White Paper

Communication Service Providers
Telco Network Operators

Delivering up to 500 Gbps Throughput for Next-Gen CDNs[1]

Achieve high CDN throughput for UHD content with Varnish Edge Cloud running on Intel architecture-based servers

Content delivery networks (CDNs) that support popular video on demand (VoD) and live video services are under increasing load. CDN nodes need to service an increased number of concurrent users, support caching for ultra-high definition (UHD) content, as well as more immersive and personalized content. To keep up with the demand, CDN nodes need greater performance across input/output (IO), memory and compute. To illustrate the growing need for CDN performance; a high-definition (1080p) stream can consume up to 6.77 Mbps per stream, and upgrading the quality of that same content to ultra-high definition (4K) would increase the bitrate to 26.82 Mbps per stream, a near 4X increase in required bandwidth.[2]

Added to VoD trends is the explosion of live streaming consumption. Live streaming traffic includes gaming platforms, new live video social media platforms, and services that are broadcasting large-scale in-person events, such as sports, plays and concerts, as well as smaller scale events—including education or business meetings – that is also growing in popularity.

This increase in live streaming can be seen in a report[3] from leading CDN service provider Akamai which showed its traffic growing from 70 Tbps in the fourth quarter of 2018 to 180 Tbps in the same quarter of 2020. The company predicts that its traffic could hit 25,000 Tbps as viewers grow to 2.5 billion and the average streaming bandwidth totals 10 Mbps.

To continue to support this growth will require higher performance CDN platforms. Varnish and Intel have teamed up to explore the performance limit of their commercial CDN software running on a broadly available Intel architecture-based server without the use of accelerators. While accelerators, such as offloading Transparent Layer Security (TLS) processing to the network adapter, can be used to improve performance, they often require code changes to the CDN software.

It wasn't long ago that a 200Gbps CDN cache node was state of the art. But as can be seen by the rapid growth in peak bandwidth demand (as shown by the Akamai report above), 200Gbps may not be enough and much higher throughput levels will soon be necessary to meet peak capacity needs at high bandwidth and high user density. These tests show that CDN software, with modern design, when used in conjunction with a server configured for high performance, is no longer a bottleneck when scaling beyond 200Gbps.

Intel and Varnish have teamed up for a series of performance white papers that examine the "better" performance of a cost-optimized platform (that paper is available here) and the "best" performance described in this document. Through both test results, Intel and Varnish seek to demonstrate the ability to provide solutions for all types of use cases.

Table of Contents

Highly Efficient Varnish Software is a Fast Caching Solution 2

CDN Performance on Single or Dual Processor Servers 3

Testing Configuration 3

Benchmarking High Throughput CDN Edge Cache Nodes 4

VoD and Live Linear Test Methodologies 4

Exceeding 500Gbps of CDN Edge-Cache Node Throughput . . . 5

Conclusion 7

Appendix A: Sample Wrk Configurations 8

Appendix B: Sample query configurations for Wrk 8

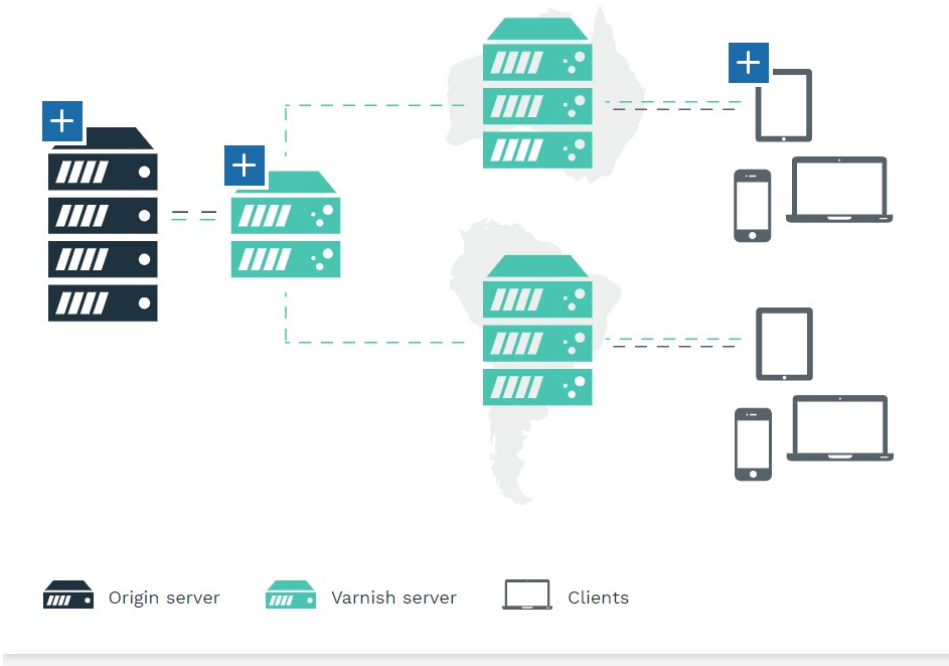

Figure 1. Two-tier Varnish solution architecture for a large-scale CDN.

Highly Efficient Varnish Software is a Fast Caching Solution

Varnish Edge Cloud is a powerful, feature-rich web cache and HTTP(S) accelerator that is used by a wide range of content and communications service providers (CoSPs) 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 low-latency content 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.

Varnish Edge Cloud 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(S) video streams in large-scale communications networks that face challenging workload demands.

CDN Performance on Single or Dual Processor Servers

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

Testing Configuration

While Varnish can cache a wide range of content types, for these tests it was optimized for video on demand, with the ability to also handle live video. Both single and dual processor configurations were tested to address a range of use cases and performance needs. Historically, non-uniform memory access (NUMA) has been associated with multi-processor systems, but this is not its only use case. In modern CPU architectures, both the single-processor and dual-processor systems can utilize NUMA. NUMA architectures enable a processor to access its local memory and I/O resources as well as those in another NUMA region and understand that a given resource may be remote.

The strong performance scaling from the single to dual processor system is due, in part, to NUMA awareness in Varnish, which was added into the latest release and 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.

In the single processor server, we are using the Sub-NUMA Clustering feature (SNC) to split the single CPU into two NUMA regions. This is beneficial as this workload can be bottlenecked by Linux page reclaim. Enabling SNC reduces the impact of that Linux behavior and allows for higher total throughput.

Bare metal servers were used for the SUTs. For these tests, requests were made randomly across the full span of the dataset, resulting in content that is all of equal popularity, to show performance under a more difficult scenario than is likely in real-world conditions.

Test traffic was generated using clients designed to emulate user devices. A typical real-world CDN client is a single-user device (TV, computer, mobile device, etc.) 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 systems that emulated the content requests of thousands of clients and consistently generated up to 100 Gbps of data traffic per system.

These client systems were connected to the CDN servers using 100 GbE links through a switch; 4x100 GbE connections for the single-processor SUT, and 8x100 GbE for the dual-processor SUT. Testing was done using Wrk, a widely recognized open-source HTTP(S) benchmarking tool. The single processor test setup can be seen in Figure 2.

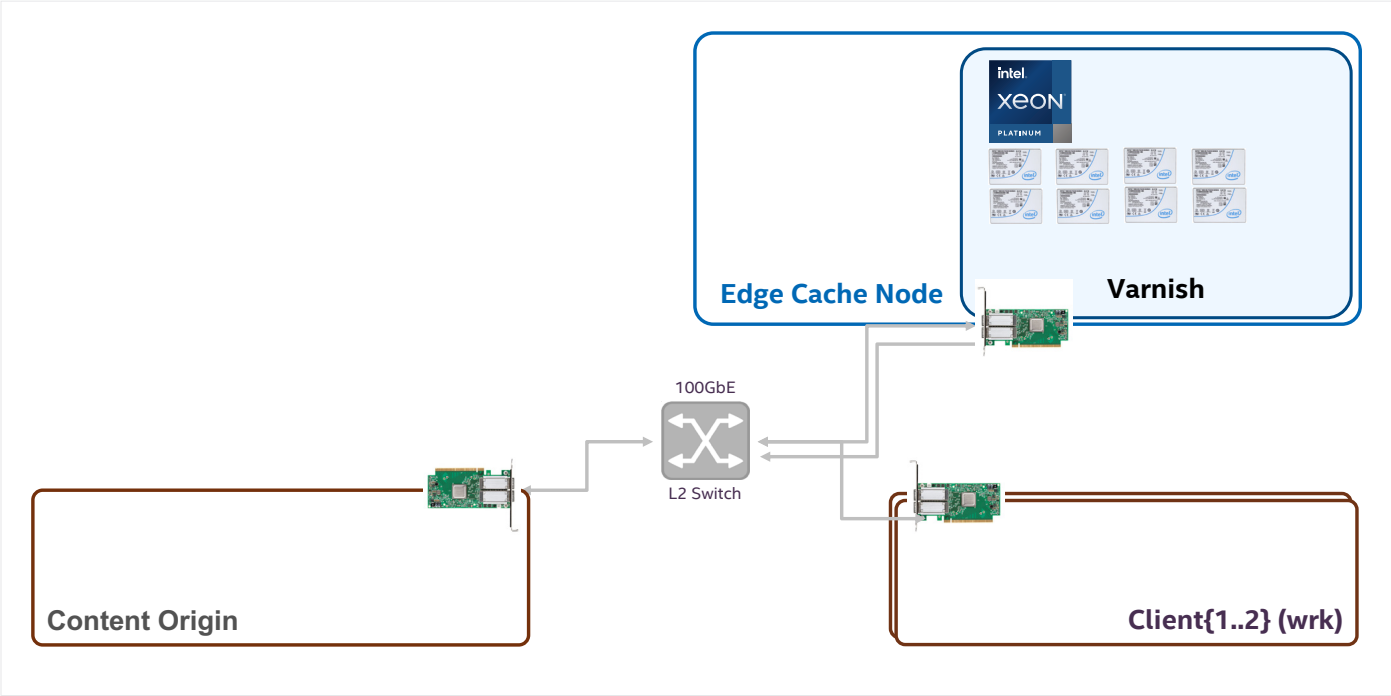

Figure 2. Single processor CDN test configuration.

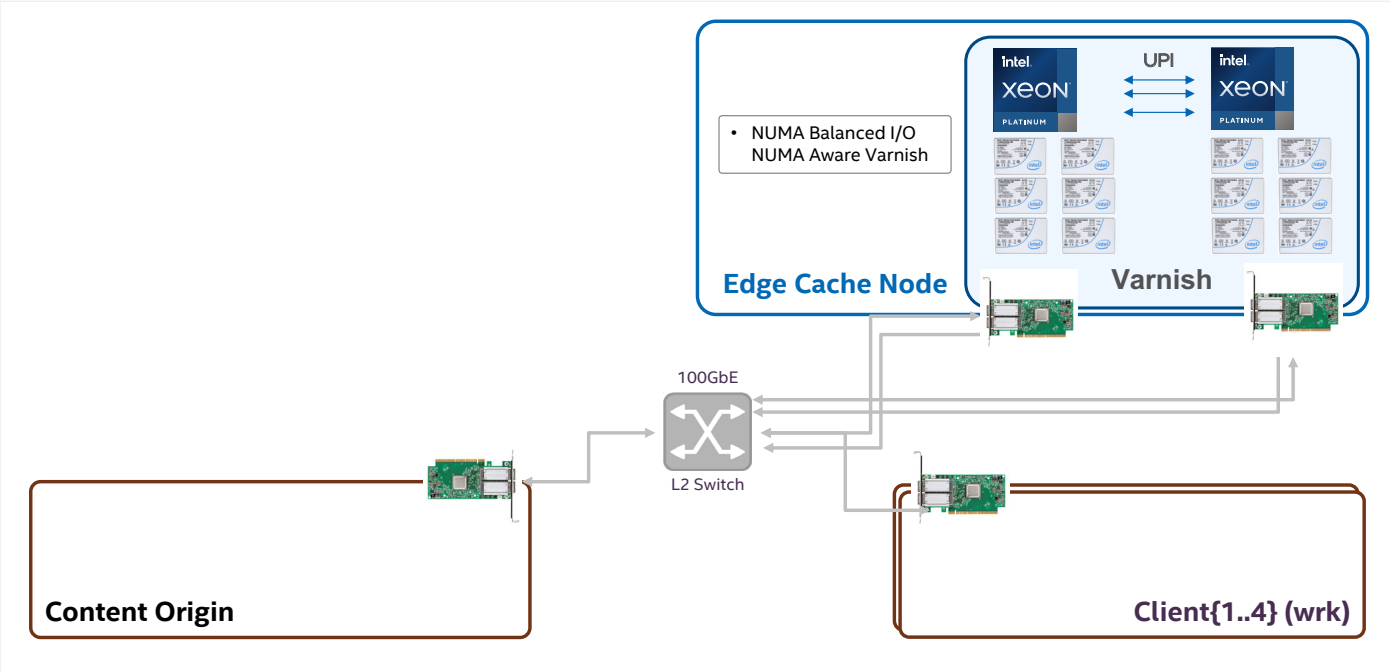

Figure 3. Dual-processor server CDN test setup

The dual-processor version of the test setup is similar but with four dual-port connections for a total of 800 Gbps into the dual processor server as shown in Figure 3.

Benchmarking High Throughput CDN Edge Cache Nodes

Two workloads were tested on these servers. The primary workload, which drove the hardware configuration, emulated a very high throughput VoD service. As such, the servers were equipped with a fan-out of PCIe Gen4 NVMe storage and PCIe Gen4 100GbE adapters.

The VoD tests were run against a cache that had been pre-filled with the test dataset so that during benchmarking all data could be found on disk and there were no requests to the origin server, similar to how a real-world cache node could potentially be pre-populated with content during off-peak hours. In a real-world application, the most popular content available on a cache node is kept in DRAM for fastest access, with the remainder of cached content being stored on NVMe disk storage.

However, in these tests, all content in the dataset is considered to be of equal popularity and is requested randomly. NVMe disk storage. However, in these tests, all content in the dataset is considered to be of equal popularity and is requested randomly, forming the worst case test scenario, increasing the probability of NVMe disk access for every request. Additionally, each physical client has its own dataset, further reducing the likelihood that data will be found in DRAM. The tests were configured for a 100 percent cache hit rate, meaning no content requests are made from the origin server during the test.

The second workload tested was a live-linear workload leveraging the existing DRAM in the server and some of the capabilities of Varnish Configuration Language (VCL) to allow these systems to also be able to handle a live workload during time periods when that content is more popular than the VoD that the systems would normally be handling, such as a very popular sporting event. VCL allows Varnish to make caching decisions based on the URL, in this case used to keep live content in DRAM only while allowing the VoD content to be cached on disk as well as in DRAM.

Given the nature of a live workload where the cache is constantly updated to maintain a buffer that is typically only a few minutes long, the live-linear tests are executed with a target cache hit ratio of 93.3%. To control the hit rate, the dataset was sized such that only 93.3% of it can fit in the memory available for Varnish to use on the system at any given instant in time. As with the VoD workload, all content was of equal popularity and requested randomly.

Each of these workloads were run on both the single and dual-processor server configurations using a total of 256, 512 and 4,096 connections per processor, with each Wrk instance generating a portion of the connections. These connection levels provided a range of performance under different test conditions. In both the VoD and live-linear workloads, data was encrypted using Transport Layer Security (TLS 1.2) for all requests between the client and the cache node.

VoD and Live Linear Test Methodologies

Executing the VoD and live video tests required multiple clients equipped with 100GbE network adapters to simulate the load that would be applied to the cache nodes by thousands of video players. To do this, each of the clients were configured similarly with benchmarking tests executed on all clients at the same time using parallel-ssh.

Wrk HTTPS requests were submitted as fast as possible up to the limits of the client CPU or adapter unless an inter-request delay is configured in the query. In addition, Wrk was configured for a uniform random access pattern. This access

pattern is a near worst-case scenario for a cache node that has the bulk of the popular content cached on NVMe disks and a very small subset in DRAM.

One consideration when using Wrk to benchmark a cache node is that as the number of total connections is increased, an inter-request delay needed to be added to maintain performance. Without this inter-request delay, Wrk itself can become the bottleneck rather than the cache node. Also, an inter-request delay emulates a real-life video player that requests video, audio and subtitle/text objects at a regular interval to keep its video buffers full.

For these tests, 256 and 512 connections per processor are executed without a delay, and 4096 connections per processor is configured to use a delay randomly selected within the range of 115 to 120 milliseconds. See Appendix A for an example of a Wrk execution script and Appendix B for details on the queries used.

The number of connections was determined by input from customers as well as experience using Wrk. The compute workload on a CDN edge cache node is comprised of the TLS encryption and managing TCP connections. Each connection has a finite compute and memory footprint, so consumption of those resources increases as the connection count grows, eventually coming to a point where the compute overhead to maintain the connections can diminish the throughput per connection.

These tests were designed to show the performance under three different test conditions rather than just at the optimal number of connections. For instance, these effects can be seen in the results where total throughput at 4096 connections is less than at 512 connections. It is important to note that due to the bursty nature of Wrk, one Wrk connection is not equivalent to one end-user connection.

Exceeding 500Gbps of CDN Edge-Cache Node Throughput

The headline performance of more than 500 Gbps was seen in tests of the live-linear emulated services as seen in Figure 4. The maximum performance for both single and dual processor systems was seen at the 256 connection per CPU test case (512 connections for the dual processor case). In a single processor configuration, the system reached 366.38 Gbps; in the dual-processor configuration the throughput totaled 509.65 Gbps.

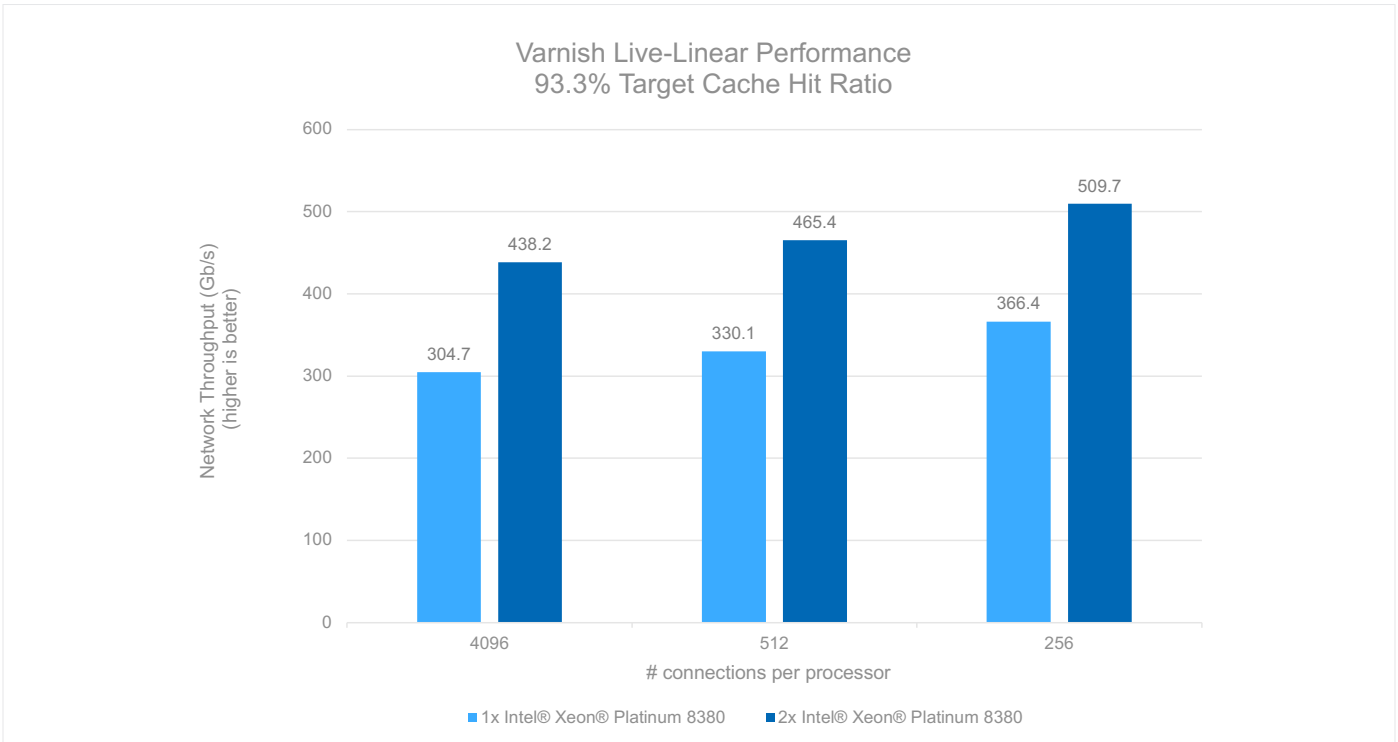

Figure 4. Live-linear video performance for Varnish CDN software running on single and dual socket servers. The highest network throughput performance came with 256 connections.

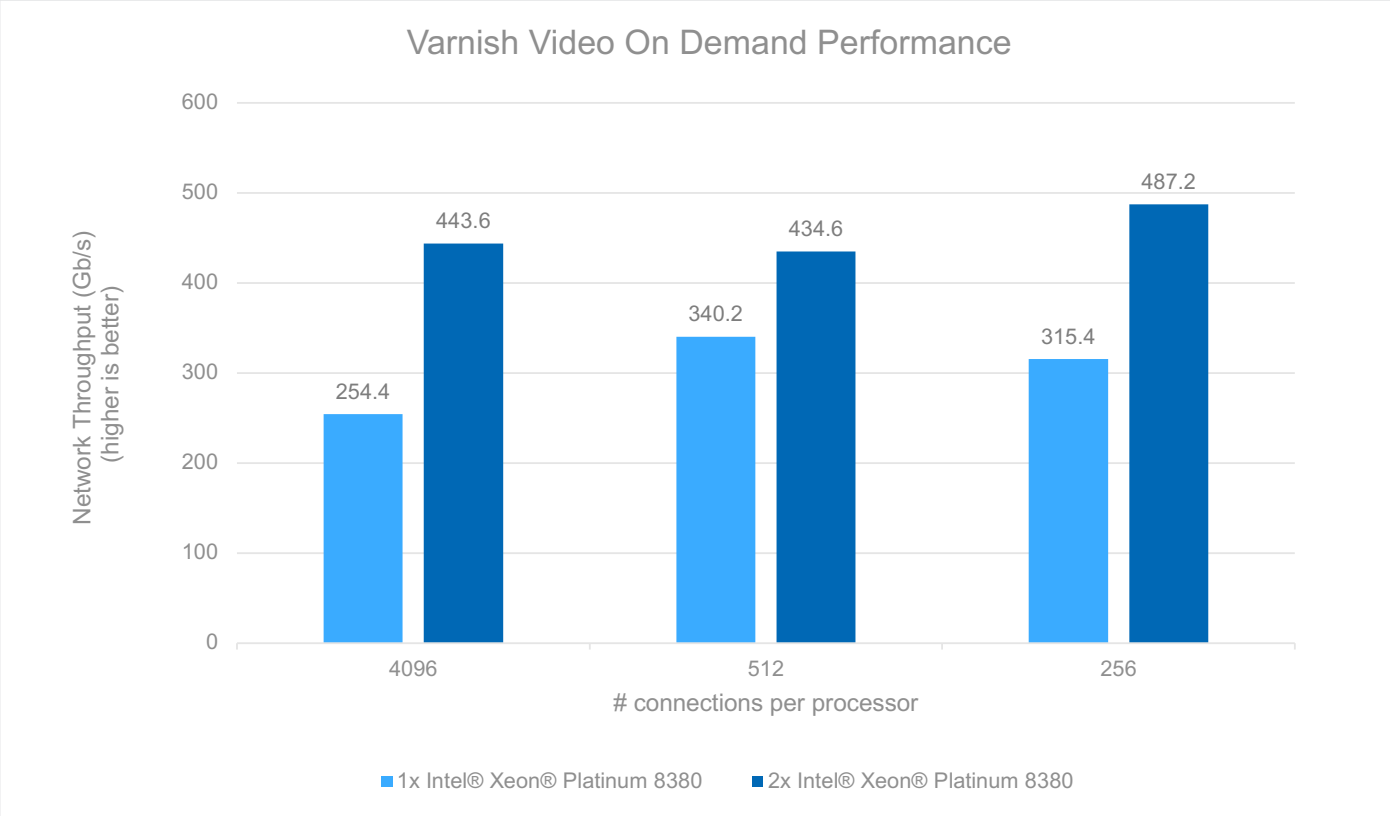

Figure 5. VoD performance for Varnish CDN software running on single and dual socket servers. The highest performance came with 256 connections.

The VoD tests (see Figure 5) show that the throughput at 256 connections per CPU was 315.4 Gbps for the single processor server reached 315.4 Gbps and the dual processor server maxed out at 487.20 Gbps. The maximum single server performance was achieved with 512 connections per processor where it totaled 340.18 Gbps.

All of these test results are based on an average of 5 test runs. These tests show that a CDN cache node running Varnish Edge Cloud can be configured for outstanding VoD throughput, and also for live-linear video streaming performance.

The configurations used in these tests benefited from the efficient and performant design of Varnish Edge Cloud, including its built-in NUMA awareness capabilities and in-core TLS. With this 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 take extra effort to ensure NUMA locality, such as creating multiple virtual machines or containers that are each isolated to a specific NUMA region, or also without the use of any function offloads (such as purpose-built accelerators for TLS processing).

Intel Reference Design for CDNs

For CoSPs and server OEMs that want a performance-validated solution for CDNs, the Intel® Select Solutions for Visual Cloud Delivery Networks is an optimized solution that combines hardware resources, open-source libraries and caching frameworks and virtualized infrastructure specification for CDNs and other visual cloud applications. This Intel Select Solution reference design details high-performance, well-balanced systems based on 3rd generation Intel Xeon Scalable processors.

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.

Conclusion

Thanks to the growing demand for streaming video[3] CDN systems need to continue to increase their performance to allow CoSPs to support this demand. With its built-in NUMA support and optimized cache node functionality, Varnish Edge Cloud CDN is designed for highly efficient and highly performant CDN services. Intel offers a full range of supporting visual cloud technologies including the 3rd generation Intel Xeon Scalable CPU along with Intel Ethernet adapters, Intel NVMe SSDs and Intel® Optane™ Persistent Memory. The record-setting CDN throughput numbers described in this paper set the stage for next-generation CDN services by demonstrating that 500Gbps throughput is possible with commercially available software, broadly available Intel based servers, and without the use of expensive and power-hungry accelerators. Knowing that their future bandwidth needs can be met, CoSPs can trust in Varnish and Intel to help them stay ahead of demand for on-demand and live video content.

Learn More

Varnish Software

Intel® Select Solutions for Visual Cloud Delivery Network

3rd generation Intel® Xeon® Scalable processor

Intel® Network Builders

Intel® Visual Cloud

Appendix A: Sample Wrk Command Lines

Video on Demand, 512 connections per WRK client

```
#!/bin/bash
date
export HOSTNAME=`hostname`
ulimit -n 65535
http_proxy= $HOME/wrk/wrk \
  -t 48 \
  -c 512 \
  -d 120s \
  --timeout 20s \
  -L --latency \
  -s $HOME/wrk/scripts/query.lua \
  https://AAA.BBB.CCC.DDD/
```

Live-Linear, 512 connections per WRK client

```
#!/bin/bash
date
ulimit -n 65535
http_proxy= $HOME/wrk/wrk \
  -t 48 \
  -c 512 \
  -d 120s \
  --timeout 20s \
  -L --latency \
  -s $HOME/wrk/scripts/query_128gb_dram_livelinear.lua \
       https://AAA.BBB.CCC.DDD/
```

Appendix B: Sample query configurations for Wrk

Video on Demand, independent dataset per Wrk client

```
# cat wrk/scripts/query.lua
-- example script that adds a query string

local threadcounter = 1
local threads = {}

function setup(thread)
  thread:set("id", threadcounter)
  table.insert(threads, thread)
  threadcounter = threadcounter + 1
end

function init(args)
  math.randomseed(os.time()*id)
end

request = function()
  -- for multiple clients with independent datasets
  local hostname = os.getenv("HOSTNAME")
  local param_value = math.random(600000)
path = "/1MB/.file?version=" .. hostname .. "_" .. param_value
return wrk.format("GET", path)
end
```

Live-Linear.

```
# cat wrk/scripts/query.lua
-- example script that adds a query string

local threadcounter = 1
local threads = {}

function setup(thread)
  thread:set("id", threadcounter)
  table.insert(threads, thread)
  threadcounter = threadcounter + 1
end

function init(args)
  math.randomseed(os.time()*id)
end

request = function()
  param_value = math.random(140400)
  path = "/1MB/.file?live=" .. param_value
  return wrk.format("GET", path)
end
```

intel®

Notices & Disclaimers

[1] 3rd generation Intel Xeon Scalable testing done by Intel in September 2021. Single processor SUT configuration was based on the Supermicro SMC 110P-WTR-TNR single socket server based on Intel® Xeon® Platinum 8380 processor (microcode: 0xd000280) with 40 cores operating at 2.3 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 two 100 GbE Intel® Ethernet Network Adapters E810. 1.2 TB of boot storage was available via an Intel SSD. Application storage totaled 3.84TB per drive and was provided by 8 Intel P5510 SSDs. The operating system was Ubuntu Linux release 20.04 LTS with kernel 5.4.0-80 generic. Compiler GCC was version 9.3.0. The workload was wrk/master (April 17, 2019), and the version of Varnish was varnish-plus-6.0.8r3. Openssl v1.1.1h was also used. All traffic from clients to SUT was encrypted via TLS.

3rd generation Intel Xeon Scalable testing done by Intel in September 2021. Dual processor SUT configuration was based on the Supermicro SMC 220U-TNR dual socket server based on Intel® Xeon® Platinum 8380 processor (microcode: 0xd000280) with 40 cores operating at 2.3 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 Mellanox MCX516A-CDAT adapters. 1.2 TB of boot storage was available via an Intel SSD. Application storage totaled 3.84TB per drive and was provided by 12 Intel P5510 SSDs. The operating system was Ubuntu Linux release 20.04 LTS with kernel 5.4.0-80- generic. Compiler GCC was version 9.3.0. The workload was wrk/master (April 17, 2019), and the version of Varnish was varnish-plus-6.0.8r3. Openssl v1.1.1h was also used. All traffic from clients to SUT was encrypted via TLS.

[2] https://www.seagate.com/video-storage-calculator/

[3] https://www.ir.akamai.com/static-files/1ef574a5-ae14-48f6-854a-47d96c4a75fe; slides 42 & 43

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.

1021/TM/H09/PDF ♻ Please Recycle 348938-001US