Synthesis with External Don't-Cares Using Shannon Entropy and Craig Interpolation

Kai-hui Chang*, Valeria Bertacco*, Igor L. Markov*†, Alan Mishchenko‡

*EECS Department, University of Michigan, Ann Arbor, MI

†Synplicity, Inc., Sunnyvale, CA

‡EECS Department, University of California, Berkeley, CA

{changkh, valeria, imarkov}@umich.edu, alanmi@eecs.berkeley.edu

ABSTRACT

Traditional digital circuit synthesis flows start from an HDL behavioral definition and assume that circuit functions are almost completely defined, making don't-care conditions rare. However, recent design methodologies do not always satisfy these assumptions. For instance, third-party IP blocks used in a system-on-chip are often over-designed for the requirements at hand. By focusing only on the input combinations occurring in a specific application, one could resynthesize the system to greatly reduce its area and power consumption. Therefore we extend modern digital synthesis with a novel technique, called SWEDE, that makes use of extensive external don't-cares. Experiments indicate that SWEDE scales to large ICs with half-million input vectors and handles practical cases well.

1. INTRODUCTION

Due to the increasing demand for integrated circuits to provide more functions while consuming less power, designing a new chip becomes more and more difficult. One way to reduce this design effort is to reuse previously designed circuits, such as Intellectual Property (IP) blocks and general-purpose processors. This approach, however, may result in designs with unnecessarily large area and power consumption because they are over-provisioned with respect to the target functionality. On the positive side, system performance and cost may be improved by customizing reused components to the target applications and environment. This novel optimization, which we call *design specialization*, poses a new synthesis challenge, which differs from traditional formulations by the abundance of external don't-cares. In fact, both academic and commercial synthesis tools available today appear to be structured and optimized to extract optimization opportunities from small, localized sets of external don't-cares. While this approach succeeds on mainstream synthesis instances, it does not perform well on the type of instances generated in the context of design specialization (and cannot handle some cases at all). Our experimental study revealed that the performance of existing tools, such as Espresso [14] and some commercial synthesis tools, greatly deteriorate when extensive don't-cares are added. In addition, several other tools, such as ABC [24] and many commercial tools, do not handle our problem instances at all, or do not provide specification formats for this situation. It is typical of these mainstream synthesis tools to explore as many ways to exploit a don't-care as possible. When the majority of the terms are indeed don't-cares, however, this approach becomes ineffective because the search space explodes.

In this work we address two types of synthesis problems in the presence of extensive external don't-care sets. The first type assumes that the care-terms are known and represented using a truth table while the circuit structure is unknown. To solve this problem, we propose CleanSlate and InterSynth algorithms that synthesize the truth table from scratch. The second problem type assumes that an initial circuit already exists for customization. To solve this problem, we developed a FastShrink algorithm, which takes as input an optimized design and *reduces* it based on the specified don't-care set. Note that FastShrink might find optimization opportunities even when applied after CleanSlate. Our approach is based on an important insight: extensive don't-cares allow simple greedy algorithms to quickly produce a reasonably small netlist, and the missed optimization opportunities can be recovered afterward using more sophisticated synthesis techniques. Since this latter step does not consider don't-cares, it can run much faster and leverages existing tools. This two-step process eliminates the need for a time-consuming don't-care optimizer, yet it is still capable of generating high-quality netlists. We integrated these techniques in our tool, called *SWEDE* (Synthesis Within an Extensive Don't-care Environment). In our experiments we performed synthesis from truth tables with large don't-care sets and observed SWEDE completing ten times faster than state-of-the-art synthesis tools while producing smaller circuits. We have also used SWEDE to customize circuits with up to 30K gates and half-a-million input vectors in under two hours on a single processor in most cases.

SWEDE's high performance enables several new synthesis applications and enhances many others, including (1) input constraint synthesis for emulation; (2) acceleration of the most-frequent computation in a unit [2, 8]; (3) customization of third-party IP components in an System-on-Chip (SoC); and (4) support for graceful wear-out of electronic devices [19]. Applications in category (1) can solve current engineering problems, while the others provide new system design paradigms. Our techniques may help address a wide range of emerging concerns in IC design, including increasing verification difficulty, unpredictability of manufacturing [19], and lower-power circuits [8]. Since our simplified circuits provide correct outputs only within the specified care set, stimuli outside this realm may not be viable. While "soft" application domains such as multimedia can tolerate these situations well, other applications may require an output flag indicating that a given input cannot be processed correctly.

The rest of the paper is organized as follows. In Section 2 we review previous work and provide necessary background. We propose our new synthesis techniques in Section 3, whose analysis and applications are given in Section 4. Experimental results are provided in Section 5, and Section 6 concludes this paper.

2. BACKGROUND

In this section we first review relevant previous work. Next, we describe three important concepts: bit-signatures, Craig interpolation, and Shannon entropy. These concepts are used in our synthesis techniques.

2.1 Prior Work on Synthesis with Don't-Cares

Much research has been developed in exploiting don't-cares in synthesis optimization. A classic tool implementing some of the most commonly-used techniques is Espresso [14]. Although other more sophisticated synthesis tools exist, such as ABC [24] and MV-SIS [29], these focus specially on synthesis problems with a small number of don't-cares. Moreover, their input specification format makes it impractical to describe a large number of don't-cares. For example, a design that could arise in our problem domain may have 50 inputs and as many as one million care terms, leaving more than 10^{15} combinations to be don't-care terms. In order to specify such a complex set of don't-cares, Gao *et al.* [6] proposed the use of an external netlist to encode them. The construction of such a netlist, however, can be challenging.

In addition to synthesizing from a truth table, it is also possible to optimize a design starting from an existing circuit and simplifying it using the don't-cares via resynthesis techniques such as rewiring [21] and node merging [12]. One major challenge in this context is the representation and manipulation of such don't-cares. For instance, Muroga *et al.* [11] proposed the concept of *Compatible Sets of Permissible Functions (CSPFs)*, which was used by Savoj *et al.* [15] to optimize multi-level networks composed of NOR gates. This representation was later improved by Yamashita *et al.* [21, 22] and became *Sets of Pairs of Functions to be Distinguished (SPFDs)*. One major drawback in these techniques is that representing the don't-cares is cumbersome and the related data structures are difficult to work with. Traditionally, these don't-cares are represented by BDDs, often exhausting all memory resources even for moderate-size designs. To address this problem, Sinha proposed an efficient representation of SPFDs based on graphs that can be used in logic resynthesis [23]. This approach improves the memory profile of SPFDs, but deteriorates the computing time. Recently, Plaza *et al.* [12] relied on bit-signatures generated by functional simulation to approximate observability don't-cares for node merging, followed by SAT-based verification. This approach is faster and more efficient in memory than other solutions. However, external don't-cares were not used in the optimization. In stead of utilizing don't-cares for circuit optimization, techniques based on logic decomposition and refactoring can also effectively reduce the size of a circuit. To this end, the greedy algorithm proposed by Rajski [13] *et al.* is used in our CleanSlate synthesis flow.

2.2 Craig Interpolation

The concept of Craig interpolation originated in mathematical logic in 1957 and has recently become popular in formal verification. In contrast, we are going to use it in logic synthesis.

DEFINITION 1. *Consider a pair of Boolean functions, $A(x,y)$ and $B(y,z)$, such that $A(x,y) \wedge B(y,z) = 0$, where x and z are variables appearing only in A and B, respectively, and y are common variables of A and B. An interpolant of $A(x,y)$ w.r.t. $B(y,z)$ is a Boolean function I over the common variables y that satisfies the following conditions: $A(x,y) \Rightarrow I(y)$ and $I(y) \Rightarrow \overline{B(y,z)}$* [4].

Consider an unsatisfiable SAT instance composed of two sets of clauses A and B. In this case, $A(x,y) \wedge B(y,z) = 0$. An interpolant of A can be computed from the proof of unsatisfiability of the SAT instance by the algorithm found in [9] (Definition 2). The resulting interpolant is a single-output multi-level logic network represented as an And-Inverter-Graph (AIG) [24]. If $A(x,y)$ is the on-set of a function, $B(y,z)$ is its off-set, and $\overline{A(x,y)} \wedge \overline{B(x,y)}$ is its don't-care set, then $I(y)$ can be seen as an optimized version of $A(x,y)$ where the don't-cares are used in a particular way to optimize representation of I.

Interpolation is used in formal verification to compute an over-approximation of the complete set of reachable states [9]. Interpolation has also been used in area- and delay-driven technology mapping into K-input LUTs [10]. When applied to technology mapping, interpolation is used to generate new functions for the node being synthesized.

2.3 Bit-Signatures and Entropy

Our FastShrink synthesis technique is based on bit-signatures generated using simulation, which are defined below. Note that a signature is essentially a signal's partial truth table. If the input vectors are applied exhaustively, then the signature of a signal is its complete truth table.

DEFINITION 2. *Given a wire (signal) w in a circuit, computing function f, and input vectors v_1, v_2 ... v_k, the* signature *of w is the bit-vector $(f(v_1),...,f(v_k))$, where $f(v_i) \in \{0,1\}$ represents the output of f given the input vector v_i.*

The second step of the FastShrink technique (see Section 3.3) exploits short-range optimization opportunities in a circuit. Intuitively, signals with less information are easier to optimize. To quickly identify such signals, we use *Shannon entropy*, which is calculated as follows [17]:

$$E_s = -\frac{\#ones}{k}\log_2(\frac{\#ones}{k}) - \frac{\#zeros}{k}\log_2(\frac{\#zeros}{k}) \qquad (1)$$

In the equation, E_s is the entropy of signature s, $\#ones$ is the number of 1s in the signature, and $\#zeros$ is the number of 0s in the signature. Variable k is the number of bits in the signature and is also the number of vectors applied to the circuit. A larger E means that the signature contains more information.

3. CIRCUIT OPTIMIZATION WITH EXTERNAL DON'T-CARES

In this section we formalize the synthesis problem described earlier and propose three circuit-optimization techniques. One shrinks an existing netlist, while the other two perform synthesis starting from a functional specification (truth table). We then illustrate our techniques by example.

3.1 Problem Formulation

We formulate the circuit-specialization problem as follows. Given a circuit, the complete set of all possible input vectors and their output responses (or, equivalently, a functional specification in the form of a truth table), we seek to produce a small netlist that generates the correct outputs for the given inputs. Our solution considers a combinational flattened circuit and performs the optimization without any structural or other information from the user. On the other hand, if structural information is available in the original netlist, it can be used to improve quality of results.

3.2 Fast Synthesis based on Truth Tables

In this section we introduce two fast synthesis techniques based on truth tables. The first one, called CleanSlate, greedily expands cubes and then performs more sophisticated resynthesis to minimize the size of the netlist. The second one, called InterSynth, is based on interpolation.

3.2.1 The CleanSlate Technique

Our specification-based synthesis technique, called CleanSlate, starts from a truth table and produces a technology-mapped netlist.

The algorithm is outlined in Figure 1: CleanSlate first greedily expands a cube, one literal at a time, similar to the heuristic used in Espresso (lines 1-3). A cube is subsumed by the expanding cube and is eliminated if its outputs are the same as those of the expanding cube. The expansion stops when the cube overlaps another cube with different outputs. After producing an optimized truth table, CleanSlate generates a two-level netlist (line 4), which is fed to ABC for further optimization. Using ABC, CleanSlate first performs fast logic sharing detection of the netlist [13], and then converts the netlist to an And-Inverter-Graph (AIG) [24]. After that, it expands 2-input ANDs in the AIG to multi-input ANDs to create more opportunities for logic sharing detection, and performs AIG resynthesis to optimize the netlist. The procedure in lines 7-10 is applied several times to achieve better optimization (three times in our implementation). At completion, we apply a technology mapping step to produce the final netlist.

The rationale behind our solution is that the large number of don't-cares enables even a greedy algorithm to generate a reasonably small two-level netlist within a short time. We then bypass a time-consuming two-level optimization process, and instead perform multi-level synthesis. As our experimental results in Section 5 indicate, CleanSlate runs 10X faster than exiting tools, handles more complex circuits, and provides better synthesis quality.

```
flow CleanSlate(TruthTable)
 1   foreach row ∈ TruthTable
 2     expand the cube of row until a different cube is reached;
 3     remove other rows in TruthTable subsumed by row;
 4   convert TruthTable to a two-level netlist;
 5   perform fast logic sharing detection of the netlist using [13];
 6   repeat N times
 7     transform the network to an AIG by 1-level structural hashing;
 8     expand 2-input ANDs in AIG to multi-input ANDs;
 9     perform fast logic sharing detection using [13];
10     perform AIG resynthesis (AIG balancing, rewriting
       and refactoring);
11   return netlist by technology mapping the AIG;
```

Figure 1: The CleanSlate synthesis flow.

3.2.2 The InterSynth Technique

Another specification-based synthesis technique is InterSynth. This approach is based on computing multi-output interpolants, as shown in the pseudo-code of Figure 2. The computation begins by dividing the input patterns into the on-set and the off-set for each output of the design. Next, the multi-output on-sets and off-sets are converted into AIGs and synthesized to reduce the total number of AIG nodes. After that, an incremental SAT problem is solved for each output, by assuming that the on-set and the off-set of this output are true at the same time. The proof of unsatisfiability of this instance is used to derive the interpolant for the output under consideration. The interpolants for all outputs are then combined into a single AIG, which is synthesized to reduce the total number of AIG nodes. Finally, the AIG is mapped into two-input gates as described in Section 3.2.1.

```
function InterSynth(TruthTable)
1   divide TruthTable into on-set and off-set for each output;
2   synthesize shared AIG F0 for off-sets of all outputs;
3   synthesize shared AIG F0 for on-sets of all outputs;
4   for each pair of outputs, f1 and f0, of AIGs F1 and F0
5     derive proof P of f1 ∧ f0 being unsatisfiable;
6     derive interpolant f from the proof P;
7   create shared AIG F from the set of interpolant AIGs {f};
8   synthesize AIG to minimize the number of nodes and levels;
9   return netlist by technology mapping the AIG;
```

Figure 2: The InterSynth synthesis flow

.

InterSynth differs from [10] in that it interpolates all primary outputs of the network rather than one node. For this, we extend the interpolation procedure to work for multi-output unsatisfiability proofs derived by solving several incremental SAT problems. The interface of a SAT solver such as MiniSAT [5] allows us to specify assumptions for each incremental SAT run. When the run is proved unsatisfiable, assumptions are lifted and the SAT solver can be reused. The assumptions used in the incremental runs express the condition that the on-set and the off-set are true simultaneously. This condition is, by construction, unsatisfiable for the on-set and the off-set. The resulting interpolant is a multi-output AIG such that the function of each output is contained in the interval defined by the on-set of this function and the complement of the off-set.

3.3 Specializing an Existing Netlist

Given an existing netlist, FastShrink uses a two-step process to produce a specialized new netlist. The first step, called *SignalMerge*, quickly merges signals in an existing circuit that are identical under the given input combinations. The second step, called *ShannonSynth*, performs further optimization using local don't-cares. The algorithm of SignalMerge is shown in Figure 3. It first simulates care-term vectors and then merges signals with identical signatures. This allows SignalMerge to leverage both external and internal satisfiability don't-cares to remove redundant gates. Our implementation selects the signal closest to primary inputs for merging to achieve smaller circuit delay. After the signals are merged, unconnected gates are removed. To expose additional merging opportunities, large cells such as AOI, OAI, etc. are decomposed into smaller gates. After signals in the netlist are merged, the netlist can be technology mapped again.

```
function SignalMerge(Circuit)
1   simulate vectors to generate signatures;
2   foreach signals with identical signatures
3     target ← the signal ∈ signals closest to primary inputs;
4     merge signals to target;
5     remove gates with no fanouts;
```

Figure 3: The SignalMerge algorithm.

Signal merging can remove redundant logic that generates identical signal functions. ShannonSynth pushes the optimization further by reimplementing subcircuits in smaller structures using don't-cares. To quickly identify subcircuits with high optimization potential, we use Shannon entropy to guide our resynthesis. Intuitively, signatures with low entropy contain less information and should be easier to optimize. In our experience we found that for a random subcircuit-extraction technique to produce the same quality as our entropy-guided approach, 50% more runtime is required.

The ShannonSynth algorithm in Figure 4 first simulates vectors in the care terms to generate a signature for each signal. Next, it computes the entropy of each signature. To make sure its resynthesis attempts are fruitful, the algorithm only tries subcircuits whose output signatures have small entropy (the bottom 20% of all signatures in our implementation). The key idea in this algorithm is that, instead of trying to resynthesize the netlist in the subcircuit,

```
function ShannonSynth(Circuit)
1   simulate vectors to generate signatures;
2   compute the entropy of each signature;
3   foreach signal whose signature has 20% smallest entropy
4     extract a subcircuit involving signal as its output;
5     build a truth table using the subcircuits' inputs and outputs;
6     resynthesize the truth table using CleanSlate;
7     if (resynthesized netlist is smaller)
8       replace the subcircuit with the resynthesized netlist;
```

Figure 4: The ShannonSynth algorithm.

we build a partial truth table using only the subcircuit's input and output signatures so that we can exploit don't-cares. ShannonSynth then synthesizes the truth table using the CleanSlate algorithm. In this step, however, we use Espresso to replace lines 1-3 of CleanSlate to achieve better resynthesis quality. This is appropriate in local resynthesis because the truth tables are small. After an optimized truth table is generated, ABC is still called for further optimization and technology mapping. If the new resynthesized netlist is smaller than the original one, ShannonSynth replaces it.

3.4 A Circuit Specialization Example

We now illustrate the FastShrink algorithm on a 3-bit ripple-carry adder. In this example, input A can only assume values 3, 4 or 5; while input B has values 1 or 7. SignalMerge first simulates all possible six input combinations on the given adder to produce 6-bit signatures for all internal signals. The circuit annotated with the signatures is shown in Figure 5(a). SignalMerge then merges signals with identical signatures and removes all the gates that are no longer connected (Figure 5(b)). At this point, only 8 out of the 15 gates are still needed, resulting in a much smaller circuit.

To further optimize the circuit, we invoke ShannonSynth. This extracts a subcircuit composed of gates g7, g8 and g9 to explore further optimizations. First, a truth table is built using the signatures of the subcircuit's inputs and outputs as follows:

A1	A0	B1	g5	g9
1	1	0	1	1
1	1	1	0	1
0	0	0	0	0
0	0	1	1	0
0	1	0	0	0
0	1	1	1	1

We then feed the truth table to CleanSlate for synthesis and obtain a new netlist, "g9=A0 & (g5 | B1)", that only uses two gates. Since this resynthesized netlist is smaller, it will replace the original one. Another ShannonSynth run replaces gate g0 with an inverter, and the final result is shown in Figure 5(c). By using the signatures of the subcircuit instead of the netlist for resynthesis, we can fully utilize don't-cares for optimization. This optimization is not performed by many traditional synthesis tools that only use function-preserving netlist transformations. Note that among the 58 don't-care input combinations, 25.9% are still added correctly.

4. ANALYSIS AND APPLICATIONS

In this section we first analyze our techniques and then outline several applications made possible by SWEDE.

4.1 Analysis

An important property of FastShrink is that every netlist modification it performs always preserves the output responses of the given input vectors. This is because we operate on signatures, which are simulated values of the input vectors. Since all the changes made by FastShrink preserve signatures, the output responses are also preserved. Moreover, we observe that FastShrink subsumes the common *constant propagation* technique, which is used when a subset of the signals are constant 0 or 1. To simplify our reasoning, we assume that the netlist is decomposed into 1- or 2-input gates, but the same holds in the general case as well.

PROPOSITION 1. *SignalMerge followed by ShannonSynth subsumes the optimizations produced by constant propagation.*

PROOF. Since the output of a 1-input gate can only be constant 0 or 1, SignalMerge connects the output signal to VCC or GND, thus eliminating the gate. Given a 2-input gate, suppose the constant

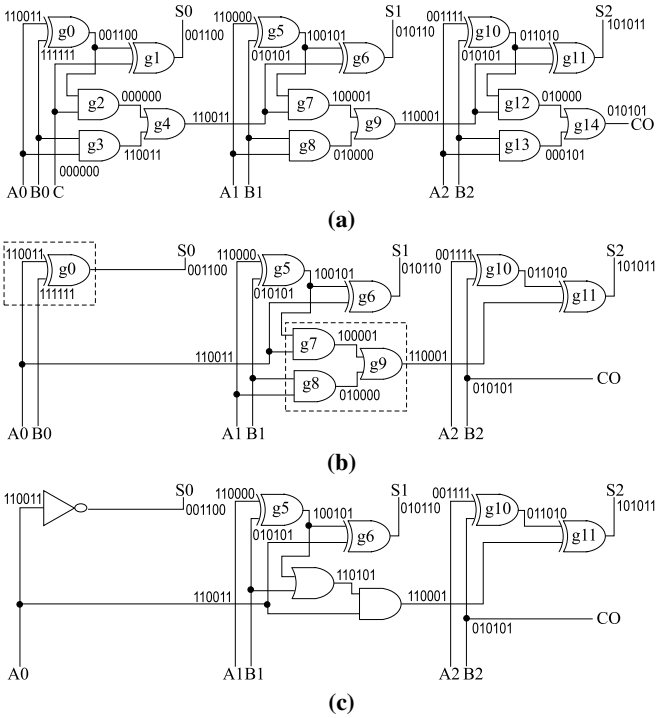

Figure 5: Ripple-carry adder specialization example: (a) original circuit, (b) after SignalMerge, and (c) after ShannonSynth. Allowed input values are 3, 4, 5 (for A) and 1 and 7 (for B).

input is the controlling value of the gate, then the output of the gate can only be constant 0 or 1. In this case, SignalMerge proceeds as the 1-input gate. Now suppose that the constant input is not the controlling value of the gate, then the output of the gate can be either identical or the complement to the other input. If the output is identical, then SignalMerge connects it directly to the non-constant input, eliminating the gate. Otherwise, we build a truth table using the gate's input and output signatures and rely on ShannonSynth to simplify the gate to an inverter. □

Finally, note also that a SignalMerge pass guarantees that no two signals are identical in the final circuit, since it merges all the signals with identical signatures.

Our analysis on how current commercial synthesis tools utilize don't-cares suggests that they perform inter-block optimizations by first dissolving the boundaries between the blocks to form a large flattened netlist, and then employing resynthesis techniques such as those introduced in Section 2.1. In other words, they convert external don't-cares into internal don't-cares before optimizations are performed. Although effective, this approach has the following drawbacks. First, the block boundaries are not preserved after optimization, which may make verification difficult, especially when dealing with third-party IP blocks in an SoC design. Second, dissolving boundaries makes it difficult to use external don't-cares because the chip's environment often depends on applications and cannot be modeled easily using a netlist. While state-of-the-art synthesis tools mostly exploit internal don't-cares, our work shows how to effectively exploit external don't-cares without viewing them as internal don't-cares and without blending multiple blocks into one netlist.

4.2 Applications

In this section we discuss some of the new applications that are enabled by our techniques, including three applications based on

circuit specialization followed by one that requires synthesizing truth tables.

Acceleration of common-case computations: certain classes of SoC designs include several instances of a computational module to improve the parallelism of the system. For instance, this is the case for multimedia SoC where the required output throughput is achieved by increasing the parallelism of the computation. Among CPU designs, a specific example is the case of the Sun Niagara T1 where 8 processor cores were sharing one Floating Point Unit (FPU). However, due to its poor performance on FP testbenches, the second generation processor has been enhanced with 8 FPUs. Often the input distribution of components embedded in a system is highly skewed for a very small set, while remaining combinations are rare [16]. For instance, it is observed that often under 10% of a program's instructions account for 90% of its execution time [8]. Hence, SWEDE can be adopted to explore a "Better Than Worst-Case Design" methodology [2], also known as "Common-Case Computation" [8], where one of several units is fully functional, and all others are optimized to only operate correctly for a few high-frequency input combinations. This approach reduces power and area of the final system. If an optimized computation fails at runtime, a fully-functional module is invoked as a back-up. Note that, for this approach to be viable, it may be necessary to deploy either a functional checker (validating the operation results) or a "valid input detection" circuit, as we are planning to explore.

Customization of third-party IP components in an SoC: in order to improve reuse, SoC designs often acquire some components from third-party vendors. In the fourth quarter of 2007, total IP revenue has reached $265.4 million, with a growth rate of 4.1% each year [25]. Such components are typically embedded in an environment that only exploits a small fraction of their functionality. It is then possible to use SWEDE to reduce the component's complexity (and power consumption) based on the specific environment in which it is embedded. For example, floating point logic in an embedded processor is redundant if the target application does not require any floating point computation. Manually removing redundant portions of the design, however, can be difficult and error-prone. While some hard IPs are difficult to modify, a large segment of the $1B/year IP market consists of soft IPs, such as ARM processors, USB and PCI-Express devices, etc. The source code is given to customers unencrypted because design companies would not agree to put unknown blocks in their chips. In addition, design houses often need to patch possible problems and better optimize their entire SoC designs in terms of placement and floorplanning. Importantly, such source code can be modified, and the techniques in our paper may lead to new business models — competing on cost by simplifying existing IPs automatically. For example, there are many USB and PCI-Express peripherals for PCs and laptops that are dedicated to a single function, like WiFi, WiMax, voice-over-IP, Dolby 7.1 sound, etc. Needless to say, such devices do not exercise the entire bus protocol, but the IP on which they are built may support it. Therefore, to reduce the cost, one may automatically customize the inherited bus IP to a given application. Whether or not the cost differential is significant, IP specialization may noticeably reduce power consumption. For example, Apple iPhone contains the S-Gold2 baseband chipset from Infineon in which Apple chose to turn off FM radio support and MMC/SD card compatibility, apparently to reduce power [26].

Graceful wear-out of electronic devices: extreme transistor scaling is leading to reduced silicon reliability, including early device and interconnect wear-out. To overcome the impact of this issue there is a growing need for low-cost reliable design solution. The use of SWEDE enables reliability through component sparing [3], where spare components can be optimized to provide only bare-bone functionality, sufficient to keep the system operational in critical aspects until it is replaced. An example of this spare-optimization application is discussed by Wagner *et al.* [19], where the authors identify a small subset of a processor design that must be kept operational in order to provide full system functionality (in this case the spare was part of the processor itself with acceleration features excluded). When the original circuit becomes unreliable, it will be replaced by the barebone spare component to avoid a system-level crash.

Synthesis for fast emulation: in the emulation domain, one common issue is the synthesis of the input constraints. Emulation systems can apply constrained-random simulation at very high performance compared to logic simulation. However, if the input constraints are not synthesizable, then at each clock cycle the emulator must communicate with a simulating host, incurring a huge performance impact on the emulation. At the same time, input constraints are often written in a high-level language (C++, Vera, etc.) and cannot be synthesized. SWEDE can be deployed by running the random simulation *only* on the design's input constraints (and not including the design itself). This simulation would be very fast and generates a set of care terms that SWEDE then synthesizes in a circuit uploaded on the emulator along with the design. Each emulation run would use a different constraint circuit, each synthesized by SWEDE based on the random stimuli. On the other hand, the design itself does not need to be resynthesized for each run.

5. EXPERIMENTAL RESULTS

In this section, we use two design examples to evaluate the capability of SWEDE in specializing circuits. One is an Alpha processor running real applications, and the other is an integer multiplier. In addition, we compare SWEDE with existing synthesis tools, including Espresso and a commercial synthesis tool, in terms of of capabilities in synthesizing truth tables with external don't-cares. Table 1 reports the numbers of primary inputs and outputs, as well as initial cell count for the benchmarks we used. Benchmarks C1908-C7552 are from ISCAS'85, and Alpha is a processor from [27] that implements a subset of the Alpha ISA. Our experiments were performed on Linux workstations with AMD Opteron 280 CPUs (2.4GHz) and 8G memory.

Benchmark	Description	#In/Outputs	#Cells
C1908	16-bit SEC/DED circuit	33/25	461
C2670	12-bit ALU and controller	233/140	484
C3540	8-bit ALU	50/22	1060
C5315	9-bit ALU	178/123	1057
C7552	32-bit adder/comparator	207/108	1187
Alpha	5-stage pipeline Alpha CPU	3054/3619	30531

Table 1: Characteristics of benchmarks.

5.1 Case Studies

In this section we study two design examples, an Alpha processor and an embedded multiplier.

Case study 1 (Alpha processor): for this study we ran five applications from the SpecINT'00 suite [30], whose characteristics are summarized in Table 2. The processor was synthesized using Cadence RTL compiler [28] with the highest optimization effort, and was mapped to a 0.18μm library. Since our Alpha processor only implements a subset of the Alpha ISA, simulation was performed in lockstep with the Simplescalar instruction set simulator [1]. We then use SignalMerge to optimize the circuit based on the stimuli from each program. Figure 6 and 7 report the final size of the optimized designs and the synthesis runtimes, respectively,

Benchmark	Description	Language
bzip2	Compression tool	C
gcc	Compiler	C
mcf	Combinatorial optimization	C
parser	Word processing	C
perlbmk	Perl programming language	C

Table 2: Characteristics of SpecINT programs [30].

achieved after simulating up to half a million instructions. They indicate that the optimization potential varies from application to application: for instance, the bzip2 application has a very small stimuli set, hence we can exploit aggressive optimizations on it; while gcc has a much wider span, hence little optimization can be extracted. This is aligned with the intuition that bzip2 is a specialized algorithm applying the same operations to arbitrary data sets, while gcc's operation is much more complex. This result suggests that if the program running on a circuit is known, SWEDE can potentially reduce its size significantly, generating a much smaller circuit that consumes less power. Figure 7 also shows that SignalMerge operates in approximately linear time on the number of input vectors in the care set, which enables it to handle complex designs efficiently. Designs can be further optimized by ShannonSynth: this step has greater runtime complexity, however, this is offset by the fact that ShannonSynth only takes into consideration small blocks in a circuit. For comparison, in the figures we also show the trend of optimizing for a constrained-random trace generated by StressTest [20] (diamond-bullet lines). Its curve indicates that with random inputs, we can only reduce the circuit by 10%, even when the number of instructions is as small as 6400. This is not surprising since, intuitively, random traces span a much larger fraction of the circuit's configurations than real applications, making optimization difficult.

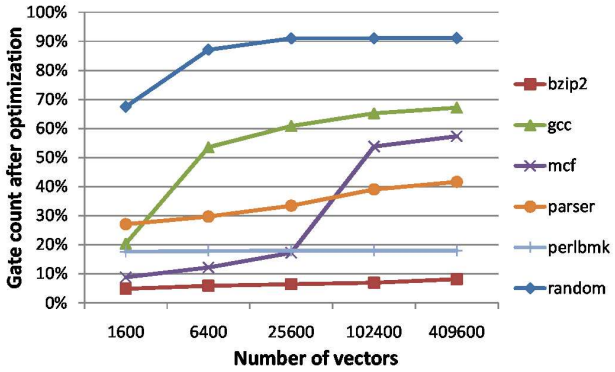

Figure 6: Gate count after specializing the Alpha CPU with SignalMerge. 30-90% of the gates can be removed for applications as long as half-million dynamic instructions.

In Figure 8 we show the results when optimizing individual components in the Alpha processor using the gcc application. The blocks we studied are the instruction fetch unit (IF), the decode unit (ID), the execute block (EX) and the memory access controller (MEM). The result indicates that the optimization potential is very block specific. In particular, the EX block cannot be optimized well because the execution unit needs to handle a wide range of input values, making don't-cares less dense. The MEM block also has very limited optimization potential because it only has 363 gates but has 195 inputs. This shallow logic structure makes signal sharing difficult.

Case study 2 (constant-coefficient multiplier): embedded systems and digital signal processors often need to perform simple operations repetitively [7]. For example, consider a portable electronic measurement device that must convert between US units and

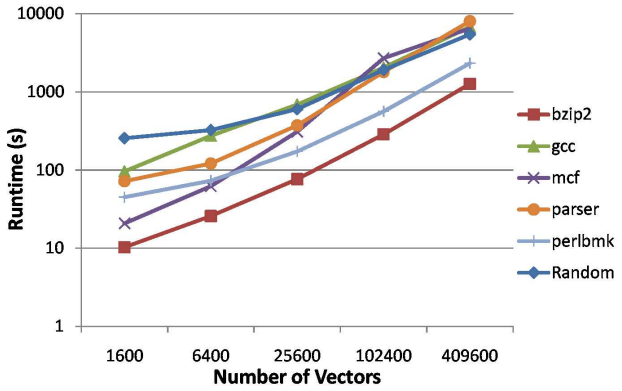

Figure 7: SignalMerge runtime to specialize Alpha. Runtime is approximately linear on the number of stimulus vectors used.

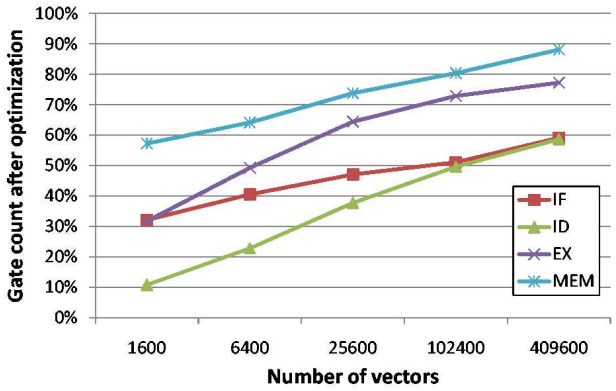

Figure 8: Gate count of Alpha blocks after specialization.

metric units while keeping power consumption low. To keep the circuit simple, an integer multiplier can be used, adjusting the decimal point afterward.

To support conversions between inches, feet, miles and meters, one needs to be able to multiply by the following six constants: 2.54, 30.4, 1.61 and their inverse. For the sake of this example, we made the assumption that the user can only compute with 5-digit decimal values. We used SWEDE to optimize the circuit starting from a 16-bit Wallace-tree multiplier. The original circuit had 1938 gates, and our care set included 393,216 patterns. For comparison, we converted external DCs into internal DCs by hard-coding the constants in the RTL code, and then we synthesized the design using two different commercial synthesis tools, Tool1 and Tool2. The results are summarized in Table 3. Since different synthesis tools may use different multiplier architectures, the reduction ratios should be compared instead of the cell counts. As the results suggest, FastShrink performs better than existing synthesis tools. For comparison with existing tools that support true external don't-care synthesis, we also attempted to synthesize the truth table of the 393216 patterns using Espresso and Tool1 (truth-table synthesis mode) but could not obtain a result netlist after 96 hours.

	Tool1		Tool2		**FastShrink**	
	Orig.	Opt.	Orig.	Opt.	Orig.	Opt.
Cell count	1387	834	2238	1440	1938	981
Reduction Ratio	39.9%		35.7%		**49.4%**	

Table 3: Comparison of two major commercial tools and SWEDE in synthesizing constant-coefficient multipliers. Original cell counts, optimized cell counts and the reduction ratios are shown.

Bench-mark	Number of cells after (re)synthesis					Runtime (s)				
	Truth table based				Netlist based	Truth table based				Netlist based
	Espresso	Tool1	CleanSlate	InterSynth	FastShrink (SignalMerge)	Espresso	Tool1	CleanSlate	InterSynth	FastShrink (SignalMerge)
C1908	2518	6891	**1352**	**828**	**284** (332)	16.19	143.76	**4.17**	**0.99**	**33.68** (0.32)
C2670	6098	T/O	**4467**	**2592**	**571** (665)	1494.51	T/O	**45.26**	**34.81**	**54.13** (1.36)
C3540	1925	6271	**1140**	**1980**	**1059** (1094)	29.12	193.69	**3.55**	**2.01**	**115.4** (1.54)
C5315	5183	T/O	**3594**	**5882**	**1238** (1312)	635.17	T/O	**27.70**	**25.04**	**179.56** (1.42)
C7552	5072	T/O	**3644**	**4923**	**1311** (1387)	911.54	T/O	**35.39**	**26.68**	**150.51** (0.71)

Table 4: Comparison of existing tools and SWEDE using one-hour time-out. All our solutions, CleanSlate, InterSynth and FastShrink, provide better synthesis quality with significantly shorter runtime.

While this multiplier only serves as a simple and intuitive example, the case study indicates that SWEDE can seamlessly handle even traditionally difficult synthesis problems, such as multipliers. This is because SWEDE is unconcerned with the complexity of the original functionality and can focus on just a few important inputs for its optimization. To further study the behavior of the specialized multiplier, we computed all the multiplications where one input ranges from 0 to 65535, and the other from 100 to 199, producing a total of 6553600 input combinations. The range for the second input was selected around the range of our specialized input constants. The results show that 29.33% of the input combinations were still multiplied correctly, while the average error over all input combinations was 9.75%. The greatest error we observed was 98.72%, produced by 56685×188.

5.2 Comparison with Existing Tools

In this experiment we compared CleanSlate and InterSynth with Espresso and a commercial tool (Tool1). We used the ABC system [24] to implement the interpolation-based procedure InterSynth for computing multi-level representations of Boolean functions that agree with the given on-set/off-set. The results are verified by checking that interpolants are implied by the on-sets and do not overlap with the off-sets. To avoid the influence of technology mapping on our experiments, we only used inverters and basic two-input gates. To evaluate Espresso, which lacks a technology mapper, we fed the optimized truth tables to ABC. We used 128 random patterns to generate the truth tables, and summarized the results in Table 4. CleanSlate and InterSynth outperform Espresso and Tool1, producing the smallest netlists in just a small fraction of the time. Moreover, in several cases Tool1 timed-out after one hour. We also tried synthesizing from care sets of 256, 512 and 1024 random patterns using the same circuits. We found that CleanSlate can finish all the benchmarks within 6.5 minutes, while Espresso and Tool1 timed-out after 1 hour for most of the benchmarks.

Although CleanSlate and InterSynth, which operate from a truth table specification, produce comparatively better results, a comparison with Table 1 shows that the generated netlists are still larger than the original ones. The reason is that the original netlists are often produced from higher-level specifications, which include conceptual structures that lead to better optimizations. On the other hand, trying to synthesize a compact netlist using only input and output values is much more difficult. Therefore, if a netlist is available, the best optimizations can be obtained through FastShrink, whose results are also shown in Table 4.

SWEDE is based on signatures, which can be calculated easily using simulation. This makes SWEDE simple to use because designers only need to provide input vectors to the circuit that belong to the care terms. Since signatures can be represented compactly using bit-vectors and allow bit-parallel computation, our solution is both fast and memory-efficient. As our experimental results show, we can handle half-million input vectors in less than three hours.

6. CONCLUSIONS

To reduce circuit design complexity in the multi-billion transistor era, SoC and embedded systems heavily rely on reuse and third-party IP components. Often, the design environment surrounding such components uses only a fraction of the functionality that these general-purpose components implement. The unused logic in those circuit blocks not only occupies valuable die area but also consumes more power, hurting the circuit's performance and quality. Hence, new synthesis optimization opportunities are available in simplifying these components to the subset of functionality required by the system they are embedded in. Surprisingly, existing synthesis tools perform poorly in this context, which typically involves a small care-set and a very large don't-care set. To address this problem, we proposed a new tool called SWEDE, and provided three new synthesis techniques which can specialize a circuit using external don't-cares: FastShrink, CleanSlate and InterSynth. Unlike traditional synthesis tools that pursue maximal use of don't-cares by explicitly branching on different don't-care assignments, our greedy algorithms (SignalMerge and the first phase of CleanSlate) implicitly exploit the fact that most terms are don't-cares and quickly generate a small netlist. Further circuit optimization is performed by our ShannonSynth technique and the second phase of CleanSlate. This novel synthesis flow allows SWEDE to scale better when massive don't-cares exist. As our empirical results indicate, SWEDE provides better synthesis quality than state-of-the-art tools while running 10X faster. In fact, SWEDE can handle designs as large as 30K cells with 0.5M care-set vectors in a few hours, demonstrating its superior scalability and efficiency.

We discussed a number of new applications enabled by SWEDE, including new system-design paradigms and solutions to current engineering problems. These new applications promise to produce circuits that run faster, consume less power, and can be used as inexpensive back-up modules for larger circuits that may fail during operation. Our future work seeks to develop compact circuit indicators for output correctness in specialized circuits.

7. REFERENCES

[1] T. Austin, E. Larson and D. Ernst, "Simplescalar: An Infrastructure for Computer System Modeling", IEEE Computer, Feb. 2002, pp. 59-67.

[2] T. Austin, V. Bertacco, D. Blaauw and T. Mudge, "Opportunities and Challenges for Better Than Worst-Case Design", ASPDAC'05, pp. 2-7.

[3] K. Constantinides, S. Plaza, J. Blome, B. Zhang, V. Bertacco, S. Mahlke, T. Austin and M. Orshansky , "BulletProof: A Defect-Tolerant CMP Switch Architecture", HPCA'06, pp. 5-16.

[4] W. Craig, "Linear Reasoning: A New Form of the Hebrand-Gentzen Theorem", Jour. of Sym. Logic, vol. 22 (3), 1957, pp. 250-287.

[5] N. Een and N. Sorensson, "An Extensible SAT-solver", SAT'03. http://minisat.se/

[6] M. Gao, J.-H. Jiang, Y. Jiang, Y. Li, A. Mishchenko, S. Sinha, T. Villa and R. Brayton, "Optimization of Multi-Value Multi-Level Networks", ISMVL'02, pp. 168-177.

[7] C.-Y. Lai, C.-Y. Huang and K.-Y. Khoo, “Improving Constant-Coefficient Multiplier Verification by Partial Product Identification”, DATE’08, pp. 813-818.

[8] G. Lakshminarayana, A. Raghunathan, K. S. Khouri and N. K. Jha, “Method for Synthesis of Common-Case Optimized Circuits to Improve Performance and Power Dissipation”, United States Patent, No. 6,308,313 B1, Oct. 2001.

[9] K. L. McMillan, “Interpolation and SAT-based Model Checking”, CAV’03, pp. 1-13, LNCS 2725, Springer, 2003.

[10] A. Mishchenko, R. Brayton, J.-H. R. Jiang, and S. Jang, “SAT-based Logic Optimization and Resynthesis”, IWLS ’07, pp. 358-364.

[11] S. Muroga, Y. Kambayashi, H. C. Lai, and J. N. Culliney, “The Transduction Method — Design of Logic Networks Based on Permissible Functions,” IEEE Computer, Oct. 1989, pp. 1404-1424

[12] S. M. Plaza, K.-H. Chang, I. L. Markov, and V. Bertacco, “Node Mergers in the Presence of Don’t Cares”, ASPDAC’07, pp. 414-419.

[13] J. Rajski, J. Vasudevamurthy, “The testability-preserving concurrent decomposition and factorization of Boolean expressions”, IEEE TCAD, June 1992, pp.778-793.

[14] R. Rudell and A. Sangiovanni-Vincentelli, “Multiple-Valued Minimization for PLA Optimization”, IEEE TCAD, Sep. 1987, pp. 727-750.

[15] H. Savoj and R. K. Brayton, “The Use of Observability and External Don’t Cares for the Simplification of Multi-Level Networks”, DAC’90, pp. 297-301.

[16] E. Schnarr and J. R. Larus, “Fast Out-of-Order Processor Simulation Using Memoization”, ASPLOS’98, pp. 283-294

[17] C. E. Shannon, “A Mathematical Theory of Communication”, The Bell System Technical Journal, Vol. 27, Oct. 1948, pp. 379-423.

[18] S. Sinha and R. K. Brayton, “Implementation and Use of SPFDs in Optimizing Boolean Networks”, ICCAD’98, pp. 103-110.

[19] I. Wagner, V. Bertacco and T. Austin, “Shielding Against Design Flaws with Field Repairable Control Logic”, DAC’06, pp. 344-347.

[20] I. Wagner, V. Bertacco and T. Austin, “StressTest: An Automatic Approach to Test Generation via Activity Monitors”, DAC’05, pp. 783-788.

[21] S. Yamashita, H. Sawada, and A. Nagoya, “A New Method to Express Functional Permissibilities for LUT Based FPGAs and Its Applications”, ICCAD, 1996, pp. 254-261.

[22] S. Yamashita, H. Sawada and A. Nagoya, “SPFD: A New Method to Express Functional Flexibility”, IEEE TCAD, Aug. 2000, pp. 840-849.

[23] Y.-S. Yang, S. Sinha, A. Veneris and R. E. Brayton, “Automating Logic Rectification by Approximate SPFDs”, ASPDAC’07, pp. 402-407.

[24] Berkeley Logic Synthesis and Verification Group, ABC: A System for Sequential Synthesis and Verification, Release 80308. `http://www-cad.eecs.berkeley.edu/~alanmi/abc/`

[25] “EDA Sales Jump in Q4”, EE Times, Apr. 03, 2008.

[26] “Europe Suppliers Score in Apple’s iPhone”, EE Times Europe, Jul. 02, 2007.

[27] Bug UnderGround, `http://bug.eecs.umich.edu/`

[28] `http://www.cadence.com/`

[29] MVSIS, `http://www-cad.eecs.berkeley.edu/Respep/Research/mvsis`

[30] SpecINT2000 benchmarks, `http://www.spec.org/`