VARNISH SOFTWARE

White paper:

VARNISH SOFTWARE AND INTEL – Optimizing content delivery network (CDN) performance

Varnish Software and Intel – Optimizing content delivery network (CDN) performance

Executive summary

This white paper outlines learnings from a joint collaboration between Varnish Software and Intel, using Varnish caching technology and Intel's next-generation CDN platform with an aim to deliver an optimized reference solution for CDN providers.

Optimizing content delivery network (CDN) performance

Millions of data centers and telecom towers strategically placed across the planet, over 885,000 kilometers of undersea cable, billions of devices, and nearly all of it in service of one thing: online video[1]. The unseen complexity of delivering content smoothly, in milliseconds, to global audiences across devices drives constant development and innovation -- both in seeking new technologies to power speed and enable scalability but also in optimizing existing hardware and software to maximize the performance it can deliver. This is particularly true in accelerating and enhancing CDN performance.

This white paper documents a collaboration between Varnish Software and Intel, using Varnish caching technology alongside Intel's next-generation CDN platform to deliver an optimized reference solution for CDN providers.

Background: Video dominates content delivery

Today, consumers expect their web and video content to be available anytime, anywhere and on any device. Video makes up at least 80% of current internet traffic, and general internet traffic per day per person is expected to reach 1.5 GB during 2020[2]. More demanding consumer trends, such as streaming high-definition video across devices, are becoming the norm. IoT-related traffic demands and high-data-intensity use cases requiring ultra-low latency transit will come to dominate the landscape, making high-throughput content delivery essential.

As this content continues to get richer, more localized, personalized, interactive and immersive, the pressure is on for the CDN providers to deliver much more with less. CDN providers are increasingly looking for ways to unlock vital new avenues for content delivery, automation, and intelligence. Their implementation needs careful thought to ensure internal systems are capable of handling the increased demand that these technologies enable.

Strengthening the backbone and expanding the scope of content delivery

The CDN is arguably the backbone of today's video delivery, and it, like the content delivery use cases CDNs must enable, will continue to evolve to meet increasing demand. To meet the demand for seamless HD and 4K streaming of videos, films, and live events, as well as to support unparalleled growth in OTT (over-the-top) and emerging edge use cases, such as cloud gaming, immersive augmented reality, virtual reality (VR), and real-time, multiplayer gaming, new approaches are needed to address more open, real-time, and cloud-based CDN solutions.

CDN providers need greater speed and capacity at the edge to provide seamless delivery of this content. To keep up with growing demand, CDN providers are increasingly using 100GbE network interface technology to serve 100Gb cache servers. These servers are driving the need for faster storage technologies (such as PCIe* storage devices). Increasingly, larger memory subsystems are needed to host the most popular content directly in memory to make it even faster and more accessible to stream out.

As CDN providers embrace the shift to higher network throughput and newer storage and customer memory technologies, software plays an essential role in helping CDN providers to realize the fullest potential that the platform has to offer.

References

[1] https://itpeernetwork.intel.com/global-video-demand/#gs.i2c3gc
[2] Cisco: https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

Varnish Software and Intel

Varnish Software and Intel joined forces to assist a mutual customer in building their own CDN. This project opened the door to further engineering-oriented discussions to foster cooperation between the two companies, leading to a joint technology-research partnership to address CDN providers' critical issues, particularly around performance.

Having previously achieved 100 Gbps streaming with a single Varnish node in testing conducted both by Varnish and Intel independently, Varnish and Intel wanted to work together to test the bounds of this capacity even further.

Testing methodology and CDN test environment

Using a series of test cases simulating different conditions, testing was conducted at Intel labs using various Intel hardware setups and Varnish cache configurations (see "Test configurations, test cases and results" for more detail). The aim was to maximize reliably replicable performance from both bare-metal CDN and containerized test environments with Varnish edge cache nodes.

How would Varnish caching technology perform when deployed in various scenarios including on bare metal, within containers using Intel's latest hardware, and with in-core TLS?

CDN test environment

Four-node bare metal setup (2 x 100GbE)

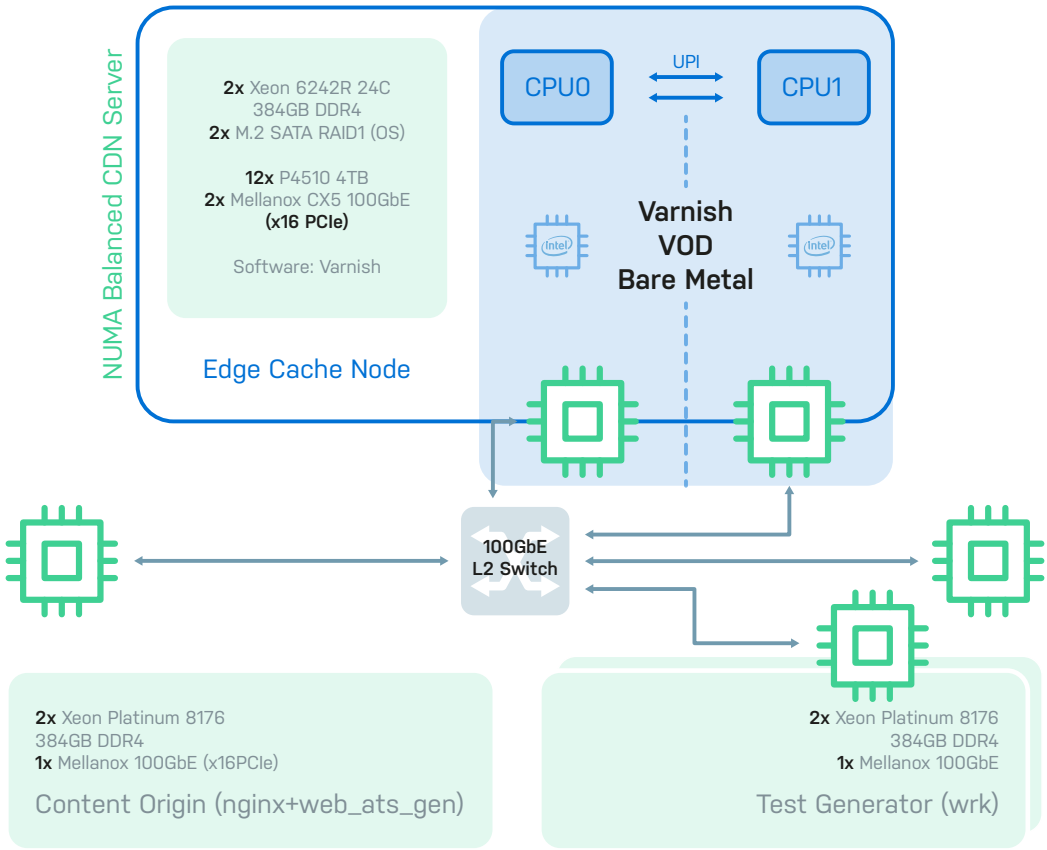

Figure 1: Bare metal CDN test environment

Four-node containerized setup (2 x 100GbE)

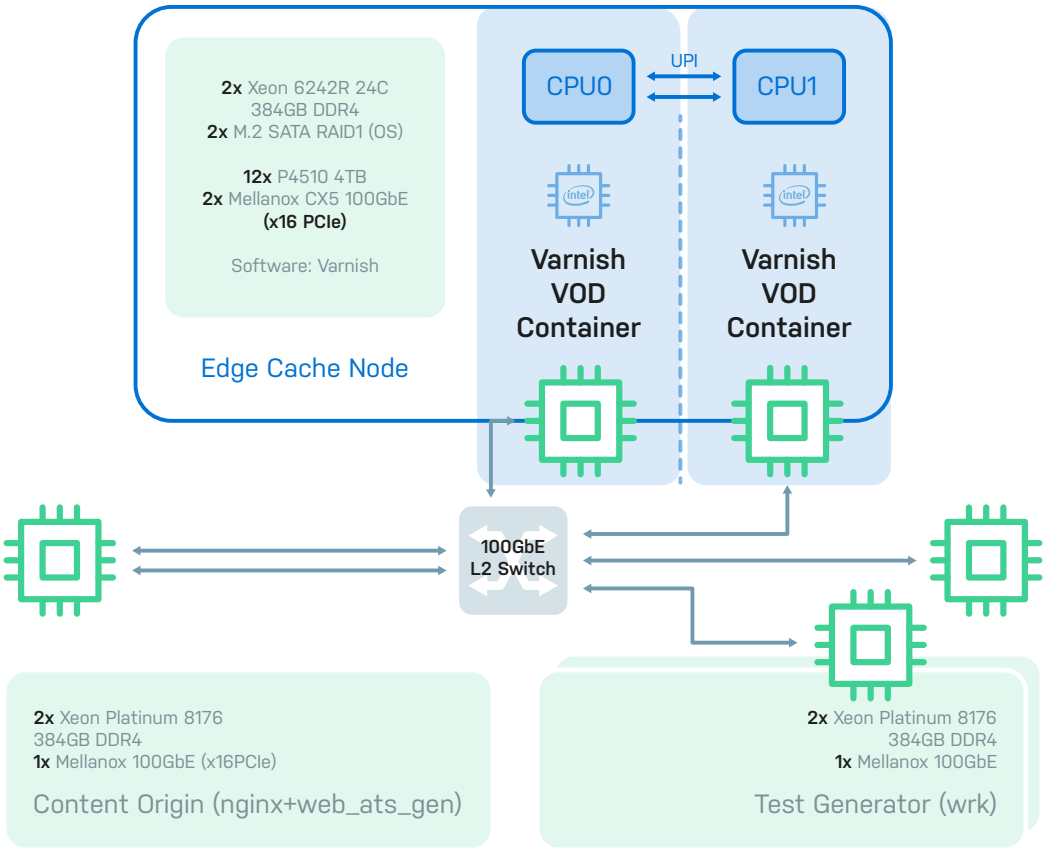

Figure 2: Containerized CDN test environment

To undertake testing under different conditions, the size of the stores were manipulated to vary the cache hit rate, i.e. in order to test performance at 100% cache hit rate, 90%, 80%, etc.

Test configurations, test cases and results

The test cases used across the test series consisted of the same configurations throughout, with one exception:

- the cache hit rate was adjusted for each new test series as indicated below

In the figures displayed in the test case result charts below, the lower the numbers reflected, the better the result. For example, in test case 1, the average time to first byte is 1 millisecond, which shows the near-instant delivery of the first byte. The lower the number, the lower the latency, which is what the Varnish-Intel set-ups aim for.

Test Case 1: Warm cache, 100% cache hit, bare metal, with built-in TLS

Bare-metal VoD/Web CDN HTTP

Single Varnish instance across sockets

2x100Gbps

5MB video objects

Average network TX throughput: 178.7 Gb/s

Average CPU usage: 82.52%, 34 cores (6242R)

Average Time To First Byte (TTFB) latency: 1ms

Maximum (3 runs) P99 TTFB latency: 6.36ms

Average Time To Last Byte (TTLB) latency: 34.8ms

Maximum (3 runs) P99 TTLB latency: 242.71ms

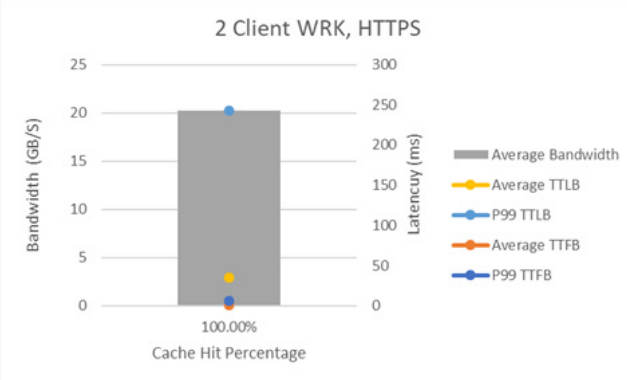

Figure 3: Bare metal CDN 100% cache hit ratio metrics, HTTPS

Test Case 2: Warm cache, 80% cache hit, bare metal, with built-in TLS

Bare-metal VoD/Web CDN HTTPS

Single Varnish instance across sockets

2x100Gbps

5MB video objects

Average network TX throughput: 149.63 Gb/s

Average CPU usage: 90.86%, 37 cores (6242R)

Average Time To First Byte (TTFB) latency: 3.83ms

Maximum (3 runs) P99 TTFB latency: 19.19ms

Average Time To Last Byte (TTLB) latency: 33.37ms

Maximum (3 runs) P99 TTLB latency: 197.88ms

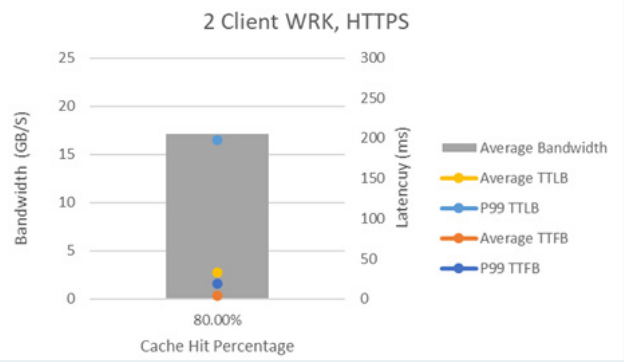

Figure 4: Bare metal CDN 80% cache hit ratio metrics, HTTPS

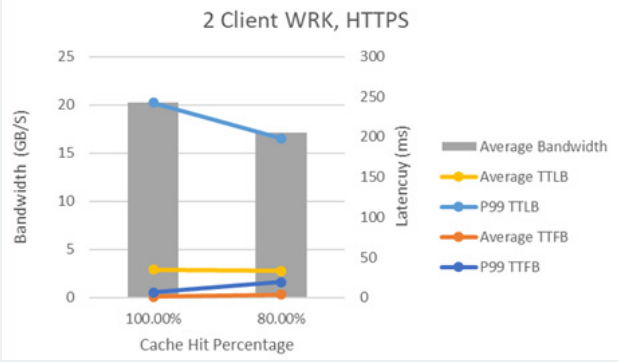

Figure 5: Bare metal bandwidth and latency results

References

[3] Test conducted by Intel on 13/08/2020

Test Case 3: Warm cache, 100% cache hit, containerized, with built-in TLS

Containerized VoD/Web CDN HTTPS

Two Varnish instances, one per socket

2x100Gbps

5MB video objects

Average network TX throughput: 193.23 Gb/s

Average CPU usage: 65.32%, 27 cores (6242R)

Average Time To First Byte (TTFB) latency: 0.94ms

Maximum (3 runs) P99 TTFB latency: 5.37ms

Average Time To Last Byte (TTLB) latency: 31.44ms

Maximum (3 runs) P99 TTLB latency: 255.32ms

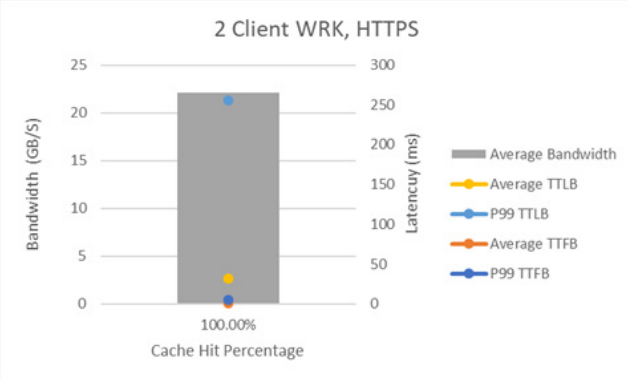

Figure 6: Containerized CDN 100% cache hit ratio metrics, HTTPS

Test Case 4: Warm cache, 80% cache hit, containerized, with built-in TLS

Containerized VoD/Web CDN HTTPS

Two Varnish instances, one per socket

2x100Gbps

5MB video objects

Average network TX throughput: 194.43 Gb/s

Average CPU usage: 72.53%, 30 cores (6242R)

Average Time To First Byte (TTFB) latency: 1.48ms

Maximum (3 runs) P99 TTFB latency: 7.7ms

Average Time To Last Byte (TTLB) latency: 30.26ms

Maximum (3 runs) P99 TTLB latency: 244.541ms

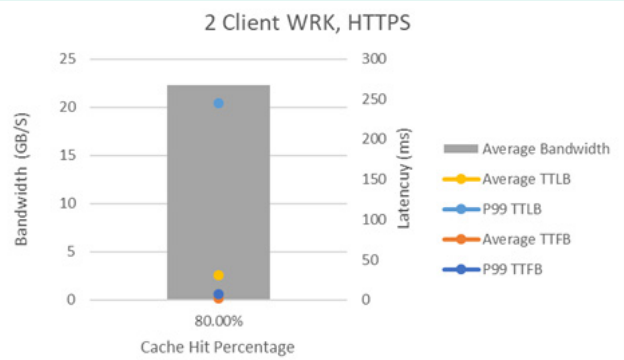

Figure 7: Containerized CDN 80% cache hit ratio metrics, HTTPS

Figure 8: Containerized bandwidth and latency results

Discussion

For an I/O intensive application, like a Varnish cache node, both raw throughput and predictable latency are critical performance metrics. The configuration of the cache node used for these experiments improves performance by enabling 12 NVMe drives and 2 100GbE NICs to be installed in a balanced manner.

Most systems do not have the ability to have as many NICs and drives installed at the same time without PCIe* oversubscription, and of those that do, most do not offer NUMA-balanced I/O. Having one NIC and five drives on each socket improves total throughput, reduces the likelihood of data being on a drive attached to the remote socket, and makes transaction latency easier to predict.

Conclusion

As illustrated in the test case results, the combination of Varnish Software and Intel hardware offers an optimized reference solution for CDN providers on both bare-metal and containerized CDN deployments.

As CDN providers begin to look at the shape of future content delivery, including the video-heavy present and near-future, achieving their strategic aims will require careful planning of their internal systems to handle the increased demand and a view on how content delivery and CDNs themselves are evolving with technology. The Varnish Software and Intel reference solution offers an immediate response to the current and changing challenges of content delivery, particularly in the area of high-performance CDN video throughput with TLS.

About Varnish Software

Varnish Software's powerful caching technology helps the world's biggest content providers deliver lightning-fast web and streaming experiences for huge audiences, without downtime or loss of performance.

Our solutions combine open-source flexibility with enterprise robustness to speed up media streaming services, accelerate websites and APIs, and enable global businesses to build custom CDNs, unlocking unbeatable content delivery performance and resilience.

Our customers are able to scale easily to match peaks in demand, protect their critical infrastructure and keep costs predictable, enabling them to deliver great web experiences for all of their users, at all times.

VARNISH SOFTWARE

New York	+1 646 586 2052
Los Angeles	+1 310 648 8474
Paris	+33 1 70 75 27 81
London	+44 20 7060 9955
Stockholm	+46 8 410 909 30
Singapore	+65 8434 8028

www.varnish-software.com