Warehouse-scale Video Acceleration: Co-design and Deployment in the Wild

Extended Abstract

Parthasarathy Ranganathan, Daniel Stodolsky, Jeff Calow, Jeremy Dorfman, Marisabel Guevara, Clinton Wills Smullen IV, Raghu Balasubramanian, Sandeep Bhatia, Prakash Chauhan, Anna Cheung, In Suk Chong, Niranjani Dasharathi, Jia Feng, Brian Fosco, Samuel Foss, Ben Gelb, Sara J. Gwin, Yoshiaki Hase, Da-ke He, C. Richard Ho, Roy W. Huffman Jr., Elisha Indupalli, Indira Jayaram, Poonacha Kongetira, Aki Kuusela, Cho Mon Kyaw, Aaron Laursen, Yuan Li, Fong Lou, Kyle A. Lucke, JP Maaninen, Ramon Macias, Maire Mahony, David Alexander Munday, Srikanth Muroor, Narayana Penukonda, Eric Perkins-Argueta, Devin Persaud, Alex Ramirez, Ville-Mikko Rautio, Yolanda Ripley, Amir Salek, Sathish Sekar, Sergey N. Sokolov, Rob Springer, Don Stark, Mercedes Tan, Mark S. Wachsler, Andrew C. Walton, David A. Wickeraad, Alvin Wijaya, and Hon Kwan Wu

Google Inc.

1. Motivation

Video processing plays a pivotal role in several key workloads in large cloud data centers. Video sharing and video streaming workloads (e.g., YouTube, Netflix, Facebook, etc) are well known as contributing to the dominant portion of internet traffic [4]. COVID-19 recently further amplified the importance of efficient video processing: beyond the surge in adoption of video-conferencing workloads (e.g., Zoom, Meet, Teams, etc) for social connection and education, medical practitioners in the front line of the pandemic relied on video platforms for life-saving procedures. Trends towards live-streaming and higher video quality/resolution (e.g., 4K/8K videos) and more immersive video (e.g., 360 degree views) further increase the computational demand for video processing. Several important emerging workloads – virtual/augmented reality, cloud gaming, cameras in IoT devices – are also very video-centric, further increasing the future importance of video processing. While the computational demand for video processing is exploding, improvements from Moore's Law are stalling with traditional approaches [1]. Future growth in this important area is not sustainable without adopting domain-specific hardware accelerators.

2. Limitations of the State of the Art

Current state-of-the-art data center video processing platforms are implemented in software running on general-purpose servers carefully optimized for scale [3]. As discussed above, this approach is unlikely to be sustainable both in supporting existing growth and costs, but also in enabling new video workloads and capabilities.

Prior work on hardware video acceleration has focused primarily on video decoding and encoding on consumer/end-user systems (mobile devices, desktops, televisions), with few specifications [2] and products available that target data center infrastructure. Introducing video transcoding accelerators at such warehouse-scale [3] is a challenging endeavor. In addition to the high-quality, high-availability, and high-throughput requirements of cloud deployments, the accelerator design needs to address the complexity of server-side video transcoding (plethora of formats and complex algorithmic and modality tradeoffs), deployment at scale (workload diversity and video serving patterns), and co-design with large-scale distributed systems. These are discussed at length in Section 2 of the main paper.

3. Key Insights

Our key insight is that warehouse-scale video *acceleration* has to be designed with the same "data-center-as-a-computer" approach that has been beneficial for warehouse-scale general compute ("data-center-as-an-accelerator"). Such an approach simultaneously imposes and relaxes constraints: the accelerator has to be co-designed with the warehouse-scale computing (WSC) stack and meet WSC constraints on power, costs, reliability, and diversity and churn in workloads, but at the same time, individual hardware blocks can be designed at the aggregate (cluster-level operation) and assuming software fall-back. This can lead to simpler, more efficient, and more adaptive designs.

As one specific application of this insight, conventional ASIC design often assumes low density; data center ASICs can operate without this constraint, leading to more amortized designs and less stranding. Additionally, transcoding from a source video to multiple resolutions and formats in parallel is a fundamental and frequent operation. Support for inline scaling and shared memory provides valuable latency and throughput improvements, especially for high resolution live streaming. Similarly, software deployments are highly disruptive in data centers (i.e. kernel and firmware releases). This suggests that hardware designs should be optimized for userspace software control. Likewise, given the distributed nature of data center computing, and distributions of video processing use-cases, systems should also be optimized for efficient work scheduling

to accelerators. Together, these principles suggest that video accelerators that target the data center need to be co-designed across the hardware block, ASIC, board, and cluster levels. Most prior work on accelerators do not take such a holistic approach.

Section 3 in the paper presents other key common design insights across our system, and lower-level insights specific to individual optimizations.

4. Main Artifacts

We discuss the design and deployment of a warehouse-scale video acceleration system built on a new hardware accelerator building block – a *video coding unit (VCU)* – for data center video transcoding at scale, supporting multiple video-centric workloads (video sharing, photos/video archival, live streaming, cloud gaming) with strict quality, throughput, latency, and cost requirements. We discuss our *hardware/software co-design methodology* for fungibility, work scheduling, rapid design and deployment, and specific *solutions around hardware and software design* (VCU system architecture, online multi-dimensional bin packing scheduler). We also present rich data from production workloads illustrating real-world deployment considerations.

5. Key Results and Contributions

- We present the design of our warehouse-scale video acceleration system including the design of a new hardware accelerator building block – a *video coding unit (VCU)* – and a system architecture that balances individual codec hardware blocks in VCUs, VCUs in boards and systems, all the way to individual systems in clusters and geographically-distributed data centers.
- We discuss our insights for system design when co-designing accelerators with large-scale distributed systems at data center scale, and how it translates to specific design optimizations: density- and stranding-optimized hardware system balance, fungibility abstractions that enable easy adaptation to change, and co-design with cluster scheduling for improved utilization and reliability.
- We present insights from our deployment at scale including fleetwide results from longitudinal studies across tens of thousands of servers. For offline two-pass encoding, the VCU systems provide 8x-20x higher encoding throughput than software encoding, for H.264 and VP9, respectively. Our accelerator system has an order of magnitude performance-per-cost improvement over our prior well-tuned baseline system with state-of-the-art CPUs while still meeting strict quality requirements. We also present results demonstrating how our careful co-design allows for real-world failure management and agility to changing requirements.
- Finally, we also discuss how our accelerator, beyond improving efficiency, enables new live/video-on-demand workloads, and increased bandwidth and storage compression, unlocking new workloads like cloud gaming – all new capabilities that were otherwise not possible.

6. Why ASPLOS

Our work is at the intersection of hardware (ASIC development, system architecture), distributed systems (data center scheduling, video processing platform), and video processing research (transcoding algorithms). We believe this is a great fit with the multidisciplinary focus of ASPLOS. Additionally, as discussed above, the key thesis of our work is around co-design across the WSC stack, and many of our results (Section 4 of the paper) are excellent showcases of the synergy between architecture and systems (e.g., system balance, failure management, responsiveness to workload diversity and churn).

7. Citation for Most Influential Paper Award

This paper is the first to discuss the design and deployment of video acceleration at scale in warehouse-scale data centers. The order-of-magnitude improvement in video transcoding from the VCU accelerator enabled new capabilities and innovations in video applications (low-latency cloud gaming, immersive high-quality video, etc). Most importantly, the insights in this paper on co-designing accelerators with large-scale distributed systems at data center scale spurred subsequent accelerator projects to also take a pragmatic full-system data-center-as-an-accelerator view of design.

References

[1] John Hennessy and David Patterson. A new golden age for computer architecture: Domain-specific hardware/software co-design, enhanced security, open instruction sets, and agile chip development. In *2018 ACM/IEEE 45th Annual International Symposium on Computer Architecture (ISCA)*, pages 27–29, 2018.

[2] Kevin Lee and Vijay Rao. Accelerating facebook's infrastructure with application-specific hardware. https://engineering.fb.com/data-center-engineering/accelerating-infrastructure/, Mar 2019. (Accessed on 2020-08-20).

[3] Urs Hölzle Luiz André Barroso and Parthasarathy Ranganathan. *The Datacenter as a Computer*. Morgan & Claypool Publishers, 3 edition, October 2018.

[4] Sandvine Internet Phenomena Report Q3 2019. https://www.sandvine.com/hubfs/Sandvine_Redesign_2019/Downloads/Internet%20Phenomena/Internet%20Phenomena%20Report%20Q32019%2020190910.pdf, 2019. (Accessed on 2020-08-19).