VideoChef: Efficient Approximation for Streaming Video Processing Pipelines

Ran Xu[α], Jinkyu Koo[α], Rakesh Kumar[α], Peter Bai[α], Subrata Mitra[β] , Sasa Misailovic[γ], Saurabh Bagchi[α]
α: Purdue University, β: Adobe Research, 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:

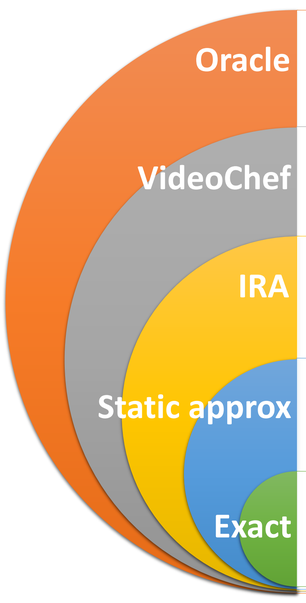
```
for (i = 0; i<n; i = i + approx_level)
  result = compute_result();
```

Loop memorization:

```
for (i = 0; i<n; i = i ++)
  if(i % approx_level == 0)
    cached_result = result = compute_result();
  else
    result = cached_result;
```

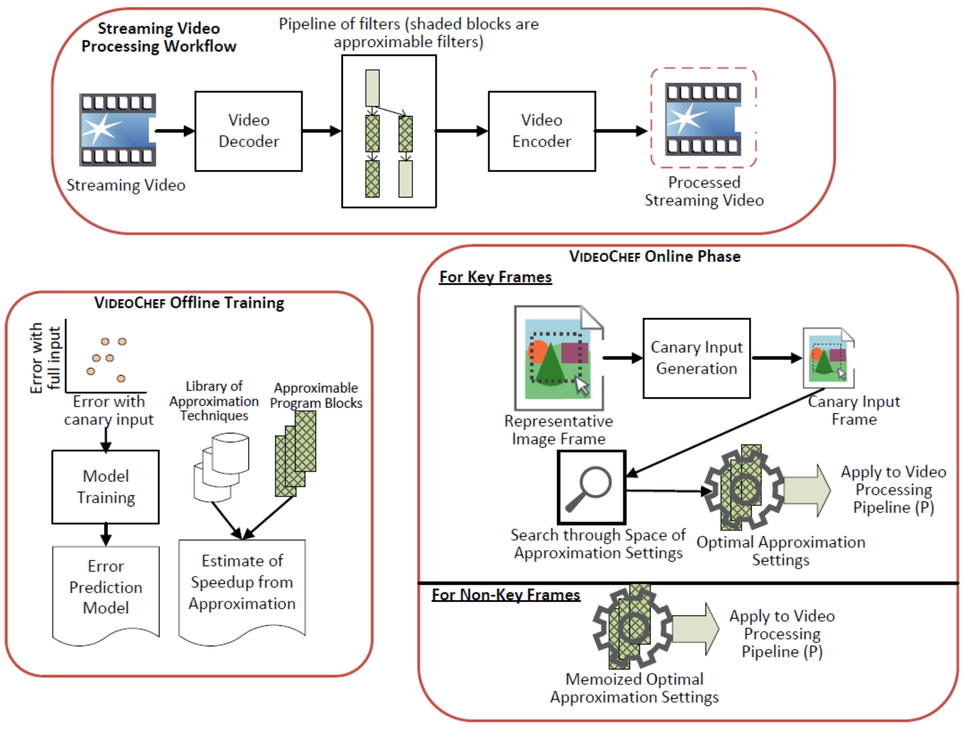
Approximation parameters:
1 = Exact execution
The higher the more approximate

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

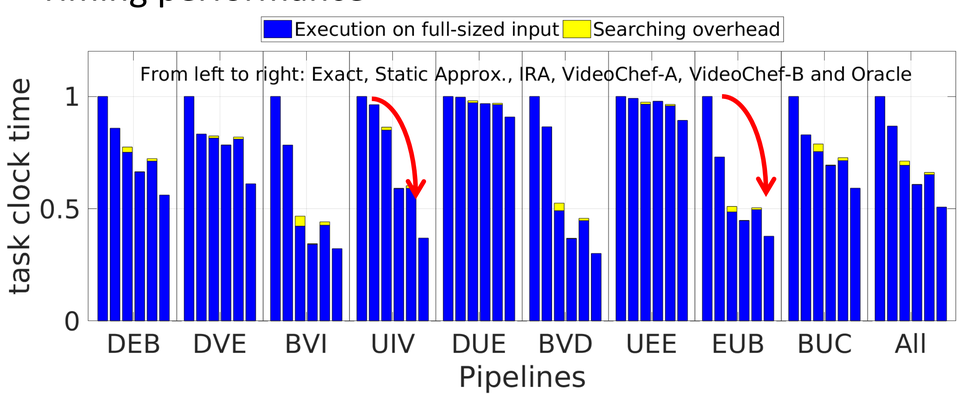


Error mapping model, Searching policy and sample policy

Evaluation

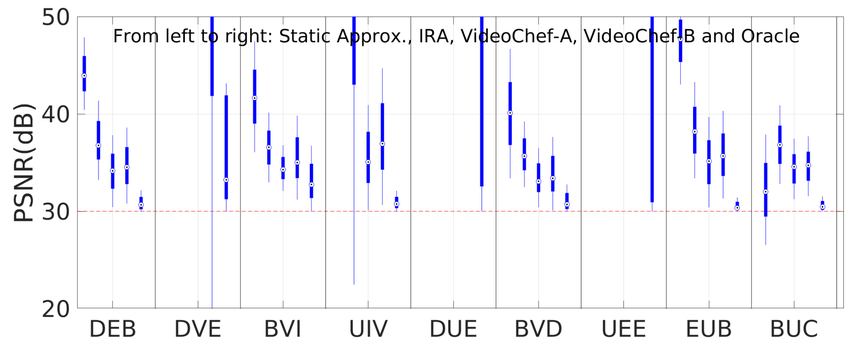
- 106 videos w/ 10 video filters and 9 3-stage filter pipelines
- 2 approximation techniques, each with 6 approximation levels
- Comparing 6 configurations and 2 PSNR thresholds

Timing performance



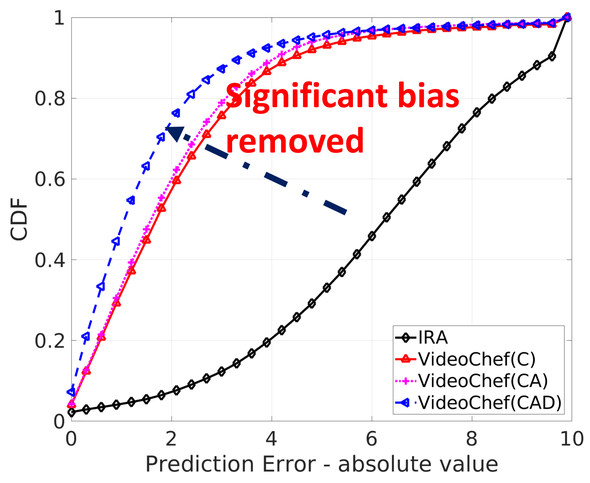
Reduce
39.1% over exact
29.9% over static
14.6% over IRA
Within
20% of Oracle

Quality performance

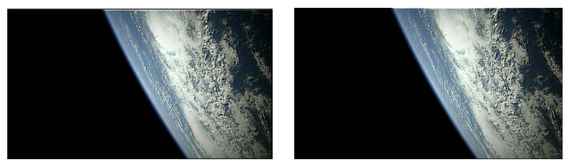


Tracks the Oracle quality and the user specified quality threshold violation < 5%

Error mapping model
– CDF of the two error metrics



User Perception Study
-- Watching VideoChef and Oracle videos side-by-side

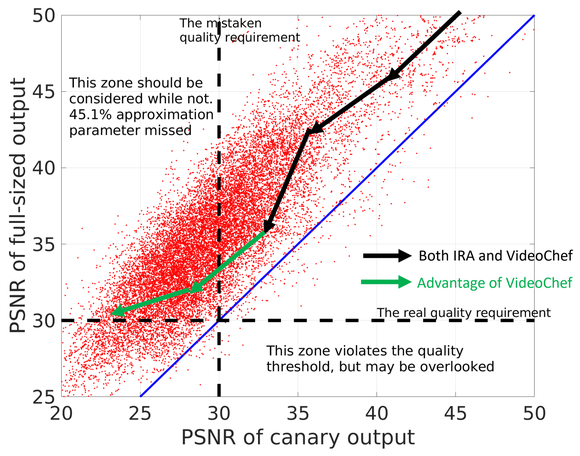


Degree of difference	Percentage
No difference	58.59%
Little difference	34.77%
Large difference	6.64%
Total difference	0

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, Searching policy and sample policy



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.

Error mapping model

- C model – aware of canary PSNR
  $F = w_0 + w_1 \times C + w_2 \times C^2$
- CA model – C model plus approximation parameters
  $F = \vec{w} \cdot (1, C, \vec{A})$
- CAD model – CA model plus feature vectors (row difference)
  $F = \vec{w} \cdot (1, C, \vec{A}, \vec{D})$

Hints to trigger a new search

- I-frames in MPEG-4 videos
- Scene change detector

References

[1] LAURENZANO, M. A., HILL, P., SAMADI, M., MAHLKE, S., MARS, J., AND TANG, L. Input responsiveness: using canary inputs to dynamically steer approximation. In PLDI (2016), ACM.
[2] Xu, R., Koo, J., Kumar, R., Bai, P., Mitra, S., Misailovic, S., & Bagchi, S. VideoChef: Efficient Approximation for Streaming Video Processing Pipelines. In USENIX ATC (2018), USENIX Association.

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.

PURDUE UNIVERSITY
Adobe Research
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN