Intel Xeon Optimization for Efficient Media Workload Acceleration

Karan Puttannaiah
Senior Member, IEEE
Cloud Software Development Engineer, Intel Corporation
karan.puttannaiah@intel.com

Rajesh Poornachandran
Senior Principal Engineer,
Intel Corporation
rajesh.poornachandran@intel.com

***Abstract*— This paper discusses key methodologies involved in performing Workload Affinity characterization along with how to characterize the power-performance tradeoff across fine granular Intel Xeon CPU parameters across variety of industry popular Media use cases. Key results from the detailed study along with business acumen helped to define first ever Media workload optimized Intel Xeon CPU [1].**

Keywords—media transcode, video compression, performance, optimization, server processor, power management.

I. Introduction

The vast production and consumption of media (video, audio, etc.) combined with ever growing trends of video resolutions and quality necessitates efficient and effective strategies for optimization of media compression/decompression. Media transcoding, the process of converting multimedia between encoding formats, plays a crucial role in achieving efficient compression and faster transmission over the network. This process, however, can be computationally intensive, leading to bottlenecks that impede real-time applications as well as offline large-scale processing. This paper explores the nuances of high-performance media transcoding challenges and the opportunities for efficient processing on Intel® Xeon® processors. We investigate workload affinity towards various computational components and the crucial need for optimization strategies to ensure seamless media experience. This paper discusses how computational resources were tuned on Xeon processors, resulting in first ever media optimized Xeon SKUs [1,2].

II. Media Transcode Workload

A. Workload Description

Media transcoding workloads encompass a wide range of use cases, each presenting distinct performance requirements and optimization challenges [3,4]. In live streaming scenarios, real-time transcoding is paramount to deliver multiple versions of the video feed (differing in resolutions, bitrates) for diverse viewing devices and network conditions. Low latency and high frames per second (FPS) are crucial for a smooth viewer experience. Traditional broadcast workflows often necessitate transcoding content from production formats to distribution-ready formats with specific codec and bitrate requirements. Here, the transcoding density (number of streams processed simultaneously) and consistency of output quality are critical. Pre-recorded VOD (Video-on-Demand) assets are typically transcoded into multiple formats and bitrates to create an adaptive streaming experience. This prioritizes the efficient use of storage while maintaining visual quality across bitrate levels. Figure 1 shows a visual representation of transcode flow. The input is typically a high-quality video (can be raw uncompressed or slightly compressed while maintaining high quality), with high resolution and large file size. The decoder first converts into raw frames to be encoded by the encoder. Depending on the requirements on resolution, preset and bitrate, it is then compressed to an output video file suitable for transmission and consumption at the users' end. Note that the media transcode is critical part of a media pipeline involving important parts such as media delivery. The focus here is primarily on media transcode.

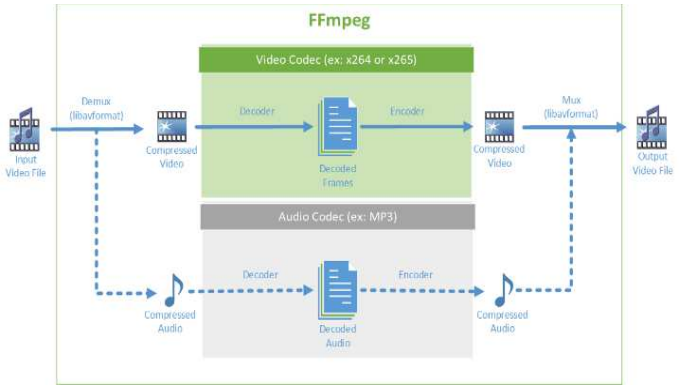

Figure 1: ffmpeg 1:1 Transcode Flow

There are several codec standards available in industry and academia. Among the popular ones are, 1) H.264 (AVC), a widely adopted, industry standard, offering widespread compatibility and mature encoding tools [5,6], 2) H.265 (HEVC) provides improved compression efficiency compared to H.264, enabling higher quality at lower bitrates [7], 3) SVT-AV1: open-source codec gaining traction, offering competitive compression efficiency especially optimized for multi-threaded scenarios [8]. Table 1 provides a summary of popular codecs and their characteristics that are in use today.

Codecs	x264	X265/SVT-HEVC	SVT-AV1

Encoding Standard	H.264	H.265	AV1
Other Names	Advanced Video Coding (AVC)	High Efficiency Video Coding (HEVC)	AOMedia Video 1 (AV1)
Output quality	Good	Better	Mostly same as HEVC
Computing power needed	Less	More	More
Hardware support	More	Less, but growing	Less, suited for many core servers

Table 1: Summary of Popular Codecs in industry [7]

Encoder/decoder settings based on user requirements define the balance between speed and quality. Faster presets reduce processing time, potentially trading off some quality. Slower presets engage more complex algorithms to improve compression but increase encoding time significantly. Higher bitrates typically result in better quality but larger files. Encoding to higher bitrates often requires more processing time. PSNR (Peak Signal-to-Noise Ratio) is an objective metric for image/video quality assessment, though it may not always perfectly reflect human perception of quality. Table 2 captures the summary of use cases based on the application.

Applications	Workloads/KPIs	ABR Ladder	Encoder Presets
Transcode Processing Time	VOD - 4K60 [1:4]	4K60 -> 4K60,1080p60, 720p60,360p60	placebo
Number of Bundles	Broadcast - 4K60 [1:4]	4K60 -> 4K60,1080p60, 720p60,360p60	slower
Number of Bundles	OTT Premium - 1080p60 [1:6]	1080p60 -> 1080p60 / 720p60 / 720p30 / 480p30 / 360p30 / 160p30	medium

Table 2: Resolutions and presets associated with popular industry applications/use-cases

B. Use cases

Table 3 highlights the various use cases capturing several codecs, resolutions and presets. This set of media transcode tests captures the wide range of media applications based on popular use cases in industry. The 'fast' preset is tailored for Over-the-Top (OTT) platforms, where the balance between encoding speed and video quality is optimized for streaming services that demand fast content delivery with decent quality. The 'medium' preset, on the other hand, is intended for broadcast use cases, focused on traditional television distribution channels. The 'veryslow' preset is the go-to choice for Video-On-Demand (VOD) services, where the highest quality is paramount and encoding time is less of a constraint, allowing for more complex compression algorithms to be employed, thereby reducing the data rate without sacrificing the viewing experience. Similarly, the preset in SVT-AV1 and SVT-HEVC are represented using encode modes 1 through 12, with preset 12 specifying fastest encode while preset 1 specifying highest quality.

Codec	Resolution	Preset	ASM
x264	1080p	fast	avx2
x264	1080p	medium	avx2
x264	1080p	veryslow	avx2
x265	1080p	medium	avx2
x265	4k	veryslow	avx2
x265	1080p	medium	avx512
svt_av1	1080p	12	avx2
svt_av1	1080p	8	avx2
svt_av1	1080p	5	avx2
svt_av1	4k	12	avx2
svt_av1	4k	8	avx2
svt_av1	1080p	12	avx512
svt_av1	1080p	8	avx512
svt_av1	1080p	5	avx512
svt_av1	4k	12	avx512
svt_av1	4k	8	avx512
svt_hevc	1080p	9	avx2
svt_hevc	1080p	5	avx2
svt_hevc	1080p	1	avx2
svt_hevc	4k	9	avx2
svt_hevc	1080p	5	avx512
svt_hevc	4k	9	avx512
svt_hevc	4k	5	avx512
svt_hevc	4k	1	avx512

Table 3: Ffmpeg Media Benchmark use cases

ffmpeg	**n4.4**
x264	5db6aa6cab1b146e07b60cc1736a01f2 1da01154
x265	3.1
SVT-AV1	v0.9.1
SVT-HEVC	1.5.1

Table 4: Codec Configurations

III. Experimental Setup and Results

This section discusses the role of key computational resources of the platform and their role in achieving efficient media transcode performance. These become the foundation for media optimized SKU definition that follow.

CPU	Intel Xeon 8592+ 128 logical core per socket | 2 sockets 350 Watts TDP per socket
Memory	16x 32GB, 5600 MT/s, DDR5
Storage	INTEL 1TB SSDPD21K015TAR
Operating System	CentOS Stream 9
Kernel	kernel-6.2
BIOS	109.D34
Microcode	0xa1000230

A. The Role of Core count and frequency

Media transcoding, the process of converting audio and video formats, is a computationally intensive task. The performance of these workloads directly correlates with the number of available processor cores and their operating frequency. Transcoding processes demonstrate impressive scaling efficiency exceeding 90% as core count and frequency increase. It is important to note that there are diminishing returns when focusing on a single instance of a transcoding job. Beyond a certain threshold of cores and frequency, performance gains become negligible. However, most real-world use cases involve customers running multiple transcoding streams simultaneously. In these scenarios, where CPU utilization is consistently high, the high scaling efficiency ensures optimal performance. By default, codecs use thread-count based on the (logical) core count available. In many core processors such as Xeon, this can lead to over-subscription, causing thread synchronization issues. Fortunately, the users are able to set the thread count for each transcode instances. For example, "-threads" in x264 and "-lp" in SVT-AV1. Figure 2 shows Core count scaling shows linear scaling with ~100% scaling efficiency of three prominent use cases 1) x264 1080p medium avx2, 2) x265 1080p medium avx2, 3) svt_av1 1080p em12 avx2. Here, other factors such as frequency of the cores and frequency of the SoC interconnect fabric are fixed to understand the impact of core-count.

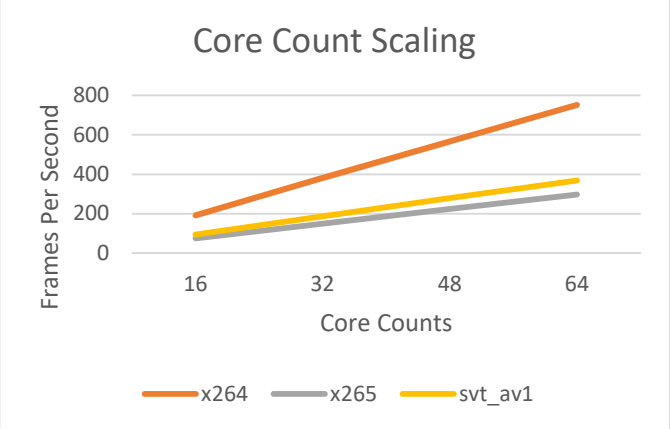

Figure 2: Core count scaling shows linear scaling with ~100% scaling efficiency.

Figure 3 shows core frequency scaling shows linear scaling with >90% scaling efficiency illustrating the affinity of the workload performance to core-frequency.

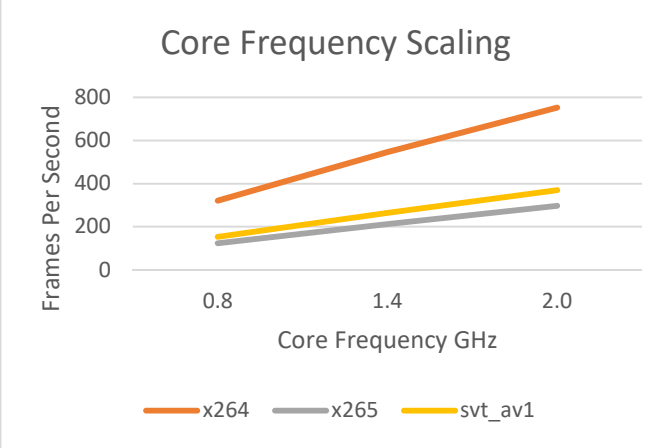

Figure 3: Core frequency scaling shows linear scaling with >90% scaling efficiency.

B. Opportunities with Soc Interconnect Fabric resources

For high-quality transcode (e.g., veryslow, encmode=1), major computational burden is on the cores. The SoC interconnect fabric frequency can be as low as All-core turbo frequency of SoC interconnect fabric for such cases. For lower-quality higher-speed transcode (e.g., fast, encmode=9), SoC interconnect fabric frequency moderately impacts the performance until base frequency of the interconnect fabric, but has negligible effect beyond that frequency. Power efficiency can potentially be improved by carefully balancing these frequencies, especially in high-quality transcoding scenarios which are very much core/compute bound.

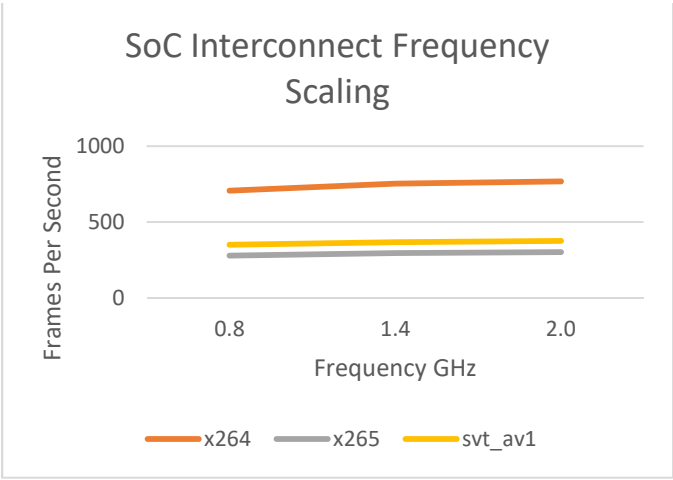

Figure 4: SoC interconnect fabric frequency scaling shows poor scaling up to about CFC P1 frequency with negligible difference beyond.

The performance of media transcoding workloads exhibits a strong dependence on core count and core frequency within the processing system. This is due to the computationally intensive nature of decoding and encoding video data. In contrast, the effect of memory speed is less pronounced, provided the memory operates within a reasonable range. Once memory bandwidth exceeds a baseline requirement to supply and store data for the transcoding process, further increases in memory speed yield minimal gains in overall transcoding performance. In our observations, reducing the speed by about 20% from maximum of 4800MT/s in Xeon 4th generation down to 4400MT/s resulted in negligible performance drop. The upside of reducing memory speed is a small amount (typically a few watts) CPU power savings, which in-turn can be consumed by the cores to increase their frequencies.

C. *Multi-socket scenario*

Customers often prefer solutions that minimize the need for software modifications, making NUMA (Non-Uniform Memory Access) pinning an attractive optimization technique for media workloads. Despite inherent multithreading, media workloads often face scalability challenges across CPU sockets due to suboptimal OS scheduling. NUMA pinning addresses this by aligning threads and memory allocation within NUMA nodes, resulting in improved performance. Furthermore, by localizing memory access after NUMA pinning, the demand for inter-socket UPI (Ultra Path Interconnect) bandwidth decreases. This presents an opportunity to optimize power consumption by reducing the number of UPI links and their operating frequency, potentially allowing power savings to be redirected towards boosting core frequency for further performance gains.

D. *Media Cdyn*

In determining the specifications of a CPU, the Cdyn of the workload that the CPU is intended for plays a key role in terms of defining the frequencies, i.e. Base frequency of SSE instructions (P1_SSE), All-core turbo frequency (P0N), base frequency of AVX2 and AVX512 instructions (P1_AVX2/512) to meet platform power, voltage and current constraints. As described in the previous sections, understanding the affinity of a given workload to these fine-granular entities, helps us to tailor a Xeon SKU definition optimizing for the workload stickiness to the socket. Additionally, this helps us to smartly balance the limited power/current budget to ration to portions of silicon (i.e. Core for Media WL given the workload scaling affinity). From our post-silicon measurements on the active Cdyn across Cores, SoC Interconnect Fabric, Memory, and IO we observed that Media Cdyn is more affinitized to the number of cores and frequency of each core in ordinal priority. Given the need for compute & energy efficient processors, we took this as an opportunity to define Media Workload customized processor having higher power budget allocation for higher number of cores with higher operational frequencies for Media Workload segments in Comms Service Providers (CoSP) e.g. Comcast and in Next Wave Cloud Service Providers (NW CSPs) towards segment optimized SKU definition. This resulted in the Icelake processor 8352M [1], and Sapphire Rapids processor 6438M [2].

IV. Process followed for Xeon Media SKU Definition

In previous sections we discussed the engineering methodology of how we understand the affinity of a workload to Xeon SKU definition knobs and how we characterize Cdyn to define optimal SKU frequency points. This is necessary but not sufficient. For a successful SKU definition, we worked very closely with the business segment leads to identify and optimize the most potential SKU of choice for this segment that fits the TCO needs of the segment customers. Based on these inputs, we worked on different What-If permutations in an iterative fashion to get the optimal specs for the Media Optimized SKU. This included trade-offs of UPI count, DDR speed and Base frequency of interconnect fabric to get a better base frequency of Cores that worked well for the segment. Given Xeon is a general purpose compute processor that can support wide variety of market segments, for scalability we took advantage of Xeon ISS [10] Intel Speed Select Profiles to fall back to traditional knobs. Figure 5 show the overall high-level process followed as an example towards definition of Media Optimized SKU.

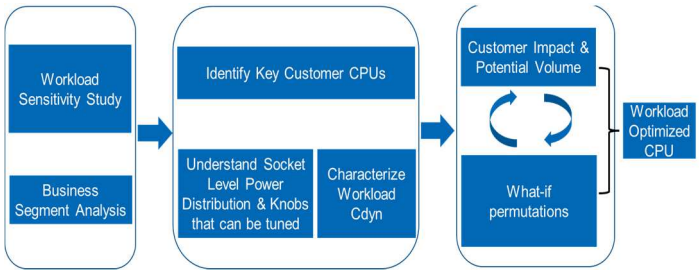

Figure 5: Workflow followed for Media SKU definition

V. CONCLUSION & FUTURE WORK

This paper discussed key methodologies involved performing Workload Affinity characterization along with how to characterize the power-performance tradeoff across fine granular Xeon SKU definition knobs across variety of industry popular Media use cases. Key results from the detailed study along with business acumen helped us to define first ever Media workload optimized Xeon CPU. As a future work, we intend to continue the differentiation to future Xeon roadmap. Additionally, we intend to expand to Xeon plus Accelerators towards a system level approach with/without Application of AI in Media Transcode and Media Analytics use cases.

ACKNOWLEDGMENT

The authors would like to acknowledge the collaborative efforts of the following people (in alphabetical order) Francisco Gonzalez Monge, Jason Crop, Juan Gonzalez Gonzalez, Kinchit Desai, Kiran Atmakuri, Lynn Comp, Marlon Cardenas, Mathew Garrison, Naga Gurumoorthy, Nagesh Puppala, Raju Yasala, Vasavee Vijayaraghavan, Vikram Krishnamachary and many more.

REFERENCES

[1] Intel® Xeon® Platinum 8352M Processor, url: https://www.intel.com/content/www/us/en/products/sku/217215/intel-xeon-platinum-8352m-processor-48m-cache-2-30-ghz/specifications.html

[2] Intel® Xeon® Gold 6438M Processor, url: https://ark.intel.com/content/www/us/en/ark/products/232398/intel-xeon-gold-6438m-processor-60m-cache-2-20-ghz.html

[3] Q. Hu, X. Zhang, Z. Gao, and J. Sun, "Analysis and optimization of x265 encoder," in 2014 IEEE Visual Communications and Image Processing Conference, IEEE.

[4] DCPerf: An open source benchmark suite for hyperscale compute applications, url: https://github.com/facebookresearch/DCPerf.

[5] L. Merritt, and R. Vanam, "x264: A high performance H.264/AVC encoder," 2006, url:http://neuron2. net/library/avc/overview_x264_v8_5. pdf.

[6] S. Hashemizadehnaeini, "Transcoding H. 264 Video via FFMPEG encoder," PhD dissertation, Politecnico Di Milano, 2014.

[7] Coding efficiency comparison of AV1/VP9, H.265/MPEG-HEVC, and H.264/MPEG-AVC encoders, Picture Coding Symposium (PCS), 2016

[8] F. Kossentini, H. Guermazi, N. Mahdi, C. Nouira, A. Naghdinezhad, H. Tmar, O. Khlif, P. Worth, and F. B. Amara, The SVT-AV1 encoder: overview, features, and speed-quality tradeoffs. Applications of Digital Image Processing XLIII, vol. 11510, pp. 469–490, 2020.

[9] SPEC CPU® 2017, Standard Performance Evaluation Corporation, url: https://www.spec.org/cpu2017/

[10] Managing Intel® Speed Select Technology (Intel® SST) Performance Profiles for Enhanced Core Configuration, url: https://www.intel.com/content/www/us/en/support/articles/000095518/processors.html