Work-in-Progress: A Machine Learning-Based Approach for Power and Thermal Management of Next-Generation Video Coding on MPSoCs

Arman Iranfar, Marina Zapater, David Atienza
Embedded Systems Laboratory (ESL), Swiss Federal Institute of Technology Lausanne (EPFL), Switzerland
{arman.iranfar, marina.zapater, david.atienza}@epfl.ch

ABSTRACT

High Efficiency Video Coding (HEVC) provides high efficiency at the cost of increased computational complexity followed by increased power consumption and temperature of current Multi-Processor Systems-on-Chip (MPSoCs). In this paper, we propose a machine learning-based power and thermal management approach that dynamically learns the best encoder configuration and core frequency for each of the several video streams running in an MPSoC, using information from frame compression, quality, performance, total power and temperature. We implement our approach in an enterprise multicore server and compare it against state-of-the-art techniques. Our approach improves video quality and performance by 17% and 11%, respectively, while reducing average temperature by 12%, without degrading compression or increasing power.

KEYWORDS

HEVC, power/thermal management, machine learning, MPSoC

ACM Reference format:
Arman Iranfar, Marina Zapater, David Atienza. 2017. Work-in-Progress: A Machine Learning-Based Approach for Power and Thermal Management of Next-Generation Video Coding on MPSoCs. In *Proceedings of CODES/ISSS '17 Companion, Seoul, Republic of Korea, October 15–20, 2017,* 2 pages.
DOI: 10.1145/3125502.3125533

1 INTRODUCTION

Video streaming services are expected to account for 80% of global traffic by 2019 [3]. Due to the great variety of devices accessing media content as well as the users' demand for higher quality video, encoding has become a key application in current High Performance Computing (HPC). To satisfy the emerging large video resolutions and frame rates, the High Efficiency Video Coding (HEVC) standard provides twice the compression of its predecessors while maintaining the same video quality, at the price of increasing the encoder complexity by several times [2].

This, together with the increase of video streaming users, poses an important challenge for power- and thermal-aware resource allocation and management of these applications when running on multiprocessor systems-on-chip (MPSoCs). Current research in the area is mostly focused on the optimization of one or several blocks of the encoding algorithm [4] shown in Figure 1. Each block contains several parameters to configure the encoder (i.e., configuration knobs), which affect encoding efficiency, power consumption, temperature, and processing time. The most important ones are shown in Figure 1. Figure 2 shows the impact of QP on encoding efficiency and performance for an HD test sequence. The same bar graphs can be showns for other encoding parameters.

Moreover, the contents of a video along with the video type (in terms of resolution, bit depth, etc.) play a major role in the obtained performance (encoding time per frame), quality (peak signal-to-noise ratio, PSNR, measured in dB), compression (bitrate, measured in bits per second, bps), power consumption, and peak temperature. Such variations in the video contents motivate a frame-by-frame power and thermal management. Therefore, the encoding configuration and the CPU frequency must be dynamically adjusted to provide the best possible outcomes. To the best of our knowledge, few works jointly consider temperature constraints as well as encoding efficiency of next generation video encoders [5–8], none of which addresses multi-streaming on MPSoCs. Moreover, the great number of different combinations of configuration knobs, in addition to content variations within a video and diversity of video

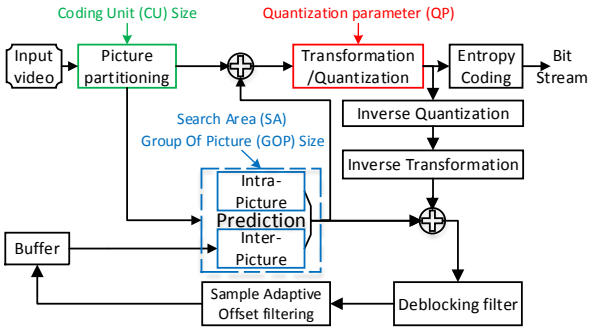

Figure 1: HEVC encoder block diagram and main configuration parameters

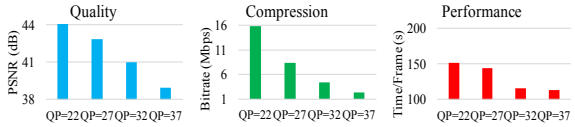

Figure 2: Impact of QP on quality, compression, and performance

CODES/ISSS '17 Companion, Seoul, Republic of Korea

978-1-4503-5185-0/17/10…$15.00
DOI: 10.1145/3125502.3125533

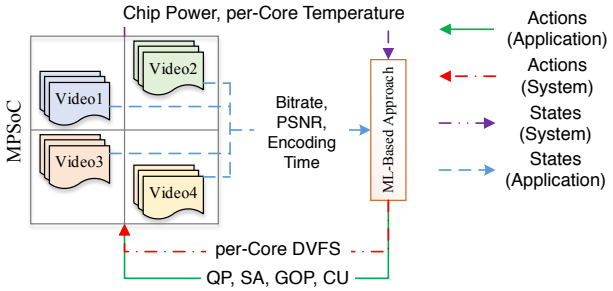

Figure 3: Proposed ML-based approach

types require a more generic solution than that proposed by previous works, which considers power, temperature, performance, and encoding efficiency. The contributions of our work are as follows:

- we propose a machine learning (ML)-based power and temperature management approach for multi-streaming on MPSoCs considering next generation video coding standards
- we address the challenge of power and thermal management for HEVC by jointly integrating application-level configuration and system-level knobs on top of any arbitrary algorithmic optimizations.

2 PROPOSED ML-BASED APPROACH

Our proposed ML-based approach uses the Q-learning algorithm to learn and apply the best encoding configuration and core frequency for each of the streams being encoded on a MPSoC in order to increase PSNR and performance, reduce temperature, and satisfy a predefined target bitrate and power cap. Figure 3 shows the proposed approach.

The proposed ML-based approach consists of 3 phases, namely *exploration*, *exploration-exploitation*, and *exploitation*. In the exploration phase, once the first frame arrives, an action is selected randomly from an action pool, which includes all available configuration knobs and operating frequencies. Once the state transitions from an initial state to a new one, the new Q-value corresponding to the selected action and the initial state is calculated.

The exploration phase for each pair of state-action continues until the learning rate, defined as a function of the number of state-action observations, decreases below a predefined threshold. Thereafter, in the exploration-exploitation phase, the ML agent exploits the learned state-action pairs, while updating the Q-values. When the learning rate again decreases to another predefined threshold, the exploitation phase starts and the agent simply exploits the learned state-action pairs without any update to the Q-values. Finally, we define the states and available actions, as well as the reward function, as follows:

States. States observed by the ML agent can be divided in system-level and application-level states, shown in Figure 3.

Actions. The proposed action pool consists of the most effective encoding configuration modes, including *Search area (SA)*, *QP*, *CU* size, *GOP* size, in conjunction with the available core frequencies, shown in Figure 3.

Reward Function. The proposed reward function provides a proper feedback from the selected action for a previous state and is a weighted average of five sub-function, one for each observed state. Each sub-function provides a higher reward for more desirable states, and it has to provide sufficiently large negative reward when temperature constraint and power cap are not satisfied.

Table 1: PSNR, deviation from the target bitrate, performance, power consumption and average temperature (θ_{avg}) of the proposed approach compared with TONE [6]

	PSNR(dB)	Bitrate	Perf.	Power	θ_{avg}
Min	**+0.3**	**-3%**	**+8%**	**-1%**	**-8%**
Max	**+0.8**	**-10%**	**+16%**	**-10%**	**-14%**
Average	**+0.7**	**-7%**	**+11%**	**-4%**	**-12%**

3 EXPERIMENTAL RESULTS

We evaluate the proposed ML-based approach in comparison with TONE [6], the most similar work to ours, by running experiments on an enterprise server while monitoring performance, power, temperature, and coding efficiency. In this work we use standard test sequences and the reference software HM 16.3 [1]. We assume a realistic case where streams are randomly released on cores and go away, simulating YouTube or Netflix servers. Table 1 shows PSNR, deviation from the target bitrate, performance, power consumption, and average temperature. TONE, unaware of the available potentials due to changes in the number of videos being processed, fails to increase the performance while maintaining the desirable encoding quality and meeting the power and thermal constraints.

4 FUTURE WORK

In the future we will investigate how to exploit more application-level knobs, as well as analyzing the impact of content variation in the memory sub-system, to further improve encoding efficiency and performance. In addition, we will explore new stream allocation and migration strategies to target multistreaming on MPSoCs in the context of a wider set of realistic scenarios.

ACKNOWLEDGMENT

This work has been partially supported by the EC H2020 MANGO project (GA No. 671668).

REFERENCES

[1] Philippe Bordes, Pierre Andrivon, Franck Hiron, Philippe Salmon, and Ronan Boitard. 2016. Joint Collaborative Team on Video Coding (JCT-VC) of ITU-T SG 16 WP 3 and ISO/IEC JTC 1/SC 29/WG 11. (2016). https://HEVC.hhi.fraunhofer.de
[2] Frank Bossen, Benjamin Bross, Karsten Suhring, and David Flynn. 2012. HEVC complexity and implementation analysis. *IEEE TCSVT* 22, 12 (2012), 1685–1696.
[3] Cisco Systems, Inc. 2016. Cisco Visual Networking Index: Forecast and Methodology 2015-2020. Cisco Whitepaper. (2016).
[4] Guilherme Correa, Pedro Assuncao, Luciano Agostini, and Luis A Silva Cruz. 2016. Complexity scalability for real-time HEVC encoders. *Journal of Real-Time Image Processing* 12, 1 (2016), 107–122.
[5] Daniel Palomino, Muhammad Shafique, Hussam Amrouch, Altamiro Susin, and Jorg Henkel. 2014. hevcDTM: Application-driven dynamic thermal management for high efficiency video coding. In *DATE, 2014*. IEEE, 1–4.
[6] Daniel Palomino, Muhammad Shafique, Altamiro Susin, and Jörg Henkel. 2014. TONE: Adaptive temperature optimization for the next generation video encoders. In *Proc. of the 2014 ISLPED*. ACM, 33–38.
[7] Daniel Palomino, Muhammad Shafique, Altamiro Susin, and Jörg Henkel. 2016. Thermal optimization using adaptive approximate computing for video coding. In *DATE, 2016*. IEEE, 1207–1212.
[8] Muhammad Shafique and Jofkrg Henkel. 2014. Low power design of the next-generation high efficiency video coding. In *ASP-DAC, 2014 19th Asia and South Pacific*. IEEE, 274–281.