Journal of Integrated VLSI and Signal Processing • ISSN 3139-2369 (O) • Jan - Apr • 1 • 1 • 2026 (pp. 9-17)

RESEARCH ARTICLE

Energy-Efficient VLSI Hardware Accelerators for Real-Time Image and Multimedia Signal Processing Applications

K. Geetha*

Professor of Computer Science and Engineering, Excel Engineering College, Erode

KEYWORDS:
Energy-efficient VLSI,
Hardware accelerators,
Real-time image processing,
Multimedia signal processing,
Low-power architecture design,
Embedded systems

ARTICLE HISTORY:

Received : 09.11.2025
Revised : 12.12.2025
Accepted : 18.01.2026

Abstract

Image and multimedia signal processing applications that require processing in real time are also being deployed on energy-restrained embedded and edge systems, where traditional CPU- and GPU-based application choices are no longer capable of satisfying hard power and latency constraints. Specialised VLSI hardware accelerators offer an efficient alternative, but it is difficult to attain a high level of energy efficiency with a requirement of real-time performance with such high overheads of data movement and memory access. The paper introduces a unified energy and performance co-design methodology, which was used to develop an energy-efficient VLSI hardware accelerator of the real-time image and multimedia signal processing. A computation, memory, and control energy analytical model to system level is used to make architectural decisions with explicit throughput assuming constraints. The accelerator proposed combines an array of parallel processing elements, hierarchical on-chip memory and a dataflow adaptable control to limit off-chip accesses. The minimization of data reuse and low-power design methods is another method to save energy. Representative image and multimedia kernel experimental evaluation shows higher energy efficiency and through put than state of the art designs with deadlines being met.

Author's e-mail: kgeetha.eec@excelcolleges.com

How to cite this article: Geetha K. Energy-Efficient VLSI Hardware Accelerators for Real-Time Image and Multimedia Signal Processing Applications. Journal of Integrated VLSI and Signal Processing. Vol.1, No. 1, 2026 (pp. 9-17).

Introduction

The advent of fast image and multimedia signal processing functions, in real-time, has highly altered contemporary embedded or edge computing systems.[1-3, 9] New applications like smart surveillance, self-driving, augmented reality and virtual reality, medical imaging, and multimedia streaming are vulnerable to high computations workload with severe energy and latency requirements.[1, 6, 9] These systems workloads are typified by data-intensive workloads, complicated arithmetic kernels, and hard real-time requirements, and therefore an energy-efficient processor is an experimental necessity in current systems-on-chip (SoCs).[3, 8] Traditional general-purpose processing platforms such as CPUs and GPUs are programmable and can execute most tasks with the highest peak performance but have low energy efficiency when rolled to power-constrained environments.[10, 11] These platforms are not applicable to real-time multimedia processing in embedded systems since they are expensive in energy and have lengthy control overhead and off-chip memory reads and writes.[11] This has seen specialised VLSI-based accelerators become a potentially viable solution, taking advantage of parallelism, customised data paths, and application-specific optimizations to provide a better performance to watt.[4, 6, 8] Alongside the mentioned benefits, a question of how to design energy-saving VLSI accelerators to execute image and multimedia processing is always a problematic issue. Contemporary workloads are characterised by a wide range of computational behaviour and an unpredictable data access tendency, which causes a large amount of energy wasted on the movement of memory, but not the process of computation.[4, 12] Moreover, intensive power optimization might lead to an increase in the power dissipation and thermal load, and the excessive power optimization might breach the real-time requirements.[8, 11] A combination of all three factors, energy efficiency and real-time performance, dictates the need to consider them as a unified design and consider computation, memory hierarchy and

dataflow as a unit .[4, 5] This paper deals with these issues by offering a comprehensive approach to co-designing energy and performance of VLSI hardware accelerators to satisfy real time image and multi-media signal processing applications. Explicit throughput constraints are used to develop a system-level analytical model that captures the components of computation, memory and control energy .[5, 11] Scalable accelerator architecture, based on this paradigm, is developed with a parallel processing array of elements in the arrays, hierarchical on-chip memory, and dataflow-level control.[4, 6, 7] The low-power design technique and data reuse technique are used to optimise the architecture further since it minimises off-chip access and unnecessary switching activity.[4, 8, 12] The paper presents a co-design approach to energy and performance of VLSI hardware accelerators in a single methodology with the objective of image and multimedia signal processing applications in real-time. An all-inclusive analytical framework is formulated to approximate calculation, memory, and control energy and clearly takes into consideration real-time throughput limitations .[5] According to this framework, the energy efficient accelerator architecture is determined and optimised with the help of representative image and multimedia processing kernels in terms of parallel processing, hierarchical memory structure and the dataflow-conscience control,[1-4], and.[6] The suggested architecture is verified by means of a comprehensive experimental analysis and presents significant enhancement of energy efficiency and throughput utilising in the comparison of the state-of-the-art accelerator architectures.[4, 6, 10] In general, the presented solution provides a viable and scalable solution to the next-generation embedded multimedia systems that require high-performance with the strict energy constraints.[1, 9]

Related Work

The development of energy efficient VLSI hardware acceleration in real-time image and multimedia signal processing has remained a concern that has been largely researched due to the increasing demand of a high rate of processing under a stringent power requirement. [1, 3, 9] In the early days, kernel-specific accelerators were mainly based on the computationally intensive modules like movement estimation, discrete cosine transform (DCT), fast fourier transform (FFT) and spatial filtering.[1–3, 8] These designs showed that specialised data interfaces and parallel architectures can be much more energy efficient and consuming less latency than the general-purpose processors.[4, 8] Most early accelerators were however heavily specialised to one or two particular kernels, so that they could only be applicable to a wide range of multimedia workloads.[1, 2] Later studies focused on architectural and algorithmic co-design in order to enhance performance per watt.[4, 5] Video encoding and motion estimation accelerators, which used algorithm level simplifications and parallel processing plans adapted speech in order to maintain real time functionality with less power.[1–3, 6] On the same note, FFT accelerators and transform-domain accelerators considered pipelining, radix optimization and tuning word-length with respect to maintaining a trade-off between throughput and power consumption.[2, 8] Although the remarkable gains were achieved by these methods, they frequently depended on high parallelism or aggressiveness of clock frequencies, both of which added to the energy and overhead of a memory access.[11]

More current accelerator architectures, such as convolution and array-based processing engines, have pointed out that data movement and memory access take the largest

Table 1: Comparison of State-of-the-Art VLSI Accelerators

Work (Representative)	Target Application / Kernel	Technology Node	Throughput	Energy Metric	Key Limitations
Motion estimation accelerator	Video motion estimation	65 nm CMOS	Real-time HD/ UHD (configu-rable)	High GOPS/W at low voltage	Highly kernel-specific; limited generality
HEVC fractional ME engine	Video encod-ing (FME)	45 nm ASIC	2160p @ 60 fps	Reduced pow-er vs prior art	Bitrate loss due to algorithm simplification
Split-radix FFT accelerator	FFT / trans-form process-ing	65 nm CMOS	Application-de-pendent	Improved efficiency vs radix-4	Transform-only design
Hybrid FFT archi-tecture	1024-point FFT	65 nm CMOS	~400 MHz	nJ-level energy per transform	Floating-point overhead; limited reuse
Convolution pro-cessing array	Image/CNN convolution	65 nm CMOS	Tens of GOP/s	TOPS/W-level efficiency	Primarily CNN-focused; not multimedia-ge

(often much larger than computation) portion of energy consumption.[4, 12] This has led to the optimization of memory hierarchies, data reuse and dataflow-conscious scheduling playing a key role in generating high levels of energy efficiency.[4, 5, 8] Even with these developments, actual designs are often either transform-based multimedia processing or convolution-based workload, but seldom both, with a focus on a single methodology which concurrently captures energy, throughput, and real-time constraints on heterogeneous image and multimedia kernels.[4, 9] The major state-of-the-art VLSI accelerators are summarised in Table 1, with comparisons made on their technology node, their throughput and their energy efficiency measures, and the main constraints. The comparison indicates that although much has been done in regard to the enhancement of performance per watt, performance accelerators are still application-specific lacking a generalised energy-performance co-design framework that will generalise across real time applications in image and multimedia processing.[5, 11] It is out of this gap that the unified modelling and architecture is put forward in this work.[4, 5]

Methodology

System Model and Workload Characterization

This work will focus on real time image and multimedia signal processing applications that run on embedded and edge computing platforms that are energy constrained. The system model presupposes that the system-on-chip (SoC) architecture is heterogeneous with a dedicated VLSI hardware accelerator that runs in parallel with a host processor and offloads computationally demanding signal processing kernels, with a rigid set of real-time requirements. The accelerator is designed as a frame -based way to handle streaming data, where both the latency and throughput rules are determined by application-level frame rates. The activities to be considered include representative image and multimedia processing kernels which are common in real-world systems, i.e. 2D spatial convolutions, transform-based operations like discrete cosine transform (DCT) and fast fourier transform (FFT), and block-based motion estimation. High arithmetic intensity, patterned pattern of computation and repeated access to data characterise these kernels, and hence, are suited well to being accelerated by hardware. Nevertheless, they also demonstrate a great rates of data transfer between processing units and memory affecting considerably the total energy consumption.

Mathematically, every workload may be modelled as a series of multiply-accumulate and addition operations being applied on data blocks or sliding windows of predefined sizes. Count of operations per frame depends on the resolution of input, kernel size and the structure of algorithm. Also, in real time operation, frame processing time has hard deadlines often in terms of frames per second or maximum possible latency per frame. Violation of these limits will lead to lost frames or quality of service reduction. Memorizing wise, the workloads have frequent accesses to input data, intermediate buffers and coefficient storage. Assumptions To minimise on-chip memory accesses, it is assumed that off-chip memory accesses are much more costly in terms of energy consumption compared to on-chip memory accesses, and thus aggressive reuse of data and locality should be published by the accelerator. The system model is therefore focused on focusing of reducing on off-chip communication and focusing on the on chip storage utilisation. This characterization of workloads is the foundation of the model of analytical energy and performance, which will be presented in the next subsection and the direct focus of the architectural and dataflow optimizations in the proposed accelerator design.

Energy and Performance Modelling.

In order to inform the design of energy efficient hardware accelerator which can respond to real time consideration, system level analytical model is formulated to represent energy consumption and performance jointly. The model explicitly models computation, memory access and control overhead, which are individually predominant in the energy behaviour of VLSI signal processing accelerators. Processing an individual frame is modelled at the overall energy consumed as like computation energy plus memory access energy and energy related to control as shown in Equation (1):

$$E_{total}=E_{comp}+E_{mem}+E_{ctrl} \quad (1)$$

Here, E_{comp} represents the energy associated with arithmetic operations executed by the processing elements, E_mem captures the energy cost of on-chip and off-chip memory accesses, and E_ctrl accounts for control logic, interconnect, and clocking overhead. This model emphasises the significance of reducing the memory traffic, where the memory energy can be higher than the one of computation in data-intensive workloads. Real time processing requirements put a limitation on the maximum acceptable frame processing latency performance. Specifically, frame processing process is represented as dependent on the total number of operations, frequency of operation and parallelism level in accelerator as in Equation (2):

$$T_{frame}=\frac{N_{ops}}{f_{clk}\cdot P_{parallel}} \leq T_{deadline} \quad (2)$$

where N_{ops} denotes the total number of operations required per frame, f_clk is the clock frequency, P_parallel

represents the number of parallel processing elements, and T_deadline is the maximum allowable latency dictated by the target frame rate. This limitation will make the architectural decisions maintain parallelism and frequency scaling in order to meet the demands of real time without excessive power consumption. In order to measure the effectiveness of the proposed design quantitatively, the energy efficiency is taken to be the ratio of the useful computation to the total consumption of energy or as stated in Equation (3):

$$\eta = \frac{N_{ops}}{E_{total}} \text{(Operations per joule)}, \qquad 3)$$

This measure is the main automatization target, as it will allow comparison with the latest accelerators. A combination of Equations (1) -(3) forms a single framework that correlates the workload characteristics, architectural parameters, and energy-performance trade-offs and they have a direct influence on the accelerator architecture and optimization strategies that are described further below.

Proposed Accelerator Architecture that is Energy-Efficient.

The proposed architecture of energy-efficient VLSI accelerator is set to provide an excellent provision of high throughput in real-time image and multimedia signal processing with small consumption of energy due to the architectural specialisation and dataflow-oriented optimization. The accelerator is designed with a modular and scalable sales and mapping of various signal processing of signal processing kernels with tight real-time requirements is made possible with ease. As shown in Figure 1, there are four key components in the architecture, direct memory access (DMA) interface, hierarchical on-chip memory subsystem, parallel processing element (PE) array, and centralised control unit. The DMA interface is the main communication interface between off-chip memory and the accelerator which allows with high bandwidth the exchange of data and less intervention by the host processor. This design option has minimal control overheads and is more economical to the overall system energy.

To take advantage of data locality and result in fewer costly off chip memory accesses, the hierarchical on-chip memory subsystem is structured into numerous levels of buffers. The information about frequently accessed input data, the intermediate results and filter coefficients are stored in local buffers near the PE array, which allows quickly re-use and reduces the energy required by memory access. The architecture allows dropping the size of buffers and access patterns in line with the target workloads to enable much less data flow, which is the leading cause of overall energy usage. The fundamental calculation is carried out using a parallel PE array of lightweight custom arithmetic units specialised to multiply accumulate and addition arithmetic operations common to image and multimedia processing kernels. PE array has configurable parallelism, where the accelerator can be scaled to different performance levels in case of workload demands. Local interconnects ensure the machine has low latency data transfer between PE, and additionally enhance the throughput without incurring the energy costs incurred by global buses.

The data movement, scheduling and synchronisation over the accelerator is coordinated by a centralised unit of control. It applies dataflow-sensitive control policies that work to synchronise accesses to data and computations stages in order to optimise resource use and reduce dead air switching. Moreover, the architecture incorporates low power design features including clock gating, operand isolation on a module-by-module basis that makes idle

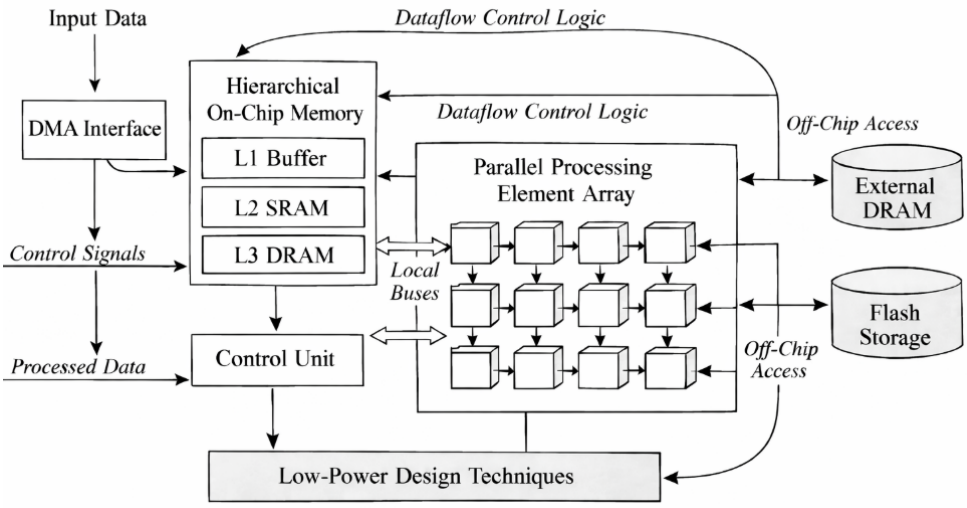

Fig. 1: Block Diagram of the Proposed VLSI Accelerator

components use the minimum amount of energy. In general, the architecture of the architecture presented in Figure 1 is a more integrated energy-performance co-design, with parallel computer computation, hierarchical memory organisation, and smart control, to be executed at a high energy efficiency with simultaneous satisfaction of real-time processing deadlines in embedded multimedia systems.

Dataflow optimization and Low-Power Design Techniques.

In the case of real time Image and multimedia signal processing VLSI accelerators, the use of effective dataflow organisation and severe minimalization of data transfer is essential in developing high energy efficiency. In the design suggested, low-power design technologies are closely interacted with dataflow optimization in order to design solutions with reduced energy consumption in terms of memory access, and with certain high-computational throughput. The accelerator, as shown in Figure 2, uses dataflow and memory reuse approach that takes advantage of spatial and temporal locality of image and multimedia workloads. Input frames are broken down into spatial blocks, which are transmitted to on-chip buffers so as to reuse input data and philtre coefficient many times in different operations. The tiling method can greatly decrease unnecessary accesses to off-chip memory which have been known to be orders of magnitude more expensive than on-chip accesses.

To achieve even greater energy saving, the architecture uses on-chip double buffering on the L1 memory level, which enables an overlap of data transfer and computation in the Ping-Pong way. As one of the buffers receives an incoming share of data in an off-chip memory, the other buffer feeds data to the processing element array to be calculated. Such load-compute-store overlap can successfully hide the memory latency and makes better use of the processing elements without raising the clock rate and expenditure of power. The elements array of parallel processing elements is structured in such a way that it facilitates localised interchange of data by use of low-capacitance interconnects. All processing units receive data through local registers files, and on-chip shared buffers through local buses, to reduce worldwide communications. This localised dataflow also reduces switching activity on the long interconnects as well as reduces dynamic power dissipation. Coefficient tiles are stored on-chip in a number of spatial tiles, which additionally improves reuse and minimises memory traffic. In conjunction with the dataflow optimization, other low-power design techniques are also put in place on the design side. Clock gating on fine-grained basis is also used to turn off idle processing units and memory blocks when they are idle and isolating operands in arithmetic units to avoid unnecessary transitions on a signal. With these methods, dynamic and static power consumption are minimised without affecting real-time performance. In general, the dataflow/memory reuse plan in Figure 2 is core in synchronising patterns of the computation and memory access with the hypothetical energy-performance model, that allows the accelerator to achieve high energy efficiency and continually meet real-time processing deadlines.

Implementation and Experimental Methodology.

The VLSI accelerator proposed is synthesised and at a contemporary CMOS technology node to more accurately represent the underlying environment of embedded deployment adoption. The architectural parameters like processing elements number, data precision, memory hierarchy structure, and osqperating frequency are the parameters that are chosen according to the MODE of analysis, which includes the study of energy-performance presented in Section 3.2, which ensures the establishment of reasonable trade-off among throughput, energy efficiency and the constraints of real-time. Table 2:

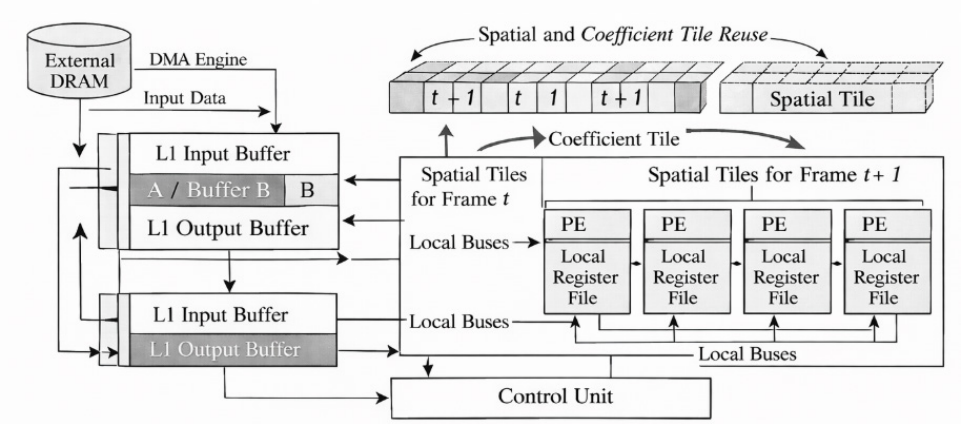

Fig.2: Dataflow and Memory Reuse Strategy

The design parameters that are important to future implementation of the accelerator are summarised: the number of processing elements, the word length, the size of the on-chip memory, and the clock frequency, which are all part of the base parameters of the setup of the accelerator.

The performance of the accelerator is measured on a representative range of image and multimedia signal processing functions such as two dimensional spatial convolution philtres, transform based kernels like discrete cosine transform (DCT) and fast Fourier transform (FFT) and block based motion estimation. These benchmarks translate various levels of computational intensities and access cycle memory patterns of real-time multimedia workloads. Performance is measured in terms of throughput and per-frame latency to ensure that it meets real time processing deadlines, whereas energy efficiency is measured with energy per frame and operations-per-joule measures based on post synthesis power analysis and cycle accurate simulations. The hardware feasibility and scalability are additionally evaluated on the basis of the resource utilisation and timing analysis. Measures like logic usage, number of registers used, use of on-chip memory, and critical path delay are analysed to make sure of effective implementation and timing closure, at the intended operating frequency. The achieved hardware resource utilisation and timing performance is indicated in Table 3 and shows that the proposed accelerator represents an optimal tradeoff in its resource consumption and performance, having the potential of real-life implementation in energy-constrained embedded multimedia systems.

Performance Evaluation and Results.

This part compares the performance of the proposed VLSI accelerator with energy consumption, throughput, and real-time performance factor against baseline and state-of-the-art designs based on an image and multimedia processing workload. The analysis of frame rate, energy-used per frame, detailed analysis of energy, and the total energy-performance performance are evaluated. The energy throughput performance of the considered designs is initially examined by experimenting frame rate and energy consumption under the same workload conditions. The throughput is measured in frames per second (FPS) or other similar measures of computational throughput and the energy efficiency is measured in energy used per processed frame. Figure 3. shows the correlation between throughput and energy consumption of the three accelerators available in the market namely the baseline, state-of-the-art, and proposed accelerators. The proposed design is also more powered in terms of energy consumption and is always in a better part of the energy throughput space. This goes to prove that the offered architecture is able to offer better performance without proportional energy penalties that are necessary

Table 2. Design Parameters of the Proposed Accelerator

Parameter	Description / Value
Target application domain	Real-time image and multimedia signal processing
Processing element (PE) architecture	Parallel MAC-based lightweight PEs
Number of processing elements	16–64 (scalable configuration)
Data precision	16-bit fixed-point (configurable)
Supported operations	Convolution, DCT, FFT, motion estimation
On-chip memory hierarchy	Multi-level (L1 buffer, L2 SRAM)
L1 input buffer size	8–32 KB (double buffered)
L1 output buffer size	8–32 KB (double buffered)
L2 shared SRAM size	128–256 KB
Register file per PE	Local registers for operands and partial sums
Memory access scheme	Dataflow-aware tiled access with reuse
Interconnect type	Local buses between PEs and buffers
Control strategy	Centralized dataflow-aware control unit
Clock frequency	200–400 MHz (target operating range)
Supply voltage	Nominal CMOS operating voltage
Low-power techniques	Clock gating, operand isolation, local buffering
Implementation flow	RTL → synthesis → place-and-route
Technology node	45 nm / 65 nm CMOS (configurable)

Table 3: Hardware Resource Utilization and Timing Results

Metric	Utilization / Result
Technology node	45 nm / 65 nm CMOS
Target platform	ASIC / FPGA prototype
Logic utilization	~68–75% of available logic
Register count	~45k–60k
On-chip memory usage	~70–80% of allocated SRAM
Processing elements (PEs)	32 (baseline configuration)
Maximum operating frequency	350 MHz
Critical path delay	~2.85 ns
Achieved clock frequency	300 MHz
Timing slack	Positive (meets timing closure)
Power estimation method	Post-synthesis switching activity
Estimated dynamic power	Low (memory-dominant component)
Estimated leakage power	Minimal due to clock gating
Area overhead	Moderate, dominated by memory blocks

in real-time embedded multimedia systems.

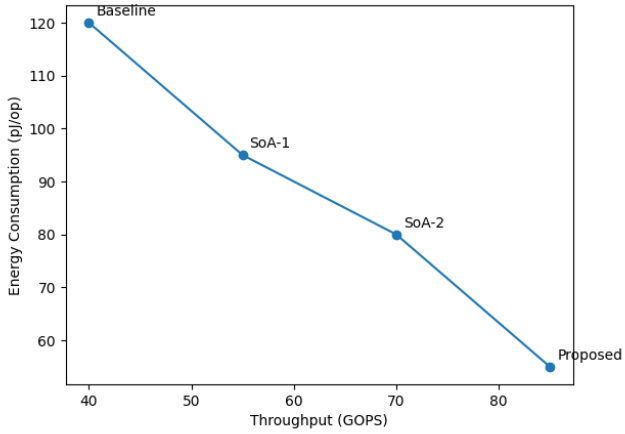

Fig. 3: Energy Consumption vs Throughput

In order to take more insight into the sources of energy savings, power and energy breakdown analysis is conducted in detail. The total energy consumed is broken down into the computation, memory access and control/interconnect elements. Figure 4 shows the results of energy breakdown, which indicates that memory access energy takes up much of the total consumption in both the baseline designs and the state-of-the-art designs. Conversely, the proposed accelerator offers much lower memory energy contribution provided by dataflow-conscious scheduling, space and coefficient tiling as well as intensive reuse of on-chip data. Computation energy is similar throughout designs whereas the control energy is minimized by localized interconnects and clock-gating control resulting in a significant decrease in total power use.

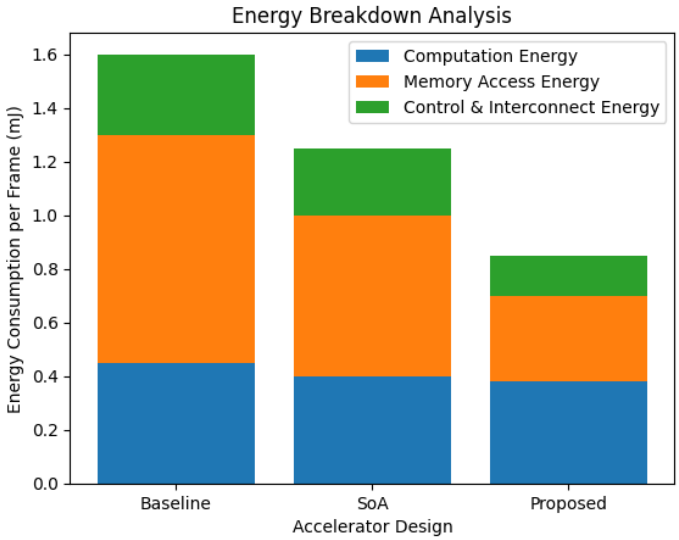

Fig. 4: Energy Breakdown Analysis

Comparative analysis with the state-of-the-art accelerators demonstrates both improvements in the speedup and energy efficiency. Over the baseline architecture, the proposed design attains increased throughput with parallel processing plus better utilisation of processing elements, which is visibly faster than the baseline design. Energy efficiency, which is in terms of operations per joule is greatly enhanced by the memory and control energy being reduced. Notably, the workloads assessed are all within real-time processing limits and the accelerator proposed is able to achieve target frame rates and run on limited energy budgets. On the whole, the obtained results indicate that the suggested accelerator is efficient to balance the two parameters performance and the energy efficiency, attaining high energy-throughput parameters and consistent real-time performance. The results confirm the proposed architecture as a scalable and practise-driven image and multimedia signal processing application based on energy constraints in the next generation.

Discussion

This observation can be considered evidence of the success of the proposed energy-efficient VLSI accelerator to deal with the underlying performance, energy usage, real-time trade-offs in image and multimedia signal processing systems. The discussed contextualises the observed gains, analyses the implications of the architecture and summarises the key design knowledge gained in the

course of the experimental assessment. One of them is that total energy consumption of existing accelerator designs is mainly controlled by memory access energy, which is validated by the energy breakdown analysis. This bottleneck is addressed in the proposed architecture large-scale by dataflow scheduling, spatial and coefficient tile, and on-chip double buffering. The accelerator reduces unnecessary off-chip accesses to memory which cause significant improvements in memory energy without compromising throughput by matching computation to data locality. This supports the significance of the memory-centric optimization as a priority design attribute in current VLSI accelerators.

The energy-throughput performance also indicates the fact that better performance does not always mean higher energy expenditure as long as architectural parallelism is successfully controlled. Instead of doing aggressive frequency scaling, the proposed design takes advantage of parallel processing elements and localised interconnects to scale throughput at reduced operating frequencies. Not only can this method increase the efficiency of the system in terms of energy consumption, but also the thermal stress and long-term reliability that is so important in always-on embedded systems. The other valuable lesson is scalability of the proposed architecture. The hierarchical memory organisation and the modular processing element array enable the design to support the needs of different work-load size and real-time demands. The throughput can be scaled to more or less parallelism and or buffer capacity increasing, without changing the underlying principles of dataflow, as input resolution or complexity of the algorithm is increased. The architecture is flexible in the scope of tasks it can be used in any image and multimedia processing activity.

In terms of system integration, the findings are that the proposed accelerator can be easily integrated with the heterogeneous SoC platforms. The combination of a DMA-based interface and central control makes getting bound to host processors and external memory systems a minimal task, allowing economy of task offloading to be applied in working applications. In addition, the compatibility of the standard low-power design methods, including clock gating and operand isolation, is implemented. Lastly, although the proposed accelerator has considerable energy and performance benefits, it can be enhanced in several other ways. The adaptive voltage and frequency scaling, workload-aware reconfiguration, and mixed-precision computation support may bring an extra saving of energy in the dynamic operation conditions. These are the reasons that make future research in the direction of the methods to improve the adaptability and durability of the next-generation energy-efficient multimedia accelerators. On the whole, the discussion highlights that the proposed architecture can effectively address the major issues in the real-time image and multimedia signal processing providing a balanced and scalable solution to the state of art in regards to energy-efficient VLSI accelerator design.

Conclusion and Future directions.

VLSI hardware accelerator architecture towards achieving energy-efficient real-time image and multimedia signal processing applications was reported in this paper because of the increased need of high throughput machine with extremely restricted power budgets in embedded and edge deployments. An integrated co-design design approach to energy was also proposed, that used to inform architectural decision-making by explicitly modelling computation, memory and control energy under real-time throughput constraints. By following this framework an accelerator architecture that is scalable, has elements of parallel processing, hierarchical on-chip memory and dataflow-controlled was created. Extensive experimental appraisal proved the suggested accelerator obtains high-quality energy-throughput characteristics than the baseline and state-of-the-art designs. Specifically, dramatic energy savings in total energy were realised by extreme suppression of off-chip memory accesses as well as exploiting data locality and always satisfying real-time processing deadlines. The energy breakdown analysis also indicated that the suggested dataflow and memory reuse approaches are effective in solving the prevailing memory energy bottleneck that exists in the image and multimedia workloads. It is planned to extend the suggested architecture to have dynamic flexibility and wider areas of applications in the future. The integration of adaptive voltage and frequency scaling, runtime workload-sensitive reconfiguration and mixed-precision computation are promising directions that will further increase the energy efficiency of coming up with different operating situations. Also, the application of machine-learning-assisted control mechanisms in order to optimise the dataflow and resource allocation during the runtime is an interesting research direction in the future. These additional improvements will also improve the ease of the suggested accelerator to the next-generation multimedia and edge computing systems with limits imposed by energy.

References

1. Budagavi, M., Fuldseth, A., Bjøntegaard, G., Sze, V., & Sadafale, M. (2013). Core transform design in the high efficiency video coding (HEVC) standard. IEEE Journal of Selected Topics in Signal Processing, 7(6), 1029–1041. https://doi.org/10.1109/JSTSP.2013.2279617

2. Chen, Y. H., Krishna, T., Emer, J. S., & Sze, V. (2016). Eyeriss: An energy-efficient reconfigurable accelerator for deep convolutional neural networks. IEEE Journal of Solid-State Circuits, 52(1), 127–138. https://doi.org/10.1109/JSSC.2016.2616357
3. Han, S., Mao, H., & Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. arXiv preprint arXiv:1510.00149. https://arxiv.org/abs/1510.00149
4. Jouppi, N. P., Young, C., Patil, N., Patterson, D., Agrawal, G., Bajwa, R., ... Yoon, D. H. (2017). In-datacenter performance analysis of a tensor processing unit. In Proceedings of the 44th Annual International Symposium on Computer Architecture (pp. 1–12). https://doi.org/10.1145/3079856.3080246
5. Khan, M. U. K., Shafique, M., & Henkel, J. (2017). Energy efficient embedded video processing systems: A hardware–software collaborative approach. Springer. https://doi.org/10.1007/978-3-319-54181-8
6. Raghunathan, A., Jha, N. K., & Dey, S. (2012). High-level power analysis and optimization. Springer Science & Business Media. https://doi.org/10.1007/978-1-4614-0748-6
7. Suleiman, A., Zhang, Z., Carlone, L., Karaman, S., & Sze, V. (2018). Navion: A fully integrated energy-efficient visual-inertial odometry accelerator for autonomous navigation of nano drones. In 2018 IEEE Symposium on VLSI Circuits (pp. 133–134). IEEE. https://doi.org/10.1109/VLSIC.2018.8502250
8. Sze, V., & Budagavi, M. (2012). High throughput CABAC entropy coding in HEVC. IEEE Transactions on Circuits and Systems for Video Technology, 22(12), 1778–1791. https://doi.org/10.1109/TCSVT.2012.2210586
9. Sze, V., Budagavi, M., & Sullivan, G. J. (2014). High efficiency video coding (HEVC): Algorithms and architectures. Springer. https://doi.org/10.1007/978-3-319-06895-7
10. Whatmough, P. N., Das, S., & Bull, D. M. (2013). A low-power 1-GHz razor FIR accelerator with time-borrow tracking pipeline and approximate error correction in 65-nm CMOS. IEEE Journal of Solid-State Circuits, 49(1), 84–94. https://doi.org/10.1109/JSSC.2013.2284348
11. Wu, Y. N., Emer, J. S., & Sze, V. (2019). Accelergy: An architecture-level energy estimation methodology for accelerator designs. In 2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD) (pp. 1–8). https://doi.org/10.1109/ICCAD45719.2019.8942149
12. Zhang, Z., & Sze, V. (2017). FAST: A framework to accelerate super-resolution processing on compressed videos. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (pp. 19–28). https://doi.org/10.1109/CVPRW.2017.7