Cache Documentation

Authors: Nikodem Gazda, Luke Saleh, Brennan Borchert, Quinlan Stewart

Table of Contents

[2 General Implementation 2](#_Toc181984411)

[2.1 Cache Operations 2](#_Toc181984412)

[2.1.1 Read Hit 2](#_Toc181984413)

[2.1.2 Read Miss 2](#_Toc181984414)

[2.1.3 Write Hit 2](#_Toc181984415)

[2.1.4 Write Miss 2](#_Toc181984416)

[2.2 Results 2](#_Toc181984417)

[3 Individual Components 2](#_Toc181984418)

[3.1 Cache Block 2](#_Toc181984419)

[3.1.1 Functionality 2](#_Toc181984420)

[3.1.2 Results 2](#_Toc181984421)

[3.1.3 Sub-component: Cache Data 2](#_Toc181984422)

[3.1.4 Sub-component: Cache Tags 4](#_Toc181984423)

[3.1.5 Sub-component: Cache Valid Bits 7](#_Toc181984424)

[3.1.6 Sub-component: Cache Hit Logic 7](#_Toc181984425)

[3.1.7 Sub-component: Cache LRU Buffer 7](#_Toc181984426)

[3.1.8 Sub-component: Cache Control 7](#_Toc181984427)

[3.2 Ram Block 7](#_Toc181984428)

[3.2.1 Functionality 7](#_Toc181984429)

[3.2.2 Results 7](#_Toc181984430)

[3.3 Miscellaneous Logic 9](#_Toc181984431)

[3.3.1 Register 9](#_Toc181984432)

[3.3.2 Delay 10](#_Toc181984433)

[4 Conclusion 11](#_Toc181984434)

# General Implementation

## Cache Operations

### Read Hit

### Read Miss

### Write Hit

### Write Miss

## Results

# Individual Components

## Cache Block

### Functionality

### Results

### Sub-component: Cache Data

#### Functionality

The Cache Data block is the physical storage for data within the cache. The data is organized into multiple ways (configurable at compile time) and the input index is used as the cache address, whose size is the number of ways specified in the cache multiplied by the number of index addresses.

The inputs for this block are the write enable, the desired way to access, the cache index, the data input, and the only output is the data output. Reads from this block are combinational/instant while the writes sequential and take 1 cycle to update.

The memory locations in the Cache Data block must be indexed with both the desired way and the index. For example, if you would like to read from bank 2 in address 0x3, you must specify 2 as the desired way and 0x2 as the desired index.

#### Results

The Cache Data testbench consists of a set of direct testcases and random testcases.

**Directed Testcases:**

In the directed testcases, we increment the data value of each bank of each index, set the write enable to 1, and wait for the rising edge of the clock (ex. bank 1 of index 2 would have a value of index\*number of banks + current bank = 9). We do this for every index and bank in the Cache Data block. The testbench itself has a reference model that mimics the behavior of the Cache Data block in an array. After all addresses have been written, we then read from every index/bank pair and compare with the data saved in the same addresses of the reference model. This testcase ensures that data that is written to the Cache Data block is maintained over longer periods of time, even as other indices and banks are written to and read from. See the figure below to see the first 5 values written to and read from the Cache Data block.

A white screen with black text

Description automatically generated

**Figure X:** Directed tests for the Cache Data block.

**Random Testcases:**

We performed 1000 random test cases with the following constrained random verification:

* Data input is all 0s 10% of the time, all 1s 10% of the time, and the rest of the values the other 80% of the time.
* Index input follows the same constraints as the data inputs.
* Way input is kept random.
* The write enable is asserted 50% of the time for 50% reads and 50% writes.
* Writes are followed by reads which must use the same index and way used for the write so that each write to a random address is checked.

A screenshot of a computer program

Description automatically generated

**Figure X:** First 5 random testcases for the Cache Data block.

In this testcase, we can see that a read from bank 1 of index 1 returns 5, which is the value that was written to that address in the directed testcase. The next operation is a write to bank 2 of index 1, and we can see the read from this index/bank pair in the next cycle return that value of 253. The remainder of the random testcases passed with no failures.

**Assertions:**

* **out\_correct\_check:** This assertion checks the data output is always equal to the contents of the reference model. This is an immediate assertion since reads from the cache data block are combinational/instant.
* **write\_check:** When the write enable is de-asserted, the write enable on the previous cycle was asserted, the index on this cycle matches the index from the previous cycle, and the way on this cycle matches the way from the previous cycle, then within the same cycle we expect that the data output is equal to the input from the previous cycle.

### Sub-component: Cache Tags

#### Functionality

The Cache Tags block is similar to the Cache Data block in that it acts as the physical storage for the tags in the cache, organized into banks and indices in the same way, with reads being combinational/instant and writes being synchronous as well.

The inputs for this block are the write enable, the desired way to access, the cache index, the target tag input, and instead of outputting the tag from just one bank, each bank from each index is output. This parallel output exists so it’s easier for the Cache Hit Logic block to determine whether there’s been a hit/miss. Reads from this block are combinational/instant while the writes sequential and take 1 cycle to update.

#### Results

The Cache Tag testbench consists of a set of direct testcases and random testcases and is very similar to the Cache Data testbench.

**Directed Testcases:**

The directed testcases for the Cache Tag block are the same, except that the printed outputs are each bank of the index. The following paragraph can be skipped if you’ve already read about the Cache Data testcase.

In the directed testcases, we increment the data value of each bank of each index, set the write enable to 1, and wait for the rising edge of the clock (ex. bank 1 of index 2 would have a value of index\*number of banks + current bank = 9). We do this for every index and bank in the Cache Tag block. The testbench itself has a reference model that mimics the behavior of the Cache Tag block in an array. After all addresses have been written, we then read from every index/bank pair and compare with the data saved in the same addresses of the reference model. This testcase ensures that data that is written to the Cache Tag block is maintained over longer periods of time, even as other indices and banks are written to and read from. See the figure below to see the first couple values written to and read from the Cache Tag block.

A screenshot of a computer program

Description automatically generated

**Figure X:** Directed tests for the Cache Tag block.

**Random Testcases:**

We performed 1000 random test cases with the following constrained random verification. If you’ve read the CRV for Cache Data, you can assume the same restrictions:

* Tag input is all 0s 10% of the time, all 1s 10% of the time, and the rest of the values the other 80% of the time.
* Index input follows the same constraints as the data inputs.
* Way input is kept random.
* The write enable is asserted 50% of the time for 50% reads and 50% writes.
* Writes are followed by reads which must use the same index and way used for the write so that each write to a random address is checked.

A screenshot of a computer program

Description automatically generated

**Figure X:** First 5 random testcases for the Cache Tag block.

In this testcase, we can see a write of tag 118 to bank 1 of index 1. In the next testcase, we read from index 1 and see all the values stayed the same from the directed testcase except for bank 1, which has changed to the input tag of 118 from the previous write. The same happens in the next random write/read pair with a value of 61 in bank 2 of index 0. The remainder of the random testcases passed with no failures.

**Assertions:**

* **These assertions are the same as the ones from Data Cache, adjusted for the tags and the full index outputs.**
* **out\_correct\_check:** This assertion checks the tag output is always equal to the contents of the reference model. This is an immediate assertion since reads from the cache tag block are combinational/instant.
* **write\_check:** When the write enable is de-asserted, the write enable on the previous cycle was asserted, the index on this cycle matches the index from the previous cycle, and the way on this cycle matches the way from the previous cycle, then within the same cycle we expect that the tag output of the specified way is equal to the input from the previous cycle.

### Sub-component: Cache Valid Bits

#### Functionality

The Cache Valid Bits block is similar to the Data and Tag blocks in that reads are combinational while writes take one cycle, however, the data stored in this block is 1 bit for each bank for each index. When the write enable is asserted, a 1 is written to the specified bank/index each time, since a first time write to each location would cause that address to become valid, and replacing any address would cause it to become valid for the new address. The valid bits are only cleared when reset is asserted.

The inputs for this block are the write enable, the desired way to access, the cache index, and the valid bit from each bank of the specified index is the output. This parallel output exists so it’s easier for the Cache Hit Logic block to determine whether there’s been a hit/miss.

#### Results

The Cache Valid testbench consists of a set of direct testcases and random testcases and is very similar to the Cache Tag testbench.

**Directed Testcases:**

The directed testcases for the Cache Valid block are the same except that each bank output has a width of only 1.

In the directed testcases, set the write enable to 1 and wait for the rising edge of the clock for every index and bank in the Cache Valid block. The testbench itself has a reference model that mimics the behavior of the Cache Valid block in an array. After all addresses have been written, we then read from every index/bank pair and compare with the data saved in the same addresses of the reference model. This testcase ensures that data that is written to the Cache Valid block is maintained over longer periods of time, even as other indices and banks are written to and read from. See the figure below to see the first couple values written to and read from the Cache Valid block.

A screenshot of a computer program

Description automatically generated

**Figure X:** Directed tests for the Cache Valid block.

**Random Testcases:**

We performed 1000 random test cases with the following constrained random verification. These slightly differ from the Cache Tag in that the reset signal is randomized as well.

* Index input is all 0s 10% of the time, all 1s 10% of the time, and the rest of the values the other 80% of the time.
* Way input is kept random.
* The write enable is asserted 50% of the time for 50% reads and 50% writes.
* The reset signal is asserted 5% of the time so that the block’s contents are cleared occasionally.
* Writes are followed by reads which must use the same index and way used for the write so that each write to a random address is checked.

A screenshot of a computer program

Description automatically generated

**Figure X:** First 5 random testcases for the Cache Valid block.

These first 5 testcases aren’t very telling since only a 1 can be written in each entry, but it does let us know that 1 is retained in the block for each index referenced. At points in the testbench, reset is asserted as well, clearing the contents of the cache and allowing for more variability.

**Assertions:**

* **out\_correct\_check:** This assertion checks the valid bit output is always equal to the contents of the reference model on all falling edges

### Sub-component: Cache Hit Logic

#### Functionality

The Cache Hit Logic block handles the logic for determining whether the incoming address is in the cache, whether it be for a read or write. This block is completely combinational.

The inputs are the target way, the full output of all the tags from the Cache Tag block, the full output of all the valid bits from the Cache Valid block, and the output of the LRU buffer in case of a miss.

The output is a bit labeled hit; when asserted, the Cache Hit Logic block is reporting a hit, and when not asserted, the logic block is reporting a miss. When there is a hit, the block also outputs the bank that corresponds to the matched tag/valid bit. When there is a miss, the block outputs the bank the LRU reports as the least recently used bank for the current index.

#### Results

The Cache Hit Logic testbench consists of a set of direct testcases and random testcases.

**Directed Testcases:**

There are 6 unique directed testcases for the Cache Hit Logic block, each repeated 5 times in our testing:

**Test 1:** all tags are 0, 1, 2, 3, target tag is 0, valid bits are 0, lru\_way is 0

This is a test case where all inputs are 0 which should return a **miss**, since all the valid bits are 0, and the output **chosen way should be 0**, since the lru\_way is 0.

**Test 2:** all tags are 0, 1, 2, 3, target tag is 0, valid bits are 0, lru way is 1

This is the same test as before with the only change being in the lru\_way from 0🡪1. This should return a **miss** as well and the output **chosen way should be 1** to reflect that the output chosen way reflects the LRU way on a miss.

**Test 3:** all tags are 0, 1, 2, 3, targt tag is 0, valid bits are 1, lru way is 2

This test should return a **hit**, since all the valid bits are asserted and there exists a tag that matches the target tag, and the **output chosen way should be 0**, since that’s the location of the tag that matches. Note that the output chosen way should not reflect the LRU way.

**Test 4:** all tags are 0, 1, 2, 3, targt tag is 1, valid bits are 1, lru way is 3

This test is the same as the previous with the only changes being in the target\_tag: 0🡪1, and the lru\_way: 2🡪3. This test should return a **hit** and the **output chosen way should be 1**. The output chosen way should not reflect the LRU way in this test case either.

**Test 5:** all inputs are 0s

This testcase is similar to Test 1 except that all input tags are 0. This case should only ever be true in the beginning of operation when no cache entries have been written to yet, and it should return a **miss** and the **output chosen way should be 0**, reflecting the LRU way.

**Test 6:** all inputs are 1s

This scenario should never happen in the actual use of the cache since the tags should never be the same in the banks of one index. Either way, we should expect a return of a **hit** and the output **chosen way should be 3**, since the Cache Hit Logic block checks each tag incrementally.

All expected outputs match the true outputs, as seen below:

A screenshot of a computer program

Description automatically generated

Figure X: Directed tests for Cache Hit Logic block.

**Random Testcases:**

We performed 100 random testcases (timed out with any more, limitation of EDA playground) with the following constrained random verification:

* **target\_tag\_ranges, lru\_way\_ranges, tags\_ranges:** make sure target tag, lru way, and tags are ‘0 10% of the time, ‘1 10% of the time, and the rest of the range 80% of the time. Since the tag input is an array, each tag is assigned independently of the others.
* **tags\_never\_same:** This constraint prevents any of the tags from being equal to the others. In the operation of the cache, the tag inputs should never be the same unless in the first couple reads/writes where the valid bits are low, since a read/write to an address with the same tag would only alter that address. This edge case is covered in the multiple iterations of the directed testcases, therefore, we prevent this behavior here.
* **target\_tag\_match:** We want to test an even split of hits and misses, therefore, we need to ensure that one of the tags matches the target tag 50% of the time. To do this, we must randomize three integers before any of the DUT signals; one to get the 50/50 chance of deciding to force a match between one of the tags and the target tag, the second to randomly choose which tag to force, and the third to get a 50% chance of that valid bit being asserted so we have an even number of hits and misses when the tags match. These random variables are solved before the DUT’s signals.

A screenshot of a computer code

Description automatically generated A screenshot of a computer program

Description automatically generated

Figure X: Randomized tests for Cache Hit Logic block.

As you can see in the randomized tests, in Test 1, no tag matches the target tag, therefore the output is a miss and the chosen way matches the lru\_way. In test 2, bank 3 matches the target tag, and the corresponding valid bit is high, meaning the output is a hit and the chosen way is bank 3. Test 3 has the same behavior as test 1. Test 4 shows a tag match with bank 0, but the corresponding valid bit is 0, resulting in a miss with the chosen way reflecting the lru\_way. Test 5 follows the same behavior as with tests 1 and 3.

The bottom output reports the number of tests, failures, and tag matches. This is meant to show that the CRV is constraining the inputs how we expect it to, with about a 50% chance of tags matching the target tag.

Assertions:

* **Miss\_check:** If a miss is reported, the chosen way must match the lru way.
* **Hit\_check:** If a hit is reported, the tag corresponding to the chosen way must match the target tag.
* **Tags\_match:** If any of the tags match the target tag, there should be a hit reported. This is kinda a duplicate of the hit\_check but I just wanted to make sure.
* **No\_tags\_match:** When no tags match the target tag, we expect a miss.

### Sub-component: Cache LRU Buffer

#### Functionality

#### Results

### Sub-component: Cache Control

#### Functionality

#### Results

## Ram Block

### Functionality

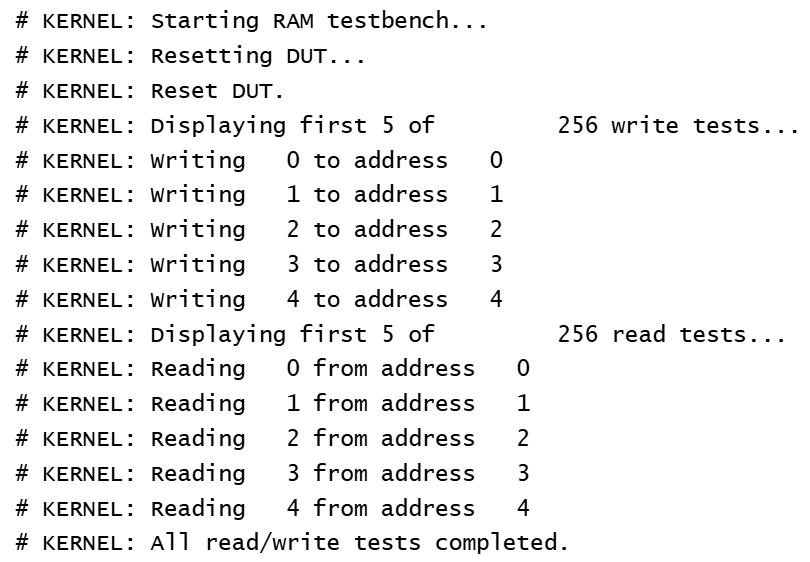
The RAM block is implemented as a DEPTH-long WIDTH-bit array of memory/register locations. The RAM takes in an address, data, and a write enable as inputs, and its only output is a WIDTH-bit data output. Both RAM reads and writes are synchronous, only updating the contents of the ram on a cycle where the write enable is asserted. Reads are only valid a cycle after write enable is not asserted.

### Results

The RAM testbench consists of a set of direct testcases and random testcases.

**Directed Testcases:**

In the directed testcases, we set the address and data input to the incrementing test number, set the write enable to 1, and wait for the rising edge of the clock (ex. Testcase 2 has addr=2, data\_in=2, we=1). We do this for every address in the RAM. The testbench itself has a reference model that mimics the behavior of the RAM in an array. After all addresses have been written, we then read from every address and compare with the data saved in the same addresses of the reference model. This testcase ensures that data that is written to RAM is maintained over longer periods of time, even as other addresses are written to and read from. See the figure below to see the first 5 values written to and read from the RAM.

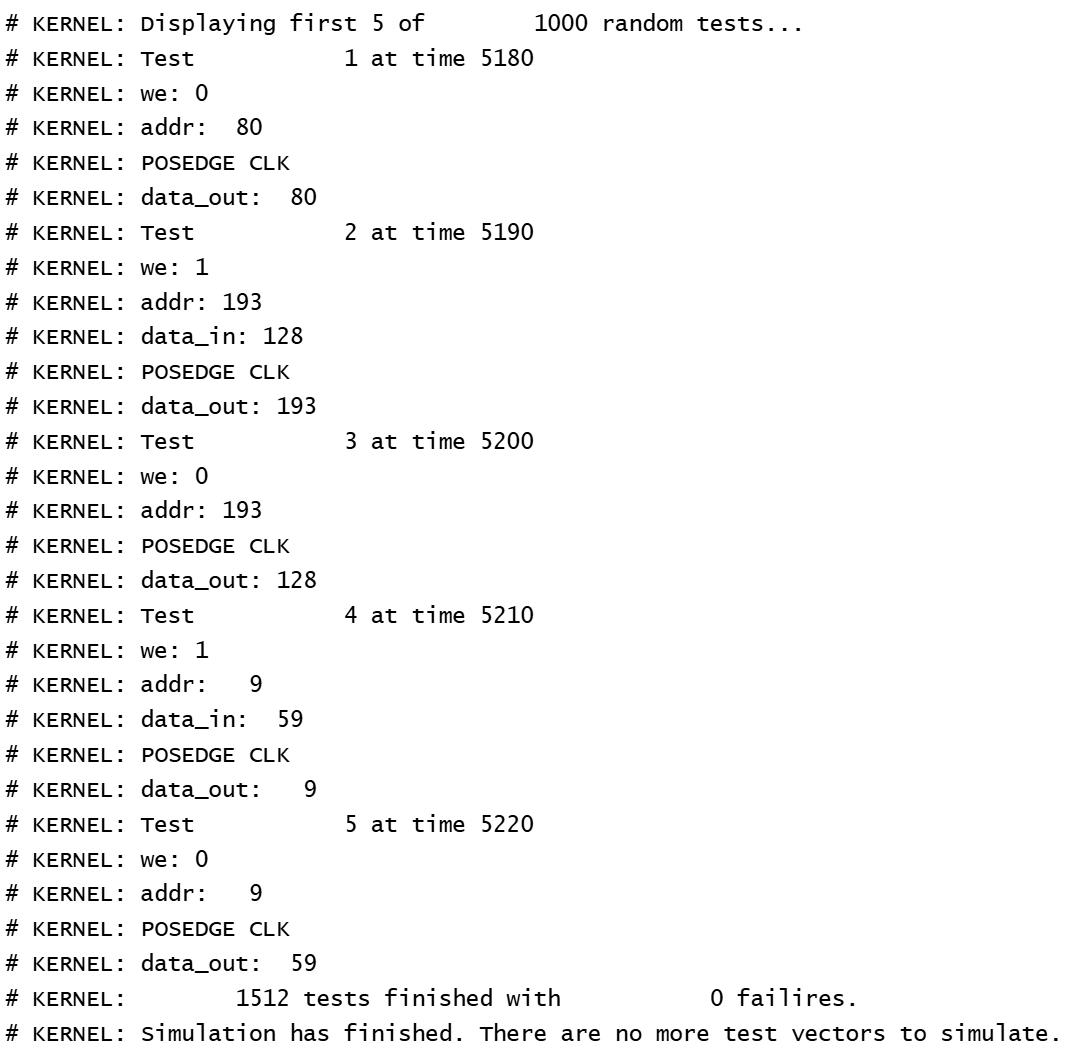


**Figure X:** Directed tests for the RAM.

**Random Testcases:**

We performed 1000 random test cases with the following constrained random verification:

* Data input is all 0s 10% of the time, all 1s 10% of the time, and the rest of the values the other 80% of the time.
* Address input follows the same constraints as the data inputs.
* The write enable is asserted 50% of the time for 50% reads and 50% writes.
* Writes are followed by reads which must use the same address used for the write so that each write to a random address is checked.



**Figure X:** First 5 random testcases for the RAM block.

In this testcase, we can see that a read from address 80 returns 80, which was the value written to that address in the directed testcase. The next operation is a write of 128 to address 193, and we can see in the next operation, which is a read from address 193, that the data out is 128, and so the data was saved correctly. The next 2 testcases exhibit the same behavior for address 9 with the data value 59.

**Assertions:**

* **out\_correct\_check:** On a read/when the write enable is de-asserted, data output is equal to the data found in the reference model using the same address.
* **write\_check:** When the write enable is asserted and followed by a read, and the address is the same for the read as for the write, then we wait a cycle for the read to happen and check that the input for the write was the data output for the read.

## Miscellaneous Logic

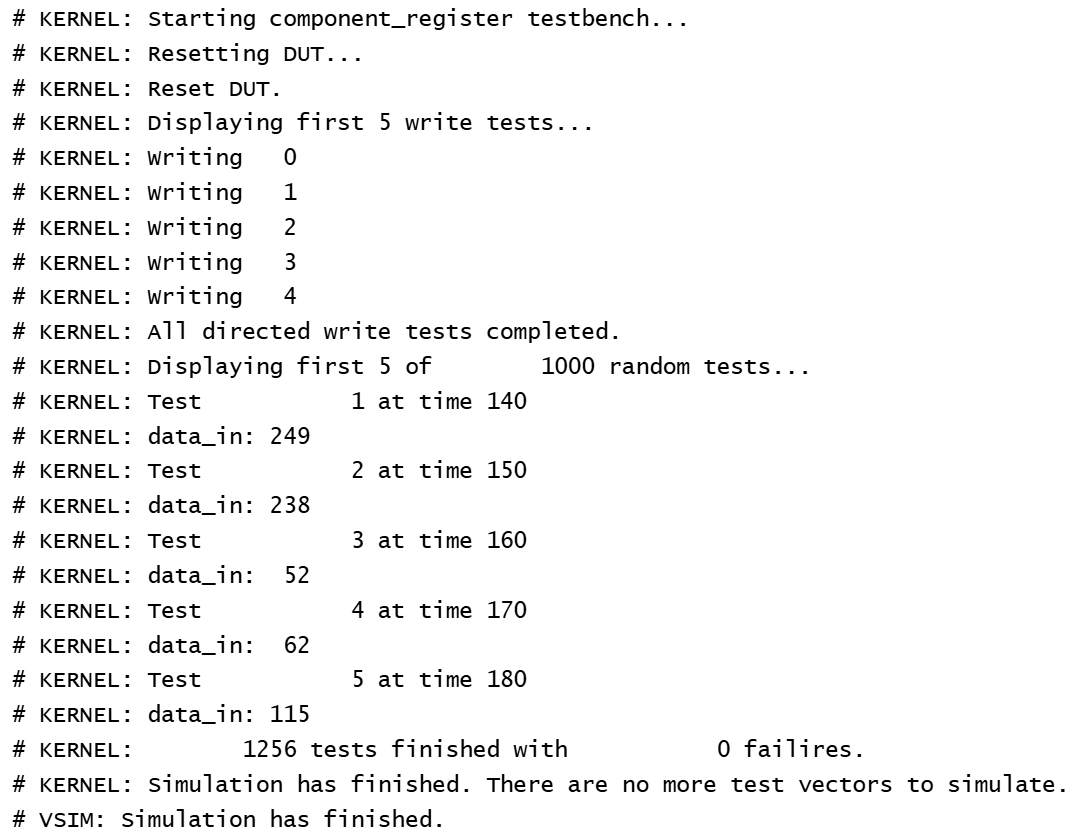
### Register

#### Functionality

The register entity is a standard register with a WIDTH-bit input and output with no enable.

#### Results

To test the register, we first did a set of directed tests where we assigned an 8-bit register each possible value from 0 to 255. See the figure below for the first 5 directed tests. We also did a set of 1000 randomized tests, with the data inputs being constrained to all 0s 10% of the time, all 1s 10% of the time, and all values in-between for the remaining 80% of the time.



**Figure X:** Testbench results for the register component.

**Assertions:**

The assertions for this testbench are just checking that on each clock, the data output from the register is the data that was on the input line in the previous cycle.

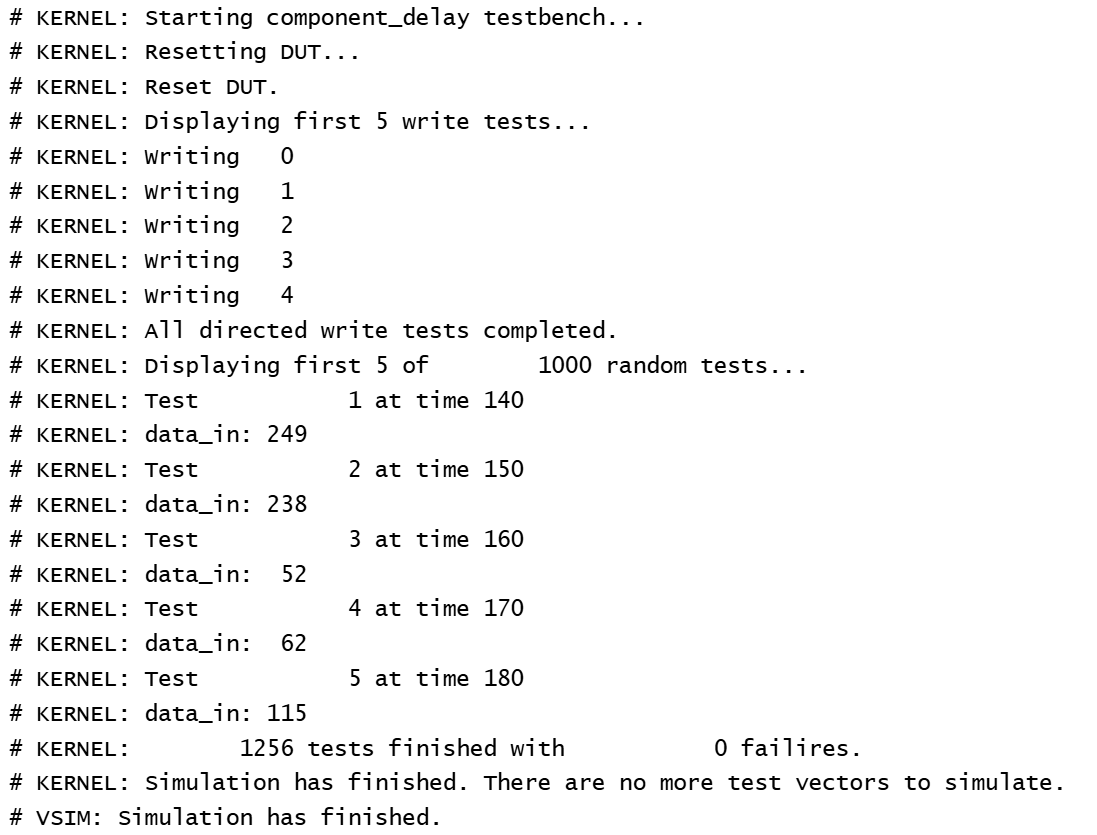
### Delay

#### Functionality

The delay entity is a CYCLE-long array of WIDTH-bit register components.

#### Results

The delay testbench is similar to the register testbench since the delay component is just a string of registers. We did the same directed tests from 0 to 255 and the same set of 1000 randomized tests, with the data inputs being constrained the same way.



**Figure X:** Testbench results for the delay component.

**Assertions:**

The assertions for this testbench are similar to the register testbench’s checking that on each clock, the data output from the register is the data that was on the input line in the previous number of cycles specified by the length of the delay, which is 4 cycles in this testbench.

# Conclusion

This was a fun and challenging project. During this project, we learned that it’s much easier to implement a design by first very clearly specifying the functionality and ONLY then implementing it in code. This is how it’s done in the industry, but we underplayed the complexity of the design and thought we could jump straight into the implementation. Since we realized we would have to go back and define the exact functionality, the design got confusing and we only started making progress once everything was set in stone.

This project was both a good practice of our understanding of caches/the algorithms that help them work and of hardware design/verification.