Review Article

Development of Multiobjective High-Level Synthesis for FPGAs

Darian Reyes Fernandez de Bulnes, Yazmin Maldonado, and Leonardo Trujillo

Tecnológico Nacional de México, IT Tijuana, Posgrado en Ciencias de la Ingeniería, Tijuana, BC, Mexico

Correspondence should be addressed to Yazmin Maldonado; yaz.maldonado@tectijuana.edu.mx

Received 2 May 2019; Revised 25 January 2020; Accepted 8 February 2020; Published 29 June 2020

Copyright © 2020 Darian Reyes Fernandez de Bulnes et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Traditionally, the High-Level Synthesis (HLS) for Field Programmable Gate Array (FPGA) devices is a methodology that transforms a behavioral description, as the timing-independent specification, to an abstraction level that is synthesizable, like the Register Transfer Level. This process can be performed under a framework that is known as Design Space Exploration (DSE), which helps to determine the best design by addressing scheduling, allocation, and binding problems, all three of which are NP-hard problems. In this manner, and due to the increased complexity of modern digital circuit designs and concerns regarding the capacity of the FPGAs, designers are proposing novel HLS techniques capable of performing automatic optimization. HLS has several conflicting metrics or objective functions, such as delay, area, power, wire length, digital noise, reliability, and security. For this reason, it is suitable to apply Multiobjective Optimization Algorithms (MOAs), which can handle the different trade-offs among the objective functions. During the last two decades, several MOAs have been applied to solve this problem. This paper introduces a comprehensive analysis of different MOAs that are suitable to perform HLS for FPGA devices. We highlight significant aspects of MOAs, namely, optimization methods, intermediate structures where the optimizations are performed, HLS techniques that are addressed, and benchmarks and performance assessments employed for experimentation. In addition, we show the analysis of how multiple objectives are optimized currently in the algorithms and which are the objective functions that are optimized. Finally, we provide insights and suggestions to contribute to the solution of major research challenges in this area.

1. Introduction

Field Programmable Gate Arrays (FPGAs) designs are made with High-Level Synthesis (HLS). HLS also is known as behavioral synthesis or architectural synthesis, the process to transform an algorithmic description to a synthesizable Register Transfer Level (RTL) netlist. HLS allows designers to work at a higher level of abstraction by using high-level languages such as C/C++ to define the hardware description. Typically, behavioral description, also known as algorithmic level design or system-level design, defines inputs, outputs, and data flow of the behavior inside the algorithm in terms of operations to be performed. Inwardly, this description is usually represented (as an intermediate structure) with an acyclic directed graph. It establishes the data dependencies indicated in the data flow and input/output relations of the design [1]. For any behavioral description, there may be many possible RTL implementations, each one with its own features.

1.1. High-Level Synthesis. HLS can be performed under a framework that is known as Design Space Exploration (DSE), which helps to compute the best design using scheduling, allocation, and binding techniques. All of these tasks are NP-hard problems [2]. Scheduling defines how the design operations will be scheduled into clock cycles. Allocation determines the type and the number of hardware resources (for instance, Functional Units (FUs), storage, or connectivity components) needed to satisfy the design constraints. Binding, also known as assignment mapping and module selection, determines how each variable (in each clock cycle) will be linked to an FU. As Coussy et al. stated, “allocation, scheduling, and binding can be performed simultaneously or in specific sequence depending on the strategy and algorithms used” ([3]; p. 5).

In [4], the evolution of HLS for FPGAs and the HLS tools with single-objective optimization are discussed; according to this review, it is clear that HLS is important because (i)
software programmers want to use FPGA devices to accelerate tasks, so without having knowledge about Hardware Description Language (HDL)—VHDL (Very High-Speed Integrated Circuit (VHSCIC) Hardware Description Language) or Verilog—they can create circuit designs; (ii) designing at a higher level of abstraction leads to increased productivity; for example, software debugging is faster than hardware debugging; and (iii) this process has a lot of potential to perform optimizations. Recently, HLS has been applied to a variety of applications with significant benefits in terms of performance and energy consumption. For instance, [5] presents a case study comparing HLS and hand-written RTL implementations, where HLS achieves a drastic reduction in delay. Another example is a convolutional neural network developed in [6], demonstrating the ability of HLS to support complex algorithms. Additionally, there are many practical applications of HLS where multiobjective optimization was applied, for example, custom processor design to find an optimized architecture [7], watermarking to provide protection of authorship in reusable Intellectual Property (IP) [8], and exploration of a low-cost Trojan security hardware [9].

1.2. Multiobjective Optimization in High-Level Synthesis.
There are several opportunities to perform optimizations in HLS during scheduling, allocation, and binding. These optimizations are highly multiobjective by nature, with conflicting objective functions. To deal with that scenario, it is necessary to apply Multiobjective Optimization Algorithms (MOAs). These algorithms maintain a trade-off between conflicting metrics. Multiobjective optimization is dedicated to solve problems in which a set of objective functions \( f_1(x), f_2(x), \ldots, f_m(x) \) must be optimized simultaneously. A multiobjective optimization problem where all objective functions should be minimized can be defined as

\[
\arg \min_{x \in D} F(x) = (f_1(x), f_2(x), \ldots, f_m(x)),
\]

where \( D \) is known as the decision space. The image set \( O \) which results from projecting \( F: D \rightarrow O \) is called the objective space, which is the space where the objective vectors belong. An objective vector \( v = (v_1, \ldots, v_m) \) dominates \( u = (u_1, \ldots, u_m) \) if and only if all the components of \( v \) are equal or better than the corresponding components of \( u \) and at least one component of \( v \) is strictly better. For a multiobjective optimization problem where all objective functions are of minimization, Pareto dominance can be defined as

\[
\forall i \in \{1, \ldots, m\}: v_i \leq u_i \land \exists i \in \{1, \ldots, m\}: v_i < u_i.
\]

A point \( x^* \) is Pareto optimal if there is no other solution \( x \in D \) that dominates it. The set of optimal Pareto solutions are the Pareto optimal set \( P^* \). The Pareto Front (PF) is the image of the Pareto optimal set in the objective space \( P^* = F(x), x \in P^* \) \([10]\). Solutions to this problem should approximate the Pareto Front, instead of a single solution. The solution quality is commonly expressed in terms of Pareto dominance.

Always, it is desired to find an approximation with good convergence and diversity. Convergence is the proximity to the set of ideal points. Figure 1 provides two examples of PF approximations (minimization of two objective functions). The first plot (left) contains a set of solutions where some regions are not covered, so this PF is not attractive because the decision maker could lose important information of the PF. The second one (right) shows a front having a very good spread of solutions (diversity).

According to the HLS literature, the authors have tried to optimize the following objective functions, as shown in Figure 2.

(i) \( \text{Delay} \) is the total number of time steps or clock cycles. It is also called control step, timing, latency, or performance. This objective can be replaced by throughput, which is given as the ratio of the operating frequency to the latency multiplied by the input size. These system-level specifications are defined by the behavioral description.

(ii) \( \text{Area} \) is the total of occupied components in the device, i.e., FUs plus registers \([11]\). It is also called memory or space.

(iii) \( \text{Power} \) is the total power consumption (dynamic power plus static power).

(iv) \( \text{Wire length} \) is the measure of the overall interconnection length plus connectivity components used by the design, which is based on a global routing step. It is also called an interconnection or data path. This measurement must be computed out after binding.

(v) \( \text{Digital noise} \) is an estimation of computational errors plus noise propagation when the design contains real numbers, considering floating-point accuracy. When real numbers are represented by a limited number of bits, this causes a loss of information, which is usually considered as noise. It is also called error propagation.

(vi) \( \text{Reliability} \) refers to the need to avoid the presence of soft error (intermittent failure caused by neutrons and alpha particles). The probability that a soft error will occur depends on which types of FUs are used for the design operations since some FUs are ideal for certain types of operations.

(vii) \( \text{Temperature} \) should be minimized for every design because temperature variations and hotspots inside an FPGA can cause electronic failures.

(viii) \( \text{Security} \) is the protection against attacks, for instance, IP protection and reverse engineering attacks. This objective is also called robustness.

Most of these metrics should be minimized; only reliability and security must be maximized. According to the review of the state of the art that we have made in this survey, we find that multiobjective optimization works assume that the objective functions are in conflict but only one work verifies that some of the objective functions are in conflict \([12]\). According to the above, Table 1 presents a summary of
the possible conflicts between the eight objective functions. With the symbol ✓, we mark the objective functions where a conflict was demonstrated through a payoff matrices [12]. The objective functions that some authors have assumed are in conflict because they use a multiobjective approach are shown with the symbol ✓. Then, with the symbol ≈, we indicate the objective functions that we hypothesize are in conflict, according to what is known of the internal structure of FPGA devices. Finally, the symbol ? means that we do not know if the two objective functions are in conflict.

1.3. Contribution. Figure 3 is an Euler diagram of the optimization methods applied to HLS and highlights in black the subject area of this paper. The intersection of the optimization methods with three main stages (HLS, logic synthesis, and layout synthesis) involved in circuits implementation into FPGA devices is shown in [13]. Considering that multiobjective optimization is a subarea of optimization, this paper focuses on the multiobjective optimization of HLS for FPGA devices. For instance, HLS with single-objective optimization is not considered in this survey.

In summary, the novel contributions of this survey include the following:

1. A review of the state of the art on HLS techniques with multiobjective optimization
2. A description and comparisons of MOAs applied to HLS, analyzing optimization methods, HLS techniques, intermediate structures where optimization is performed, objective functions, the cost assignment strategies, and the benchmarks employed for experimentation
3. Identification of major research challenges in this area that should be studied in the near future and notes on how to tackle them, including a hypothetical grand challenge to carry out HLS as a many-objective optimization problem with eight objective functions

The rest of the paper is organized as follows. Section 2 discusses the related survey, while Section 3 provides an overview of multiobjective optimization techniques in HLS. In Section 4, open issues are presented. Finally, we discuss our conclusions in Section 5 and outline future work in this area.

2. Related Surveys

The origins of HLS can be traced back to the ALERT system [14], developed by IBM at the T. J. Watson Research Center in 1969, but it was not until 2003 [15] that this task was studied as a combinatorial multiobjective problem for FPGA devices. Since then, several surveys concerning
optimizations (regardless of the number of objective functions) in HLS for FPGA devices have been published. The work of [16] provides a taxonomy of optimization in HLS on the basis of the intermediate representation used, such as a Data Flow Graph (DFG) or Control DFG (CDFG), and the performed tasks in HLS, namely, scheduling, allocation, and binding. It also enumerates research based on transformations of initial behavioral descriptions. The survey of [17] includes several approaches and frameworks for HLS optimization. For the first time in this area, a manuscript mentions multiobjective optimization and even explains some objective functions. It also presents details inside the optimization techniques, for example, how different types of internal structures are used to perform the optimizations. Four years later, [18] describes a retrospective of HLS and also explains the algorithms and academic software to apply optimization approaches.

Reference [19] presents a survey of memory, power, and temperature optimization techniques in HLS, explaining how these objective functions had been handled and the importance of analyzing the relationships (trade-off) between them. They also wrote notes about open issues, such as the order of optimization and code generation for low power. The survey presented in [20] deals with the three most popular objective functions: delay, area, and power. The paper also presented methodologies for multiobjective optimization and a classification of metaheuristics that were used. A review of bioinspired optimization techniques was presented in [21], including a few evolutionary multiobjective approaches. They presented details about using both evolutionary computation and hardware design. The state of the art of HLS software tools is investigated in [22], which includes comparisons and evaluations of some software tools. The authors also present a taxonomy of input languages in software tools. Although the survey provides a comprehensive analysis of HLS software tools (commercial and academic), it does not mention which tools perform multiobjective optimization. The overview presented in [23] mentions strategies to solve the DSE problem by reducing the design space-time. The techniques are compared based on their performance improvement. They also include a few multiobjective approaches and performance metrics formulations.

In summary, it is important to note that none of the previous papers are completely focused on the subject area of MOAs in HLS for FPGA devices, the main contribution of this paper.

### 3. Multiobjective Approaches in High-Level Synthesis for FPGA Devices

In this section, the state of the art of MOAs in HLS for FPGA devices is presented.

In order to provide a visual representation of this survey, we created an online relational graph available at http://201.174.122.25/moo_hls_fpga [24].

The graph was created with the [25] library and arranged by the edge-weighted force-directed algorithm; the graph is shown in Figure 4. This graph allows you to search papers on multiobjective optimization in HLS. Circular gray nodes are papers in the state of the art and the number of citations is calculated using Google Scholar and is represented by the size of the circle. When a paper is selected, the paper is a blue node. The multiobjective methods are classified in the red box. In the purple box, the MOAs are organized. In the light blue box, the cost assignment strategies are shown; the objective functions are shown in the yellow box. In the green box, the benchmarks are classified; finally, in the blue box, the compiler techniques are shown.

The edges are the connection between the papers with the multiobjective method, the cost assignment, the objective functions, the benchmarks used in the paper, the

<table>
<thead>
<tr>
<th>Objective functions</th>
<th>Delay</th>
<th>Area</th>
<th>Power</th>
<th>Wire length</th>
<th>Digital noise</th>
<th>Reliability</th>
<th>Security</th>
<th>Temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delay</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
</tr>
<tr>
<td>Area</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
</tr>
<tr>
<td>Power</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>Wire length</td>
<td></td>
<td>=</td>
<td>=</td>
<td></td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
</tr>
<tr>
<td>Digital noise</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
</tr>
<tr>
<td>Security</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>≈</td>
<td>=</td>
<td>=</td>
<td>=</td>
</tr>
<tr>
<td>Temperature</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>≈</td>
<td>≈</td>
<td>=</td>
<td>≈</td>
</tr>
</tbody>
</table>
compiler techniques, and the MOA. Figure 5 is an example of the graph, the paper [26] is the gray node, the multi-objective method is a branch and bound, Pareto dominance is the cost assignment strategy, DFG and CDFG are the compiler techniques, while branch and X is the MOA, area and power are the objective functions, and finally, experiments were carried out on the CDFG toolset benchmark. In this case, the gray nodes [27, 28] are the papers that have at least one author in common with respect to the blue node.

3.1. Optimization Approach. The optimization approaches can be classified into the following two categories [29]:

1) Compiler Techniques. Behavioral description is represented by an acyclic directed graph, such as DFG, CDFG, Synchronous-DFG (SDFG), Loop-Array Dependency Graph (LADG), Timed Marked Graph (TMG), Sequencing and Binding Graph (SBG), Prefix Graph, Problem Graph, and Specification Graph (CDFG and DFG are the most used, see Figure 6). All these intermediate structures have the same intention, to represent the semantics of the behavioral description. This technique requires, before optimization, converting (compilation) the behavioral description into the structure and, after the optimization, converting the optimized structure to RTL (RTL generation). Figure 7 presents the general framework of compiler techniques. The behavioral description and the components library are inputs, where the latter describes the characteristics of the FPGA device. The multiobjective optimization process must perform scheduling, allocation, and binding. Generally, the output is an
HDL code that is ready for an EDA software tool to perform logic and layout synthesis.

(2) **HLS Tool as a Black Box.** These are approaches that explore the design space using commercial and academic HLS tools as black boxes. These approaches invoke the software tool to choose the objective functions. This technique is more comfortable to code because there is no need to worry about compilation, RTL generation, and estimations; but it is strongly dependent on the selected software tool. The variations of the scheduling, allocation, and binding are made through simulation tasks, knob settings, pragma directives, or profiling annotations inside the behavioral description. Figure 8 presents the general framework of approaches that use HLS tools as a black box. This methodology has a higher computational cost because in each iteration the selected HLS tool has to recompile the behavioral description and regenerate the RTL.

Figure 9 shows a taxonomy of multiobjective methods based on [30]. In this survey, we focus on the highlighted boxes which are MOAs in HLS.

Six multiobjective methods have been studied by authors in this domain, organized as exact or approximate methods. According to Figure 9, these methods are branch and X, problem-specific heuristics, single-solution-based heuristics, learning-based methods, evolutionary algorithms, and swarm intelligence systems. Figure 10 shows six multiobjective methods represented in our relational graph in the state of the art. For instance, in the graph, we can see that the branch and X method is the least used since it has the least amount of edges. On the other hand, the most used multiobjective method is the swarm intelligence system.

Next, we explain each approach highlighted in Figure 9.

1. Among exact methods [30], branch and X searches over the whole solution space, which is explored by dynamically building a tree whose root node represents the problem being solved. The optimization is performed by subdividing the problem into simpler subproblems.

2. **Problem-specific heuristics** are, as the name implies, methods that are based specifically on the problem. They can achieve good results but cannot be applied generically to other problems.

3. **Single-solution based metaheuristics** function as a classifier or regression problem using a training set of instances. Then, the model acts over the decision-making process.

4. **Learning-based methods** approximate the PF using machine learning models that learn by posing a classification or regression problem using a training set of instances. Then, the model acts over the decision-making process.

5. **Evolutionary algorithms** are population-based metaheuristics, where solutions are selected and reproduced using variation operators (for instance, mutation and recombination). The main components to design an evolutionary algorithm are the following: representation, selection strategy, reproduction strategy, and replacement strategy. Population-based metaheuristics share common concepts. They start with a random initial population. Later, a new population is created in each generation that replaces the current population. This process iterates until a stop criterion is met.

6. **Swarm intelligence systems** are another population-based metaheuristic, and these systems are inspired by the collective behavior of species such as ants, bees, and wasps. The key features of these algorithms are simple and nonsophisticated agents; they move in the search space and cooperate with each other by an indirect communication [30].

In [30], it is argued that it is better to use population-based metaheuristics than exact methods for multiobjective optimization problems. The reason is that, with exact methods, if the number of objective functions increases, then the algorithm design is more complex. In the same way, population-based metaheuristics are better than single-solution based metaheuristics because the population of solutions helps with diversity preservation on the PF, and consequently the convergence too.

On the other hand, in multiobjective optimization, to be able to compare solutions, it is necessary to apply cost assignment strategies. For a given solution, a cost assignment strategy maps to a cost vector (several objective functions) into a single value. Figure 11 shows a taxonomy of cost assignment strategies based on [30], where we are highlighting works of HLS in the literature.

Next, we describe in chronological order specific works of HLS for FPGA devices.

3.2. **Branch and X Approaches.** In [26, 27], a branch and bound algorithm was developed, which is capable of
generating a nondominated solution with the CDFG toolset [31]. Over a CDFG, a multiobjective optimization is carried out with a Pareto dominance technique considering area and power as metrics (see Figure 5). One year later, in [28], published by the same authors, a biobjective proposal with similar characteristics was presented, taking into account FUs that support dynamic voltage and frequency. The publications of the branch and X method are scarce, due to their ineffectiveness dealing with multiobjective problems and their high probability of being stuck in a local optimum.

3.3. Problem-Specific Heuristic Approaches. The paper [32] was the first application of a Fuzzy Inference System (FIS) to optimize this problem with a multiobjective focus. Three proposals based on the DFG are presented: a module selection scheme in the HLS using fuzzy logic, an allocation process of DFG, and scheduling of DFG with processing times characterized by fuzzy sets. Two years later, [33] presented another problem-specific heuristic based on the decomposition of an Architecture Configurations Graph (ACG). In [34], a greedy algorithm to optimize delay and area was studied. The authors analyzed in detail the estimations of the objective functions. Two years later, [35] explored power-area trade-offs in HLS through dynamic FU allocation with a network flow rebanding using a DFG

![Figure 6: Graph of the intermediate structures used in the state of the art.](image)

![Figure 7: General framework of compiler techniques.](image)
representation. In [36], a hierarchy factor method to simultaneously optimize delay, area, and power was studied. The authors in [37] studied a greedy algorithm to minimize the area, power, and digital noise as objective functions. The authors introduce an analytical precision analysis approach based on a quantization error propagation model. Sengupta et al. presented several papers with a priority factor-based heuristic [1, 38–41]. The proposed approaches try to resolve several issues related to the DSE, such as the precision from evaluation, when the time is exhausted during the evaluation, and also automation of the exploration process. Furthermore, scheduling, allocation, and binding were tested with several DSP benchmarks and real-world problems. At the same time, [42] introduced a hybrid priority factor-based heuristic and FIS employing an aggregation method and fuzzy dominance to optimize delay, area, and power. The proposed hybrid exploration was applied to different DSP benchmarks, and these methods provide acceleration compared to some DSE approaches. A DSE by hybrid priority factor-based heuristic and FIS is presented in [43]. It is a combination of the priority factor method and fuzzy search technique that is rapid and accurate, used in the evaluation and selection in the architecture design space. Other hybrid approaches that use an aggregation method are presented in [44], which is a combination of a priority factor-based heuristic and a dependency matrix algorithm. This iterative heuristic method has a considerably good exploration runtime while delay and area are used as the objective functions. Krishna et al. [45] proposed a different hybrid heuristic, which is a combination of a priority factor-based heuristic and greedy algorithm to optimize delay and power. This work also has a design with less execution time, providing increased acceleration when it is compared with other iterative proposals. Another FIS was presented in [46], this time with the cost assignment strategies of fuzzy dominance. It achieves significant improvement in speedup with a real benchmark. A brute force search based on adders and multipliers is presented by [47]. The authors considered the code level transformations together with the architectural level optimizations and their impact on the scheduled data path. The same authors optimized delay and area again, but this time

Figure 8: General framework of the HLS tool as a black box.
with a gradient-based heuristic pruning [48, 49]. The work [50] presents a clustering method that acts over pragma directives to optimize delay and area using a PF approximation. In [51], a schedule and binding heuristic with network flow rebinding is described. Their work employs a dynamic FU allocation strategy in HLS to achieve a compromise between power and area. References [52–54] proposed an algorithm to explore the design space using binary search employing an ACG. Alternatively, the problem of DSE was addressed in [55, 56] by a D-logic based exploration. These are mathematical models for the power, delay, and area metrics that deterministically prune the vast design space into a subset of valid design variants without compromising the speed and the quality of the design.

The HLS design requires an efficient exploration approach with the ability to determine optimal/near-optimal scheduling solutions and module selection with significant speed and precision. Based on this idea, [57] introduced a heuristic based on the primacy selector (s-value) metric which is common with the matrix topology methods. Most of the research has focused on using an HLS tool as a black box with pragma directives, and [58] is another example. In this case, the divide and conquer algorithm with the CHStone benchmark was used [59]. By profiling annotations, in [60, 61], a greedy algorithm to optimize delay and area with an aggregation method is presented. This methodology is completely autonomous and it incorporates area and frequency like constraints. The work of [62] presented a fully automated C-to-FPGA framework to address this problem. This technique can satisfy hardware resource constraints (scratchpad size) while still aggressively exploiting data reuse. This approach can also be used to reduce the on-chip buffer size subject to bandwidth constraints. In [29], an iterative method with pruning that can deal with the DSE of multiple loops on FPGAs is described.

Many methodologies have been introduced which are capable of drastically reducing the number of variants to be analyzed for the selection of the optimized design using the minimal execution time. The paper [63] presented a problem-specific heuristic based on a graph merging approach to deal with delay, area, and power. Allocation and scheduling of reconfigurable arrays are implemented in Verilog HDL and synthesized by an RTL representation using the Xilinx ISE Design Suite. The graph merging approach is validated by the results which showed that the area allocated is less for the graph merging technique than the reconfigurable array using multiplexers. Concerning the objective

![Figure 9: Taxonomy of multiobjective methods in HLS for FPGA devices.](image-url)
function digital noise (as well as area), [64] studied a bit-width optimization by a divide and conquer algorithm for fixed points. In [65], the authors present a hierarchical DSE method that can speed up the exploration and can also perform incremental DSE avoiding rerunning a full exploration—by an HLS tool—each time that the changes in the source are made, a Cyclic Redundancy Check- (CRC-) based method is used to detect changes at the behavioral description (source code).

Pham et al. [66] proposed a heuristic based on an access pattern simulator by a LADG to reduce the dimensions of the design space. A scheduling and binding heuristic for HLS of fault-tolerant FPGA applications is presented in [67]. The authors stated that integrating redundancy into the HLS is an attractive approach that enables synthesis to rapidly explore different trade-offs at no cost to the designer. In [68], the authors present a multiobjective optimization with quick estimates of cycle count and FPGA area usage for designs in the Delite Hardware Definition Language (DHDL). Their estimations take into account available off-chip memory bandwidth and on-chip resources for data path and routing, as well as effects from low-level optimizations like LUT packing and logic duplication. A year later, linear programming for multiobjective optimization is studied by [69, 70] and a colored interval graph approach is studied by [71, 72].

3.4. Single-Solution Based Metaheuristics. Aggregation method consists of changing a multiobjective optimization problem into a monoobjective one or a set of such problems. It consists of using an aggregation function combining various objective functions $f_i$ into a single-objective function $f$ generally in a linear way:

$$F(x) = \sum_{i=1}^{n} \lambda_i f_i(x), \quad x \in D,$$

where the weights $\lambda_i \in [0...1]$ and $\sum_{i=1}^{n} \lambda_i = 1$. However, the use of scalarization approaches is only justified when they generate Pareto optimal solutions [30]. Zwolinski and Gaur [15] optimized the delay and area by an aggregation method, scaling from multiobjective to monoobjective with a simple weight vector (see equation (3)). Within the next two years, another three approaches from this emerging research field were published, and one of them was [73] with single-solution based metaheuristics. On this occasion, simulated annealing, random search Pareto, and tabu search algorithms were used, with the peculiarity of selecting weak dominance as the cost assignment strategy. In [74], a similar approach employing simulated annealing was used. This time with pragma directives instead of simulation configurations. The paper [75] studied the trade-offs between power and security estimations on a CDFG. This paper considers IP protection as a new objective function of the DSE.

3.5. Learning-Based Methods. Machine learning methods have been used in recent years, almost all of them using HLS tools as a black box (see Figure 8). These techniques always perform scheduling, allocation, and binding because HLS tools are responsible for carrying them out. In [76],
A machine learning algorithm is presented, where the authors determine the PF approximations by only sampling and synthesizing a fraction of the design space. A DSE to derive PF approximations of the design configurations for a set of targeted metrics (in this case delay and area) is developed in [49]. That work used a response surface method with Pareto dominance to perform scheduling, allocation, and binding. In the same year, [77] investigated a methodology based on random forest and results compared favorably with other black box alternatives. This research optimizes simultaneously the same objective functions (delay and area), but this time using knob settings to create variations on the search process. One year later, a machine learning approach based on simulated annealing was created for HLS of the DSE [78] using pragma directives. This approach employs a standard simulated annealer to generate a training set and uses this set to implement a decision tree. The delay and area optimization developed by [79] used an Adaptive Threshold Non-Pareto Elimination (ATNE). This approach focuses on understanding and estimating the inaccuracy, instead of focusing on regression accuracy improvement. They employed five OpenCL applications as behavioral descriptions to perform experiments.

An alternative strategy was proposed in [80], called cluster-based heuristic, an open-source project. The exploration methodology is divided into five steps: initial sampling, clustering, cluster selection, intracluster exploration, and intercluster exploration. Ma et al. [81] presented a Gaussian process regression to optimize simultaneously delay, area, and power. Machine learning is applied to predict the PF approximation of the adders in the physical domain, because it is infeasible to exhaustively run the HLS tools for many architectural solutions. On the other hand, [82] developed HyperMapper 2.0, a methodology and corresponding software framework, which handles multi-objective optimization in the DSE for FPGAs. This methodology also can incorporate prior knowledge from the user in the search. Another random forest approach is presented in [83], which focuses on hardware loop unrolling with an HLS directive.

3.6. Evolutionary Algorithms. Evolutionary algorithms have been good candidates to tackle DSE. The first one was in [84], making use of a Weighted Sum Genetic Algorithm (WSGA). This is the first proposal where area and digital noise are objective functions. Additionally, the same authors propose an extension with a similar DFG-based methodology [85–88], but with power as an additional objective function. One of the most important contributions in this field was offered in [89], since they explained the use of a multichromosome approach, and consequently, it was
more feasible to represent the scheduling and allocation tasks concurrently.

Strength Pareto Evolutionary Algorithm 2 (SPEA2), using Pareto dominance, is an algorithm that performs a much more intelligent multiobjective search. In 2006, it was used for the first time in this problem by [90] with two objective functions and in [91] with three objective functions. Another evolutionary algorithm that uses Pareto dominance is the Nondominated Sorting Genetic Algorithm II (NSGA-II), which uses the crowding distance as a diversity preservation technique. This algorithm was used for the first time in HLS for FPGA devices in [92–94] to optimize delay and area. One year later, these same proposals were improved, in terms of the representation of the solutions (encoding) in [95–97].

In [98, 99], a dynamic combination of WSGA and Ant Colony Optimization (ACO) is presented. In this method, the initial pheromone distribution is generated with WSGA, and then ACO is used to obtain the solutions. Dynamic switching conditions are also discussed. In [100, 101], the SystemCoDesigner software tool was presented, which offers a fast DSE and rapid prototyping of behavioral SystemC models. The work [102] presents a multiobjective evolutionary algorithm for hardware-software partitioning of embedded systems, and the MediaBench benchmark [103] was selected for testing. Anderson and Khalid [104] applied the Simple Evolutionary Algorithm for Multiobjective Optimization (SEAMO), a genetic-based algorithm to prune the design space of the parametrized core and determine a PF approximation by simulation. Speeding-up expensive evaluations in HLS using solution modeling and fitness (cost) inheritance are presented in [105]. They use NSGA-II with CDFG for delay and area optimization. The works [11, 106, 107] present a different approach with respect to the previous ones. The research employs the multichromosome representation presented in [89] but incorporates an accurate power estimation. The methodology based on NSGA-II was evaluated through the MediaBench benchmark in DFG, and the results indicate that it yields improved solutions with better diversity compared to a WSGA approach.

In [108–111], the authors solved scheduling, allocation, and binding using WSGA. The presented approach incorporates a new seeding process for two special parent chromosomes as well as a load factor heuristic, which guarantees that the final solution will always be near-optimal in terms of the user-specified constraints. In [112], a fully automated design flow that exploits multiobjective DSE to enable runtime resource management is studied. They developed a technique that identifies the most promising operating points by using profiling information coming from both software simulation and hardware synthesis. The optimization is done by using the Greedy Evolutionary Multiobjective Optimization (GEMO) algorithm. Schafer and Wakabayashi [113] demonstrate the feasibility to apply NSGA-II in conjunction with a machine learning-based predictive model. It is an HLS tool based on a black box method that creates a predictive model from a training set until a given error threshold is reached. Then, it continues with the exploration using the predictive model avoiding time-consuming synthesis and simulations of new configurations. HLS for FPGA devices by Learning Automata Genetic Algorithm (LAGA) is studied in [114]. According to this work, the scheduling and allocation are performed over a DFG, optimizing delay and area simultaneously.

In another example, [115] presented a technique for area-delay trade-off using residual load decoding heuristics with genetic algorithms for integrated DSE of scheduling and allocation. They employed the aggregation method as a cost assignment strategy. The work [116] summarizes a set of techniques that were presented in previous papers, the main one being [11]. This work explains how to deal with the simultaneous optimization of delay, area, and power. In the same year, [117] released another chromosome representation along with a driven integrated exploration of loop unrolling factor and data path by WSGA for scheduling of the CDFG. In [118], a DSE methodology for the optimization of delay and area by an evolutionary algorithm based on pragma directives is presented. One year later, [119] described another application of NSGA-II for the optimization of delay and power with the NCBI BLASTP benchmarks [120]. Other methods use a predictive model to avoid having to resynthesize each new configuration to be explored. In [121], a dedicated DSE for FPGAs is presented that is based on a pruning algorithm with an adaptive windowing method to extract the design candidates to be further (logic) synthesized after HLS. The adaptive windowing is based on a learning method inspired by the Rival Penalized Competitive Learning (RPCL) model in order to classify which designs need to be synthesized.

In [12], an approach to apply two optimizations consecutively is presented. As the first optimization, several metaheuristic algorithms for multiobjective optimization were applied in HLS based on [116]. As a second optimization, reductions of LUTs at the logic synthesis stage were carried out. The paper showed how several optimizations belonging to different design stages can coexist. One year later, as an extension, a many-objective optimization algorithm—Nondominated Sorting Genetic Algorithm III (NSGA-III)—was applied in [122] for the first time to this problem. In [123], a delay and power optimization is proposed. In this case, an SDFG is employed for modeling DSP applications. In [124], the authors focused on finding the smallest microarchitecture for a specific target latency. They used pragma directives with the S2CBench benchmark [125]. The authors of [126] incorporated a new dimension to the multiobjective optimization of this problem, the reliability. This methodology is composed of two main phases. The first one performs HLS for DSE leading to a trade-off curve of designs with delay, area, and reliability. The second phase finds the most reliable system given delay and area constraints by either implementing time or space redundancy, or a mixture of both using any combinations of microarchitectures found by the explorer.

3.7. Swarm Intelligence Systems. This family of algorithms did not appear in this domain until 2006, when [127, 128] implemented an ACO to perform scheduling and allocation.
taking into account the objective functions delay and area. A comparison between Particle Swarm Optimization (PSO) and the evolutionary algorithms NSGA-II and WSGA was made in [11]. According to their work, it is observed that, compared to WSGA, PSO shows considerable improvement in runtime with a comparable solution quality. The proposed integrated approach in [129] comprises a comprehensive mapping process and a sophisticated strategy for evaluating solutions. They introduced a PSO driven DSE methodology for delay and power trade-off over a CDFG.

An adaptive DSE framework called integrated Particle Swarm Optimization (i-PSO) for delay and power as objective functions in HLS is presented in [130], including a sensitivity analysis of the algorithm. The use of PSO for the DSE of data paths in HLS is also proposed in [131–134], and the MediaBench benchmark and another DSP benchmark (the paper does not provide details of the benchmark name) to measure the optimization quality of the simultaneous exploration of data path and loop unrolling factor are used. Other authors published a similar strategy, but delay and area were optimized in [135, 136]. The authors in [137] describe an approach to solve the DSE problem, which is based on the Bacterial Foraging Optimization Algorithm (BFOA). They also study BFOA in a similar way, where delay and power were optimized in [138–144]. The proposed exploration approach is simulated to operate in the feasible temperature range of an Escherichia coli bacterium in order to mimic its biological life-cycle. Mishra and Sengupta [7] studied the trade-offs between delay and power proposing MOPSE, an adaptive multiobjective PSO based on DSE. Sengupta and Mishra [145] described an approach to solve the DSE problem based on Weighted Sum Particle Swarm Optimization (WSPSO) with two variants of the acceleration coefficient, a hierarchical time-varying acceleration coefficient and a constant acceleration coefficient.

A compiler approach performing delay, area, and power optimization is presented in [146], where a better behavior of the firefly algorithm over simulated annealing (single-solution based metaheuristic) stands out. This metaheuristic has a competitive execution time, compared, for instance, with an evolutionary algorithm. Research in [136] described a methodology based on automating DSE and loop unrolling factor using the high-level transformation during area-delay trade-off using the PSO. Using CDFG, [147–155] described approaches based on the k-cycle transient fault secured data path during the HLS. Bhuvaneswari [116] studied a multi-chromosome structure on a DFG to optimize delay, area, and power using several algorithms, including swarm intelligence and evolutionary algorithms. Multiobjective optimization is performed in [9, 156], considering an interesting topic, the secure information processing against transient fault secured design through a hybrid transformation with delay and area objective functions by a PSO. In [168], a BFOA to achieve low-cost (delay and area) IP design is performed. And [169] studied a PSO to achieve delay and power minimization combined with an IP functional locking.

Other examples within the use of this type of metaheuristic can be found in [8, 159], and a multivariable signature encoding for embedding a dynamic watermark in an IP design was presented. These investigations used the same DSE framework with the PSO optimizing delay and area. The authors of [160, 161] proposed a firefly algorithm for scheduling and allocation on the DFG using the MediaBench benchmark and another DSP benchmark (in the paper, no details of the benchmark name are given). Besides, these papers report a sensitivity analysis that provides a good tuning of the algorithm control parameters for performing the DSE that leads to faster convergence.

Obfuscation is the process of transforming an original application or design into a functionally equivalent form to make the reverse engineering process significantly more complex. The authors in [162] provided a structural obfuscation methodology for protecting IP core at the HLS design stage. The proposed approach specifically targets the protection of IP cores that involve complex loops. The authors of [163, 164] created a multiobjective optimization (delay and area) that can deal with low-cost functional obfuscation of reusable IP cores. The work in [165] was the first to incorporate the switching device and the storage element delay from scheduling during the delay estimation. They provide a BFOA that gives a balanced DSE methodology and includes comprehensive delay estimation by considering the combined delay of FUs, the switching devices, and the storage elements directly from scheduling. Results indicate improvement in achieving a more realistic delay estimation process than previous approaches.

In [166], the authors presented an optimization of delay and area of the obfuscated JPEG CODEC IP core design using particle swarm based on the DSE. And [167] introduced an obfuscation of fault secured design through a hybrid transformation with delay and area objective functions by a PSO. In [168], a BFOA to achieve low-cost (delay and area) IP design is performed. And [169] studied a PSO to achieve delay and power minimization combined with an IP functional locking.

3.8. Analysis, Comparisons, and Main Findings. Figure 12 shows all the MOAs used over the years. This chart evidences, in addition to the increase of papers over the years, that swarm intelligence systems have been the most studied. Evolutionary algorithms have also been used, due to the simple way in which chromosomes can be generated.

Analyzing the cost assignment strategies in HLS, scalar approaches have been the most used. Among these strategies, the aggregation (or weighted) method has been the only one studied, due to its simplicity. During the last decade, more methods such as dominance-based approaches and indicator-based approaches have been used.

In Figure 13, the cost assignment strategies used over the years are shown. In Figure 14, the cost assignment strategies are shown, in the proposed relational graph, where the aggregation method and Pareto dominance are the most used.
Objective functions have been estimated, represented, and calculated in different ways (especially delay, area, and power, as seen in Figure 15). The authors have proposed many ways to represent the circuit design, and therefore, the estimations have to be coupled to the data structure of the representation (for instance, chromosome representation in evolutionary algorithms). At least one of the delay, area, and power metrics is present in almost all works covered in this survey. For that reason, their estimation methods have become sophisticated over time. In Figures 16 and 17, we can see the objective functions using compiler techniques and an HLS tool as a black box; it is important to note that delay, area, and power stand out from the rest.

Furthermore, thanks to a technique called payoff matrices, [12] showed how the objective functions of delay, area, and power are in conflict. It demonstrates the importance to solve this problem with a multiobjective approach. However, until now, the optimization process has not been solved considering all eight objective functions simultaneously. The papers that used the most objective functions have been [86–88] dealing with four (considered to be a many-objectives optimization problem). In the case of the optimizations with HLS tools as a black box, the objective functions have been delay, area, power, and reliability because those are the ones that can be obtained from the software tools.

Regarding the optimization method, diversity is as important as convergence. Therefore, MOAs should have techniques for diversity preservation with statistical density estimations. In this sense, the following techniques have been applied in HLS for FPGA devices: nearest neighbor and histogram ([30]; p. 343). These techniques are implicit inside many MOAs mechanisms. For example, the NSGA-II uses the nearest neighbor technique (with crowding distance) and the NSGA-III uses the histogram (with reference points).

The benchmarks used to evaluate these techniques are very important for the experiments, comparisons, and validation of the results. In the state of the art, we can find
that MediaBench, also called Express benchmark, is the most used. MediaBench was introduced in [103] for performance evaluation of solutions on microprocessor architectures applied to multimedia and communication systems. Figure 18 shows the benchmarks used in the state of the art considered in this survey.

Nonetheless, many papers have used DSP benchmarks like [170] or real-world benchmarks like [33]. The benchmark proposed in [31] is used by the authors who proposed using the branch and X approach. The S2C Bench benchmark [125] was employed for optimization proposals that use pragma directives for the search process. On the other hand, the PERFECT benchmark [171] is referenced in [69, 70] for an accelerator of Wide-Area Motion Imagery (WAMI) applications for SystemC specifications. Schafer et al. have used the [172] in [50, 74, 113] with pragma
directives to optimize delay and power. The experiments in [62, 66, 123] are performed for five applications from the polyhedral benchmark suite (PolyBench) [173], a benchmark for testing loop and array related problems. Another benchmark used with pragma directives and profiling annotations was CHStone, a benchmark program suite for practical C-based HLS [59]. It was used by learning-based methods and problem-specific heuristics. Other benchmarks of less use were the ACM/SIGDA benchmarks [174] in [7, 116], the Linpack benchmark [175] in [75], the NCBI BLASTP [120] in [119], the BDTI DSP [176] in [162], and the SHOC benchmark suite [177] in [83].

In addition to the benchmarks, a better way to measure the performance of the optimization method is by quality indicators. However, quality indicators have been studied in a few papers, [12, 29, 49, 58, 65, 66, 73, 77, 79–82, 91, 102, 113, 118, 121–123, 134, 145, 178]. Some quality indicators used are Average Distance from Reference Set (ADRS) in [179], Epsilon in [180], Hypervolume in [181], and R in [182]. ADRS was the most frequent quality indicator and is usually represented by percentage. It is based on the normalized distance between two PF approximations \((\mathcal{P}, \mathcal{P})\):

\[
\text{ADRS}(\mathcal{P}, \mathcal{P}) = \frac{1}{|\mathcal{P}|} \sum_{\bar{\mathcal{P}} \in \mathcal{P}} \left\{ \min_{p \in \mathcal{P}} y(\bar{\mathcal{P}}, p) \right\},
\]

where

\[
y(\bar{\mathcal{P}}, p) = \max_{i=1, \ldots, m} \left\{ 0, \frac{f_i(p) - f_i(\bar{\mathcal{P}})}{f_i(\bar{\mathcal{P}})} \right\},
\]

and \(m\) is the number of objective functions. A high value of ADRS reports a low-quality approximation, while a low one indicates that \(\bar{\mathcal{P}}\) has good approximates to \(\mathcal{P}\).

The main findings of these works can be summarized as follows. After reviewing and analyzing the state of the art about MOAs in HLS for FPGAs, we found only one paper that demonstrates that some of the objective functions are in conflict [12]. This is an important aspect; in the rest of the publications, the authors assume that the objective functions are in conflict. On the other hand, with the review of the state of the art, we conclude that there is no survey that allows researchers to contextualize all of the related works related to MOAs in HLS for FPGAs. This paper is intended to be of help to carry out new research in this area. In this survey, we have focused on organizing the papers according to the

---

**Figure 16:** Objective functions using compiler techniques considered in the literature.

**Figure 17:** Objective functions using HLS tools as a black box considered in the literature.
MOAs, the cost assignment strategies, the objective functions, the benchmarks, and the compiler techniques. With this analysis, we have detected that swarm intelligence systems and evolutionary algorithms are the most used. The most used graphs of the intermediate structures are DFG and CDFG. The aggregation method and Pareto dominance are the most used cost assignment strategies. Moreover, of the 8 objective functions studied, the most optimized functions are area, power, and delay. Regarding the benchmarks, MediaBench and other DSPs are the most used in these studies.

4. Open Issues

In this section, future challenges are presented. In addition, some ideas about how to approach them are mentioned:

(1) It is important to use quality indicators to measure the convergence and diversity of the PF, instead of observing the convergence of only some solutions within the PF approximation, as is done in most of the papers. The Hypervolume quality indicator is a good option because it measures the volume of the dominated space bounded from below by a reference point and it is capable of measuring convergence and diversity at the same time [183, 184].

(2) Temperature has been studied in a few papers [185, 186] with a single-objective approach. This objective function should be studied with a multi-objective approach, since the temperature is in conflict with the wire length objective function, because if the use of FUs is increased, then more interconnections will be needed.

(3) The grand challenge is the optimization problem of HLS with 8 objective functions: delay, area, power, wire length, digital noise, reliability, security, and temperature (see Figure 2). We want to push the FPGA designers and researchers to create a new representation for solutions that includes scheduling, allocation, and binding, with which all these objective functions can be estimated. Then, we verify by payoff matrices that these eight objective functions are in conflict. Later, it is necessary to use many-objective optimization algorithms like NSGA-III [187, 188] or MOEA/D [189] to solve the problem. Finally, with many-objective optimization, the results obtained can be analyzed.

(4) To develop more estimations methods for the objective functions: wire length, digital noise, reliability, security, and temperature, this is an area of opportunity where researchers can develop estimations of these metrics with the intention of increasing their potential. One possibility is to use machine learning for this task.

(5) The HLS software tools with multiobjective optimizations should show the PF approximation. Also, these tools should let the designer select the optimization method and configure the most important parameters which are most convenient to him, so the

![Figure 18: Benchmarks used in the state of the art of multiobjective HLS.](image-url)
designer can choose which solution will be implemented into the FPGA device. In [190], visualization techniques are presented which can be used to improve the HLS software tools. In this challenge, the runtime of Multiobjective Optimization Algorithms could be considered a weakness relative to modern tools, such as Vivado HLS. Therefore, to improve this point, we pose the following challenge.

(6) Since multiobjective optimizations require large execution times, it is desirable that the executions of the algorithms are performed on a web server with high-performance computing and parallelization potential, instead of the user side. This can be achieved by developing an HLS web-based software tool with a microservices-based architecture or service-oriented architecture, instead of a monolithic application [191] or using cloud computing to streamline the process.

5. Conclusions

This paper presented the state of the art of multiobjective optimization methods in HLS. An online graph was designed with the aim of creating a visual representation of this survey. In summary, an analysis of the convergence of two fields was carried out: HLS and MOAs. The optimizations methods were identified and classified, as well as internal aspects within them, such as the intermediate structures where the optimizations are performed; HLS techniques; and the benchmarks employed for experimentation. Moreover, this work also studied what cost assignment strategies have been used in the algorithms and which are the objective functions to be optimized. In addition, it was demonstrated that multiobjective HLS is a knowledge area that has been in constant growth since 2003, where a wide range of algorithms and specific details in the scheduling, allocation, and binding techniques have been addressed. To finish, we identified open issues and we mentioned some ideas about how to approach them. The main one is that this problem must be visualized as a many-objective optimization problem with eight objective functions to optimize simultaneously.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work has been supported by Tecnológico Nacional de México/IT Tijuana with the project titled “Identificación de dispositivos de internet de las cosas usando aprendizaje máquina en VHDL”, with the number 7924.20-P. Darian Reyes Fernandez de Bulnes was supported by CONACYT scholarship no. 433536. Special thanks are due to Rogelio Valdez PhD student. We also thank Dr. Daniel E. Hernández Morales and the Instituto Tecnológico de Tijuana for providing financial support in the publication of this manuscript.

References


Scientific Programming


C. Haubelt, T. Schlichter, J. Keinert, and M. Meredith, "SystemCoDesigner: automatic design space exploration and rapid prototyping from behavioral models," in Proceedings of...


