Massively Parallel Encoding

July 31, 2018

TPX Technology Product Xperience

COMCAST

Why parallelize?

GROWING CODEC COMPLEXITY

- Computational complexity is a price paid for better compression efficiency
  - Successive generations of video coding standards have increasing computational complexity
  - Moore's law, hardware accelerations, properly optimized implementations, help
- High quality UltraHD encoding is still excruciatingly slow
  - 1-2 days per movie is a realistic speed
  - Growing problem as amount of UltraHD content grows

NEED FOR PARALLELIZATION

- Video encoding is CPU-bound
- Newer CPUs have similar speeds, but more cores
- Parallelization reduces turnaround time

Fine-grained parallelism

CODEC TOOLS

- Fine-grained: wavefront parallel processing (HEVC)
  - CTU-level parallel encoding without loss of coding efficiency
- Coarse: slices and tiles
  - Large independent areas within a picture encoded in parallel

TEMPORAL

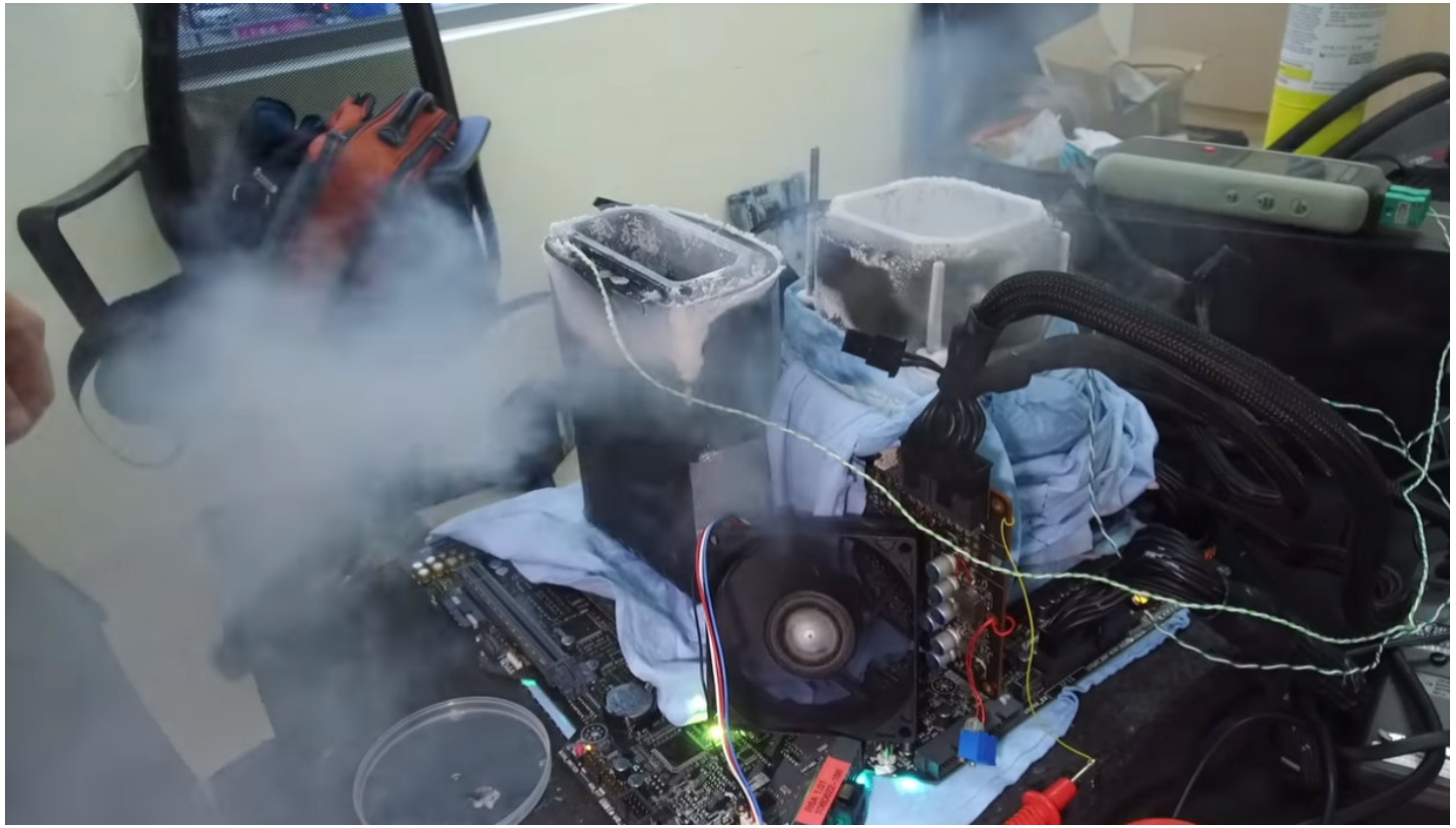
- Fine-grained: frame threads
  - A separate thread encodes a complete frame (or slice)
  - Inter-thread communication used to coordinate CPB state

LIMITATIONS

- Large number of slices reduces compression efficiency and may introduce artefacts
- Large number of frame threads may decrease rate control precision

Why parallelize more?

WE ARE STILL CPU FREQUENCY-BOUND



TPX Technology Product Xperience

COMCAST

Massively parallel encoding

BRUTE FORCE APPROACH

- Video broken into separately encoded “chunks”
  - Each “chunk” encode runs independently
  - Chunks may not run on the same machine, same data center, or same continent
- Chunks “stitched” together for egress
  - Either physically or conceptually via an ABR manifest

BENEFITS

- Better CPU utilization
  - Single serial encode does not always saturate CPU
- Better resource utilization
  - Short chunk encodes can run on cloud infrastructure and “burst” when needed.
- Reduced turnaround time
  - Given enough servers / instances, turnaround is as short as a single chunk encode

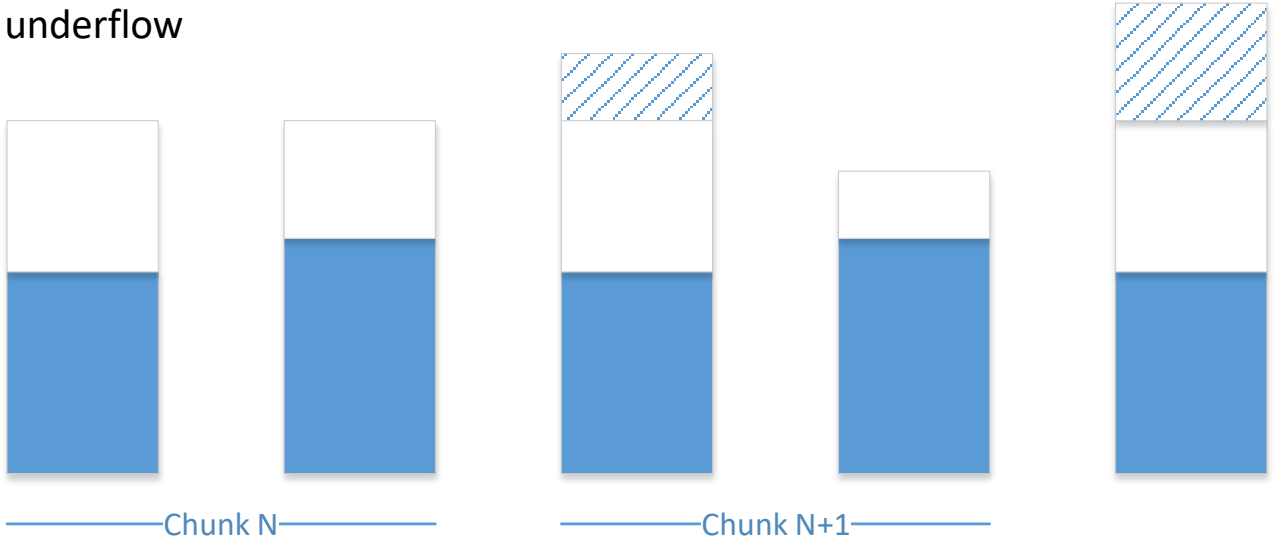
TPX Technology Product Xperience

COMCAST

HRD and Chunked Encoding

PROBLEM

- Rate control of a chunk encode assumes an initial CPB occupancy level
  - It is unaware of the CPB occupancy level at the end of the previous chunk
- Drift between real (“post-stitch”) and perceived CPB occupancy level
  - May result in overflow / underflow



TPX Technology Product Xperience

COMCAST

HRD and Chunked Encoding

X265 SOLUTIONS

- Encoder may be provided with explicit CPB start and end occupancy
- Chunk N is guaranteed that it has at least N unoccupied bits in the beginning
  - leaves at least N bits free in the end of the encode
- Side effect: underutilization of CPB at the end of a chunk
  - Observed when chunk duration is a few seconds
- Alternative solution: encode extra (overlapping) GOPs from each side
  - Allows more precise estimation of CPB end state at a price of extra processing
  - May have mismatches when different rates are used to encode different parts of the system.

Creating chunks

STITCH MARKS

- Naïve chunking by number of frames may cut a scene
- Significant change of quantizer mid-scene may be noticeable
  - More probable in cases of content-adaptive (“per-title”) encoding
- Chunks containing complete scenes reduce the “stitch mark” effect

MINIMAL DURATION

- Encoder initialization takes time
- Shorter chunks imply more encoder initializations and teardowns.
  - Throughput penalty

Multi-representation encodes

ADAPTIVE STREAMING REQUIRES MULTIPLE ENCODES AT DIFFERENT RATES

- Typically each encode is done independently
- Objects boundaries and movements are same across resolutions
- Chances are that different encodes will make similar mode and rate control decisions
  - From rate-distortion standpoint, distortion is similar.

HIERARCHICAL ENCODING

- Low-resolution encodes are performed first and export analysis decision
  - Quadtree structure
  - Motion vectors
- High-resolution encodes reuse information from low-resolution encodes
  - Recent literature suggested the other direction – however this approach results in significantly less savings
- Conceptually similar to hierarchical motion estimation

TPX Technology Product Xperience

COMCAST

Hierarchical encoding

HOW DOES HIERARCHICAL ENCODING WORK?

- “Blueprint” per bitrate ladder
- Implemented in x265
  - `--analysis-[save|load], --reuse-level, --scale`
- Analysis prediction for 2Nx2N uses highest bitrate from the NxN resolution
  - RDO-wise, rate contribution is smaller
- For same resolution prediction, closest higher bitrate is typically best predictor
  - Significantly different rate contribution may make prediction worse

FURTHER DIRECTIONS

- Multi-codec (HEVC -> AVC) encoding may result in speeding up AVC encoding
- HDR -> SDR encoding may result in some speed-ups

TPX Technology Product Xperience

COMCAST

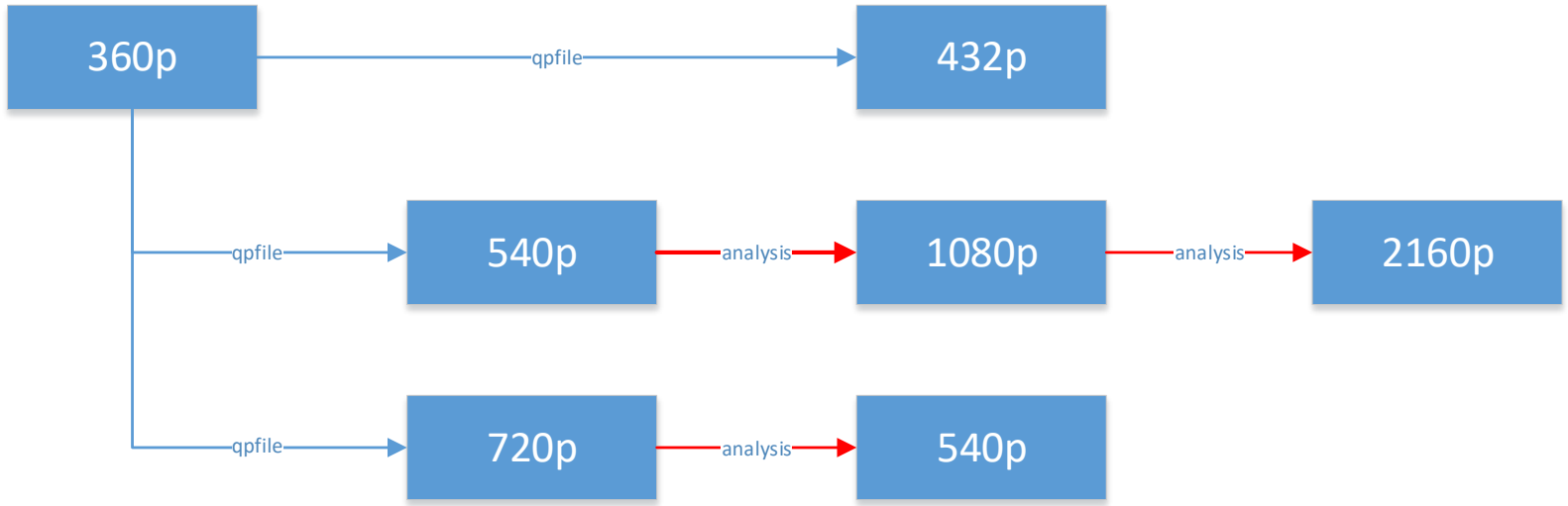
Example blueprint: Apple HLS

HLS HEVC HDR10

- 12-step ladder
  - Most resolutions are multiples of two
  - Several steps per resolution

DEPENDENCIES

- Qpfile: frame types, any resolutions
- Analysis: CU-level analysis, only for multiples of 2



TPX Technology Product Xperience

COMCAST

Results for 2160p

SPEED-UP

- >2x when CU, PU and MV information refined and not recalculated
- Multiple degrees of reuse represent different speed/quality trade-offs
  - Dynamic reuse switches between these strategies on the fly
- Similar results for lower resolutions and three-step 540p->1080p->2160p approach

QUALITY IMPROVEMENT

- Minor but consistent

		Standalone (Values)	Difference (vs. Standalone)		
			reuse-level=2	reuse-level=3	Dynamic
Lucy	Speed-up	100% (0.76fps)	2.4x	1.55x	1.83x
	SSIM+	95.54	-0.93	-0.97	-0.95
Lone Survivor	Speed-up	100% (0.76fps)	2.33x	1.53x	1.83x
	SSIM+	98.31	-0.41	-0.43	-0.42

TPX Technology Product Xperience

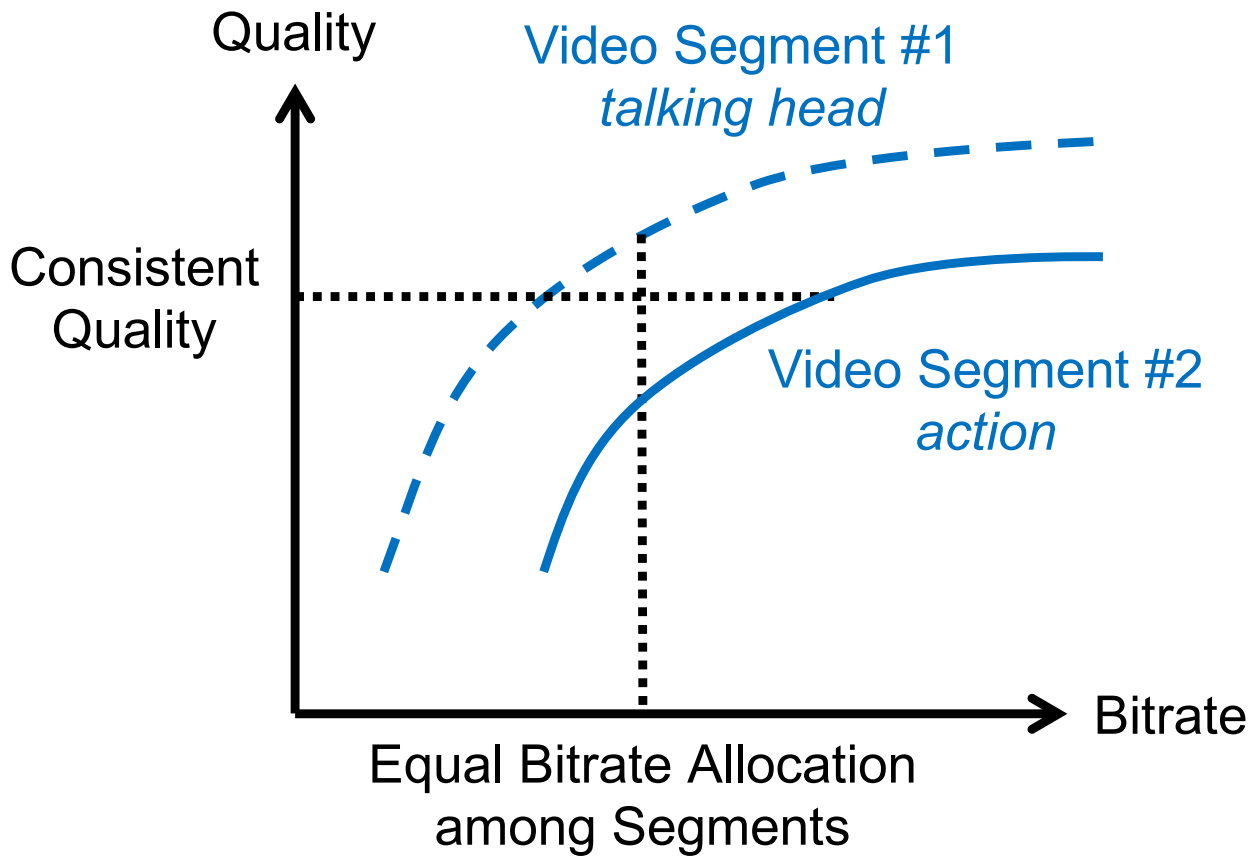
COMCAST

Hierarchical and chunked encoding

“FATTER” ENCODE JOBS

- All steps of a ladder encoded jointly
- All steps are aligned and have same frame types
  - Side benefit: aligned variable GOPs
- Works well with “overlapping” chunked encoding

Complexity differs



TPX Technology Product Xperience

COMCAST

Content adaptive encoding

CONTENT-ADAPTIVE ENCODING

- Each scene gets as many bits as needed to maintain quality
- Popularized by Netflix as “per-title” encoding
- New guidelines published by UltraHD Forum
  - 40% bitrate savings mentioned

CHUNKED HIERARCHICAL ENCODING WORKS WELL WITH CAE

- Lowest resolution is a predictor for higher-resolution bitrates
  - Per scene, not chunk
- Optimal bitrate values derived from low resolution results
  - YouTube paper described use of “trial encode” and machine learning to estimate higher-resolution bitrates
- `--zones` option is your friend in x264/x265

TPX Technology Product Xperience

COMCAST

Q&A

TPX Technology Product Xperience

COMCAST