Noise Budgeting in Multiple-Kernel Word-Length Optimization

Van-Phu Ha, Tomofumi Yuki, Olivier Sentieys

Noise Budgeting in Multiple-Kernel Word-Length Optimization

Van-Phu HA
Univ Rennes, Inria, CNRS, IRISA
Rennes, France
van-phu.ha@inria.fr

Tomofumi Yuki
Univ Rennes, Inria, CNRS, IRISA
Rennes, France
tomofumi.yuki@inria.fr

Olivier Sentieys
Univ Rennes, Inria, CNRS, IRISA
Rennes, France
olivier.sentieys@inria.fr

***Abstract*—Word-Length Optimization (WLO) is a key step in the implementation of Digital Signal Processing applications on hardware platforms. Existing approaches face scalability problems for applications with several kernels. In this paper, we present our work-in-progress to address the scalability problems when performing multi-kernel WLO. Our approach uses application-wide analysis to derive noise budgets for each kernel, followed by independent WLO. The main idea is to characterize the impact of approximating each kernel to the accuracy/cost through simulation and regression analysis. The constructed models can be used to determine the appropriate noise budget for each kernel. When applied to WLO for two kernels in Image Signal Processor, the noise budgets given by our analysis closely match those found empirically with global WLO.**

***Index Terms*—Fixed-Point refinement, Word-Length Optimization, Multiple Kernel Optimization**

I. Introduction and Motivation

Fixed-Point arithmetic is widely used for the implementation of Digital Signal Processing (DSP) systems on electronic devices. The Float-to-Fix conversion in hardware design flow always demands to optimize the word-length of variables in the system to find a good trade-off between the cost and quality requirement. The process of word-length determination is typically called Word-Length Optimization (WLO). WLO is known as an NP-Hard problem with the complexity growing exponentially when more variables are involved. In fact, this process may take up to $25-50\%$ of design time [1]. Therefore, WLO is still a problem of interest in reducing the time-to-market in the electronics industry.

A common approach to WLO involves iterative search algorithms based on simulations [2]–[7]. Others have proposed relaxing the problem with convex functions to directly solve for the optimal solution [8]–[10]. However, these methods cannot be directly extended to handle applications with several kernels. The accuracy evaluation by simulations is nearly impossible because the exponential growth of the problem leads to a huge number of simulations. Methods based on convex optimizations require the accuracy to be modeled as convex functions, which cannot be analytically constructed in general. In particular, existing analytical models focus on modeling the noise power, and cannot be used for other quality metrics such as Structural Similarity (SSIM).

This work was supported in part by the French National Research Agency under ARTEFaCT project.

Recently, some techniques to address the scalability of WLO have been proposed [11]–[13]. The work by Novo et al. [12] first performs local WLO with different accuracy constraints and later combine the local solutions with an iterative search. The others have explored hierarchical decomposition of the original system to reduce the problem size at each step in the exploration [11], [13]. These work employ a large number of local WLO, which limits the scalability when considering applications with several kernels.

II. Specific Problem and Solution

In this study, we discuss an approach to overcome scalability issues when applying WLO for applications with several kernels. We use Image Signal Processor, a post-processing pipeline for digital cameras illustrated in Figure 1, as our running example. The ISP consists of several kernels such as noise cancelling (Non-Local Means), linearization recovering, light-fall-off correcting (Vignetting Correction), color adjusting and quality improving by the remaining kernels. The challenges in performing WLO for applications with several kernels include:

- The number of operations is too high to perform global optimization across all kernels.
- The kernels often contain diverse topologies, e.g., including non-linear structures and unsmooth operators, making it difficult to analytically model the impact on quality.
- Analytically modeling quality metrics, other than noise power, is difficult. For instance, in ISP, the commonly used metric is SSIM, which is hard to model analytically.

In this work, we view an application as a composition of kernels; a small enough unit of computation where local WLO is feasible. For our ISP example, the functional components shown in Figure 1 is directly used as kernels, but the granularity of a kernel can be arbitrarily changed.

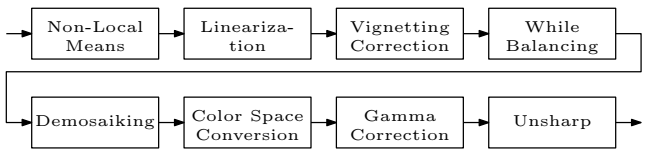

Fig. 1. Image Signal Processor

The loss of quality at the outputs of each kernel after global optimization indicates the amount of approximation

that should be applied for each kernel to obtain the minimum cost. We call this amount of approximation to be applied for each kernel as noise budget. However, performing global optimization has scalability issues, and is not feasible for large application. We assume that performing local optimization to each kernel using the noise budget found in global solution gives a similar solution. Then, finding the noise budget without global optimization allows the global optimization problem to be simplified into local optimization of each kernel. For example, consider that we approximate two kernels, NLM and Unsharp, in the ISP with the target quality being SSIM ≥ 0.99. Applying global optimization, which is feasible for two kernels, gives a solution with the output quality being 0.99 SSIM. Applying the wordlengths in this solution independently to the two kernels gives output quality 0.9990 for NLM and 0.9923 for Unsharp. These are the noise budget given by the global optimization, which we would like to predict. Our approach consists of 3 steps described below:

1) The relation between cost and quality is found through exploration and regression. Each kernel is explored once with a heuristic search targeting high output quality to obtain cost and quality data samples. The cost is estimated from wordlength configurations based on the cost of each operator determined by synthesizing the operators targeting ASIC. These collected data points are used as inputs to non-linear regression with the function $a \times x^b + c$ as a template. This gives us a *continuous* function that directly relates quality and cost without going through wordlengths. The use of power function in the template is motivated by the non-linear behavior of SSIM with respect to the wordlengths.
2) The noise budgets that gives minimal cost are found by performing constrained optimization over the functions.
3) The noise budgets found are used as constraints for independent WLO of each kernel. Then, the results are combined to form the final solution.

Our approach is scalable because (i) both the exploration and WLO are independent to each kernel, and hence has a linear complexity with respect to the number of kernels; and (ii) the functions of the constrained optimization are convex and hence convex optimization tools can be used.

Our preliminary result is summarized in Table I. We first perform local WLO using Tabu search [6] to collect data points for regression. Figure 2 illustrates the constructed models and the data points from initial exploration. Note that only Pareto optimal points are used for regression. The predicted noise budgets are used to perform another local WLO for each kernel and the solutions are combined.

We are able to predict the noise budgets that are close to those inferred from global search results. The execution time is not significantly shorter than the global search for our experiments with two kernels, but we expect the difference to increase when more kernels are included in the exploration.

TABLE I
THE GLOBAL SOLUTION VS. OUR SOLUTION BY CONSTRAINTS GIVEN BY THE NOISE BUDGETING PROCESS, AT REQUIRED SSIM = 0.99

Kernel	SSIM			Cost (μm^2)		Exec. Time (minutes)	
	Ref.	Ours (Predicted)	Ours (Measured)	Ref.	Ours	Ref	Ours
NLM	0.9936	0.9961	0.9978	11570	12576	-	44
Unsharp	0.9923	0.9939	0.9939	8251	8317	-	31
Both	0.9900	-	0.9920	19821	20893	229	75 (+89)

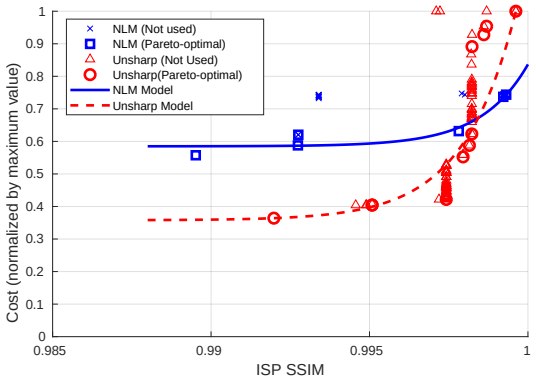

Fig. 2. Model of Cost-SSIM interaction for NLM and Unsharp

III. Conclusion

WLO in the context of large applications remains challenging. In this paper, we present our preliminary work towards improving the scalability of WLO in such contexts. The key idea in our approach is to analyze accuracy models constructed from sampling the design space to derive noise budgets for each kernel. Then, each kernel may be optimized independently, avoiding the combinatorial explosion of the search space as larger applications are considered.

The preliminary results indicate that the approach is feasible and effective for the ISP example. We are currently working on experimenting with approximating more than the two kernels in ISP, other examples, and different accuracy constraints to further evaluate our approach. we are also exploring how the model construction time can be further reduced.

References

[1] M. Clark, M. Mulligan, D. Jackson, and D. Linebarger, "Accelerating fixed-point design for mb-ofdm uwb systems," *CommsDesign, january*, vol. 4, 2005.

[2] M.-A. Cantin, Y. Savaria, and P. Lavoie, "A comparison of automatic word length optimization procedures," in *Circuits and Systems, 2002. ISCAS 2002. IEEE International Symposium on*, vol. 2, pp. II–II, IEEE, 2002.

[3] K. Han, I. Eo, K. Kim, and H. Cho, "Numerical word-length optimization for cdma demodulator," in *Circuits and Systems, 2001. ISCAS 2001. The 2001 IEEE International Symposium on*, vol. 4, pp. 290–293, IEEE, 2001.

[4] H. Choi and W. Burleson, "Search-based wordlength optimization for vlsi/dsp synthesis," in *VLSI Signal Processing, VII, 1994.,[Workshop on]*, pp. 198–207, IEEE, 1994.

[5] G. A. Constantinides, P. Y. Cheung, and W. Luk, "Wordlength optimization for linear digital signal processing," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 22, no. 10, pp. 1432–1442, 2003.

[6] H.-N. Nguyen, D. Ménard, and O. Sentieys, "Novel algorithms for word-length optimization," in *Signal Processing Conference, 2011 19th European*, pp. 1944–1948, IEEE, 2011.

[7] D.-U. Lee, A. A. Gaffar, R. C. Cheung, O. Mencer, W. Luk, and G. A. Constantinides, "Accuracy-guaranteed bit-width optimization," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 25, no. 10, pp. 1990–2000, 2006.

[8] P. D. Fiore, "Efficient approximate wordlength optimization," *IEEE Transactions on Computers*, vol. 57, no. 11, pp. 1561–1570, 2008.

[9] S.-C. Chan and K. M. Tsui, "Wordlength optimization of linear time-invariant systems with multiple outputs using geometric programming," *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 54, no. 4, pp. 845–854, 2007.

[10] K. N. Parashar, D. Menard, and O. Sentieys, "A polynomial time algorithm for solving the word-length optimization problem," in *Computer-Aided Design (ICCAD), 2013 IEEE/ACM International Conference on*, pp. 638–645, IEEE, 2013.

[11] J. Chung and L.-W. Kim, "Bit-width optimization by divide-and-conquer for fixed-point digital signal processing systems," *IEEE Transactions on Computers*, vol. 64, no. 11, pp. 3091–3101, 2015.

[12] D. Novo, I. Tzimi, U. Ahmad, P. Ienne, and F. Catthoor, "Cracking the complexity of fixed-point refinement in complex wireless systems," in *Signal Processing Systems (SiPS), 2013 IEEE Workshop on*, pp. 18–23, Ieee, 2013.

[13] K. N. Parashar, D. Menard, and O. Sentieys, "Accelerated performance evaluation of fixed-point systems with un-smooth operations," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 33, no. 4, pp. 599–612, 2014.