraging Intel-supported data centers, companies can develop infrastructure with an optimal balance between price, power, and performance. They also gain the ability to continuously optimize and evolve their solution by building on top of existing infrastructure, future-proofing the solution's IT. Taboola, the largest content recommendation platform and readership database,[1] built their data centers with Intel components to gain these benefits while enabling scalable, real-time inferencing for their large computational requirements. The following analysis explores how Taboola leverages Intel technology to maximize the utility of their data centers to achieve exceptional scalable performance while circumventing the need for costly third-party cloud computing.



Taboola: Driving Meaningful Customer Engagement with AI-Driven Content Recommendations

Taboola is the world's largest content recommendation engine, delivering tailored recommendations to more than a billion unique people every month.[1] Taboola's content recommendations are leveraged by thousands of publishers and advertisers, including CBS Interactive, Euronews, Pandora, Samsung,[2] and more to reach their target audience, increase engagement, and monetize traffic. Meanwhile, readers benefit from an enhanced online experience with organic and sponsored content recommendations that are curated just for them.



To enable these content recommendations, Taboola's solution crawls all of a publisher's web pages to extract the meaningful context of each page. When a user organically reaches a page, Taboola performs AI inferencing on two levels. First, the solution leverages classic machine learning to update its readership dataset with new user data, such as time of day, recently viewed content, trending topics, and context to predict appropriate content on an individual basis. Then, Taboola's AI algorithms, trained on vast amounts of data, perform real-time inferencing that narrows down billions of content recommendations to the most optimized, personalized combination within a specific page for each user.

Taboola's solution not only enables publishers to drive business results by reaching people genuinely, effectively, and at the right time, but also provides exceptional, personalized user experiences that keep people coming back to these sites time and time again.

Optimizing the World's Largest Content Recommendation Engine for AI

As Taboola expands its online footprint with new users, publishers, and advertisers and evolves the accuracy of its recommendation engine based on new ingested data, it needs steady increases in power and capacity to provide the speed and accuracy its customers expect. To accomplish this, Taboola must maximize the performance and utility of its current infrastructure while staying up to date with rapid advancements in technology. In pursuit of these objectives, Taboola needs to meet three long-term goals for its data centers.

1 Solving the Cost of Driving Billions of Content Recommendations

Taboola's solution delivers thousands of content recommendations a second.[1] As Taboola continues to acquire new publishers and their readers, its AI algorithms must process an increasing volume of written content and user data to provide more accurate content recommendations, all in real time. From an IT perspective, controlling the cost of producing these recommendations is a top priority. These computing requirements would result in exorbitant and escalating costs if done by a third-party cloud provider. To avoid that, Taboola must maintain an efficient IT infrastructure in their data centers to enable the most powerful inferencing at the lowest price. This approach ensures the provision of top-notch service without imposing high costs on its customers.

2 Maintaining a Responsible Ecological Footprint

Within the broader climate, Taboola strives to maintain a responsible ecological footprint while delivering optimal performance. Since processing units are producing heat in the physical world as they produce recommendations in the virtual world, Taboola prioritizes hardware selection to maximize the number of recommendations per second, per CPU, and per unit of electricity or power consumed. This entails procuring CPUs that are best suited for Taboola's inferencing tasks with the most relevant power pool. Taboola also considers broader aspects such as server cooling strategies and server cycle management during less busy periods to further enhance energy efficiency. By implementing these measures, Taboola aims to uphold its commitment to a responsible ecological footprint while delivering exceptional performance.

3 Minimizing the Space Footprint of High-Performance Data Centers

Another objective Taboola aims to achieve with its data centers is to minimize the physical footprint of its infrastructure by maximizing density. By strategically increasing the server and CPU count per space unit, Taboola aims to optimize space utilization. This is accomplished through the strategic selection of hardware solutions that provide high density, enabling more servers to be housed within racks without excessive empty space. By addressing the challenge of space utilization, Taboola ensures efficient use of its physical infrastructure while maintaining optimal performance.

Taboola Leverages the Capabilities of Intel® technology to Maximize Their Data Centers

Based on these objectives, Taboola's data centers must be strategically managed to offer the innovative, scalable, and cost-effective performance of their solution. To build high-performance data centers that position Taboola as a leading content recommendation engine and readership database, Taboola partners with technology providers like Intel. This partnership allows Taboola's IT team to optimize the performance of their current infrastructure and software and build upon it to innovate more advanced features with new generations of hardware. By strategically managing their data centers in partnership with Intel® Technologies, Taboola ensures the continued success and competitiveness of their solution.

To deliver its content recommendation service globally, Taboola runs six front-end data centers and four back-end data centers. Through its longstanding partnership with Intel, Taboola has been able to not only optimize their current infrastructure, but also continuously evolve these data centers to accommodate growing compute workloads while maintaining and improving Taboola's capabilities. Specifically, Taboola leveraged Intel to gain the following hardware and process improvements:

Sustainable and Buildable Infrastructure

Taboola runs their content recommendations on Intel® Xeon™ Processors, an advanced line of Central Processing Units (CPUs) that are designed to offer high-performance computing in Taboola's data centers. This line of hardware is optimal for Taboola's solution because it supports inferencing for each unique user, as opposed to batch inferencing, to provide the most accurate content recommendations.

Taboola prefers Intel® Xeon™ Processors because they are highly versatile and can easily be repurposed and built on top of one another. Taboola has approximately 12,000 servers, a significant capital investment that Taboola wants to continue leveraging far into the future. For this infrastructure to be financially sustainable, Taboola must extend the life cycle of existing servers while leveraging the latest capabilities offered by new generations of Intel® Xeon™ processors.

Intel® Xeon™ processors are not solution specific and continue to stay relevant as new generations of hardware are released. When upgrading servers to accommodate increased computing load, Taboola's IT department can distribute the workload across different generations of processors, future-proofing the hardware. For example, they can utilize the new servers with updated instruction sets for inferencing, while repurposing older generations of servers for various tasks such as database work, load balancing, networking, and other essential computing operations.

The older CPUs might have faster clock speeds, while newer CPUs have updated instruction sets and so on. Taboola mixes and matches as needed to find the optimal purpose for each generation of Intel® Xeon™ processors, extending the lifecycle of their infrastructure and resulting in significant benefits for Taboola and their customers who benefit from those cost savings.

Maximized Infrastructure with Minimal Space

Considering Taboola's six front-end data centers and four back-end data centers, achieving higher density in their data centers is a key goal to maximize their current infrastructure and minimize their space footprint in the future. This approach brings significant benefits for both Taboola and their customers:

Cost Optimization

By constructing high-density data centers, Taboola can reduce costs by minimizing the need for additional space, power, and cooling infrastructure. This translates into substantial cost-savings for their customers.

Scalability and Future-Proofing

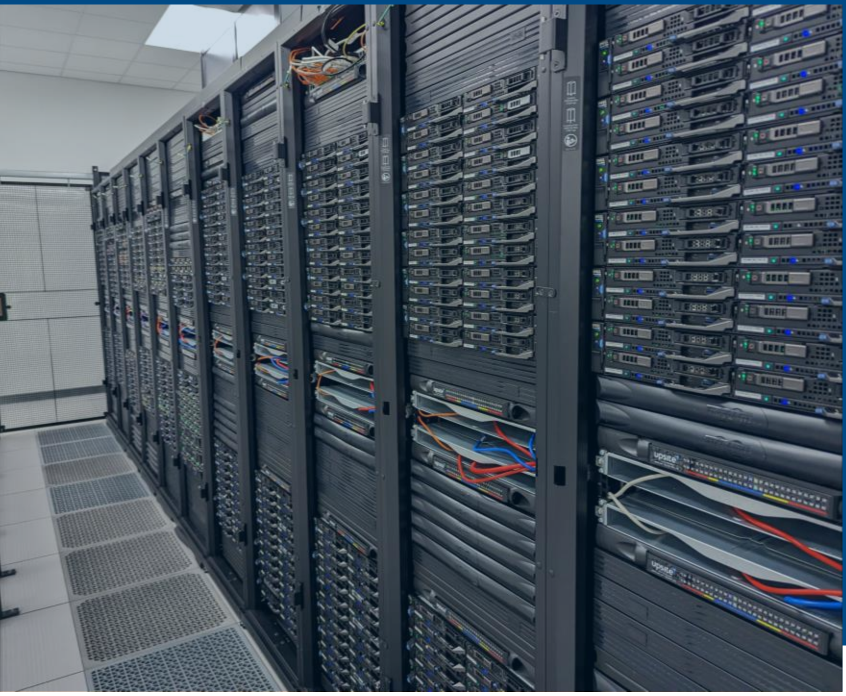
As Taboola's data demands continue to grow, driven by the increasing number of publishers and readers, strategically planning dense data centers accommodates additional servers and equipment within their existing infrastructure. This enables easier scalability as their computing requirements expand.

Energy Efficiency

Building higher density deployments allows Taboola to create further energy efficiency improvements, such as better results with power and cooling systems that dispose of heat more efficiently and require less electricity.

Performance

Denser data centers enable Taboola's IT department to reduce network latency and improve data processing speed, resulting in faster compute times for their services.



Taboola leverages Intel® Technologies to achieve the best infrastructure design that minimizes space utilization without compromising operational performance. Leveraging the small form factor of Intel® processors, Taboola can fit two CPUs in each server and two servers in each level of the rack, enabling approximately 48 CPUS in just the top of one cabinet with a capacity for more servers. This level of density is rare and unique to Taboola, as larger CPUs typically limit density to only two servers within a larger 2U compute rack unit. This means that while others may have only one server in the same space, Taboola can accommodate four servers due to their use of smaller form factor Intel® CPUs.

Optimization of Prediction Throughput

As an AI-based content prediction engine, the Taboola solution runs on a neural network based on the open-source TensorFlow framework that uses deep learning to infer visitor preferences. This AI infrastructure is fundamental to meeting Taboola's speed and accuracy requirements while analyzing a variety of data for each website visitor. The self-learning power of AI also ensures that Taboola's AI algorithms continue to learn from new data sources and from the way individual consumers respond to the recommendations, enhancing recommendation accuracy without the need for complex programming.

Specifically, Taboola uses the TensorFlow Serving (TFS) framework, which is architected on top of TensorFlow and employs a client-server workflow to deliver recommendations. Each TFS server hosts a pretrained model of the Taboola neural network. When TFS receives a prediction request from a client, it runs the client data in a forward pass through the model and returns the result. Throughput is critical to scalability to match individuals with brand and editorial content that's interesting and relevant to them.

To offer the same speed and accuracy across a fast-growing database of website visitors and published content, Taboola needs to ensure it can gradually increase TFS throughput while maximizing the utility of its current infrastructure. For instance, to achieve higher performance without expanding their infrastructure footprint, Taboola engaged with Intel software engineers to optimize their TFS code.

The following three enhancements were completed in just a few weeks, resulting in a 2.5x improvement in performance over the original, unoptimized code.[3]

1 Integration of Intel MKL-DNN

TFS commonly relies on the open-source Eigen* C++ template library to perform Taboola's inferencing. Although TFS itself has been highly optimized for Intel® architecture, Eigen has not. Intel® Match Kernel Library for Deep Neural Networks (Intel® MKL-DNN) provides primitives for neural network processing that are all highly optimized for performance on the latest Intel® microarchitecture. The integration of the Intel ® MKL-DNN library delivered 1.15x the performance of the unoptimized version of TFS.[3]

2 Two instances of TFS with CPU and NUMA pinning

Intel engineers have found that hosting two instances of TFS per server and allocating the processor and memory resources efficiently to each instance improves performance.

3 Vectorized tensor broadcast operations

To take performance to the next level, Intel engineers used Intel® VTune™ Amplifier to identify performance bottlenecks by profiling Taboola's application during runtime. With the Intel® VTune™ Amplifier, engineers can visualize the contribution of each software module to the overall runtime of the application. They can also look more closely to identify the precise lines of source code within those modules that are impairing performance and are good candidates for optimization.

Using this tool, Taboola learned that, unsurprisingly, the most time-consuming operation was a tensor operation known as broadcasting. As revealed by the Intel® VTune™ Amplifier analysis, the Eigen implementation of tensor broadcasting relies heavily on scalar instructions that do not take advantage of vector processing capabilities available in Intel® Xeon™ Scalable Processors. The software optimization team vectorized the tensor broadcast functions in Eigen using Intel AVX-512 instructions. As part of the optimization effort, the engineering team created two new tensor broadcasting member functions based on the types of input tensors identified in Taboola's application. For both types of tensors, the engineering team was able to significantly reduce the number of operations in a typical broadcast operation, accelerating operations.

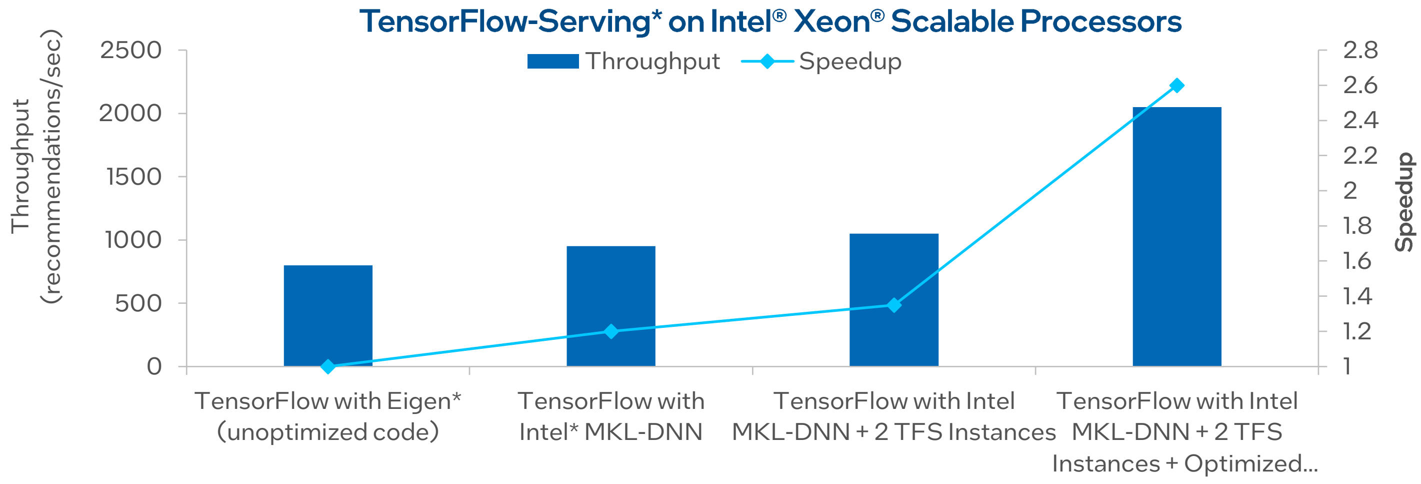


Figure 1. TFS Performance Gains: Performance comparisons for the optimized versions of TFS versus the unoptimized baseline version.

Taboola used these performance gains to deliver more and better recommendations at higher speeds. With upwards of ten thousand servers across multiple data centers, the benefits in cost savings, efficiency, and growth potential were substantial.

Continuous Updates that Integrate Seamlessly with Current Infrastructure

Given the rapid growth Taboola's success and consequently its workload demands, more performance was needed over time. This is where Taboola's IT team found themselves when they wanted to serve more content recommendations while maintaining client latency below 100ms. In this case, Intel and Taboola leveraged higher core counts and memory bandwidth on the 4th Generation Intel® Xeon™ Processors and applied the latest TFS with Intel optimizations to increase Taboola's prediction throughput by 1.74x.[4] This significant increase in throughput enabled Taboola's solution to perform more content recommendations within the same given time frame on existing infrastructure, helping publishers and advertisers benefit from the same lower costs and from the best return on their ad spend that they are accustomed to.

Conclusion

Taboola has rapidly emerged as one of the world's most trusted and comprehensive content recommendation platforms by matching individuals with online brand and editorial content that's thoughtfully curated for their interests. Speed and accuracy are crucial to Taboola's content discovery services, and Taboola relies on Intel-powered data centers deliver the scalable infrastructure and performance that grows with the solution.

Over time, Taboola's partnership with Intel has enabled them to continuously optimize their solution with new capabilities offered with new generations of processors, while future-proofing their existing infrastructure by repurposing it for other tasks that contribute to optimizing Taboola's overall workload. With Intel's support, Taboola has the toolset needed to not only optimize their solution's application performance today, but to ensure it remains highly competitive and relevant far into the future.

Learn More

- Taboola Website
- Taboola Demo Video
- Intel® Xeon® Scalable Processors Product Page
- Intel® Optimization for TensorFlow

> “Taboola delivers tailored recommendations to more than a billion unique Internet users every month to help them explore what’s interesting and new. Optimizing on latest-gen Intel platforms will help us in multiple ways. This includes reducing response time and improving model accuracy. This will also help reduce operational costs as the server efficiency goes up.”
>
> **— Ariel Pisetzky, VP IT & Cyber, Taboola**

Sources

1. Internal Taboola Estimate, Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy.
2. For these and other Taboola case studies; visit the Taboola website at Taboola Case Studies
3. Performance results are based on Taboola and Intel testing as of 6 August, 2018 and may not reflect all publicly available security updates. System Configuration: Two-socket server configured with 2 x Intel® Xeon® Platinum processor 8180 (2.50 GHz, 28 cores), 192 GP DDR4@2666MHz memory (12 x 16 GP DIMMS), 1.5 TB Intel® SSD (SC2BX01), CentOS Linux* release 7.5.1804 (Core) (3.10.0-862.9.1.el7.x86_64); Baseline software application: TensorFlow-Serving r1.9 (https://github.com/tensorflow/service); Intel Optimized software application: TensorFlow-Serving r1.9 + Intel MKL-DNN (https://mirror.bazel.build/github.com/intel/mkl-dnn/archive/0c1cf54b63732e5a723c5670f66f6dfb19b64d20.tar.gz) +optimizations (availability of optimizations expected in TensorFlow-Serving release 1.10)
4. Configuration: BASELINE (8380): Tested by Intel as of Nov 2022. 2 socket Intel® Xeon® Platinum 8380 CPU @ 2.30GHz Processor(ICX), 40 cores/socket, HT On, Turbo ON, Total Memory 256 GP (16 slots/16GB/3200MT/s DDR4), BIOS: SE5C620.86B.01.01.0006.2207150335, ucode 0xd000375, Ubuntu 20.04.2 LTS, 5.4-113-generic, gcc 9.4.0 compiler, Inference Framework: TensorFlow, Topology: TensorFlow Serving, 1 instance/2 socket, Multiple streams, Datatype: FP32.

   OPTIMIZED (8480): Tested by Intel as of 2022. 2 socket Intel® Xeon® Platinum 8480 CPU @ 2.0GHz Processor(ICX), 56 cores/socket, HT On, Turbo ON, Total Memory 256 GP (16 slots/16GB/4800 MT/s DDR5), BIOS: SE5C7411.86B.8713.D03.2209091345 (09/09/2022), ucode 0x2b000070, Ubuntu 22.04.1 LTS, 5.15.0-52-generic, gcc 11.3.0 compiler, Inference Framework: TensorFlow, Topology: TensorFlow Serving, 1 instance/2 socket, Multiple streams, Datatype: FP32.

Notices & Disclaimers

Intel is committed to respecting human rights and avoiding complicity in human rights abuses. See Intel’s Global Human Rights Principles. Intel® products and software are intended only to be used in applications that do not cause or contribute to a violation of an internationally recognized human right.

Intel technologies may require enabled hardware, software or service activation. No product or component can be absolutely secure. Your costs and results may vary. Intel does not control or audit third-party data. You should consult other sources to evaluate accuracy. Code names are used by Intel to identify products, technologies, or services that are in development and not publicly available. These are not "commercial" names and not intended to function as trademarks.

You may not use or facilitate the use of this document in connection with any infringement or other legal analysis concerning Intel products described herein. You agree to grant Intel a non-exclusive, royalty-free license to any patent claim thereafter drafted which includes subject matter disclosed herein.