Ultra-low latency software encoding — VVC and HEVC — capabilities and applications

Mauricio Alvarez-Mesa, PhD (Spin Digital),
Chi Ching Chi (Spin Digital) and
Benjamin Bross (Fraunhofer HHI)

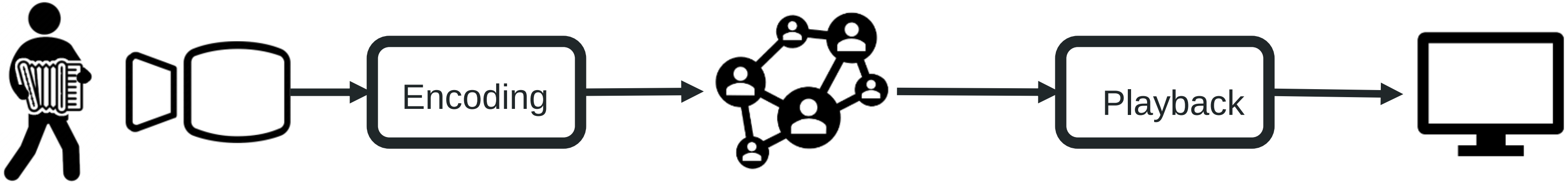
Mile High Video. Denver. CO.
Feb 3rd 2026

Content

1. What is ultra-low latency (ULL)
2. Challenges and applications of ULL
3. Hardware vs software implementations
4. A modern ULL software implementation
5. Latency results
6. Conclusions and future directions

Video Streaming Latency

- **Latency:** cumulative effect of delays (processing & buffering) at every stage
  - **Encoding latency:** lookahead, reordering, encoding, frame parallelism
- **End-to-end latency**
  - Standard: 15-30+ seconds (HLS / DASH)
  - Low: unspecified way of referring to latencies lower than *standard*
    - LL-HLS: 3-10 seconds
  - Ultra-low: required for real-time and interactive applications



Applications of ultra-low latency

Real-time communication

Remote monitoring and teleoperation

Cloud gaming and remote desktop

Broadcast contribution

Application / Latency	Excellent	Acceptable	Unusable
Voice - video call (ITU-T G.114)	< 150 ms	150 ms - 400 ms	> 400 ms
Remote desktop	< 50 ms	50 ms - 150 ms	> 150 ms
Cloud gaming	< 40 ms	40 ms - 100 ms	> 100 ms

Ultra-low latency encoder implementations

Software (CPU-based)

Can use more complex encoding algorithms, frequent updates

Better quality for ultra-low latency at low bitrates

Ultra-low latency encoding on minipcs and laptops, lowest latency limited by CPU speed

Hardware (GPU - ASIC)

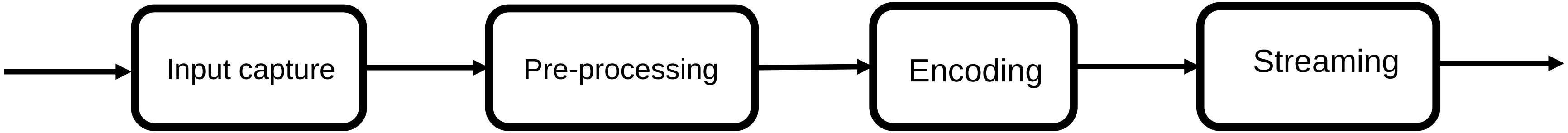
Simple and fast encoding algorithms, feature updates require new hardware platform

Low quality at low bitrates, designed for high bitrates

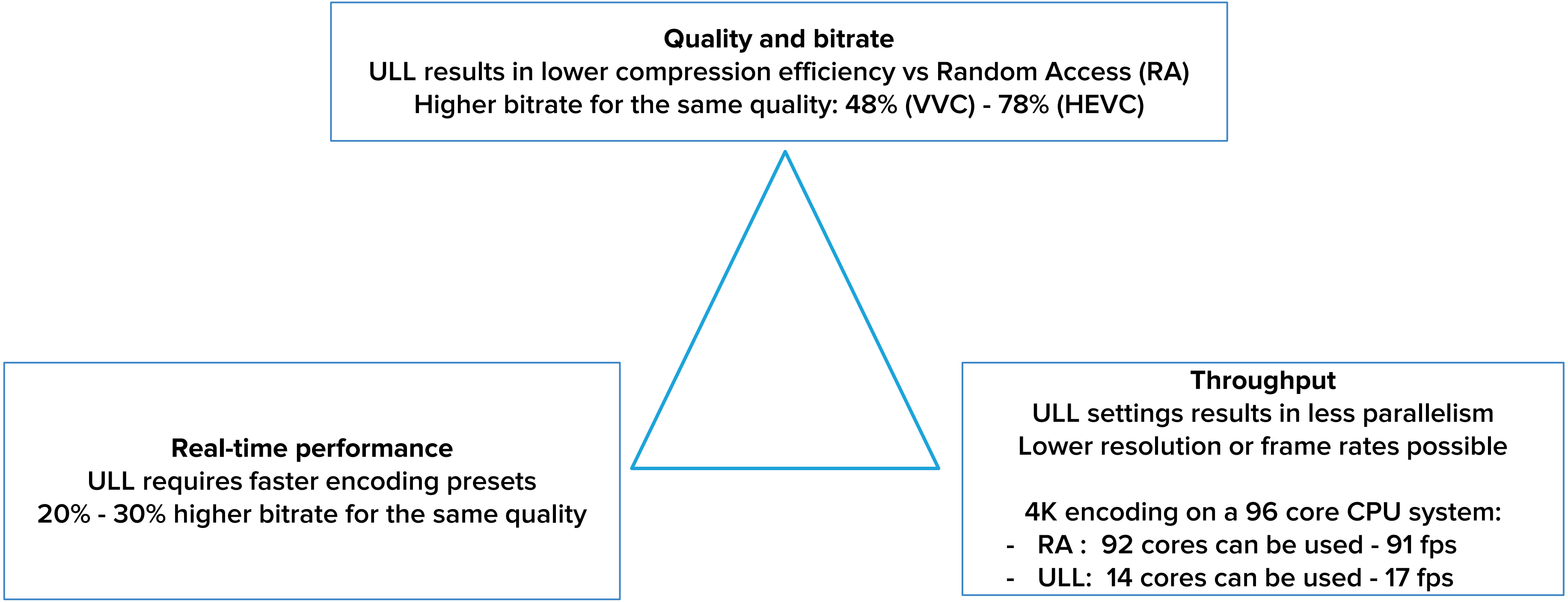
Lowest latency at potentially lowest power and silicon area

A software encoder for ultra-low latency

- 1 frame latency, no frame buffers: ~~lookahead, B-frames, frame parallelism~~
- Gradual decoder refresh (GDR)
- CBR with small HRD buffer
- Optimized HEVC and VVC implementation
- Same core encoder for high efficiency and low latency
- Complete framework for real-time streaming



Tradeoff bitrate, quality, latency and performance



Quality and bitrate
ULL results in lower compression efficiency vs Random Access (RA)
Higher bitrate for the same quality: 48% (VVC) - 78% (HEVC)

Real-time performance
ULL requires faster encoding presets
20% - 30% higher bitrate for the same quality

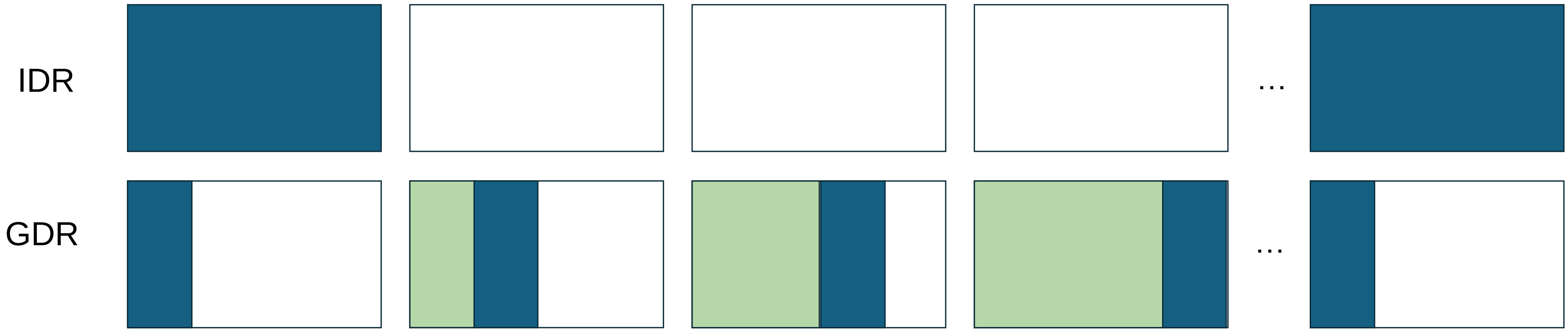
Throughput
ULL settings results in less parallelism
Lower resolution or frame rates possible

4K encoding on a 96 core CPU system:

- RA : 92 cores can be used - 91 fps
- ULL: 14 cores can be used - 17 fps

Low-latency, intra-frames, and GDR

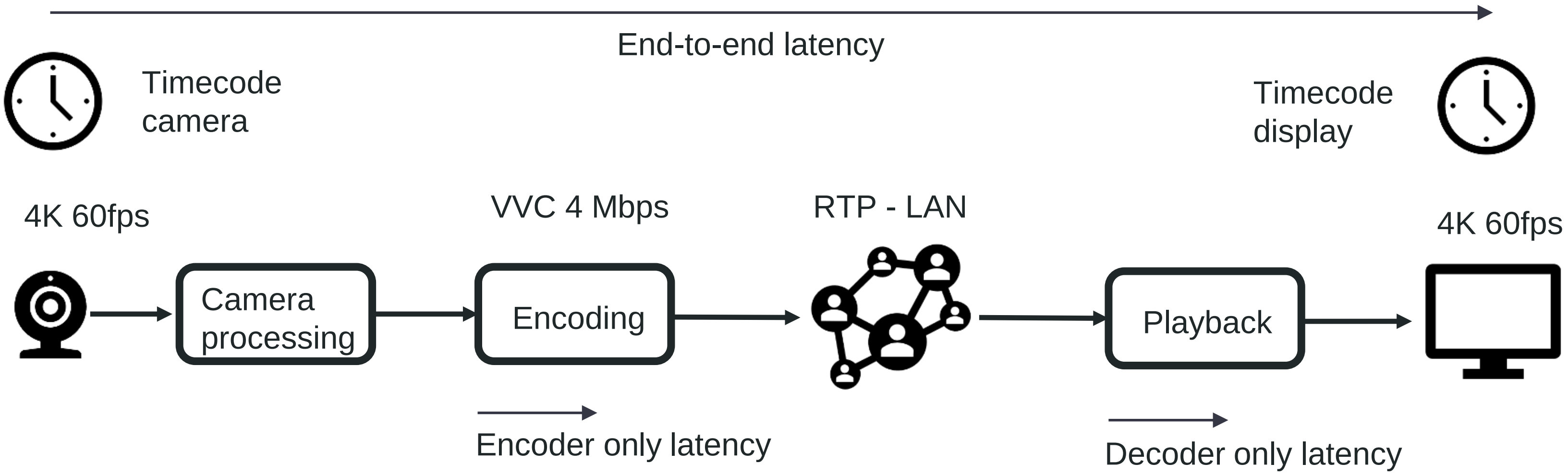
- **Intra frame problem in low delay**
  - Small HRD buffer and low bitrate results in low quality intra frames
  - Affects other frames that reference the IDR
- **GDR (gradual decoding refresh)**
  - Progressively refreshes pictures by spreading intra coded areas over several pictures
  - VVC GDR implementation is more efficient (virtual boundaries vs tiles/slices)



Encoding settings: high efficiency vs low latency

Parameter	**High efficiency**	**Ultra-low latency**
GOP structure	Random access GOP-16 (Hierarchical)	Low-delay P GOP-1 (No B frames)
Decoder refresh	IDR	GDR
HRD buffer	1 second	100 ms
Frame parallelism	9 frames	1 - 2 frames
Lookahead	60 frames	1 frame
Total Encoder latency	**85 frames**	**1-2 frames**

Latency measurements



CPU: AMD Ryzen AI Max+ 395 (16 cores)

CPU: AMD Ryzen 7 8745H (8 cores)

Latency results

	Encoding only	Decoding only	Encoder + decoder	End-to-end
VVC ultra-low latency GOP-1, Lookahead 1f, FiF 2f HRD 100 ms GDR *faster* preset - CBR - 4Mbps	19.02 ms	8.24 ms	27.26 ms	140.00 ms
VVC high efficiency GOP-16 - Lookahead 60f - FiF 9f HRD 1000 ms IDR *faster* preset - CBR - 4Mbps	1,370.00 ms	279.00 ms	1,649.00 ms	3,000.00 ms

Note: Camera processing takes around 100ms

Summary and future directions

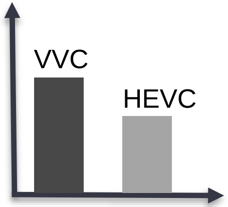
Ultra-low latency applications emerging and growing

Ultra-low latency is possible with software encoders

30 ms VVC encoding and decoding latency possible today

Small form factor devices (minipcs, laptops) can do ULL VVC encoding

GDR is needed for low latency and low bitrate scenarios



VVC outperforms HEVC in ULL with the same computing resources

We are working on further optimizations for ULL encoding and decoding on smaller and lower power platforms

Thank you!

Time for questions

http://spin-digital.com

mauricio@spin-digital.com