RETROSPECTIVE: Performance of Image and Video Processing with General-purpose Processors and Media ISA Extensions

Parthasarathy Ranganathan∗, Sarita Adve†, Norman P. Jouppi∗
∗Google, †University of Illinois at Urbana-Champaign

I. Motivation

Twenty four years ago, in ISCA1999, we published a paper titled *"Performance of image and video processing with general-purpose processors and media ISA extensions."* At that time, the default in the architecture community was to study SPEC and SPLASH benchmarks. But a new wave of applications were emerging. Driven by advances in digitization and multimedia, and the early internet, *media processing* workloads – encompassing image, video, audio, and graphics – were becoming dominant. New architectures were needed to solve the computational demands of these workloads, but we did not understand these workloads well. Simultaneously, we were also in another transition – architectures that tried to more aggressively leverage instruction-level parallelism (ILP) through techniques like out-of-order execution, speculative execution, etc. Were these new architectures useful for media processing? Did they compete with, or complement, other ideas like new media-specific instruction-set extensions? We wrote this paper to address these questions.

II. Understanding begets Optimization

Key insights. We pulled together a new benchmark suite, and extended the RSIM simulator to support media processing instructions. Our results presented a systematic and detailed understanding of the architectural implications from media processing workloads. Contrary to prevailing intuition at that time, we found that complex ILP techniques were in fact effective for these workloads, and also that general-purpose processors with media instructions were on track to replace DSPs (digital signal processing chips) in achieving some of the ambitious targets of that time (like real-time 30 frames per second).

The groundwork for new optimizations. The insights from our paper led to several interesting follow-on optimizations. Building on our observations about memory hierarchy for media processing workloads, at the very next ISCA (2000), we published a nice idea around reconfigurable caches [7], a design that today is commonplace in the cache partitioning and quality-of-service (QoS) support in processors from Intel and AMD. Another observation around how *"out-of-order issue can better exploit the non-blocking loads feature of the system by allowing the latency of multiple long-latency load misses to be overlapped with one another"* reinforced observations made in other work at that time around MLP (Memory level Parallelism) [5], again now commonplace in the architecture community.

And even more longer-term ripples. Interestingly, beyond immediate follow-on papers, the intuition built from this work also continued to influence broad research directions for the authors. Initially, our work on energy-aware user interfaces [6], robotic telepresence [2], and cross-layer optimization for mobile multimedia systems [11], but the inspiration from this work has continued well past two decades, even in our most recent work (e.g., extended reality [1] and VCU acceleration for media transcoding at scale [9]).

Benchmarks and simulators make the field. This paper also demonstrated the importance of appropriate benchmarks in jump-starting research on a new area. We spent a lot of time creating a combination of kernels and real-world benchmarks that were representative, yet amenable to detailed computer architecture simulation. This is a lesson that we find relevant even today (e.g., release of MLPerf [10], ILLIXR [1], vBench [4], Google traces [8]). Similarly, our original paper was the first to work with a detailed out-of-order simulator (RSIM) with careful modeling of SIMD instruction support. As part of this, we developed a new methodology of cycle-time breakdown for out-of-order processors that continues to be relevant even today [3].

III. Back to the future: an evergreen problem

Our original paper examined how the architecture community should respond to a new class of workloads, viz. media processing. With the recent emergence of new workloads like machine learning and extended reality, the lessons from our experience continue to be more relevant than ever.

The architectural toolbox: then and now. In our original paper, we examined two main classes of architectural optimizations – the instruction-set architecture and the micro-architecture/memory system. These continue to be important levers. The support for bfloat16 in the TPU ML accelerator, the support for vector instructions (e.g., AVX512), or the ongoing momentum in the RISCV open hardware efforts show the continued value of exploring innovations in the former, while innovations like the use of new stacked memory hierarchies in the TPU or accelerator achitecture and system balance tradeoffs in the VCU validate the latter.

However, in retrospect our work missed two key levers that are now more relevant. First, it was not just about the design

of individual chips, but about building larger-scale distributed systems with these chips as building blocks. This means that hardware design needs to deeply consider additional issues like networking, scheduling, scaling, virtualization, quality-of-service, etc. Second, co-design across the entire ecosystem –the hardware, firmware, the drivers, the libraries, the compilers, the tools, all the way to the applications– is important. Whether it is our recent work on extended reality in the ILLIXR project, or video acceleration at warehouse-scale with the VCU project, these two aspects are key to the design of current systems. Another notable missing aspect in our original paper was the lack of discussion on power[1]. Today, understanding power/energy efficiency is a critical pre-requisite when evaluating new optimizations for new workloads, as is also the corresponding focus on environmental sustainability.

A plethora of future opportunities. Interestingly, nearly two decades later, media processing and architectural exploration to enable new use-cases in this area, continue to be as relevant and critical as when we wrote our original paper. Video sharing and video streaming workloads dominate internet traffic; video conferencing is transforming traditional definitions of workplace and collaboration. Immersive computing (including virtual/augmented/mixed/extended reality, metaverse, spatial computing, digital twins, etc.) will process new and multimodal media data, including visual, audio, haptic, and olfactory sensory data, potentially transforming most industries and human activities. There are, however, still orders-of-magnitude power/performance/quality-of-experience gaps between what is available in systems today and what is desired to achieve the full potential of immersive technologies. Blurring the boundaries of continuous ego-centric sensing and computing further opens up questions of trust – how to ensure security, privacy, and other ethical behavior in such environments. Enabling this rich agenda for co-designed hardware and software systems research will require collaborations not just among techologists, but also with applications and human factors researchers (e.g., the IMMERSE Center for Immersive Computing at Illinois - https://immerse.illinois.edu).

IV. Closing remarks

As we conclude our retrospective, we share a few additional reflections. We recall our heated discussions on whether modeling a 1GHz processor was the right future-looking choice. Today, not only does that number look small, but we have also gone beyond the speed wars to more scaling-efficient multicore architectures. We also recall our discussions on implementing the Sun VIS media instruction set. Today, Sun is no longer a company, and other ISAs like x86 and ARM, and more recently RISCV, dominate the conversation. These serve as a reminder of how lucky we are to work in an area that sees tremendous innovation and change, but also how that also comes with a responsibility to focus more beyond immediate fads to fundamentals. The approach in our paper, on focusing on *understanding* before optimizing, continues to be very relevant.

We also recall the serendipitous conversations and collaborations that led to our work in this area. Re-readng the acknowledgements in our original paper reminded us how our interests in researching architectures for media processing were triggered by cross departmental lunch conversations with our colleagues working on digital signal processing, image processing, and graphics. As we look to the future and the next generation of exciting computer architecture innovations ahead, such cross-stack cross-area codesign and cross-pollination will continue to be key.

We are humbled and grateful at being featured in this ISCA50 retrospective. We continue to be excited by the opportunities in this area and hope that the insights and contributions from our work continue to pave the way for future advances.

[1]In defense of the original paper, the end of Dennard scaling happened only several years later, but also to our earlier point on infrastructure making the field, tools like Wattch were developed much later as well.

References

[1] M. Huzaifa, R. Desai, S. Grayson, X. Jiang, Y. Jing, J. Lee, F. Lu, Y. Pang, J. Ravichandran, F. Sinclair, B. Tian, H. Yuan, J. Zhang, and S. V. Adve, "ILLIXR: An Open Testbed to Enable Extended Reality Systems Research," *IEEE Micro*, vol. 42, no. 4.

[2] N. P. Jouppi, S. Iyer, S. Thomas, and A. Slayden, "BiReality: Mutually-immersive Telepresence," in *ACM Multimedia*, 2004, pp. 860–867.

[3] S. Kanev, J. P. Darago, K. Hazelwood, P. Ranganathan, T. Moseley, G.-Y. Wei, and D. Brooks, "Profiling a Warehouse-Scale Computer," in *Proceedings of the 42nd Annual International Symposium on Computer Architecture*, ser. ISCA '15. New York, NY, USA: Association for Computing Machinery, 2015, p. 158–169. [Online]. Available: https://doi.org/10.1145/2749469.2750392

[4] A. Lottarini, A. Ramírez, J. Coburn, M. A. Kim, P. Ranganathan, D. Stodolsky, and M. Wachsler, "vbench: Benchmarking Video Transcoding in the Cloud," in *Proceedings of the Twenty-Third International Conference on Architectural Support for Programming Languages and Operating Systems, ASPLOS 2018, Williamsburg, VA, USA, March 24-28, 2018*, X. Shen, J. Tuck, R. Bianchini, and V. Sarkar, Eds. ACM, 2018, pp. 797–809. [Online]. Available: https://doi.org/10.1145/3173162.3173207

[5] V. S. Pai and S. V. Adve, "Code Transformations to Improve Memory Parallelism," in *Proceedings of the Annual IEEE/ACM International Symposium on Microarchitecture*, 1999, pp. 147–155.

[6] P. Ranganathan, E. Geelhoed, M. Manahan, and K. Nicholas, "Energy-aware user interfaces and energy-adaptive displays," *Computer*, vol. 39, no. 3, pp. 31–38, 2006.

[7] P. Ranganathan, S. Adve, and N. P. Jouppi, "Reconfigurable Caches and Their Application to Media Processing," in *Proceedings of the 27th Annual International Symposium on Computer Architecture*, ser. ISCA '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 214–224. [Online]. Available: https://doi.org/10.1145/339647.339685

[8] P. Ranganathan and V. Lee, "Advancing systems research with open-source Google workload traces," in *Google Cloud Blog*, 2022. [Online]. Available: https://cloud.google.com/blog/topics/systems/workload-traces-for-google-warehouse-scale-computers

[9] P. Ranganathan, D. Stodolsky *et al.*, "Warehouse-scale video acceleration: co-design and deployment in the wild," in *ASPLOS '21: 26th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Virtual Event, USA, April 19-23, 2021*, T. Sherwood, E. D. Berger, and C. Kozyrakis, Eds. ACM, 2021, pp. 600–615. [Online]. Available: https://doi.org/10.1145/3445814.3446723

[10] V. J. Reddi, C. Cheng *et al.*, "MLPerf Inference Benchmark," in *Proceedings of the ACM/IEEE 47th Annual International Symposium on Computer Architecture*, ser. ISCA '20. IEEE Press, 2020, p. 446–459. [Online]. Available: https://doi.org/10.1109/ISCA45697.2020.00045

[11] W. Yuan, K. Nahrstedt, S. V. Adve, D. L. Jones, and R. Kravets, "GRACE-1: Cross-Layer Adaptation for Multimedia Quality and Battery Energy," *IEEE Trans. Mob. Comput.*, vol. 5, no. 7, pp. 799–815, 2006. [Online]. Available: https://doi.org/10.1109/TMC.2006.98