The first lab assignment is a warmup lab where you will use a standard-cell-based ASIC toolflow to quantitatively analyze the area, energy, and performance of various integer multiplier designs. You will push three given baseline designs through the ASIC toolflow: a fixed-latency iterative design that always takes the same number of cycles, a variable-latency iterative design that exploits properties of the input operands to reduce the execution time, and a simple single-cycle design. The alternative design is a pipelined integer multiplier that is parameterized by the number of pipeline stages. You are required to implement both a cycle-level (CL) and register-transfer-level (RTL) model of the alternative design, verify the designs using an effective testing strategy, push all designs through the ASIC toolflow, and perform an evaluation comparing the various implementations. As with all lab assignments, the majority of your grade will be determined by the lab report. You should consult the course lab assignment assessment rubric for more information about the expectations for all lab assignments and how they will be assessed.

This lab is designed to give you experience with the PyMTL hardware modeling framework and with a state-of-the-art standard-cell-based ASIC toolflow for quantitatively analyzing the area, energy, and timing of a design. Your experience in this lab assignment will create a solid foundation for completing the rest of the lab assignments and the five-week design project.

This handout assumes that you have read and understand the course tutorials. To get started, you should access the course computing resources and perform the following steps:

```bash
% source setup-ece5745.sh
% mkdir -p ${HOME}/ece5745
% cd ${HOME}/ece5745
% git clone git@github.com:cornell-ece5745/lab1-groupXX.git
% cd lab1-groupXX/sim
% mkdir build
% cd build
% py.test .. --tb=no
```

Where XX is your lab group number. All of the tests should pass except for the tests related to the pipelined multiplier you will be implementing in this lab.

For this lab you will be working in the lab1-imul subproject which includes the following files:

- `IntMulMag.py`: BitStruct message types for multiplier
- `IntMulFL.py`: functional-level multiplier
- `IntMulFixedLatCL.py`: CL model of fixed-latency iterative multiplier
- `IntMulFixedLatRTL.py`: RTL model of fixed-latency iterative multiplier
- `IntMulVarLatCL.py`: CL model of variable-latency iterative multiplier
- `IntMulVarLatCalcShamtRTL.py`: RTL model to calculate shift amount
- `IntMulVarLatRTL.py`: RTL model of variable-latency iterative multiplier
1. Introduction

In ECE 4750, you learned how to evaluate baseline and alternative designs according to several key metrics: execution time (in cycles), area, energy, and cycle time. Using RTL modeling, we were able to quantitatively measure the execution time (in cycles) for a given input dataset for each design, but we relied on qualitative estimates to analyze the area, energy, and cycle time of each design. In this lab, we will learn how to use a standard-cell-based ASIC toolflow to quantitatively measure the area, energy, and cycle time of a design; we can then combine these measurements with our understanding of the cycle-level performance of each design to conduct a much more detailed design-space exploration.

We have provided you with a functional-level model of an integer multiplier as shown in Figure 1. You can find this functional-level (FL) implementation in IntMulFL.py and the associated unit tests in IntMulFL_test.py. This implementation always takes a single-cycle and uses a higher-level modeling style compared to the cycle-level (CL) and register-transfer-level (RTL) modeling you will be using in your designs. The interface for the FL model and indeed for all of the integer multiplier designs is identical. Each multiplier should take as input a 64-bit BitStruct with two fields containing the 32-bit operands. The message type is defined in IntMulMsg.py. All implementations should treat the input operands and the result as two’s complement numbers and thus should be able to handle both signed and unsigned multiplication. In addition, both the input and output interfaces use the val/rdy micro-protocol to control when new inputs can be sent to the multiplier and when the multiplier has a new result ready to be consumed. Another module should be able to send request messages to the multiplier and never explicitly be concerned with how many cycles the implementation takes to execute a multiply transaction and return the response message.

2. Baseline Design

There are three baseline designs for this lab assignment: a fixed-latency iterative integer multiplier, a variable-latency iterative multiplier, and a simple single-cycle multiplier. Both CL and RTL models of each design are provided for you.

![Figure 1: Cloud Diagram for Integer Multiplier FL Model](image-url)

Input and output use latency-insensitive val/rdy interfaces; input/output queue interface adapters simplify interacting with these interfaces. The input message includes two 32-bit operands; output message is a 32-bit result. Clock and reset signals are not shown.
2.1. Fixed-Latency Iterative Integer Multiplier

The fixed-latency iterative multiplier will always take approximately 32 cycles. Figure 2 illustrates the fixed-latency iterative multiplication algorithm using “pseudocode” which is really executable Python code. Each iteration we check the least significant bit of the \( b \) operand; if this bit is zero then we shift the \( b \) operand to right and the \( a \) operand to the left, but if this bit is one then we add \( a \) to the result before shifting. Each iteration is essentially calculating a partial product for the multiplication. Note that we say “approximately” because there may be a few extra cycles of overhead in handling the val/rdy micro-protocol (although a more optimized design can remove this overhead).

Figure 3 shows a cloud diagram for the fixed-latency iterative multiplier CL model. The CL model uses interface adapters to simplify interacting with the val/rdy interfaces. The CL model accepts a request, calculates the result in the first cycle, and then simply waits for 35 cycles using a counter before returning the response. We chose 35 cycles to better approximate the cycle-level performance of the RTL model which includes three cycles of overhead. The PyMTL code for the fixed-latency iterative multiplier CL model is in `IntMulFixedLatCL.py`.

The fixed-latency iterative multiplier RTL model uses a control/datapath split. The datapath for the fixed-latency design is shown in Figure 4. The blue signals represent control/status signals for communicating between the datapath and the control unit. The datapath model uses a structural design style by instantiating a child module for each of the blocks in the datapath diagram. The control unit for the fixed-latency design uses the simple finite-state-machine shown in Figure 5. The IDLE state is responsible for consuming the message from the input interface and placing the input operands into the input registers; the CALC state is responsible for iteratively using adds and shifts to calculate the multiplication; and the DONE state is responsible for sending the message out the output interface. Note that if you implement the FSM exactly as shown each multiply should take 35 cycles: one for the IDLE state, 32 for iterative calculation, one cycle to realize the calculation is done, and one cycle for the DONE state. The extra cycle to realize the calculation is done is because we are not using Mealy transitions from the CALC to DONE states. We need to wait for the counter to reach 32 and then we move into the idle state. The control unit is structured into three parts: a sequential concurrent block for just the state element, a combinational concurrent block for state transitions, and a combinational concurrent block for state outputs. The PyMTL code for the fixed-latency iterative multiplier RTL model is in `IntMulFixedLatRTL.py`.

---

**Figure 2:** Fixed-Latency Iterative Multiplication Algorithm – Assumes a and b are 32-bit Bit objects. Always uses 32 shifts and subtractions to calculate the partial-products over time. This is executable Python code.

**Figure 3:** Cloud Diagram for Fixed-Latency Iterative Integer Multiplier CL Model – CL model uses input/output queue adapters to simplify interacting with val/rdy interfaces. A counter delays the response to create a cycle-level timing model.

```python
def imul_fixed_algo( a, b):
    result = Bits( 32, 0 )
    for i in range(32):
        if b[0] == 1:
            result += a
        a = a << 1
        b = b >> 1
    return result
```
2.2. Variable-Latency Iterative Integer Multiplier

The fixed-latency iterative integer multiplier will always take 35 cycles to compute the result. While it is possible to optimize away the cycle to realize the calculation is done and to eliminate the IDLE/DONE states, fundamentally this algorithm is limited by the 32 cycles required for the iterative calculation. The variable-latency iterative multiplier takes advantage of the structure in some pairs of input operands to improve performance and energy efficiency. Figure 6 illustrates the variable-latency iterative multiplication algorithm using “pseudocode”. If the b operand has many consecutive zeros we don’t need to shift one bit per cycle; instead we can shift the B register multiple bits in one step and directly jump to the next required addition. The calc_shamt function calculates the shift amount based on the number of trailing zeros in the b operand. Various different implementations of calc_shamt are possible: considering more bits will improve the performance but likely increase area and energy.

```python
def imul_var_algo( a, b ):
    result = Bits( 32, 0 )
    while b != 0:
        if b[0] == 1:
            result += a
            shamt = calc_shamt( b )
            a = a << shamt
            b = b >> shamt
        else:
            counter = 32
            shamt = calc_shamt( b )
            a = a << shamt
            b = b >> shamt
    return result
```

Figure 6: Variable-Latency Iterative Multiplication Algorithm – Assumes a and b are 32-bit Bits objects. Shifts by more than one to skip over sequences of zeros in the b operand. Various different calc_shift functions are possible. This is executable Python code.
The cloud diagram in Figure 3 is still applicable for the variable-latency iterative multiplier CL model; the only difference is how we set the counter. We use a relatively optimistic timing model where the counter is set to the number of ones in the b operand; this is the best case given the simple algorithm shown in Figure 6. Our RTL model is more conservative, and we could eventually refine our CL model to more accurately capture the cycle-level performance of the RTL model. The PyMTL code for the variable-latency iterative multiplier CL model is in `IntMulVarLatCL.py`.

The variable-latency iterative multiplier RTL model is very similar to the fixed-latency RTL model with a couple key differences. The datapath for the variable-latency design is shown in Figure 7. Notice that we have added a a new model that takes the b operand and input and calculates the variable shift amount for both the left and right shifters. The PyMTL code for this new model is in `IntMulVarLatCalcShamt.py`. Refactoring this logic into a separate model enables unit testing the logic before integrating it into the overall design. The control unit for the variable latency design uses the simple finite-state-machine shown in Figure 8. The only difference from the fixed-latency design is that we finish the calculation when the b operand is zero. The PyMTL code for the variable-latency iterative multiplier RTL model is in `IntMulVarLatRTL.py`.

### 2.3 Single-Cycle Integer Multiplier

While the iterative multipliers will likely require minimal area, they also require many cycles to calculate each result. As a point of comparison, we will also consider a simple single-cycle integer multiplier. An CL model is not necessary, since the integer multiplier FL model already captures the single-cycle timing behavior. Figure 9 shows the single-cycle RTL model. We use registered inputs for both the operands and the valid bit. If the response interface is not ready, then we stall the multiplier by disabling the register enable signals and combinatorially propagating the response.
ready signal to the request ready signal. For the actual multiplier we simply use the \(*\) operator; we will rely on the ASIC toolflow to choose the most appropriate multiplication hardware. The PyMTL code for the single-cycle multiplier RTL model is in `IntMulScycleRTL.py`.

3. Alternative Design

For your alternative design, you should implement a pipelined integer multiplier with the same val/rdy interface used in the baseline designs. You should implement both a CL and RTL model. Both models should be parameterized by the number of pipeline stages; you can assume that the number of stages will be a power of two between 1 and 32 inclusive.

The pipelined integer multiplier CL model should capture the cycle-approximate performance of the RTL model. If the pipeline in your RTL model is “inelastic” (i.e., if the response is not ready then all transactions in the pipeline stall) your CL model should capture this. However, if the pipeline in your RTL model is “elastic” (i.e., if the response is not ready then transactions in the pipeline can continue down the pipeline by essentially “squeezing” out bubbles in the pipeline) then again your CL model should capture this. We recommend you use CL interface adapters to simplify interacting with the val/rdy interfaces. We also recommend you use a standard Python `deque` container with each item in the `deque` modeling a stage in the pipeline. If a pipeline stage contains a valid result, then the corresponding item in the `deque` will be a valid `Bits` object. If a pipeline stage contains an invalid result, then the corresponding item in the `deque` could be the `None` constant. The basic approach for a CL model of an “inelastic” pipeline might involve the following four steps: (1) if the response queue is full then do nothing since the pipeline is supposed to stall; (2) if the response queue is not full then pop a value from the `deque` and if this value is not `None` then enqueue the value onto the response queue; (3) if the response queue is not full and the request queue is not empty, then perform the multiplication and append the result to the `deque`; and (4) if the response queue is not full and the request queue is empty, then append a `None` value to the `deque`. Note that if you do not append a `None` value when the request queue is empty, the number of items in the `deque` will not equal the number of pipeline stages and your CL model will poorly model the cycle-level performance of the pipelined multiplier.

The pipelined integer multiplier RTL model should use a child model that implements a single “partial product step” of the fixed-latency iterative multiplication algorithm. Each partial product step would involve three inputs: one for \(a\), \(b\), and `result` and three outputs also for \(a\), \(b\), and `result`. Each partial product step will perform two shifts, an addition, and then mux the result of the addition conditionally based on the least-significant bit of the \(b\) input. As always, you should unit test the partial product step model, and potentially even build a composition of a few steps for integration testing. Once you are confident in your step model, implement the parameterized RTL model by using elaboration to insert the pipeline registers between the steps. Note that there are many optimizations you might want to try that may lead to higher quality hardware; you should feel free to explore different approaches if you like.
4. Testing Strategy

We have provided you with all of the unit tests for the integer multipliers, although you will need to add your own test scripts for any new models you create. You are also certainly welcome to add more tests for your alternative design if you would like. Most of the unit tests are defined in IntMulFL_test.py; the other test scripts reuse these unit tests as much as possible. The following commands illustrate how to run all of the tests for the entire project, how to run just the tests for this lab, and how to run just the tests for the FL model, fixed-latency CL model, and fixed-latency RTL model.

% cd ${HOME}/ece5745/lab1-groupXX/sim/build
% py.test .. --tb=no
% py.test ../lab1-imul --tb=no
% py.test ../lab1-imul/IntMulFL_test.py
% py.test ../lab1-imul/IntMulFixedLatCL_test.py
% py.test ../lab1-imul/IntMulFixedLatRTL_test.py

Once your alternative design passes all of the provided tests, then you should verify that your RTL model can be successfully translated using the following command:

% cd ${HOME}/ece5745/lab1-groupXX/sim/build
% py.test ../lab1-imul/IntMulNstageRTL_test.py --test-verilog

5. Evaluation

Once you have verified the functionality of your alternative design, you should then use the provided simulator to evaluate the cycle-level performance of the various multiplier designs. You can run the simulator for the first baseline design like this:

% cd ${HOME}/ece5745/lab1-groupXX/sim/build
% ../lab1-imul/imul-sim --impl cl-fixed --input small --stats
% ../lab1-imul/imul-sim --impl rtl-fixed --input small --stats

The simulator will display the total number of cycles to execute the specified input dataset. You should study the line traces (with the --trace option) and possibly the waveforms (with the --dump-vcd option) to understand the reason why each design performs as it does on the various patterns. For your pipelined multiplier, you can use the --nstages option to set the number of pipeline stages. Be careful to factor out the pipeline startup overhead for your pipelined multipliers for a fair comparison.

Once you have explored the cycle-level performance of the various multiplier designs, you should push the RTL models through the standard-cell-based ASIC toolflow to quantify the area, energy, and timing of each design. We would like you to push the following designs through the flow:

- Fixed-latency iterative multiplier
- Variable-latency iterative multiplier
- Single-cycle multiplier
- Pipelined multiplier with one stage
- Pipelined multiplier with two stages
- Pipelined multiplier with four stages
- Pipelined multiplier with eight stages
For each design we would like to analyze the power/energy when executing three specific input datasets: small (small random numbers), large (large random numbers), and sparse (higher probability of zero bits in the input operands). You can generate the Verilog and VCD files for a given design and input dataset using the simulator like this:

```bash
% cd ${HOME}/ece5745/lab1-groupXX/sim/build
% ../lab1-imul/imul-sim --impl rtl-fixed --input small --translate --dump-vcd
```

This will generate a Verilog file with the prefix IntMulFixedLatRTL and a unique hash along with a VCD file named imul-rtl-fixed-small.verilator.vcd. You will need to update the ASIC toolflow Makefrag to point to these two files in order to push this specific design through the flow and then analyze the power of this design executing this specific input dataset. For the baseline iterative designs you can set the clock constraint to be 0.6 ns, and for the baseline single-cycle design you can set the clock constraint to be 1.5 ns. For your pipelined multiplier you will need to experiment to find a good clock constraint. If you have no negative slack you are not pushing the tools hard enough. If you have too much negative slack you are pushing the tools too hard. About 5–10% negative slack is usually reasonable.

**Acknowledgments**

This lab was created by Christopher Batten and Derek Lockhart as part of the course ECE 5745 Complex Digital ASIC Design at Cornell University.