Verifying High-Level Latency-Insensitive Designs with Formal Model Checking

Steve Dai, Alicia Klinefelter, Haoxing Ren, Rangharajan Venkatesan, Ben Keller, Nathaniel Pincney, Bruce Khailany
NVIDIA

Abstract—Latency-insensitive design mitigates increasing interconnect delay and enables productive component reuse in complex digital systems. This design style has been adopted in high-level design flows because untimed functional blocks connected through latency-insensitive interfaces provide a natural communication abstraction. However, latency-insensitive design with high-level languages also introduces a unique set of verification challenges that jeopardize functional correctness. In particular, bugs due to invalid consumption of inputs and deadlocks can be difficult to detect and debug with dynamic simulation methods. To tackle these two classes of bugs, we propose formal model checking methods to guarantee that a high-level latency-insensitive design is unaffected by invalid input data and is free of deadlock. We develop a well-structured verification wrapper for each property to automatically construct the corresponding formal model for checking. Our experiments demonstrate that the formal checks are effective in realistic bug scenarios from high-level designs.

I. INTRODUCTION

As modern SoC design challenges continue to motivate reuse of existing design blocks, latency-insensitive (LI) design has emerged as a practical methodology for synchronizing pre-assembled modules under increasing pressure of lengthening interconnect delay [1], [2]. By exposing a valid-ready interface from each module, LI design decouples the timing of intra-module computation from that of inter-module communication to ensure robust functionality while tolerating arbitrary communication latency between modules. This methodology enables flexible physical design implementation without impacting verification of individual components.

In parallel with this trend, hardware designers have embraced high-level languages for high-productivity VLSI design. In particular, high-level synthesis (HLS) compilers can automatically synthesize RTL from C++ models. Because an HLS compiler translates untimed functional blocks in software into interconnected cycle-accurate hardware modules with customized throughput and latency, it is natural for HLS to adopt an LI-based composition of the modules to take advantage of the modularity and relaxed timing requirement of LI design. The confluence of HLS and LI design has enabled rapid design of large-scale chips using high-level languages [3], [4].

Figure 1 shows a typical HLS design flow (on the left) for generating an RTL model from a C++ description. Designers generally rely on dynamic simulation of C++ and RTL models for verifying their designs. The HLS design flow allows designers to leverage C++ simulation for the bulk of their design verification tasks and promises orders of magnitude speedup over a conventional RTL flow [5]. Nevertheless, due to inherent differences in timing models between C++ and RTL, C++ simulation must be complemented by RTL simulation to expose bugs that require cycle-accurate introspection.

Figure 1. Typical HLS design and verification flow with our proposed formal extensions — Design is typically verified with dynamic simulations in C++ and RTL. We extend verification with automated formal model checking.

The quality of dynamic simulation depends heavily on the effectiveness of the set of chosen stimuli. These stimuli realistically represent only a small subspace of all possible inputs and risk excluding difficult-to-anticipate corner cases. In addition, the LI property allows for arbitrary timing of inter-module interface, further requiring expansion of the verification coverage space to include a full range of input arrival times and relative input ordering. With dynamic simulation, it is especially difficult to detect and eliminate design issues resulting in invalid consumption of inputs and deadlock, which constitute two commonly occurring bugs in LI-based HLS design. To address these two classes of bugs, we propose augmenting a typical HLS design flow with formal verification extensions (right side of Figure 1 in bold). The major contributions of this work are:

- We propose RTL formal model checking methods that guarantee an LI-based HLS design is not affected by invalid input data and is free of deadlock for all intended use cases.
- We develop an automated flow to generate the verification wrapper for each property and implement the corresponding formal checks without human intervention.
- We demonstrate the effectiveness of the proposed verification models and flow in realistic industry and academic bug scenarios, as well as on a range of design blocks.

The remainder of the paper is organized as follows: Section II reviews related work; Section III establishes the basics for LI design in HLS; Section IV describes the two classes of common bugs we target and proposes two formal models to detect these bugs; Section V demonstrates the effectiveness of our models, followed by conclusions in Section VI.

II. RELATED WORK

LI design refers to the correct-by-construction composition of storable computational processes that exchange data through communication channels in accordance with an appropriate LI protocol such as valid/ready or request/acknowledge [1], [6]. As long as each computational process (i.e., functional module) is implemented correctly, the composition will also behave correctly regardless of the latency of the channels. The introduction of the methodology has led to
the emergence of a family of LI protocols, followed by a set of dynamic simulation and formal verification techniques to validate the correctness of these protocols [7], [8]. Our work is concerned with verifying designs implemented using LI methodology, rather than validating the correctness of the methodology and protocols as in these previous works.

Designers typically verify LI-based HLS designs with dynamic simulation in C++ and RTL as they iterate on various design changes. Conventionally, formal verification methods have been applied to prove the correctness of C-to-RTL transformations by checking the equivalence of the model before and after transformation, rather than validating the correctness of the C++ implementation itself [9], [10]. KAiros leverages formal equivalence checking to verify whether a modified HLS design is equivalent to the original (golden) design after incremental code modification or change in HLS optimizations [11]. Unlike previous formal techniques for HLS, our methods do not target bugs caused by the transformations of the HLS tool. Instead, we target designer-induced bugs while not requiring a golden model as ground truth.

III. LI IMPLEMENTATION IN HLS

In this work, LI-based HLS designs are realized using MatchLib, a high-level library of synthesizable port and channel primitives in C++ implementing LI valid-ready protocols for HLS tools [12]. For an LI design in HLS, each functional block exposes an interface of directional ports from the MatchLib library, shown as InA, InB, and Out in Figure 2(a). These ports are then connected to MatchLib ports on other functional blocks via MatchLib channels. As shown in Figure 2(a), each port can read (pop) data from a channel (e.g., InA and InB) or write (push) data to a channel (e.g., Out).

Transactions acting on a MatchLib port can be either blocking or non-blocking. Blocking communication prevents subsequent transactions from executing until the current read or write has succeeded, and can block forward progress. In contrast, non-blocking communication allows subsequent transactions to execute regardless of whether the current read or write is successful, and cannot block forward progress. Figure 2(b) implements the same adder design as in Figure 2(a), except with non-blocking input ports InA and InB. statusA and statusB in (b) are used to indicate whether DataA and DataB respectively contain valid input data from the channels. Because of the use of a blocking pop in (a), Line 4 cannot be executed until the Pop in Line 3 successfully reads valid data from port InA into the local variable DataA. However, with non-blocking pop in (b), Line 6 can be executed after Line 5 regardless of whether the PopNB in Line 5 is successful.

```c
1 void adder_blocking() { 1
2 while (1) { 2
3     DataA = InA.Pop(); 3
4     DataB = InB.Pop(); 4
5     Out.push(DataA+DataB); 5
6 } 6
7 } 7
```

```c
1 void adder_nonblocking() { 1
2 while (1) { 2
3     statusA = InA.PopNB(DataA); 3
4     if (statusA) 4
5         statusA = InA.PopNB(DataA); 5
6         statusB = InB.PopNB(DataB); 6
7         if (statusA && statusB) 7
8             Out.push(DataA+DataB); 8
9         statusA = false; statusB = false; 9
10     statusA = false; statusB = false; 10
11 } 11
```

Fig. 2. Simple adder design in C++ — (a) Input ports are blocking. (b) Input ports are non-blocking.

The corresponding valid-ready interfaces in RTL for input port InA and output port Out generated from the C++ descriptions in Figure 2 are shown in Figure 3 respectively. Blocking and non-blocking ports share the same RTL interface but differ in their internal blocking logic.

It is important for an LI implementation to include non-blocking communication so that realistic and scalable designs can be expressed. Figure 4 illustrates a three-input one-output process $p_0$. If $p_0$ implements a multiplier, all three inputs must be available before the output is valid; however, if $p_0$ is an arbiter, only one input needs to be available before producing a valid output. The multiplier could be implemented with blocking communication, while the arbiter must include non-blocking reads to be functional.

While the decision to use blocking versus non-blocking communication is dependent on internal process details, each comes with its own pitfalls that result in the two classes of bugs targeted by this paper. Blocking communication creates an inherent wait-for relationship between design blocks and can lead to deadlock. Non-blocking communication requires custom bookkeeping logic (e.g., Lines 4, 6, and 8 in Figure 2(b)) to prevent undesirable consumption of invalid input data that can increase the chance of designer error. Mixing blocking and non-blocking communication further complicates the design. Our work provides safeguards from these kinds of unintended and undesirable consequences through automation, without extra burden on the HLS designer or verification engineer.

IV. FORMAL MODEL CHECKING FOR LI DESIGN

The first class of bugs we target involves consumption of invalid inputs at the LI interface. When non-blocking communication is needed, a user typically writes custom bookkeeping logic to manage the communication. This is a common cause of errors, since improperly constrained non-blocking reads and writes can result in tainted updates to stateful elements by invalid input data. The risk of error increases with additional non-blocking ports as designers attempt to manage the complex interaction among multiple instances of custom bookkeeping logic while keeping track of how the sequential C++ design entry will translate into parallel hardware in generated RTL. This class of bugs is difficult to detect with dynamic simulation because the designer-imposed constraints are often buggy only under limited corner cases that are non-trivial to conceive ahead of time during test planning.

The second class of bugs we target involves deadlock, which may arise due to a multitude of factors, including incorrect capacities for communication channels, improper application or combination of blocking and non-blocking ports, latency-sensitive bookkeeping logic, or circular dependencies among
formal model that checks for invalid input consumption. On the output side, the model verifies that any but may receive different invalid inputs, as modeled by the instances are set up to always receive the same valid inputs, inputs and outputs as shown. On the input side, both DUT of the same DUT, with corresponding reference and test This model consists of a reference as well as a test instance conditions that constrain the inputs and check the outputs. Formally, finding buggy combinations of signal arrivals can require many iterations to expose bugs requiring complex input patterns.

Formal model checking [13] is commonly applied to more thoroughly verify various hardware components and protocols [14], [15]. Therefore, it is uniquely positioned to address the verification gap by proving the absence of the two classes of bugs in our designs without limitation to a specific subset of stimulus and input arrival timing. In particular, we apply RTL-based formal model checking on the HLS-generated RTL design blocks to verify properties associated with these two classes of bugs. Figures 5 and 6 present an overview of the two corresponding formal models.

A. Invalid Input Consumption Check

Our invalid input consumption check proves whether valid output data of the design are unaffected by invalid input data. In other words, invalid input data must not assert influence on any valid output data. Recall the non-blocking adder in Figure 2(b) in which input data are read from non-blocking ports InA and InB. If Line 9 in Figure 2(b) is not guarded by the conditional statement in Line 8, the adder will perform addition regardless of whether DataA and DataB contain valid input data. In this case, valid output at Out is affected by invalid inputs, and fails the property of being unaffected by invalid input. While this hypothetical bug represents a relatively contrived case of incorrect bookkeeping logic, there are many examples of improperly constrained non-blocking operations that can be discovered by this check.

Figure 5 shows how we wrap a design under test (DUT) into a formal model that checks for invalid input consumption. This model consists of a reference as well as a test instance of the same DUT, with corresponding reference and test inputs and outputs as shown. On the input side, both DUT instances are set up to always receive the same valid inputs, but may receive different invalid inputs, as modeled by the multiplexers. On the output side, the model verifies that any valid outputs are equivalent by comparing the corresponding output valid and output data signals. $Val_{ref} = Val_{test}$ and $Val_{ref} \& Val_{test} \implies D_{ref} = D_{test}$ define the conditions that constrain the inputs and check the outputs. Note that the model is set up such that both the reference and test DUTs receive the same external ready signal.

B. Deadlock Detection

Our deadlock detection proves whether a design is absent of deadlock. Figure 7 shows a simple system with two interacting functional blocks that contains a potential deadlock. If cond in the Consumer function initializes to false, consumer never accepts the data from producer, but instead immediately tries to push an acknowledgment to producer. However, producer would stall during push because its data are not accepted and cannot move on to popping the acknowledgment from consumer. As a result, both producer and consumer end up in a stalled state. When all interacting modules are stalled, no module can trigger forward progress, and the system is in deadlock. This represents one of many scenarios for which an LI design could deadlock.

To detect a deadlock under any of these scenarios, we can make use of global stall signals that a typical HLS tool generates for individual RTL modules. For instance, Mentor Catapult HLS assigns a wait controller for each stallable interface (e.g., blocking port) of a module. These wait controllers then communicate with the staller of the module to form the global stall signal for clock gating the module [16]. Figure 6 provides an abstract illustration of the producer and consumer modules in RTL generated from the C++ descriptions in Figure 7. As shown in Figure 6, our verification wrapper constructs the formal model by aggregating the global stall signals from individual modules. During the formal check, our model ensures that any input or output to the system is not blocked. The system contains a deadlock if the formal check determines that it is possible for all aggregated global stall signals to be asserted at the same time. If not, the system is free of deadlock. The deadlock-free property for an N-module design is formally expressed as $\neg \left( \bigwedge_{i=1}^{N} GlobalStall_{i} \right)$.

We extend our deadlock detection method to support non-blocking ports, which do not include explicit stall signals because they cannot be blocked by definition. In this case, we devise a custom global stall signal for each RTL module from the ready signals of its non-blocking ports. This custom global stall signal is asserted if none of the relevant ready signals have been asserted for N clock cycles, where N is a known constant at verification time based on the HLS-applied optimizations. This type of custom global stall signal can be

```c++
1 void Producer() {
2 while (1) {
3 Data.Push(1);
4 if (cond) Data.Pop();
5 }
6 }
7 void Consumer() {
8 while (1) {
9 Ack.Pop();
10 Ack.Push(1);
11 }
12 }
```

Fig. 7. Simple system in C++ with two interacting functional blocks — Both blocks will stall if Consumer does not accept data from Producer.
used in lieu of the global stall signal in Figure 5 to implement the same deadlock check.

V. EXPERIMENTS

Our formal models are implemented using SystemVerilog assertions and verified with bounded model checking [17] using Synopsys VC Formal 2018.10 running on an Intel Xeon CPU at 3GHz. Formal model checking is performed on the RTL synthesized from C++ using HLS tool. Because HLS is an automated process that compiles C++ into predictably-structured RTL well-suited for the extraction of relevant signals outlined in Sections IV-A and IV-B, our formal flow can be fully automated. Although the bugs we target originate in C++ during design entry, we formally verify the designs in RTL because the full scopes of the bugs only manifest under the cycle-accurate timing model of RTL.

We first validate our formal models using known bug cases, listed in Table I abstracted from real industry and academic HLS designs. We abstract the design names to indicate the primary cause of the bugs. Each of these bug cases consists of the initial (buggy) version of the corresponding design followed by one or more patched (but possibly still buggy) versions of the same design. The initial versions of all the designs are written without knowledge of our formal models. Likewise, the models are developed without specific knowledge of the initial designs. As such, our abstracted bug cases provide a minimally viable but faithful reproduction of the specific bugs to help narrow down the root causes of the bugs and to understand and validate the results of our formal models.

Table I details the findings of the formal engine against our proposed models for different versions of each design, along with post-logic-synthesis gate count and runtime of the formal engine. The evolution of each design from the initial version to the final fix demonstrates the effectiveness and correctness of our formal models. As shown in the table, our formal models have extracted bugs even from purportedly fixed and verified designs, which speaks to the shortcoming of the existing verification methodology. The counterexamples provided by these proofs are instrumental in quickly identifying the root cause and devising the appropriate fix.

To further demonstrate the applicability of our approach, we apply our formal models on a set of open-source HLS library components from MatchLib [12]. These library components are meant to be reused across a large number of designs, and therefore constitute good candidates for extensive verification.

We also experiment with full applications of our own and from open-source HLS benchmark suite Rosetta [18] to demonstrate the general applicability of our models. As shown in Table II we apply the invalid input check on the five design components and successfully prove that they are unaffected by invalid inputs. On the other hand, we apply deadlock detection on a 2x2 array of network-on-chip (NoC) router components, an optical flow accelerator, and a machine-learning accelerator for spam filtering, where we identify certain deadlock states.

Results in Table II shows that we are able to prove exhaustively that the five components are unaffected by invalid input, and prove at the user-defined bound that specific versions of the applications are free of deadlock. Specifically, our deadlock check supports the NoCRouterArray application which makes use of non-blocking ports exclusively. We discover an incorrect protocol constraint in NoCRouterArray.v1 that results in a deadlock and apply the appropriate fix in NoCRouterArray.v2 after examining the counterexample trace. We also detect a deadlock in OpticalFlow.v2 that escapes the test cases in the supplied test bench. Compared to OpticalFlow.v1, OpticalFlow.v2 is buggy because it contains one FIFO with a reduced capacity, resulting in a potential cyclic wait scenario from insufficient FIFO capacity due to the re-convergence pattern in the benchmark’s dataflow [19]. SpamFilter.v2 does not incur the same problem even with reduced FIFO sizes compared to SpamFilter.v1 because the benchmark’s dataflow pattern contains no branches and thus no re-convergence.

VI. CONCLUSIONS

While the LI design methodology simplifies the composition of synthesized functional blocks in HLS, this latency-tolerant design style also introduces additional difficulty in ensuring the correctness of high-level designs. In this paper, we provide an automated flow based on formal model checking to guarantee that a high-level LI design is not affected by invalid input data and is free of deadlock. The proposed verification techniques are effective and generally applicable to a range of LI-based HLS designs, and lead to promising improvement in the quality of verification. We believe that closing the verification gap is key to mainstream adoption of HLS tools, and our formal verification extensions play a crucial role as part of static sign-off toward this direction [20]. We expect wider adoption of our approach as formal tools and their underlying engines become increasingly scalable [21], [22].
REFERENCES

[1] L. P. Carloni, K. L. McMillan, A. Saldanha, and A. L. Sangiovanni-Vincentelli, “A Methodology for Correct-by-construction Latency Insensitive Design,” Int’1 Conf. on Computer-Aided Design (ICCAD), 2003.

[2] L. P. Carloni, “From Latency-Insensitive Design to Communication-based System-Level Design,” Proceedings of the IEEE, 2015.

[3] R. Venkatesan, Y. S. Shao, B. Zimmer, J. Clemons, M. Fojitik, N. Jiang, B. Keller, A. Klinefelter, N. Pinckney, P. Raina et al., “A 0.11 PJ/OP, 0.32-128 TOPS, Scalable Multi-Chip-Module-Based Deep Neural Network Accelerator Designed with A High-Productivity VLSI Methodology,” Symp. on High Performance Chips (Hot Chips), 2019.

[4] T. Ajayi, K. Al-Hawaj, A. Amarnath, S. Dai, S. Davidson, P. Gao, G. Liu, A. Lotfi, J. Puscar, A. Rao et al., “Celerity: An Open Source RISC-V Tiered Accelerator Fabric,” Symp. on High Performance Chips (Hot Chips), 2017.

[5] J. Cong, B. Liu, S. Neundorffer, J. Nogueria, K. Vissers, and Z. Zhang, “High-Level Synthesis for FPGAs: From Prototyping to Deployment,” IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2011.

[6] L. P. Carloni, K. L. McMillan, and A. L. Sangiovanni-Vincentelli, “Theory of Latency-Insensitive Design,” IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2001.

[7] S. Suhail, D. Mathaikutty, D. Berner, and S. Shukla, “Validating Families of Latency Insensitive Protocols,” IEEE Transactions on Computers (TC), 2006.

[8] C.-H. Li, R. Collins, S. Sonalkar, and L. P. Carloni, “Design, Implementation, and Validation of a New Class of Interface Circuits for Latency-Insensitive Design,” Int’1 Conf. on Formal Methods and Models for Codesign (MEMOCODE), 2007.

[9] C. Karfa, D. Sarkar, C. Mandal, and P. Kumar, “An Equivalence-Checking Method for Scheduling Verification in High-Level Synthesis,” IEEE Trans. on Computer-Aided Design of Integrated Circuits and Systems (TCAD), 2008.

[10] Y. Kim, S. Kopuri, and N. Mansouri, “Automated Formal Verification of Scheduling Process using Finite State Machines with Datapath (FSMD),” Int’1 Symp. on Signals, Circuits and Systems (SCS), 2004.

[11] L. Piccolboni, G. Di Guglielmo, and L. P. Carloni, “KAIROS: Incremental Verification in High-Level Synthesis through Latency-Insensitive Design,” Formal Methods in Computer Aided Design (FMCAD), 2019.

[12] B. Khailany, E. Krimer, R. Venkatesan, J. Clemons, J. S. Emer, M. Fojitik, A. Klinefelter, M. Pellauer, N. Pinckney, Y. S. Shao et al., “A Modular Digital VLSI Flow for High-Productivity SoC Design,” Design Automation Conf. (DAC), 2018.

[13] E. M. Clarke Jr, O. Grumberg, D. Kroening, D. Peled, and H. Veith, Model Checking. MIT Press, 2018.

[14] B. Bingham, M. Greenstreet, and J. Bingham, “Parameterized Verification of Deadlock Freedom in Symmetric Cache Coherence Protocols,” Formal Methods in Computer-Aided Design (FMCAD), 2011.

[15] D. Kaufmann, A. Biere, and M. Kauers, “Verifying Large Multipliers by Combining SAT and Computer Algebra,” Formal Methods in Computer Aided Design (FMCAD), 2019.

[16] Catapult Synthesis User and Reference Manual, Software Version v10.5a, Mentor Graphics, 2020.

[17] A. Biere, A. Cimatti, E. M. Clarke, O. Strichman, Y. Zhu et al., “Bounded Model Checking,” Advances in Computers, 2003.

[18] Y. Zhou, U. Gupta, S. Dai, R. Zhao, N. Srivastava, H. Jin, J. Featherston, Y.-H. Lai, G. Liu, G. A. Velasquez et al., “Rosetta: A Realistic High-Level Synthesis Benchmark Suite for Software Programmable FPGAs,” Int’1 Symp. on Field-Programmable Gate Arrays (FPGA), 2018.

[19] M. Fingeroff, High-Level Synthesis Blue Book. Xlibris Corporation, 2010.

[20] P. Ashar and V. Viswanath, “Closing the Verification Gap with Static Sign-off,” Int’1 Symp. on Quality Electronic Design (ISQED), 2019.

[21] M. Mann and C. Barrett, “Partial Order Reduction for Deep Bug Finding in Synchronous Hardware,” Int’1 Conf. on Tools and Algorithms for the Construction and Analysis of Systems (TACAS), 2020.

[22] C. Barrett, C. L. Conway, M. Deters, L. Hadarean, D. Jovanović, T. King, A. Reynolds, and C. Tinelli, “CVC4,” Int’1 Conf. on Computer Aided Verification (CAV), 2011.