WHITE PAPER

Content Delivery Network (CDN) Providers

Network Edge

Increase CDN Channel Density at a Reduced TCO with Intel® Optane^{TM} Technology

Authors

Timothy Miskell

Tushar Gohad

Ai Bee Lim

Nehal Mehta

Chris Cavigioli

Felipe Pastor

Greg Smith

Jeremy Payne

1 Abstract

A Content Delivery Network (CDN) refers to a geographically distributed network of caching servers at the network edge that provides fast and efficient delivery of popular content[1]. Video makes up at least 80% of the current Internet traffic. Between 2020 and 2025 the cloud CDN market is expected to grow at a CAGR of 28%[2]. Consumer trends, such as high-definition video streaming across devices, growing IoT traffic, and data-intensive use cases requiring ultra-low latency transit are expected to make high-throughput content delivery increasingly important. As this content continues to get richer, more localized, personalized, interactive, and immersive, the volume of data required grows substantially. Therefore, the pressure is on CDN providers to deliver more content at a lower cost. CDN providers are increasingly looking for ways to optimize platform performance and cost to deliver to customer expectations for a great experience without breaking the bank. Fortunately, advances in processor, memory and storage architectures are available to make the cost/efficiency equation easier to solve. This paper will show how AT&T* and Intel® have collaborated to demonstrate a foundational architecture for the next generation of open standards-based high performance CDNs; this will include benchmarks on 2nd and 3rd Generation Intel® Xeon® Scalable Processors and Intel® Optane™ Persistent Memory (PMem) technology as the elements to increase video stream density per node, while lowering Total Cost of Ownership (TCO).

Keywords: Content Delivery Network, CDN, Live Linear, linear video, live streaming, video delivery, caching, high endurance storage, throughput, scalability, channel density, 3D XPoint, Intel® Optane™ Persistent Memory 100 Series, Intel® Optane™ Persistent Memory 200 Series, PMem, hot VoD, 2nd Generation Intel® Xeon® Scalable Processors , 3rd Generation Intel® Xeon® Scalable Processors, TCO.

Table of Contents

Figures

Tables

2 Introduction

Content that users request more frequently is typically moved physically closer to the user on regional or edge CDN servers. Per Cisco VNI index[3], CDNs serve most of the Internet traffic today, 82% of which is video traffic delivered over IP. Video content delivered by CDNs falls in two broad categories: 1) Static, pre-recorded content such as Video on Demand (VoD) and 2) Live Linear streamed video for IP Television (IPTV), AR/VR, cloud gaming, social media live broadcasting and other rich content usages. Over the last 2 years, Live Linear traffic has enjoyed a meteoric growth as a percentage of all the Internet traffic and the trend is expected to continue through 2023. Add to this trend the explosion in the number of consumer devices connected to the Internet and content resolution moving towards 4K UHD; CDNs have been forced to evolve rapidly to keep up with the growing volume of data. With infrastructure budgets remaining relatively flat, CDN system architectures are being rethought to support higher video stream densities and increasingly dynamic and ephemeral content versus the traditional, static, limited stored-file model.

In order to keep up with the growth in volume, 100-200 Gigabit/s (Gbps) egress bandwidth is now common on CDN nodes at edge locations, often featuring fast PCIe*-based flash storage (NVMe) nearing 100TB per server storage densities. But that is not the only change. CDN nodes that are deployed at the edge tier of telco networks have unique requirements on the cache storage aspect of the architecture. Unlike the historical static file model, content cached or buffered at these nodes' changes in popularity often and increasingly is short lived. Because the content is becoming so transient, it imposes a high endurance requirement on the media that it's stored temporarily on. Also, a large memory capacity is needed to buffer each of the many individual streams a server needs to handle in these ad-hoc and on-demand scenarios.

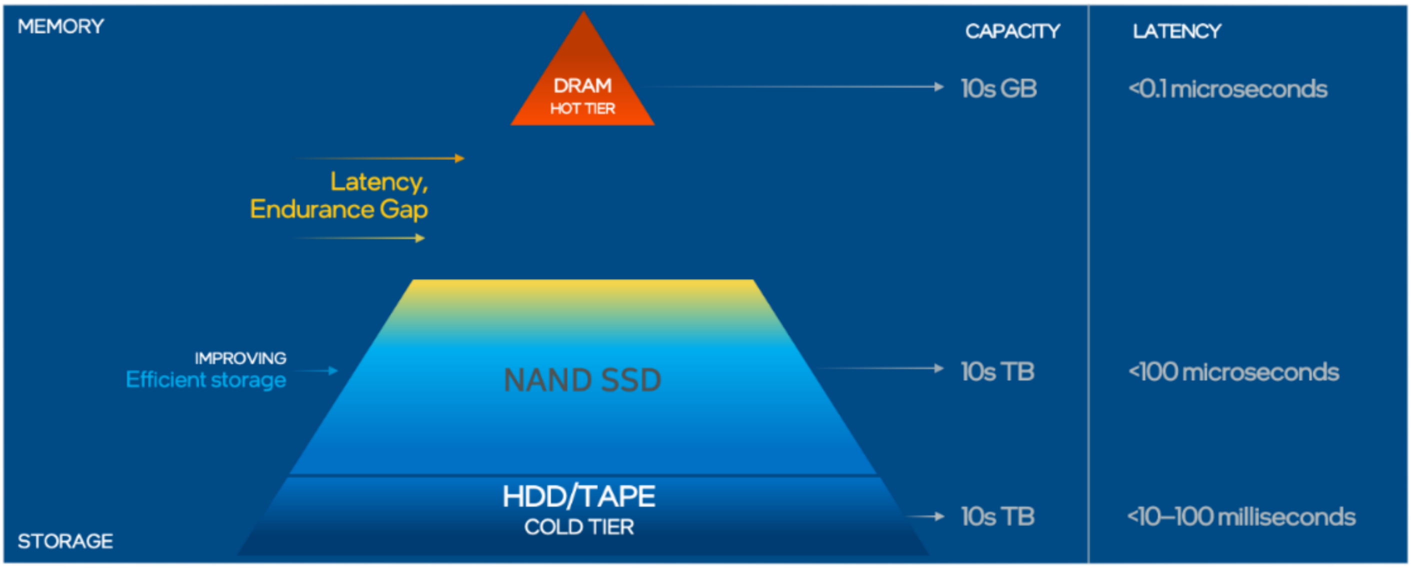

Figure 1. Traditional Memory/Storage Hierarchy

In the traditional storage media hierarchy shown in Figure 1, a NAND-based Solid-State Drive (SSD) offers endurance in the range of 3-5 Drive Writes Per Day (DWPD) for products commonly available on the market. This is likely acceptable for a static video-on-demand stored file model, but this endurance level is insufficient to support Live Linear video buffers that are typically overwritten tens to hundreds of times a day. This pushes the choice of media for Live Linear buffers to the next high endurance tier in the hierarchy, which is Dynamic Random Access Memory (DRAM). DRAM is volatile but high endurance media which makes it suitable for caching transient Live Linear content. DRAM also provides the lowest access latency when compared to the SSD/HDD tiers, making it even more attractive for caching Live Linear content for the CDN server.

Unfortunately, large capacity DRAM modules are quite expensive relative to SDD or HDD technologies on a price/capacity perspective-especially for the newest and largest capacity DRAM DIMMs can often account for up to 25% of the cost of a typical edge cache node. The high cost factor often results in limitation on the amount of DRAM memory capacity that is affordable. Since both the volume of streams and number of unique Live Linear channels supported is directly correlated to the cache capacity, smaller DRAM capacity also puts limit on number channels and streams that can be supported by a single CDN server. If there were a technology that helped bridge the latency and endurance gap in the memory/storage pyramid shown in Figure 1 and provided higher density media, better channel density could be supported per server.

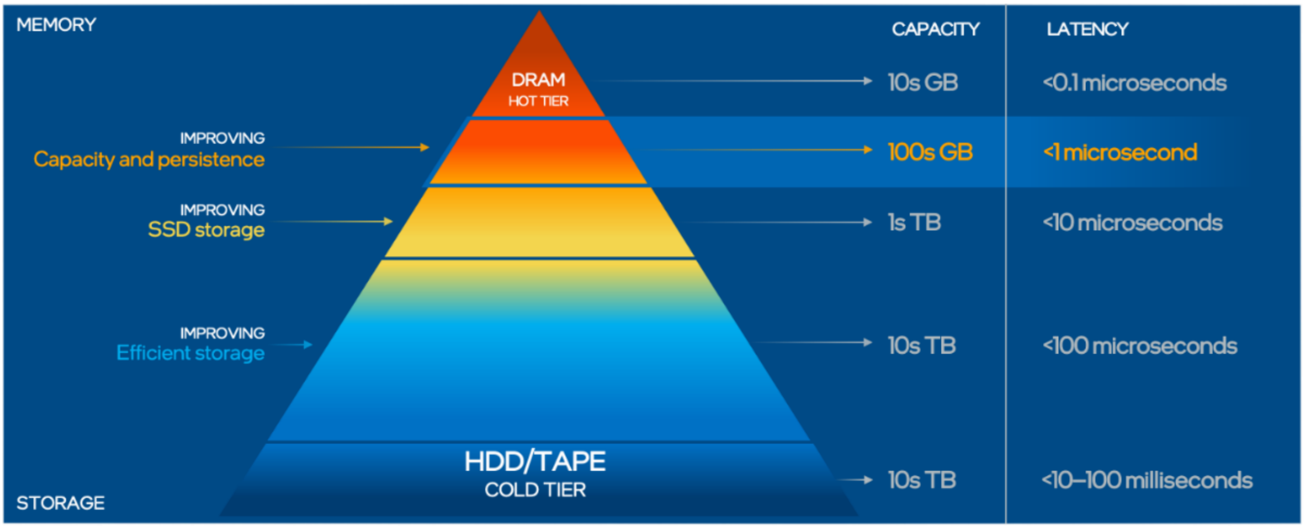

Figure 2. Evolved Memory/Storage Hierarchy Featuring the New Persistent Memory (PMem) Tier

Intel Optane PMem is one such technology (Figure 2). It offers a new tier between traditional DRAM and NAND SSD storage tiers in terms of performance, endurance, scalability, and cost. This new technology provides higher memory capacity (up to 4TB per socket) at a significantly lower cost than DRAM (approx. 20-40% savings at large capacities) thus enabling higher channel density for Live Linear CDNs.

Another relevant use case here is caching for extremely popular VoD content where a large subset of users request the same content. Much of this usage pattern has been driven by the ability to proliferate "viral" content easily by social media channels. Because viral content changes often, "hot VoD" usage can result in frequent overwrites, and thus can quickly wear out traditional NAND SSD storage media. This is the reason why CDN providers typically use DRAM to cache extremely popular content. But high cost of DRAM memory makes traditional "add more memory" approaches challenging.

What's the best way to optimize CDN server cost using Intel Optane PMem? Replace a DRAM-only tier with a mix of DRAM and Intel Optane PMem—a small amount of DRAM paired with Intel Optane PMem as bulk memory. With this approach, for a similar total memory cost, extra memory capacity is made available that can be used to cache additional video streams and Live-Linear channels on the same server. More cache capacity can also mean additional video and audio profiles cached for adaptive streaming, thus positively impacting user experience.

This paper demonstrates that for Live Linear streaming usages as well as hottest VoD content, Intel Optane PMem can deliver similar performance as DRAM in meeting Quality of Service (QoS) requirements, at a lower Total Cost of Ownership (TCO). Architectural recommendations and configurations developed through extensive and rigorous testing by AT&T* and Intel technologists are also provided here. These recommendations are aimed at taking the guesswork out of developing a highly scalable, highly efficient, next generation CDN.

2.1 CDN Edge Caching – Key Performance Indicators

Key Performance Indicators (KPIs) are the critical metrics used to define requirements for performance, latency, throughput, reliability, or other factors necessary to provide a suitable service to a set of customers. Following are some of the key KPIs we used when evaluating the value Intel Optane PMem brings to the Live Linear CDNs.

2.1.1 Average Response Time Per Object Size

Response time here is defined as the average time for the test client to repeatedly download an object of fixed size over a given single connection. Naturally, response time in this sense varies with object size. Response time includes the time from the client http request being made by the test client (last byte of request leaves the client) until the full object response has been received by the test client and assumes the network is not a bottleneck.

2.1.2 Peak Number of Simultaneous Client Connections

The peak number of simultaneous client connections is defined as the peak number of client-to-cache TCP (with TLS if HTTPS is being tested) connections that are simultaneously active, pulling objects from the cache, with no delay between one HTTP or HTTPS request and the next. This peak is assumed concurrent with the maximum obtainable throughput in Gbps, after which notable degradation of response times occur.

The client pulls as fast as it can over each given test connection, and the test set and network themselves are not a bottleneck in terms of achieving this. This peak condition must be maintained for a period of 5 minutes or longer. For our purposes, the number of simultaneous client connections will be gradually increased until a definite degradation in average response time is observed, as described earlier. Additionally, the connections will be modeled to mimic production loads as closely as possible, in terms of channel popularity and bit rate popularity within channels.

2.1.3 Maximum Throughput

Maximum throughput is defined as the total Gbps outbound from the cache server to the test set, summed over all simultaneous client connections at peak (defined above).

2.1.4 Byte Hit Ratio Percentage

The Byte Hit Ratio Percentage is defined for our purposes as the percentage of bytes served directly from cache (RAM, RAM disk, PMem, SSD or HDD), without recourse to pulling those bytes from the origin server (e.g., a different server in a central or regional data center), divided by the total number of bytes served, multiplied by 100. For these tests, the hit ratio is a controlled variable that will be adjusted to mimic what is observed in production. The desired target hit ratio for this testing will be 85%, representative of a best-case child Live Linear cache hit ratio.

Byte Hit Ratio Percentage = (Bytes served from cache server/ total bytes) * 100

2.1.5 Number of Simultaneous Channels Supported

Groups of connections will be arranged to mimic groups of clients accessing the same unique channel, with each channel offering a variety of bit rates and formats (e.g., HLS vs. DASH). In this case, we select a subset of bit rates based on representative popularity curve snapshots derived from a production environment. Specifically, all the clients (users) within a group will be arranged to access the same channel. The number of such groupings represents the number of simultaneous channels being accessed by clients. Each such grouping can be simulated by the clients in that group using the same manifest file or playlist to access a series of URLs that are unique from one another and unique from all other URLs of other channels and channel variants (bit rates and formats). Channels represent different Live Linear stations, and channel variants represent different quality levels or bit rates, and formats (e.g., HLS vs. DASH) available to clients accessing a specific channel.

2.2 Testing Memory Alternatives for CDN Caching

As mentioned earlier, CDNs are often implemented with DRAM as a write buffer for live content streaming (or hot VoD content hosting). Here we are exploring testing of memory alternatives with Intel Optane PMem to evaluate if the increase in the memory footprint closer to the CPU can provide additional channel density without impact to cache hit ratio, total sustainable throughput, and response time for the content.

The next section will illustrate the test setup for the CDN cache server, the server with different memory alternatives topology, the test methodology used and the traffic profile configuration to emulate client consumption for different live channels.

Figure 3. At Left: Intel® Optane™ PMem and At Right: Standard DDR4 DRAM DIMMs

DRAM memory is closest to a standard server systems' processor. Intel Optane PMem is populated in the same manner as the DRAM memory on the same DIMM form factor as illustrated in Figure 4 above. Intel Optane PMem is physically populated into (and electrically compatible with) the DIMM socket on the server board. The server board firmware supports initialization of DRAM and Intel Optane PMem devices and thus there are numerous population rules.

The below web link provides a typical Intel server board technical specification. Pages 55 and 56 in this document include a configuration guide for DRAM memory and PMem (Persistent Memory) population.

https://www.intel.com/content/dam/support/us/en/documents/server-products/server-boards/S2600WF_TPS.pdf

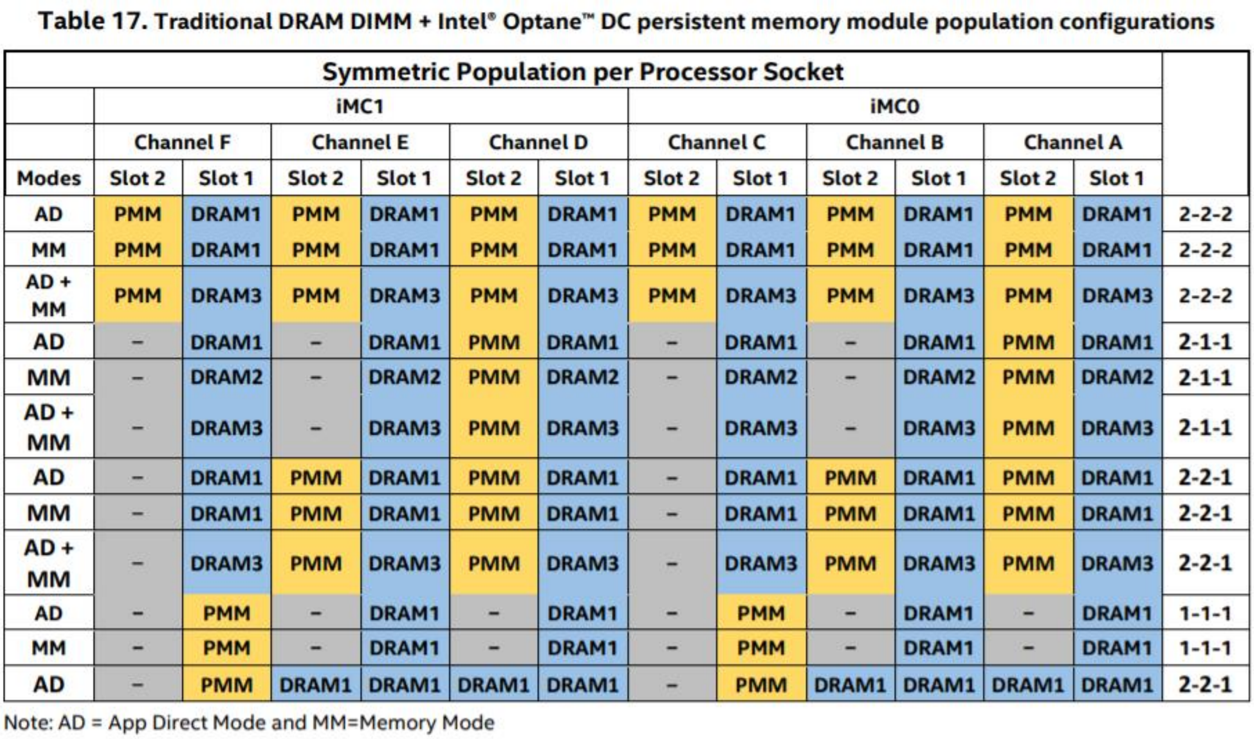

Table 17. Traditional DRAM DIMM + Intel® Optane™ DC persistent memory module population configurations

Symmetric Population per Processor Socket													
	iMC1						iMC0						
	Channel F		Channel E		Channel D		Channel C		Channel B		Channel A		
Modes	Slot 2	Slot 1	Slot 2	Slot 1	Slot 2	Slot 1	Slot 2	Slot 1	Slot 2	Slot 1	Slot 2	Slot 1	
AD	PMM	DRAM1	PMM	DRAM1	PMM	DRAM1	PMM	DRAM1	PMM	DRAM1	PMM	DRAM1	2-2-2
MM	PMM	DRAM1	PMM	DRAM1	PMM	DRAM1	PMM	DRAM1	PMM	DRAM1	PMM	DRAM1	2-2-2
AD + MM	PMM	DRAM3	PMM	DRAM3	PMM	DRAM3	PMM	DRAM3	PMM	DRAM3	PMM	DRAM3	2-2-2
AD	–	DRAM1	–	DRAM1	PMM	DRAM1	–	DRAM1	–	DRAM1	PMM	DRAM1	2-1-1
MM	–	DRAM2	–	DRAM2	PMM	DRAM2	–	DRAM2	–	DRAM2	PMM	DRAM2	2-1-1
AD + MM	–	DRAM3	–	DRAM3	PMM	DRAM3	–	DRAM3	–	DRAM3	PMM	DRAM3	2-1-1
AD	–	DRAM1	PMM	DRAM1	PMM	DRAM1	–	DRAM1	PMM	DRAM1	PMM	DRAM1	2-2-1
MM	–	DRAM1	PMM	DRAM1	PMM	DRAM1	–	DRAM1	PMM	DRAM1	PMM	DRAM1	2-2-1
AD + MM	–	DRAM3	PMM	DRAM3	PMM	DRAM3	–	DRAM3	PMM	DRAM3	PMM	DRAM3	2-2-1
AD	–	PMM	–	DRAM1	–	DRAM1	–	PMM	–	DRAM1	–	DRAM1	1-1-1
MM	–	PMM	–	DRAM1	–	DRAM1	–	PMM	–	DRAM1	–	DRAM1	1-1-1
AD	–	PMM	DRAM1	DRAM1	DRAM1	DRAM1	–	PMM	DRAM1	DRAM1	DRAM1	DRAM1	2-2-1

Note: AD = App Direct Mode and MM=Memory Mode

Figure 4. Traditional DRAM DIMM + Intel® Optane™ DC Persistent Memory Module Population Configurations for 2nd Gen Intel® Xeon® Scalable Processor

The Intel Optane PMem can be initialized as Memory Mode (MM) which is a volatile memory use case like DRAM; alternatively, it can also be initialized as App Direct Mode (AD) which support the non-volatile memory use case like any storage devices.

Intel Optane PMem DIMM is offered in sizes of 128GB, 256GB, and 512GB which are higher capacity than the typical commodity DRAM DIMM readily available in 2021. In this paper, 128GB Intel Optane PMem DIMMs are used in 2-2-1 and 2-2-2 modes (see **Error! Reference source not found.** f or details of those modes).

2.2.1 Test Setup

The test methodology, testing topologies and key performance metrics carried out in this paper are defined collaboratively between Intel and AT&T*.

The baseline and Device Under Test (DUT) CDN cache servers are setup with the following configurations:

- DRAM server - populated with DRAM memory only - this is the reference baseline configuration
- DUT1 server - populated with DRAM and Intel Optane PMem in a 2-2-1 topology
- DUT2 server - populated with DRAM and Intel Optane PMem in a 2-2-2 topology

The "baseline" DRAM configuration does not use Intel Optane PMem DIMMs. Performance of this baseline setup compared with DUT1 and DUT2 configurations which do use Intel Optane PMem DIMMs for CDN caching.

As Intel Optane PMem is a new technology, we are evaluating both 2-2-1 topology and 2-2-2 topology as there are cost implications (DUT2 is a higher cost and capacity configuration compared to DUT1).

Apart from the system memory topology, all processors, peripherals, system firmware, operating system and software application are the same for each configuration.

2.2.1.1 DRAM Server

Memory population as shown in Figure 5 has the memory topology of 12x 64GB DDR4 DIMMs operating at 2666 MT/s resulting in a total memory of 768GB per server.

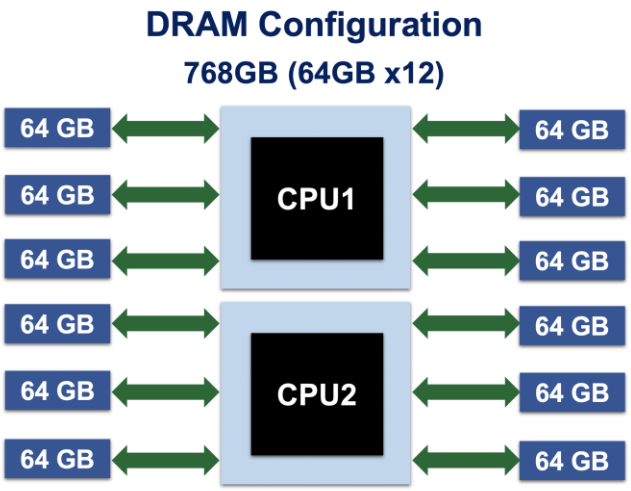

Figure 5. DRAM Server Memory Topology

2.2.1.2 DUT1 Server (2-2-1 Configuration)

Memory population for the DUT1 server as shown in Figure 5,**Error! Reference source not found.** has the memory topology of 12x 16GB DDR4 DIMMs and 8x 128GB Intel Optane PMem operating at 2666 MT/s. In this case a 2-2-1 memory topology is used. This configuration provides 192GB of DRAM memory and 1TB of Intel Optane PMem close to the processors

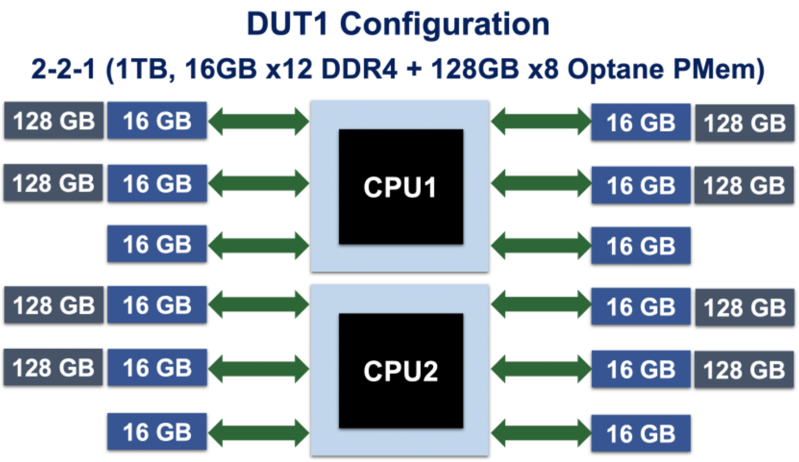

Figure 6. DUT1 Server Memory Topology

2.2.1.3 DUT2 Server (2-2-2 Configuration)

Memory population for the DUT2 server as shown in Figure 7 has the memory topology of 12x 16GB DDR4 DIMMs and 12x 128GB Intel Optane PMem operating at 2666 MT/s. In this case a full 2-2-2 memory topology is applied to achieve maximum memory bandwidth. This configuration provides 192GB of total DRAM memory and 1.5TB of persistent memory close to the processors. Note that other capacity points for the Intel Optane PMem modules are also available, but not used in this testing.

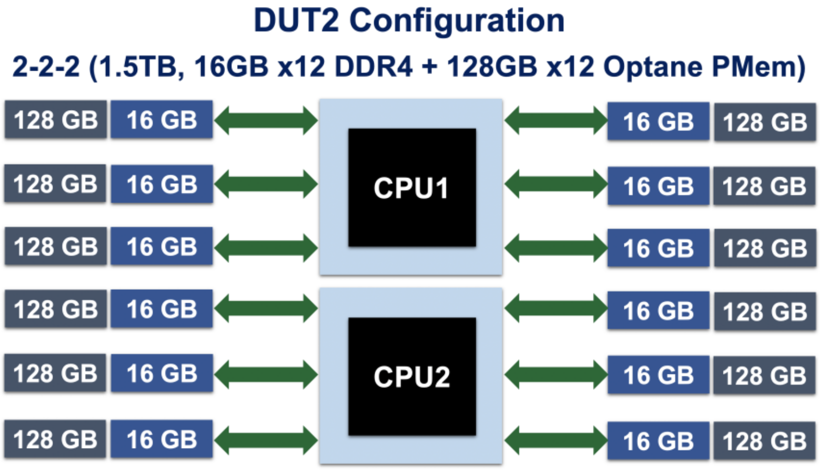

Figure 7. DUT2 Server Memory Topology

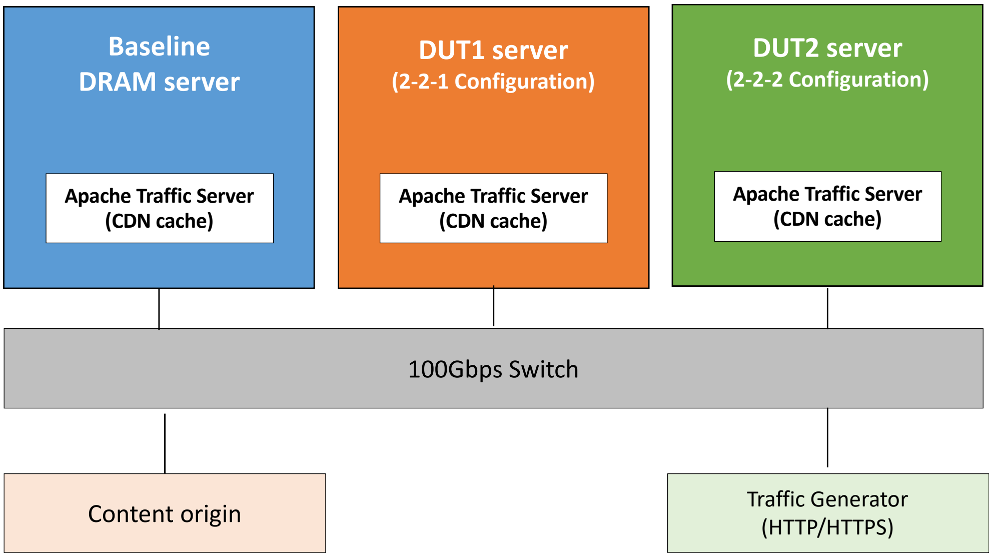
Baseline DRAM server
Apache Traffic Server (CDN cache)
DUT1 server (2-2-1 Configuration)
Apache Traffic Server (CDN cache)
DUT2 server (2-2-2 Configuration)
Apache Traffic Server (CDN cache)
100Gbps Switch
Content origin
Traffic Generator (HTTP/HTTPS)

Figure 8. Server Configuration for CDN Tests

Figure 8 shows the three test configurations with a CDN cache server application, specifically Apache Traffic Server (ATS) connected via a 100 Gbps Arista switch to a L4-L7 traffic generator; and an origin HTTP server also connected. The DRAM server, DUT1 server and DUT2 server were benchmarked separately.

During testing for each of the three test configurations, two virtual machines per socket are serving the requests from the packet generator under a 100 Gbps network connection. Due to testing environment network limitations, only one socket is tested. The other socket would be configured similarly in production.

Figure 9 and Figure 10 show the application test topology.

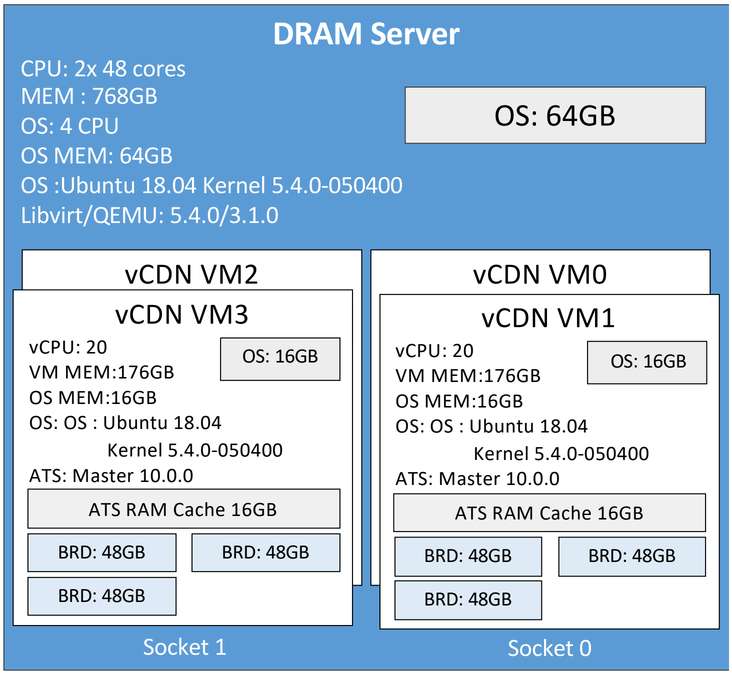

Figure 9. Test Topology: DRAM Server (Baseline)

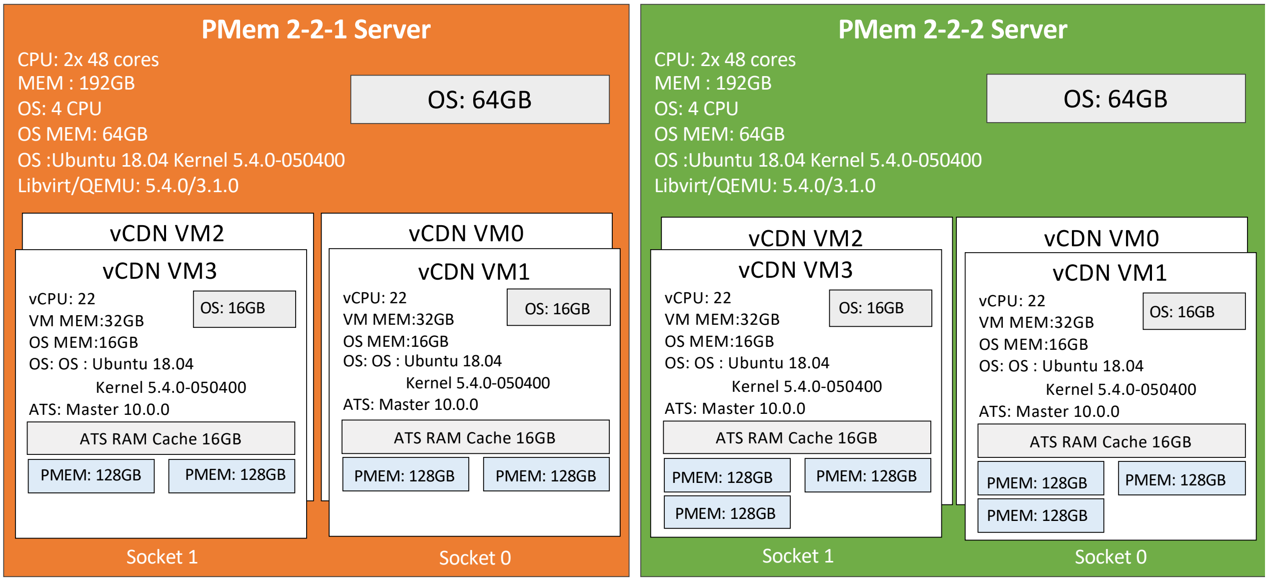

Figure 10. Test Topology: DUT1 and DUT2

Table 1. Host and Guest Resource Distribution

Configuration	Per Server Resources					Per Virtual Machine Resources					
	Number of Cores	Total DRAM Memory (GB)	Intel® Optane™ PMem (GB)	OS Kernel Mem (GB)	OS cores	VM cores	VM Total Memory (GB)	VM Kernel Mem (GB)	ATS Ram Cache (GB)	ATS Storage Cache	Number of Channels[4] 6 min buffer
DRAM	96	768	0	64	4	22	176	16	16	48GB x 3	100
PMem 2-2-1	96	192	1024	64	4	22	32	16	16	128GB x 2	175
PMem 2-2-2	96	192	1536	64	4	22	32	16	16	128GB x 3	250

For the given flavor, we configure the guest to use 500K file descriptors per VM, for which we observe that each VM is roughly capable of supporting up to approximately 2000 transactions/s at 200 connections/s, while generating approximately 42.5 Gbps of HTTPS throughput. For the given test case, Apache Traffic Server is the sole application configured to run within the guest. Moreover, we note that Intel Optane PMem can be configured in a mixed mode configuration with a certain percentage of the available persistent memory within the system set aside for memory mode and the remaining percentage set aside for App Direct Mode. Outside the scope of the present study, we reserve the exploration of mixed mode (volatile/non-volatile) configurations as future work.

For each of the test cases, there are three profiles used in the packet generator for each virtual machine:

- 100 Channels
- 175 Channels
- 250 Channels

The details of how many channels can be fit into each virtual machine will be described in the next section. Table 1 summarizes the resource allocated for each virtual machine for each DUT configuration. Each DUT server hosts 4 virtual machines, with 2 virtual machines aligning to each CPU socket.

Intel Optane PMem supports two modes of operation: Memory Mode and App Direct Mode. In Memory Mode, the DRAM acts as a cache while the Optane PMem provides large, volatile memory capacity; the functionality is implemented by the CPU memory controllers and is transparent to applications. Cache management operations are handled by the memory controller integrated in the Intel Xeon CPU. When data persistence is required, App Direct Mode is used. In App Direct Mode applications can access Optane PMem directly from user space as a byte-addressable memory, or as a block device (also known as the Storage over App Direct Mode).

Apache Traffic Server CDN application does not need to retain cache contents across reboots, so persistence is not necessary, however for ease of integration, we chose Storage over App Direct Mode. For the CDN video workload signature, this mode performs similarly to DRAM. For CDN stacks that have the ability to turn off disk based caching and support (main) memory as the primary tier, Memory Mode configuration is an equally viable option. It has been shown to yield similar results, but not covered in this paper.

2.2.2 Intel Optane PMem Technology Reliability

Thanks to Intel Optane PMem endurance levels, this technology is able to augment DRAM-based solutions offering increased number of streams and unique channels, increased sustainable network throughput and lower rack level TCO, all without compromising solution reliability.

Table 2. Endurance Data For 128GB Intel® Optane™ Persistent Memory Series 100 Modules

PRODUCT FAMILY FORM FACTOR	Intel® Optane™ Persistent Memory (PMem), Series 100
PMem SKU	128 GB
TECHNOLOGY	Intel® Optane™ technology
LIMITED WARRANTY	5 years
AFR	≤ 0.44
ENDURANCE 100% WRITES 15W 256B	292 PBW
ENDURANCE 100% WRITES 15W 64B	91 PBW

For detailed endurance info for Intel Optane PMem devices please refer to Intel Optane Persistent Memory 200 Series Brief[5].

2.2.3 Test Methodology

The testing included a Spirent C200 hardware appliance L4 through L7 traffic generator for performance benchmarking purposes. The traffic generator is configured in 4x 100 Gbps functional mode, with two out of the four 100 Gbps ports being used to load the current device under test. Traffic is load balanced across the two ports such that each port generates approximately 50 Gbps of HTTP throughput. Figure 11 represents the load profile for the number of simulated users per second that generate HTTP traffic over the course of a given test. The load profile is represented by a step function consisting of five distinct intervals. For each interval, the number of clients increases by approximately 50 simulated users per second. In addition, each interval within the load profile is sustained for a duration of 4 minutes, with each test case completing in the span of 20 minutes.

The aforementioned KPIs are measured under steady state conditions for each of the sustained intervals. All simulated users request channels according to a uniform distribution. In order to maintain the target cache, hit ratio, 85% of the simulated users request segments from the set of known, available channels. The remaining 15% of the simulated users request random segments outside of the set of available channels. The randomized segments are constructed by prepending a dynamically generated alphanumeric string to the URL.

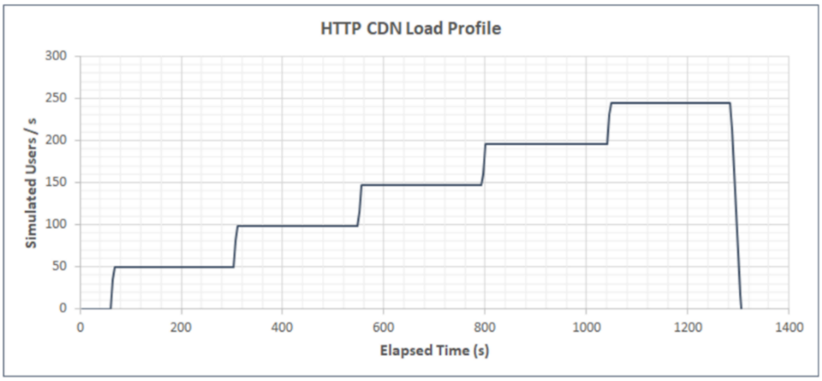

Figure 11. Load Profile Representing the Number of Simulated Users Per Second During the Test

Table 3 below represents the set of objects that a given simulated user may request, including DASH video, DASH audio, HLS video, as well as HLS audio content. Each object has an associated quality descriptor, variant, and size. The URL consists of the IP address, the channel number, the variant number, the media type, the object size, along with the segment number. In this case, <IP> represents the IP address of the target CDN cache, <channel-#> represents the channel number, and <segment-#> represents the segment number. Each media type includes up to 3 variants, with each representing a distinct bit rate available to the client for a given channel. A subset of the available variants is included within the table below. Specifically, the set represents the top ten most popular requested objects for a production environment and includes an additional 4K video object in order to account for potential future popular streaming profiles.

Collectively, all eleven objects constitute an overarching "segment", representing "all" available bit rates and formats clients may request within a given channel. The overall 'object size' of the given "segment" is represented by the sum of the individual object sizes, in this case totaling 25 MB. Based upon this 25 MB "segment", accounting for a 6 minute buffer size consisting of 60 segments where each segment is 6 seconds long, we determine that our DRAM server can store a total of approximately 200 channels across two VMs allocated to a single socket. We use the following set of equations in order to determine the total number of channels that can be supported by a given server topology on a per socket and on a per VM basis.

$$n_{total\ segments} = total\ disk\ cache\ size\ /\ segment\ size\ = total\ disk\ cache\ size\ /\ 25\ MB$$

$$n_{segments}\ /\ channel\ = 60$$

$$n_{channels} = n_{total\ segments}\ /\ (n_{segments}\ /\ channel) = n_{total\ segments}\ /\ 60$$

$$n_{channels}\ /\ socket\ = n_{channels}\ /\ 2$$

$$n_{channels}\ /\ VM\ = (n_{channels}\ /\ socket)\ /\ 2\ = n_{channels}\ /\ 4$$

In a similar manner, we determine that DUT1 and DUT2 can store a total of approximately 350 channels and 500 channels, respectively per socket under test for the server configuration.

Table 3. Simulated user HTTP request list including the URL pattern and corresponding object size

Media Type	Variant	URL	Object Size (KB)
HLS Audio	1	http://<ip>/<channel-#>/variant01/hlsha/405KB/<segment-#>	384
HLS Video	2	http://<ip>/<channel-#>/variant02/hlshv/4770KB/<segment-#>	4096
DASH Medium Quality Video	1	http://<ip>/<channel-#>/variant01/dshmv/443KB/<segment-#>	512
DASH Medium Quality Video	3	http://<ip>/<channel-#>/variant03/dshmv/1106KB/<segment-#>	1024
DASH Medium Quality Video	2	http://<ip>/<channel-#>/variant02/dshmv/680KB/<segment-#>	512

Media Type	Variant	URL	Object Size (KB)
DASH Medium Quality Audio	1	http://<ip>/<channel-#>/variant01/dshma/72KB/<segment-#>	72
DASH High Quality Audio	1	http://<ip>/<channel-#>/variant01/dshha/268KB/<segment-#>	256
DASH High Quality Video	3	http://<ip>/<channel-#>/variant03/dshhv/3536KB/<segment-#>	4096
DASH High Quality Video	2	http://<ip>/<channel-#>/variant02/dshhv/2418KB/<segment-#>	2048
DASH High Quality Video	1	http://<ip>/<channel-#>/variant01/dshhv/1633KB/<segment-#>	1024
4K Video (DASH High Quality)	1	http://<ip>/<channel-#>/variant01/4K/15360KB/<segment-#>	12000

2.2.4 Test Results

Results show the maximum number of simultaneous channels that can be achieved for a given cache hit ratio for each of the server configurations.

Figure 12 shows the number of channels versus sustainable throughput with target 85% hit ratio over time.

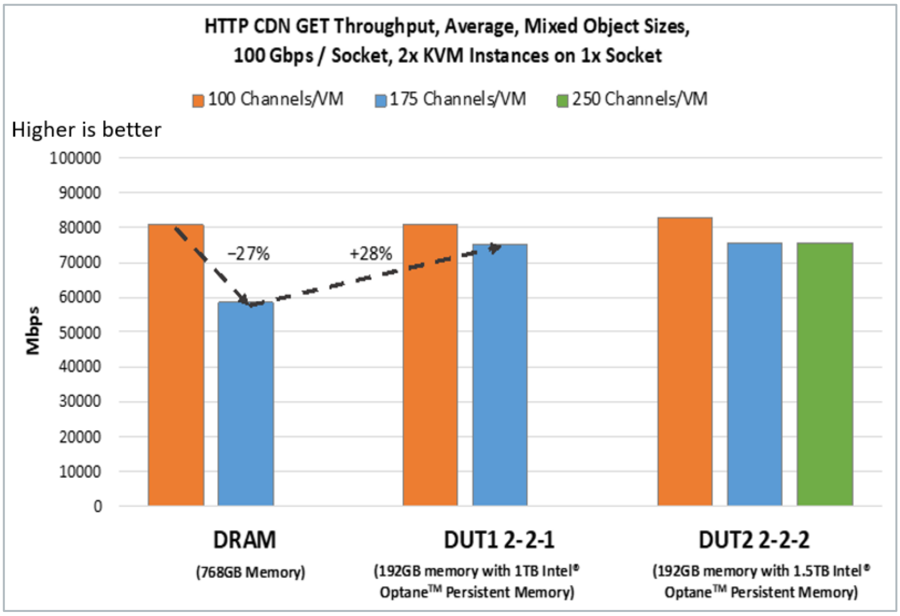

Figure 12. Sustainable Throughput vs. Number of Channels

The figure above plots the number of channels supportable vs max throughput for all three configurations: 1) DRAM-only, 2) DUT1 (PMem 2-2-1 configuration) and 3) DUT2 (PMem 2-2-2 configuration). In comparison to the DRAM-only case, both DUT1 with 1TB Intel Optane PMem and DUT2 with 1.5TB Intel Optane PMem display a throughput increase of approximately 28% while simultaneously being able to cache 175 channels per VM (or 350 channels per CPU socket).

Meanwhile DUT2 is the only configuration that is able to achieve and sustain the same throughput with 250 channels per VM, for a total of 500 channels for each CPU socket. This is a 250% increase in channel density from 200 channels per CPU socket supported by the DRAM-only configuration.

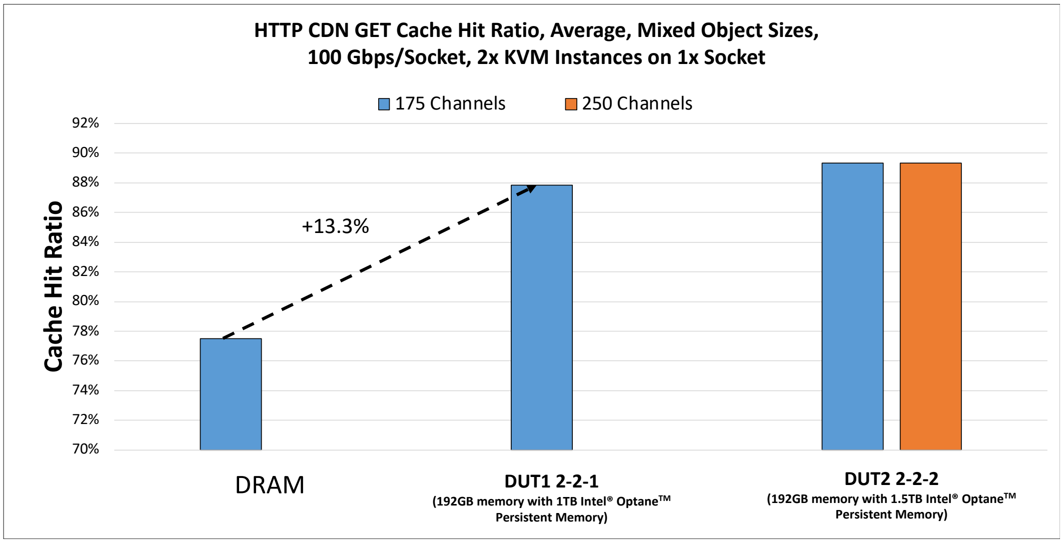

Figure 13. Number of Channels vs. Cache Hit Ratio

The above figure visualizes the cache hit ratio vs the number of channels supported by each of the configurations. In the case of the DRAM server, the average cache hit ratio falls below 85% for 175 simultaneous channels per VM, total of 350 channels per CPU socket. However, DUT1 is able to showcase an average cache hit ratio above 85% for streaming 175 simultaneous channels per VM while the DUT2 configuration can scale to 250 simultaneous channels per VM while maintaining the target cache hit ratio.

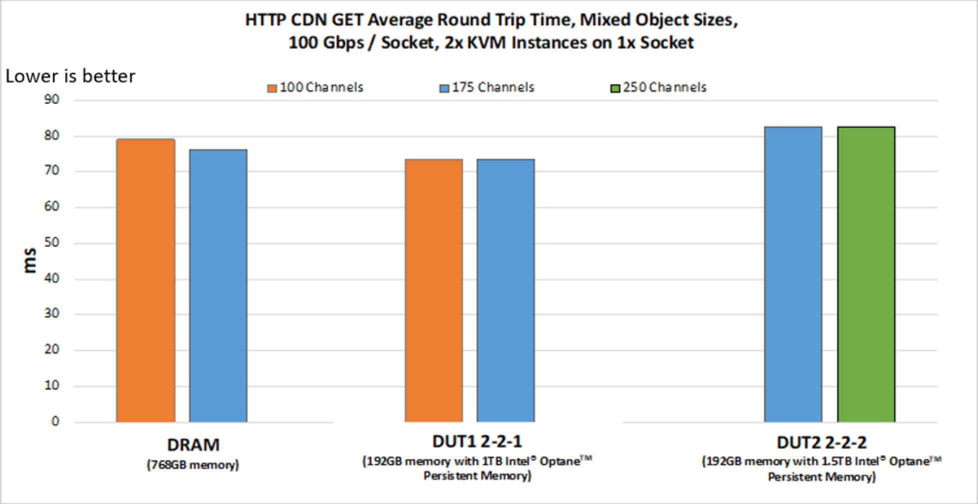

Figure 14. Number of Channels vs. Average Response Time

The above figure displays the round-trip time for 4MB object requests. DUT1 is able to show lower average response time for 175 and 250 simultaneous channels per VM compared to DRAM configuration. DUT2 shows less than 5% higher average response time performance than DRAM server for 175 channels per VM, total of 350 channels per socket. DUT2 is able to maintain the same response time for 250 channels versus 175 channels per VM.

2.3 Evolving vCDN Platform with 3rd Gen Intel Xeon Scalable Processor

Intel launched the 3rd Gen Intel Xeon Scalable Processor[6] supporting the next generation Intel Optane Persistent Memory 200 series[7]. This next generation provides higher memory bandwidth with 8 memory channels of DDR4 and supporting higher IO bandwidth via Peripheral Component Interconnect Express (PCIe*) generation 4, enabling multiple 200Gbps IO bandwidth with a single x16 add-in card. On top of that, we have the processor's core providing enhanced instructions sets i.e., VPMADD52, vectorized AES-NI, Vectorized Cumulative Multiplication VCLMUI and SHA-NI to accelerate RSA, block ciphers, hashing and TLS, SSL workloads. The section that follows highlights some of the expected incremental benefits enabled by this next-generation server and PMem platform.

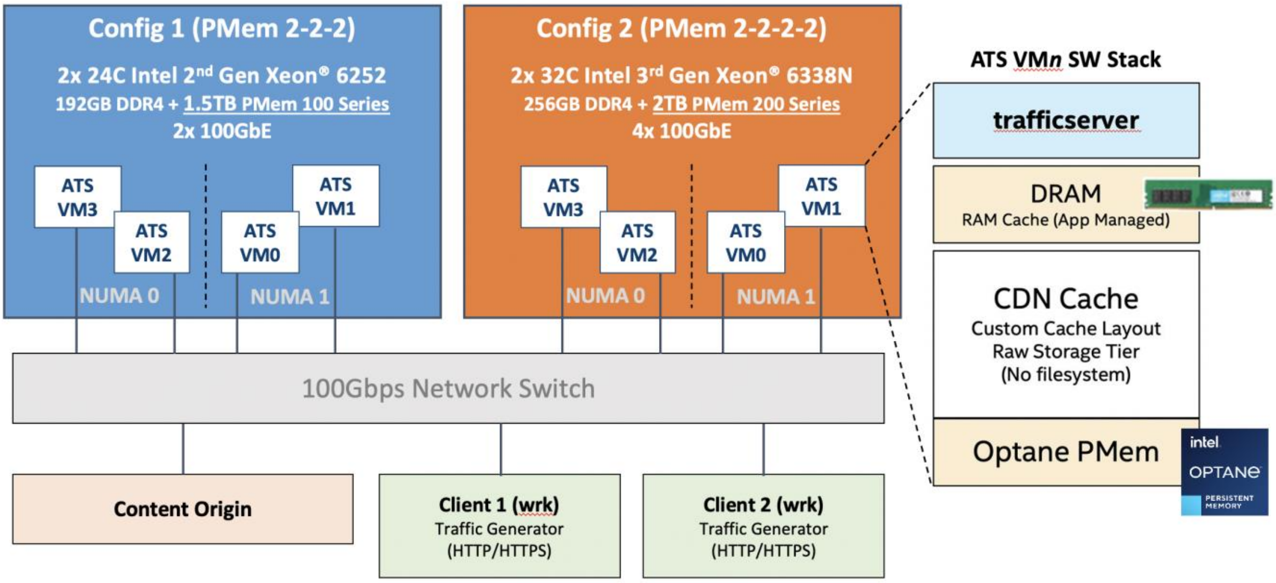

Figure 15. Test Setup for Comparing 2nd and 3rd Gen Intel® Xeon® Scalable Processor Servers

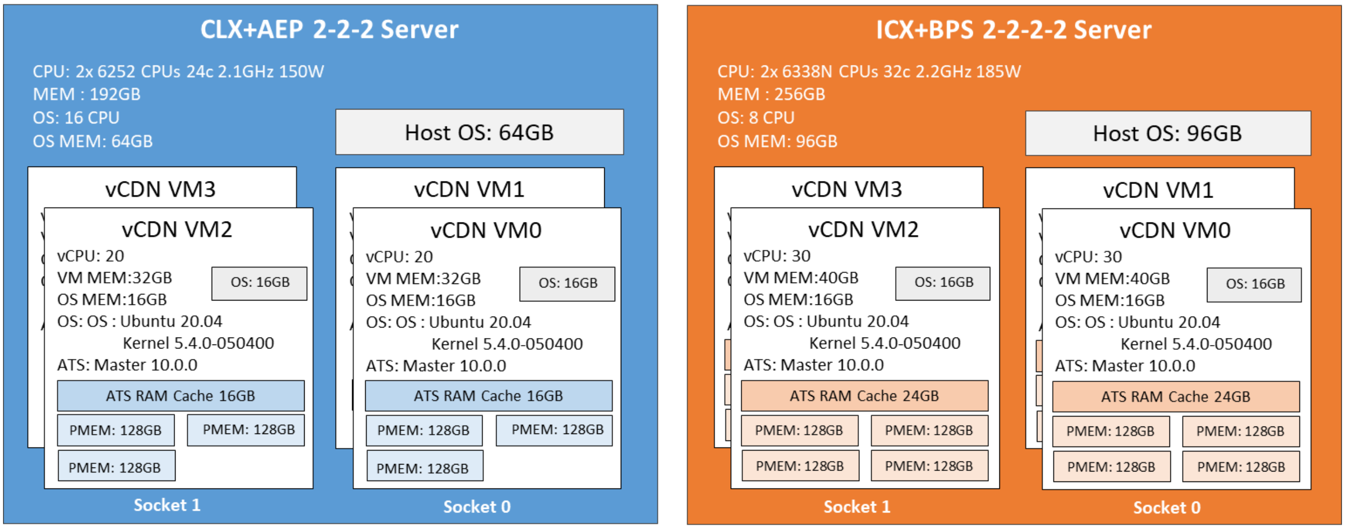

Figure 16. Resource Allocation for the 2nd and 3rd Gen Intel® Xeon® Scalable Processor Servers

Table 4. Summary of Resources for Comparison and Calculation for per Virtual Machine (VM) Channel Density

Configuration	Per Sever Resources					Per Virtual Machine Resources					
	Number of Threads	Total DRAM Memory (GB)	Intel® Optane™ PMem (GB)	OS Kernel Memory (GB)	OS Threads	VM Threads	VM Total Memory (GB)	VM Kernel Memory (GB)	ATS RAM Cache (GB)	ATS Storage Cache (PMem)	Number of Channels[8] Supported
2nd Gen Intel® Xeon Processor PMem 2-2-2	96	192	1536	64	16	20	32	16	16	128GB x 3	250
3rd Gen Intel® Xeon Processor PMem 2-2-2-2	128	256	2048	96	8	30	40	16	24	128GB x 4	325

2.3.1 Test Case

For this test case, we have ATS RAM Cache set to 'enabled'. As shown in the table above, we have calculated that the 3rd Generation Intel® Xeon® Scalable Processor 2-2-2-2^9 representing the Next Generation server is able to fit 100% of 325 channels per VM, or with two VMs per socket we can get 100% cache fit with 650 channels. In order to test for 90/10 cache, hit ratio, we are generating 715 channels/socket for this test case, with mixed object sizes up to segmented 4KB objects against the two servers. HTTPS requests are generated using multiple client nodes running WRK load generator.

2.3.2 Test Results

Test results show a 67% throughput improvement for the 3rd Gen Intel Xeon Scalable Processor platform with the Intel Optane Persistent Memory 200 series.

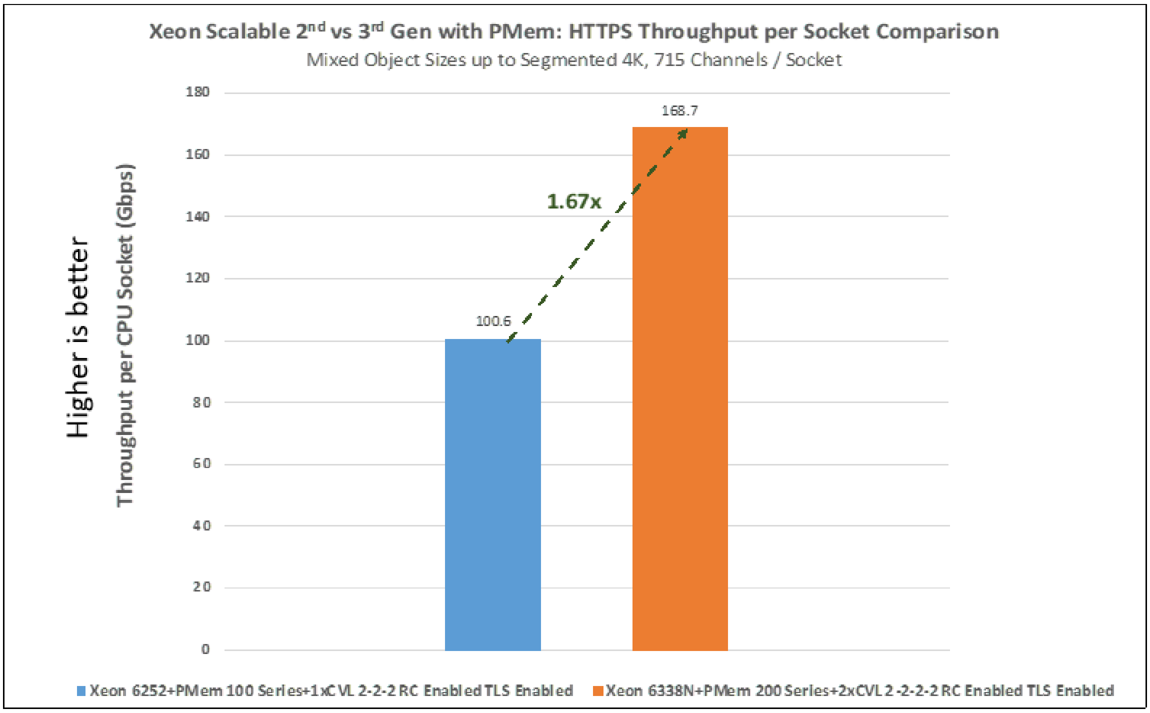

Figure 17. Throughput Improvement with 3rd Gen Xeon® over 2nd Gen Xeon®

Based on the results shown in

Figure 17, we observe an improvement in overall throughput of up to 67% when comparing a 2nd Gen Intel Xeon Scalable Processor platform with Intel Optane Persistent Memory 100 Series, against a 3rd Gen Intel Xeon Scalable Processor platform with Intel Optane Persistent Memory 200 Series. Specifically, we observe that the 3rd Gen Intel Xeon Scalable Processor platform is able to reach up to 168.7 Gbps of TLS throughput per socket with 2x Columbiaville NICs and a 2-2-2-2^9 Intel Optane PMem configuration.

2.4 Impact of Intel Optane PMem on CDN Service Delivery TCO

We have demonstrated how CDN Live Linear video service delivery benefits from incorporating Intel Optane PMem into the solution. In this section we discuss the impact of this technology to overall service TCO.

For the TCO analysis, focus is placed on the rack level economics to enable CDN service delivery. Communications service providers deploy services based on a preconfigured design, which is defined at the rack level and characterized to define a rack level SLA and mapping the required traffic against the validated configuration SLA.

For this exercise, we will assume the datacenter rack level power feeds are under 14kW, which limits the number of servers per rack, and will define 8 servers per rack to be conservative based on traditional CDN vendor deployment models. Individual server configuration corresponds to what is defined on Figure 9, Figure 10 and Figure 16.

Based on the sustainable throughput vs number of channels presented in Figure 12, and assuming 6 seconds video chunks, 6 minutes write buffer, 25MB buffer per segment, 60 segments per channel and a 5 year cost analysis, these are the TCO results at rack level:

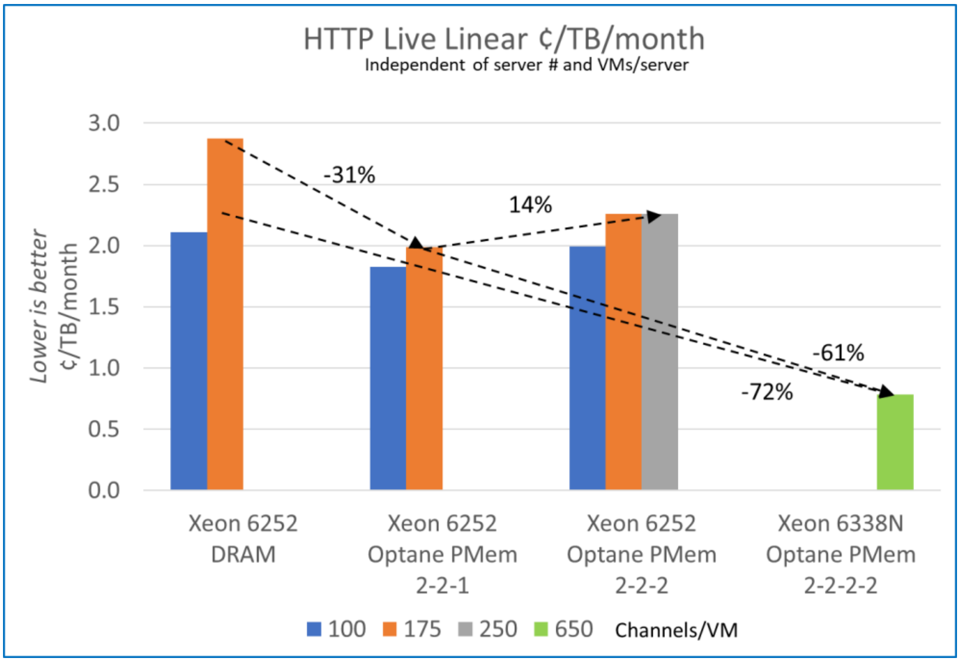

Figure 18. TCO Results at Rack Level

The key findings of this TCO analysis are:

- For the 2nd Gen Intel® Xeon® Scalable Processors based configurations analyzed, data shows that Intel® Optane™ Persistent Memory (PMem) delivers better KPIs than DRAM:
  - Intel Optane PMem 2-2-1 (1.0 TB density) config offers 30% lower ¢/TB/month than DRAM config
  - Intel Optane PMem 2-2-2 (1.5 TB density) config offers 40% increased channel/VM density at 14% higher ¢/TB/month vs. 2-2-1 (1 TB) maintaining a similar throughput level
- The 3rd Gen Intel® Xeon® Scalable Processors based configuration delivers significantly better TCO than all 2nd Gen Intel® Xeon® Scalable Processors based configurations previously analyzed, showing an up to 72% TCO reduction when compared with the DRAM based config.
- The 3rd Gen Intel® Xeon® Scalable Processors based configuration offers increased channel density, up to 700 channels per server, without having a TCO penalty, making this architecture especially suited for edge deployments where low numbers of CDN servers are required.
- The 3rd Gen Intel® Xeon® Scalable Processors based configuration, with its increased Intel Optane PMem Series 200 memory density, will be able to support streaming profiles for higher video resolutions (4K, 8K and beyond) that are becoming increasingly popular.

2.5 Conclusion

The study shows how Intel Optane PMem based on Intel 3D XPoint technology, offers near-DRAM performance and latency for CDN workloads, higher channel density and dramatically improved TCO. Intel Optane PMem creates a new memory tier that fills the wide gap between DRAM memory and SSD storage. It targets high capacity workloads that require near-DRAM performance, offering higher cache capacity for Live Linear content delivery.

AT&T* found that the increased channel capacity and reduced TCO offered by the 3rd Gen Intel Xeon Scalable Processors and Intel Optane PMem 200 Series brought a significant increase to their options for production environment deployment, offering significant value to their vCDN, that will help them to be more competitive in the marketplace.

2.6 References

1. Content Delivery Network, https://en.wikipedia.org/wiki/Content_delivery_network
2. Intricately Market Analysts on July 8, 2020
3. Cisco Visual Network Index (VNI) 2017-2022, https://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html
4. Calculation based on 6 seconds video, 6 minutes write buffer, 25MB buffer per segments, 60 segments per channel.
5. Intel Optane Persistent Memory 200 Series Brief, https://www.intel.com/content/www/us/en/products/docs/memory-storage/optane-persistent-memory/optane-persistent-memory-200-series-brief.html
6. Intel Corporation, “3rd Gen Intel Xeon Scalable Processors”, 2021 https://www.intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html
7. Intel Corporation, “Achieve Greater Insight from Your Data with Intel Optane Persistent Memory”, 2021

https://www.intel.com/content/www/us/en/products/docs/memory-storage/optane-persistent-memory/optane-persistent-memory-200-series-brief.html

8. Total Channels / VM at 100% cache hit; 60 Segments / Channel; 25 MB "segment" based on sample of popular media types from production including 4K video.
9. The terminology of a 2-2-2-2 topology for Intel Optane PMem 200 Series is used here strictly for the purposes of consistency with the use of the terminology for memory topologies with respect to Intel Optane PMem 100 Series. The officially accepted terminology for the corresponding memory topology with Intel Optane PMem 200 Series in this case would be 8+8.
10. Content Delivery Network, https://en.wikipedia.org/wiki/Content_delivery_network
11. 2020 State of The CDN Industry: Trends, Size, And Market Share, https://blog.intricately.com/2020-state-of-the-cdn-industry-trends-market-share-customer-size
12. Cisco Visual Network Index (VNI) 2017-2022, https://www.cisco.com/c/en/us/solutions/service-provider/visual-networking-index-vni/index.html
13. Intel® Corporation, "Intel® Optane™ Technology: Revolutionizing Memory and Storage," 2019, https://www.intel.com/content/www/us/en/architecture-and-technology/intel-optane-technology.html
14. Intel® Corporation, "Intel® Optane™ DC Persistent Memory: Big Breakthrough for your Biggest Data Challenges," 2019, https://www.intel.com/content/www/us/en/architecture-and-technology/optane-dc-persistent-memory
15. Intel® Corporation, "Intel® Optane™ DC Persistent Memory: The Challenge of Keeping Up with Data," 2019, https://www.intel.com/content/www/us/en/products/docs/memory-storage/optane-persistent-memory/optane-dc-persistent-memory-brief.html
16. Intel® Corporation, "3rd Gen Intel® Xeon® Scalable Processors", 2021, https://www.intel.com/content/www/us/en/products/docs/processors/xeon/3rd-gen-xeon-scalable-processors-brief.html
17. Intel® Corporation, "Achieve Greater Insight From Your Data with Intel® Optane™ Persistent Memory", 2021, https://www.intel.com/content/www/us/en/products/docs/memory-storage/optane-persistent-memory/optane-persistent-memory-200-series-brief.html

2.7 Test System Configurations

Config 1 - Test by Intel as of 3/16/2021.

1 node: 2x Intel Xeon Gold 6252 Processor, 24 core HT ON Turbo ON, Total Memory 1728GB (12 slots/16GB/2666MT/s, 12 slots/128GB Optane PMem/2666MT/s), BIOS SE5C620.86B.02.01.0010.010620200716 (ucode: 0x5003003), 1x Intel E810-C-QDA2, 1x Intel XXV710-DA2, Ubuntu* 20.04.1, kernel 5.4.0-050400-generic, gcc 9.3.0 compiler, OpenSSL 1.1.1i, ATS 10.0.

2 clients: 2x Intel Xeon Gold 6238R Processor, 28 core HT ON Turbo ON, Total Memory 768GB (12 slots/64GB/2933MT/s), BIOS: SE5C620.86B.02.01.0010.010620200716 (ucode: 0x5003003), Ubuntu* 20.04.1, kernel 5.4.0-050400-generic, wrk 4.0.2, 2x Intel Xeon Gold 6252N Processor, 24 core HT ON Turbo ON, Total Memory 384GB (12 slots/32GB/2666MT/s), BIOS: SE5C620.86B.0D.01.0374.013120191835 (ucode: 0x5003003), Ubuntu* 20.04.1, kernel 5.4.0-050400-generic, wrk 4.0.2.

Config 2 - Test by Intel as of 3/12/2021.

1 node: 2x Intel Xeon Gold 6338N Processor, 32 core HT ON Turbo ON, Total Memory 2048GB (16 slots/16GB/3200MT/s, 16 slots/128GB Optane PMem/3200MT/s), BIOS WLYDCRB1.SYS.0020.P86.2103050636 (ucode: 0x8d055260), 2x Intel E810-C-QDA2, 1x Intel XXV710-DA2, Ubuntu* 20.04.1, kernel 5.4.0-050400-generic, gcc 9.3.0 compiler, OpenSSL 1.1.1i, ATS 10.0.

2 clients: 2x Intel Xeon Gold 6238R Processor, 28 core HT ON Turbo ON, Total Memory 768GB (12 slots/64GB/2933MT/s), BIOS: SE5C620.86B.02.01.0010.010620200716 (ucode: 0x5003003), Ubuntu* 20.04.1, kernel 5.4.0-050400-generic, wrk 4.0.2, 2x Intel Xeon Gold 6252N Processor, 24 core HT ON Turbo ON, Total Memory 384GB (12 slots/32GB/2666MT/s), BIOS: SE5C620.86B.0D.01.0374.013120191835 (ucode: 0x5003003), Ubuntu* 20.04.1, kernel 5.4.0-050400-generic, wrk 4.0

2.8 TCO Cost Assumptions

TCO calculations by Intel as of Q1'2021.

Server unit cost calculated as average across the industry: config 1 DRAM only - $25053, config 1 Optane PMem 2-2-1 - $21032, config 1 Optane PMem 2-2-2 - $26868 and config 2 - $30068

8 servers in a 10KW rack, with the following server wall power requirement: config 1 DRAM only – 1.1KW, config 1 Optane PMem 2-2-1 – 1.1KW, config 1 Optane PMem 2-2-2 – 1.16KW and config 2 - 1.26KW

Average cost/KWh used: $0.12

Assumed PUE factor of 1.3

Assumed annual rack cost per RU of $75.76

Streaming compute node average to peak ratio of 50%

Notices & Disclaimers

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.