VideoChef: Efficient Approximation for Streaming Video Processing Pipelines
Ran Xu\textsuperscript{α}, Jinkyu Koo\textsuperscript{α}, Rakesh Kumar\textsuperscript{β}, Peter Bai\textsuperscript{β}, Subrata Mitra\textsuperscript{β}, Sasa Misailovic\textsuperscript{γ}, Saurabh Bagchi\textsuperscript{γ}
\textsuperscript{α}: Purdue University, \textsuperscript{β}: Adobe Research, \textsuperscript{γ}: University of Illinois at Urbana - Champaign

Premise of approximate computing
- Video streaming applications require low-latency processing
- Devices are resource constrained

Approximation techniques and parameters
- Loop perforation: for (i = 0; i < n; i = i + approx_level)
  result = compute_result();
- Loop memorization: for (i = 0; i < n; i = i + approx_level)
  if(i % approx_level == 0)
    cached_result = result = compute_result();
  else
    result = cached_result;

Progress in approximation in video processing
- Video proc. w/ approx.
  Optimal parameters: \( (+) \) Final goal
- Video proc. w/ approx.
  Canary + Error mapping + Sampling
  \( (+) \) Unbiased error metric
  \( (+) \) Overhead controlled
  \( (+) \) Close to optimal parameters
- Video proc. w/ approx.
  Canary Input to search
  \( (+) \) Parameters for each input
  \( (+) \) Biased error metric
- Video proc. w/ approximation
  \( (+) \) Too conservative para. for all input.
- Video processing
  \( (+) \) Slow

End-to-end system workflow

Error mapping model, Searching policy and sample policy

Evaluation
- 106 videos w/ 10 video filters and 9 3-stage filter pipelines
- 2 approximation techniques, each with 6 approximation levels
- Comparing 6 configurations and 2 PSNR thresholds

Timing performance
- Reduce 39.1% over exact
- 29.9% over static
- 14.6% over IRA

Quality performance
- Tracks the Oracle quality and the user specified quality threshold violation < 5%

Conclusion and contribution
- A system for performance and accuracy optimization of video streaming pipelines in a data-dependent manner.
- Predictive model to accurately estimate the quality degradation in the full output from the error generated when using the canary input.
- Efficient and incremental search technique for the optimal approximation setting that takes hints from the video encoding parameters to reduce the overhead of the search process.
- Quantitative evaluation and user study

Error mapping model
- \( F = w_0 + w_1 \times C + w_2 \times C^2 \)
- CA model – C model plus approximation parameters
- CAD model – CA model plus feature vectors (row difference)

Searching policy
- Start from (1,1,1), then increase by 1 in each dimension and follow the least-error path until PSNR of full-sized output reaches error threshold.

References

Acknowledgement
This material is based in part upon work supported by the National Science Foundation under Grant Numbers CNS-1527262 and CCF-1703637 and by Sandia National Lab under their Academic Alliance Program.