Implementing a VVC software live encoder: lessons learned and looking ahead.

Dr. Mauricio Alvarez-Mesa
Chi Ching Chi

Spin Digital Labs

ITU. Geneva. January 17th 2025

Content

1. About us
2. Software encoders for live applications
3. Lessons learned by implementing a VVC encoder
4. Cost, complexity and next-gen codecs

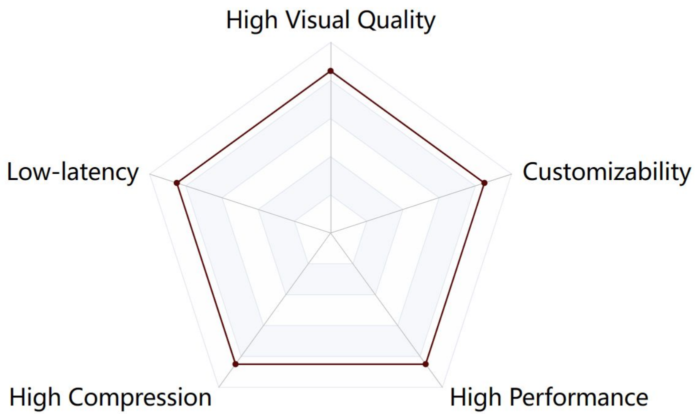
About us: Spin Digital Labs

- Develop high performance video codecs
- Based in Berlin, 10+ years of experience
- Software SDK and applications for HEVC and VVC
  - Live encoding
  - 4K, 8K, 120 fps, HDR, VR-360°
- Broadcast, streaming, immersive media

8K live streaming presented by Intel at Paris Olympics

Software encoders for live applications

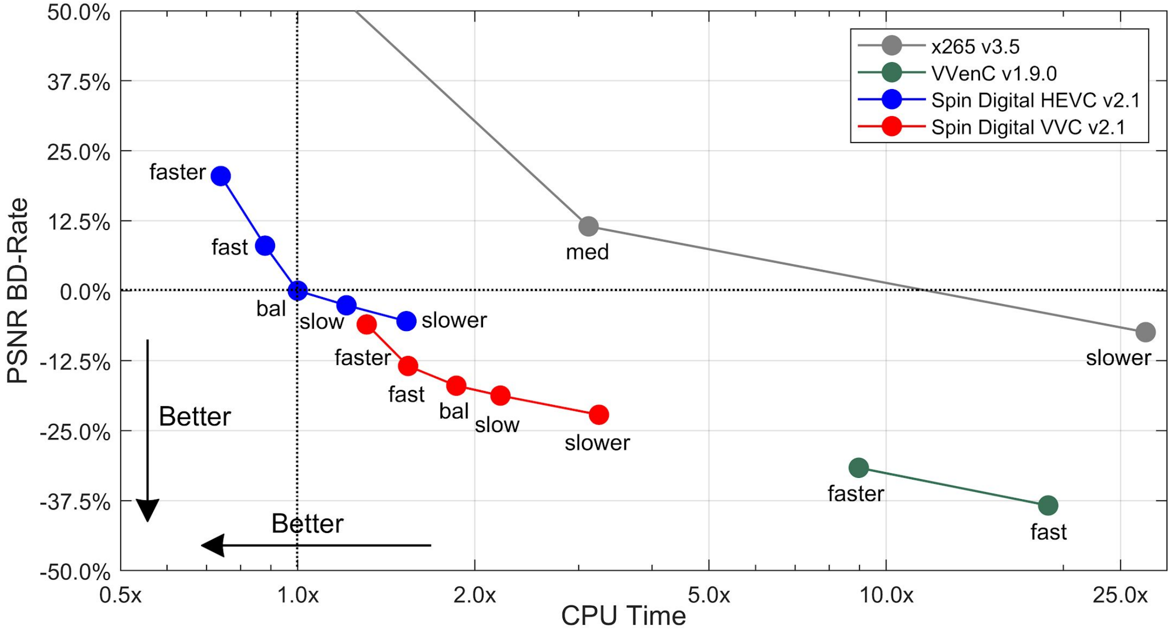
- Our research questions:
  - What is the potential of VVC for live applications?
  - What is the practical bitrate reduction compared to HEVC?
  - What is possible with affordable computing?
- Our starting point:
  - An existing real-time HEVC encoder
  - Intended use cases: live broadcasting and streaming
- Our achievements:
  - A highly optimized software live VVC encoder
  - Compression gains compared to optimized HEVC
  - Using standard CPU server platforms



Lessons learned:

1. Practical VVC bitrate gains are ~20% not ~40%.
2. VVC performance is limited by complexity and cost.
3. Tradeoff: Parallel processing vs compression efficiency
4. Complex (RDO intensive) coding tools limit performance

1. Practical VVC gains are ~20% (not ~40%)

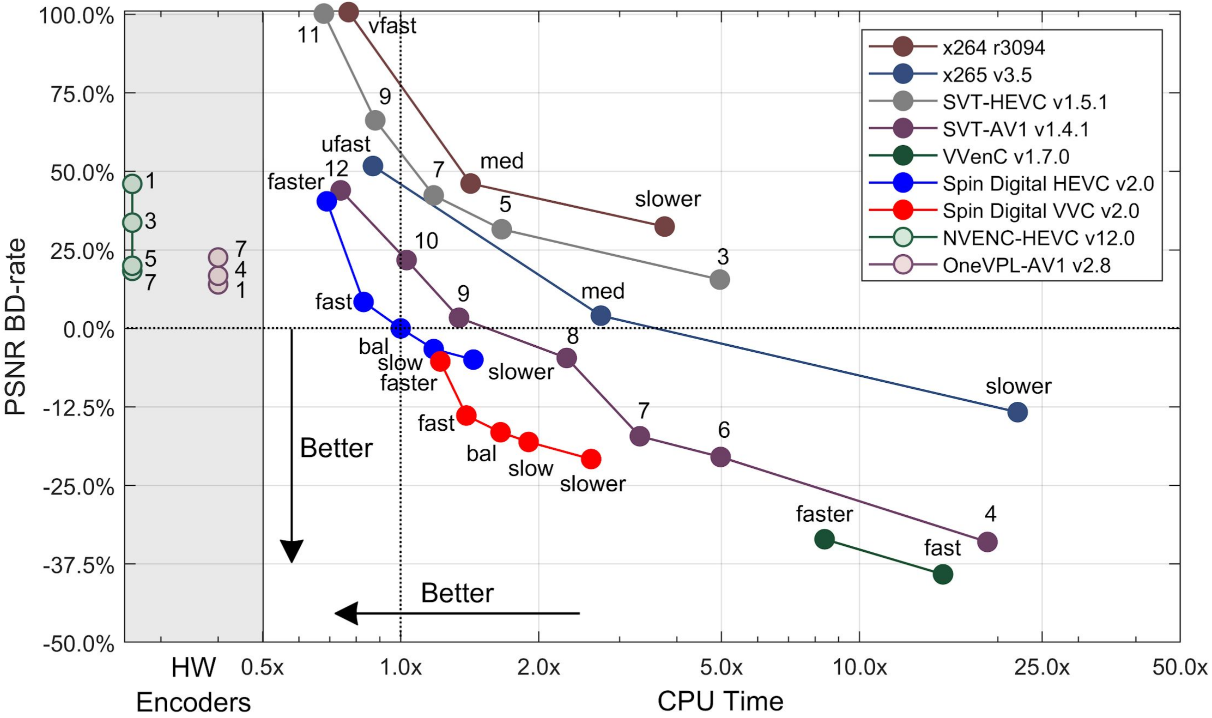


- We implemented the tools with the best coding gains and suitability for real time.
- VVC reference: 39% savings based on BD-Rate PSNR. 50% savings based on MOS

Test conditions

- 11 1-minute 4K videos
- Random access
- Rate control: CBR
- BD-rate PSNR
- Single threaded CPU complexity
- Spin Digital HEVC/VVC Encoders: Dec 2023 - SDK v6.1
- Reference: Spin Digital HEVC

2. VVC performance is limited by complexity and cost



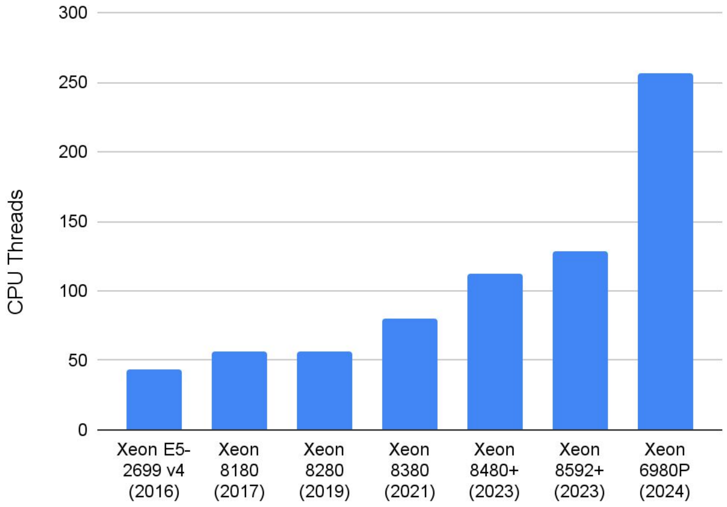
- HEVC replaces AVC
- VVC extends HEVC
- The set of usable coding tools in VVC is limited by complexity and cost

Test conditions

- 11 1-minute 4K videos
- Random access
- Rate control: CBR
- BD-rate PSNR
- Single threaded CPU complexity
- Spin Digital HEVC/VVC Encoders
  - Feb 2023 - SDK v6.0
- Reference: Spin Digital HEVC
- Download full report

3. Parallel processing vs compression efficiency

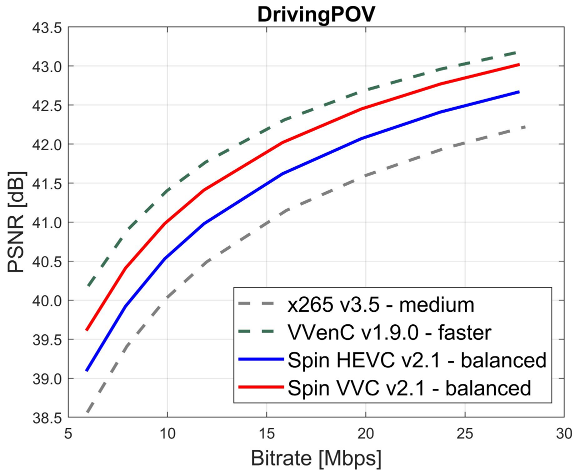
- Modern CPU architectures
  - Single threaded improvements are limited
  - SIMD instructions stable (AVX 512)
  - Growing number of CPU cores
- Challenges of implementation
  - How to use many cores efficiently, achieve compression gains, and real-time performance
  - Tradeoff parallelism vs compression, latency, quality



CPU threads over different generations of Intel Xeon server CPUs

4. RDO intensive coding tools limit performance

- RDO intensive coding tools
  - RDO at the sub-block level
  - Complex sequential evaluation
  - Not clear intrinsic correlation with image statistics
- Technical limitations
  - Relies on single threaded performance
  - Cannot be used for live encoding
- Business / cost limitations
  - Compression efficiency comes with high cost
  - Industry is becoming more cost-sensitive (stable market)
  - Costs only affordable for highly viewed VoD streams

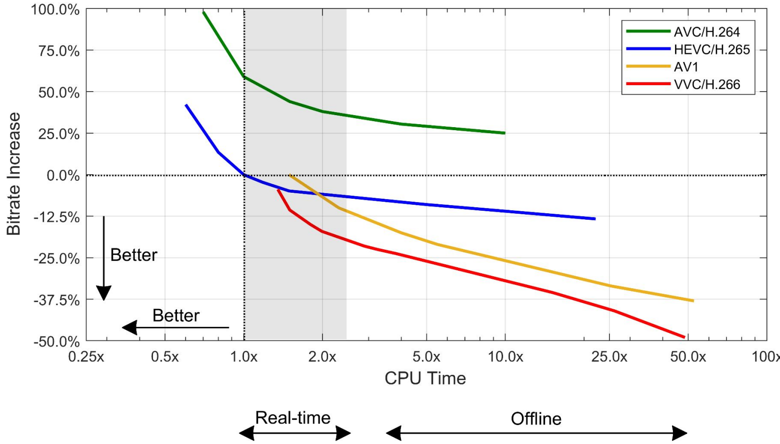


Test conditions

- 4Kp59.94 HDR
- Random access
- Rate control: CBR
- Spin Digital HEVC/VVC Encoders: Dec 2023 - SDK v6.1

Cost, complexity and next-gen codecs

- If next generation codec continues trend on more reliance on complex RDO extensive coding tools
  - **Live**: Practical real-time software encoders will not be significantly better than VVC
  - **VoD**: Economic benefit limited to high watch time offline streaming
  - Predict slower adoption than VVC as the use cases will be more niche
- Can we rethink complexity?
  - Coding tools that can use of many core CPUs GPU, and Matrix/NN extensions



Thank you!

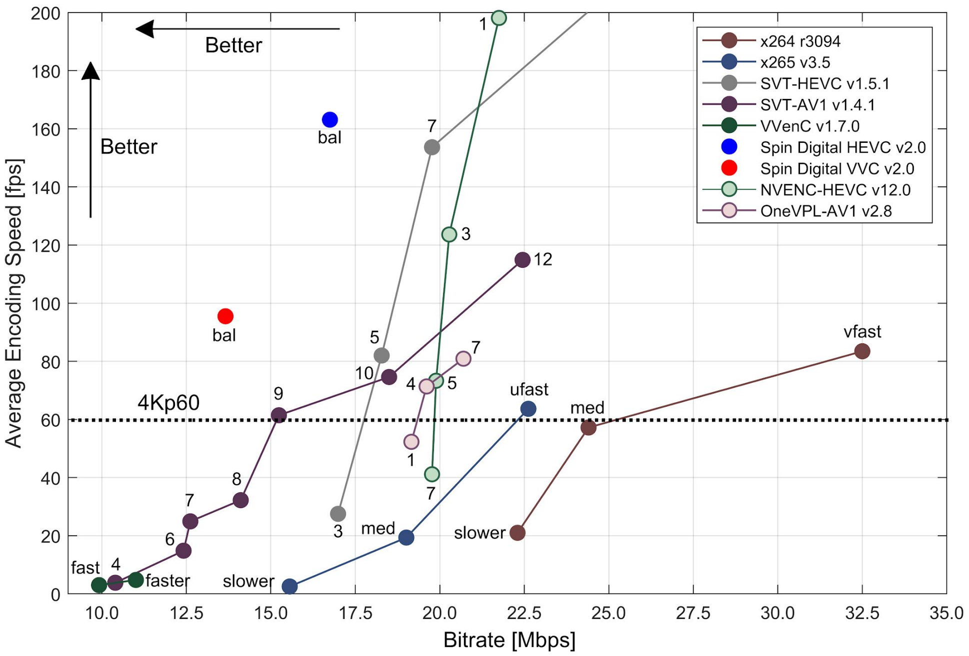
http://spin-digital.com

Backup slides

Modern CPU architectures: example Intel server CPUs

CPU model	Year	Num cores / threads	SIMD	Base frequency [GHz]	TDP [Watt]
Xeon 6980P (Granite Rapids)	2024	128 / 256	AVX 512 + VNNI + AMX	2.0	500
Xeon 8592+ (Emerald Rapids)	2023	64/ 128	AVX 512 + VNNI + AMX	1.9	350
Xeon 8480+ (Sapphire Rapids)	2023	56 / 112	AVX 512 + VNNI + AMX	2.0	350
Xeon 8380 (Ice Lake)	2021	40 / 80	AVX 512 + VNNI	2.3	270
Xeon 8280 (Cascade Lake)	2019	28 / 56	AVX 512 + VNNI	2.7	205
Xeon 8180 (Skylake)	2017	28 / 56	AVX-512	2.5	205
Xeon E5-2699 v4 (Broadwell)	2016	22 / 44	AVX2	2.2	145

Multithreaded performance



Test conditions

- Frame rate at the same quality (PSNR of 41.5 dB)
- When encoding DrivingPOV (4K 10-bit HDR)
- using 2x Intel Xeon Platinum 8368 CPU (2x 38 cores)
- GPU encoders:
  - RTX3070 GPU for NVENC,
  - ARC A770 GPU for OneVPL

Possible directions

- Improve CPU architecture to be more capable of extracting “micro” parallelism available in RDO process
    - Cluster multi-threading
- Tools that fix gaps in current expressive capabilities
    - For example, Film grain or noise in general
- Coding tools that correlate more with video properties
    - Examples
        - New hardware advances image segmentation and object recognition, able to do this in real time
        - If new coding tools would correlate strongly with some of these inputs, encoder complexity increase is manageable