Catapult HLS User View Scheduling Rules

Stuart Swan, Platform Architect, Siemens EDA

17 June 2025

Introduction

This document describes the user view scheduling rules for C++ and SystemC models that are synthesized through Catapult HLS. The goals of this document are:

1. To provide easy-to-understand rules to HLS users.
2. To ensure precise and consistent rules in both SystemC and C++.
3. To offer rules that are effectively compatible with how Catapult currently operates.
4. To enable the best possible *quality of results* (QOR) in Catapult synthesis.
5. To cover all known user requirements and scenarios.
6. To serve as a suitable starting point for a standardization proposal (e.g., in Accellera SWG).

To illustrate the goals of this document more specifically, consider what an engineer writing a testbench for an HLS model needs to understand about how HLS tools operate. This engineer may be using SystemVerilog UVM or may be writing a testbench in C++/SystemC. They likely are not an expert in any specific HLS tool (and may not want to be), but their testbench needs to work for both the pre-HLS model as well as the post-HLS model. Thus, it is crucial for the DV engineer to have a precise understanding of how the HLS tool will transform the design while still enabling it to be fully verified. This document describes what transformations the HLS tool is allowed to perform so that the pre-HLS and post-HLS models can be effectively verified with the same testbench. The overarching philosophy of the scheduling rules is to present "no surprises" to such a DV engineer, while still giving the HLS tool ample freedom to optimize the design.

Background

HLS tools generate RTL from C++ models. Broadly speaking, this conversion takes a sequential C++ model and turns it into concurrent hardware that maintains the same behavior. HLS tools identify concurrent processes within the C++/SystemC model and then independently synthesize each process. Briefly, some of the techniques that HLS tools use to achieve good HW QOR when synthesizing each process include:

- Optimized scheduling based on the selected silicon target technology.
- Automatic HW pipeline construction according to the user's specifications.
- Automatic HW resource sharing.
- Automatic scheduling of memory accesses.

The internal behavior of each process is specified by the control and dataflow behavior of the C++ code within the process. However, the external communication that each process has with other processes and HW blocks is specified via IO operations that are coded within the model. To enable a reliable, scalable, and verifiable HLS flow that generates high quality hardware, the scheduling behavior of these

IO operations needs to be precisely handled at all steps of the flow. This document specifies the rules that govern the scheduling behavior for these IO operations within HLS models. These rules are specified with respect to an individual process, but the intent of the rules is to enable reliable and verifiable behavior of large sets of interacting processes in real-world designs.

By default, HLS tools can insert additional clock cycles (or latency) anywhere within a process -- for example, when pipelining a loop or to enable HW resource sharing. The overall approach used in this document is to make the entire design and testbench *latency insensitive* to the maximum extent possible, while still fully enabling key HW optimizations.

In some cases, designs or testbenches may use protocols which are latency sensitive. These situations can be handled by isolating the latency sensitive portions to small, self-contained parts of the design or testbench, and then keeping the rest of the design and testbench latency insensitive. See Appendix D for more information.

In many cases the overall design will need to satisfy end-to-end latency requirements. For these designs it is still highly advantageous to use a latency-insensitive modeling approach and verify in the post-HLS model that the overall design latency requirements have been satisfied, since this is typically easy to do.

This document only covers sequential HW processes. Combinational HW processes are omitted since their synthesis is straightforward. All the examples and discussion in this document are for SystemC processes that are sensitive only to a single rising clock edge.

This document distinguishes between the following:

1. The *conceptual model* for the scheduling rules.
2. The simulation behavior of a model using the rules in C++ or SystemC.
3. The synthesis of a C++/SystemC model using the rules in an HLS tool such as Catapult.

The goal is to align each of these three cases as closely as possible, so that the user has easy to understand rules, while simulation and synthesis work without surprises. However, as we will see, there are practical considerations which may in certain cases cause small deviations from the conceptual model in either simulation or HLS.

A simple real-world example that illustrates the motivation for the scheduling rules is provided in appendix A of this document.

The examples referred to in this document are available here:

https://github.com/Stuart-Swan/Matchlib-Examples-Kit-For-Accellera-Synthesis-WG/tree/master

The most up-to-date version of this document is available here:

https://github.com/Stuart-Swan/Matchlib-Examples-Kit-For-Accellera-Synthesis-WG/blob/master/matchlib_examples/doc/catapult_user_view_scheduling_rules.pdf

An Analogy from RTL Synthesis

To better understand the specific purpose of this document, let's consider how RTL synthesis works. Say you have a sequential block that you are modeling in Verilog RTL, and it has an output port coded like:

```
Out1 <= #some_delay new_val;
```

In Verilog simulation, if some_delay is less than the clock period of the block, then it will probably not affect the overall cycle level behavior of the system during simulation. However, if some_delay is more than the clock period, it probably will.

During RTL synthesis, all RTL synthesis tools will ignore all delays in the input model, in this case even if some_delay is greater than the clock period. Some RTL synthesis tools might give a warning or error for the code above like "Simulation and synthesis results are likely to mismatch because the delay in model is greater than clock period." Some RTL synthesis tools might outright reject a model containing such delays.

One might argue that RTL synthesis tools should always match the Verilog simulation behavior of the input model. But the overall approach works well because RTL is a good and simple *conceptual model* that users and tool vendors can align around. The slight differences between the *simulation model* and the *conceptual model* used by RTL synthesis tools can be easily managed.

We'll return to this example later in this document.

Next let's clarify some terminology related to signal IO. Say you have a Verilog model like:

```
forever begin
  @(posedge clk);  // wait for 1st rising clock edge
  output1 <= input1 + 10;
  @(posedge clk);  // wait for 2nd rising clock edge
end
```

In this example we say that the read of the input1 signal occurs at the first wait statement, and the write of the output1 occurs at the second wait statement, since that is what you would observe in Verilog simulation when you ran this model. To generalize this, in Verilog designs within implicit state machines like this model which are only sensitive to a clock edge, signal reads occur at the closest preceding wait statement, and signal writes occur at the closest succeeding wait statement.

Catapult HLS Status Concerning Rules in this Document

A separate document provides a list of clarifications related to support within Catapult HLS for the rules described in this document. The items in the list are named *Cat#*, so that each item has a unique number.

This document annotates certain rules with *Cat#* to refer to the items in the separate document.

Terms Used in this Document

Latency Insensitive: In digital hardware design, *latency insensitive* refers to a system that operates correctly despite variable communication delays between components. This is achieved by using mechanisms to decouple computation from communication timing. Such designs improve scalability and reliability in complex systems with unpredictable or variable latencies. ARM's AXI4 and APB are examples of latency insensitive protocols. ARM's AMBA 5 CHI credit-based NOC protocol is also an example of a latency insensitive protocol.

Process: In Verilog, a process is an *always block* and its equivalent constructs. In SystemC, a process is an instance of an SC_THREAD or SC_METHOD.

Message Passing Interface: A *message passing interface* reliably delivers messages (or transactions) from one process to another. This document uses this term to denote the type of communication found in Kahn Process Networks. See https://en.wikipedia.org/wiki/Kahn_process_networks
Message passing read interfaces are always separate from message passing write interfaces – there are no bidirectional message passing interfaces.

Synchronization Interface: A *synchronization interface* synchronizes one process with another and/or with a global clock. For an example of a synchronization interface, see https://en.wikipedia.org/wiki/Barrier_(computer_science)

Signal IO: In digital HW design, signals are the fundamental communication mechanism. Signals enable communication between two HW blocks/processes, but communication with signals in real HW always incurs at least some delay because communication cannot be faster than the speed of light.
In HDLs and in SystemC, signal delays are modeled with the *delayed update* semantics.

blocking / non-blocking: A *blocking* message passing interface suspends the execution of the calling process until the message is either sent or received. A *non-blocking* message passing interface never suspends execution of the calling process: instead, a return code is provided to indicate whether the operation completed or not.

One-way / two-way handshake protocols: A one-way handshake protocol only has one signal between a sender and receiver to synchronize communication. A two-way handshake protocol has two signals (one in each direction) between a sender and receiver to synchronize communication.

shall: This term indicates that a compliant tool or flow is required to follow the indicated rule.

may: This term indicates that a compliant tool or flow is allowed to follow the indicated rule but is not required to do so.

Classes of Operations Involved in Scheduling Rules

There are three classes of operations involved in the scheduling rules:

1. Calls to message passing interfaces (which are all ac_channel methods, all SystemC MIO calls except calls to SyncChannel)
2. Calls to synchronization interfaces (which are calls to ac_sync and Matchlib SyncChannel, also ac_wait and SystemC wait)
3. Signal IO (which are SystemC signal reads and writes, also C++ model *direct inputs*)

All the operations above are referred to as *IO operations*.

Basic Conceptual Model

The basic conceptual model encompasses processes that have no loop pipelining but may have preserved loops. If a process has a preserved loop, then the user may place a wait statement in the loop, or the HLS tool may automatically add one into the body of the loop. For a preserved loop the HLS tool may also add a wait statement at the loop termination. Such automatically added wait statements are called *implicit wait statements* in this document. Wait statements explicitly placed in the model are called *explicit wait statements* in this document.

Note that both *implicit wait statements* and *explicit wait statements* are classified as calls to a *synchronization interface*.

The basic conceptual model rules are:

1. Synchronization interface calls within a process always remain in the source code order.

2. Signal read operations occur at the closest preceding call to a synchronization interface. (Cat1)

3. Signal write operations occur at the closest succeeding call to a synchronization interface.

4. Message passing operations are free to be reordered subject to the following constraints:
   - All message passing operations before a call to a synchronization interface shall be completed when the call to the synchronization interface has completed.
   - All message passing operations after a call to a synchronization interface shall not start until the call to the synchronization interface has completed.
   - Two message passing operations on separate interfaces which appear in sequence in the model may be executed either in the same sequence or in parallel in simulation and in synthesis, but they shall not be executed in the reverse sequence. (Cat2)

Some explanation for the very last point: While pure message passing systems with unbounded FIFOs are immune to reordering issues, real-world hardware implementations have finite buffer sizes. Reversing the order of message passing calls in a system with bounded FIFOs can introduce deadlocks that were not present in the original model. The last rule ensures that HLS tools cannot introduce such deadlock cases.

Pipelined Loops

When a loop is pipelined in HLS, the body of the loop is split into pipeline stages. HLS may start the next iteration of the loop before the current iteration has completed.

The user may place wait statements in the body of a pipelined loop to manually separate operations into their respective pipeline stages. Alternatively, the HLS tool may implicitly add these wait statements into the model to separate the pipeline stages. We call the former *explicit pipeline stage wait statements*, and the latter *implicit pipeline stage wait statements*.

Both explicit and implicit pipeline stage wait statements are classified as calls to a *synchronization interface*.

The scheduling rules for pipelined loops are the same as the rules given in the *basic conceptual model*, with the addition of these pipeline stage wait statements into the set of calls to synchronization interfaces. However, one rule from the conceptual model is relaxed when loops are pipelined: During loop pipelining, message passing read operations from later loop iterations may be scheduled before or in parallel with message passing write operations from the current loop iteration, and SystemC wait() statements and ac_wait statements, if present in the loop body, will not preclude this from occurring. Potential deadlocks are avoided by having the pipeline automatically flush. See Appendix B for further discussion.

When HLS pipelines a loop, multiple iterations of the loop are overlapped and execute at the same time. During loop pipelining, for all IO operations, HLS shall ensure that an access to a specific message passing interface, signal, or synchronization interface shall not be moved over or in parallel with an access to the same interface from a different loop iteration.

When HLS pipelining occurs, any signal reads and writes and associated synchronization interface calls become embedded in specific pipeline stages. With pipelining, the post-HLS model may begin execution of the next loop iteration before the current iteration has completed. Considering the entire set of signals read or written by the process, if HLS pipelining would cause the order of signal IO to differ between the pre-HLS and post-HLS models, or if any two signal IO operations that occur on the same clock edge in the pre-HLS model would not occur on the same clock edge in the post-HLS model, then the HLS tool shall detect such a situation and require the user to explicitly indicate in the model source (e.g., via a pragma) that HLS pipelining can still be used. (Cat7) Another way to state this requirement is the following: If a process communicates with other processes using only latency-insensitive protocols, then HLS pipelining by default shall not break any of the protocols.

Direct Inputs

Normal SystemC signal read operations occur at the closest preceding synchronization interface call (e.g., wait statement) in both the pre-HLS and post-HLS models. (Cat1). If the HLS tool adds states to the process, or if it pipelines the process, then this implies that the HLS tool must add registers for each such read operation such that the read occurs where specified in the pre-HLS model, and the value is stored until the point where it is consumed in the post-HLS model. The area cost of such registers may be high if there are a lot of signals, and in some cases, it may be unneeded area since the value of the signals may not actually need to be stored internally to the process.

The simplest case is if such external signals are held stable after the process comes out of reset. In this case, HLS may assume that it is free to read the signal values as late as possible, with no need for register storage. This case is handled with the following pragma on SystemC signals and ports (Cat6):

#pragma hls_direct_input

To ensure that there are no pre-HLS versus post-HLS simulation mismatches, the environment that drives the signal shall hold it stable after all receiving processes that use this signal with this pragma come out of reset. Note that with this approach it is allowable to have *dynamic resets*, i.e. activation of block resets and associated resetting of direct input signals as part of the normal operation of the HW.

A related but somewhat more complex case is where input signals to the HW block may only be changed at "agreed upon" times, typically while the portion of the HW block that relies on them is temporarily idle. For example, a block may process 2D images. At the start of each new image, it may be desirable for the TB or external environment to update the control signals for how the block will process the next image. This needs to be done precisely since typically HLS designs are pipelined, and the HW pipeline for the current iteration must be fully *ramped down* before the input signals can be updated to affect the next iteration. In this case we can use the SystemC *SyncChannel* or C++ *ac_sync* primitives to precisely synchronize the DUT with the TB/environment to enable the input signals to be updated at the correct time. The `#pragma hls_direct_input_sync` directive shown below associates the sync operation with the direct inputs that it controls. The precise synchronization scheme shown here ensures that there are no pre-HLS versus post-HLS simulation mismatches even though we are using direct inputs and also changing their values while the design is executing.

```
  // This is example 61* in Catapult Matchlib examples
  sc_in<bool> SC_NAMED(clk);
  sc_in<bool> SC_NAMED(rst_bar);

  Connections::Out<uint32_t>  SC_NAMED(out1);
  Connections::In<uint32_t>   sample_in[num_samples];
  Connections::SyncIn         SC_NAMED(sync_in);
#pragma hls_direct_input
  sc_in<uint32_t> direct_inputs[num_direct_inputs];


  void main() {
    out1.Reset();
    sync_in.Reset();

#pragma hls_unroll yes
    for (int i=0; i < num_samples; i++) {
      sample_in[i].Reset();
    }

    wait();  // reset state

    while (1) {
#pragma hls_direct_input_sync all
     sync_in.sync_in();

#pragma hls_pipeline_init_interval 1
#pragma hls_stall_mode flush
     for (uint32_t x=0; x < direct_inputs[0]; x++) {
       for (uint32_t y=0; y < direct_inputs[1]; y++) {
        uint32_t sum = 0;
```

```
#pragma hls_unroll yes
        for (uint32_t s=0; s < num_samples; s++) {
           sum += sample_in[s].Pop() * direct_inputs[2 + s];
        }
        ac_int<32, false> ac_sum = sum;
        ac_int<32, false> sqrt = 0;
        ac_math::ac_sqrt(ac_sum, sqrt);  // internal loop unrolled in cat .tcl file
        if (sqrt > direct_inputs[7])
          out1.Push(sqrt);
      }
     }
    }
  }
};
```

From the perspective of the testbench or the environment, the updating of the signals controlled by the `hls_direct_input_sync` directive shall only occur at a precise point. The TB shall first wait for the `rdy` signal for the sync to be asserted by the DUT, and then the TB shall update all the input signals it wishes to change while simultaneously driving the sync `vld` signal high for one cycle.

It is important to note that the only safe operation to use to synchronize the updating of direct inputs is the sync operation as shown above. Other operations such as Push/Pop or ac_channel operations should not be used for this.

In the example above the DUT block that is being synthesized determines when to call sync, and thus it determines when the direct inputs will be updated. In some cases, it may be necessary for the environment around the DUT to determine when the direct updates should be updated. In this case the same approach as shown above should be used, however a separate input from the environment to the DUT (either using a signal or a message passing interface) should request that the DUT call sync as soon as feasible. This will ensure that the DUT has properly ramped down its pipeline and is ready to receive the newly updated direct inputs as per the overall synchronization scheme described above.

Additional Options for Scheduling Message Passing Interfaces

The following option may be added during HLS (Cat2):

```
STRICT_IO_SCHEDULING=relaxed
```

when this is specified, the HLS tool is allowed to reorder message passing interface calls freely. However, it is still not allowed to move these calls across synchronization interface calls.

It is recommended that this relaxed option only be used when the user wants to see what order the HLS tool prefers to schedule message passing interface calls (e.g., to achieve best QOR). Once the user knows the preferred order, it is recommended that the user modify the pre-HLS source code to reflect the preferred order, and then return to the use of the default scheduling modes. This methodology ensures that HLS cannot introduce any new deadlocks into the system as described earlier.

Scheduling of Array Accesses

Arrays may appear in HLS models, and they may be preserved through synthesis and mapped to RAMs. Pointers may also appear in HLS models, and pointer dereferences are resolved to array accesses during HLS.

There are two cases to consider for arrays for the purposes of the scheduling rules:

1. Array instantiation in the HW is internal to the process
   - The array accesses are not visible external to the process, and thus their scheduling is also not visible externally.
   - All the scheduling rules described elsewhere in this document remain unaffected in this case.

2. Array instantiation in the HW is external to the process
   - In this case the user model shall indicate how array accesses are mapped onto IO operations that are external to the process.
   - We call this the *array access mapping layer*. The *array access mapping layer* maps array accesses onto IO operations described above (signal IO, message passing interface calls, and synchronization calls).
   - The user model may indicate that it is allowable for HLS to transform array accesses, for example, to cache, merge, split, or reorder array accesses (e.g., to improve QOR). These transformed operations, if allowed, are an outcome from the use of the *array access mapping layer*.
   - In all cases the scheduling rules described elsewhere in this document for the core IO operations (signal IO, message passing calls, synchronization calls) remain unaffected.
   - In all cases array accesses shall remain constrained by any explicit synchronization interface calls present in the process in the source model.
   - Note that if transformed operations occur and array accesses are visible externally in both pre-HLS and post-HLS model, then comparison of pre-HLS and post-HLS behaviors may need to account for the transformed operations.

When a loop is pipelined, multiple iterations of the loop are overlapped and execute at the same time. During loop pipelining, an access to a memory interface (or array) may be moved over or in parallel with an access to the same memory if the HLS tool can prove the reordering is conflict free. If the array/memory is external to the process, the *array access mapping layer* shall indicate that such reordering is allowable if such reordering is to occur during loop pipelining.

Additional States Added by HLS Synthesis

By default, HLS synthesis tools may add additional states to processes (e.g. add latency to enable resource sharing), which may introduce latency differences in the interface behavior between the pre-HLS and post-HLS models. These additional states are never included in the set of *synchronization interface calls* as described above.

When the directive IMPLICIT_FSM=true is set on a process, the HLS synthesis tool shall ensure that the cycle level behavior of the interfaces of the pre-HLS and post-HLS models shall be identical. With this option, the internal state machines of the pre-HLS and post-HLS models will be the same.

When the directive IO_MODE=FIXED is set on a process, the HLS synthesis tool shall ensure that the cycle level behavior of the interfaces of the pre-HLS and post-HLS models shall be identical. With this option, it is still possible that the state machine internal to the process is different between the pre-HLS and post-HLS models (e.g. the post-HLS model may choose to use a pipelined multiplier where the pre-HLS model did not.)

Avoiding Pre-HLS and Post-HLS Simulation Mismatches

The scheduling rules described in this document are designed to be easy to understand, while providing good QOR via HLS and generally avoiding any mismatches between the pre-HLS and post-HLS simulation behaviors.

Non-blocking message passing operations (PushNB/PopNB in SystemC, C++ ac_channel nb_read/nb_write) are a potential source of mismatches between pre-HLS and post-HLS simulations since their behavior is inherently dependent on the latency within the model, which often changes during HLS. Because of this, non-blocking message passing interfaces should only be used when no alternative approach is possible. For example, non-blocking message passing interfaces are required to model time-based arbitration of multiple message streams which access a shared resource. A full discussion of recommended guidelines on the use and verification of non-blocking message passing interfaces is beyond the scope of this document. (However, see discussion in Appendix C). Note that the scheduling rules described previously in this document fully specify how HLS tools are required to schedule such operations.

Unidirectional message passing between two processes should not be relied on to achieve synchronization between the two processes, since in general the message latency and storage capacity between the processes may be variable. Also, it is possible that prefetch behavior of the message reader may vary. Bidirectional message passing can be relied upon for synchronization between two processes. (An example of this is the AXI4 ar and r, and aw and b, channels). Note that such synchronization is weaker than explicit synchronization like SyncChannel or signal IO that uses a two-way handshake.

SystemC signal IO operations are a potential source of pre-HLS versus post-HLS simulation mismatches since timing behaviors may change between the two models. The following section provides guidance and rules to help avoid potential mismatches due to signal IO.

When signal IO operations are synchronized with a wait statement, there generally should be a proper two-way handshake associated with the wait statement so that the signal IO is latency insensitive. (Note that this statement does not apply to the signal synchronization approaches described in the Direct Inputs section.)

For example, here's a simple two-way handshake protocol when writing the signal `out_dat`:

```
out_dat = value;
out_vld = 1;
do {
  wait();
} while (out_rdy != 1);
out_vld = 0;
```

And here's a two-way handshake example when reading signal `in_dat`:

```
in_rdy = 1;
do {
  wait();
} while (in_vld != 1);
value = in_dat;
in_rdy = 0;
```

Some signal-level protocols have different two-way handshaking approaches (e.g. ARM APB), but they are still latency insensitive.

If signal IO operations are associated with a wait statement and that wait statement does not have a proper two-way handshake, then the signal IO is likely to be latency sensitive and may result in pre-HLS versus post-HLS simulation mismatches. In some systems a one-way signal handshake is sufficient for reliable system operation. See Appendix E for further discussion.

The scheduling rules state that signal IO operations occur at either SystemC wait statements or SyncChannel calls (sync_in and sync_out). In the remainder of this section, we will use *wait* statement to refer to both.

RULE 1: It is always best coding style to group signal write operations just before their corresponding wait statement, and signal read operations just after their corresponding wait statement. (Cat3). An example is below:

```
sc_in<int> i1;
sc_in<bool> go;
sc_out<int> o1;
void my_thread() {
  int new_val=0;
  while (1) {
    o1.write(new_val);
    do {
      wait();
    } while (!go.read());
    new_val = i1.read();
    new_val = some_function(new_val); // function has no internal IO
  }
}
```

By placing the signal IO operations as close as possible to their corresponding wait statement, the HW intent is very clear. And there is no benefit either in terms of simulation performance or HLS QOR if they are placed further away from their corresponding wait statement.

Let's look at another similar example, which now also uses a Matchlib Connections blocking Pop operation:

```
sc_in<int> i1;
sc_in<bool> go;
sc_out<int> o1;
Connections::In<int> pop1;
void my_thread() {
  int new_val=0;
  while (1) {
    o1.write(new_val);
```

```
    do {
      wait();
    } while (!go.read());
    int pop_val = pop1.Pop();
    new_val = i1.read();
    new_val = some_function(new_val + pop_val); // function has no internal IO
  }
}
```

According to the *Conceptual Model scheduling rules* part of this document, the Pop operation does not affect the scheduling of the i1.read() operation. However, it is possible that in the pre-HLS SystemC simulation, the Pop operation may block for a clock cycle or more (only if no items are available to Pop). This means that it is possible in the pre-HLS simulation that the value of the signal i1 may change between the time before the Pop operation starts and the time it completes. If this occurs, there may be a pre-HLS and post-HLS simulation mismatch if the HLS tool schedules the i1.read() operation at the wait statement. The proper fix to this issue (to reiterate) is to move the i1.read() operation as close as possible to its corresponding wait statement. This will make the potential simulation mismatch disappear, and it will not affect QOR or simulation performance.

To automatically avoid all such potential pre-HLS versus post-HLS simulation mismatches, HLS tools may provide error or warning messages in cases where models have the pattern shown above. Precisely speaking: if a blocking message passing operation separates a signal read or write operation from its corresponding synchronization interface call, then the HLS tool may emit an error or warning indicating that reordering the signal IO operation and the message passing operation in the source text is advisable.

Another scenario in which RULE 1 applies is shown below:

```
sc_in<bool> go;
sc_out<int> o1;
void my_thread() {

  while (1) {
    wait();              WAIT 1
    o1.write(some_value);
    if (go.read()) {
      some_value = some_function();
    }
    else {
      wait();            WAIT 2
    }
    some_other_function();
    wait();              WAIT 3
  }
}
```

The signal read of `go` is clearly and uniquely associated with WAIT 1. However, the signal write of `o1` associates with WAIT 2 if `go` is false and WAIT 3 if it is not. This is a violation of RULE 1 and should be flagged as an error during HLS. The fix, as before, is to move the signal IO operation as close as possible to its intended wait statement so that the association is unconditional.

Next, let's consider *rolled* (or *preserved*) loops that perform signal IO within the loop body. Consider the following example:

```
 sc_in<int> i1;
 sc_out<int> o1;
 void my_thread() {
     wait(); // reset state
     while (1) {
       wait(); // start of while loop
       #pragma hls_unroll no
       for (int i=0; i < 10; i++) {
         o1.write(i1.read() * i);
       }
   }
}
```

Note that the `i1.read()` operation is located inside the `for` loop, so presumably the user's intent is that it should be read as the loop iterates. *If that is not the user's intent, then he simply should move the i1.read() operation before the loop start.*

In the post-HLS simulation, each iteration of the loop will consume at least one clock cycle, and a new value for i1 will be read (and a new value for o1 written) on each iteration. Again, this is the user's intent as per the code.

In the pre-HLS simulation, the for loop body will execute in zero time, and only the last write to o1 will have any effect. The solution to avoid this mismatch is to manually place a `wait()` statement within the for loop body so that the signal IO synchronization is explicit in the pre-HLS simulation.

RULE 2: If you have signal IO operations within *rolled* (or *preserved* loops), manually place a wait statement within the body of the loop to avoid pre-HLS versus post-HLS simulation mismatches, and while doing so also follow RULE 1.

To automatically prevent these types of pre-HLS versus post-HLS simulation mismatches, HLS tools may emit warning or error messages if they encounter a rolled loop which has signal IO operations within the loop body, and the loop body does not have a wait statement included within the loop body. (Cat4)

If a pre-HLS model adheres to RULE 1 and RULE 2, then all signal IO in the post-HLS model shall occur only at clock cycles that correspond to explicit wait statements or explicit synchronization statements. User designs that do not adhere to both RULE 1 and RULE 2 are *ill-formed*.

Returning to the Analogy from RTL Synthesis

At the beginning of this document, we presented the example of a Verilog sequential block with an output coded like:

```
Out1 <= #some_delay new_val;
```

Recall that in Verilog simulation, if some_delay is less than the clock period of the block, then it will probably not affect the overall cycle level behavior of the system during simulation. However, if some_delay is more than the clock period, it probably will.

During RTL synthesis, all RTL synthesis tools will ignore all delays in the input model, in this case even if some_delay is greater than the clock period. Some RTL synthesis tools might give a warning for the code above like "Simulation and synthesis results are likely to mismatch because delay in model is greater than clock period."

HLS tools that choose to adhere very closely to the *conceptual model* presented in this document should automatically provide errors or warnings for violations of RULE 1 and RULE 2 as described in the section above. This is analogous to the error message that the RTL synthesis tool would provide in the example directly above.

However, it is possible also that HLS synthesis tools may choose to adhere in these cases more closely to the pre-HLS SystemC simulation behavior. In this case such HLS tools might not provide any errors or warnings for violations of RULE 1 and RULE 2. This is analogous to an RTL synthesis tool being very smart (maybe even too smart) about synthesizing matching HW based on the actual value of some_delay in the example directly above.

Summary

At the beginning of this document, we said that the intent was to present "no surprises" to a DV engineer who is using a single testbench to verify both the pre-HLS and post-HLS models. The key aspects of the document which support this are:

- Three groups of IO operations are defined (message passing, signal IO, and synchronization calls) and each is treated uniformly. These IO operations are easy for verification engineers to understand because they are already using them in their testbenches.

- The document specifically avoids complex constructs such as *protocol regions* used in some HLS tools.

- The document preserves the ability of the pre-HLS SystemC model to be *throughput accurate* by using a library such as Matchlib.

- Synchronization calls can affect the scheduling of signal IO operations, and synchronization calls can affect the scheduling of message passing calls, but message passing calls cannot affect the scheduling of signal IO operations and vice-versa.

- HLS cannot by default reverse the order of message passing calls, so it cannot introduce new deadlocks into the post-HLS model.

- HLS pipelining is largely a *don't care* from the perspective of the verification engineer. If the design and the testbench are insensitive to changes in latency, and if external array accesses are not reordered or rearranged during loop pipelining, then the possible use of HLS pipelining will not affect verification, provided the pipelines flush automatically. Even if the design or testbench are sensitive to changes in latency, or if they are sensitive to reordering or rearranging of external memory accesses due to the use of HLS loop pipelining, then the behavior of the DUT will only change in expected (rather than unexpected) ways.

- Signal IO operations in the post-HLS model by default always occur exactly at synchronization points (e.g., wait statements) that are either explicit in the pre-HLS model or easily deducible based on the use of loop roll/unroll or loop pipelining directives. The HLS tool can check that every signal IO operation is tightly coupled with exactly one wait statement / synchronization operation in the pre-HLS model.

Appendix A – Factory Analogy

The scheduling rules described in this document apply to pre-HLS and post-HLS HW models. Although the rules may seem abstract and perhaps even arbitrary, they are shaped by the need to model systems in the real world that are required to have predictable behavior.

To understand the motivation behind the rules, it might be helpful to draw a simple real-world analogy and its correspondence to the rules described earlier.

Consider a factory that produces various types of wooden furniture:

The factory consists of people (processes) stationed at workbenches with various tools.

Each person is given written instructions about the specific tasks they are to perform (C++ code within a process).

People are instructed to send or receive objects (messages) to or from other people in the factory.

Sending or receiving objects may be blocking or non-blocking from the perspective of a person.

There is a clock with a second hand on the factory wall that everyone can see. (HW clock).

People have colored flags they can raise or lower to communicate with other people (signals).

If someone raises or lowers a flag, this is only seen by others the next time the clock second hand is at the top of the clock. (propagation delay of signals between sequential processes)

A person can choose to pipeline the tasks that he was assigned by hiring subordinates (pipeline stages) and having each one do a subtask. In general, this will improve the throughput for the tasks that the person was assigned.

It is possible for two people to explicitly synchronize their work by communicating via a synchronization protocol such as a barrier (synchronization).

It is possible to restart the work of some or all the people by raising a reset flag (HW resets).

Assume:

1. That the time that people take to complete their various tasks is in general variable, and similarly that the time that objects (messages) take to pass between people is in general variable.
2. That each person in general wants to complete their tasks as quickly and efficiently as possible.
3. That the factory needs to be able to make multiple types of furniture at the same time. For example, it might make chairs that need to be different colors or have different styles.

To reliably produce output, each person in the factory will need to adhere to rules like those described in this document.

Here's a sketch of a specific way the above example relates to the scheduling rules:

Person 1 sends chair seats and chair backs to Person 2. This is done with blocking operations, and there is no storage capacity between the people when the objects are sent. (This means that a Push operation cannot complete until the corresponding Pop operation is performed.) Assume that the fastest an object can be sent between people is 1 minute.

The written instructions that person 1 is given are:

```
while (1) {
 // internal processing code for seats and backs ...
 seats.Push(seat_object);
 backs.Push(back_object);
}
```

The written instructions that person 2 is given are:

```
while (1) {
 seat_object = seats.Pop();
 back_object = backs.Pop();
 // internal processing code for seats and backs ...
}
```

If both person 1 and person 2 choose to perform their tasks sequentially as written, then objects will be passed over time as:

```
            Person 1            Person 2
Minute 1:   seats.Push(seat1);
Minute 2:   backs.Push(back1);  seat1 = seats.Pop();
Minute 3:   seats.Push(seat2);  back1 = backs.Pop();
Minute 4:   backs.Push(back2);  seat2 = seats.Pop();
Minute 5:                       back2 = backs.Pop();
```

If both person 1 and person 2 choose to perform their IO operations in parallel, then objects will be passed over time as:

```
           Person 1                                Person 2
Minute 1:  seats.Push(seat1); backs.Push(back1);
Minute 2:  seats.Push(seat2); backs.Push(back2);   seat1 = seats.Pop(); back1 = backs.Pop();
Minute 3:                                          seat2 = seats.Pop(); back2 = backs.Pop();
```

If person 1 chooses to perform his IO operations in parallel, and person 2 chooses to perform his IO operations sequentially as written, then objects will be passed over time as:

```
           Person 1                                  Person 2
Minute 1:  seats.Push(seat1); backs.Push(back1);
Minute 2:                                            seat1 = seats.Pop();
Minute 3:  seats.Push(seat2); backs.Push(back2);     back1 = backs.Pop();
Minute 4:                                            seat2 = seats.Pop();
Minute 5:                                            back2 = backs.Pop();
```

If person 1 chooses to perform his IO operations sequentially as written, but person 2 chooses to perform his IO operations sequentially in the reverse order as it was written, then objects will be passed over time as:

```
            Person 1            Person 2
Minute 1:   seats.Push(seat1);
Minute 2:   backs.Push(back1);  back1 = backs.Pop();
Minute 3:                       seat1 = seats.Pop();
```

In this case the person 1 seats.Push(seat1) operation at minute 1 will not complete at the start of minute 2. This means that the person 1 backs.Push(back1) operation will never start, and thus the Person 2 back1 backs.Pop() operation will never complete. So, the system will be in deadlock.

This example directly corresponds to the ordering rules related to message passing operations in the *basic conceptual model* within this document, and in particular the specific rule that disallows reversing the order of message passing operations.

Appendix B – Pipelined Loops and Message Passing

The scheduling rules state that during loop pipelining message passing read operations from later loop iterations may be scheduled before or in parallel with message passing write operations from the current loop iteration. This might seem to be in contradiction with the rules for message passing operations described in the *basic conceptual model*. However, it is important to remember that each message passing channel is *self-synchronized*. If a process has a pipelined loop that automatically flushes in the post-HLS model, then from an external perspective the only effect of allowing the message passing read operations to occur earlier is like a *pre-fetch* operation. If the incoming message is not available, the pipelined process in the post-HLS model will flush, and its behavior from the external perspective will appear exactly as if it is not pipelined.

Appendix C – Verification of Designs that are Partially Latency Sensitive

If designs are completely latency insensitive, verification of the pre-HLS versus post-HLS models is straightforward. However, if the design contains some components such as arbiters which have latency-sensitive behavior, and if that latency sensitive behavior is in some cases externally visible to the DUT, then verification becomes somewhat more complex. Techniques can be applied to simplify verification.

Consider a design which has an arbiter like Matchlib toolkit example 09*. The arbiter uses non-blocking PopNB() on its inputs, and blocking Push() on its output to emit the winner of the arbitration. Since the latency between the pre-HLS and post-HLS designs will differ, the order of transactions presented to the arbiter in the two scenarios will differ, and thus the order of the winners will differ. This may result in verification mismatches between the two models.

To force the pre-HLS and post-HLS simulations to match, we can run the two designs side-by-side. We can snoop the inputs to the arbiter in the post-HLS model, and only allow the inputs to the pre-HLS arbiter to proceed when the corresponding inputs are seen in the post-HLS model. This will force the order of the inputs to be equivalent between the pre-HLS and post-HLS models, and thus both arbiters will pick the same winners. When this technique is used, the pre-HLS simulation will be throttled by the post-HLS simulation, but the overall verification will still work properly.

Another related example involves interrupt request signals feeding into an interrupt controller within a CPU model. Typically, each interrupt request signal is a single bit sc_signal<>, indicating that an interrupt request is pending. If the requests originate from accelerator blocks that are being synthesized through HLS, then because of the latency differences between the pre-HLS and post-HLS models, the order of interrupt requests arriving at the interrupt controller will differ between the two models, and this will likely result in verification mismatches. To force the order of the requests to match, we snoop the requests arriving at the controller in the post-HLS model, and only then allow the requests to be seen in the pre-HLS model.

Appendix D – Modeling Latency Sensitive Protocols

In some cases, designs or testbenches may use protocols which are latency sensitive. These situations can be handled by isolating the latency sensitive portions to small, self-contained parts, and then keeping the rest of the design and testbench latency insensitive.

Consider a case where a signal-level protocol is latency sensitive. The protocol will have specific timing behavior that it must meet. To handle this, we create a transactor that has two sides: the first side interacts with the signals and handles the detailed timing requirements, and the second side sends and receives messages with the rest of the system. The message passing side is latency insensitive.

To properly model the detailed timing behavior, the transactor is modeled at the cycle-accurate level in SystemC. There is an example of such a transactor model in the following document and example:

https://github.com/Stuart-Swan/Matchlib-Examples-Kit-For-Accellera-Synthesis-WG/blob/master/matchlib_examples/examples/53_transactor_modeling/transactor_modeling.pdf

Appendix E – One-Way Handshake Protocols

In some systems a one-way signal handshake is sufficient for reliable system operation. When using a one-way handshake, the implicit assumption is that the “missing” handshake isn’t required since it is always true.

For example, in time-domain signal processing hardware designs, signal processing HW blocks may input new samples at a fixed rate. The HW blocks are always ready to receive new samples on each clock, but they need to know if the samples are valid. These types of designs can be modeled with the one-way handshake dat/vld protocol demonstrated in this example:

https://github.com/Stuart-Swan/Matchlib-Examples-Kit-For-Accellera-Synthesis-WG/tree/master/matchlib_examples/examples/32_dat_vld

Appendix F – Scheduling Rules: Modeling Guidelines Summary

1) Prefer to use Connections::In/Out + SyncChannel over signal IO and wait().

2) Prefer to use Pop()/Push() over PopNB()/PushNB().

3) When pipelining a loop, prefer to use `hls_stall_mode flush`.

When using signal IO:

4) If modeling a cycle-accurate process, use `disable_spawn` and follow example 53_transactor_modeling style and do not use Push/Pop in the process.

5) If modeling a *direct input*, use `hls_direct_input` and possibly `hls_direct_input_sync`, and only change the signal at allowed times.

6) If combining signal IO with Connections::In/Out in same process, use proper signal IO handshake that does not rely on In/Out ports for process synchronization. Place signal IO operations very close to their wait() statements.

7) Unless you are modeling a cycle-accurate process, you should expect that latency will change between the pre-HLS and post-HLS models.

Appendix G –Equivalence Rules

This document informally states that a primary goal is to present “no surprises” to a DV engineer using the same testbench to verify the pre-HLS and post-HLS models.

Somewhat more formally, we can show how the scheduling rules within this document establish a set of equivalence rules between the pre-HLS and post-HLS models that provide strong guarantees about their behaviors.

The *basic conceptual model* establishes the base behavior for a single process:
1) Synchronization interface calls always remain in source code order.
2) Signal reads occur at closest preceding synchronization interface call.
3) Signal writes occur at closest succeeding synchronization interface call.

4) All message passing calls after a synchronization interface will not start until the synchronization interface call has completed.
5) All message passing calls before a synchronization interface shall be completed when the synchronization interface call has completed.
6) Two message passing operations on separate interfaces which appear in sequence in the model may be executed either in the same sequence or in parallel in simulation and in synthesis, but they shall not be executed in the reverse sequence.
7) Synchronization calls can affect the scheduling of signal IO operations, and synchronization calls can affect the scheduling of message passing calls, but message passing calls cannot affect the scheduling of signal IO operations and vice-versa.
8) The STRICT_IO_SCHEDULING=relaxed option is not used.
9) In the pre-HLS model, the storage capacity of message channels is zero (rendezvous-style communication).

Conceptually, systems are composed of many such processes which follow the *basic conceptual model,* and which communicate using only *synchronization interface calls* and *latency insensitive* signal level protocols and *blocking message passing* calls, and no *non-blocking message passing* interface calls. Here we call this system of processes the *conceptual system*.

Practically, the modeling approach described above is too restrictive for real-world systems with optimized HW, so the scheduling rules in this document allow for key HW optimizations. But this is done in a carefully controlled manner, such that the HW optimizations do not break equivalent behavior with the *conceptual system*.

Here is a summary of the HW optimizations that the scheduling rules allow, and a brief description of how we maintain equivalent behavior with the *conceptual system*:

1. Pipelined loops: Message passing read operations from later loop iterations can execute before message write operations from the current loop iteration. Assuming the pipeline automatically flushes, this behavior change can be viewed as a *prefetch* operation. If the message to be read is unavailable, the pipeline will flush, and the process will appear exactly as if it is not pipelined.

2. Pipelined loops: HLS by default will not break any latency-insensitive signal IO protocols when pipelining loops.

3. Direct inputs: Signals that do not change after a process comes out of reset can be read as late as possible using *hls_direct_input*, saving register area.

4. Direct inputs: When *hls_direct_input_sync* is used, signals read by a process are updated at the precise point at which the HW pipeline is guaranteed to be ramped down, so the signal values do not need to be saved within pipeline registers, saving register area.

5. Memory accesses: HLS can reorder memory accesses within a process only if it can prove that the reordering is conflict-free.

6. Shared memories: Memories shared between processes are modeled as simple C arrays but always include explicit synchronization between processes in both the pre-HLS and post-HLS models.

7. Non-blocking message passing interfaces: Equivalent behavior between pre-HLS and post-HLS systems can be maintained by snooping the post-HLS system and using this to delay pre-HLS message arrival.

8. Latency-sensitive global signal IO: Equivalent behavior between pre-HLS and post-HLS systems can be maintained by snooping the post-HLS system and using this to delay pre-HLS signals.

9. Latency-sensitive local signal IO: Isolate local latency-sensitive protocols to small, dedicated transactors, and use latency-insensitive signal IO or message passing to communicate with the rest of the system.

10. One-way signal handshake protocols: Only use in cases where backpressure is not possible and embed assertions into the model to check that this is always true.

11. Combinational processes: If the pre-HLS model contains combinational processes, they will have identical behavior in the post-HLS model and thus they cannot introduce any differences between the models.

Taken as a whole, these rules allow full system verification to be performed on the pre-HLS system model. Full system verification does not need to be repeated on the post-HLS RTL system, provided the equivalence rules described above are followed.

Formalization of the Equivalence Rules in Appendix G

The following notation is used in this section.

Symbol Meaning

P	A process (thread or method) in the design
Σ	The alphabet of observable IO actions
τ_P	A (finite or infinite) trace of observable actions executed by process P, ordered by global clock cycles
S	The subset of Σ that are synchronization calls (explicit wait, SyncChannel, implicit pipeline-stage waits)
R	The subset of Σ that are signal reads
W	The subset of Σ that are signal writes
M	The subset of Σ that are message-passing calls (Push, Pop, etc.)

A *system trace* is the tuple
τ = (τ_P)_P, one component per process.
For any action a ∈ Σ, let clk(a) be the clock cycle at which a is *committed*.

---

1. Conceptual Model for a Single Process

For every process P the pre-HLS trace τ_P must satisfy the following *partial-order constraints*:
R1 (Program order of S).
∀ s1, s2 ∈ S : s1 <_src s2 ⇒ clk(s1) < clk(s2)

R2 (Location of R/W).
$\forall r \in R : clk(r) = clk(pred\_S(r))$
$\forall w \in W : clk(w) = clk(succ\_S(w))$
where pred_S(r) (resp. succ_S(w)) is the closest synchronization call that precedes (resp. follows) the statement in source order.
If no lexical predecessor (successor) exists, a virtual Sync at time 0 (resp. ∞) is assumed.
R3 (Isolation around S).
Let pref_S(s) (resp. suff_S(s)) be the set of message calls lexically before (resp. after) synchronization call s. Then
$\forall m \in pref\_S(s) : clk(m) \le clk(s)$ and $\forall m \in suff\_S(s) : clk(s) < clk(m)$.
R4 (Non-inversion of independent M).
For m1,m2 ∈ M operating on distinct channels,
$m1 <\_src\ m2 \Rightarrow clk(m1) \le clk(m2)$.
R5 (Rendezvous channel capacity in pre-HLS model).
For every pre-HLS channel c, at any clock cycle t, the number of unmatched Push_c actions committed is zero, and no Push_c may commit unless a matching Pop_c also commits at t.
The pre-HLS *conceptual* semantics for a process is thus a labeled partial order
$(\Sigma, \le\_P)$, where $\le\_P$ is generated by R1–R5.

---

2. Equivalence Relation

Let τ_pre and τ_post be the traces of the same process before and after HLS.
We say $\tau\_pre \approx \tau\_post$ iff all of the following hold:
E1 (Synchronization-order preservation).
$\forall s1,s2 \in S : clk\_pre(s1) < clk\_pre(s2) \Rightarrow clk\_post(s1) < clk\_post(s2)$
E2 (Signal-visibility preservation).
$\forall r \in R : clk\_post(r) = clk\_post(pred\_S(r))$
$\forall w \in W : clk\_post(w) = clk\_post(succ\_S(w))$
E3 (Safe message re-ordering).
For every pair m1,m2 ∈ M on distinct channels,
$m1 <\_src\ m2 \Rightarrow clk\_post(m1) \le clk\_post(m2)$.
FIFO order is always preserved on same channel:
$\forall m_1,m_2 \in M$ on the same channel: $clk\_pre(m_1) < clk\_pre(m_2) \Rightarrow clk\_post(m_1) < clk\_post(m_2)$
E4 (Per-channel FIFO semantics).
For every channel c, the sequence of Push_c and Pop_c actions in τ_post is a legal FIFO schedule (no dropped or duplicated messages).
E5 (Messages cannot cross syncs).
For every message m and synchronization call s,
$clk\_pre(m) \le clk\_pre(s) \Rightarrow clk\_post(m) \le clk\_post(s)$
$clk\_pre(m) > clk\_pre(s) \Rightarrow clk\_post(m) > clk\_post(s)$

See *Appendix I – Per Process Trace Equivalence Proof* for formal proof of the following:
For any single pre-HLS process P that obeys the conceptual rules R1–R5 and communicates over channels of finite capacity $B(c) \ge 0$, every finite observable trace produced by P has a matching post-HLS RTL trace that satisfies the equivalence properties E1–E5.

---

3. Allowed Transformations and Why They Preserve ≈

Transformation permitted by the rules	Formal justification
Loop pipelining (overlaps iterations)	Creates additional S actions (implicit stage waits). Rules R1–R5 are applied with the larger S set. E1–E5 hold because actions of different iterations that refer to the same channel or signal remain separated by at least one S.
Re-ordering of message reads vs. later writes inside a pipeline	Allowed only when the two actions are on *different* channels and the pipeline auto-flushes. This preserves FIFO property per channel (E4) and does not violate E3 and E5.
#pragma hls_direct_input (late sampling of static signals)	This pragma signals a designer-guaranteed exception to Rule E2, where functional equivalence is preserved due to the asserted stability of the signal's value after reset.
#pragma hls_direct_input_sync (controlled updates)	This pragma establishes a specific timing exception to Rule E2 for the controlled input signals. The update of these signals is explicitly bound to the specific, preceding synchronization call s*, creating a new, localized rule where clk(w_update) = clk(s*). Functional equivalence is preserved because this explicit timing contract replaces the default rule and is applied consistently to both the pre-HLS and post-HLS models.
Memory-access reordering proven conflict-free	Let addr(a) be the symbolic address of access a. If the tool proves $addr(a_1) \neq addr(a_2)$ for the overlapped window, then the commutation of a_1 and a_2 is observationally silent; no external signal/message depends on the internal order.
Implicit FSM states / added latency	Extra states introduce *silent* ε-steps between observable actions; the partial order on Σ is unchanged, so E1–E5 are unaffected.
Shared-memory arrays with explicit synchronisation	When a memory is accessed by multiple processes, the *designer* explicitly coordinates access with a sync operation modeled as an S-action. Because these S-actions appear in both traces, the read/write order is fixed by E1–E2, port-level FIFO legality is preserved by E4, and the overall equivalence relation ≈ still holds.
Non-blocking message-passing (PushNB/PopNB) verified by post-HLS snooping	A verification wrapper delays the *pre-HLS* arrival of every nb-message until the identical event is observed on the *post-HLS* channel. This wrapper is itself purely synchronising (S), so it cannot violate R1–R5. Once the wrapper is assumed, τ_pre and τ_post coincide on all M-actions, so E3–E5 are preserved. This is not a transformation that inherently preserves ≈, but rather a verification technique to enforce equivalence for inherently latency-sensitive operations.
Latency-sensitive *global* signal I/O matched by snooping	The same “mirror-and-delay” wrapper used for NB channels is applied to any globally-visible signal whose timing matters. Because the wrapper inserts only S-actions, the partial-order constraints are unchanged. This is not a transformation that

Transformation permitted by the rules	Formal justification
	inherently preserves ≈, but rather a verification technique to enforce equivalence for inherently latency-sensitive operations.
Latency-sensitive *local* signal I/O isolated in cycle-accurate transactors	Local timing-exact protocols are confined to dedicated transactor processes that expose only latency-insensitive M or S operations to the rest of the design. Since the transactor boundary is now the observable interface, R1–R5 apply *outside* the timing-sensitive region, so system-level ≈ still holds.
One-way signal-handshake protocols with run-time assertions	For single-direction vld or rdy signals, an assertion proves that back-pressure can *never* occur. That assertion establishes a refinement in which the missing handshake is equivalent to a permanent logical '1', so the protocol is observationally identical to a two-way handshake that trivially satisfies E2.
Combinational processes	These have no S, R, W, or M actions, so their τ_P is the empty trace. The equivalence relation therefore holds vacuously.

4. Causal Dependency Between Processes

For the purpose of system-level equivalence, we must distinguish between incidental timing and functionally significant ordering.

Formal Definition of Causal Dependency

To formalize the notion of functionally significant ordering and resolve ambiguity, we define the happens-before relation, denoted by ->, on the set of all observable IO actions (Σ) across all processes in the system. This relation establishes a strict partial order of events, where a -> b is read as "a happens before b".

The *happens-before* relation is the smallest relation satisfying the three conditions below:

1. Intra-Process Order: If actions a and b occur within the same process P and a precedes b in the program's execution trace, then a -> b. This directly reflects the sequential nature of the code within a single thread.

2. Inter-Process Communication: The relation is established by the direct exchange of information or synchronization between processes.

    - Message Passing: If a is a blocking Push or non-blocking PushNB action on a channel in process P1 that sends a message, and b is the corresponding Pop or PopNB action in process P2 that receives that same message, then a -> b.

    - Signal Handshake: If a is a signal write action w(sig) in P1 and b is a signal read action r(sig) in P2 that reads the value written by a as part of an explicit handshake protocol, then a -> b. A complete two-way handshake (e.g., vld/rdy) creates a causal chain, such as w(vld)@P1 -> r(vld)@P2 -> w(rdy)@P2 -> r(rdy)@P1.

- Explicit Synchronization: If a set of actions {s_1, s_2, ... s_n} across different processes participates in a single, atomic synchronization event (e.g., a SyncChannel barrier), then the completion of that event for the group happens-before any subsequent, dependent action in any of the participating processes.

3. Transitivity: The relation is transitive. If a -> b and b -> c, then a -> c.

Using this relation, we now provide a formal definition for causal dependency between synchronization events, which are the fundamental ordering anchors in the scheduling model.

Definition: Causal Dependency
A synchronization event s2 in process P2 is causally dependent on a synchronization event s1 in process P1 if and only if s1 -> s2.
If neither s1 -> s2 nor s2 -> s1 holds true, the events s1 and s2 are considered causally independent (or concurrent). Any change in the observed temporal ordering of causally independent events between the pre-HLS and post-HLS simulations is considered an incidental, functionally insignificant variation in latency. A well-formed design, under this methodology, shall not rely on a specific ordering of causally independent events for functional correctness.

---

5. System-Level Theorem

Theorem (Compositional Equivalence).
For every process P in the design, let the HLS compiler apply only transformations listed in Section 3. Then
$\forall P : \tau\_pre,P \approx \tau\_post,P \quad \Rightarrow$ the entire system trace satisfies
$\tau\_pre,system \approx \tau\_post,system$.
Proof sketch. (Full proof is in Appendix J – System-Level Equivalence Proof)
Equivalence is defined per observable channel or signal.
By E1–E5 each process preserves:

1. ordering of synchronizing actions visible to other processes,
2. FIFO legality of every channel,
3. atomic association between signal IO and its bounding sync.
4. the *side-of-sync* predicate for all message actions (E5).

Because the composition of partial-order–preserving relationships is itself preserving, and because channels/signals are the only inter-process observables, system-level behavior is bisimilar under the relation ≈.

---

6. Practical Implication (“No Surprises” Guarantee)

If a verification environment exercises only the alphabet Σ (signals, message ports, synchronization calls) and does not test internal latency, then any testbench that passes on the pre-HLS model is *provably* guaranteed to pass on the post-HLS RTL, provided the design obeys the scheduling rules. This justifies single-testbench methodologies.

Appendix G thus establishes that the scheduling rules create a *trace-equivalence* relation between the high-level model and the synthesized RTL: every legal compiler optimization is a morphism of the labeled partial-order structure defined by R1–R5, ensuring functional indistinguishability at all observable interfaces.

Appendix H – Possible Criticisms

This section highlights several potential challenges in the verification approach of Appendix G.

1. Designer-Managed Shared-Memory Synchronization

Appendix G requires that any memory shared between processes use explicit synchronization primitives inserted by the designer. The HLS tool does not perform static verification of those primitives. In practice, we mitigate this risk by encapsulating shared memories within parameterized library classes (for example, see examples 12_ping_pong_mem and 15_native_ram_fifo) that are pre-verified for correct memory access coordination in both the pre-HLS and post-HLS models. Typically such classes will be parameterized for aspects such as element type and memory size.

2. Latency-Sensitive Signal Alignment ("Snooping")

Another possible concern about the overall verification argument in Appendix G relates to the reliance on snooping the RTL model to verify non-blocking message passing interfaces and latency-sensitive global signals. Note that this technique only needs to be used when the effects of the latency-sensitive behavior are visible externally to the DUT, even given the overall interaction between the testbench and DUT is latency insensitive. Note also that this document recommends that non-blocking message passing interfaces and latency-sensitive global signals are only used when no other approaches are possible.

Consider a DUT which is very complex, and which has an arbiter model embedded deep within the design hierarchy. This arbiter performs PopNB operations to determine which inputs have requests pending and then picks a winner. (This design is similar example 09_axi_fabric).

In the most general case, if the input delays to the arbiter are difficult to determine via analysis, the entire DUT RTL must be simulated alongside the pre-HLS model so that the RTL arbiter input delays can be used to delay the pre-HLS arbiter inputs.

Sometimes there are simpler cases where the input delays to the arbiter in the RTL DUT are easier to determine. For example, each input to the arbiter might arrive a fixed number of clock cycles after one of the primary inputs to the RTL DUT is pushed by the testbench. In this case, a much simpler model can be used to align the pre-HLS model with the post-HLS model. We can simply monitor the primary inputs to the pre-HLS model and then apply the fixed cycle delays to the pre-HLS arbiter inputs.

The two cases above illustrate two ends of a spectrum of possible approaches for extracting the delays from the RTL DUT. Between these two points there exist other possible approaches.

3. Matchlib Connections by Default Model a Skidbuffer inside In<> Ports

(Note that this overall document and specifically Appendix G are not intended to be strictly tied to Matchlib. Instead, this document is intended to be applicable to any pre-HLS model written in SystemC using signals, message passing channels, and synchronization calls.)

Matchlib Connections supports throughput accurate modeling in the pre-HLS simulation. Currently the throughput accurate modeling mechanism relies on the Pre() and the Post() methods using a 1-place buffer to store transactions that are in flight on Connections::In<> ports. This means that by default In<> ports function like a skidbuffer. Connections::Out<> ports do not introduce any additional storage capacity into the pre-HLS simulation.

By default, Catapult includes skidbuffers on input ports during HLS, so the formal equivalence relationship described in this document holds.

It is possible to remove some or all the skidbuffers during Catapult HLS. In this case, to make the formal equivalence relationship hold, the pre-HLS simulation must use the
`-DFORCE_AUTO_PORT=Connections::SYN_PORT`
compile flag. This will remove all the skidbuffers from the pre-HLS model. If some of the skidbuffers are still inserted during HLS, the formal equivalence relationship will still hold.

In this case the recommended methodology is to use both approaches:

- Use the default throughput accurate mode in Matchlib for most analysis and debug, and for pre-HLS performance verification.

- To enable the pre-HLS simulation to strictly conform to the formal model presented in Appendix G, then use -DFORCE_AUTO_PORT=Connections::SYN_PORT.

Common Formal Definitions for Appendices I-K

Symbol / Rule	Definition — concise but complete
Σ	Set of *observable* actions: • channel operations Push_c, Pop_c • synchronisation events Sync, Start/Finish pairs • single-cycle signal Write/Read actions.
ε-step	An *internal*, unobservable RTL state transition (pipeline advance, FSM micro-state, etc.).
B(c)	Compile-time *capacity* of channel *c* chosen by the HLS tool. Finite, B(c) **≥** 0. • B(c)=0 ⇒ rendezvous (capacity-zero) channel. • B(c)>0 ⇒ bounded FIFO.
occ_c(t)	*Occupancy* of channel *c* after cycle *t*: number of committed Push_c minus committed Pop_c.
P_push(c)	*Finite-progress invariant (producer)*: if Push_c is continuously enabled while occ_c(t)=B(c), then some future cycle t'>t commits a Pop_c.
P_pop(c)	*Finite-progress invariant (consumer)*: if Pop_c is continuously enabled while occ_c(t)=0, then some future cycle t'>t commits a Push_c.
Equivalence rules (E1–E5)	*E1 Clock correspondence*: pre- and post-HLS traces share the same global clock order. *E2 Signal visibility*: every Write/Read pair appears in the same relative order in both traces. *E3 No cross-channel inversions*: actions on distinct channels never overtake each other between traces. *E4 FIFO legality*: for each channel *c*, the post-HLS (Push_c, Pop_c) history is a legal

Symbol / Rule	Definition — concise but complete
	bounded-capacity execution consistent with the pre-HLS rendezvous history. *E5 Side-of-Sync guarantee*: a message action never moves across its surrounding synchronisation pair (Start/Finish, explicit Sync).

Pre-HLS Basic Process Semantics Rules (B1-B5)

B1 — Global-clock, step-wise model
All observable events occur at discrete global clock edges; intermediate ε-steps are silent and do not change any signal that could be sampled by another process.

B2 — Signal edge semantics
- Stability: signal values can only be written at the end of a cycle.
- Atomicity: all bits of a multi-signal transaction appear in that same edge.

B3 —Sync Isolation of Messages
A synchronization action commits only when all earlier message operations have completed, and no later message operation may start until it has committed.

B4 — Deterministic observable schedule
For a given input trace, the sequence and timing of *observable* actions (Σ) produced by a process is deterministic.

B5 — Zero-capacity rendezvous channels
A Push on a channel completes only when the matching Pop occurs in the same cycle; no data can be buffered in the channel itself.

Bridging Relationships between B rules and R rules

The B rules presented above guarantee the premise of the R rules. This table shows how the R rules can be derived from the B rules.

B rule – Common Definitions	Related R rule(s) – Appendix G	How B rules relate to R rules
B1 Global-clock, step-wise model	R1, R2, R3	One-clock completion of every Read/Write action lets R2 fix each signal access to a specific edge and lets R3 rely on "before/after-sync" timing, and anchors the global sync order required by R1.
B2 Signal edge semantics	R1, R2	Edge-aligned signal updates ensure every read/write can be anchored to a unique clock edge (R2) and that a sync call sees a stable signal value when it fires (supports R1).

B rule – Common Definitions	Related R rule(s) – Appendix G	How B rules relate to R rules
B3 Sync Isolation of Messages	R3	Rules are identical.
B4 Deterministic observable schedule	R1, R4	Determinism fixes per-process sync order (R1) and enforces repeatable ordering; it tolerates tie-cycles permitted by R4 (≤) so long as the tie is deterministic.
B5 Zero-capacity rendezvous channels	R5 Rendezvous capacity	Both state the same property: a Push commits only when the matching Pop commits in the same cycle. R5 restates B5 so later appendices can reference it directly.

Note on R4: R4 (non-inversion of independent message calls) does not introduce a brand-new semantic premise; rather, it leverages B1 – B4 to forbid a specific re-ordering that could deadlock bounded FIFOs.

The table shows that B1–B5 are sufficient conditions for R1–R5; the reverse implication has *not* been proven here. Therefore, Appendices I–K do *not* need to reference R rules directly. The R rules are *not* re-proved here, they enter the formalism only through B1–B5 via this table. R-rules are used only in Appendix G and the table above.

Appendix I – Per Process Trace Equivalence Proof

I.1 Overview

For any single pre-HLS process P that obeys the basic process semantics rules B1-B5 and communicates over channels of finite capacity $B(c) \geq 0$, every finite observable trace produced by P has a matching post-HLS RTL trace that satisfies the equivalence properties E1–E5.
The construction is explicit and does not rely on unbounded buffers; rendezvous channels ($B(c)=0$) are handled directly.

I.2 Formal Preconditions

1. Observable alphabet:
   Σ = { Push_c, Pop_c, Sync, Start / Finish, Write, Read }.

2. Silent step:
   Any internal RTL micro-state transition is written ε.

3. Finite-pipeline-depth premise (FPD):
   There exists a finite constant
   pipe_depth(P) = D_P < ∞
   such that between any two consecutive observable actions of P the scheduler inserts at most *D_P* clock cycles of ε-steps. (Commercial HLS tools emit finite scheduling tables, making this premise true in practice.)

4. Weak-fairness premise (WF):
   If a micro-operation remains continuously enabled, the scheduler eventually issues it. Clock-

synchronous RTL execution with either rendezvous channels or bounded FIFOs satisfies this property.

5. Finite-progress invariants:
   For every channel *c*, producer and consumer obligations *P_push(c)* and *P_pop(c)* hold, guaranteeing that data (or space) eventually becomes available.

I.3 Auxiliary Lemmas

Lemma I.1 (Bounded **ε**-chain).
Starting from any state of P, the next observable action appears after at most *D_P* clock cycles.
Proof. Direct from FPD. ■

Lemma I.2 (Eventual space / data).
Assume P is blocked on

- Push_c with *occ_c = B(c)*, or
- Pop_c with *occ_c = 0*.

Then a complementary Pop_c (respectively Push_c) commits within finite time.
Proof. In both situations the guard of the complementary action is continuously true.
By the invariants *P_push / P_pop* and weak fairness, that action must eventually fire, unblocking P.
Here weak fairness is interpreted system-wide (all enabled micro-ops in any process eventually fire). ■

I.4 Inductive Construction for Finite Traces

Let

$\tau\_pre = \tau\_pre[0..n-1] \circ e$ (where $|\tau\_pre| = n + 1$)

be the next pre-HLS prefix.
Assume by induction that we already have a matching post-HLS prefix $\tau\_post[0..n-1]$ satisfying E1–E5.
We extend it with a finite ε-chain (written ε*) followed by an observable action *e′* so that

$\tau\_post \circ \varepsilon^* \circ e'$ matches τ_pre.

Kind of event *e*	Why **ε*** is finite
Sync, Start, Finish, Write, Read	Scheduler may idle for ≤ *D_P* cycles (Lemma I.1).
Push_c	• If *B(c) > 0* and *occ_c < B(c)*: commit after ≤ *D_P* cycles. • If *B(c) > 0* and *occ_c = B(c)*: Lemma I.2 frees space. • If *B(c) = 0*: rendezvous fires when both sides are enabled; Lemma I.2 guarantees the peer.
Pop_c	Symmetric to Push_c.

Each case ensures ε* terminates, so the extension preserves E1–E5. ■

I.5 Extension to Infinite Traces

Because every finite prefix of a pre-HLS trace can be extended in bounded time and the construction is prefix-consistent, König's lemma on finitely branching trees yields an infinite post-HLS trace whose every finite prefix matches the corresponding pre-HLS prefix. Hence

$\tau\_pre \approx \tau\_post$ (for countably infinite traces as well).

Appendix J – System-Level Equivalence Proof

Theorem J.1 (Compositional Equivalence)
Let every process P satisfy τ_pre,P ≈ τ_post,P by Appendix I, and let every channel *c* have finite capacity B(c) ≥ 0 obeying P_push(c) & P_pop(c).
Then the aggregate traces of the whole design are equivalent: **τ**_pre,system **≈ τ**_post,system (rules E1–E5).
Proof Sketch

1. E1, E2 (clock & signals). Per-process preservation lifts directly to system scope.
2. E3 (no cross-channel inversions). An inversion would have to occur *inside* some channel *c*.
   If B(c)=0 the rendezvous forbids overtaking.
   If B(c)>0 the FIFO maintains order and P_push/Pop(c) ensures delivery, so overtaking cannot be observed.
3. E4 (FIFO legality). For each *c* the (Push_c, Pop_c) history in τ_post is a legal bounded-capacity schedule; τ_pre history (capacity 0) is trivially legal for any B(c) > 0.
4. E5 (side-of-sync). Appendix I ties every Push/Pop to its surrounding synchronisation, independent of capacity.

These partial-order preservations compose to form a bisimulation. **∴ τ**_pre,system **≈ τ**_post,system. □

Note: The concepts of causally dependent and causally independent synchronization relations were introduced in Appendix G. The Appendix J system level equivalence proof implicitly incorporates these concepts such that causally dependent synchronization is always preserved.

Appendix K – Proof of Liveness-Preservation

K.1 Scope and Notation

- System S is a collection of pre-HLS processes that obey rules B1–B5 and communicate over channels of finite capacity B(c) **≥** 0.
- Post-HLS RTL(S) is the hardware implementation produced by HLS.
- Observable events are Push_c, Pop_c, Sync, Start, Finish, Read and Write, exactly as in Appendix I.
- **ε**-step is any silent RTL micro-transition.
- A wait-for graph (WFG) at run-time has
    - nodes = blocked actions (a Push waiting for space, a Pop waiting for data, or any Sync/Wait operation),

- edges = “action A cannot proceed until action B occurs”.
  A cycle in the WFG implies a deadlock.

- Liveness (in this appendix) means absence of system-wide deadlock. Starvation of an individual process without a WFG cycle is not covered.

K.2 Premises

1. Finite-Pipeline-Depth (FPD) — already established in Appendix I (§ I.2-3).
2. Weak-Fairness (WF) — if an operation’s guard stays true, the scheduler eventually issues it.
3. Pre-HLS Liveness Assumption — the original system S is deadlock-free under WF.
4. Trace Equivalence Theorem — Appendix I constructs, for every finite pre-HLS trace, a matching post-HLS trace that preserves the order of observable events (E1–E5).

K.3 Lemma K.1 — Wait-For Graph Preservation

For any global state reached after a common observable prefix
$\tau\_pre \approx \tau\_post$
the set of blocked actions and the edge relation between them are isomorphic in the two models.

Proof Sketch

- A blocked Push_c (resp. Pop_c) in RTL must appear at the same program point as in the source model by E1 (program-order) and E3 (in-channel order).
- The guard (space or data) depends only on the channel’s occupancy count. Because every Push and Pop is reflected 1-for-1 in the other model (E4), the counts—and therefore the blocked/not-blocked status—agree.
- A Sync/Wait action is enabled by the same Boolean condition in both models; E2 states that observable control events are neither duplicated nor dropped.
- Hence nodes correspond exactly. Edges correspond because every dependency “A awaits B” is expressed by the same source-level condition; those conditions evaluate identically above.

K.4 Lemma K.2 — No New Cycles from Finite Buffering

Let C be any directed cycle of processes and channels in S.
Adding finite buffer capacity $B(c) > 0$ to any subset of the rendezvous channels on C cannot create a deadlock if the rendezvous (B=0) system was live.

Proof

We reason by contradiction.

1. Suppose the rendezvous system is live but the buffered system deadlocks.
2. At the deadlock each channel c on C has an occupancy $occ\_c \in [0, B(c)]$.
3. Advance time backward (conceptually drain buffers) by repeatedly applying the following monotone reduction until some channel hits occupancy 0:

If occ_c > 0 and its consumer is blocked, virtually deliver one message to that consumer.
Because capacities are finite, the reduction terminates.

4. The resulting state is a deadlock of the rendezvous system: every Push/Pop on C is still blocked, but now each channel has occ_c = 0.
5. This contradicts the assumed liveness of the rendezvous system.

K.5 Theorem K.1 — Liveness Preservation
If the pre-HLS system S is deadlock-free under Weak-Fairness, then RTL(S) generated by HLS is also deadlock-free under the same fairness assumption.
Proof
Assume, toward contradiction, that RTL(S) deadlocks after some finite observable prefix τ_post.

1. By Appendix I there exists a pre-HLS trace τ_pre that matches τ_post.
2. At the deadlock, construct the WFG G_post. By Lemma K.1 there is an isomorphic WFG G_pre for the source model. In particular, G_pre contains a cycle.
3. If all channels in that cycle have B(c)=0, the rendezvous system is deadlocked—contradicting the pre-HLS liveness assumption.
4. Otherwise, pick any channel on the cycle with B(c)>0. Replace every buffered channel on the cycle by capacity 0 and apply Lemma K.2 to obtain a rendezvous-cycle deadlock, again contradicting liveness.
   Therefore the assumed RTL deadlock cannot occur.

Implementation Note

Because the Appendix I, J, and K proofs above are parametric in B(c) ≥ 0, designers are free to instantiate *any* finite depth—zero included—on every channel without jeopardising functional correctness or liveness, provided the environment upholds the weak fairness assumption.