## Contents

1. Intel® HLS Compiler Pro Edition Best Practices Guide ................................................................. 4  
2. Best Practices for Coding and Compiling Your Component ............................................................ 5  
3. FPGA Concepts .................................................................................................................................. 7  
   3.1. FPGA Architecture Overview ........................................................................................................ 7  
      3.1.1. Adaptive Logic Module (ALM) .................................................................................................. 8  
      3.1.2. Digital Signal Processing (DSP) Block .................................................................................... 10  
      3.1.3. Random-Access Memory (RAM) Blocks ................................................................................. 10  
   3.2. Concepts of FPGA Hardware Design ........................................................................................... 10  
      3.2.1. Maximum Frequency ($f_{MAX}$) ............................................................................................. 10  
      3.2.2. Latency .................................................................................................................................. 11  
      3.2.3. Pipelining .............................................................................................................................. 11  
      3.2.4. Throughput ........................................................................................................................... 12  
      3.2.5. Datapath ............................................................................................................................... 12  
      3.2.6. Control Path ......................................................................................................................... 12  
      3.2.7. Occupancy ........................................................................................................................... 13  
   3.3. Methods of Hardware Design ...................................................................................................... 14  
      3.3.1. How Source Code Becomes a Custom Hardware Datapath ...................................................... 14  
      3.3.2. Scheduling ............................................................................................................................ 17  
      3.3.3. Mapping Parallelism Models to FPGA Hardware ................................................................... 23  
      3.3.4. Memory Types ...................................................................................................................... 32  
4. Interface Best Practices ...................................................................................................................... 39  
   4.1. Choose the Right Interface for Your Component ......................................................................... 40  
      4.1.1. Pointer Interfaces .................................................................................................................. 41  
      4.1.2. Avalon Memory Mapped Host Interfaces ............................................................................. 43  
      4.1.3. Avalon Memory Mapped Agent Memories ............................................................................ 46  
      4.1.4. Avalon Memory Mapped Agent Registers ............................................................................ 48  
      4.1.5. Avalon Streaming Interfaces .................................................................................................. 48  
      4.1.6. Pass-by-Value Interface ......................................................................................................... 50  
   4.2. Control LSUs For Your Variable-Latency MM Host Interfaces ..................................................... 53  
   4.3. Avoid Pointer Aliasing .................................................................................................................. 54  
5. Loop Best Practices ............................................................................................................................. 56  
   5.1. Reuse Hardware By Calling It In a Loop ....................................................................................... 57  
   5.2. Parallelize Loops .......................................................................................................................... 58  
      5.2.1. Pipeline Loops ........................................................................................................................ 59  
      5.2.2. Unroll Loops .......................................................................................................................... 60  
      5.2.3. Example: Loop Pipelining and Unrolling .............................................................................. 61  
   5.3. Construct Well-Formed Loops ...................................................................................................... 63  
   5.4. Minimize Loop-Carried Dependencies ......................................................................................... 64  
   5.5. Avoid Complex Loop-Exit Conditions ....................................................................................... 65  
   5.6. Convert Nested Loops into a Single Loop .................................................................................... 66  
   5.7. Place if-Statements in the Lowest Possible Scope in a Loop Nest .................................................. 67  
   5.8. Declare Variables in the Deepest Scope Possible ....................................................................... 67  
   5.9. Raise Loop II to Increase $f_{MAX}$ .............................................................................................. 68  
   5.10. Control Loop Interleaving .......................................................................................................... 68

The Intel® HLS Compiler Pro Edition Best Practices Guide provides techniques and practices that you can apply to improve the FPGA area utilization and performance of your HLS component. Typically, you apply these best practices after you verify the functional correctness of your component.

In this publication, `<quartus_installdir>` refers to the location where you installed Intel Quartus® Prime Design Suite.

The default Intel Quartus Prime Design Suite installation location depends on your operating system:

- **Windows**: `C:\intelFPGA_pro\21.2`
- **Linux**: `/home/<username>/intelFPGA_pro/21.2`

About the Intel HLS Compiler Pro Edition Documentation Library

Documentation for the Intel HLS Compiler Pro Edition is split across a few publications. Use the following table to find the publication that contains the Intel HLS Compiler Pro Edition information that you are looking for:

<table>
<thead>
<tr>
<th>Table 1. Intel High Level Synthesis Compiler Pro Edition Documentation Library</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title and Description</strong></td>
</tr>
<tr>
<td><strong>Release Notes</strong></td>
</tr>
<tr>
<td>Provide late-breaking information about the Intel HLS Compiler.</td>
</tr>
<tr>
<td><strong>Getting Started Guide</strong></td>
</tr>
<tr>
<td>Get up and running with the Intel HLS Compiler by learning how to initialize your compiler environment and reviewing the various design examples and tutorials provided with the Intel HLS Compiler.</td>
</tr>
<tr>
<td><strong>User Guide</strong></td>
</tr>
<tr>
<td>Provides instructions on synthesizing, verifying, and simulating intellectual property (IP) that you design for Intel FPGA products. Go through the entire development flow of your component from creating your component and testbench up to integrating your component IP into a larger system with the Intel Quartus Prime software.</td>
</tr>
<tr>
<td><strong>Reference Manual</strong></td>
</tr>
<tr>
<td>Provides reference information about the features supported by the Intel HLS Compiler. Find details on Intel HLS Compiler command options, header files, pragmas, attributes, macros, declarations, arguments, and template libraries.</td>
</tr>
<tr>
<td><strong>Best Practices Guide</strong></td>
</tr>
<tr>
<td>Provides techniques and practices that you can apply to improve the FPGA area utilization and performance of your HLS component. Typically, you apply these best practices after you verify the functional correctness of your component.</td>
</tr>
<tr>
<td><strong>Quick Reference</strong></td>
</tr>
<tr>
<td>Provides a brief summary of Intel HLS Compiler declarations and attributes on a single two-sided page.</td>
</tr>
</tbody>
</table>
2. Best Practices for Coding and Compiling Your Component

After you verify the functional correctness of your component, you might want to improve the performance and FPGA area utilization of your component. Learn about the best practices for coding and compiling your component so that you can determine which practices can help you optimize your component.

As you optimize your component, apply the best practices techniques in the following areas, roughly in the order listed. Also, review the example designs and tutorials provided with the Intel High Level Synthesis (HLS) Compiler to see how some of these techniques can be implemented.

- **Understand FPGA Concepts** on page 7
  
  A key best practice to help you get the most out of the Intel HLS Compiler is to understand important concepts about FPGAs. With an understanding of FPGA architecture, and some FPGA hardware design concepts and methods, you can create better designs that take advantage your target FPGA devices.

- **Interface Best Practices** on page 39
  
  With the Intel High Level Synthesis Compiler, your component can have a variety of interfaces: from basic wires to the Avalon Streaming and Avalon Memory-Mapped Host interfaces. Review the interface best practices to help you choose and configure the right interface for your component.

- **Loop Best Practices** on page 56
  
  The Intel High Level Synthesis Compiler pipelines your loops to enhance throughput. Review these loop best practices to learn techniques to optimize your loops to boost the performance of your component.

- **Memory Architecture Best Practices** on page 72
  
  The Intel High Level Synthesis Compiler infers efficient memory architectures (like memory width, number of banks and ports) in a component by adapting the architecture to the memory access patterns of your component. Review the memory architecture best practices to learn how you can get the best memory architecture for your component from the compiler.

- **System of Tasks Best Practices** on page 87
  
  Using a system of HLS tasks in your component enables a variety of design structures that you can implement including executing multiple loops in parallel and sharing an expensive compute block.

- **Datatype Best Practices** on page 94
  
  The datatypes in your component and possible conversions or casting that they might undergo can significantly affect the performance and FPGA area usage of your component. Review the datatype best practices for tips and guidance how best to control datatype sizes and conversions in your component.
• Alternative Algorithms

The Intel High Level Synthesis Compiler lets you compile a component quickly to get initial insights into the performance and area utilization of your component. Take advantage of this speed to try larger algorithm changes to see how those changes affect your component performance.
3. FPGA Concepts

A key best practice to help you get the most out of the Intel HLS Compiler is to understand important concepts about FPGAs. With an understanding of FPGA architecture, and some FPGA hardware design concepts and methods, you can create better designs that take advantage your target FPGA devices.

3.1. FPGA Architecture Overview

A field-programmable gate array (FPGA) is a reconfigurable semiconductor integrated circuit (IC).

FPGAs occupy a unique computational niche relative to other compute devices, such as central and graphics processing units (CPUs and GPUs), and custom accelerators, such as application-specific integrated circuits (ASICs). CPUs and GPUs have a fixed hardware structure to which a program maps, while ASICs and FPGAs can build custom hardware to implement a program.

While a custom ASIC generally outperforms an FPGA on a specific task, ASICs take significant time and money to develop. FPGAs are a cheaper off-the-shelf alternative that you can reprogram for each new application.

An FPGA is made up of a grid of configurable logic, known as adaptive logic modules (ALMs), and specialized blocks, such as digital signal processing (DSP) blocks and random-access memory (RAM) blocks. These programmable blocks are combined using configurable routing interconnects to implement complete digital circuits.

The total number of ALMs, DSP blocks, and RAM blocks used by a design is often referred to as the FPGA area or area that the design uses.
The following image illustrates a high-level architectural view of an FPGA:

3.1.1. Adaptive Logic Module (ALM)

The basic building block in an Intel FPGA device is an adaptive logic module (ALM). A simplified ALM consists of a lookup table (LUT) and an output register from which the compiler can build any arbitrary Boolean logic circuit.
The following figure illustrates a simplified ALM:

### 3.1.1.1. Lookup Table (LUT)

A *lookup table (LUT)* that implements an arbitrary Boolean function of N inputs is often referred to as an *N-LUT*.

### 3.1.1.2. Register

A *register* is the most basic storage element in an FPGA. It has an input (*in*), an output (*out*), and a clock signal (*clk*). It is synchronous, that is, it synchronizes output changes to a clock. In an ALM, a register may store the output of the LUT.

The following figure illustrates a register:

Note: The clock signal is implied and not shown in some figures.
The input data propagates to the output on every clock cycle. The output remains unchanged between clock cycles.

### 3.1.2. Digital Signal Processing (DSP) Block

A digital signal processing (DSP) block implements specific support for common fixed-point and floating-point arithmetic, which reduces the need to build equivalent logic from general-purpose ALMs.

The following figure illustrates a simplified three-input DSP block consisting of a multiplier (×) and an adder (+):

The following figure illustrates a simplified DSP block:

![DSP Block Diagram](image)

### 3.1.3. Random-Access Memory (RAM) Blocks

A random-access memory (RAM) block implements memory by using a high density of memory cells.

For more information, refer to Memory Types.

### 3.2. Concepts of FPGA Hardware Design

#### 3.2.1. Maximum Frequency (f\text{MAX})

The maximum clock frequency at which a digital circuit can operate is called its f\text{MAX}. The f\text{MAX} is the maximum rate at which the outputs of registers are updated.

The physical propagation delay of the signal across Boolean logic between two consecutive register stages limits the clock speed. This propagation delay is a function of the complexity of the combinational logic in the path.

The path with the most combinational logic elements (and the highest delay) limits the speed of the entire circuit. This speed limiting path is often referred to as the critical path.

The f\text{MAX} is calculated as the inverse of the critical path delay. You may want to have high f\text{MAX} since it results in high performance in the absence of other bottlenecks.
3.2.2. Latency

*Latency* is the measure of how long it takes to complete one or more operations in a digital circuit. You can measure latency at different granularities. For example, you can measure the latency of a single operation or the latency of the entire circuit.

You can measure latency in time (for example, microseconds) or in clock cycles. Typically, clock cycles are the preferred way to express latency because measuring latency in clock cycles disconnects latency from your circuit clock frequency. By expressing latency independent of circuit clock frequency, it is easier to discern the true impact of circuit changes to the performance of the circuit.

You may want to have low latency, but lowering latency might result in decreased $f_{\text{MAX}}$.

For more information and an example, refer to Pipelining.

3.2.3. Pipelining

*Pipelining* is a design technique used in synchronous digital circuits to increase $f_{\text{MAX}}$. Pipelining involves adding registers to the critical path, which decreases the amount of logic between each register. Less logic takes less time to execute, which enables an increase in $f_{\text{MAX}}$.

The critical path in a circuit is the path between any two consecutive registers with the highest latency. That is, the path between two consecutive registers where the operations take the longest to complete.

Pipelining is especially useful when processing a stream of data. A pipelined circuit can have different stages of the pipeline operating on different input stream data in the same clock cycle, which leads to better data processing throughput.

**Example**

Consider a simple circuit with operations A and B on the critical path. If operation A takes 5 ns to complete and operation B takes 15 ns to complete, then the time delay on the critical path is 20 ns. This results in an $f_{\text{MAX}}$ of 50 MHz ($1/\text{max\_delay}$).

**Figure 1.** Unpipelined Logic Block with an $f_{\text{MAX}}$ of 50 MHz and Latency of Two Clock Cycles

If a pipeline register is added between A and B, the critical path changes. The delay on the critical path is now 15 ns. Pipelining this block results in an $f_{\text{MAX}}$ of 66.67 MHz, and the maximum delay between two consecutive registers is 15 ns.
Figure 2. Pipelined Logic Block with an $f_{\text{MAX}}$ of 66.67 MHz and Latency of Three clock cycles

While pipelining generally results in a higher $f_{\text{MAX}}$, it increases latency. In the previous example, the latency of the block containing A and B increases from two to three clock cycles after pipelining.

Related Information
Pipeline Loops on page 59

3.2.4. Throughput

Throughput of a digital circuit is the rate at which data is processed.

In the absence of other bottlenecks, higher $f_{\text{MAX}}$ results in higher throughput (for example, samples/second).

Throughput is a common measure of the performance of a circuit, and throughput and performance are often used interchangeably when discussing a circuit.

3.2.5. Datapath

A datapath is a chain of registers and Boolean logic in a digital circuit that performs computations.

For example, the datapath in Pipelining on page 11 consists of all of the elements shown, from the input register to the last output register.

Memory blocks are outside the datapath and reads and writes to memory are also considered to be outside of the datapath.

3.2.6. Control Path

While the datapath is the path on which computations occur, the control path is the path of logic that the compiler adds to manage the flow of data through the datapath.

Control paths include controls such as the following:

- Handshaking flow control
  Handshaking ensures that one part of your design is ready and able to accept data from another part of your design.
• Loop control
  Loop controls control the flow of data through the hardware generated for loops in your code, including any loop carried dependencies.

• Branch control
  Branch controls implement conditional statements in your code. Branch control can include parallelizing parts of conditional statements to improve performance.

The control path also consumes FPGA area, and the compiler uses techniques like clustering the datapath to help reduce the control path and save area. To learn about clustering, refer to Clustering the Datapath on page 18.

### 3.2.7. Occupancy

The **occupancy** of a datapath at a point in time refers to the proportion of the datapath that contains valid data.

The occupancy of a circuit over the execution of a program is the average occupancy over time from the moment the program starts to run until it has completed.

Unoccupied portions of the datapath are often referred to as **bubbles**. Bubbles are analogous to a "no operation" (no-op) instructions for a CPU that have no effect on the final output.

Decreasing bubbles increases occupancy. In the absence of other bottlenecks, maximizing occupancy of the datapath results in higher throughput.
3.3. Methods of Hardware Design

Traditionally, you program an FPGA using a hardware description language (HDL) such as Verilog or VHDL. However, a recent trend is to use higher-level languages. Higher levels of abstraction can reduce the design time and increase the portability of your design.

The sections that follow discuss how Intel HLS Compiler maps high-level languages to a hardware datapath.

3.3.1. How Source Code Becomes a Custom Hardware Datapath

Based on your source code and the following principles, the Intel HLS Compiler builds a custom hardware datapath in the FPGA logic:

- The datapath is functionally equivalent to the C++ program described by the source. Once the datapath is constructed, the compiler orchestrates work items and loop iterations such that the hardware is effectively occupied.
- The compiler builds the custom hardware datapath while minimizing area of the FPGA resources (like ALMs and DSPs) used by the design.
3.3.1.1. Mapping Source Code Instructions to Hardware

For fixed architectures, such as CPUs and GPUs, a compiler compiles code into a set of instructions that run on functional units that have a fixed functionality. For these fixed architectures to be useful in a broad range of applications, some of their available functional units are not useful to every program. Unused functional units mean that your program does not fully occupy the fixed architecture hardware.

FPGAs are not subject to these restrictions of fixed functional units. On an FPGA, you can synthesize a specialized hardware datapath that can be fully occupied for an arbitrary set of instructions, which means you can be more efficient with the silicon area of your chip.

By implementing your algorithm in hardware, you can fill your chip with custom hardware that is always (or almost always) working on your problem instead of having idle functional units.

The Intel HLS Compiler maps statements from the source code to individual specialized hardware operations, as shown in the example in the following image:

\[ c = a + b; \]

In general, each instruction maps to its own unique instance of a hardware operation. However, a single statement can map to more than one hardware operation, or multiple statements can combine into a single hardware operation when the compiler finds that it can generate hardware that is more efficient.

The latency of hardware operations is dependent on the complexity of the operation and the target \( f_{\text{MAX}} \).

The compiler takes these hardware operations and connects them into a graph based on their dependencies. When operations are independent, the compiler automatically infers parallelism by executing those operations simultaneously in time.

The following figure shows a dependency graph created for the hardware datapath. The dependency graph shows how the instruction is mapped to hardware operations and how the hardware operations are connected based on their dependencies. The loads in this example instruction are independent of each other and can therefore run simultaneously.
3.3.1.2. Mapping Arrays and Their Accesses to Hardware

Similar to the mapping of statements to specialized hardware operations, the compiler can map arrays (and structs) to hardware memories based on memory access patterns and variable sizes.

The datapath interacts with this memory through load/store units (LSUs), which are inferred from array accesses in the source code.

The following figure illustrates a simple example of mapping arrays and their accesses to hardware:

A RAM can have a limited number of read ports and write ports, but a datapath can have many LSUs. When the number of LSUs does not match the available number of read and write ports, the compiler uses techniques like replication, double pumping, sharing, and arbitration. For descriptions of these techniques, refer to Component Memory on page 32.

For more details about configuring memories, refer the following topics:

- Memory Architecture Best Practices on page 72
FPGAs provide specialized hardware block RAMs that you can configure and combine to match the size of your arrays. Customizing your memory configuration for your design can provide terabytes-per-second of on-chip memory bandwidth because each of these memories can interact with the datapath simultaneously.

Arrays might also be implemented in your component datapath. In this case, the array contents are stored as registers in the datapath when your algorithm is pipelined (as discussed in Pipelining on page 25). Storing array contents as registers in the datapath can improve performance in some cases, but it is a design decision whether to implement an array as registers or as memories.

When you access an array that is implemented as registers, LSUs are not used. The compiler might choose to use a select or a barrel shifter instead.

3.3.2. Scheduling

Scheduling refers to the process of determining the clock cycles at which each operation in the datapath executes.

Pipelining is the outcome of scheduling.

Related Information

Pipelining on page 11

3.3.2.1. Dynamic Scheduling

The Intel HLS Compiler generates pipelined datapaths that are dynamically scheduled.

A dynamically scheduled portion of the datapath does not pass data to its successor until its successor signals that it is ready to receive it.

This signaling is accomplished using handshaking control logic. For example, a variable latency load from memory may refuse to accept its predecessors’ data until the load is complete.

Handshaking helps remove bubbles in the pipeline, which increases occupancy. For more information about bubbles, refer to Occupancy.
The following figure illustrates four regions of dynamically scheduled logic:

**Figure 4. Dynamically Scheduled Logic**

Black arrows represent data and valid signals and red arrows represent signals to stall incoming valid data flow.

### 3.3.2.2. Clustering the Datapath

Dynamically scheduling all operations adds overhead in the form of additional FPGA area needed to implement the required handshaking control logic.

To reduce this overhead, the compiler groups fixed latency operations into *clusters*. A cluster of fixed latency operations, such as arithmetic operations, needs fewer handshaking interfaces, thereby reducing the area overhead.
If A, B, and C from Figure 4 on page 18 do not contain variable latency operations, the compiler can cluster them together, as illustrated in Figure 5 on page 19.

Clustering the logic reduces area by removing the need for signals to stall data flow in addition to other handshaking logic within the cluster.
Cluster Types

The Intel HLS Compiler can create the following types of clusters:

- **Stall-Enable Cluster (SEC):** This cluster type passes the handshaking logic to every pipeline stage in the cluster in parallel. If the cluster is stalled by logic from further down in the datapath, all logic in the SEC stalls at the same time.

![Stall-Enable Cluster Diagram]

- **Stall-Free Cluster (SFC):** This cluster type adds a first in, first out (FIFO) buffer to the end of the cluster that can accommodate at least the entire latency of the pipeline in the cluster. This FIFO is often called an exit FIFO because it is attached to the exit of the cluster datapath.

Because of this FIFO, the pipeline stages in the cluster do not require any handshaking logic. The stages can run freely and drain into the capacity FIFO, even if the cluster is stalled from logic further down in the datapath.
Cluster Characteristics

The exit FIFO of the stall free cluster results in some tradeoffs:

- **Area**: Because an SEC does not use an exit FIFO, it can save FPGA area compared to an SFC.

  If you have a design with many small, low-latency clusters, you can save a substantial amount of area by asking the compiler to use SECs instead of SFCs with the `hls_use_stall_enable_clusters` component attribute. For details, refer to the `hls_use_stall_enable_clusters Component Attribute` in the Intel HLS Compiler Reference Manual.

- **Latency**: Logic that uses SFCs might have a larger latency than logic that uses SECs because of the write-read latency of the exit FIFO.

- **fMAX**: In an SFC, the `oStall` signal has less fanout than in an SEC.

  For a cluster with many pipeline stages, you can improve your design `fMAX` by using an SFC.
• **Handshaking**: The exit FIFO in SFCs allow them to take advantage of hyper-optimized handshaking between clusters. For more information, refer to Hyper Optimized Handshaking.

SECs do not support this capability.

• **Bubble Handling**: SECs remove only leading bubbles. A leading bubble is a bubble that arrives before the first piece of valid data arrives in the cluster. SECs do not remove any arriving afterwards.

SFCs can use the exit FIFO to remove all bubbles from the pipeline if the SFC gets a downstream stall signal.

• **Stall Behavior**: When an SEC receives a downstream stall, it stalls any logic upstream of it within one clock cycle.

When an SFC receives a downstream stall, the exit FIFO allows it to consume additional valid data depending on how deep the exit FIFO is and how many bubbles are in the cluster datapath.

### 3.3.2.3. Handshaking Between Clusters

By default, the handshaking protocol between clusters is a simple stall/valid protocol. Data from the upstream cluster is consumed when the stall signal is low and the valid signal is high.

**Figure 8. Handshaking Between Clusters**

![Handshaking Between Clusters Diagram](image_url)

**Hyper-Optimized Handshaking**

If the distance across the FPGA between these two clusters is large, handshaking may become the critical path that affects peak $f_{MAX}$. In the design

To improve these cases, the Intel HLS Compiler can add pipelining registers to the stall/valid protocol to ease the critical path and improve $f_{MAX}$. This enhanced handshaking protocol is called *hyper-optimized handshaking*. 


Hyper-Optimized Handshaking Data Flow

Cluster 1

- iValid
- oStall
- oValid
- iStall

Cluster 2

- iValid
- oStall
- oValid
- iStall

Valid data will not stop flowing until 4 cycles after Cluster 2 asserts \texttt{stall}.

The following timing diagram illustrates an example of upstream cluster 1 and downstream cluster 2 with two pipelining registers inserted in-between:

Hyper-Optimized Handshaking

Figure 10.

Restriction: Hyper-optimized handshaking is currently available only for the Intel Agilex™ and Intel Stratix® 10 device families.

3.3.3. Mapping Parallelism Models to FPGA Hardware

This section describes how to map parallelism models to FPGA hardware:

3.3.3.1. Data Parallelism

Traditional instruction-set-architecture-based (ISA-based) accelerators, such as GPUs, derive data parallelism from vectorized instructions and by executing the same operation on multiple processing units.

In comparison, FPGAs derive their performance by taking advantage of their spatial architecture. FPGA compilers do not require you to vectorize your code. The compiler vectorizes your code automatically whenever it can.

The generated hardware implements data parallelism in the following ways:
3.3.3.1.1. Executing Independent Operations Simultaneously

As described in Mapping Source Code Instructions to Hardware, the compiler can automatically identify independent operations and execute them simultaneously in hardware.

This simultaneous execution of independent operations combined with pipelining is how performance through data parallelism is achieved on an FPGA.

The following image illustrates an example of an adder and a multiplier, which are scheduled to execute simultaneously while operating on separate inputs:

![Figure 11. Automatic Vectorization in the Generated Hardware Datapath](image)

This automatic vectorization is analogous to how a superscalar processor takes advantage of instruction-level parallelism, but this vectorization happens statically at compile time instead of dynamically, at runtime.

Because determining instruction-level parallelism occurs at compile time, there is no hardware or runtime cost of dependency checking for the generated hardware datapath. Additionally, the flexible logic and routing of an FPGA means that only the available resources (like ALMs and DSPs) of the FPGA restrict the number of independent operations that can occur simultaneously.

**Unrolling Loops**

You can unroll loops in the design by using loop attributes. Loop unrolling decreases the number of iterations executed at the expense of increasing hardware resource consumption corresponding to executing multiple iterations of the loop simultaneously.

Once unrolled, the hardware resources are scheduled as described in Scheduling.

The Intel HLS Compiler never attempts to unroll any loops in your source code automatically. You must always control loop unrolling by using the corresponding pragma. For details, refer to *Loop Unrolling* (*unroll Pragma*) in the Intel High Level Synthesis Compiler Reference Manual.
**Conditional Statements**

The Intel HLS Compiler attempts to eliminate conditional or branch statements as much as possible.

Conditionally executed code becomes predicated in the hardware. Predication increases the possibilities for executing operations simultaneously and achieving better performance. Additionally, removing branches allows the compiler to apply other optimizations to the design.

**Figure 12. Conditional Statements**

In this example, the function foo can be run unconditionally. The code that cannot be run unconditionally, like the memory assignments, retain a condition.

```c
int result;
if (x > 0)
    result = foo (a);
else
    result = foo (b);
```

```c
int result;
/* foo(a) and foo(b) both execute */
int ra = foo(a);
int rb = foo(b);
/* result is multiplexed */
bool cond (x > 0);
if (cond) result = ra;
if (!cond) result = rb;
```

**Related Information**

Unroll Loops on page 60

**3.3.3.1.2. Pipelining**

Similar to the implementation of a CPU with multiple pipeline stages, the compiler generates a deeply-pipelined hardware datapath. For more information, refer to Concepts of FPGA Hardware Design and How Source Code Becomes a Custom Hardware Datapath.

Pipelining allows for many data items to be processed concurrently (in the same clock cycle) while making efficient use of the hardware in the datapath by keeping it occupied.
Pipelining and Vectorizing a Pipelined Datapath

Consider the following example of code mapping to hardware:

**Figure 13. Example Code Mapping to Hardware**

```
Mem[100] += 42 * Mem[101];
```

Multiple invocations of this code when running on a CPU would not be pipelined. The output of an invocation is completed before inputs are passed to the next invocation of the code.

On an FPGA device, this kind of unpipelined invocation results in poor throughput and low occupancy of the datapath because many of the operations are sitting idle while other parts of the datapath are operating on the data. The following figure shows what throughput and occupancy of invocations looks like in this scenario:

**Figure 14. Unpipelined Execution Resulting in Low Throughput and Low Occupancy**
The Intel HLS Compiler pipelines your design as much as possible. New inputs can be sent into the datapath each cycle, giving you a fully occupied datapath for higher throughput, as shown in the following figure:

**Figure 15. Pipelining the Datapath Results in High Throughput and High Occupancy**

You can gain even further throughput by vectorizing the pipelined hardware. Vectorizing the hardware improves throughput, but requires more FPGA area for the additional copies of the pipelined hardware:
Understanding where the data you need to pipeline is coming from is key to achieving high performance designs on the FPGA. You can use the following sources of data to take advantage of pipelining:

- Components
- Loop iterations

**Pipelining Loops Within A Component**

Within a component, loops are the primary source of pipeline parallelism.

When the Intel HLS Compiler pipelines a loop, it attempts to schedule the loop execution such that the next iteration of the loop enters the pipeline before the previous iteration has completed. This pipelining of loop iterations can lead to higher throughput.

The number of clock cycles between iterations of the loop is called the *Initiation Interval* (II).

For the highest performance, a loop iteration would start every clock cycle, which corresponds to an II of 1.
Data dependencies that are carried from one loop iteration to another can affect the ability to achieve II of 1. These dependencies are called *loop-carried dependencies*.

The II of a loop must be high enough to accommodate all loop carried dependencies.

*Tip:* The II required to satisfy this constraint is a function of the $f_{\text{MAX}}$ of the design. If the $f_{\text{MAX}}$ is lower, the II might also be lower. Conversely, if the $f_{\text{MAX}}$ is higher, a higher II might be required.

The Intel HLS Compiler automatically identifies these dependencies and tries to build hardware to resolve them while minimizing the II, subject to the target $f_{\text{MAX}}$.

Naively generating hardware for the code in Figure 17 on page 29 results in two loads: one from memory $b$ and one from memory $c$. Because the compiler knows that the access to $c[i-1]$ was written to in the previous iteration, the load from $c[i-1]$ can be optimized away.

**Figure 17. Pipelining a Datapath with Loop Iteration**

```c
for (int i=1; i < n; i++)
    c[i] = c[i-1] + b[i];
```

The dependency on the value stored to $c$ in the previous iteration is resolved in a single clock cycle, so an II of 1 is achieved for the loop even though the iterations are not independent.

For additional information about pipelining loops, refer to *Pipeline Loops* on page 59.

When the Intel HLS Compiler cannot initially achieve II of 1, it chooses from several optimization strategies:

- **Interleaving** on page 29
- **Speculative Execution** on page 30

These optimizations are applied automatically by the Intel HLS Compiler, and additionally can be controlled through pragma statements in the design.
When a loop nest has an inner loop II that is greater than 1, the Intel HLS Compiler can attempt to interleave iterations of the outer loop into iterations of the inner loop to better utilize the hardware resources and achieve higher throughput.

**Figure 18. Interleaving**

```
// >Column-wise partial reduction
for (int i=0; i < N; i++) {
    for (int j=0; j < M; j++) {
        b[i] += foo( a[i][j] );
    }
}
```

For additional information about controlling interleaving in your component, refer to **Loop Interleaving Control (max_interleavingPragma)** in the **Intel High Level Synthesis Compiler Reference Manual**.

**Speculative Execution**

When the critical path that affects II is the computation of the exit condition and not a loop-carried dependency, the Intel HLS Compiler can attempt to relax this scheduling constraint by speculatively continuing to execute iterations of the loop while the exit condition is being computed.

If it is determined that the exit condition is satisfied, the effects of these extra iterations are suppressed.

Typically, the exit condition for a loop iteration must be evaluated before the program determines whether to start the next loop iteration or continue into the rest of the function. This requirement means that the loop initiation interval (II) cannot be lower than the number of cycles required to compute the exit condition. Speculated iterations can help lower the loop II because operations within the loop can occur in the function pipeline at the same time as the exit condition is evaluated.

This speculative execution can achieve lower II and higher throughput, but it can incur additional overhead between loop invocations (equivalent to the number of speculated iterations). A larger loop trip count helps to minimize this overhead.

For any speculated iteration, instructions with side effects outside of the loop (like writing to memory or a stream) are not completed until the loop exit condition for the iteration has been evaluated. For loop iterations that are in flight but incomplete when the loop exit condition is met, side effect data is discarded.
While speculated iterations can improve loop II, they occupy the pipeline until they are completed. A new loop invocation cannot start until all of the speculated iterations have completed. For example, the next iteration of an outer loop cannot start until all the speculated iterations of an inner loop have completed.

**Terminology**
A loop invocation is what starts a series of loop iterations. One loop iteration is one execution of the body of a loop.

**Figure 19. Loop Orchestration Without Speculative Execution**

**Figure 20. Loop Orchestration With Speculative Execution**

**Related Information**
"Loop Iteration Speculation (speculated_iterationsPragma)" in the *Intel High Level Synthesis Compiler Reference Manual*
Pipelining Across Component Invocations

The pipelining of work across component invocations is similar to how loops are pipelined.

The Intel HLS Compiler attempts to schedule the execution of component invocations such that the next invocation of a component enters the pipeline before the previous invocation has completed.

3.3.3.2. Task Parallelism

The compiler achieves concurrency by scheduling independent individual operations to execute simultaneously, but it does not achieve concurrency at coarser granularities (for example, across loops).

For larger code structures to execute in parallel with each other, you must write them as separate components or tasks that launch simultaneously. These components or tasks then run independently, and synchronize and communicate using pipes or streams, as shown in the following figure:

Figure 21. Multiple Task Functions Running Asynchronously

```
for(i=0..N) {
  ...
  mypipe::write(x);
  ...
}
```

```
for(i=0..N) {
  ...
  y = mypipe::read();
  ...
}
```


3.3.4. Memory Types

The compiler maps user-defined arrays in source code to hardware memories. You can classify these hardware memories into the following categories:

- **Component Memory** on page 32
  Component memory is memory allocated from memory resources (such as RAM blocks) available on the FPGA.

- **External Memory** on page 37
  External memory is memory resources that are outside of the FPGA.

3.3.4.1. Component Memory

If you declare an array inside your component, the Intel HLS Compiler creates component memory in hardware. Component memory is sometimes referred to as local memory or on-chip memory because it is created from memory resources (such as RAM blocks) available on the FPGA.
The following source code snippet results in the creation of a component memory system, an interface to an external memory system, and access to these memory systems:

```c
#include <HLS/hls.h>
constexpr int SIZE = 128;
constexpr int N = SIZE - 1;
using HostInterface = ihc::mm_host<int, ihc::waitrequest<true>,
                                  ihc::latency<0>>;

component void memoryComponent(HostInterface &hostA)
{
    hls_memory int T[SIZE]; // declaring an array as a component memory
    for (unsigned i = 0; i < SIZE; i++)
    {
        T[i] = i; // writing to component memory
    }
    for (int i = 0; i < N; i += 2)
    {
        // reading from a component memory and writing to a external
        // Avalon memory-mapped agent component through an Avalon
        // memory-mapped host interface
        hostA[i] = T[i] + T[i + 1];
    }
}
```

The compiler performs the following tasks to build a memory system:

- Build a component memory from FPGA memory resources (such as block RAMs) and presents it to the datapath as a single memory.
- Map each array access to a load-store unit (LSU) in the datapath that transacts with the component memory through its ports.
- Automatically optimizes the component memory geometry to maximize the bandwidth available to loads and stores in the datapath.
- Attempts to guarantee that component memory accesses never stall.

To learn more about controlling memory system architectures, review the following topics:

- **Component Memories (Memory Attributes)** in the *Intel High Level Synthesis Compiler Pro Edition Reference Manual*
- **Memory Architecture Best Practices** on page 72

### Stallable and Stall-Free Memory Systems

Accesses to a memory (read or write) can be stall-free or stallable:

- **Stall-free memory access**
  - A memory access is stall-free if it has contention-free access to a memory port. A memory system is stall-free if each of its memory operations has contention-free access to a memory port.

- **Stallable memory access**
  - A memory access is stallable if it does not have contention free access to a memory port. When two datapath LSUs try to transact with a memory port in the same clock cycle, one of those memory accesses is delayed (or stalled) until the memory port in contention becomes available.
As much as possible, the Intel HLS Compiler tries to create stall-free memory systems for your component.

**Figure 22. Examples of Stall-free and Stallable Memory Systems**

This figure shows the following example memory systems:

- **A**: A stall-free memory system
  This memory system is stall-free because, even though the reads are scheduled in the same cycle, they are mapped to different ports. There is no contention for accessing the memory ports.

- **B**: A stall-free memory system
  This memory system is stall-free because the two reads are statically-scheduled to occur in different clock cycles. The two reads can share a memory port without any contention for the read access.

- **C**: A stallable memory system
  This memory system is stallable because two reads are mapped to the same port in the same cycle. The two reads happen at the same time. There reads require collision arbitration to manage their port access requests, and arbitration can affect throughput.
**Port**  
A memory port is a physical access point into a memory. A port is connected to one or more load-store units (LSUs) in the datapath. An LSU can connect to one or more ports. A port can have one or more LSUs connected.

**Bank**  
A memory bank is a division of the component memory system that contains a subset of the data stored. That is, all the data stored for a component is split across banks, with each bank containing a unique piece of the stored data.

A memory system always has at least one bank.

**Replicate**  
A memory bank replicate is a copy of the data in the memory bank with its own ports. All replicates in a bank contain the same data. Each replicate can be accessed independent of the others.

A memory bank always has at least one replicate.

**Private Copy**  
A private copy is a copy of the data in a replicate that is created for nested loops to enable concurrent iterations of the outer loop.

A replicate can comprise multiple private copies, with each iteration of an outer loop having its own private copy. Because each outer loop iteration has its own private copy, private copies are not expected to all contain the same data.
The following figure illustrates the relationship between banks, replicates, ports, and private copies:

**Figure 23. Schematic Representation of Component Memories Showing the Relationship between Banks, Replicates, Ports, and Private Copies**

A) Simplest memory configuration

- Memory Size = 4096 bytes
- Bank 0 (1024x4)
- Replicate 0 (1024x4)
- Copy 0 (1024x4)
- Depth = 1024 words
- Width = 4 bytes

B) Memory with two banks

- Memory Size = 8192 bytes
- Bank 0 (1024x4)
- Replicate 0 (1024x4)
- Copy 0 (1024x4)
- Bank 1 (1024x4)
- Replicate 0 (1024x4)
- Copy 0 (1024x4)
- Depth = 2048 words
- Width = 4 bytes

C) Memory with two replicates

- Memory Size = 8192 bytes
- Bank 0 (2048x4)
- Replicate 0 (1024x4)
- Copy 0 (1024x4)
- Replicate 1 (1024x4)
- Copy 0 (1024x4)
- Depth = 2048 words
- Width = 4 bytes

D) Memory with two private copies

- Memory Size = 8192 bytes
- Bank 0 (2048x4)
- Replicate 0 (2048x4)
- Copy 0 (1024x4)
- Replicate 1 (2048x4)
- Copy 1 (1024x4)
- Depth = 2048 words
- Width = 4 bytes
Strategies that Enable Concurrent Stall-Free Memory Accesses

The compiler uses a variety of strategies to ensure that concurrent accesses are stall-free including:

- Adjusting the number of ports the memory system has. This can be done either by replicating the memory to enable more read ports or by clocking the RAM block at twice the component clock speed, which enables four ports per replicate instead of two.

  Clocking the RAM block at twice the component clock speed to double the number of available ports to the memory system is called *double pumping*.

  All of a replicate’s physical access ports can be accessed concurrently.

- Partitioning memory content into one or more banks, such that each bank contains a subset of the data contained in the original memory (corresponds to the top-right box of Schematic Representation of Local Memories Showing the Relationship between Banks, Replicates, Ports, and Private Copies).

  The banks of a component memory can be accessed concurrently by the datapath.

- Replicating a bank to create multiple coherent replicates (corresponds to the bottom-left box of Schematic Representation of Local Memories Showing the Relationship between Banks, Replicates, Ports, and Private Copies). Each replicate in a bank contains identical data.

  The replicates are loaded concurrently.

- Creating private copies of an array that is declared inside of a loop nest (corresponds to the bottom-right box of Schematic Representation of Local Memories Showing the Relationship between Banks, Replicates, Ports, and Private Copies).

  These private copies enable loop pipelining because each pipeline-parallel loop iteration accesses its own private copy of the array declared within the loop body. Private copies are not expected to contain the same data.

Despite the compiler’s best efforts, the component memory system can still be stallable. This might happen due to resource constraints or memory attributes defined in your source code. In that case, the compiler tries to minimize the hardware resources consumed by the arbitrated memory system.

3.3.4.2. External Memory

If the component accesses memory outside of the component, the compiler creates a hardware interface through which the datapath accesses this external memory. The interface is described using a pointer or Avalon® memory-mapped host interface as a function argument to the component. One interface is created for every pointer or memory-mapped host interface component argument.

The code snippet in Component Memory on page 32 shows an external memory described with an Avalon memory-mapped host interface and its accesses within the component.

Unlike component memory, the compiler does not define the structure of the external memory. The compiler instantiates a specialized LSU for each access site based on the type of interface and the memory access patterns.
The compiler also tries various strategies to maximize the efficient use of the available memory interface bandwidth such as eliminating unnecessary accesses and statically coalescing contiguous accesses.
4. Interface Best Practices

With the Intel High Level Synthesis Compiler, your component can have a variety of interfaces: from basic wires to the Avalon Streaming and Avalon Memory-Mapped Host interfaces. Review the interface best practices to help you choose and configure the right interface for your component.

Each interface type supported by the Intel HLS Compiler Pro Edition has different benefits. However, the system that surrounds your component might limit your choices. Keep your requirements in mind when determining the optimal interface for your component.

Demonstrating Interface Best Practices

The Intel HLS Compiler Pro Edition comes with a number of tutorials that illustrate important Intel HLS Compiler concepts and demonstrate good coding practices.

Review the following tutorials to learn about different interfaces as well as best practices that might apply to your design:

<table>
<thead>
<tr>
<th>Tutorial</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>You can find these tutorials in the following location on your Intel Quartus Prime system:</td>
<td>&lt;quartus_installdir&gt;/hls/examples/tutorials</td>
</tr>
<tr>
<td>interfaces/overview</td>
<td>Demonstrates the effects on quality-of-results (QoR) of choosing different component interfaces even when the component algorithm remains the same.</td>
</tr>
<tr>
<td>best_practices/const_global</td>
<td>Demonstrates the performance and resource utilization improvements of using const qualified global variables. Also demonstrates the type of interface created when you access global variables.</td>
</tr>
<tr>
<td>best_practices/parameter_aliasing</td>
<td>Demonstrates the use of the __restrict keyword on component arguments</td>
</tr>
<tr>
<td>best_practices/lsu_control</td>
<td>Demonstrates the effects of controlling the type of LSUs instantiated for variable-latency Avalon Memory Mapped Host interfaces</td>
</tr>
<tr>
<td>interfaces/explicit_streams_buffer</td>
<td>Demonstrates how to use explicit stream_in and stream_out interfaces in the component and testbench.</td>
</tr>
<tr>
<td>interfaces/explicit_streams_packets_empty</td>
<td>Demonstrates how to use the usesPackets, usesEmpty, and firstSymbolInHighOrderBits stream template parameters.</td>
</tr>
<tr>
<td>interfaces/explicit_streams_packets_ready_valid</td>
<td>Demonstrates how to use the usesPackets, usesValid, and usesReady stream template parameters.</td>
</tr>
<tr>
<td>interfaces/mm_host_testbench_operators</td>
<td>Demonstrates how to invoke a component at different indices of an Avalon Memory Mapped (MM) Host (mm_host class) interface.</td>
</tr>
<tr>
<td>Tutorial</td>
<td>Description</td>
</tr>
<tr>
<td>-----------------------------------</td>
<td>-----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>interfaces/mm_agents</td>
<td>Demonstrates how to create Avalon-MM Agent interfaces (agent registers and agent memories).</td>
</tr>
<tr>
<td>interfaces/mm_agents_csr_volatile</td>
<td>Demonstrates the effect of using volatile keyword to allow concurrent agent memory accesses while your component is running.</td>
</tr>
<tr>
<td>interfaces/mm_agents_double_buffering</td>
<td>Demonstrates the effect of using the hls_readwrite_mode macro to control how memory hosts access the agent memories</td>
</tr>
<tr>
<td>interfaces/multiple_stream_call_sites</td>
<td>Demonstrates the tradeoffs of using multiple stream call sites.</td>
</tr>
<tr>
<td>interfaces/pointer_mm_host</td>
<td>Demonstrates how to create Avalon-MM Host interfaces and control their parameters.</td>
</tr>
<tr>
<td>interfaces/stable_arguments</td>
<td>Demonstrates how to use the stable attribute for unchanging arguments to improve resource utilization.</td>
</tr>
</tbody>
</table>

**Related Information**
- Avalon Memory-Mapped Interface Specifications
- Avalon Streaming Interface Specifications

### 4.1. Choose the Right Interface for Your Component

Different component interfaces can affect the quality of results (QoR) of your component without changing your component algorithm. Consider the effects of different interfaces before choosing the interface between your component and the rest of your design.

The best interface for your component might not be immediately apparent, so you might need to try different interfaces for your component to achieve the optimal QoR. Take advantage of the rapid component compilation time provided by the Intel HLS Compiler Pro Edition and the resulting High Level Design reports to determine which interface gives you the optimal QoR for your component.

This section uses a vector addition example to illustrate the impact of changing the component interface while keeping the component algorithm the same. The example has two input vectors, vector \( \mathbf{a} \) and vector \( \mathbf{b} \), and stores the result to vector \( \mathbf{c} \). The vectors have a length of \( N \) (which could be very large).

The core algorithm is as follows:

```c
#pragma unroll 8
for (int i = 0; i < N; ++i) {
    c[i] = a[i] + b[i];
}
```

The Intel HLS Compiler Pro Edition extracts the parallelism of this algorithm by pipelining the loops if no loop dependency exists. In addition, by unrolling the loop (by a factor of 8), more parallelism can be extracted.

Ideally, the generated component has a latency of \( N/8 \) cycles. In the examples in the following section, a value of 1024 is used for \( N \), so the ideal latency is 128 cycles (1024/8).
The following sections present variations of this example that use different interfaces. Review these sections to learn how different interfaces affect the QoR of this component.

You can work your way through the variations of these examples by reviewing the tutorial available in `<quartus_installdir>/hls/examples/tutorials/interfaces/overview`.

### 4.1.1. Pointer Interfaces

Software developers accustomed to writing code that targets a CPU might first try to code this algorithm by declaring vectors `a`, `b`, and `c` as pointers to get the data in and out of the component. Using pointers in this way results in a single Avalon Memory-Mapped (MM) Host interface that the three input variables share.

Pointers in a component are implemented as Avalon Memory Mapped (Avalon-MM) host interfaces with default settings. For more details about pointer parameter interfaces, see Intel HLS Compiler Default Interfaces in *Intel High Level Synthesis Compiler Pro Edition Reference Manual*.

The vector addition component example with pointer interfaces can be coded as follows:

```c
component void vector_add(int* a,
                          int* b,
                          int* c,
                          int N) {
    #pragma unroll 8
    for (int i = 0; i < N; ++i) {
        c[i] = a[i] + b[i];
    }
}
```

The following diagram shows the Function View in the Graph Viewer that is generated when you compile this example. Because the loop is unrolled by a factor of 8, the diagram shows that `vector_add.B2` has 8 loads for vector `a`, 8 loads for vector `b`, and 8 stores for vector `c`. In addition, all of the loads and stores are arbitrated on the same memory, resulting in inefficient memory accesses.
The following Loop Analysis report shows that the component has an undesirably high loop initiation interval (II). The II is high because vectors $a$, $b$, and $c$ are all accessed through the same Avalon-MM Host interface. The Intel HLS Compiler Pro Edition uses stallable arbitration logic to schedule these accesses, which results in poor performance and high FPGA area use.

In addition, the compiler cannot assume there are no data dependencies between loop iterations because pointer aliasing might exist. The compiler cannot determine that vectors $a$, $b$, and $c$ do not overlap. If data dependencies exist, the Intel HLS Compiler cannot pipeline the loop iterations effectively.
Compiling the component with an Intel Quartus Prime compilation flow targeting an Intel Arria® 10 device results in the following QoR metrics, including high ALM usage, high latency, high II, and low f_MAX. All of which are undesirable properties in a component.

Table 2. QoR Metrics for a Component with a Pointer Interface

<table>
<thead>
<tr>
<th>QoR Metric</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALMs</td>
<td>15593.5</td>
</tr>
<tr>
<td>DSPs</td>
<td>0</td>
</tr>
<tr>
<td>RAMs</td>
<td>30</td>
</tr>
<tr>
<td>f_MAX (MHz)</td>
<td>298.6</td>
</tr>
<tr>
<td>Latency (cycles)</td>
<td>24071</td>
</tr>
<tr>
<td>Initiation Interval (II) (cycles)</td>
<td>~508</td>
</tr>
</tbody>
</table>

1The compilation flow used to calculate the QoR metrics used Intel Quartus Prime Pro Edition Version 17.1.
2The f_MAX measurement was calculated from a single seed.

4.1.2. Avalon Memory Mapped Host Interfaces

By default, pointers in a component are implemented as Avalon Memory Mapped (Avalon MM) host interfaces with default settings. You can mitigate poor performance from the default settings by configuring the Avalon MM host interfaces.

You can configure the Avalon MM host interface for the vector addition component example using the ihc::mm_host class as follows:
component void vector_add(
  ihc::mm_host<int, ihc::aspace<1>, ihc::dwidth<8*8*sizeof(int)> >& a,
  ihc::mm_host<int, ihc::aspace<2>, ihc::dwidth<8*8*sizeof(int)> >& b,
  ihc::mm_host<int, ihc::aspace<3>, ihc::dwidth<8*8*sizeof(int)> >& c,
  int N) {
    #pragma unroll 8
    for (int i = 0; i < N; ++i) {
      c[i] = a[i] + b[i];
    }
  }

The memory interfaces for vector a, vector b, and vector c have the following attributes specified:

- The vectors are each assigned to different address spaces with the `ihc::aspace` attribute, and each vector receives a separate Avalon MM host interface. With the vectors assigned to different physical interfaces, the vectors can be accessed concurrently without interfering with each other, so memory arbitration is not needed.
- The width of the interfaces for the vectors is adjusted with the `ihc::dwidth` attribute.
- The alignment of the interfaces for the vectors is adjusted with the `ihc::align` attribute.

The following diagram shows the Function View in the System Viewer that is generated when you compile this example.
The diagram shows that `vector_add.B2` has two loads and one store. The default Avalon MM Host settings used by the code example in Pointer Interfaces on page 41 had 16 loads and 8 stores.

By expanding the width and alignment of the vector interfaces, the original pointer interface loads and stores were coalesced into one wide load each for vector a and vector b, and one wide store for vector c.
Also, the memories are stall-free because the loads and stores in this example access separate memories.

Compiling this component with an Intel Quartus Prime compilation flow targeting an Intel Arria 10 device results in the following QoR metrics:

<table>
<thead>
<tr>
<th>QoR Metric</th>
<th>Pointer</th>
<th>Avalon MM Host</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALMs</td>
<td>15593.5</td>
<td>643</td>
</tr>
<tr>
<td>DSPs</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RAMs</td>
<td>30</td>
<td>0</td>
</tr>
<tr>
<td>$f_{\text{MAX}}$ (MHz)$^2$</td>
<td>298.6</td>
<td>472.37</td>
</tr>
<tr>
<td>Latency (cycles)</td>
<td>24071</td>
<td>142</td>
</tr>
<tr>
<td>Initiation Interval (II) (cycles)</td>
<td>~508</td>
<td>1</td>
</tr>
</tbody>
</table>

$^1$The compilation flow used to calculate the QoR metrics used Intel Quartus Prime Pro Edition Version 17.1.

$^2$The $f_{\text{MAX}}$ measurement was calculated from a single seed.

All QoR metrics improved by changing the component interface to a specialized Avalon MM Host interface from a pointer interface. The latency is close to the ideal latency value of 128, and the loop initiation interval (II) is 1.

**Important:** This change to a specialized Avalon MM Host interface from a pointer interface requires the system to have three separate memories with the expected width. The initial pointer implementation requires only one system memory with a 64-bit wide data bus. If the system cannot provide the required memories, you cannot use this optimization.

### 4.1.3. Avalon Memory Mapped Agent Memories

Depending on your component, you can sometimes optimize the memory structure of your component by placing your component parameters in Avalon Memory Mapped (Avalon MM) agent memories.

Agent memories are owned by the component and expose an MM agent interface for an MM Host to read from and write to.

When you allocate an agent memory, you must define its size. Defining the size puts a limit on how large a value of $N$ that the component can process. In this example, the RAM size is 1024 words. This RAM size means that $N$ can have a maximal size of 1024 words.

The vector addition component example can be coded with an Avalon MM agent interface as follows:

```c
component void vector_add(
    hls_avalon_agent_memory_argument(1024*sizeof(int)) int* a,
    hls_avalon_agent_memory_argument(1024*sizeof(int)) int* b,
    hls_avalon_agent_memory_argument(1024*sizeof(int)) int* c,
    int N) {
    #pragma unroll 8
    for (int i = 0; i < N; ++i) {
```
$$c[i] = a[i] + b[i];$$

The following diagram shows the Function View in the System Viewer that is generated when you compile this example.

Figure 26. **System Viewer Function View of vector_add Component with Avalon MM Agent Interface**

Compiling this component with an Intel Quartus Prime compilation flow targeting an Intel Arria 10 device results in the following QoR metrics:

**Table 4. QoR Metrics Comparison for Avalon MM Agent Interface**

<table>
<thead>
<tr>
<th>QoR Metric</th>
<th>Pointer</th>
<th>Avalon MM Host</th>
<th>Avalon MM Agent</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALMs</td>
<td>15593.5</td>
<td>643</td>
<td>490.5</td>
</tr>
<tr>
<td>DSPs</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

*continued...*
QoR Metric | Pointer | Avalon MM Host | Avalon MM Agent
--- | --- | --- | ---
RAMs | 30 | 0 | 48
$f_{MAX}$ (MHz)\(^2\) | 298.6 | 472.37 | 498.26
Latency (cycles) | 24071 | 142 | 139
Initiation Interval (II) (cycles) | ~508 | 1 | 1

\(^1\)The compilation flow used to calculate the QoR metrics used Intel Quartus Prime Pro Edition Version 17.1.
\(^2\)The $f_{MAX}$ measurement was calculated from a single seed.

The QoR metrics show by changing the ownership of the memory from the system to the component, the number of ALMs used by the component are reduced, as is the component latency. The $f_{MAX}$ of the component is increased as well. The number of RAM blocks used by the component is greater because the memory is implemented in the component and not the system. The total system RAM usage (not shown) should not increase because RAM usage shifted from the system to the FPGA RAM blocks.

4.1.4. Avalon Memory Mapped Agent Registers

If you want your component to be controlled and configured by an Avalon MM Host in your system, consider placing your parameters in Avalon MM Agent registers by using the `hls_avalon_agent_register`_argument attribute.

For more information about the control-and-status register, refer to Agent Interfaces in the [Intel High Level Synthesis Compiler Pro Edition: Reference Manual](https://example.com).

4.1.5. Avalon Streaming Interfaces

Avalon Streaming (Avalon ST) interfaces support a unidirectional flow of data, and are typically used for components that drive high-bandwidth and low-latency data.

The vector addition example can be coded with an Avalon ST interface as follows:

```c
struct int_v8 {
    int data[8];
};
component void vector_add(
    ihc::stream_in<int_v8>& a,
    ihc::stream_in<int_v8>& b,
    ihc::stream_out<int_v8>& c,
    int N) {
    for (int j = 0; j < (N/8); ++j) {
        int_v8 av = a.read();
        int_v8 bv = b.read();
        int_v8 cv;
        #pragma unroll 8
        for (int i = 0; i < 8; ++i) {
            cv.data[i] = av.data[i] + bv.data[i];
        }
        c.write(cv);
    }
}
```

An Avalon ST interface has a data bus, and ready and busy signals for handshaking. The `struct` is created to pack eight integers so that eight operations at a time can occur in parallel to provide a comparison with the examples for other interfaces. Similarly, the loop count is divided by eight.
The following diagram shows the Function View in the System Viewer that is generated when you compile this example.

**Figure 27.** System Viewer Function View of `vector_add` Component with Avalon ST Interface

The main difference from other versions of the example component is the absence of memory.
The streaming interfaces are stallable from the upstream sources and the downstream output. Because the interfaces are stallable, the loop initiation interval (II) is approximately 1 (instead of exactly 1). If the component does not receive any bubbles (gaps in data flow) from upstream or stall signals from downstream, then the component achieves the desired II of 1.

If you know that the stream interfaces will never stall, you can further optimize this component by taking advantage of the `usesReady` and `usesValid` stream parameters.

Compiling this component with an Intel Quartus Prime compilation flow targeting an Intel Arria 10 device results in the following QoR metrics:

<table>
<thead>
<tr>
<th>QoR Metric</th>
<th>Pointer ALMs</th>
<th>Avalon MM Host ALMs</th>
<th>Avalon MM Agent ALMs</th>
<th>Avalon ST ALMs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALMs</td>
<td>15593.5</td>
<td>643</td>
<td>490.5</td>
<td>314.5</td>
</tr>
<tr>
<td>DSPs</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RAMs</td>
<td>30</td>
<td>0</td>
<td>48</td>
<td>0</td>
</tr>
<tr>
<td>$f_{\text{MAX}}$ (MHz)$^2$</td>
<td>298.6</td>
<td>472.37</td>
<td>498.26</td>
<td>389.71</td>
</tr>
<tr>
<td>Latency (cycles)</td>
<td>24071</td>
<td>142</td>
<td>139</td>
<td>134</td>
</tr>
<tr>
<td>Initiation Interval (II) (cycles)</td>
<td>~508</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

$^1$The compilation flow used to calculate the QoR metrics used Intel Quartus Prime Pro Edition Version 17.1.

$^2$The $f_{\text{MAX}}$ measurement was calculated from a single seed.

Table 5. QoR Metrics Comparison for Avalon ST Interface$^1$

Moving the `vector_add` component to an Avalon ST interface, further improved ALM usage, RAM usage, and component latency. The component II is optimal if there are no stalls from the interfaces.

4.1.6. Pass-by-Value Interface

For software developers accustomed to writing code that targets a CPU, passing each element in an array by value might be unintuitive because it typically results in many function calls or large parameters. However, for code that targets an FPGA device, passing array elements by value can result in smaller and simpler hardware on the FPGA device.

The vector addition example can be coded to pass the vector array elements by value as follows. A `struct` is used to pass the entire array (of 8 data elements) by value.

When you use a `struct` to pass the array, you must also do the following things:

- Define element-wise copy constructors.
- Define element-wise copy assignment operators.
- Add the `hls_register` memory attribute to all `struct` members in the definition.

```
struct int_v8 {
    hls_register int data[8];
    //copy assignment operator
    int_v8 operator=(const int_v8& org) {
```
This component takes and processes only eight elements of vector \(a\) and vector \(b\), and returns eight elements of vector \(c\). To compute 1024 elements for the example, the component needs to be called 128 times \((1024/8)\). While in previous examples the component contained loops that were pipelined, here the component is invoked many times, and each of the invocations are pipelined.

The following diagram shows the Function View in the System Viewer that is generated when you compile this example.
Figure 28. System Viewer Function View of vector_add Component with Pass-By-Value Interface

The latency of this component is one, and it has a loop initiation interval (II) of one.

Compiling this component with an Intel Quartus Prime compilation flow targeting an Intel Arria 10 device results in the following QoR metrics:

Table 6. QoR Metrics Comparison for Pass-by-Value Interface

<table>
<thead>
<tr>
<th>QoR Metric</th>
<th>Pointer</th>
<th>Avalon MM Host</th>
<th>Avalon MM Agent</th>
<th>Avalon ST</th>
<th>Pass-by-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALMs</td>
<td>15593.5</td>
<td>643</td>
<td>490.5</td>
<td>314.5</td>
<td>130</td>
</tr>
<tr>
<td>DSPs</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>RAMs</td>
<td>30</td>
<td>0</td>
<td>48</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(f_{\text{MAX}}) (MHz)(^2)</td>
<td>298.6</td>
<td>472.37</td>
<td>498.26</td>
<td>389.71</td>
<td>581.06</td>
</tr>
<tr>
<td>Latency (cycles)</td>
<td>24071</td>
<td>142</td>
<td>139</td>
<td>134</td>
<td>128</td>
</tr>
<tr>
<td>Initiation Interval (II) (cycles)</td>
<td>(~508)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
The compilation flow used to calculate the QoR metrics used Intel Quartus Prime Pro Edition Version 17.1.

The \( f_{\text{MAX}} \) measurement was calculated from a single seed.

The QoR metrics for the vector_add component with a pass-by-value interface shows fewer ALM used, a high component \( f_{\text{MAX}} \), and optimal values for latency and II. In this case, the II is the same as the component invocation interval. A new invocation of the component can be launched every clock cycle. With an initiation interval of 1, 128 component calls are processed in 128 cycles so the overall latency is 128.

4.2. Control LSUs For Your Variable-Latency MM Host Interfaces

Controlling the type of load-store units (LSUs) that the Intel HLS Compiler Pro Edition uses to interact with variable-latency Memory Mapped (MM) Host interfaces can help save area in your design. You might also encounter situations where disabling static coalescing of a load/store with other load/store operations benefits the performance of your design.

Review the following tutorial to learn about controlling LSUs:
<quartus_installdir>/hls/examples/tutorials/best_practices/lsu_control.

To see if you need to use LSU controls, review the High-Level Design Reports for your component, especially the Function Memory Viewer, to see if the memory access pattern (and its associated LSUs) inferred by the Intel HLS Compiler Pro Edition match your expected memory access pattern. If they do not match, consider controlling the LSU type, LSU coalescing, or both.

Control the Type of LSU Created

The Intel HLS Compiler Pro Edition creates either burst-coalesced LSUs or pipelined LSUs.

In general, use burst-coalesced LSUs when an LSU is expected to process many load/store requests to memory words that are consecutive. The burst-coalesced LSU attempts to "dynamically coalesce" the requests into larger bursts in order to utilize memory bandwidth more efficiently.

The pipelined LSU consumes significantly less FPGA area, but processes load/store requests individually without any coalescing. This processing is useful when your design is tight on area or when the accesses to the variable-latency MM Host interface are not necessarily consecutive.

The following code example shows both types of LSU being implemented for a variable-latency MM Host interface:

```cpp
component void
dut(mm_host<int, dwidth<128>, awidth<32>, aspace<4>, latency<0>> &Buff1,
    mm_host<int, dwidth<32>, awidth<32>, aspace<5>, latency<0>> &Buff2) {
    int Temp[SIZE];

    using pipelined = lsu<style<PIPELINED>>;
    using burst_coalesced = lsu<style<BURST_COALESCED>>;

    for (int i = 0; i<SIZE; i++) {
        Temp[i] = burst_coalesced::load(&Buff1[i]); // Burst-Coalesced LSU
    }
```
for (int i = 0; i<SIZE; i++) {
    pipelined::store(&Buff2[i], 2*Temp[i]); // Pipelined LSU
}

Disable Static Coalescing

Static coalescing is typically beneficial because it reduces the total number of LSUs in your design by statically combining multiple load/store operations into wider load/store operations.

However, there are cases where static coalescing leads to unaligned accesses, which you might not want to occur. There are also cases where multiple loads/stores get coalesced even though you intended for only a subset of them to be operational at a time. In these cases, consider disable static coalescing for the load/store operations that you did not want to be coalesced.

For the following code example, the Intel HLS Compiler does not statically coalesce the two load operations into one wide load operation:

    component int
dut(mm_host<int, dwidth<256>, awidth<32>, aspace<1>, latency<0>> &Buff1, int i, bool Cond1, bool Cond2) {
        using no_coalescing = lsu<style<PIPELINED>, static_coalescing<false>>;
        int Val = 0;
        if (Cond1) {
            Val = no_coalescing::load(&Buff1[i]);
        }
        if (Cond2) {
            Val = no_coalescing::load(&Buff1[i + 1]);
        }
        return Val;
    }

If the two load operations were coalesced, an unaligned LSU would be created, which would hurt the throughput of your component.

Related Information

Avalon Memory-Mapped Host Interfaces and Load-Store Units

4.3. Avoid Pointer Aliasing

Add a restrict type-qualifier to pointer types whenever possible. By having restrict-qualified pointers, you prevent the Intel HLS Compiler Pro Edition from creating unnecessary memory dependencies between nonconflicting read and write operations.

The restrict type-qualifier is __restrict.

Consider a loop where each iteration reads data from one array, and then it writes data to another array in the same physical memory. Without adding the restrict type-qualifier to these pointer arguments, the compiler must assume that the two arrays might overlap. Therefore, the compiler must keep the original order of memory accesses to both arrays, resulting in poor loop optimization or even failure to pipeline the loop that contains the memory accesses.

You can also use the restrict type-qualifier with Avalon memory-mapped (MM) host interfaces.
For more details, review the parameter aliasing tutorial in the following location:

<quartus_installdir>/hls/examples/tutorials/best_practices/parameter_aliasing
5. Loop Best Practices

The Intel High Level Synthesis Compiler pipelines your loops to enhance throughput. Review these loop best practices to learn techniques to optimize your loops to boost the performance of your component.

The Intel HLS Compiler Pro Edition lets you know if there are any dependencies that prevent it from optimizing your loops. Try to eliminate these dependencies in your code for optimal component performance. You can also provide additional guidance to the compiler by using the available loop pragmas.

As a start, try the following techniques:

- Manually fuse adjacent loop bodies when the instructions in those loop bodies can be performed in parallel. These fused loops can be pipelined instead of being executed sequentially. Pipelining reduces the latency of your component and can reduce the FPGA area your component uses.

- Use the `#pragma loop_coalesce` directive to have the compiler attempt to collapse nested loops. Coalescing loops reduces the latency of your component and can reduce the FPGA area overhead needed for nested loops.

- If you have two loops that can execute in parallel, consider using a system of tasks. For details, see System of Tasks Best Practices on page 87.

Tutorials Demonstrating Loop Best Practices

The Intel HLS Compiler Pro Edition comes with a number of tutorials that illustrate important Intel HLS Compiler concepts and demonstrate good coding practices.

Review the following tutorials to learn about loop best practices that might apply to your design:

<table>
<thead>
<tr>
<th>Tutorial</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>divergent_loops</code></td>
<td>Demonstrates a source-level optimization for designs with divergent loops</td>
</tr>
<tr>
<td><code>loop_coalesce</code></td>
<td>Demonstrates the performance and resource utilization improvements of using</td>
</tr>
<tr>
<td><code>loop_fusion</code></td>
<td>Demonstrates the latency and resource utilization improvements of loop fusion.</td>
</tr>
<tr>
<td><code>loop_memory_dependency</code></td>
<td>Demonstrates breaking loop-carried dependencies using the <code>ivdep pragma</code>.</td>
</tr>
</tbody>
</table>

Continued...
5. Loop Best Practices

<table>
<thead>
<tr>
<th>Tutorial</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>loop_controls/max_interleaving</td>
<td>Demonstrates a method to reduce the area utilization of a loop that meets the following conditions: • The loop has an II &gt; 1 • The loop is contained in a pipelined loop • The loop execution is serialized across the invocations of the pipelined loop</td>
</tr>
<tr>
<td>best_practices/optimize_ii_using_hls_register</td>
<td>Demonstrates how to use the hls_register attribute to reduce loop II and how to use hls_max_concurrency to improve component throughput</td>
</tr>
<tr>
<td>best_practices/parallelize_array_operation</td>
<td>Demonstrates how to improve f_{MAX} by correcting a bottleneck that arises when performing operations on an array in a loop.</td>
</tr>
<tr>
<td>best_practices/relax_reduction_dependency</td>
<td>Demonstrates a method to reduce the II of a loop that includes a floating point accumulator, or other reduction operation that cannot be computed at high speed in a single clock cycle.</td>
</tr>
<tr>
<td>best_practices/remove_loop_carried_dependency</td>
<td>Demonstrates how to improve loop performance by removing accesses to the same variable across nested loops.</td>
</tr>
<tr>
<td>best_practices/resource_sharing_filter</td>
<td>Demonstrates the following versions of a 32-tap finite impulse response (FIR) filter design: • optimized-for-throughput variant • optimized-for-area variant</td>
</tr>
<tr>
<td>best_practices/speculated_iterations</td>
<td>Demonstrates how to use #pragma speculated_iterations to control when speculated iterations are used.</td>
</tr>
<tr>
<td>best_practices/triangular_loop</td>
<td>Demonstrates a method for describing triangular loop patterns with dependencies.</td>
</tr>
</tbody>
</table>

5.1. Reuse Hardware By Calling It In a Loop

Loops are a useful way to reuse hardware. If your component function calls another function, the called function will be the top-level component. Calling a function multiple times results in hardware duplication.

For example, the following code example results in multiple hardware copies of the function \( foo \) in the component \( myComponent \) because the function \( foo \) is inlined:

```c
int foo(int a) {
    return 4 + sqrt(a) /}
}

component
void myComponent() {
    ...
    int x =
    x += foo(0);
    x += foo(1);
    x += foo(2);
    ...
}
```

If you place the function \( foo \) in a loop, the hardware for \( foo \) can be reused for each invocation. The function is still inlined, but it is inlined only once.
component
void myComponent()
{
... 
  int x = 0;
#pragma unroll 1
  for (int i = 0; i < 3; i++)
  {
    x += foo(i);
  }
... 
}

You could also use a switch/case block if you want to pass your reusable function different values that are not related to the loop induction variable i:

component
void myComponent()
{
... 
  int x = 0;
#pragma unroll 1
  for (int i = 0; i < 3; i++)
  {
    int val = 0;
    switch(i)
    {
    case 0:
      val = 3;
      break;
    case 1:
      val = 6;
      break;
    case 2:
      val = 1;
      break;
    }
    x += foo(val);
  }
... 
}

You can learn more about reusing hardware and minimizing inlining be reviewing the resource sharing tutorial available in `<quartus_installdir>/hls/examples/tutorials/best_practices/resource_sharing_filter`.

### 5.2. Parallelize Loops

One of the main benefits of using an FPGA instead of a microprocessor is that FPGAs use a spatial compute structure. A design can use additional hardware resources in exchange for lower latency.

You can take advantage of the spatial compute structure to accelerate the loops by having multiple iterations of a loop executing concurrently. To have multiple iterations of a loop execute concurrently, unroll loops when possible and structure your loops so that dependencies between loop iterations are minimized and can be resolved within one clock cycle.

These practices show how to parallelize different iterations of the same loop. If you have two different loops that you want to parallelize, consider using a system of tasks. For details, see System of Tasks Best Practices on page 87.
5.2.1. Pipeline Loops

Pipelining is a form of parallelization where multiple iterations of a loop execute concurrently, like an assembly line.

Consider the following basic loop with three stages and three iterations. A loop stage is defined as the operations that occur in the loop within one clock cycle.

![Basic loop with three stages and three iterations](image)

If each stage of this loop takes one clock cycle to execute, then this loop has a latency of nine cycles.

The following figure shows the pipelining of the loop from Figure 29 on page 59.

![Pipelined loop with three stages and four iterations](image)

The pipelined loop has a latency of five clock cycles for three iterations (and six cycles for four iterations), but there is no area tradeoff. During the second clock cycle, Stage 1 of the pipeline loop is processing iteration 2, Stage 2 is processing iteration 1, and Stage 3 is inactive.

This loop is pipelined with a loop initiation interval (II) of 1. An II of 1 means that there is a delay of 1 clock cycle between starting each successive loop iteration.

The Intel HLS Compiler Pro Edition attempts to pipeline loops by default, and loop pipelining is not subject to the same constant iteration count constraint that loop unrolling is.

Not all loops can be pipelined as well as the loop shown in Figure 30 on page 59, particularly loops where each iteration depends on a value computed in a previous iteration.
For example, consider if Stage 1 of the loop depended on a value computed during Stage 3 of the previous loop iteration. In that case, the second (orange) iteration could not start executing until the first (blue) iteration had reached Stage 3. This type of dependency is called a loop-carried dependency.

In this example, the loop would be pipelined with II=3. Because the II is the same as the latency of a loop iteration, the loop would not actually be pipelined at all. You can estimate the overall latency of a loop with the following equation:

\[
\text{latency}_{\text{loop}} = (\text{iterations} + \text{speculated iterations} - 1) \times \text{II} + \text{latency}_{\text{body}}
\]

where \( \text{latency}_{\text{loop}} \) is the number of cycles the loop takes to execute and \( \text{latency}_{\text{body}} \) is the number of cycles a single loop iteration takes to execute.

The Intel HLS Compiler Pro Edition supports pipelining nested loops without unrolling inner loops. When calculating the latency of nested loops, apply this formula recursively. This recursion means that having II>1 is more problematic for inner loops than for outer loops. Therefore, algorithms that do most of their work on an inner loop with II=1 still perform well, even if their outer loops have II>1.

**Related Information**

Speculative Execution on page 30

### 5.2.2. Unroll Loops

When a loop is unrolled, each iteration of the loop is replicated in hardware and executes simultaneously if the iterations are independent. Unrolling loops trades an increase in FPGA area use for a reduction in the latency of your component.

Consider the following basic loop with three stages and three iterations. Each stage represents the operations that occur in the loop within one clock cycle.

**Figure 31. Basic loop with three stages and three iterations**

If each stage of this loop takes one clock cycle to execute, then this loop has a latency of nine cycles.

The following figure shows the loop from Figure 31 on page 60 unrolled three times.
Three iterations of the loop can now be completed in only three clock cycles, but three times as many hardware resources are required.

You can control how the compiler unrolls a loop with the `#pragma unroll` directive, but this directive works only if the compiler knows the trip count for the loop in advance or if you specify the unroll factor. In addition to replicating the hardware, the compiler also reschedules the circuit such that each operation runs as soon as the inputs for the operation are ready.

For an example of using the `#pragma unroll` directive, see the `best_practices/resource_sharing_filter` tutorial.

### 5.2.3. Example: Loop Pipelining and Unrolling

Consider a design where you want to perform a dot-product of every column of a matrix with each other column of a matrix, and store the six results in a different upper-triangular matrix. The rest of the elements of the matrix should be set to zero.
The code might look like the following code example:

```
1. #define ROWS 4
2. #define COLS 4
3. component void dut(...) {
4.     float a_matrix[COLS][ROWS]; // store in column-major format
5.     float r_matrix[ROWS][COLS]; // store in row-major format
6.     // setup...
7.     for (int i = 0; i < COLS; i++) {
8.         for (int j = i + 1; j < COLS; j++) {
9.             float dotProduct = 0;
10.            for (int mRow = 0; mRow < ROWS; mRow++) {
11.                dotProduct += a_matrix[i][mRow] * a_matrix[j][mRow];
12.            }
13.            r_matrix[i][j] = dotProduct;
14.        }
15.     }
16.     // continue...
17. }
18. }
```

You can improve the performance of this component by unrolling the loops that iterate across each entry of a particular column. If the loop operations are independent, then the compiler executes them in parallel.

Floating-point operations typically must be carried out in the same order that they are expressed in your source code to preserve numerical precision. However, you can use the `-ffp-contract=fast` compiler flag to relax the ordering of floating-point operations. With the order of floating-point operations relaxed, the following conditions occur in this loop:

- The multiplication operations can occur in parallel.
- The addition operations can be composed into an adder tree instead of an adder chain.

To learn more, review the tutorial: `<quartus_installdir>/hls/examples/tutorials/best_practices/floating_point_ops`

The compiler tries to unroll loops on its own when it thinks unrolling improves performance. For example, the loop at line 14 is automatically unrolled because the loop has a constant number of iterations, and does not consume much hardware (ROWS is a constant defined at compile-time, ensuring that this loop has a fixed number of iterations).

You can improve the throughput by unrolling the `j`-loop at line 11, but to allow the compiler to unroll the loop, you must ensure that it has constant bounds. You can ensure constant bounds by starting the `j`-loop at `j = 0` instead of `j = i + 1`. You must also add a predication statement to prevent `r_matrix` from being assigned with invalid data during iterations 0,1,2,… of the `j`-loop.
for (int i = 0; i < COLS; i++) {
    #pragma unroll
    for (int j = 0; j < COLS; j++) {
        float dotProduct = 0;
        #pragma unroll
        for (int mRow = 0; mRow < ROWS; mRow++) {
            dotProduct += a_matrix[i][mRow] * a_matrix[j][mRow];
        }
        r_matrix[i][j] = (j > i) ? dotProduct : 0; // predication
    }
}

Now the j-loop is fully unrolled. Because they do not have any dependencies, all four iterations run at the same time.

Refer to the resource_sharing_filter tutorial located at <quartus_installdir>/hls/examples/tutorials/best_practices/resource_sharing_filter for more details.

You could continue and also unroll the loop at line 10, but unrolling this loop would result in the area increasing again. By allowing the compiler to pipeline this loop instead of unrolling it, you can avoid increasing the area and pay about only four more clock cycles assuming that the i-loop only has an II of 1. If the II is not 1, the Details pane of the Loops Analysis page in the high-level design report (report.html) gives you tips on how to improve it.

The following factors are factors that can typically affect loop II:
- loop-carried dependencies
  See the tutorial at <quartus_installdir>/hls/examples/tutorials/best_practices/loop_memory_dependency
- long critical loop path
- inner loops with a loop II > 1

5.3. Construct Well-Formed Loops

A well-formed loop has an exit condition that compares against an integer bound and has a simple induction increment of one per iteration. The Intel HLS Compiler Pro Edition can analyze well-formed loops efficiently, which can help improve the performance of your component.

The following example is a well-formed loop:

for(int i=0; i < N; i++)
{
    //statements
}

Well-formed nested loops can also help maximize the performance of your component.
The following example is a well-formed nested loop structure:

```c
for(int i=0; i < N; i++)
{
    //statements
    for(int j=0; j < M; j++)
    {
        //statements
    }
}
```

### 5.4. Minimize Loop-Carried Dependencies

Loop-carried dependencies occur when the code in a loop iteration depends on the output of previous loop iterations. Loop-carried dependencies in your component increase loop initiation interval (II), which reduces the performance of your component.

The loop structure that follows has a loop-carried dependency because each loop iteration reads data written by the previous iteration. As a result, each read operation cannot proceed until the write operation from the previous iteration completes. The presence of loop-carried dependencies reduces the pipeline parallelism that the Intel HLS Compiler Pro Edition can achieve, which reduces component performance.

```c
for(int i = 1; i < N; i++)
{
}
```

The Intel HLS Compiler Pro Edition performs a static memory dependency analysis on loops to determine the extent of parallelism that it can achieve. If the Intel HLS Compiler Pro Edition cannot determine that there are no loop-carried dependencies, it assumes that loop-dependencies exist. The ability of the compiler to test for loop-carried dependencies is impeded by unknown variables at compilation time or if array accesses in your code involve complex addressing.

To avoid unnecessary loop-carried dependencies and help the compiler to better analyze your loops, follow these guidelines:

**Avoid Pointer Arithmetic**

Compiler output is suboptimal when your component accesses arrays by dereferencing pointer values derived from arithmetic operations. For example, avoid accessing an array as follows:

```c
for(int i = 0; i < N; i++)
{
    int t = *(A++);
    *A = t;
}
```

**Introduce Simple Array Indexes**

Some types of complex array indexes cannot be analyzed effectively, which might lead to suboptimal compiler output. Avoid the following constructs as much as possible:
• Nonconstants in array indexes.
  For example, \( A[K + i] \), where \( i \) is the loop index variable and \( K \) is an unknown variable.

• Multiple index variables in the same subscript location.
  For example, \( A[i + 2 \times j] \), where \( i \) and \( j \) are loop index variables for a double nested loop.
  The array index \( A[i][j] \) can be analyzed effectively because the index variables are in different subscripts.

• Nonlinear indexing.
  For example, \( A[i \& C] \), where \( i \) is a loop index variable and \( C \) is a nonconstant variable.

**Use Loops with Constant Bounds Whenever Possible**

The compiler can perform range analysis effectively when loops have constant bounds.

You can place an `if`-statement inside your loop to control in which iterations the loop body executes.

**Ignore Memory Dependencies**

If there are no implicit memory dependencies across loop iterations, you can use the `ivdep` pragma to tell the Intel HLS Compiler Pro Edition to ignore possible memory dependencies.

For details about how to use the `ivdep` pragma, see Loop-Carried Dependencies (`ivdep Pragma`) in the Intel High Level Synthesis Compiler Pro Edition Reference Manual.

### 5.5. Avoid Complex Loop-Exit Conditions

If a loop in your component has complex exit conditions, memory accesses or complex operations might be required to evaluate the condition. Subsequent iterations of the loop cannot launch in the loop pipeline until the evaluation completes, which can decrease the overall performance of the loop.

Use the `speculated_iterations` pragma to specify how many cycles the loop exit condition can take to compute.

**Related Information**

Loop Iteration Speculation (`speculated_iterations Pragma`)
5.6. Convert Nested Loops into a Single Loop

To maximize performance, combine nested loops into a single loop whenever possible. The control flow for a loop adds overhead both in logic required and FPGA hardware footprint. Combining nested loops into a single loop reduces these aspects, improving the performance of your component.

The following code examples illustrate the conversion of a nested loop into a single loop:

<table>
<thead>
<tr>
<th>Nested Loop</th>
<th>Converted Single Loop</th>
</tr>
</thead>
<tbody>
<tr>
<td><code>for (i = 0; i &lt; N; i++)</code>&lt;br&gt;<code>{</code>&lt;br&gt;<code>  //statements</code>&lt;br&gt;<code>  </code>for (j = 0; j &lt; M; j++)<code>&lt;br&gt;</code>  //statements<code>&lt;br&gt;</code>  <code>//statements</code>&lt;br&gt;<code>}</code>&lt;br&gt;<code>//statements</code>&lt;br&gt;`}</td>
<td><code>for (i = 0; i &lt; N*M; i++)</code>&lt;br&gt;<code>{</code>&lt;br&gt;<code>  //statements</code>&lt;br&gt;<code>}</code></td>
</tr>
</tbody>
</table>

You can also specify the `loop_coalesce` pragma to coalesce nested loops into a single loop without affecting the loop functionality. The following simple example shows how the compiler coalesces two loops into a single loop when you specify the `loop_coalesce pragma`.

Consider a simple nested loop written as follows:

```c
#pragma loop_coalesce
for (int i = 0; i < N; i++)
  for (int j = 0; j < M; j++)
    sum[i][j] += i+j;
```

The compiler coalesces the two loops together so that they run as if they were a single loop written as follows:

```c
int i = 0;
int j = 0;
while(i < N){
  sum[i][j] += i+j;
  j++;
  if (j == M){
    j = 0;
    i++;
  }
}
```

For more information about the `loop_coalesce pragma`, see "Loop Coalescing (loop_coalesce Pragma)" in Intel High Level Synthesis Compiler Pro Edition Reference Manual.

You can also review the following tutorial: `<quartus_installdir>/hls/examples/tutorials/best_practices/loop_coalesce`
5.7. Place if-Statements in the Lowest Possible Scope in a Loop Nest

If you have a nests of loops, avoid placing loops within conditional statements. These conditions can cause the outer loop to take different paths (divergent loops), which can reduce the QoR of your component because these condition prevent the Intel HLS Compiler from pipelining the loops.

For example, the following code example results in divergent loops:

```c
for (int row = 0; row < outerTripCount; row++) {
    if (loopCondition)  {
        for (int col = 0; col < innerTripCount; col++) {
            foo();
        }
    } else {
        for (int col = 0; col < innerTripCount; col++) {
            bar();
        }
    }
}
```

This code example is better rewritten as follows:

```c
for (int row = 0; row < outerTripCount; row++) {
    for (int col = 0; col < innerTripCount; col++) {
        if (loopCondition) {
            foo();
        } else {
            bar();
        }
    }
}
```

For more details, review the divergent loops tutorial available in the following location:

<quartus_installdir>/hls/examples/tutorials/best_practices/divergent_loops

5.8. Declare Variables in the Deepest Scope Possible

To reduce the FPGA hardware resources necessary for implementing a variable, declare the variable just before you use it in a loop. Declaring variables in the deepest scope possible minimizes data dependencies and FPGA hardware usage because the Intel HLS Compiler Pro Edition does not need to preserve the variable data across loops that do not use the variables.

Consider the following example:

```c
int a[N];
for (int i = 0; i < m; ++i) {
    int b[N];
    for (int j = 0; j < n; ++j) {
        // statements
    }
}
```
The array \(a\) requires more resources to implement than the array \(b\). To reduce hardware usage, declare array \(a\) outside the inner loop unless it is necessary to maintain the data through iterations of the outer loop.

**Tip:** Overwriting all values of a variable in the deepest scope possible also reduces the resources necessary to represent the variable.

### 5.9. Raise Loop II to Increase \(f_{\text{MAX}}\)

If you have a loop that does not affect the throughput of your component, you can raise the initiation interval (II) of the loop with the `ii` pragma to try and increase the \(f_{\text{MAX}}\) of your design.

**Example**

Consider a case where your component has two distinct sequential pipelineable loops: an initialization loop with a low trip count and a processing loop with a high trip count and no loop-carried memory dependencies. In this case, the compiler does not know that the initialization loop has a much smaller impact on the overall throughput of your design. If possible, the compiler attempts to pipeline both loops with an II of 1.

Because the initialization loop has a loop-carried dependence, it will have a feedback path in the generated hardware. To achieve an II with such a feedback path, some clock frequency might be sacrificed. Depending on the feedback path in the main loop, the rest of your design could have run at a higher operating frequency.

If you specify `#pragma ii 2` on the initialization loop, you tell the compiler that it can be less aggressive in optimizing II for this loop. Less aggressive optimization allows the compiler to pipeline the path limiting the \(f_{\text{MAX}}\) and could allow your overall component design to achieve a higher \(f_{\text{MAX}}\).

The initialization loop takes longer to run with its new II. However, the decrease in the running time of the long-running loop due to higher \(f_{\text{MAX}}\) compensates for the increased length in running time of the initialization loop.

### 5.10. Control Loop Interleaving

The initiation interval (II) of a loop is the statically determined number of cycles between successive iteration launches of a given loop invocation. However, the statically scheduled II may differ from the realized dynamic II when considering interleaving.

With loop interleaving, the dynamic II of a loop can be approximated by the static II of the loop divided by the degree of interleaving, that is, by the number of concurrent invocations of the loop that are in flight.

Interleaving allows the iterations of more than one invocation of a loop to execute in parallel, provided that the static II of that loop is greater than 1. By default, the maximum amount of interleaving for a loop is equal to the static II of that loop.

In the presence of interleaving, the dynamic II of a loop can be approximated by the static II of the loop divided by the degree of interleaving, that is, by the number of concurrent invocations of the loop that are in flight.
Review the following tutorial to learn more about loop interleaving and how to control it:
<quartus_installdir>/hls/examples/tutorials/loop_controls/max_interleaving.
6. \( f_{\text{MAX}} \) Bottleneck Best Practices

The \( f_{\text{MAX}} \) of your component can be limited for various reasons. Review these best practices to understand some common \( f_{\text{MAX}} \) bottlenecks and how to mitigate them.

**Tutorials Demonstrating \( f_{\text{MAX}} \) Bottleneck Best Practices**

The Intel HLS Compiler Pro Edition comes with a number of tutorials that illustrate important Intel HLS Compiler concepts and demonstrate good coding practices.

Review the following tutorials to learn about \( f_{\text{MAX}} \) bottleneck best practices that might apply to your design:

<table>
<thead>
<tr>
<th>Tutorial</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>best_practices/fpga_reg</td>
<td>Demonstrates how manually adding pipeline registers can increase ( f_{\text{MAX}} )</td>
</tr>
<tr>
<td>best_practices/overview</td>
<td>Demonstrates how ( f_{\text{MAX}} ) can depend on the interface used in your component.</td>
</tr>
<tr>
<td>best_practices/parallelize_array_operation</td>
<td>Demonstrates how to improve ( f_{\text{MAX}} ) by correcting a bottleneck that arises when performing operations on an array in a loop.</td>
</tr>
<tr>
<td>best_practices/reduce_exitfifo_width</td>
<td>Demonstrates how to improve ( f_{\text{MAX}} ) by reducing the width of the FIFO belonging to the exit node of a stall-free cluster</td>
</tr>
<tr>
<td>best_practices/relax_reduction_dependency</td>
<td>Demonstrates how ( f_{\text{MAX}} ) can depend on the loop-carried feedback path.</td>
</tr>
</tbody>
</table>

6.1. Balancing Target \( f_{\text{MAX}} \) and Target II

The compiler attempts to optimize the component for different objectives for the scheduled \( f_{\text{MAX}} \) depending on whether the \( f_{\text{MAX}} \) target is set and whether the \#pragma II is set for each of the loops.

The \( f_{\text{MAX}} \) target is a strong suggestion and the compiler does not error out if it is not able to achieve this \( f_{\text{MAX}} \) whereas the \#pragma II triggers an error if the compiler cannot achieve the requested II. The \( f_{\text{MAX}} \) achieved for each block of code is shown in the Loops report.
The following table outlines the behavior of the scheduler in the Intel HLS Compiler:

<table>
<thead>
<tr>
<th>Explicitly Specify $f_{\text{MAX}}$?</th>
<th>Explicitly Specify II?</th>
<th>Compiler Behavior</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>No</td>
<td>Use heuristic to achieve best $f_{\text{MAX}}$/II trade-off.</td>
</tr>
<tr>
<td>No</td>
<td>Yes</td>
<td>Best effort to achieve the II for the corresponding loop (may not achieve the best possible $f_{\text{MAX}}$).</td>
</tr>
<tr>
<td>Yes</td>
<td>No</td>
<td>Best effort to achieve $f_{\text{MAX}}$ specified (may not achieve the best possible II).</td>
</tr>
<tr>
<td>Yes</td>
<td>Yes</td>
<td>Best effort to achieve the $f_{\text{MAX}}$ specified at the given II. The compiler errors out if it cannot achieve the requested II.</td>
</tr>
</tbody>
</table>

**Note:** If you are using an $f_{\text{MAX}}$ target in the command line or for a component, use `#pragma II = <N>` for performance-critical loops in your design.
7. Memory Architecture Best Practices

The Intel High Level Synthesis Compiler infers efficient memory architectures (like memory width, number of banks and ports) in a component by adapting the architecture to the memory access patterns of your component. Review the memory architecture best practices to learn how you can get the best memory architecture for your component from the compiler.

In most cases, you can optimize the memory architecture by modifying the access pattern. However, the Intel HLS Compiler Pro Edition gives you some control over the memory architecture.

Tutorials Demonstrating Memory Architecture Best Practices

The Intel HLS Compiler Pro Edition comes with a number of tutorials that illustrate important Intel HLS Compiler concepts and demonstrate good coding practices.

Review the following tutorials to learn about memory architecture best practices that might apply to your design:

Table 7. Tutorials Provided with Intel HLS Compiler Pro Edition

<table>
<thead>
<tr>
<th>Tutorial</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>attributes_on_mm_agent_arg</td>
<td>Demonstrates how to apply memory attributes to Avalon Memory Mapped (MM) agent arguments.</td>
</tr>
<tr>
<td>exceptions</td>
<td>Demonstrates how to use memory attributes on constants and struct members.</td>
</tr>
</tbody>
</table>
| memory_bank_configuration  | Demonstrates how to control the number of load/store ports of each memory bank and optimize your component area usage, throughput, or both by using one or more of the following memory attributes:  
  • hls_max_replicates  
  • hls_singlepump  
  • hls_doublepump  
  • hls_simple_dual_port_memory |
| memory_geometry           | Demonstrates how to control the number of load/store ports of each memory bank and optimize your component area usage, throughput, or both by using one or more of the following memory attributes:  
  • hls_bankwidth  
  • hls_numbanks  
  • hls_bankbits |

continued...
## 7. Memory Architecture Best Practices

### Tutorial Description

<table>
<thead>
<tr>
<th>Tutorial</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>memory_implementation</td>
<td>Demonstrates how to implement variables or arrays in registers, MLABs, or RAMs by using the following memory attributes:</td>
</tr>
<tr>
<td></td>
<td>• hls_register</td>
</tr>
<tr>
<td></td>
<td>• hls_memory</td>
</tr>
<tr>
<td></td>
<td>• hls_memory_impl</td>
</tr>
<tr>
<td>memory_merging</td>
<td>Demonstrates how to improve resource utilization by implementing two logical memories as a single physical memory by merging them depth-wise or width-wise with the hls_merge memory attribute.</td>
</tr>
<tr>
<td>non_trivial_initialization</td>
<td>Demonstrates how to use the C++ keyword constexpr to achieve efficient initialization of read-only variables.</td>
</tr>
<tr>
<td>non_power_of_two_memory</td>
<td>Demonstrates how to use the force_pow2_depth memory attribute to control the padding of memories that are non-power-of-two deep, and how that impacts the FPGA memory resource usage.</td>
</tr>
<tr>
<td>static_var_init</td>
<td>Demonstrates how to control the initialization behavior of statics in a component using the hls_init_on_reset or hls_init_on_powerup memory attribute.</td>
</tr>
</tbody>
</table>

### 7.1. Example: Overriding a Coalesced Memory Architecture

Using memory attributes in various combinations in your code allows you to override the memory architecture that the Intel HLS Compiler Pro Edition infers for your component.

The following code examples demonstrate how you can use the following memory attributes to override coalesced memory to conserve memory blocks on your FPGA:

- hls_bankwidth(N)
- hls_numbanks(N)
- hls_singlepump
- hls_max_replicates(N)

The original code coalesces two memory accesses, resulting in a memory system that is 256 locations deep by 64 bits wide (256x64 bits) (two on-chip memory blocks):

```c
component unsigned int mem_coalesce_default(unsigned int raddr,
                                          unsigned int waddr,
                                          unsigned int wdata){
    unsigned int data[512];
    data[2*waddr] = wdata;
    data[2*waddr + 1] = wdata + 1;
    unsigned int rdata = data[2*raddr] + data[2*raddr + 1];
    return rdata;
}
```

The following images show how the 256x64 bit memory for this code sample is structured, as well how the component memory structure is shown in the high-level design report (report.html)

**Figure 33. Memory Structure Generated for mem_coalesce_default**
The modified code implements a single on-chip memory block that is 512 words deep by 32 bits wide with stallable arbitration:

```c
component unsigned int mem_coalesce_override(unsigned int raddr,
                                             unsigned int waddr,
                                             unsigned int wdata)
{
    //Attributes that stop memory coalescing
    hls_bankwidth(4) hls_numbanks(1)
    //Attributes that specify a single-pumped single-replicate memory
    hls_singlepump hls_max_replicates(1)
    unsigned int data[512];
    data[2*waddr] = wdata;
    data[2*waddr + 1] = wdata + 1;
    unsigned int rdata = data[2*raddr] + data[2*raddr + 1];
    return rdata;
}
```

The following images show how the 512x32 bit memory with stallable arbitration for this code sample is structured, as well how the component memory structure is shown in the high-level design report (report.html).

**Figure 34. Memory Structure Generated for mem_coalesce_override**
While it might appear that you save hardware area by reducing the number of RAM blocks needed for the component, the introduction of stallable arbitration increases the amount of hardware needed to implement the component. In the following table, you can compare the number ALMs and FFs required by the components.

<table>
<thead>
<tr>
<th></th>
<th>ALMs</th>
<th>FFs</th>
<th>RAMs</th>
<th>DSPs</th>
<th>MLABs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full design (all components)</td>
<td>1164.5</td>
<td>1022</td>
<td>3</td>
<td>0</td>
<td>25</td>
</tr>
<tr>
<td>mem_coalesce_OVERRIDE</td>
<td>1074.5</td>
<td>1694</td>
<td>1</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>mem_coalesce_DEFAULT</td>
<td>90</td>
<td>208</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>

### 7.2. Example: Overriding a Banked Memory Architecture

Using memory attributes in various combinations in your code allows you to override the memory architecture that the Intel HLS Compiler Pro Edition infers for your component.

The following code examples demonstrate how you can use the following memory attributes to override banked memory to conserve memory blocks on your FPGA:

- `hls_bankwidth(N)`
- `hls_numbanks(N)`
- `hls_singlepump`
- `hls_doublepump`
The original code creates two banks of single-pumped on-chip memory blocks that are 16 bits wide:

```c
component unsigned short mem_banked(unsigned short raddr,
    unsigned short waddr,
    unsigned short wdata){
    unsigned short data[1024];
    data[2*waddr] = wdata;
    data[2*waddr + 9] = wdata + 1;
    unsigned short rdata = data[2*raddr] + data[2*raddr + 9];
    return rdata;
}
```

To save banked memory, you can implement one bank of double-pumped 32-bit-wide on-chip memory block by adding the following attributes before the declaration of `data[1024]`. These attributes fold the two half-used memory banks into one fully-used memory bank that is double pumped, so that it can be accessed as quickly as the two half-used memory banks.

```c
hls_bankwidth(2) hls_numbanks(1)
hls_doublepump
unsigned short data[1024];
```

Alternatively, you can avoid the double-clock requirement of the double-pumped memory by implementing one bank of single-pumped on-chip memory block by adding the following attributes before the declaration of `data[1024]`. However, in this example, these attributes add stallable arbitration to your component memories, which hurts your component performance.

```c
hls_bankwidth(2) hls_numbanks(1)
hls_singlepump
unsigned short data[1024];
```

### 7.3. Merge Memories to Reduce Area

In some cases, you can save FPGA memory blocks by merging your component memories so that they consume fewer memory blocks, reducing the FPGA area your component uses. Use the `hls_merge` attribute to force the Intel HLS Compiler Pro Edition to implement different variables in the same memory system.

When you merge memories, multiple component variables share the same memory block. You can merge memories by width (width-wise merge) or depth (depth-wise merge). You can merge memories where the data in the memories have different datatypes.
7. Memory Architecture Best Practices

Figure 35. Overview of width-wise merge and depth-wise merge

The following diagram shows how four memories can be merged width-wise and depth-wise.

7.3.1. Example: Merging Memories Depth-Wise

Use the `hls_merge("<mem_name>", "depth")` attribute to force the Intel HLS Compiler Pro Edition to implement variables in the same memory system, merging their memories by depth.

All variables with the same `<mem_name>` label set in their `hls_merge` attributes are merged.

Consider the following component code:

```c
component int depth_manual(bool use_a, int raddr, int waddr, int wdata) {
    int a[128];
    int b[128];
    int rdata;
    // mutually exclusive write
    if (use_a) {
        a[waddr] = wdata;
    } else {
        b[waddr] = wdata;
    }
    // mutually exclusive read
    if (use_a) {
        rdata = a[raddr];
    } else {
        rdata = b[raddr];
    }
    return rdata;
}
```

The code instructs the Intel HLS Compiler Pro Edition to implement local memories `a` and `b` as two on-chip memory blocks, each with its own load and store instructions.
Because the load and store instructions for local memories \(a\) and \(b\) are mutually exclusive, you can merge the accesses, as shown in the example code below. Merging the memory accesses reduces the number of load and store instructions, and the number of on-chip memory blocks, by half.

```c
component int depth_manual(bool use_a, int raddr, int waddr, int wdata) {
    int a[128] hls_merge("mem","depth");
    int b[128] hls_merge("mem","depth");

    int rdata;
    // mutually exclusive write
    if (use_a) {
        a[waddr] = wdata;
    } else {
        b[waddr] = wdata;
    }

    // mutually exclusive read
    if (use_a) {
        rdata = a[raddr];
    } else {
        rdata = b[raddr];
    }

    return rdata;
}
```

Figure 37. Depth-Wise Merge of Local Memories for Component depth_manual
There are cases where merging local memories with respect to depth might degrade memory access efficiency. Before you decide whether to merge the local memories with respect to depth, refer to the HLD report (result.prj/reports/report.html) to ensure that they have produced the expected memory configuration with the expected number of loads and stores instructions. In the example below, the Intel HLS Compiler Pro Edition should not merge the accesses to local memories a and b because the load and store instructions to each memory are not mutually exclusive.

```c
component int depth_manual(bool use_a, int raddr, int waddr, int wdata) {
    int a[128] hls_merge("mem","depth");
    int b[128] hls_merge("mem","depth");
    int rdata;
    // NOT mutually exclusive write
    a[waddr] = wdata;
    b[waddr] = wdata;
    // NOT mutually exclusive read
    rdata = a[raddr];
    rdata += b[raddr];
    return rdata;
}
```

In this case, the Intel HLS Compiler Pro Edition might double pump the memory system to provide enough ports for all the accesses. Otherwise, the accesses must share ports, which prevent stall-free accesses.

**Figure 38.** Local Memories for Component depth_manual with Non-Mutually Exclusive Accesses

![Local Memories Diagram](image)

### 7.3.2. Example: Merging Memories Width-Wise

Use the hls_merge("<mem_name>","width") attribute to force the Intel HLS Compiler Pro Edition to implement variables in the same memory system, merging their memories by width.

All variables with the same <mem_name> label set in their hls_merge attributes are merged.
Consider the following component code:

```c
component short width_manual (int raddr, int waddr, short wdata) {
    short a[256];
    short b[256];
    short rdata = 0;
    // Lock step write
    a[waddr] = wdata;
    b[waddr] = wdata;
    // Lock step read
    rdata += a[raddr];
    rdata += b[raddr];
    return rdata;
}
```

**Figure 39. Implementation of Local Memory for Component width_manual**

In this case, the Intel HLS Compiler Pro Edition can coalesce the load and store instructions to local memories `a` and `b` because their accesses are to the same address, as shown below.

```c
component short width_manual (int raddr, int waddr, short wdata) {
    short a[256] hls_merge("mem","width");
    short b[256] hls_merge("mem","width");
    short rdata = 0;
    // Lock step write
    a[waddr] = wdata;
    b[waddr] = wdata;
    // Lock step read
    rdata += a[raddr];
    rdata += b[raddr];
    return rdata;
}
```
7.4. Example: Specifying Bank-Selection Bits for Local Memory Addresses

You have the option to tell the Intel HLS Compiler Pro Edition which bits in a local memory address select a memory bank and which bits select a word in that bank. You can specify the bank-selection bits with the `hls_bankbits(b_0, b_1, ..., b_n)` attribute.

The \((b_0, b_1, ..., b_n)\) arguments refer to the local memory address bit positions that the Intel HLS Compiler Pro Edition should use for the bank-selection bits. Specifying the `hls_bankbits(b_0, b_1, ..., b_n)` attribute implies that the number of banks equals \(2^{\text{number of bank bits}}\).

**Table 8. Example of Local Memory Addresses Showing Word and Bank Selection Bits**

This table of local memory addresses shows an example of how a local memory might be addressed. The memory attribute is set as `hls_bankbits(3,4)`. The memory bank selection bits (bits 3, 4) in the table bits are in bold text and the word selection bits (bits 0-2) are in italic text.

<table>
<thead>
<tr>
<th>Bank 0</th>
<th>Bank 1</th>
<th>Bank 2</th>
<th>Bank 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word 0</td>
<td>000000</td>
<td>010000</td>
<td>100000</td>
</tr>
<tr>
<td>Word 1</td>
<td>000001</td>
<td>010001</td>
<td>100001</td>
</tr>
<tr>
<td>Word 2</td>
<td>000010</td>
<td>010010</td>
<td>100010</td>
</tr>
<tr>
<td>Word 3</td>
<td>000111</td>
<td>010111</td>
<td>101111</td>
</tr>
<tr>
<td>Word 4</td>
<td>001000</td>
<td>011000</td>
<td>101000</td>
</tr>
<tr>
<td>Word 5</td>
<td>001011</td>
<td>011011</td>
<td>101011</td>
</tr>
<tr>
<td>Word 6</td>
<td>001100</td>
<td>011100</td>
<td>101100</td>
</tr>
<tr>
<td>Word 7</td>
<td>001111</td>
<td>011111</td>
<td>101111</td>
</tr>
</tbody>
</table>

**Restriction:** Currently, the `hls_bankbits(b_0, b_1, ..., b_n)` attribute supports only consecutive bank bits.

**Example of Implementing the `hls_bankbits` Attribute**

Consider the following example component code:
As illustrated in the following figure, this code example generates multiple load and store instructions, and therefore multiple load/store units (LSUs) in the hardware. If the memory system is not split into multiple banks, there are fewer ports than memory access instructions, leading to arbitrated accesses. This arbitration results in a high loop initiation interval (II) value. Avoid arbitration whenever possible because it increases the FPGA area utilization of your component and impairs the performance of your component.

(1) For this example, the initial component was generated with the `hls_numbanks` attribute set to 1 (`hls_numbanks(1)`) to prevent the compiler from automatically splitting the memory into banks.
By default, the Intel HLS Compiler Pro Edition splits the memory into banks if it determines that the split is beneficial to the performance of your component. The compiler checks if any bits remain constant between accesses, and automatically infers bank-selection bits.

Now, consider the following component code:

```c
component int bank_no_arbitration (int raddr, int waddr, int wdata) {
    #define DIM_SIZE 4
```
// Adjust memory geometry by preventing coalescing and splitting memory
hls_bankbits(4, 5)
hls_bankwidth(sizeof(int)*DIM_SIZE)

// Force each memory bank to have 2 ports for read/write
hls_singlepump
hls_max_replicates(1)

int a[DIM_SIZE][DIM_SIZE][DIM_SIZE];

// initialize array a...
int result = 0;

#pragma unroll
for (int dim1 = 0; dim1 < DIM_SIZE; dim1++)
#pragma unroll
for (int dim3 = 0; dim3 < DIM_SIZE; dim3++)
a[dim1][waddr&(DIM_SIZE-1)][dim3] = wdata;

#pragma unroll
for (int dim1 = 0; dim1 < DIM_SIZE; dim1++)
#pragma unroll
for (int dim3 = 0; dim3 < DIM_SIZE; dim3++)
result += a[dim1][raddr&(DIM_SIZE-1)][dim3];

return result;

The following diagram shows that this example code creates a memory configuration with four banks. Using bits 4 and 5 as bank selection bits ensures that each load/store access is directed to its own memory bank.
In this code example, setting `hls_numbanks(4)` instead of `hls_bankbits(4, 5)` results in the same memory configuration because the Intel HLS Compiler Pro Edition automatically infers the optimal bank selection bits.

In the Function Memory Viewer (inf the High-Level Design Reports), the **Address bit information** shows the bank selection bits as \(b_6\) and \(b_7\), instead of \(b_4\) and \(b_5\):

<table>
<thead>
<tr>
<th>Address bit information</th>
<th>Byte address</th>
<th>(b_7)</th>
<th>(b_6)</th>
<th>(b_5)</th>
<th>(b_4)</th>
<th>(b_3)</th>
<th>(b_2)</th>
<th>(b_1)</th>
<th>(b_0)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-word bits</td>
<td></td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Word address bits</td>
<td>(b_7)</td>
<td>(b_6)</td>
<td>(b_5)</td>
<td>(b_4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bank bits</td>
<td>(b_7)</td>
<td>(b_6)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
This difference occurs because the address bits reported in the Function Memory Viewer are based on byte addresses and not element addresses. Because every element in array `a` is four bytes in size, bits `b4` and `b5` in element address bits correspond to bits `b6` and `b7` in byte addressing.
8. System of Tasks Best Practices

Using a system of HLS tasks in your component enables a variety of design structures that you can implement.

Common uses for a system of tasks include the following cases:

- Executing multiple loops in parallel
- Sharing an expensive compute block
- Designing your HLS system hierarchically and testing it in the Intel HLS Compiler Pro Edition simulation environment.

After you implement a system of tasks, you might need to balance the capacity of your task functions to improve performance. For details, review the advice in Balancing Capacity in a System of Tasks on page 88.

8.1. Executing Multiple Loops in Parallel

By using HLS tasks, you can run sequential loops in a pipelined manner within the context of the loop nest.

For example, in the following code sample, the first and second loops can be executing different invocations of the component `foo()` if the invocations can be pipelined by the Intel HLS Compiler Pro Edition:

```c
component void foo() {
    // first loop
    for (int i = 0; i < n; i++) {
        // Do something
    }
    // second loop
    for (int i = 0; i < m; i++) {
        // Do something else
    }
}
```

However, the same invocation of the component `foo()` cannot execute the two loops in parallel. System of tasks provides a way to achieve this by moving the loops into asynchronous tasks. With the first loop in an asynchronous task, the second loop can run concurrently with the first loop.

```c
void first_loop() {
    for (int i = 0; i < n; i++) {
        // Do something
    }
}

void second_loop() {
    for (int i = 0; i < m; i++) {
        // Do something else
    }
}
```
component void foo() {
    ihc::launch<first_loop>();
    ihc::launch<second_loop>();
    ihc::collect<first_loop>();
    ihc::collect<second_loop>();
}

Review the tutorial <quartus_installdir>/hls/examples/tutorials/system_of_tasks/parallel_loop to learn more about how to run multiple loops in parallel.

8.2. Sharing an Expensive Compute Block

With a system of tasks, you can share hardware resources at a function level. A component or another HLS task can invoke an HLS task multiple times. All \texttt{ihc::launch} and \texttt{ihc::collect} calls to the same function share the same hardware.

To allow for calls from multiple places to a task, the Intel HLS Compiler Pro Edition generates arbitration logic to the called task function. This arbitration logic can increase the area utilization of the component. However, if the shared logic is large, the trade-off can help you save FPGA resources. The savings can be especially noticed when your component has a large compute block that is not always active.

Review the tutorial <quartus_installdir>/hls/examples/tutorials/system_of_tasks/resource_sharing to see a simple example of how to share a compute block in component.

8.3. Implementing a Hierarchical Design

When you use a system of tasks, you can implement your design hierarchically, which allows for bottom-up design.

If you do not use a system of tasks, function calls in your HLS component are in-lined and optimized together with the calling code, which can be detrimental in some situations. Use a system of tasks to prevent smaller blocks of your design from being affected by the rest of the system.

The hierarchical design pattern implemented by using a system of tasks can give you the following benefits:

- Modularity similar to what a hardware description language (HDL) might provide
- Unpipelineable or poorly pipelined loops can be isolated so that they do not affect an entire loop nest

8.4. Balancing Capacity in a System of Tasks

If your component contains parallel task paths with different latencies, you might experience poor performance, and in some cases, deadlock.
Typically, these performance issues are caused by a lack of capacity in the datapath of the functions calling task function using the `ihc::launch` and `ihc::collect` calls. You can improve system throughput in these cases by adding a buffer to the explicit streams to account for the latency of the task functions.

Review the following tutorials to learn more about avoiding potential performance issues in a component that uses a system of tasks:

- `<quartus_installdir>/hls/examples/tutorials/system_of_tasks/balancing_pipeline_latency`
- `<quartus_installdir>/hls/examples/tutorials/system_of_tasks/balancing_loop_delay`
- `<quartus_installdir>/hls/examples/tutorials/system_of_tasks/launch_and_collect_capacity`

The Intel HLS Compiler Pro Edition emulator models the size of the buffer attached to a stream. However, the emulator does not fully account for hardware latencies, and it might exhibit different behavior between simulation and emulation in these cases.

In addition to the techniques outlined in the tutorials, follow the practices that follow to try to maximize the data throughput of your design.

### 8.4.1. Enable the Intel HLS Compiler to Infer Data Path Buffer Capacity

**Requirements**

In many situations, the Intel HLS Compiler can add buffer capacity automatically to the data path in a system of tasks design to achieve maximum throughput for your design. Follow a few best practices to help the Intel HLS Compiler effectively add data path buffer capacity to your design when needed.

As an example, consider the following design that runs two independent tasks. This kind of structure can be generated by code like the following example:

```c
component foo() {
    // Parse/compute data for tasks
    ihc::launch<task1>(data1);
    ihc::launch<task2>(data2);

    auto r1 = ihc::collect<task1>();
    auto r2 = ihc::collect<task2>();
    // Usage of r1, r2
}
```

The following diagram shows the state of the system of tasks at the start of the third invocation of the component, and the location of data in the overall pipeline from previous invocations.
Figure 43. Data Flow of Multiple Component Invocations Through a System of Tasks

The circles represent pipelined stages of the component, while the number indicate the location of data from different invocations of component `foo`. This diagram shows three invocations of the component underway.

In this diagram, **Entry** represents the two independent launch calls, and the **Exit** represents the two independent collect calls.

**Entry** provides work to both tasks only if both tasks can take in data (that is, both task have available buffer capacity). Similarly, **Exit** consumes the results only when both results are available.

If **Task1** and **Task2** have the same number of pipeline stages, then the data path performs at full throughput. Some data path buffer capacity is needed in the caller function to ensure that the caller can continue issuing launch calls while the collect calls wait for the task functions to complete. The compiler adds this data path buffer capacity automatically.

If the two tasks have different pipeline depths, then the design encounters a bottleneck because the task with the smaller pipeline depth lacks the buffer capacity to store finished results while waiting for the other task to finish. In this case, you can add buffer capacity to either launch or the collect call of the task with the smaller pipeline depth. For details about adding launch/collect buffer capacity, see Explicitly Add Buffer Capacity to Your Design When Needed on page 91.

The Intel HLS Compiler tries to balance data path buffer capacity automatically, but it can only add data path capacity automatically when your design follows certain practices.

Use the following best practices to obtain the maximum throughput for your system of tasks design:
A component or task function should do one of the following things:
— Do all of the work by itself without launching other tasks.
— Act as an orchestrator for issuing \texttt{ihc::launch} or \texttt{ihc::collect} calls and do none of the work.

If throughput is a priority for your design, avoid using multiple \texttt{ihc::launch} or \texttt{ihc::collect} calls to the same task function unless you are reusing the calls to the function by iterating in a loop.

Keep \texttt{ihc::launch} and \texttt{ihc::collect} calls to the same task function within the same block.

Review the block structure of your design with the Graph Viewer in the High-Level Design Reports to confirm that your calls are in the same block.

Avoid guarding your \texttt{ihc::launch} and \texttt{ihc::collect} calls with an if-condition.

If you are guarding your \texttt{ihc::launch} and \texttt{ihc::collect} calls with an if-condition, use the same if-condition for both the \texttt{ihc::launch} and \texttt{ihc::collect} calls.

### 8.4.2. Explicitly Add Buffer Capacity to Your Design When Needed

When the Intel HLS Compiler cannot infer the optimal capacity requirements, you can explicitly add buffer capacity to your design by specifying a value for the \texttt{capacity} parameter of the \texttt{ihc::launch} and \texttt{ihc::collect} functions.

#### Adding Capacity When Launching Task Functions

Consider specifying the \texttt{capacity} parameter of the \texttt{ihc::launch} call if you see stall patterns in your simulation waveforms that indicate an imbalance between the following things:

— Any back-pressure introduced by the task function
— How often the caller launches the task function

Figure 43 on page 90 shows the block diagram of such a design.

#### Adding Capacity When Collecting Task Functions

Consider specifying the \texttt{capacity} parameter of the \texttt{ihc::collect} call if you see stall patterns in your design waveforms that indicate a difference in the following things:

— The cadence of data production in the task function
— The cadence reading that data by the caller function

#### Deciding When To Specify The \texttt{capacity} Parameter

When you decide whether to add \texttt{ihc::launch} or \texttt{ihc::collect} capacity, ask yourself the following questions:
- **Do any tasks spend cycles stalled waiting for input data to reach them?**
  If yes, these tasks require launch capacity equal to the number of cycles it takes input data to reach the task.
  Adding launch capacity allows stalled tasks to buffer their start signals so that they do not stall the other tasks that are scheduled to launch on the same cycle.
  Because the consumer task depends on data from the producer task function, the consumer task stalls until data from the producer task reaches it, so you should add capacity to the `ihc::launch` call for the consumer task.

- **Do any tasks finish their first execution before the slowest task in the design (that is, the task that produces its initial `return` signal last) finishes its first execution?**
  If yes, the tasks that finish before the slowest task in the design finishes require additional collect capacity.
  Add capacity equal to the number of cycles between when the task produces its first `return` signal and when the slowest task in the design produces its first `return` signal.
  When you add collect capacity, tasks can buffer their `return` signals. Buffering the `return` signal consumes it from the task, which allows the task to produce more `return` signals without stalling.

Typically, tasks that communicate only through a return value do not require buffer capacity. The top-level component handles the synchronization of the communication of task functions, as the following figure shows:

![Diagram of task communication](https://via.placeholder.com/150)

When task functions communicate via streams, you might need to add capacity when there is a chance that a task might be launched before its inputs are ready. In the following diagram, you might need to add launch capacity to the consumer task:
For an example of adding buffer capacity to a design, refer the following tutorial:

<quartus_installdir>/hls/examples/tutorials/system_of_tasks/launch_and_collect_capacity
# 9. Datatype Best Practices

The datatypes in your component and possible conversions or casting that they might undergo can significantly affect the performance and FPGA area usage of your component. Review the datatype best practices for tips and guidance on how best to control datatype sizes and conversions in your component.

After you optimize the algorithm bottlenecks of your design, you can fine-tune some datatypes in your component by using arbitrary precision datatypes to shrink data widths, which reduces FPGA area utilization. The Intel HLS Compiler Pro Edition provides debug functionality so that you can easily detect overflows in arbitrary precision datatypes.

Because C++ automatically promotes smaller datatypes such as `short` or `char` to 32 bits for operations such as addition or bit-shifting, you must use the arbitrary precision datatypes if you want to create narrow datapaths in your component.

## Tutorials Demonstrating Datatype Best Practices

The Intel HLS Compiler Pro Edition comes with a number of tutorials that illustrate important Intel HLS Compiler concepts and demonstrate good coding practices.

Review the following tutorials to learn about datatype best practices that might apply to your design:

<table>
<thead>
<tr>
<th>Tutorial</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>You can find these tutorials in the following location on your Intel Quartus Prime system:</td>
<td>&lt;quartus_installdir&gt;/hls/examples/tutorials</td>
</tr>
<tr>
<td>best_practices/ac_datatypes</td>
<td>Demonstrates the effect of using <code>ac_int</code> datatype instead of <code>int</code> datatype.</td>
</tr>
<tr>
<td>ac_datatypes/ac_fixed_constructor</td>
<td>Demonstrates the use of the <code>ac_fixed</code> constructor where you can get a better QoR by using minor variations in coding style.</td>
</tr>
<tr>
<td>ac_datatypes/ac_int_basic_ops</td>
<td>Demonstrates the operators available for the <code>ac_int</code> class.</td>
</tr>
<tr>
<td>ac_datatypes/ac_int_overflow</td>
<td>Demonstrates the usage of the <code>DEBUG_AC_INT_WARNING</code> and <code>DEBUG_AC_INT_ERROR</code> keywords to help detect overflow during emulation runtime.</td>
</tr>
<tr>
<td>best_practices/single_vs_double_precision_mat</td>
<td>Demonstrates the effect of using single precision literals and functions instead of double precision literals and functions.</td>
</tr>
<tr>
<td>ac_datatypes/ac_fixed_constructor</td>
<td>Demonstrates the use of the <code>ac_fixed</code> math library functions.</td>
</tr>
</tbody>
</table>

...continued
## Tutorial Description

<table>
<thead>
<tr>
<th>Tutorial</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>hls_float/1_reduced_double</td>
<td>Demonstrates how your applications can benefit from changing the underlying type from double to hls_float&lt;11,44&gt; (reduced double).</td>
</tr>
<tr>
<td>hls_float/2_explicit_arithmetic</td>
<td>Demonstrates how to use explicit versions of hls_float binary operations to perform floating-point arithmetic operations based on your needs.</td>
</tr>
<tr>
<td>hls_float/3_conversions</td>
<td>Demonstrates when conversions appear in designs that use the hls_float data type and how to take advantage of different conversion modes to generate compile-type constants using hls_float types.</td>
</tr>
</tbody>
</table>

### 9.1. Avoid Implicit Data Type Conversions

Compile your component code with the `-Wconversion` compiler option, especially if your component uses floating point variables.

Using this option helps you avoid inadvertently having conversions between double-precision and single-precision values when double-precision variables are not needed. In FPGAs, using double-precision variables can negatively affect the data transfer rate, the latency, and resource utilization of your component.

Additionally, constants are treated as `signed int` or `signed double`. If you want efficient operations with narrower constants, cast constants to other, narrower data types like `ac_int<>` or `float`.

If you use the Algorithmic C (AC) arbitrary precision datatypes, pay attention to the type propagation rules.

### 9.2. Avoid Negative Bit Shifts When Using the `ac_int` Datatype

The `ac_int` datatype differs from other languages, including C and Verilog, in bit shifting. By default, if the shift amount is of a signed datatype `ac_int` allows negative shifts.

In hardware, this negative shift results in the implementation of both a left shifter and a right shifter. The following code example shows a shift amount that is a signed datatype.

```cpp
int14 shift_left(int14 a, int14 b) {
    return (a << b);
}
```

If you know that the shift is always in one direction, to implement an efficient shift operator, declare the shift amount as an unsigned datatype as follows:

```cpp
int14 efficient_left_only_shift(int14 a, uint14 b) {
    return (a << b);
}
```
10. Advanced Troubleshooting

As you develop components with the Intel HLS Compiler Pro Edition, you might encounter issues whose solution is unclear. The issues typically fall into the following categories:

- Your component behaves differently in simulation and emulation.
- Your component has unexpectedly poor performance, resource utilization, or both.

(Windows only) Long Path Issues

Components kept in directories that result in a long path name might not compile properly or fail in simulation. Check your compilation log or simulation log (debug.log) to determine if the path length is a cause of the failures. Errors that indicate that a file could not be found can indicate that your paths might be too long.

The Intel HLS Compiler uses the component name in many of the directories that it creates. Long component names can introduce long path issues even if your component is in a relatively shallow location in your directory structure.

To resolve a long path issue, try the following actions:

- Move your component up the directory hierarchy.
- Use a shorter name for your component.

10.1. Component Fails Only In Simulation

Discrepancies between the results of compiling your component in emulation (\(-march=x86-64\) mode or simulation (\(-march=FPGA\_name\_or\_part\_no\) mode are typically caused by relying on undefined or implementation-defined behavior in your component or testbench. However, there are some common cases where the discrepancies are caused by something else.

Comparing Floating Point Results

Use an epsilon when comparing floating point value results in the testbench. Floating points results from the RTL hardware are different from the x86 emulation flow.

Using \#pragma ivdep to Ignore Memory Dependencies

The \#pragma ivdep compiler pragma can cause functional incorrectness in your component if your component has a memory dependency that you attempted to ignore with the pragma. You can try to use the safelen modifier to control how many memory accesses that you can permit before a memory dependency occurs.

To see an example of using the ivdep pragma, review the tutorial in `<quartus_installdir>/hls/examples/tutorials/best_practices/loop_memory_dependency`.

**Check for Uninitialized Variables**

Many coding practices can result in behavior that is undefined by the C++ specification. Sometimes this undefined behavior works one way in emulation and a different way in simulation.

A common example of this situation occurs when your design reads from uninitialized variables, especially uninitialized `struct` variables.

Check your code for uninitialized values with the `-Wuninitialized` compiler flag, or debug your emulation testbench with the valgrind debugging tool. The `-Wuninitialized` compiler flag does not show uninitialized `struct` variables.

You can also check for misbehaving variables by using one or more stream interfaces as debug streams. You can add one or more `ihc::stream_out` interfaces to your component to have the component write out its internal state variables as it executes. By comparing the output of the emulation flow and the simulation flow, you can see where the RTL behavior diverges from the emulator behavior.

**Non-blocking Stream Accesses**

The emulation model of `tryRead()` is not cycle-accurate, so the behavior of `tryRead()` might differ between emulation and simulation.

If you have a non-blocking stream access (for example, `tryRead()` from a stream with a FIFO (that is, the `ihc::depth<>` template parameter), then the first few iterations of `tryRead()` might return `false` in simulation, but return `true` in emulation.

In this case, invoke your component a few extra times from the testbench to guarantee that it consumes all data in the stream. These extra invocations should not cause functional problems because `tryRead()` returns `false`.

**Invoking Component With `ihc_hls_enqueue()` Or `ihc_hls_enqueue_noret()`**

If you invoke your component with `ihc_hls_enqueue()` Or `ihc_hls_enqueue_noret()`, ensure that you are also calling `ihc_hls_component_run_all()`.

If you do not call `ihc_hls_component_run_all()`, simulation of your component fails.

**10.2. Component Gets Poor Quality of Results**

While there are many reasons why your design achieves a poor quality of results (QoR), bad memory configurations are often an important factor. Review the Function Memory Viewer report in the High Level Design Reports, and look for stallable arbitration nodes and unexpected RAM utilization.

The information in this section describes some common sources of stallable arbitration nodes or excess RAM utilization.
Component Uses More FPGA Resource Than Expected

By default, the Intel HLS Compiler Pro Edition tries to optimize your component for the best throughput by trying to maximize the maximum operating frequency (f(MAX)).

A way to reduce area consumption is to relax the f_MAX requirements by setting a target f_MAX value with the --clock i++ command option or the hls_scheduler_target_fmax_mhz component attribute. The HLS compiler can often achieve a higher f_MAX than you specify, so when you set a target f_MAX to a lower value than you need, your design might still achieve an acceptable f_MAX value, and a design that consumes less area.

To learn more about the behavior of f_MAX target value control see the following tutorial: <quartus_installdir>/hls/examples/tutorials/best_practices/set_component_target_fmax

Loops Do Not Achieve II=1

If you specify a target f_MAX, the compiler might conservatively increase II in order to achieve your target f_MAX.

If you specify a target f_MAX and require II=1, you should use #pragma ii 1 on your loops that require II=1. For more details, refer to Balancing Target fMAX and Target II on page 70.

Incorrect Bank Bits

If you access parts of an array in parallel (either a single- or multidimensional array), you might need to configure the memory bank selection bits.

See Memory Architecture Best Practices on page 72 for details about how to configure efficient memory systems.

Conditional Operator Accessing Two Different Arrays of struct Variables

In some cases, if you try to access different arrays of struct variables with a conditional operator, the Intel HLS Compiler Pro Edition merges the arrays into the same RAM block. You might see stallable arbitration in the Function Memory Viewer because there are not enough Load/Store site on the memory system.

For example, the following code examples show an array of struct variables, a conditional operator that results in stallable arbitration, and a workaround that avoids stallable arbitration.

```c
struct MyStruct {
    float a;
    float b;
}

MyStruct array1[64];
MyStruct array2[64];

The following conditional operator that uses these arrays of struct variables causes stallable arbitration:

MyStruct value = (shouldChooseArray1) ? array1[idx] : array2[idx];
```
You can avoid the stallable arbitration that the conditional operator causes here by removing the operator and using an explicit `if` statement instead.

```c
MyStruct value;
if (shouldChooseArray1)
{
   value = array1[idx];
} else
{
   value = array2[idx];
}
```

**Cluster Logic**

Your design might consume more RAM blocks than you expect, especially if you store many array variables in large registers.

You can use the `hls_use_stall_enable_clusters` component attribute to prevent the compiler from inserting stall-free cluster exit FIFOs.

The Area Analysis of System report in the high-level design report (`report.html`) can help find this issue.

The three matrices are stored intentionally in RAM blocks, but the RAM blocks for the matrices account for less than half of the RAM blocks consumed by the component.
If you look further down the report, you might see that many RAM blocks are consumed by Cluster logic or State variable. You might also see that some of your array values that you intended to be stored in registers were instead stored in large numbers of RAM blocks.

![RAM block usage table]

Notice the number of RAM blocks that are consumed by Cluster Logic and State.

In some cases, you can reduce this RAM block usage by with the following techniques:

- Pipeline loops instead of unrolling them.
- Storing local variables in local RAM blocks (hls_memory memory attribute) instead of large registers (hls_register memory attribute).

**Component with a System of Tasks Hangs or has Poor Throughput**

If your component contains a system of tasks, you might need to add launch/collection capacity.

Incorrectly specifying launch/collection capacity can result in hangs or poor throughput.

For details, refer to Balancing Capacity in a System of Tasks on page 88.

### Archives

<table>
<thead>
<tr>
<th>Intel HLS Compiler Version</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>20.3</td>
<td>Intel HLS Compiler Pro Edition Best Practices Guide</td>
</tr>
<tr>
<td>19.3</td>
<td>Intel HLS Compiler Best Practices Guide</td>
</tr>
<tr>
<td>19.2</td>
<td>Intel HLS Compiler Best Practices Guide</td>
</tr>
<tr>
<td>19.1</td>
<td>Intel HLS Compiler Best Practices Guide</td>
</tr>
<tr>
<td>18.1.1</td>
<td>Intel HLS Compiler Best Practices Guide</td>
</tr>
<tr>
<td>18.1</td>
<td>Intel HLS Compiler Best Practices Guide</td>
</tr>
<tr>
<td>18.0</td>
<td>Intel HLS Compiler Best Practices Guide</td>
</tr>
<tr>
<td>17.1.1</td>
<td>Intel HLS Compiler Best Practices Guide</td>
</tr>
<tr>
<td>17.1</td>
<td>Intel HLS Compiler Best Practices Guide</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>2021.06.21</td>
<td>21.2</td>
<td>• Revised references to Avalon Interfaces to align with new Avalon Interconnect terminology. Avalon master interfaces are now Avalon host interfaces, and Avalon slave interfaces are now Avalon agent interfaces.</td>
</tr>
<tr>
<td>2021.03.29</td>
<td>21.1</td>
<td>• Added Deciding When To Specify The capacity Parameter on page 91 to Explicitly Add Buffer Capacity to Your Design When Needed on page 91. • Replaced references to Graph Viewer with references to System Viewer to reflect the new name of the viewer.</td>
</tr>
<tr>
<td>2020.12.14</td>
<td>20.4</td>
<td>• Added Place if-Statements in the Lowest Possible Scope in a Loop Nest on page 67. • Revised Clustering the Datapath on page 18 to improve the description of stall-free clusters and their exit FIFOs. • Added Balancing Target fMAX and Target II on page 70. This content was available previously in the “Loop Analysis Report” section of the Intel HLS Compiler User Guide.</td>
</tr>
<tr>
<td>2020.09.28</td>
<td>20.3</td>
<td>• Added FPGA Concepts on page 7. • Revised Pass-by-Value Interface on page 50.This revision includes added requirements you must meet when using a struct to pass an array to your component and an updated code example showing how to implement these requirements. • Added Control Loop Interleaving on page 68.</td>
</tr>
<tr>
<td>2020.06.22</td>
<td>20.2</td>
<td>• Added information about capacity balancing in a system of tasks: — Balancing Capacity in a System of Tasks on page 88 (formerly Avoiding Potential Performance Pitfalls) — Enable the Intel HLS Compiler to Infer Data Path Buffer Capacity Requirements on page 89 — Explicitly Add Buffer Capacity to Your Design When Needed on page 91 • Updated Loop Best Practices on page 56 to add remove_loop_carried_dependency tutorial to list of loop best practices tutorials. • Added Raise Loop II to Increase fMAX on page 68. This information was available previously in the Intel HLS Compiler Reference Manual in the Loop Initiation Interval (i.e Pragma) section.</td>
</tr>
<tr>
<td>------------------</td>
<td>--------------------------------------</td>
<td>---------</td>
</tr>
</tbody>
</table>
| 2020.04.13       | 20.1                                 | • Added new tutorials to Loop Best Practices on page 56.  
• Added new tutorials to Interface Best Practices on page 39.  
• Updated Executing Multiple Loops in Parallel on page 87 to reflect new syntax of `ihc::launch` and `ihc::collect` functions. |
| 2020.01.27       | 19.4                                 | • Corrected the spelling of the `-ffp-contract=fast` command option in Example: Loop Pipelining and Unrolling on page 61. |
• Added information to Example: Specifying Bank-Selection Bits for Local Memory Addresses on page 81 to explain the difference between the element-address bank-selection bits selected with the `hls_bankbits` attribute and the byte-address bank-selection bits reported in the Function Memory Viewer in the High-Level Design Reports.  
• References to the Component Viewer have been replaced with references to the Function View of the Graph Viewer.  
• Reference to the Component Memory Viewer have been replaced with references to the Function Memory Viewer. |

**Document Revision History for Intel HLS Compiler Best Practices Guide**


<table>
<thead>
<tr>
<th>Document Version</th>
<th>Intel Quartus Prime Version</th>
<th>Changes</th>
</tr>
</thead>
</table>
| 2019.09.30       | 19.3                        | • Added Control LSUs For Your Variable-Latency MM Host Interfaces on page 53.  
• Updated Memory Architecture Best Practices on page 72 to list updated and improved tutorials and new memory attributes.  
• Split memory architecture examples for overriding coalesced memory architectures and overriding banked memory architectures into the following sections:  
  - [PRO] Example: Overriding a Coalesced Memory Architecture on page 73  
  - [STD] Example: Overriding a Coalesced Memory Architecture  
  - [PRO] Example: Specifying Bank-Selection Bits for Local Memory Addresses on page 81  
  - [STD] Example: Specifying Bank-Selection Bits for Local Memory Addresses |
| 2019.07.01       | 19.2                        | • Maintenance release. |
• Moved some content from Loop Best Practices on page 56 into a new section called Reuse Hardware By Calling It in a Loop on page 57.  
• [PRO] Revised Component Uses More FPGA Resource Than Expected on page 98 to include information about the `hls_scheduler_target_fmax_mhz` component attribute. |
<table>
<thead>
<tr>
<th>Document Version</th>
<th>Intel Quartus Prime Version</th>
<th>Changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>2018.12.24</td>
<td>18.1</td>
<td>• Updated to Loop Best Practices on page 56 to include information about function inlining in components and using loops to minimize the resulting hardware duplication.</td>
</tr>
</tbody>
</table>
| 2018.09.24       | 18.1                      | • **PRO** The Intel HLS Compiler has a new front end. For a summary of the changes introduced by this new front end, see Improved Intel HLS Compiler Front End in the Intel High Level Synthesis Compiler Version 18.1 Release Notes.  
• **PRO** The --promote-integers flag and the best_practices/integer_promotion tutorial are no longer supported in Pro Edition because integer promotion is now done by default. References to these items were adjusted to indicate that they apply to Standard Edition only in the following topics:  
  — Component Fails Only In Simulation on page 96  
  — Datatype Best Practices on page 94 |
| 2018.07.02       | 18.0                      | • Added a new chapter, Advanced Troubleshooting on page 96, to help you troubleshoot when your component behaves differently in simulation and emulation, and when your component has unexpectedly poor performance, resource utilization, or both. |
| 2018.05.07       | 18.0                      | • Starting with Intel Quartus Prime Version 18.0, the features and devices supported by the Intel HLS Compiler depend on what edition of Intel Quartus Prime you have. Intel HLS Compiler publications now use icons to indicate content and features that apply only to a specific edition as follows:  
  • **PRO** Indicates that a feature or content applies only to the Intel HLS Compiler provided with Intel Quartus Prime Pro Edition.  
  • **STD** Indicates that a feature or content applies only to the Intel HLS Compiler provided with Intel Quartus Prime Standard Edition.  
• Added best_practices/loop_coalesce to the list of tutorials in Loop Best Practices on page 56.  
• Added interfaces/explicit_streams_packets_ready_empty to list of tutorials in Interface Best Practices on page 39.  
• Revised Example: Specifying Bank-Selection Bits for Local Memory Addresses on page 81 with improved descriptions and new graphics that reflect what you would see in the high-level design reports (report.html) for the example component.  
• Updated Example: Overriding a Coalesced Memory Architecture on page 73 with new images to show the memory structures as well as how the FPGA resource usage differs between the two components |
| 2017.12.22       | 17.1.1                     | • Added Choose the Right Interface for Your Component on page 40 section to show how changing your component interface affects your component QoR even when the algorithm stays the same.  
• Added interface overview tutorial to the list of tutorials in Interface Best Practices on page 39. |
| 2017.11.06       | 17.1                      | Initial release.  