AN ANALYSIS OF THE HEVC HARDWARE ENCODER TOOLS IN AN NVIDIA RTX 4070 TI GPU

ALLAN SCHUCH[1]; VITOR COSTA[2]; DANIEL PALOMINO[3]; MARCELO PORTO[4]

[1]Universidade Federal de Pelotas – anschuch@inf.ufpel.edu.br
[2]Universidade Federal de Pelotas – vscosta@inf.ufpel.edu.br
[3]Universidade Federal de Pelotas – dpalomino@inf.ufpel.edu.br
[4]Universidade Federal de Pelotas – porto@inf.ufpel.edu.br

1. INTRODUCTION

The massive growth of high-resolution video streaming demands efficient real-time encoding solutions. While software encoders are flexible, dedicated hardware accelerators in Graphics Processing Units (GPUs), such as NVIDIA's NVENC (NVIDIA, 2024), are essential for achieving the required performance and power efficiency demanded for such tasks. However, to operate within strict physical constraints (power, area, thermal), these hardware encoders implement simplified or restricted versions of complex video coding standards like the High Efficiency Video Coding (HEVC) (ITU-T, 2024).

While studies have investigated these constraints on mobile devices (COSTA, V.; PERLEBERG, M.; AGOSTINI L.; PORTO M., 2024), a similar investigation for GPUs is less common. This paper addresses this gap by presenting an analysis of the HEVC toolset implemented in the NVENC encoder of the NVIDIA RTX 4070 Ti GPU (NVIDIA, 2023). The findings are compared with the HEVC tools present in the HEVC Test Model (HM) reference software version 18.0 (HHI, 2023) using the Random-Access (RA) profile, which serves as an upper-bound reference, and the constrained Apple A15 Bionic (APPLE, 2021) mobile encoder to highlight the design trade-offs inherent to different platforms.

2. METHODOLOGY

The methodology adopted in this work is based on the analysis of bitstreams generated from UHD 4K video encodings to infer the utilization or absence of specific HEVC tools, a process that resembles reverse engineering. The analysis focused on a key set of HEVC tools that was possible to observe, including: Coding Tree Unit (CTU) size, Group of Pictures (GOP) size and structure, intra period, number of reference pictures, support for biprediction and Asymmetric Partition Mode (APM), the Motion Vector (MV) search range, support for Fractional Motion Estimation (FME), limitations on Prediction Blocks (PBs) sizes, the Transform Unit (TU) hierarchy, and support for the Discrete Sine Transform (DST), among others.

The encodings were performed on an Ubuntu 22.04.5 LTS (LTD., C, 2023) system with an NVIDIA GeForce RTX 4070 Ti (12GB) GPU. The FFmpeg software (version N-117204-gc65a294f79) with the hevc_nvenc encoder was used (FFMPEG, 2025). The test sequences were selected from Classes A1 and A2 of the Common Test Conditions (CTCs) (SHARMAN, K.; SÜHRING, K., 2017), comprising UHD 4K videos, to subject the encoder to a high computational demand scenario. To identify the most comprehensive toolset, all combinations of NVENC presets were tested, and the least restrictive configuration (P7 preset, UHQ tuning) was chosen for this experiment (NVIDIA, 2025). The Split Frame feature was disabled to ensure an isolated evaluation of a single encoder instance (NVIDIA, 2025).

The final step involved decoding the generated bitstreams using a modified version of the HM-18.0 software. The modifications implemented in the decoder's source code allowed for the extraction and display of information contained in the bitstream header, as well as decisions made by the encoder during the encoding process, making it possible to infer which tools of the HEVC standard were utilized by the NVENC encoder.

3. RESULTS AND DISCUSSION

The results of the enabled tools are summarized in Table I, which contrasts the NVENC implementation with the Apple A15 Bionic mobile chipset and the HM-18.0 reference software. Despite its focus on high performance, the NVENC implementation demonstrates clear design trade-offs to manage hardware complexity. The restriction of small, non-square inter-prediction PBs (16x4, 4x16, 8x4, and 4x8) and the limitation of the CTU size to 32x32 (instead of the 64x64 supported by the standard) are choices that likely reduce the number of partition tests the encoder must perform, optimizing for real-time throughput.

However, NVENC retains a rich feature set to ensure high quality. It operates with a hierarchical B-frame GOP structure of size six and allows up to five reference frames, indicating a prioritization of coding efficiency. The support for both biprediction and APM offers flexibility and greater precision in inter-frame prediction. Furthermore, the wide MV search range and the full support for all TU sizes (from 4x4 to 32x32) with all possible depths reinforce the encoder's ability to adapt to complex content.

The comparison with the Apple A15 Bionic reveals the distinct design priorities for each platform. As expected for a high-performance GPU, NVENC supports a more comprehensive toolset. For example, its MV search range is considerably larger than that of the A15 Bionic. In contrast, the A15 Bionic employs a smaller GOP (size 4) with only two reference frames, a strategy that reduces memory demand and processing complexity, which are critical factors in mobile devices. The HM-18.0 reference software, on the other hand, shows a configuration that prioritizes maximum coding efficiency (64x64 CTU, GOP size 16) without concern for the complexity and resource constraints typical of hardware implementations.

4. CONCLUSIONS

This paper conducted an analysis of the HEVC encoding tools implemented in the NVENC hardware encoder on the Nvidia RTX 4070Ti GPU. Through bitstream analysis, it was possible to identify the set of enabled functionalities and adopted design restrictions. The comparison with the Apple A15 Bionic and the HM-18.0 reference software revealed distinct design choices influenced by the characteristics of each platform, highlighting the contrast between high-performance GPU hardware and a resource-limited mobile device. As future work, a more in-depth investigation will be conducted, complemented by an analysis of energy consumption, video quality, and performance to provide a more comprehensive evaluation of the trade-offs involved in the implementation choices of the NVENC encoder.

Table I. Comparison of HEVC encoding tools between NVENC (RTX 4070 Ti), HM-18.0, and Apple A15 Bionic

Tools/Parameters	***HM-18.0 (default RA)***	***Apple A15 Bionic****	***NVENC (RTX 4070 Ti)***
CTU Size	64x64	32x32	32x32
GOP Size	16	4	6
GOP Structure	1 P-Frame + 15 B-Frames	-	P-B-B-B-B-B
Intra Period	32	56	250
Reference pictures	Up to 5	2	Up to 5
Biprediction support	Yes	Yes	Yes
APM support	Yes	No	Yes
MV interval on X	[-1823, 1605]	[-316, 315]	[-1361, 1523]
MV interval on Y	[-1279, 1031]	[-188, 187]	[-512, 507]
FME support	Yes	-	Yes (quarter pixel precision)
Intra modes	All	Without odd directional modes for 4x4 PBs (UHD@60 fps videos)	All
Inter PB limitations	No	No 8x4 and 4x8 block support	No 16x4, 4x16, 8x4 and 4x8 block support
Intra PB limitations	No	No 4x4 luma PBs	No
Luma TU sizes (max depth relative to CU)	4x4 (2), 8x8 (2), 16x16 (2), 32x32 (1)		4x4 (3), 8x8 (2), 16x16 (1), 32x32 (0)
Chroma TU sizes (max depth relative to CU)	4x4 (2), 8x8 (2), 16x16 (2)	-	4x4 (3), 8x8 (2), 16x16 (1)
DST available	-	-	Yes (for 4x4 TUs)
Slices per frame	-	-	1
Tiles per frame	-	-	1

Fields filled with "-" indicate the absence of available value for the analyzed parameter.
*Data retrieved from (COSTA, V.; PERLEBERG, M.; AGOSTINI L.; PORTO M., 2024).

5. REFERENCES

APPLE. iPhone 13 - Tech Specs. 2021. Acesso em: 06 de agosto de 2025, Página da Web. Disponível em: <https://support.apple.com/en-us/111872>.

ARUNRUANGSIRILERT, K.; KATTO, J. Evaluation of Hardware-based Video Encoders on Modern GPUs for UHD Live-Streaming. In: INTERNATIONAL CONFERENCE ON COMPUTER COMMUNICATIONS AND NETWORKS (ICCCN), 2024., 2024. Anais. . . [S.l.: s.n.], 2024. p.1–9.

COSTA, V.; PERLEBERG, M.; AGOSTINI, L.; PORTO, M. Coding Efficiency and Time Evaluation of Apple A15 Bionic Chipset HEVC Encoder. In: IEEE 15TH LATIN AMERICA SYMPOSIUM ON CIRCUITS AND SYSTEMS (LASCAS), 2024., 2024. Anais. . . [S.l.: s.n.], 2024. p.1–5.

FFMPEG. FFmpeg Multimedia Framework. 2025. Acesso em: 06 de agosto de 2025, Site do Projeto. Disponível em: <https://ffmpeg.org/>.

HHI, F. High Efficiency Video Coding (HEVC) Test Model (HM). 2023.
Acesso em: 06 de agosto de 2025, Software Release. Disponível em:
<https://vcgit.hhi.fraunhofer.de/jvet/HM/-/releases/HM-18.0>.32

ITU-T. Recommendation H.265: High efficiency video coding. 2024. Disponível em: <https://www.itu.int/itu-t/recommendations/rec.aspx?rec=14107>.

LTD., C. Ubuntu 22.04.5 LTS (Jammy Jellyfish). 2023. Acesso em: 06 de agosto de 2025, Versão de Sistema Operacional. Disponível em: <https://releases.ubuntu.com/jammy/>.

NVIDIA. GeForce RTX 4070 Family. 2023. Acesso em: 06 de agosto de 2025, Página da Web. Disponível em: <https://www.nvidia.com/en-us/geforce/graphics-cards/40-series/rtx-4070-family>.

NVIDIA. NVENC Application Note. 2024. Acesso em: 06 de agosto de 2025. Disponível em: <https://docs.nvidia.com/video-technologies/video-codec-sdk/13.0/nvenc-application-note/index.html>.

NVIDIA. NVENC Video Encoder API Programming Guide. 2025. Acesso em: 06 de agosto de 2025. Disponível em: <https://docs.nvidia.com/video-technologies/video-codec-sdk/13.0/nvenc-video-encoder-api-prog-guide/index.html>.

SHARMAN, K.; SÜHRING, K. Common Test Conditions for HM video coding experiments. Tech. Rep., document JCTVC-AC1100 of JCT-VC, [S.l.], 2017.

SZE, V.; BUDAGAVI, M.; SULLIVAN, G. J. High Efficiency Video Coding (HEVC): Algorithms and Architectures. [S.l.]: Springer Publishing Company, Incorporated, 2014.