White Paper

Content Delivery Networks

Bridging the CDN Capacity Gap with Near 400 Gbps Video Delivery

The combination of Varnish software and Intel technology provides the throughput needed to keep up with the growing popularity and evolution of video streaming.

Content delivery networks (CDNs) need greater input/output (I/O) and compute performance to keep up with growing demand for streamed content, new types of streaming, and higher resolution content. Video on demand (VoD) services have been the primary driver of the growth in streamed content and resolution quality. The popularity of established VoD leaders combined with newer, niche streaming services, is resulting in more consumers making a daily habit of streaming VoD services. These services also provide more high-definition (1080p) and ultra-high definition (4k) programming that increases the demand for performance in a CDN.

Added to these VoD trends is the explosion of live streaming consumption, estimated by one industry analyst group to total nearly 4 billion hours[1] watched throughout 2020. Live streaming traffic includes gaming platforms, new live video social media platforms, and services that are broadcasting large-scale in-person events, such as sporting events, plays and concerts. Live streaming of smaller scale events—from religious services to business meetings—takes advantage of live streaming technologies ranging from simple video conference calls to more highly produced events using social media platforms.

All these use cases drive streaming video providers to develop faster, yet cost-effective CDN solutions to handle more users at the same resolution, or to serve higher-resolution content to a similar number of users. The growing availability of 100 GbE and 400 GbE network cards helps to consolidate users and traffic on a server. But the server CPU is still a critical element needed for higher CDN performance.

Varnish is a content caching/CDN solution that is known for its performance. The company is an Intel® Network Builders ecosystem partner and worked with Intel to test the performance of its Varnish Enterprise CDN solution. The tests used 3rd generation Intel® Xeon® Scalable processor-based servers that conform to the Intel® Select Solutions for Visual Cloud Delivery Network reference. For the tests, these servers were further optimized for VoD performance.

Varnish Delivers Fast Caching Solution

Varnish Enterprise is a powerful, feature-rich web cache and HTTP(S) accelerator that is used by a wide range of content and service providers to solve challenges related to video streaming, CDN, and website acceleration. Varnish is built on top of open source CDN frameworks with enterprise resiliency features and is designed with robust features for high performance and scalability. It allows companies to deliver content with low latency even during periods of peak demand.

Figure 1 shows a sample Varnish deployment. The origin servers host the streaming video catalog, pushing the most popular content to the CDN cache nodes for distribution to users. Depending upon the number of users and geographical dispersion of the network, the CDN can use a single tier of caching servers or, as shown in Figure 1, two or more cache tiers, geographically distributed to improve performance for users around the world.

Table of Contents

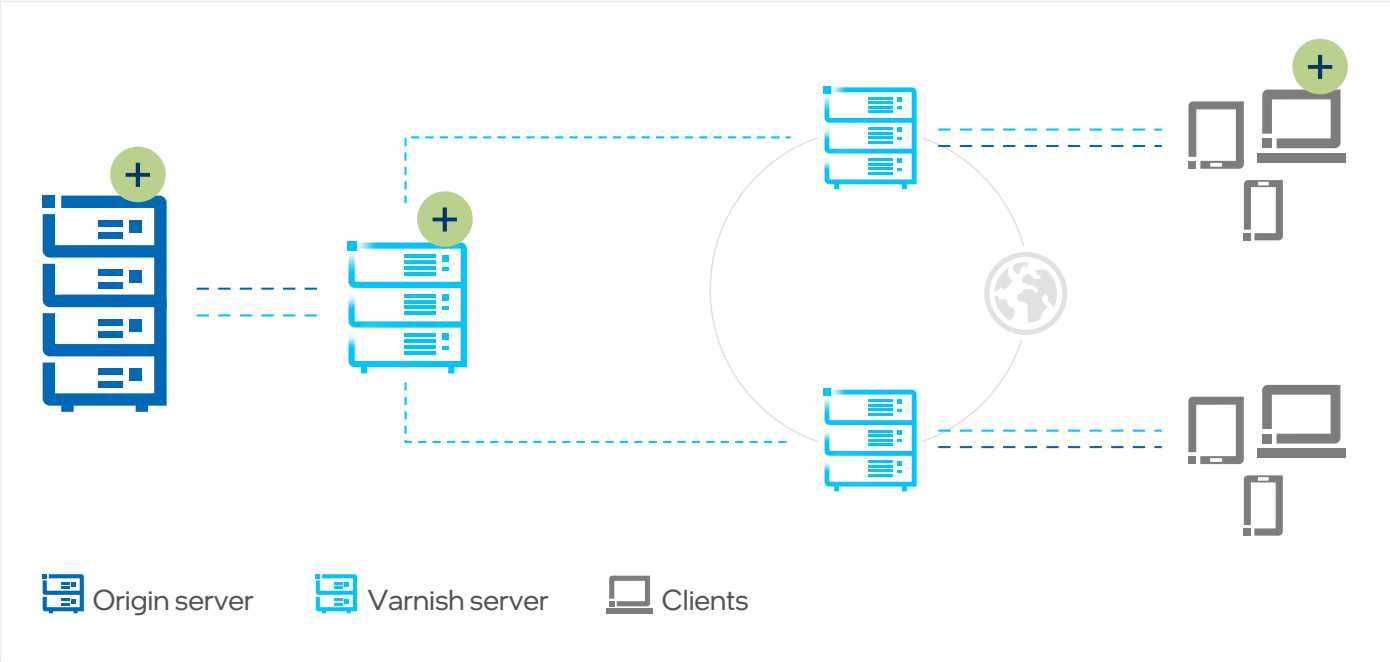

Figure 1. Two-tier Varnish solution architecture for a large-scale CDN.

Varnish Enterprise is available across bare metal, virtualized, containerized and cloud environments, packaged into separate software products that are optimized for specific content delivery challenges. These include web and API acceleration, streaming, and private CDN deployments, including in vRAN and NFVI environments.

The testing described in this paper used the Varnish Edge Cloud CDN solution, which is optimized for delivering high quality, low latency HTTP video in large-scale communications networks that face challenging workload demands.

Testing CDN Performance on Single and Dual Processor Servers

Varnish and Intel teamed up to test CDN performance on a single processor system under test (SUT) and a dual-processor SUT, both of which used 3rd generation Intel Xeon Scalable processors.

While Varnish can cache a wide range of content types, for this test it was optimized for streaming video. The servers were configured to show performance of the single-processor server and demonstrate performance scalability using a dual-processor server. Dual-processor systems utilize non-uniform memory access (NUMA) architectures to enable each processor to access its local memory and I/O resources as well as those on the other CPU. The linear performance scaling is due, in part, to NUMA awareness in Varnish, which allows the software to make better decisions about what resources in the system to use for a particular transaction, improving the locality of accesses to memory as well as I/O devices. Varnish has added NUMA awareness into the latest Varnish software release.

Bare metal servers were used for the SUTs. The tests utilized a warm cache, where CDN servers are connected to the origin server to load the popular content into the cache server. The video catalog used for this test included a 4.921 TB data set. In a real-world application the most popular content available on a cache node is stored in DRAM for fastest access, with the remainder of cached content being stored on NVMe. But for these tests, requests were made randomly across the full span of the dataset, resulting in content that is all of equal popularity, forcing increased reads from NVMe to show the performance under a more difficult scenario than is likely in real-world conditions.

In addition to the single- and dual-processor SUTs, client servers designed to emulate user devices were also used in the tests. A typical real-world CDN client is a single user device (TV, computer, mobile device) that consumes a single content stream and generates a data load on the network of between 1 Mbps to 20 Mbps, depending upon exact encoding and content resolution. The clients used in this test, however, were servers that emulated the content requests of thousands of clients and consistently generated up to 100 Gbps of data traffic per client. The tests were configured for the optimal use case of cached content at a 100 percent hit rate, meaning no content requests are made from the origin server. In all test cases, the clients and cache nodes were using HTTPS.

These client servers were connected to the CDN servers using 100 GbE links through a switch; 2x100 GbE connections for the single-processor SUT, and 4x100 GbE for the dual-processor SUT. Testing was done using wrk, a widely recognized open-source HTTP(S) benchmarking tool.

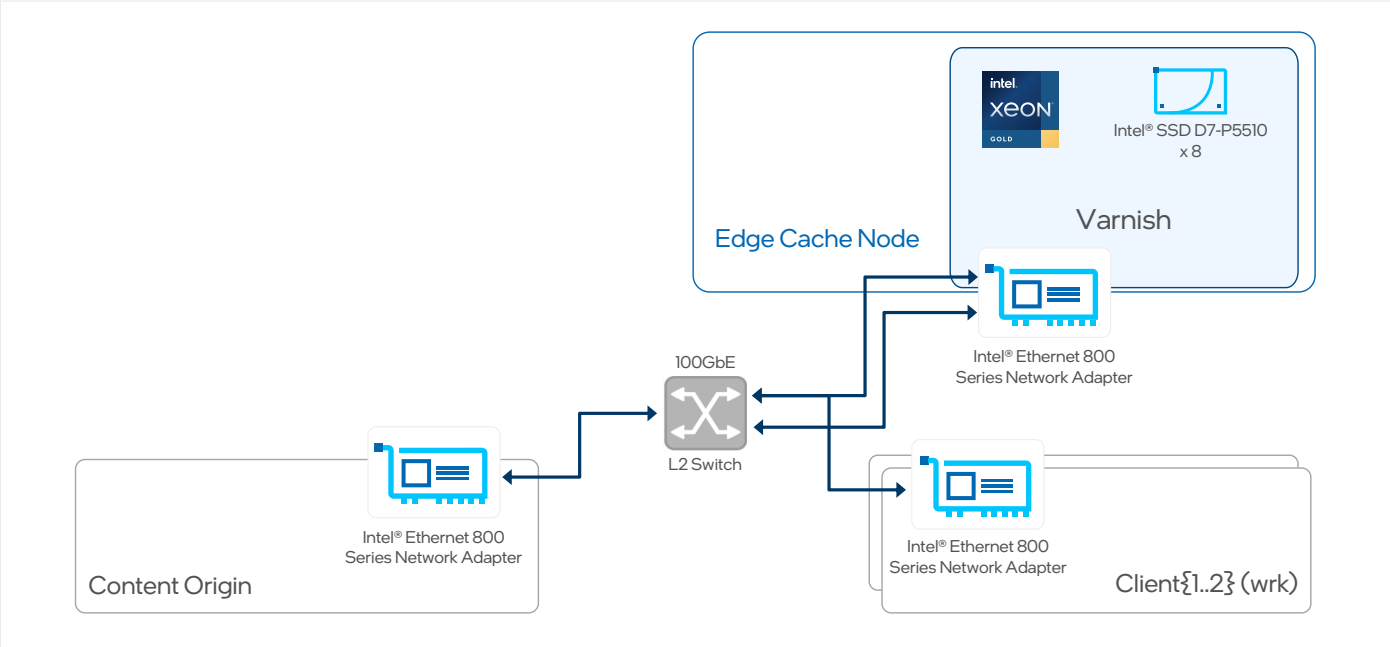

Figure 2. Single processor CDN test configuration.

The single processor test setup can be seen in Figure 2. The origin server is shown in the diagram for completeness but was not accessed during benchmarking.

The dual-processor version of the test setup is similar but with two dual-port connections for a total of 400 Gbps into the dual processor server as shown in Figure 3.

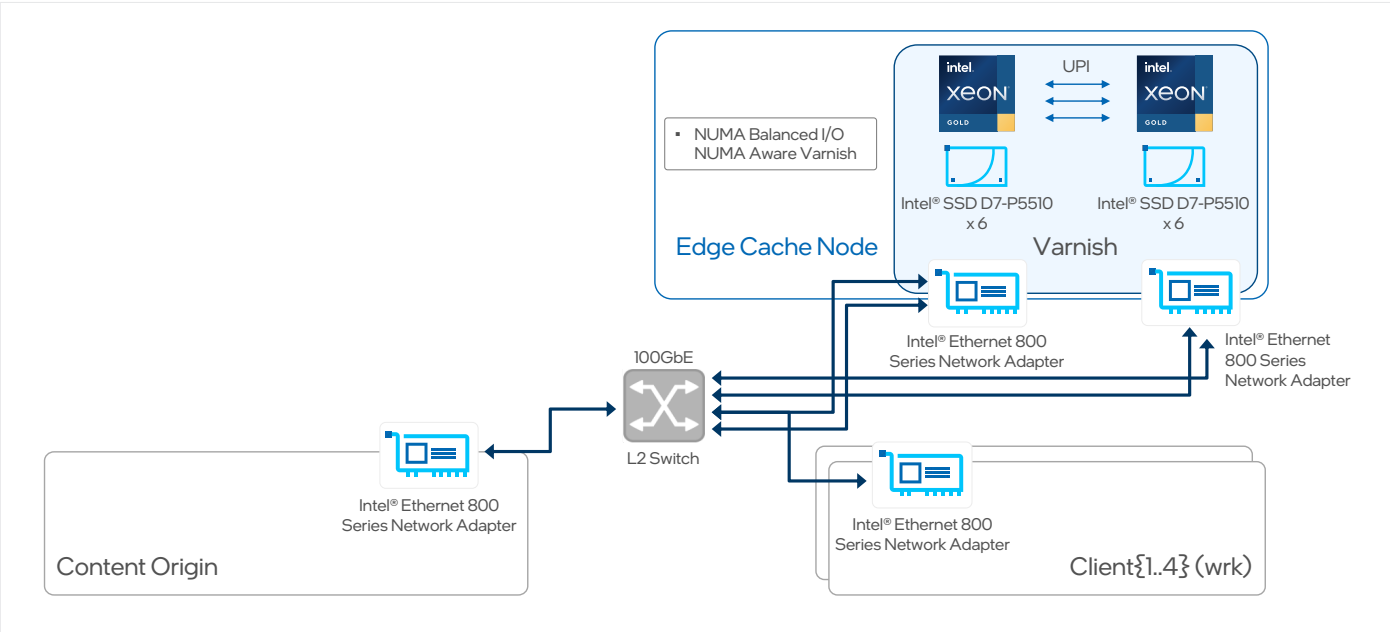

Figure 3. Dual-processor server CDN test setup.

The clients requested a series of 1 MB objects from the cache, equivalent to between one and five seconds of video, depending upon the exact encoding and resolution. In this testing, the focus was the performance of the cache node, so these clients requested chunks of data as fast as possible, instead of at the regular pace that a video player would use to maintain a buffer. Additionally, in this test, the clients were requesting video at random, and all content was of equivalent popularity. This content popularity distribution covered near worst-case for CDNs, thus pushing the SUT to its limit.

Intel Provides Reference Design for High-Performance CDNs

The servers used in these tests meet the optimized hardware resources, open-source libraries and caching frameworks and virtualized infrastructure specification for the **Intel® Select Solutions for Visual Cloud Delivery Network**.

This Intel Select Solution reference design provides high-performance, well-balanced systems based on 3rd generation Intel Xeon Scalable processors with flexible configuration options to meet different visual cloud application requirements. For CDN applications, these processors offer increased memory bandwidth, support for PCIe Gen4, increased core count and cache, new extensions to Intel® Advanced Vector Extensions 512, enhanced crypto processing acceleration, and Intel® Software Guard Extensions (Intel® SGX) for protected execution.

Intel Select Solutions are rigorously benchmarked and utilize NUMA-balanced I/O to ensure maximum throughput and consistent latency in real-world conditions. They include a tightly specified set of hardware components, including new Intel® Optane™ persistent memory 200 series, Intel® Solid State Drive Data Center Family (Intel® SSD D7-P5510 Series), Intel® Server GPU, and Intel® Ethernet 800 Series Network Adapter for improved scalability, reduced latency, and cost savings.

Near Line Rate Throughput

The tests show that the bandwidth for the single processor server averaged 192 Gbps[2] – nearly saturating the two 100 GbE connections (right column, Figure 4). The bandwidth results for this server and the dual-processor server are based on an average of five tests. These test results show that a server based on a single 3rd generation Intel Xeon Scalable processor is a good solution for a video on demand (VoD) provider to use in a mid-sized town serving tens of thousands of customers where the lower power consumption and reduced capital expense of a single-processor server, compared to a dual-processor system, are important considerations.

As a point of comparison, Figure 4 also shows performance from a set of tests that were conducted by Intel in 2020 (left). The SUT for these tests used the same cache hit rate, two 100 Gbps network adaptors, and were conducted using bare metal servers and ten P4510 NVMe drives. But the server in these tests used dual 2nd generation Intel Xeon Scalable processors, which achieved a maximum throughput of 179 Gbps.[3]

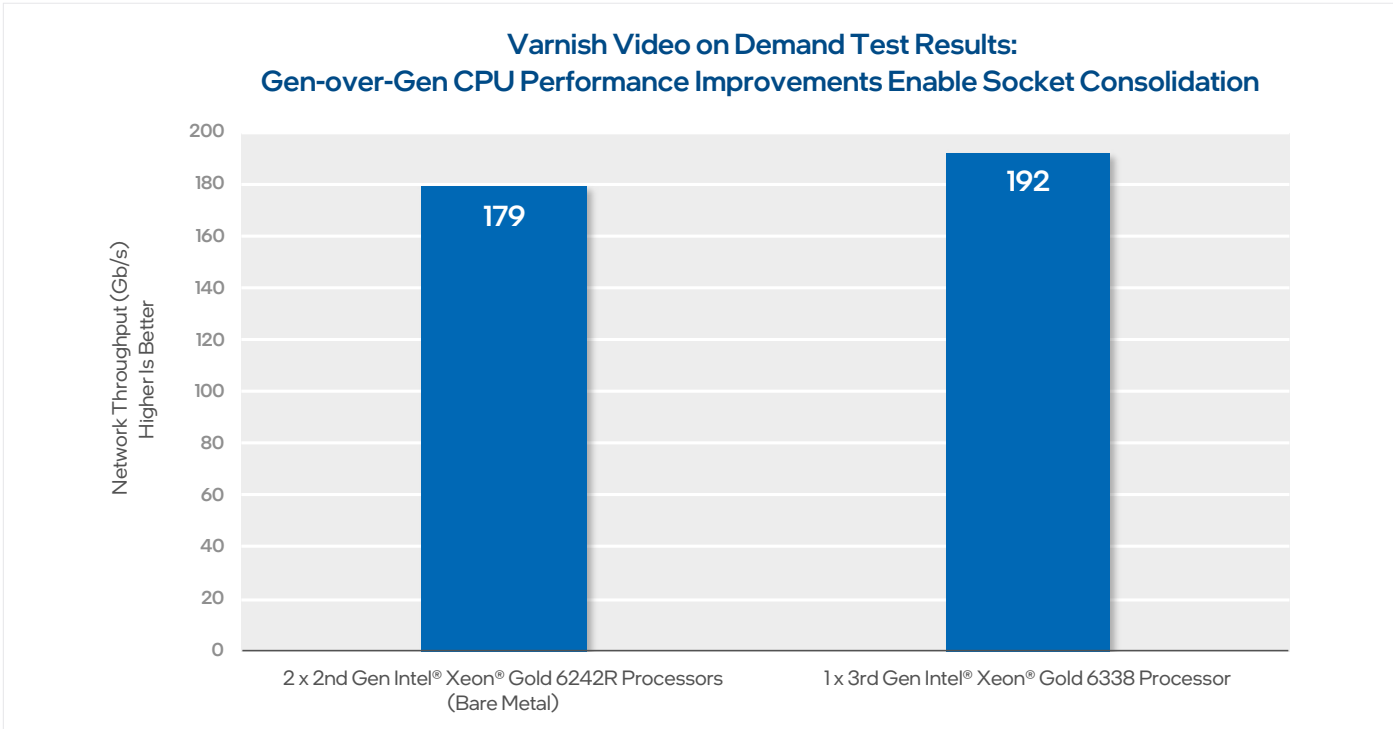

Figure 4. Testing with a single 3rd Gen Intel® Xeon® Scalable processor achieves higher performance than previous testing with dual 2nd Gen Intel® Xeon® Scalable processors.[2,3]

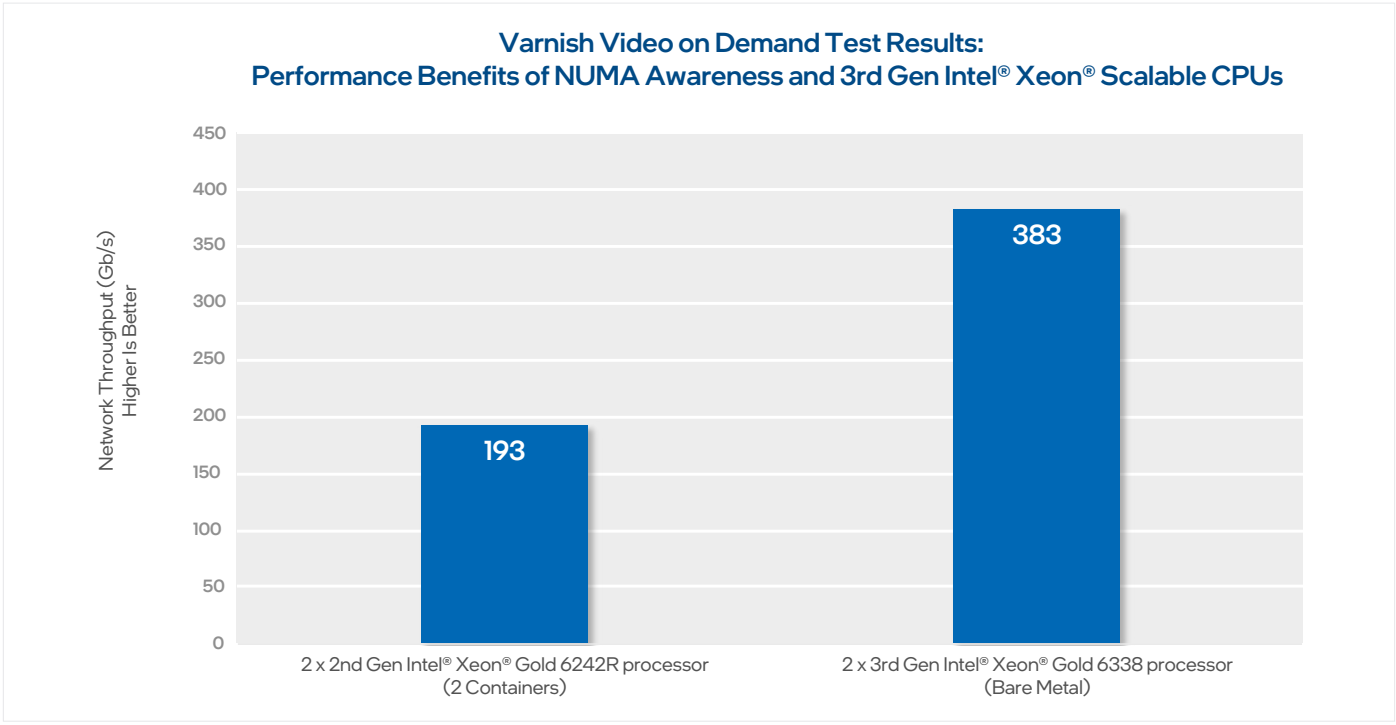

Figure 5. The performance achieved on the new two-processor system is only possible with both the gen-over-gen hardware improvements and the software improvements from Varnish.[3,4]

The dual processor server test utilizing 3rd generation Intel Xeon Scalable processors showed throughput of 383 Gbps[4] (Figure 5, right), about double the performance of the single processor-based server test results. This performance benefited from the efficient and performant design of Varnish Enterprise, including its built-in NUMA awareness capabilities and in-core TLS.

This performance is a dramatic improvement over a similar test conducted by Intel in 2020 using a dual-processor server based on 2nd generation Intel Xeon Scalable processors.

In these older tests,[3] two Varnish instances were run in containers, where each container was only able to access NUMA local resources and was also configured for a 100% cache hit rate. Using these test criteria, the system was able to achieve throughput of 193 Gbps (Figure 5, left). With the new testing, using 3rd generation Intel Xeon Scalable Processors, the NUMA awareness features in Varnish enable the benefits of NUMA-local resources without having to explicitly create two containers that are only able to access certain cores, memory, networking, and drives.

Conclusion

To keep pace with the growing demand for streamed content, CDNs need higher throughput. Varnish has developed its caching server technology to take advantage of the performance of 3rd generation Intel Xeon Scalable processors. The tests with Varnish Enterprise software show single-processor server performance of 192 Gbps and dual-processor server performance of 383 Gbps—a near wire-speed performance. The combination of Varnish high performance software and next-generation CPUs from Intel helps deliver higher throughput per node, resulting in fewer nodes needed to meet the growing capacity needs. The combined solution from Varnish software and Intel enables CDN providers to keep up with the growing popularity and evolution of video streaming.

Learn More

Varnish Software

Intel® Select Solutions for Visual Cloud Delivery Network

3rd generation Intel® Xeon® Scalable processor

Intel® Network Builders

Intel® Visual Cloud resource page

Notices & Disclaimers

[1] https://www.digitaltveurope.com/2020/05/14/almost-4-billion-hours-watched-as-live-streaming-industry-benefits-from-lockdown/

[2] 3rd generation Intel Xeon Scalable testing done by Intel in May 2021. Single processor SUT configuration was based on the Supermicro SMC 110P-WTR-TNR single socket server based on Intel® Xeon® Gold 6338 processor (microcode: 0xd000280) with 32 cores operating at 2.0 GHz. The server featured 256 GB of RAM. Intel® Hyper-Threading Technology was enabled, as was Intel® Turbo Boost Technology 2.0. Platform controller hub was the Intel C620. NUMA balancing was enabled. BIOS version was 1.1. Network connectivity was provided by a single 100 GbE Intel® Ethernet Network Adapters E810. 1.2 TB of boot storage was available via an Intel SSD. Application storage totaled 3.84TB and was provided by 8 Intel P5510 SSDs. The operating system was Ubuntu Linux release 20.04 LTS with kernel 5.4.0-65 generic. Compiler GCC was version 9.3.0. The workload was wrk/master (April 17, 2019), and the version of Varnish was varnish-plus-6.0.8r1. Openssl v1.1.1h was also used.

[3] Tests using 2nd generation Intel Xeon Scalable processors conducted by Intel 8/13/2020: Servers included 2x Intel® Xeon® Gold 6242R Processor, 20 Cores. Intel® Hyper-Threading Technology was enabled, as was Intel® Turbo Boost Technology 2.0. Total Memory 384GB (12 slots/32GB/2666MT/s), BIOS 3B14.RTN01 (microcode: 0x5003003), 12x Intel® P4510, 2x Mellanox CX5, Ubuntu 20.04, kernel 5.4.0-42-generic, varnish-plus-6.0.6r7 revision 7671b56368fe49cb39e36e6e55db3b7c1646d2cf, Docker version 19.03.12, build 48a66213fe. 2 clients, wrk 4.1.0-4-gd2efada-dirty (keep alive on, 200 total connections). Throughput measured with 100% Transport Layer Security (TLS) traffic with indicated target cache hit ratios (either 100% or 80%).

[4] 3rd generation Intel Xeon Scalable testing done by Intel in May 2021. Dual processor SUT configuration was based on the Supermicro SMC 22OU-TNR dual socket server based on Intel® Xeon® Gold 6338 processor (microcode: 0xd000280) with 32 cores operating at 2.0 GHz. The server featured 256 GB of RAM. Intel® Hyper-Threading Technology was enabled, as was Intel® Turbo Boost Technology 2.0. Platform controller hub was the Intel C620. NUMA balancing was enabled. BIOS version was 1.1. Network connectivity was provided by four 100 GbE Intel® Ethernet Network Adapters E810. 1.2 TB of boot storage was available via an Intel SSD. Application storage totaled 3.84TB and was provided by 12 Intel P5510 SSDs. The operating system was Ubuntu Linux release 20.04 LTS with kernel 5.4.0-65 generic. Compiler GCC was version 9.3.0. The workload was wrk/master (April 17, 2019), and the version of Varnish was varnish-plus-6.0.8r1. Openssl v1.1.1h was also used.

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.

Your costs and results may vary.

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.

© Intel Corporation. Intel, the Intel logo, and other Intel marks are trademarks of Intel Corporation or its subsidiaries. Other names and brands may be claimed as the property of others.

0721/DO/H09/PDF ♻ Please Recycle 347710-001US