HIGH-LEVEL LINEAR PROJECTION CIRCUIT DESIGN OPTIMIZATION FRAMEWORK FOR FPGAS UNDER OVER-CLOCKING

Rui Policarpo Duarte[†]*, Christos-Savvas Bouganis*

Department of Electrical and Electronic Engineering
Imperial College London
SW7 2AZ London, U.K.
email: r.duarte09@imperial.ac.uk, christos-savvas.bouganis@imperial.ac.uk

ABSTRACT

Frequently, the high-level algorithm parameter selection and its mapping into hardware are considered to be independent processes, often leading to suboptimal solutions. When DSP applications with real-time constraints are targeted, it is often desirable the resulting hardware system to be clocked at as high frequency as possible. Even though the trend in modern devices is to provide a fabric that can support higher frequencies, its variability makes the design tools to be pessimistic about maximum clock frequency estimates. The proposed framework optimizes and mitigates the probabilistic behaviour of digital circuits, by trying to expose the impact of variability of the fabric into high-level algorithmic specifications. FPGAs are well positioned to tackle this problem because they can be reconfigured, allowing an off-line characterization of the specific device before implementing the complete optimized circuit on the same device. Circuits generated by the proposed framework outperform typical implementations, by minimizing area, errors, and maximizing its operating clock frequency. An example of a linear projection circuit, over-clocked by 232%, shows savings up to 39% in hardware resources for the same target PSNR over traditional implementation.

1. INTRODUCTION

The most frequent strategy to implement Digital Signal Processing (DSP) designs, with high throughput requirements, is to deeply pipeline the design and clock it at its maximum frequency. However synthesis tools are conservative in the estimates for the delays of the critical paths. They keep a guard margin between the expected clock frequency and the reported maximum clock frequency. However, as the silicon technology scales down, process variation increases, leading to synthesis tools to impose larger guard margins, reporting clock frequencies lower than those expected for a particular fabrication process. Moreover, synthesis tools rely on models, and don't have information about the actual device that is being targeted. Experience has shown that it is possible to operate a circuit correctly beyond the maximum frequency reported. It is possible to determine the maximum operating frequency via experimentation, but this isn't practical when circuit optimization and algorithm parameter selection are considered simutaneously.

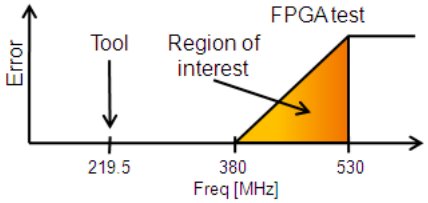

Fig. 1. Error *vs* clock frequency for a Z^6 to Z^3 linear projection circuit.

As an example to demonstrate the above, a linear projection Z^6 to Z^3 circuit has been selected. Figure 1 shows the error at the output of the system as function of frequency when DE0 board [1] is targeted. In this particular example, the clock frequency can be increased up to 380 MHz without any errors at the output of the system. Beyond that frequency, errors gradually appear at the output of the system until 530 MHz, when it stops producing meaningful results.

The key idea in this work is to have a prior characterization of the device with respect to the degradation of the performance of the computational units, and combine this information with high-level parameter selection of an algorithm. Field-Programmable Gate Arrays (FPGAs) were considered because of their reconfigurability property. This is of great importance as it enables to have components of the system characterized on a device, and later to have the device reconfigured with the complete design.

The proposed framework aims to automatically generate optimal circuit designs for linear projections, taking into account the characterization of devices to operate beyond the maximum clock frequency determined for correct operation, shown as *region of interest*. For the linear projection

[†]The author would like to thank Fundação para a Ciência e Tecnologia (Lisbon) for the support through grant SFRH/BD/69587

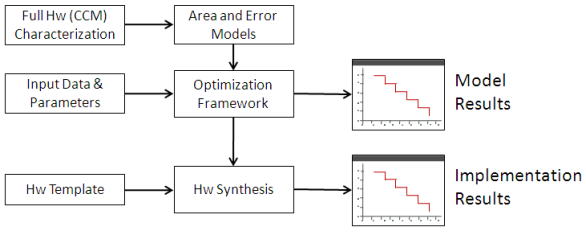

Fig. 2. Design optimization flow and results using the proposed optimization framework.

problem being considered, the proposed framework computes the orthogonal basis matrix Λ to minimize the hardware resources, the errors of the arithmetic units under over-cloking, while minimizing data reconstruction error in the original space. A Bayesian framework to optimize linear projections designs was first introduced in [2] and enhanced in [3] and [4]. In the work here presented, the framework was modified to consider information about errors in the arithmetic units when over-clocked.

2. OPTIMIZATION FRAMEWORK

Implementation and optimization of DSP systems usually consists of three steps: 1. determine the appropriate subspace to better represent data (e.g. fixed-point); 2. quantization of data to be represented in hardware; 3. assignment of the most computational demanding units to elementary elements of the device (e.g. LUTs). Often this approach leads to suboptimal solutions because the mapping of the algorithm into hardware doesn't take into account restrictions imposed by the device.

The flow and the output of the proposed framework is illustrated in Figure 2. The first step is to characterize the arithmetic units to create the models for circuit area and errors when these units are clocked at certain frequency. The second step is to create the designs from the input data and the parameters for the problem being considered. For each design, the optimization framework samples the elements of the Λ matrix using the information from the circuit models. The result is a set of designs and estimates for them from the characterization models, but assuming no errors in the calculations. In the third step, the designs are synthesized and implemented on an FPGA, and the actual results collected.

Exhaustive search for all the combinations of the variables in the design is impractical, therefore Bayesian inference methods were considered to extract the optimized designs. The framework inputs are the highly-dimensional input data X, the models for hardware *cost* (area and errors due to over-clocking), the requirements in terms of noise in the reconstructed data, and the dimensionality of the reduced space. The output is the basis matrix Λ. The framework also searches the space for possible bases using different word-lengths for the magnitude of the Constant Coefficient Multipliers (CCMs) in the Λ matrix, up to 8 bits.

3. LINEAR PROJECTION

Linear projection, also known as Karhunen-Loeve transformation (KLT) [5], or PCA is formulated as follows. The original input data vector x with P elements, where $x \in R^P$, Λ the orthogonal basis of the new space with dimensions $P \times K$, and by $F \in R^K$ the factors vector, which is the result of projecting the original data to the new space,

$$F = \Lambda^T X. \tag{1}$$

If the original data can be described in the reduced space without any errors, the original data can be recovered from the subspace via $X = \Lambda F$. If not, the original data is then obtained via (2):

$$X = \Lambda F + E \tag{2}$$

where E is the error of the approximation, which the framework iteratively tries to minimize. It assumes the data has zero mean. The columns for the Λ matrix is computed using (3):

$$\lambda_i = arg\ max\ \varepsilon\{(\lambda_i^T X_{i-1})^2\} \tag{3}$$

$$X_i = X - \sum_{k=1}^{i-1} \lambda_k \lambda_k^T X \tag{4}$$

where $X = [x^1 x^2 ... x^N]$, $X_0 = X$, $\|\lambda_i\| = 1$ and $\varepsilon\{.\}$ refers to expectation.

4. BAYESIAN FORMULATION

The Bayesian framework considers the subspace estimation and hardware implementation simultaneously, allowing the framework to efficiently explore the heterogeneity properties of modern FPGAs, generating area optimized DSP designs. The work here proposed is an extension of previous work [4]. It now considers the stochastic behaviour of the arithmetic units, when over-clocked.

The framework estimates the basis matrix Λ, the noise covariance Ψ, and the factors using Gibbs sampling algorithm [6] from the posterior distribution of the variables.

Since neither Λ and F are known, solving (2), to minimize

$$min \sum \sum (\Lambda F - X)^2 \tag{5}$$

is an ill-conditioning problem. The main advantage of the proposed framework is the ability to obtain F, Λ and Ψ according to their prior and posterior probabilities, without having to adopt an heuristic method to explore the solution space that could even lead to suboptimal solutions.

The probability distribution of the Λ matrix maps the *cost* of the circuit, and it is expressed as the product of the probabilities of the individual elements,

$$p(\Lambda) = \Pi_{p=1}^{P} \Pi_{k=1}^{K} p(\lambda_{pk}). \quad (6)$$

5. MAPPING CIRCUIT COST TO PRIOR DISTRIBUTION

5.1. Cost Matrix

The *Cost Matrix* holds the information about the *cost* of each arithmetic unit, in terms of hardware resources and errors when over-clocked. Frequently the critical-path is in the multiplication circuits, as they are the most complex arithmetic units in DSP circuits. In this work, the arithmetic units considered were the CCMs. It was assumed the adder tree generates no errors for the investigated frequency.

The hardware resources are quantified in Logic Cells (LCs), and they are the actual area required to implement the arithmetic units in the test circuit. In this work, this information is read from the report generated by the synthesis tool.

Errors were obtained from the characterization test on the FPGA. Each arithmetic unit was operated with clock frequencies higher than the maximum limit stated by the synthesis tool, and stimulated with an input sequence of 6000 random samples generated from a uniform distribution. The errors at the output of each arithmetic unit, under over-clocking, are quantified in variance and mean error.

A sign-magnitude representation was adopted as it reduces the time required to characterize the CCMs.

5.2. Prior Distribution

The prior distribution, in (6), for the Λ matrix has to meet the following properties:

$$p(\lambda_{pk}) \geq 0, \forall \lambda_{pk}, \quad (7)$$

$$\sum_{\lambda_{pk}} p(\lambda_{pk}) = 1. \quad (8)$$

The *cost* is a function of the CCMs area and errors due to over-clocking. Due to the stochastic nature of errors it is important to have the optimization framework selecting the CCMs with less variability under over-clocking. Consequently, the metric chosen to quantify errors was the variance as it measures by how much the values spread. The prior probabilities of the individual elements of the Λ matrix, expressed as function of the *cost*, is given by (11).

$$A = \left(\frac{1}{area}\right)^{\alpha} \quad (9)$$

$$E = \left(\frac{1}{1+variance}\right)^{\beta} \quad (10)$$

$$g(A(\lambda_{pk}), E(\lambda_{pk})) = c(A(\lambda_{pk}))^{-\alpha}(E(\lambda_{pk}))^{-\beta} \quad (11)$$

c is a constant used to ensure that

$$\sum_{\lambda_{pk}} g(A(\lambda_{pk}), E(\lambda_{pk})) = 1. \quad (12)$$

Hyper-parameters α and β control the weight of area and errors in the probability distribution.

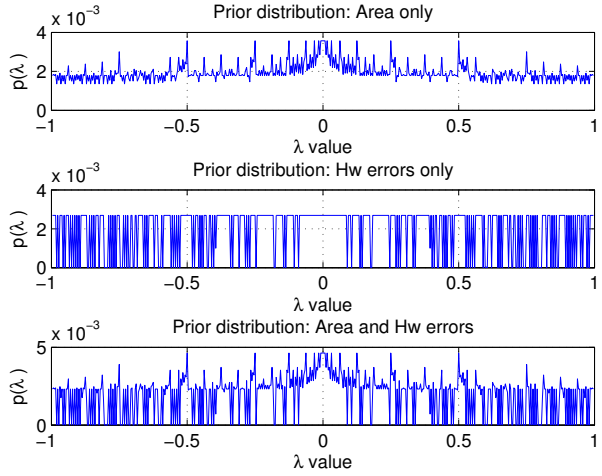

Fig. 3. Mapping of area and error information to a probability distribution for $\alpha = 1$ and $\beta = 1$ at 510 MHz.

Figure 3 exemplifies the prior distributions for $(\alpha, \beta) = [(1,0),(0,1),(1,1)]$. The top plot represents the prior distribution taking only into account the area of CCMs, whereas the middle plot shows the information about the errors due to over-clocking. The bottom plot shows the combination of both prior probability distributions.

6. PERFORMANCE EVALUATION

To demonstrate the effectiveness of the framework, a test circuit to implement a Z^6 to Z^3 linear projection was created, which is the largest linear projection circuit fully pipelined the FPGA on the DE0 board can accommodate. The data in the higher dimension space was generated randomly, sampled from a uniform distribution between -1 and 1 and quantized with 8 bits. It assumes that the covariance matrix follows a gamma distribution, and has zero mean. The implementation targets real-time linear projection designs, assuming high throughput requirements. The circuit implements an unrolled and fully pipelined linear projection.

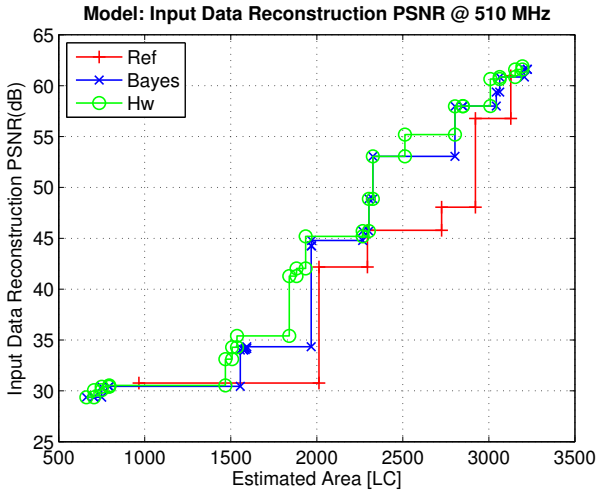

Fig. 4. Estimated circuit area *vs* model reconstruction PSNR at 510 MHz.

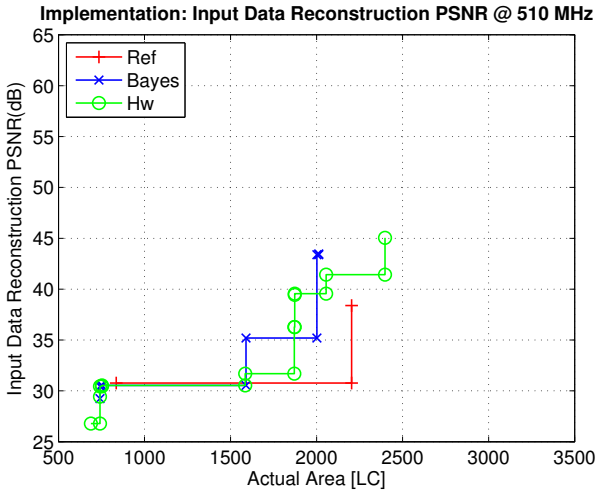

Fig. 5. Actual circuit area *vs* implementation reconstruction PSNR at 510 MHz.

We have compared the performance of the proposed optimization framework (*Hw*) with the typical KLT implementation (*Ref*) and area optimization without information about reliability (*Bayes*). The results are presented as Pareto curves and they show the optimal sets of designs, in terms of area and reconstruction Peak Signal-to-Noise Ratio (PSNR) at 510 MHz. The clock frequency considered is 2.32 times faster than the maximum specified by the synthesis tool. Figures 4 and 5 show the results for the model and the implementation, respectively.

The model results show the estimated hardware resources *vs* the estimated reconstruction PSNR in the original space. The framework considers the information about the CCMs operating at 510 MHz, but it assumes the arithmetic units don't have errors in their calculations. The results show that *Hw* designs are always better than the *Ref* designs. Compared to *Bayes* designs, the *Hw* designs are better or at least the same.

The implementation results show the actual hardware resources required *vs* the results from the board when operated under over-clocking at 510 MHz. For a PSNR above 31 dB the *Ref* design requires 39% more hardware resources than the *Hw* design. Below the limit of 2000 LCs, the *Hw* design obtained more 4.35 dB and more 8.79 dB PSNR than *Bayes* and *Ref* designs, respectively. The deviations between the model and the implementation happen because the adder tree wasn't modeled, and their errors haven't been considered in the optimization process. Nevertheless, *Hw* provides the design with the best reconstruction PSNR in both cases.

7. CONCLUSION

This paper proposes an optimization framework for implementation of linear projection designs on FPGAs. It couples the problem of data approximation, area optimization and error minimization under over-clocking. It was demonstrated that is possible to optimize a linear projection design for area, reconstruction data PSNR and resilience to operation under over-clocking simultaneously, by inserting information regarding the area and performance of the arithmetic units. It was demonstrated, at 510 MHz, that for a target PSNR it is possible to reduce resource usage and increase the clock frequency by 290 MHz.

Future work involves characterizing other arithmetic units and finding applications suitable to be optimized by the proposed framework.

8. REFERENCES

[1] T. Technologies. (2009) Terasic DE0 board user manual v. 1.3. [Online]. Available: http://www.terasic.com.tw

[2] C.-S. Bouganis, I. Pournara, and P. Y. K. Cheung, "A statistical framework for dimensionality reduction implementation in FPGAs," in *Proc. IEEE Int. Conf. Field Programmable Technology FPT 2006*, 2006, pp. 365–368.

[3] C. S. Bouganis, I. Pournara, and P. Y. K. Cheung, "Efficient mapping of dimensionality reduction designs onto heterogeneous FPGAs," in *Proc. 15th Annual IEEE Symp. Field-Programmable Custom Computing Machines FCCM 2007*, 2007, pp. 141–150.

[4] C.-S. Bouganis, I. Pournara, and P. Cheung, "Exploration of heterogeneous FPGAs for mapping linear projection designs," *IEEE Trans. VLSI Syst.*, vol. 18, no. 3, pp. 436–449, 2010.

[5] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933.

[6] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PAMI-6, no. 6, pp. 721 –741, nov. 1984.