Fast 2D Convolutions and Cross-Correlations Using Scalable Architectures

Cesar Carranza*, Daniel Llamocca†, and Marios Pattichis‡

*Sección Electricidad y Electrónica, Pontificia Universidad Católica del Perú, Lima-32, Perú

†Electrical and Computer Engineering Department, Oakland University, Rochester, MI, 48309, USA

‡Department of Electrical and Computer Engineering, The University of New Mexico, Albuquerque, NM, USA

Emails: acarran@pucp.edu.pe, llamocca@oakland.edu, pattichis@ece.unm.edu

Abstract

The manuscript describes fast and scalable architectures and associated algorithms for computing convolutions and cross-correlations. The basic idea is to map 2D convolutions and cross-correlations to a collection of 1D convolutions and cross-correlations in the transform domain. This is accomplished through the use of the Discrete Periodic Radon Transform (DPRT) for general kernels and the use of SVD-LU decompositions for low-rank kernels.

The approach uses scalable architectures that can be fitted into modern FPGA and Zynq-SOC devices. Based on different types of available resources, for $P \times P$ blocks, 2D convolutions and cross-correlations can be computed in just $O(P)$ clock cycles up to $O(P^2)$ clock cycles. Thus, there is a trade-off between performance and required numbers and types of resources. We provide implementations of the proposed architectures using modern programmable devices (Virtex-7 and Zynq-SOC). Based on the amounts and types of required resources, we show that the proposed approaches significantly outperform current methods.

Background

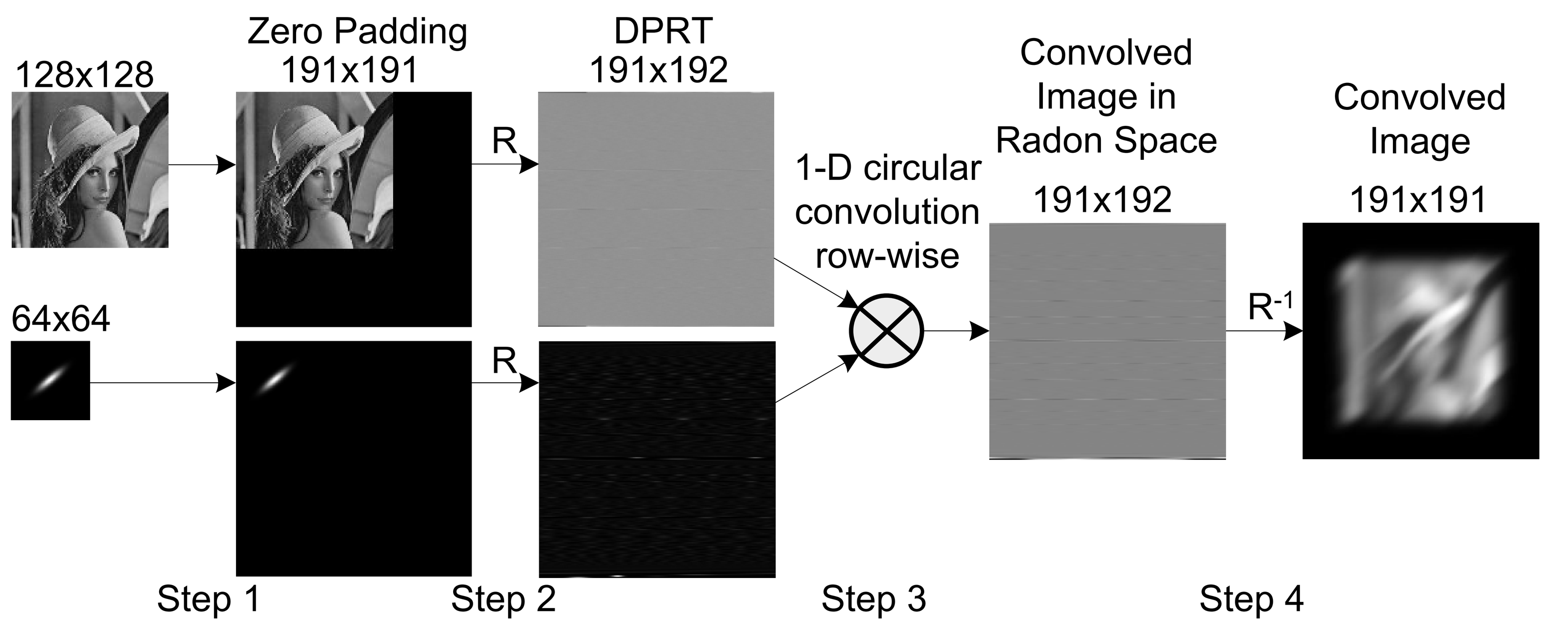

Fig. 1: 2-D Linear convolution using the Discrete Periodic Radon Transform and 1-D Circular convolutions

Proposed Methods

Method	Hardware components
FastConv / *FastXCorr*	1D Circular convolver, FDPRT/iFDPRT [1]
FastScaleConv / *FastScaleXCorr*	1D Circular convolver, SFDPRT/iSFDPRT [1]
FastRankConv / *FastRankXCorr*	1D Linear convolver, Custom SRAM

Methods

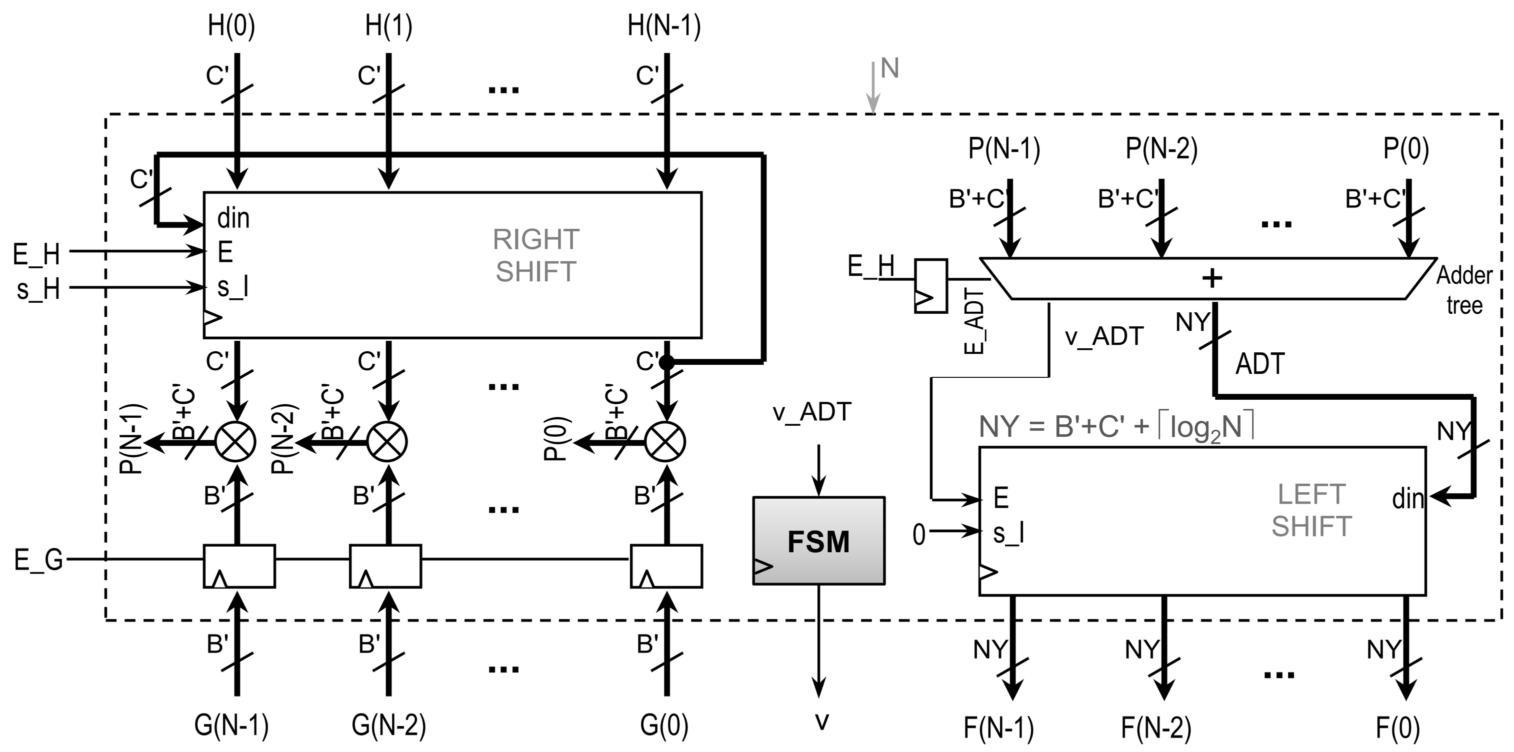

Fig. 2: Architecture for computing the 1D circular convolution.

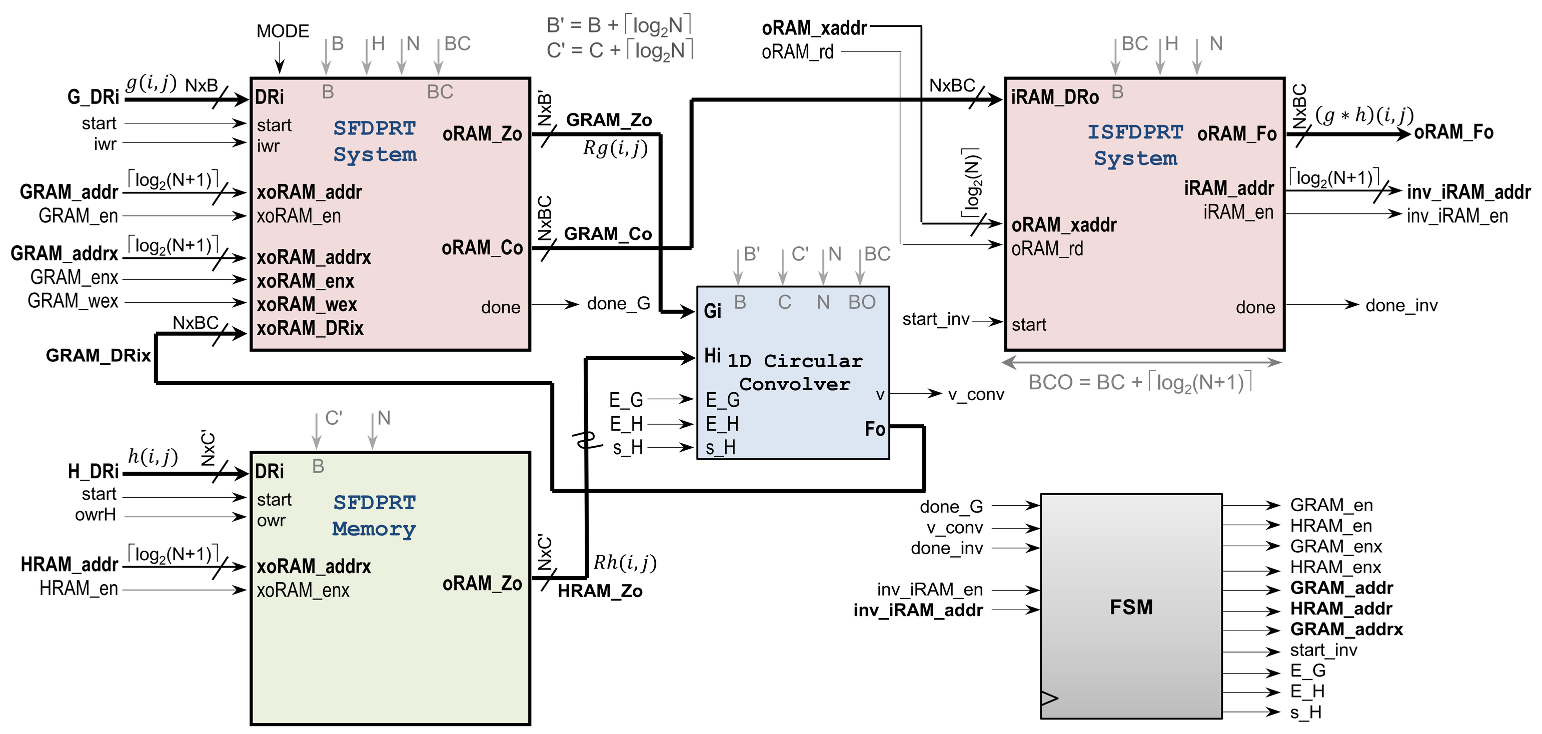

Fig. 3: *FastScaleConv* and *FastScaleXCorr*: Fast and scalable architecture system for computing 2D convolutions and cross-correlations based on the DPRT.

Conclusion

The manuscript introduced fast and scalable architectures for computing 2D cross-correlations and convolutions. *FastConv* architectures deliver the best performance by computing convolutions in $O(P)$ clock cycles. The *FastScaleConv* family of architectures allows us to implement efficient architectures that can be restricted to the architectures of different devices. The *FastRankConv* family of architectures allows us to consider low-rank approximations that can significantly reduce the number of required resources. Overall, for the same level of performance, *FastRankConv* and *FastScaleConv* require significantly fewer hardware resources than alternative approaches.

Results

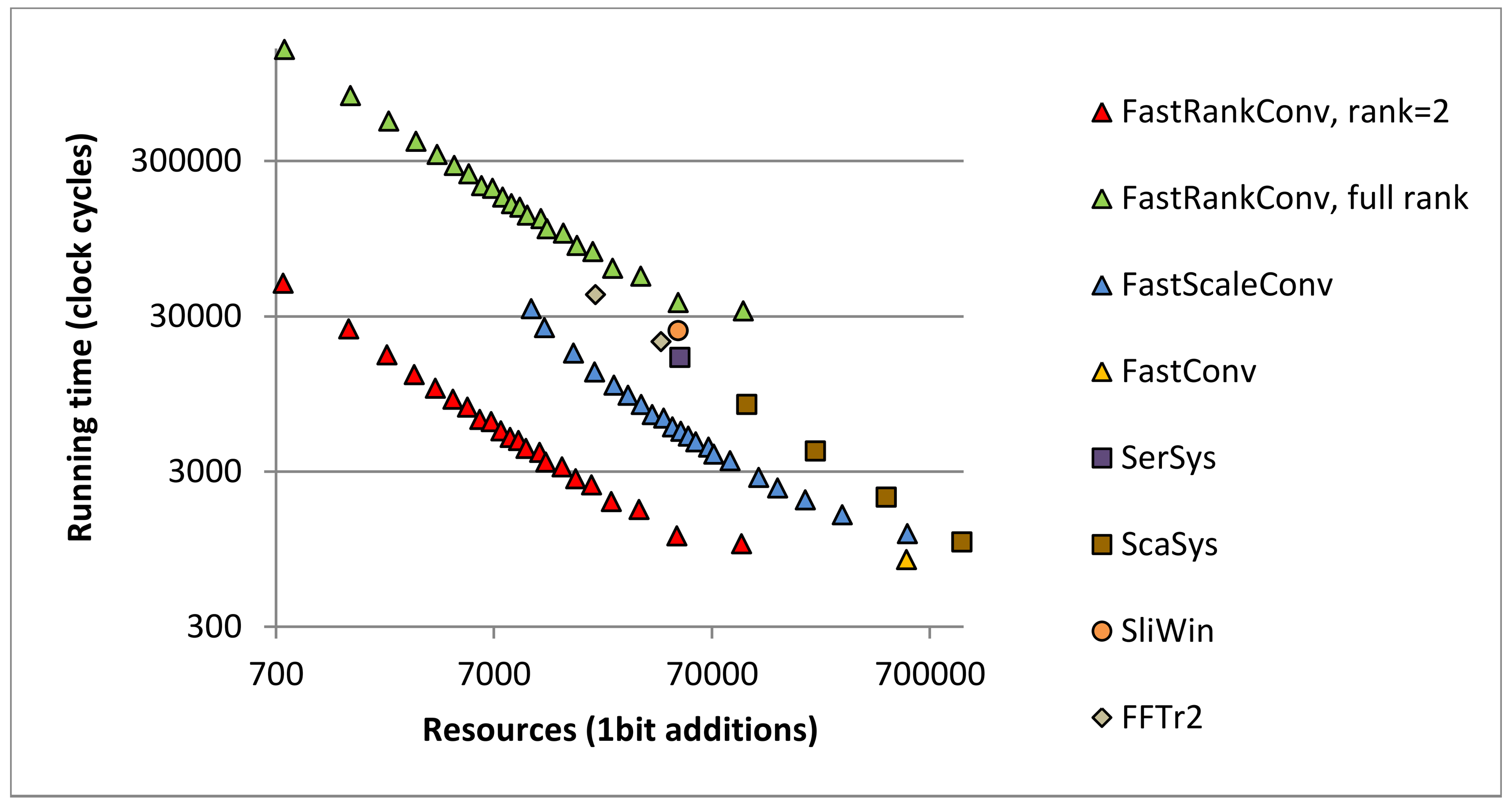

Fig. 4: Family of fast and scalable architectures for $N = 127$ ($N = 128$ for FFTr2) in terms of Running time versus the required number of 1-bit additions. Similar plots are obtained for other resources.

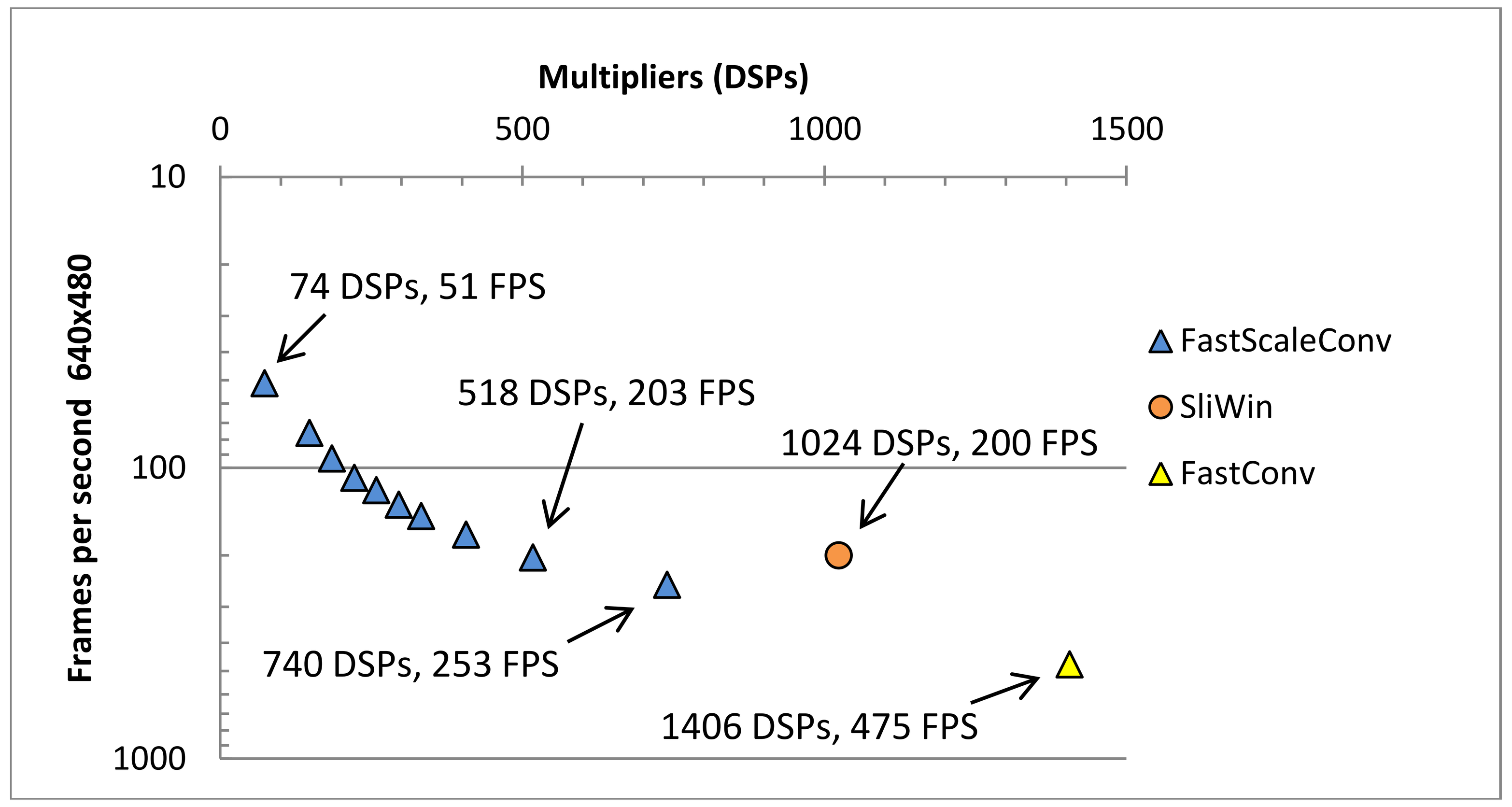

Fig. 5: Performance comparison between *SliWin* (see references), *FastConv* and *FastScaleConv*. To measure performance, we consider the number of Frames Per Second (FPS) to perform the convolution between an image of 480*p* (640×480) and a kernel of size 19×19.

References

[1] C. Carranza, D. Llamocca, and M. Pattichis, Fast and scalable computation of the forward and inverse discrete periodic radon transform, *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 119–133, Jan 2016.

[2] P. Cooke, J. Fowers, G. Brown, and G. Stitt, “A tradeoff analysis of fpgas, gpus, and multicores for sliding-window applications,” *ACM Trans. Reconfigurable Technol. Syst.*, vol. 8, no. 1, pp. 2:1–2:24, Mar. 2015.