Energy-Efficient NoC-Based Systems for Real-Time Multimedia Applications using Approximate Computing

Wagner I. Penny[12], Daniel M. Palomino[1], Marcelo S. Porto[1], Bruno Zatt[1]

[1]Video Technology Research Group (Vitech) – Programa de Pós-Graduação em Computação (PPGC) – Universidade Federal de Pelotas (UFPel) – Pelotas – RS – Brazil

[2]Instituto Federal Sul-rio-grandense (IFSUL) – Campus Pelotas – Pelotas – RS – Brazil

{wi.penny, dpalomino, porto, zatt}@inf.ufpel.br

Abstract. *This work presents an energy-efficient NoC-based system for real-time multimedia applications employing approximate computing. The proposed video processing system, called SApp-NoC, is efficient in both energy and quality (QoS), employing a scalable NoC architecture composed of processing elements designed to accelerate the HEVC Fractional Motion Estimation (FME). Two solutions are proposed: HSApp-NoC (Heuristc-based SApp-NoC), and MLSApp-NoC (Machine Learning-based SApp-NoC). When compared to a precise solution processing 4K videos at 120 fps, HSApp-NoC and MLSApp-NoC reduce about 48.19% and 31.81% the energy consumption, at small quality reduction of 2.74% and 1.09%, respectively. Furthermore, a set of schedulability analysis is also proposed in order to guarantee the meeting of timing constraints at typical workload scenarios.*

1. Introduction

Over the past few years, video content has been widely spread at the consumer market, as a result of the fast popularization of social media and video streaming services. Considering our current life context, seriously affected by the worldwide COVID-19 pandemic, such expansion has gained remarkable prominence. For instance, social media content broadcasting has grown 32.6% since the pandemic has been started [Statista 2020]. Besides, according to Cisco (2020), in 2022 up to 90% of all global Internet traffic will be related to video broadcasting and 70% of this sharing will be handled by mobile devices. New features as higher resolutions and frame rates required by current applications have demanded ever-increasing compression ratios. The High Efficiency Video Coding (HEVC) [Sze, Budagavi and Sullivan 2014] standard has been released in 2013, replacing older standards, as an alternative to improve the overall delivered compression ratio and support UHD (Ultra-High Definition) resolutions. Indeed, HEVC is capable of reaching up to 40% higher compression efficiency when compared with the previous H.264 encoder. However, such enhancement makes HEVC up to 500% more complex than H.264 [Vanne et al. 2012], which leads to higher energy consumption, raising new challenges for systems design, especially when dealing with battery-powered devices.

In general, complex signal processing applications like video coding are suitable for parallelism exploration. Parts of the video encoder have enormous opportunities for parallelization, such as the Fractional Motion Estimation (FME), which is responsible for up to 60% of total HEVC encoding effort [Grellert, Bampi and Zatt 2016]. Systems-on-Chip (SoCs) built upon a Network-on-Chip (NoC) infrastructure arise as an alternative to allow exploiting this inherent application parallelism. Packet switched NoCs are considered as scalable interconnections to support the growing communication demands of SoC components [Bokhari et. al. 2015], while allowing heterogeneous processing elements, tiling and partial power/clock gating that enable performance/energy scalability. Scalability is desirable for systems supporting distinct throughput demands such as multiple video resolutions and frame rates.

Furthermore, the usage of GPPs (General-Purpose Processors) as a processing element (PE) may not give real-time performance neither attend energy consumption/power dissipation constraints or high QoS (Quality-of-Service) levels, demanded by real-time multimedia systems. Indeed, considering their severe constraints, especially for embedded battery-powered systems, dedicated hardware acceleration becomes mandatory. Moreover, in order to seek for greater energy efficiency, researchers have been allying the hardware acceleration design to the usage of approximate computing, which has been seen as an alternative to improve performance/energy efficiency by compromising (in acceptable ranges) the quality of the applications [Venkataramani et al. 2015]. Approximate computing exploits the intrinsic error resilience of applications to reach improvements in performance or energy efficiency. Generally, multimedia processing is a suitable application to apply approximate computing since the resilience of the human visual system (HVS) to errors can be explored, by compromising the quality of the application in tolerable ranges, aiming for the computational effort/power dissipation reduction.

The main goal of this work is to research and propose architectural solutions that lead to energy-efficient and high-performance application-specific systems with scalable real-time support for error-resilient applications. We introduce the solution called *Scalable Approximate Network-on-Chip* (SApp-NoC): an energy-efficient real-time multimedia system built on top of a NoC employing hardware acceleration using multiples levels of approximate computing and leveraging application-specific properties/behavior through the use of heuristics and machine learning techniques. Two SApp-NoC were developed: *Heuristic-based SApp-NoC* (HSApp-NoC) and *Machine Learning-based SApp-NoC* (MLSApp-NoC).

2. Scalable Approximate Network-on-Chip (SApp-NoC)

In order to achieve the main goal of this work, we follow a strategy exploiting *parallelism* to provide scalability and performance improvement, exploiting *hardware acceleration* and *approximate computing* to provide high-performance with energy efficiency, and exploring *application-specific behavior* in order to keep reasonable QoS.

The parallelism is exploited by building the solution onto a NoC-based system, capable to explore the application inherent communication parallelism and to provide scalability across different demanding throughput. Approximate computing is exploited by the development of hardware accelerators employing the simplifications at different levels of approximation. Furthermore, the application QoS is kept at reasonable values by leveraging application-specific behavior, exploring error-resilient areas/steps by the employment of heuristics and machine learning-based solutions. All these assumptions lead to the development of SApp-NoC – used as our main case study and focusing on the HEVC FME step. SApp-NoC is presented in Figure 1 (a), as well as an overview of the main contributions of this work. We have tagged in Figure 1 (a) each contribution with a small circle having the corresponding chapter/section number where such a contribution is detailed (from the complete thesis manuscript). Furthermore, Figure 1 (b) shows the adopted methodology framework of this work.

Our NoC is organized in multiple neighbor Tiles to allow the NoC effective size to scale according to throughput requirements, i.e., the number of active processing elements varies according to video resolution and frame rate. The processing elements of the NoC are based on hardware acceleration, which could accelerate any application step, but in our case study targets the HEVC FME, presenting multiple levels of approximation. Hardware accelerators were named Approximate FME Filters – $FAPP_j$, where j = [0, 1, 2, 3], varying the approximation level from precise (0) to most aggressive (3). The hardware design takes into account the application behavior, focusing on low QoS degradation. At design time, algorithms are proposed to define the PE type (i.e., the approximation level), amount and placement. Finally, at run-time, tasks are smartly allocated on SApp-NoC, following application behavior-based statistics.

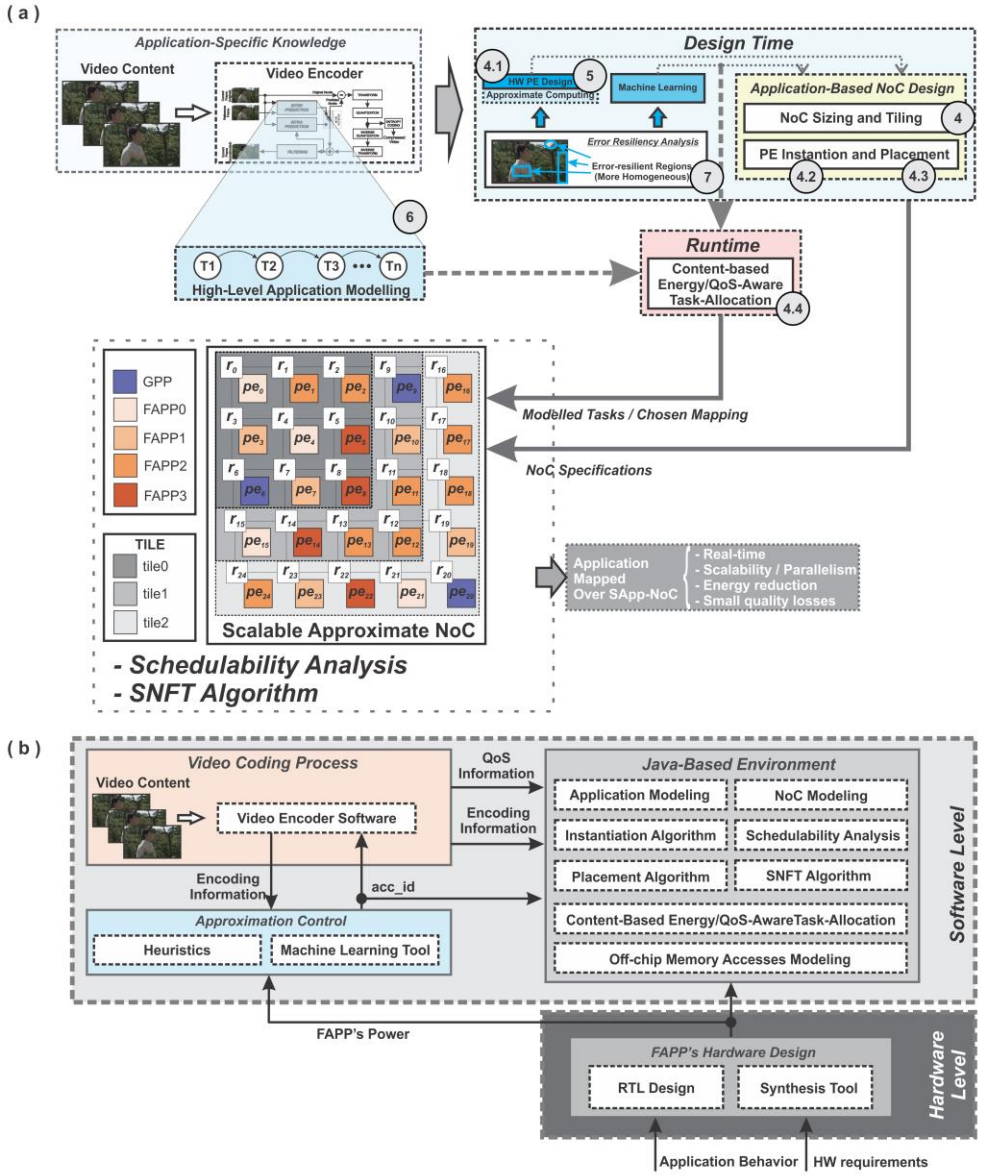

Figure 1. Detailed methodology: (a) overview of the novel contributions of this work and (b) adopted framework.

All nodes of SApp-NoC are interconnected, and each one has a processing element *PE*, linked internally to a local cache, which stores local information, and a router *r*, which routes the data packets towards their destinations. The communication between processing elements and the router is made by two unidirectional links (one from *PE* to *r* and other from *r* to *PE*). In this work we have applied the widely used square-shaped 2D-mesh topology, considering wormhole NoC with priority-preemptive arbitration, widely studied in the literature due to its ability to provide resources for hard real-time guarantees. In order to determine whether application tasks being executed and communicating over SApp-NoC can fulfill the required timing constraints, schedulability analysis [Indrusiak, Burns and Nikolic 2018] was applied.

In Figure 1 (b) we present the adopted development framework. Our contributions are present at both software and hardware levels. At hardware level, we follow a Register Transfer Level (RTL) design, making the FME hardware development and employing a synthesis tool to estimate the power and chip area. At software level, the video encoder reference software is employed to analyze the behavior of the proposed simplifications, verifying the resulted QoS information for each proposed scenario. In addition, encoding information are also gathered from the encoder software, in order to allow the usage of approximation control. Such a control

is also performed at software level by employing heuristics and a machine learning tool. The generated approximation control algorithm is inserted in the encoder software code, also aiming the analysis of the QoS behavior. All the other contributions are made employing a Java-based environment, were the application (HEVC FME), the NoC, and the off-chip memory communication are modeled in high level of abstraction. All the other proposed algorithms and the schedulability analysis are performed considering the same Java-based environment.

3. Results and Discussion

In this work, different sorts of experiments were performed to enable the whole system modeling and achieve the final results. In this section we present the obtained results and a discussion of our main contributions.

On the one hand, QoS results were obtained using the main profile of the HEVC reference software, HM 16.18 [Boyce 2014], according to recommendations of video community. Our simulations reproduced the behavior of the approximate hardware accelerators being selected by heuristics (HSApp-NoC) and by a decision tree (MLSApp-NoC), in order to check the impacts on coding efficiency. On the other hand, the PEs hardware design were described in VHDL and synthesized using a 40 nm TSMC standard cell library with 0.9V using the Cadence RTL Compiler tool. To perform the power estimation of the developed hardware, we have employed the default tool switching activity (20%). Note that the synthesis of each FAPP was made individually, therefore, when analyzing the power of the PEs of SApp-NoC we need to consider the number of instantiated PEs and their working behavior.

The energy estimation of SApp-NoC considers the power of each PE and the time each one is *active*, or *idle*, although the power gating control was not physically implemented (since the modeling of SApp-NoC is made at a high level of abstraction), we consider the usage of power gating over each *idle* PE, disregarding its power dissipation when not active. Note that we considered energy consumption evaluation only of the processing elements, without taking into account the energy consumed at the communication across the NoC.

Table 1 presents HSApp-NoC and MLSapp-NoC synthesis results, making a comparison with related works that also propose ASIC hardware accelerators designs for the HEVC FME. He et al. (2015) presented a power-efficient architecture for HEVC FME with power dissipation of 198.6 mW and throughput of 4K@120fps. In Lung and Shen (2019) it is proposed a VLSI architecture and implementation of the HEVC FME. Their design achieves a maximum throughput of 4K@39fps, with a power dissipation of 304.8 mW. When comparing the normalized FME power of the related works with our solutions, it can be noticed that our solutions presented smaller power dissipation.

When analyzing the energy results we need to consider the average FAPP selection at HSApp-NoC and MLSApp-NoC, in terms of how many times each FAPP is selected, considering the four recommended QPs, for each video sequence, in order to find the *active* time of each PE. When comparing with the energy consumed by the precise solution ($FAPP_0$), HSApp-NoC is able to save up to 48.19% in energy consumption whereas MLSApp-NoC saves up to 31.81%. Figure 2 shows the QoS results for each video class, regarding the individual FAPPs, HSApp-NoC and MLSApp-NoC. $FAPP_3$ presents the worse results, since it employs a more aggressive approximation. Regarding this scenario, class B presented the worse QoS among all classes being executed by $FAPP_3$. Such a behavior happens due to the nature of some videos, considered as high complexity sequence (HCS) and presenting the highest impacts on

Table 1. SApp-NoC synthesis results and comparison with related works.

Parameter	[He et al. 2015]	[Lung & Shen 2019]	HSApp-NoC	MLSApp-NoC
Power (mW)	198.6	304.8	60.30	47.34
Norm. Power (mW)	91.94	101.6		
Max. Throughput	4K@120fps	4K@39fps	4K@120fps	4K@120fps
Technology (nm)	65	90	40	40

QoS when approximation is applied. In fact, when compared to solutions that cannot adapt to the content, HSApp-NoC and MLSApp-NoC drastically reduce QoS degradation in relation to $FAPP_3$ (class B makes the QoS difference clearer), with MLSApp-NoC performing close to $FAPP_1$ and $FAPP_2$ in terms of QoS, while reducing the energy consumption, since only homogeneous regions of the frame are explored. In average, MLSApp-NoC presents a QoS of 1.09% while HSApp-NoC presents 2.74%.

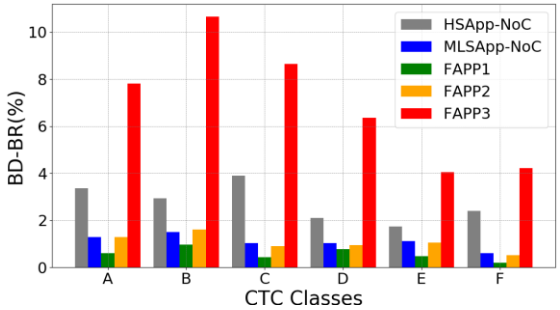

Figure 2. QoS results in terms of BD-BR for each class.

4. Main Achievements: Awards, Collaborations and Publications

Firstly, this PhD thesis was meaningfully improved by the opportunity allowed by an international collaboration, performed during the sandwich doctorate, with the Real-Time Systems Group of the University of York, UK, with the supervision of professor Leandro Indrusiak. Additionally, the work developed in this PhD thesis was awarded with the *Best Poster Award* in the *10th IEEE CASS Rio Grande do Sul Workshop*. Furthermore, it has also led to other important publications, listed below.

Direct contributions of this thesis:

- **W. Penny**, D. Palomino, M. Porto, B. Zatt. *Power/QoS-Adaptive HEVC FME Hardware using Machine Learning-Based Approximation Control*. In: VCIP 2020.
- **W. Penny**, G. Correa, L. Agostini, D. Palomino, M. Porto, G. Nazar, B. Zatt. *Low-Power and Memory-Aware Approximate Hardware Architecture for Fractional Motion Estimation Interpolation on HEVC*. In: ISCAS 2020.
- **W. Penny**, D. Palomino, M. Porto, B. Zatt, L. Indrusiak. *Design Space Exploration of HEVC RCL Mapped onto NoC-Based Embedded Platforms*. In: ReCoSoC 2019.
- **W. Penny**, D. Palomino, M. Porto, B. Zatt, L. Indrusiak. *Performance evaluation of HEVC RCL applications mapped onto NoC-based embedded platforms*. In: SBCCI 2019.
- **W. Penny**, M. Ucker, I. Machado, L. Agostini, D. Palomino, M. Porto, B. Zatt. Power-Efficient and Memory-Aware Approximate Hardware Design for HEVC FME Interpolator. In: ICECS 2018.
- **W. Penny**, J. Goebel, G. Paim, M. Porto, L. Agostini, B. Zatt. *High-throughput and power-efficient hardware design for a multiple video coding standard sample interpolator*. In: JRTIP 2018.
- G. Paim, J. Goebel, **W. Penny**, B. Zatt, M. Porto, L. Agostini. *High-throughput and memory-aware hardware of a sub-pixel interpolator for multiple video coding standards*. In: ICIP 2016.
- G. Paim, **W. Penny**, J. Goebel, V. Afonso, A. Susin, M. Porto, B. Zatt, L. Agostini. *An efficient sub-sample interpolator hardware for VP9-10 standards*. In: ICIP 2016.

Indirect contributions of this thesis:

- R. Domanski, J. Goebel, **W. Penny**, M. Porto, D. Palomino, B. Zatt, L. Agostini. *High-Throughput Multifilter Interpolation Architecture for AV1 Motion Compensation*. In: TCAS-I 2019.

- **W. Penny**, J. Goebel, D. Correa, A. Martins, G. Nazar, L. Agostini, M. Porto, B. Zatt. *Energy-Efficiency Exploration of Memory Hierarchy using NVMs for HEVC Motion Estimation*. In: ICECS 2019.
- A. Martins, **W. Penny**, M. Weber, L. Agostini, M. Porto, D. Palomino, J. Mattos, B. Zatt. *Configurable Cache Memory Architecture for Low-Energy Motion Estimation*. In: ISCAS 2018.
- I. Machado, **W. Penny**, M. Porto, L. Agostini, B. Zatt. *Characterizing Energy Consumption in Software HEVC Encoders: HM vs x265*. In: LASCAS 2017.
- A. Martins, **W. Penny**, M. Weber, D. Palomino, J. Mattos, M. Porto, L. Agostini, B. Zatt. *Cache Memory Energy Efficiency Exploration for the HEVC Motion Estimation*. In: SBESC 2017. (*Obs.: BEST PAPER AWARD of the conference*)
- **W. Penny**, I. Machado, M. Porto, L. Agostini, B. Zatt. *Pareto-based energy control for the HEVC encoder*. In: ICIP 2016.

Besides the direct contributions of this thesis, present in several publications at qualified vehicles, other supported works, which have led to the indirect contribution publications, were also very important, since it involved the partnership/co-advisement of undergraduate students like *Italo Machado*, *Mariana Ucker*, and *Douglas Correa*; as well as the partnership with master/PhD colleagues *Anderson Martins*, *Jones Goebel* and *Robson Domanski*, leading to a prospective and collaborative work within the research group.

5. Conclusions

This work presented a summary of the main achievements of the thesis, presenting the exploration of energy-efficient NoC-based systems for real-time multimedia applications using approximate computing. We proposed an energy/QoS-aware video processing system featuring scalable NoC topology and multi-level approximate hardware acceleration, called SApp-NoC, proposing a scalable NoC architecture targeting heterogeneous processing elements, designed to accelerate the HEVC FME exploring application properties to smartly define the multiple levels of approximation, using heuristics (HSApp-NoC) and machine learning (MLSApp-NoC).

References

Bokhari, H. et al. (2015). SuperNet: Multimode interconnect architecture for manycore chips. In *ACM/IEEE DAC.*

Boyce, J. (2014), HM16: High Efficiency Video Coding Test Model (HM16) Encoder Description, JCTVC-R1002, Sapporo.

Cisco (2020) "Cisco Annual Internet Report (2018–2023) White Paper", https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html, June.

Grellert, M., Bampi, and Zatt, B. (2016). Complexity-scalable HEVC encoding. In *IEEE PCS*.

He, G. et al. (2015). High-throughput power-efficient VLSI architecture of fractional motion estimation for ultra-HD HEVC video encoding. In *IEEE TVLSI*, pages 3138-3142.

Indrusiak, L., Burns, A. and Nikolic, B. (2018). Buffer-aware bounds to multi-point progressive blocking in priority-preemptive NoCs. In *IEEE DATE.*

Lung, C. and Shen, C. (2019). Design and implementation of a highly efficient fractional motion estimation for the HEVC encoder. In *JRTIP*, pages 1541-1557.

Statista (2020) "Coronavirus impact on online traffic of selected industries worldwide in week ending April 26, 2020", http://www.statista.com/statistics/1105486/coronavirus-traffic-impact-industry, June.

Sze, V., Budagavi, M. and Sullivan, G. (2014), High efficiency video coding (HEVC), Springer, USA, 1st edition.

Vanne, J. et al. (2012). Comparative Rate-Distortion-Complexity Analysis of HEVC and AVC Video Codecs. In *IEEE TCSVT*, pages 1885-1898.

Venkataramani, S. et al. (2015). Computing approximately, and efficiently. In *IEEE DATE.*