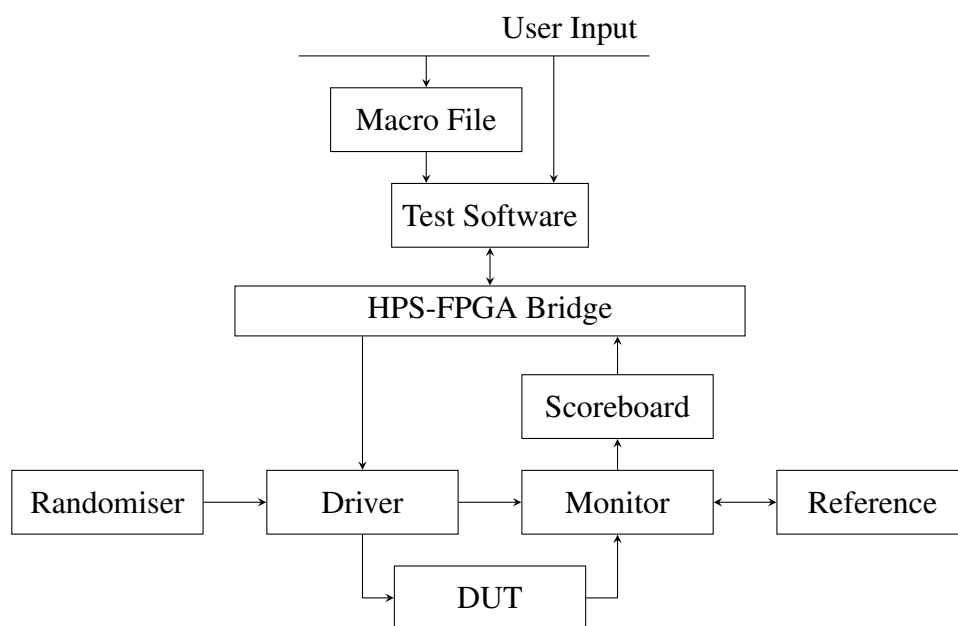


Imperial College London

Department of Electrical and Electronic Engineering

Final Year Project Report



Project Title: **An Extensible Framework for
At-speed Evaluation of Arithmetic Hardware**

Student: **Zifan Wang**

CID: **01077639**

Course: **EEE4**

Project Supervisor: **Dr. James J. Davis**

Second Marker: **Dr. Christos Bouganis**

Acknowledgements

Dr. James J. Davis for guidance

Should this be in references or here? Junjie Lu for providing designs used during testing

Maybe the testing volunteer as well?

Abstract

In this project, we aim to provide an customisable evaluation framework for arithmetic hardware. Initially, the project is conceived to perform at-speed testing of a set of newly designed high-radix online arithmetic units. The key benefits of this exotic configuration are its high throughput and ability to fail gracefully. As such, the testbench is designed to have a high maximum bandwidth and a method of monitoring the precision of the errors. However, we soon realised that in general, researchers would build their own ad-hoc testbench on FPGAs when experimenting with a new operator design. To improve the efficiency of their research and developments, we propose a flexible evaluation system for arithmetic units with this project. The framework includes both the testing software and the hardware architecture to minimise work for the user. Using a Cyclone V SoC development board, the system is implemented to demonstrate that it is indeed easy to setup and use, and can be modified to accommodate a variety of testing situations.

Contents

1	Introduction	3
2	Background Research	4
2.1	Online Arithmetic	4
2.2	High-radix Arithmetic	5
2.3	Summary	6
3	Project Specification	7
3.1	Project Organisation	7
3.2	Deliverables	7
4	System-level Design	9
4.1	Testbench Architecture	9
4.2	User Interface	10
4.3	Hardware Choice	11
4.4	Software Choice	12
5	Hardware Design	13
5.1	Providing Test Data	13
5.2	Randomiser	15
5.3	Driver	17
5.4	Monitor	19
5.5	Simplified Clock Domains	21
5.6	Scoreboard	21
6	Software Design	23
6.1	Interface Design	23
7	System Implementation	26
7.1	Project Hierarchy	26

8	Hardware Implementation	29
8.1	Randomiser	29
8.2	Driver	29
8.3	Monitor	30
8.4	Scoreboard	32
8.5	Summary	34
9	Software Implementation	36
10	Testing	38
10.1	Functional Correctness	38
10.2	Maximum Frequency	39
10.3	Out-of-the-box Testing	39
11	Evaluation	42
11.1	Product Metrics	42
11.2	Project Metrics	43
12	Conclusion	44
13	Further Work	45
13.1	Verilog Preprocessor	45
13.2	Automatic Delay Reconfiguration	45
A	User Guide	47
A.1	Introduction	47
A.2	Configuring Hardware	47
A.3	Setting up the Board	49
A.4	Running Tests	49

Chapter 1

Introduction

With the right number representation system, it is possible to perform arithmetic operations MSD first. Consequently, these online arithmetic operators are attractive for hardware implementation in both serial and parallel forms. When computing digits serially, they can be chained such that subsequent operations begin before the preceding ones complete. Parallel implementations tend to be most sensitive to failure in their LSDs, making them more friendly to overclocking than their LSD first counterparts, for which the opposite is true. In the past, online operators have typically been implemented in binary. Although Radix-2 modules are the simplest to design and has the shortest cycle time per digit, it has the highest online delay and requires the largest number of cycles to complete calculations [20]. As such, the choice of binary is not absolute.

The initial goal of this project is thus to build a testbench that can investigate the operators' suitability for FPGA implementation and examine the resultant tradeoffs between performance, area and power. However, after some time researching and working on the project, we realised that the testbench can be extended to a more general testing framework with some effort. This makes the project much more meaningful in the long term, as researchers working on other arithmetic units can also utilise this testbench after some configuration. The focus of the project thus shifted to delivering a customisable and extensible verification system while retaining the at-speed testing capabilities needed for the starting goal.

Maybe include chapter numbers. In this report, we will first discuss the motivations of investigating high-radix online arithmetic hardware on FPGAs. Following which the design of the evaluation framework will be put forth, and the design process of each individual module will be examined in detail. After determining the preferred designs, we will present how each module were built on a FPGA development board. With the implementation complete, the framework itself needs to be evaluated to see if it fulfilled its purposes. This is done with an out-of-the-box testing, where a volunteer unfamiliar with this framework is tasked to evaluate provided designs, to see if the framework is as user-friendly and customisable as we designed it to be. The results of this test will be subsequently analysed, and the report will conclude after proposing a few ideas on further improving the product.

Chapter 2

Background Research

2.1 Online Arithmetic

Traditional arithmetic operators have two common characteristics. Firstly, their order of operation may be different depending on the operation itself. A traditional adder, parallel or serial, generates its answers from the LSD to the MSD. A traditional divider design, on the other hand, generates its answer from the MSD to the LSD [3] [16].

Due to this inconsistency, arithmetic operators may be forced to compute word-by-word, waiting for all digits to finish in the previous operator before the next can start [23]. Therefore, if a divider follows an adder, the divider has to wait until the adder has completed its computation before it can begin its own.

The other commonality of traditional designs is that their precisions are specified at design-time. Once built, a 32-bit adder always adds 32 bits together, adding 16-bit numbers usually involves masking the unused bits. A possible way of making it less inefficient would be using SIMD instructions [6], splitting a large register into a few smaller ones, to execute the same instruction on them in parallel. This, however, has the tradeoff of being harder to program, and the applications must have sufficient parallelism to exploit.

Online arithmetic does not suffer from the first issue as it performs all arithmetic operations from MSD first [8] [9]. Furthermore, pipelining can be used with online serial arithmetic operators. Thus the output digit of an earlier operation can be fed into the next operator before the earlier one completes its computation.

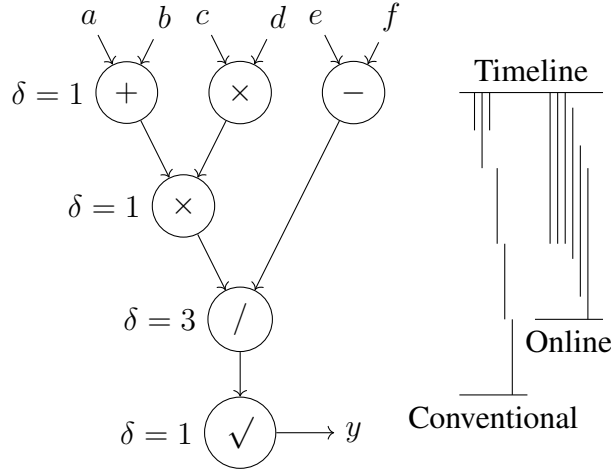


Figure 2.1: Computing $y = \sqrt{(a+b)cd/(e-f)}$ with serial online operators [8]

As illustrated in figure 2.1, while each individual operation may take longer than its conventional counterpart, online arithmetic can provide a speedup if the operators are chained in serial. In addition to the tradeoff in time, individual online arithmetic operators also uses more memory. To perform all computation from the MSD to the LSD, the use of a redundant number system is compulsory. However, this redundancy also has its advantage in making the operators scalable. The time required per digit can be made independent of the length of the operands [21].

A recently proposed architecture allows the precision of online arithmetic to be controlled at runtime [23]. Traditionally, this runtime control was restricted due to the parallel adders present in the multipliers and dividers. This architecture reuses a fixed-precision adder and stores residues in on-chip RAM. As such, a single piece of hardware can be used to calculate to any precision, limited only by the size of the on-chip RAM.

The way online arithmetic alleviates the second problem of fixed precision falls out directly from its MSD-first nature. Suppose the output of a conventional ripple adder is sampled before it has completed its operation. In this case, the lower digits would have been completed, but the carry would not have reached the higher ones. This means the error on the result would be significant, as the top bits were still undetermined [17]. However, if the output of a parallel online adder is sampled before its completion, the lower bits would be the undetermined ones. This means the error of the operation would be small. With overclocking, online arithmetic operators fail gracefully, losing their precision gradually from the lowest bits first. Thus, it allows for a runtime tradeoff between precision and frequency [18].

2.2 High-radix Arithmetic

Conventional designs of arithmetic operators use binary representations. The additional concerns of high-radix operators did not provide justifiable improvements as clock speed of processors kept in-

creasing. In recent years, the clock speed increase effectively ended, and semiconductor dies shrunk to extremely small sizes. This means the relative processing time available in a clock period increased. This enabled and drove the desire for accomplishing more per clock cycle, and high-radix arithmetic is one of them. It has been shown that, high-radix offers power saving and/or reasonable speedups to the arithmetic operations [4] [2] [5].

However, the savings are not without trade-offs. If the radix chosen is not a power of 2, then this trade-off can become unfavourable if the specification requires much I/O and little computation. This is because overhead of radix conversion would be significant [22]. It is also unwise to use high-radix representations when the numbers are unusually small, thus making the savings offered by the high-radices negligible [4]. The radix also cannot be too high, as the time in a clock period is still limited, if there is too many logic gates for the signal to propagate through, it might become the critical path and slow down the overall design.

The construct of FPGAs might make high-radices more attractive than it is on ICs. As FPGAs contain small fast carry-ripple adders, high-radix adders may be able to exploit them to obtain significant speedups [11].

2.3 Summary

Using high-radix number representations for online arithmetic is a relatively novel concept. While there has been some research with similar premises [13] [14], We take a more direct approach with this project by implementing custom operators made for high-radix online arithmetic on an FPGA. This will provide empirical results on the method, and will hopefully reveal practical insights along the way.

Furthermore, benchmarking this exotic arithmetic system with popular FPGA applications such as neural networks would be interesting, as there is not much precedence for it.

Chapter 3

Project Specification

3.1 Project Organisation

This project is a part of a larger project investigating the effect of using high-radix number representation with online arithmetic operators. The overarching aim involves implementing such a system on an FPGA and quantifying its performance improvements. This is achieved through two individual projects, vertically split from the enveloping project. One shall design the arithmetic operator modules, while the other shall design a system from the top-level to test and evaluate these operators. This project deals with the system-level issues.

As this project progresses in parallel with the designing of the operator modules, it is necessary to decouple the two projects so that, being individual projects, they can be evaluated individually. The success of one project should not be restricted by the status of the other. To this end, the goal of the system-level design is more focussed on its functionalities and robustness. This relationship and its effect on the evaluation will be examined further in the evaluation chapter of this report.

To ensure the two products will work together once they are both complete, a common interface is agreed upon. The interface will be done using Qsys. The unit-level project will build different operators, which can have varying arithmetic functions and designs. These can be packaged into individual Qsys modules, as adders, multipliers, or dividers. Alternatively they can also be delivered as a single module taking two operands and an instruction that is one of the four basic arithmetic operations. These will then become the DUTs of the testbench.

3.2 Deliverables

At the end of the project, the system should be able to perform the following:

1. Connect to the arithmetic modules as its input;

2. Generate and run tests on these modules;
3. Vary the frequency of the FPGA;
4. Evaluate its performance.

Chapter 4

System-level Design

4.1 Testbench Architecture

The design of the verification system is the major engineering challenge of this project. While there have been many similar performance analyses done on hybrid SoCs before, each of them used their own, usually ad hoc, testbench design [17] [12]. As such, most testbench are not designed to be scalable or portable, serving only what they are built for. In this project, I shall use a generic structure inspired by that of an agent in Universal Verification Methodology (UVM).

Before UVM, integrated circuit designs were verified with methodologies developed independently by stimulator vendors such as Cadence, Mentor Graphics, and Synopsys. In an effort to unify for greater efficiency, the standards organisation of the Electronic Design Automation (EDA) industry, Accellera, established UVM with support from multiple vendors. It provided a common structure for verification, with class libraries that made building and running a testbench a significantly smoother experience. The agent is a container in UVM that emulates and verifies DUTs [24]. While this project is in no position to achieve what UVM has done, I do hope that this testbench would have an easily modifiable structure that will make the process of testing similar future designs slightly simpler.

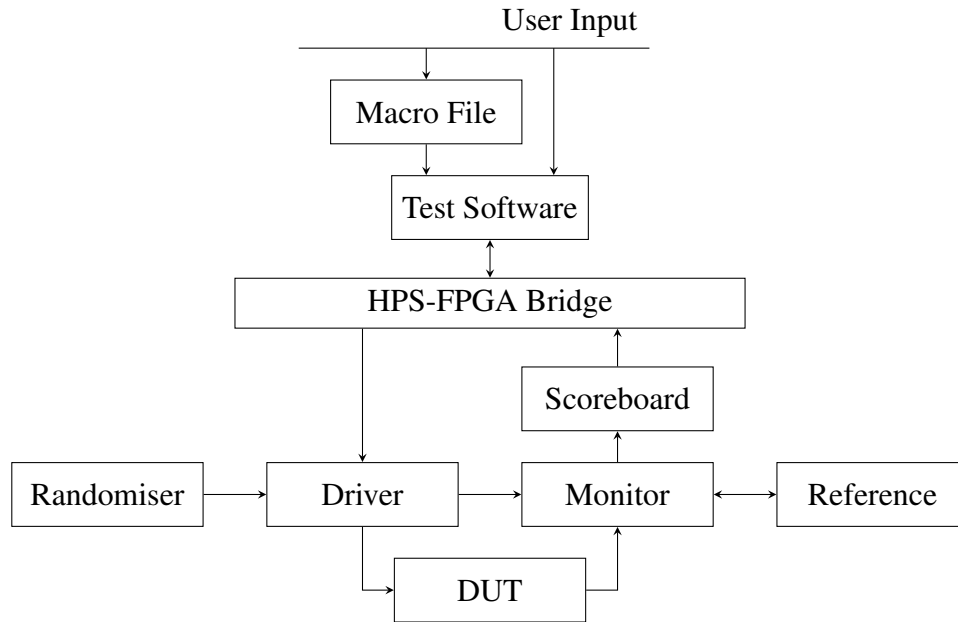


Figure 4.1: Block diagram of the testbench design

The test software running on the HPS will read instructions from either a user-written macro file, or straight from the user through command line, and sends the corresponding commands to the hardware. The driver pulls a stream of random data generated by the randomiser, and convert them to meaningful test inputs according to specification. The test output will be watched by the monitor, reporting the results to the scoreboard, which keeps track of them. The Monitor make uses of reference designs functionally identical to the DUT, which allows identification of false outputs from the DUT. Multiple instances of the reference designs means multiple test data can be processed in parallel, so that the reference design can have a relaxed frequency requirement. The scoreboard collects the results from the monitor and writes them to memory locations accessible from the other side of the bridge. These memory locations are read by the test scripts, providing results and other useful information to the user. The design and implementation of each individual hardware and software module will be elaborated in detail in the following chapters.

4.2 User Interface

For this evaluation framework to be meaningful, it has to attract users by being easy to configure and run. To achieve this, in addition to the functional designs, both hardware and software need to be designed with the user's experience in mind.

4.3 Hardware Choice

The system itself will be built on a Cyclone V SX SoC Development Board from Intel [32].

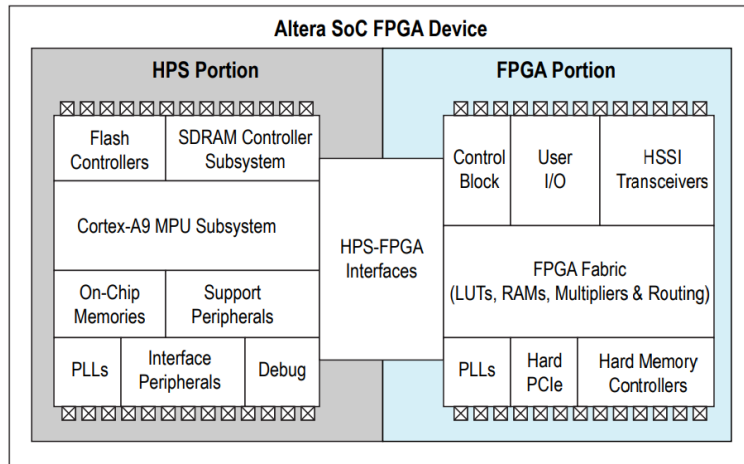


Figure 4.2: Structure of the System-on-Chip

The 5CSXFC6D6F31C6N SoC has an Arm Cortex-A9 MPCore accompanied by Intel’s 28nm FPGA fabric [25]. The FPGA is necessary for implementing the hardware design and obtaining empirical results for the project. While an FPGA without an embedded CPU will be enough for this project to work, having an Hard Processor System (HPS) on the same chip is useful as the test software can run on it. The HPS is a separate piece of hardware that distinguishes itself from a soft processor, such as the Nios II, a processor programmed onto the FPGA itself. With this additional capacity, a better user interface can thus be constructed with more detailed, on-the-fly control of the FPGA. This means setting up the testbench will only require programming the design into the FPGA, followed by running the test script on the HPS. The product will thus be self-contained. It will be more accessible as no additional setup is required for the user.

It should be noted that Xilinx offers similar boards as well. Its Zynq SoC family has a very comparable structure as they too integrate the software programmability of an Arm processor with the hardware possibility of an FPGA. For example, similar to the Cyclone V SX, Zynq-7000S features an Arm Cortex-A9 coupled with a Xilinx 28nm FPGA [36]. As such, a board like the ZedBoard [37] could be just as viable for this project.

As there are very few significant functional differences between the two brands, I shall initially explore with the Intel board, simply for its availability and my familiarity with their development tools. Due to the architectural differences between the logic elements between Xilinx and Altera FPGAs [19], the performances on the two boards are not necessarily identical. Once the project has progressed to a point where the system design is mature and tested, the Xilinx alternative can be explored as an extension.

4.4 Software Choice

The software choice follows closely with the hardware choice in this project. To develop for Intel FPGAs, Quartus has to be used. The version picked is arbitrary as there are not many functional differences between the versions that will be critical to the project. As Quartus Prime 16.0 is the version installed in the computers in the department, I will use the same version simply for convenience. This naturally means the hardware system will be built with the system integration tool that comes with Quartus – Qsys.

The Qsys software is designed to be used for integrating different hardware modules into a system. As such, it will be used as the interface for the two parallel projects.

While an HLS language could be used, in this design it suffers from a few problems and does not offer enough benefits to justify its use. Usually HLS is preferred for developing complex algorithms, because compilers can optimise them into RTL much better than humans. However, the resulting RTL would be unreadable, making directly controlling or debugging at the hardware level nearly impossible. The interfaces require detailed control of the actual hardware and the rest of the testbench has a lot of control path work and direct manipulation on the data bits. It is therefore not worth it to use HLS and as such, this design will be written in Verilog.

Other than the hardware design tools, there is some freedom of choice on the HPS side of the project. The test will be built with Python, which will be running on an Ubuntu system that is installed on the HPS. This choice is made as there are previous unrelated projects on the same development board, which means a lot of time can be saved on tedious setup works such as getting an operating system booting.

Git is used as the version control system for this project. A list of repositories on GitHub holds all files related to this project. Readme files on the repositories and the commit histories will serve as digital logbooks to this project. A copy of the user guide will also be made available on the GitHub repository.

Chapter 5

Hardware Design

5.1 Providing Test Data

In order for the driver to stress the DUT, the verification system must perform at a much higher frequency than the expected frequency of the DUT. Assuming the DUT is to run at 300MHz, to fully explore the effect of overclocking, the testbench must be able to run at double the frequency. This gives an ambitious target frequency of 800MHz. Assuming a data width of 32-bit, the target data transfer rate is then estimated to be 25.6Gbps. With this rough estimate, we can start considering different design options.

5.1.1 HPS-FPGA Bridge

As the testing is to be controlled by the HPS, the HPS-FPGA bridge will be the immediate bottleneck if the test data is to flow from HPS to FPGA. While the HPS can easily generate test data with a piece of software, there is a large amount of overhead as data crosses from one architecture to another. This overhead exists in the form of both decreased bandwidth and increased delay. Thus, it is not be sensible for the HPS to send out data during runtime.

5.1.2 Off-chip DDR SDRAM

Another thought may be to first populate the off-chip DDR SDRAM on the FPGA side, then feed that data to the DUT during test. This is already much faster than passing the data directly from HPS. The 1GB, 32-bit wide DDR3 on the FPGA side is rated at 400MHz. With double rate transfer, this gives a maximum transfer rate of 25.6Gbps.

Although using the off-chip RAM may theoretically achieve the targets, it still has its disadvantages. Firstly, the process of filling up the memory takes time. Thus, the testing would be broken up into bursts, with time in between for checking results and filling in new data. The complexity of the SDRAM interface

also requires an SDRAM controller to be used to manage SDRAM refresh cycles, address multiplexing and interface timing. These all add up to significant access latency. While it could be overcome with burst and pipelined accesses, it would further complicate the SDRAM controller. A controller is provided by Intel [27], but it would consume a non-negligible amount of the limited FPGA resources while adding unnecessary complexities to the design. Customising or building a new SDRAM controller to fit this project is possible, but needlessly time-consuming.

5.1.3 On-chip Memory

The on-chip memory is much faster and simpler to use. In comparison, this memory is implemented on the FPGA itself, and thus needs no external connections for accesses. It has higher throughput and lower latency than the SDRAM. The memory transactions can also be pipelined, giving one transaction per clock cycle. With an on-chip FIFO accessed in dual-port mode, the write operations at one end and the read operations at the other end can happen simultaneously. This feature is useful as tests are prepared and fed into the DUT, or when test results are collected and fed to the monitor.

On-chip memory is not without its drawbacks. It is volatile like SDRAM and very limited in capacity. SDRAMs can store about 1GB, while on-chip memory could only hold a few MB [26]. Volatility is not exactly of concern in this project, but its small capacity means not much test data can be held before it needs more fed in.

5.1.4 Distributed RAM / Registers

On-chip memory has a minimum latency of 1 clock cycle as the R/W access gets processed. If a even faster memory is desired, we can use LUTs or registers to store them. This option would eliminate the latency but takes up much more FPGA resources. The capacity is even more limited as LUTs are usually used for logic. There will be a significant amount of data generated during testing, and the testbench should be as lightweight as possible to allow flexibility in the DUTs. As such, distributed RAM will not be used in this project for data transfer. Registers will still be used as they are essential for many other purposes.

5.1.5 Real Time Data Generation

As seen from the analysis, the best design option here should be able to exploit the benefits of on-chip memory, and circumvent the drawback of buffering testing data generated from the HPS. Generating testing data at runtime, on the FPGA will be such a method. As arithmetic operators have a vast set of valid inputs, it is necessary to have cost-effective test generation.

A good choice here is to use random testing. With relatively low effort, random testing can provide

significant coverage and discover relatively subtle errors [7]. The main drawback of random testing is the possible lack of coverage for corner cases, for which the usual solution is to provide handwritten tests to complement it. However, as the main goal of this testbench is gauging the performance of the module, and not necessarily verifying the correctness of the module, having uncovered testing holes is acceptable during stress testing. As the project progress, special tests could be written and run separately with a relaxed timing restriction to cover the holes. It should be noted that certain corner cases may represent critical paths in the design. To combat this, the testbench provides the option to run handwritten inputs alongside random tests.

5.2 Randomiser

LFSRs are a reliable way of generating pseudorandom numbers quickly with low cost [10]. Fulfilling the design requirements, they will thus form the starting point of data generation. While it is possible for data generated to be invalid as inputs to the DUT, this should not be the case for most arithmetic units. Even if this is the case, they can be dealt by the filter in the driver. On the flip side, LFSRs go through every single possible value except for one before repeating itself in a loop, so it is more efficient than a purely random data set. The one impossible value can be covered manually, and knowing that there is an impossible value from the randomiser can be turned into a design advantage later on when we make the driver.

Following this approach, the software would only need to configure the generation at the beginning, and test data no longer needs to pass through the HPS-FPGA bridge. Thus, the testbench can provide fast and constant data to stress the DUT.

5.2.1 LFSR Configurations

While LFSRs are simple hardware modules, there are still a few design options we should explore before implementing them. To compare, we can examine an 8-bit LFSR with taps on bit [7,5,4,3].

In a Fibonacci LFSR, the taps are pulled and fed into a cascade of XOR gates. The output of the final XOR gate is then the lowest bit of the next random number. The higher bits are obtained by one left bitwise shift.

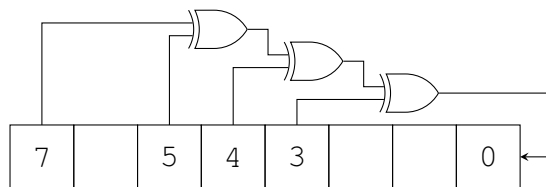


Figure 5.1: Fibonacci Configuration

In a Galois LFSR, the new bits in the taps are obtained by a XOR operation between the lowest bit and the bit on the left of each tap. The highest bit is simply the previous lowest bit, and all other bits are obtained by one right bitwise shift.

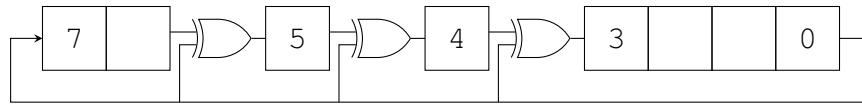


Figure 5.2: Galois Configuration

Other LFSR configurations such as Xorshift [15] exists, but they are mostly designed and optimised as pieces of software, thus being less appropriate for this design.

By examining the two configurations, we can see that Fibonacci LFSRs have to XOR multiple bits together through a cascade of 2 input XOR gates, or a single XOR gate with multiple inputs. On the other hand, Galois LFSRs have multiple XOR gates working independently. On an FPGA, the cascade of gates is usually implemented with a LUT, so while limiting LUT input to 2 might have some minor improvements, this increased delay of the Fibonacci configuration should not be obvious.

In terms of implementation, Fibonacci LFSRs are slightly easier to code if width configurability is desired. However, building a configurable Galois LFSR is only slightly more complex.

As such, we need to take a step back and examine the overall structure of the randomiser to help us make this decision.

5.2.2 Randomiser Structure

A horizontally structured randomiser uses all bits in the LFSR as output.

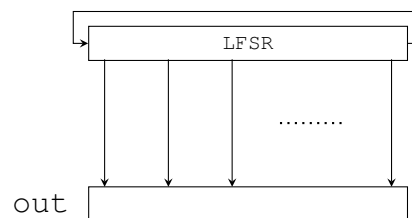


Figure 5.3: Horizontal Structure

A vertically structured randomiser uses multiple LFSRs, and combines one bit from each LFSR for its output.

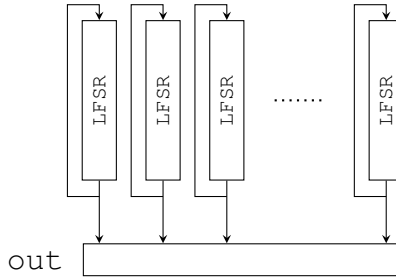


Figure 5.4: Vertical Structure

The horizontal option is easy to construct, but changing the width of the output value requires writing another wider LFSR since the tap positions would change. The vertical options is much more scalable as more or less LFSR can be instantiated depending on the required output width. As the widths of individual LFSRs are not related to the width of the output, a series cheap, 2-tap LFSRs can be used for it, making the earlier point of additional delay for Fibonacci LFSRs a non-issue.

However, the vertical structure is not without downsides. Each LFSR needs a unique seed for its initialisation, making increasing the width not completely automatic unless we also build something that generates these seeds. More importantly, the structure reduces the test efficiency introduced by LFSRs. A single LFSR will go through every non-zero value before repeating itself, the vertically arranged randomiser will have early repeats.

If we allow early repeats, then the horizontal structure can be easily scaled. This is achieved by building a long LFSR and taking a truncated version of the output value when fewer bits are required for the tests.

As such, there is no real advantage of using the vertical structure. With truncation providing the configurability in the horizontal structure, the slight advantage of the Fibonacci LFSRs in its ease to write is nullified. Having the slight advantage in terms of speed for Galois LFSRs, they will be chosen as the randomiser design in this project.

5.3 Driver

The driver should have two main types operation when feeding data into the DUT. One is the stress testing mode, where the driver tries to pushes a new piece of data into the DUT at every clock tick. The alternative is a slow manual mode, where the driver reads from the HPS-FPGA bridge and changes its output to the DUT whenever a new test point is specified by the software. The stress testing mode will expose the DUT to as much random test points as possible in the test duration. The manual mode is used when the user has a special interest in a limited list of inputs.

5.3.1 Stress Testing Mode

The vanilla way of providing data to the DUT is to for the driver to simply instantiate the same number of randomisers as the number of inputs of the DUT. Then the randomisers' outputs can be directly connected to the inputs of the DUT.

While this fast and cheap method fulfils most of the requirements of the testbench, it suffers from a few issues. One, there is no user control for the test data. If we consider the LFSR pseudo-random number sequence to be unpredictable, then the only thing the user can do will be forcing the LFSRs to initialise with same or different seeds and therefore the DUT will receive identical or different inputs. However, this level of customisation mostly meaningless.

5.3.2 Input Filtering

To introduce some level of non-trivial user control in the test data without losing speed, a filtering system is included in the driver design. The filtering system needs to be fast since this is still a stress test, yet it should provide as much utility as possible to the users.

A possible design here is to allow the user to specify a maximum and a minimum bound to the value of the input data. Each output from the randomiser is compared to these values and either sent forward if they passed or replaced if they failed the comparisons. If a higher level of control is desired, there can also be a list of invalid inputs within the bounds of validity. However, the latency can get high if the list is long and the comparisons can get slow if the high or low bound is irregular in its binary form. The replacement system can also get complex if the bounds are strict.

A better alternative to achieve this is a bit manipulation system. The user can force individual bits in the data to be cleared or set. This is less flexible than the first design, but with some tricks the user is still able to perform a great level of input control. Having only odd or even test inputs will be trivial under this system. To set maximums or minimums for the test data, the user can simply set or clear the higher bits. Certainly this imposes a strong preference to arithmetic units with regular binary representations, but most interesting designs for high-radix arithmetic units, including the ones that spawned this project, use a radix that is a power of 2. As such, the binary manipulation system will always be helpful for the user to filter out uninteresting test points or to focus in on more meaningful ones.

5.3.3 Manual Input Mode

As discussed in the randomiser section, random testing is surprisingly useful, but they do have their limits. If the user has a list of inputs that will trigger key logic paths in their design, they should be able to investigate them under this framework. This would serve as a good compliment to the random testing. The manual input mode is designed for this scenario. In this mode, the user first provides a list of numbers

when configuring the test in software, and enable the manual input mode. Then, the test software will read through this list and write them to a set of memory locations on the HPS-FPGA bridge. A simple transfer protocol will be used for the driver to read these locations and then forward them to the DUT. Due to the limitations of the bridge, the DUT cannot be fully saturated with these data. As such, each manual test input will be repeatedly sent to the DUT before the next one becomes available.

5.3.4 Synchronised Monitor Inputs

In addition to controlling what gets sent to the DUT, the driver has the responsibility to ensure that the monitor receives the test output from the DUT and the test inputs from the driver at the same time. After going through the filtering logic, the stream of test input will not only be sent to the DUT, but also sent to a shift register before reaching the monitor. The shift register will provide the delay required for the DUT to finish its operations. Since the number of cycle delay from input to output should be consistent and known by the user, the length of this delay can be configured before compiling the testbench.

5.4 Monitor

Another concern in the system design is of the different clock domains that must exist on the FPGA. Since it is not sensible to require the reference design to run as fast as the DUT, there needs to be two clock domains in the system. The initial idea is to have one domain surrounds the DUT and another that supports the rest of the control logic around the DUT. These clock frequencies can be generated with PLLs, which are provided as IP Cores in the Quartus software [28]. A clock tree will distribute them to the individual modules. Data crossing clock domains will be fed through FIFOs to prevent loss.

The proposed structure will have the bulk of the control logic running in a separate clock domain to the DUT. Only an interface with FIFOs will be running in synchronicity with the DUT. Therefore, the test controls can run at a slower frequency without bottlenecking the system, allowing the DUT to be stressed further. The problem now is to ensure the monitor can handle the stream of DUT output coming in at a higher frequency that it is running at. As the monitor needs to calculate the correct data before it can check if the DUT output is correct, it cannot keep up with the speed of the DUT. This report consider three alternatives.

5.4.1 Partial Monitors

A lightweight idea is to implement a parity checker instead of a full model inside the monitors. For example, to check an adder, the monitor can just check if the final bit with a LUT acting as a XOR gate.

Although this is reasonably fast, it cannot be extended once the DUT is faster than a parity calculation followed by a comparison. More critically, it provides no additional information once the DUT fails, and

it has a 50% rate of ignoring an error. If this is to be solved by increasing the number of bits checked, the problem returns back to its initial state. Thus this method will not be experimented.

5.4.2 Lazy Monitors

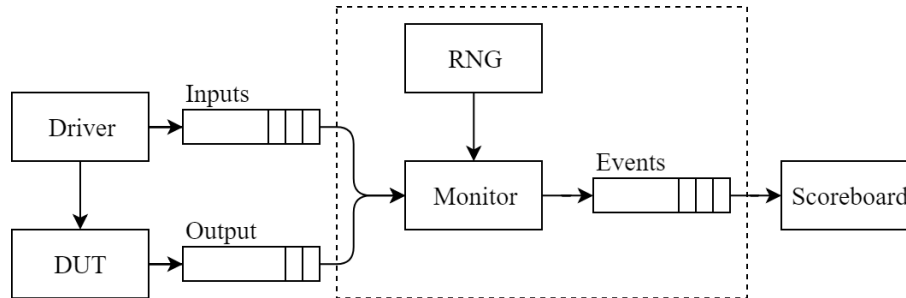


Figure 5.5: Structure of a Lazy Monitor

An more scalable alternative is to have the monitor only check a selection of data sets. For example, if the monitor is programmed to check every third test point, statistically it will make little difference to the final result. In case the DUT is aware of this and only produce correct outputs on every third operation, this process can be randomised too.

This method can be extended if the DUT get fast simply by skipping more checks, and it has the full information when it detects an error. However, this method needs the extra logic in the random controller, making the monitor slightly more complex than it probably should be.

5.4.3 Parallel Monitors

These diagrams needs to be rebuilt to better reflect the idea. Or maybe even rebuilt with the simplified clock domain thing but that could also be done in the implementation chapter.

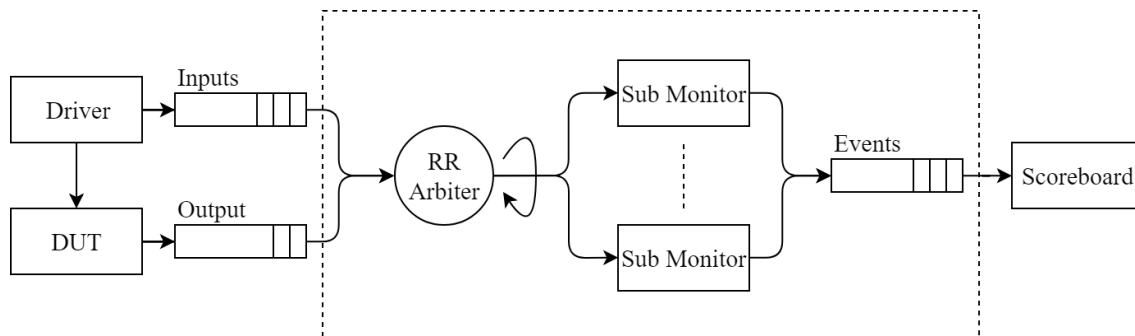


Figure 5.6: Structure of a Parallel Monitor

As the test data is uniform, the monitor can be parallelised in to a number of sub-monitors. The sub-monitors is connected to a distributor that is connected to three FIFOs. The FIFOs are the inputs and

the output of the DUT. A round robin demultiplexer distributes the data to the sub-monitors equally. The results from each sub-monitor are then sent to a single scoreboard. To avoid potential hazards, the output from the sub-monitors will be buffered before processed by the scoreboard.

This does not have data dependency on a random controller, and it can fully guarantee the correctness of the DUT. It is also scalable as more sub-monitors can be added if the DUT fills up its output buffer. As a downside, this method takes up the most FPGA resources to implement as it scales.

Comparing across the three methods, the parallel monitors will be used for this project, as it offers the best functionalities.

5.5 Simplified Clock Domains

During the implementation of the parallel monitors, it is realised that picking the parallel structure has enabled a simpler way for us to realise the hardware design regarding clock domains. So far, the assumption has been that during frequency testing, the testbench would hold on to a consistent frequency, while the frequency of the DUT is varied. However, this causes unnecessary complications as the clock ticks of the two domains would shift in and out of phase during testing, which needs to be handled with extreme care since there is heavy data moving through the domains.

Looking back at the overall structure of the framework, the slowest block is the reference designs in the sub-monitors. The priority of a reference design is to be functionally correct. Since the operation it carries out can still be complex, it should have a relaxed timing requirement in order to avoid additional burdens on the designer. This would hopefully make the reference relatively easy to produce and difficult to make mistakes on. On the other hand, the rest of the testbench should be able to operate at the speed of the DUT, as they are relatively light in terms of the logical operations that they perform. If a slow signal path arises in the system, it is also relatively harmless to sacrifice a few cycles in terms of latency to keep its maximum frequency high.

Therefore, instead of having a fast domain surrounding the DUT, the new design would have a slow domain surrounding the sub-monitor, while the rest of the testbench is clocked at the same speed as the DUT. The number of sub-monitors is a parametrised value, but it has to be an integer. As such, the slow domain will always have a clock frequency that is a factor of the frequency of the rest of the testbench. This way, they can stay in phase, which may have made the design of the monitor slightly more complicated, but vastly simplified the rest of the system.

5.6 Scoreboard

One of the simplifications has to do with the connection from the monitor to the scoreboard. Previously, there was no guarantee that the scoreboard would be synchronous with the sub-monitors producing

the results of the tests. This necessitated the event driven system, which would reduce the amount of traffic produced by the sub-monitors, and allows the scoreboard to keep track of more test data points than it normally could. Now that the scoreboard runs on the fast clock, the monitor can simply produce one piece of information for each data point, and the scoreboard would be able to handle it.

Since the precision of results coming from the DUT is of interest to us, one possible use of this bandwidth is to pass on the precision of each test data to the scoreboard. The scoreboard can then produce more detailed statistics from the test set. Instead of only counting how many points are correct and how many are wrong, it will also be able to determine more interesting values such as the maximum and minimum precision of a test set.

These output values will be stored in registers which are exposed and can be read by the HPS. If further statistics and insights are desired, it would be more sensible to perform these operations in the testing software.

Chapter 6

Software Design

6.1 Interface Design

The software should provide a layer of abstraction so that the user can run tests without worrying about too many technical details. The abstraction needs to be intuitive, but it should not compromise on the level of control given to the user. Based on the hardware design, a list of items that should be controllable by the user at the software level is drafted.

1. Operating mode (manual or auto)
2. Manual inputs in manual mode
3. Bit set/clear control in auto mode
4. Test duration
5. PLL frequency

All other variables such as the exact memory locations accessed and the binary values being read and written should be hidden away by the interface. For example, while the `freeze` signal in the scoreboard is exposed to the HPS-FPGA Bridge, the user will not have to manually assert the signal. The signal should be automatically asserted by the software before results were collected for the scoreboard, ensuring the scoreboard registers are not still counting during the read process. Allowing manual control of this signal would only require unnecessary effort from the user, thus it should be one of the items abstracted away by the software. However, the code still needs to be constructed in a way, such that an expert user attempting to modify the bridge interface should be able to do so without much trouble.

Two options were considered here.

6.1.1 Configuration File

One possible arrangement is to set up a configuration file for the software. We can have a YAML file that lists all the default values for the controls. These values can then be edited by the user to their liking. YAML is chosen as its Python-like appearance and its easiness to read and edit. Python, the language which the software is written in, also has good support for parsing YAML files with the `pyyaml` library.

	mode: manual
	bitset:
mode: auto	a: 00000000
bitset:	b: 00000000
a: 00000001	bitclr:
b: 00000000	a: 00000000
bitclr:	b: 00000000
a: 00000000	input:
b: 00000001	- a: deadf00d
input:	b: fadef00d
- a: 00000000	- a: feedf00d
b: 00000000	b: cafef00d
freq: 200	freq: 100
runtime: 60	runtime: 5

Figure 6.1: Config File 1

Figure 6.2: Config File 2

Looking at the configuration file excerpts, Figure 6.1 will make the test run in auto mode for 60 seconds at 200MHz. Input a will be always odd, and input b will always be even. Figure 6.2 on the other hand will make the test run in manual mode, in which the testbench will go through the list of inputs stated under the input key. Once the list is exhausted, the test will terminate.

This method is great for setting up one or two tests, but to scale this up to series of tests, the user may have to generate a collection of such configuration files. The test software can then scan a folder instead of just a single file and run all tests.

If there happened to be significant extensions to the interface in the future, the user will also have to go through many configuration options that maybe irrelevant to the test case. This increases the cognitive load unnecessarily, reducing the user-friendliness of the design.

6.1.2 Read-eval-print Loop

Alternatively, the software can be structured into a read-eval-print loop (REPL). This is a simple interactive command line interface, where the program reads a single user expression, evaluates it, and prints out a response. The software then waits for the next user input, and so the loop continues until the user decides to stop.

```
> mode auto
> bitset a 00000001
> bitclr b 00000001
> freq 200
PLL Configured to 200.00MHz
> run 60
Results: ...
```

Figure 6.3: CLI Excerpt 1

```
> mode manual
> input a 00010001
> input b 000a000c
> input a feedf00d
> input b cafef00d
> freq 100
PLL Configured to 100.00MHz
> run 1
Results: ...
```

Figure 6.4: CLI Excerpt 2

This allows the user to intuitively command the software. Figure 6.3 and Figure 6.4 shows possible commands the user may use to achieve the same effect as the previous option with configuration files. The user can now go directly to the option that needs to be changed, and if a mistake with the configuration was found when looking at the test results, the user can immediately make adjustments within the software's command line interface (CLI). This means the user can be more exploratory, as REPL is more interactive and direct than having to reset, edit a file, run again, and then debug.

In order to scale this option, the software can be easily modified to also accept a macro file as well. Instead of typing in individual commands in the CLI, the user can enter all the necessary commands as lines of a file, and feed that into the software. The macro file can then be scaled and automated so series of tests can be easily ran with a single instruction from the user.

Being the more user-friendly and more scalable option, this option will be implemented in the next phase of the project.

Chapter 7

System Implementation

7.1 Project Hierarchy

Programming the FPGA to communicate with the HPS is no trivial task. Luckily, there exists a golden system reference design [35] for the board in use for this project. Unfortunately, support for certain versions of Quartus are missing from the GSRD download database, including the version used for this project, 16.0. While the design can be opened with a different version of the software, it causes a series of conflicts usually related to using IP cores that have changed over the iterations. To circumvent this issue cleanly, GSRD version 14.1 was downloaded and compiled on a separate install of Quartus II 14.1. This allowed the reference design to be studied in detail, and the sections required for this project to be rebuilt with Quartus Prime 16.0.

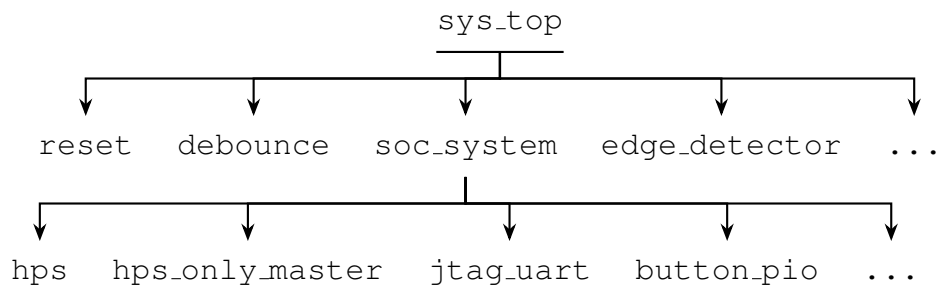


Figure 7.1: Hierarchy of the Golden Reference Design

By examining the structure of the reference design, we see that it has a top-level wrapper called the `sys_top`, which instantiates the Qsys system `soc_system` and a few IP blocks that handles the low level hardware controls on the development board. This Qsys system is of the most interest to us as it contains the module `hps`. In `hps`, there are 3 ports named `h2f_axi_master`, `h2f_axi_slave`, and `h2f_lw_axi_master`, corresponding to the bridges exposed by the HPS for connections [30].

With this knowledge, we can insert the testbench design into this hierarchy by having it wrapped into

a Qsys module, with an open port that works as a slave to the AXI bridge. As the traffic passing through the HPS-FPGA bridge is minimal in our design, the lightweight bridge will be used for its simplicity.

As the testbench only requires a list of registers to be sparingly read and written to, the logic required for the signals on the Avalon slave interface can be handwritten according to the interface specifications [34] without much trouble.

By following the naming conventions the signals allows Qsys Component Editor to automatically detect the Avalon slave from this module at analysis. This saves the troubles of editing the `_hw.tcl` file. This is done in the module `test_wrapper`.

Since Qsys is chosen as the user interface, it makes sense to also put the DUT and the reference design at this level in the hierarchy. This means that the user only needs to use Qsys to swap in new designs. As such, other than handling the interface to the HPS, the wrapper module also exposes conduits that connects to the DUT and the references designs.

In addition, the PLL and its reconfiguration module are also instantiated at this level since as packaged IP cores, Qsys is designed for such integrations.

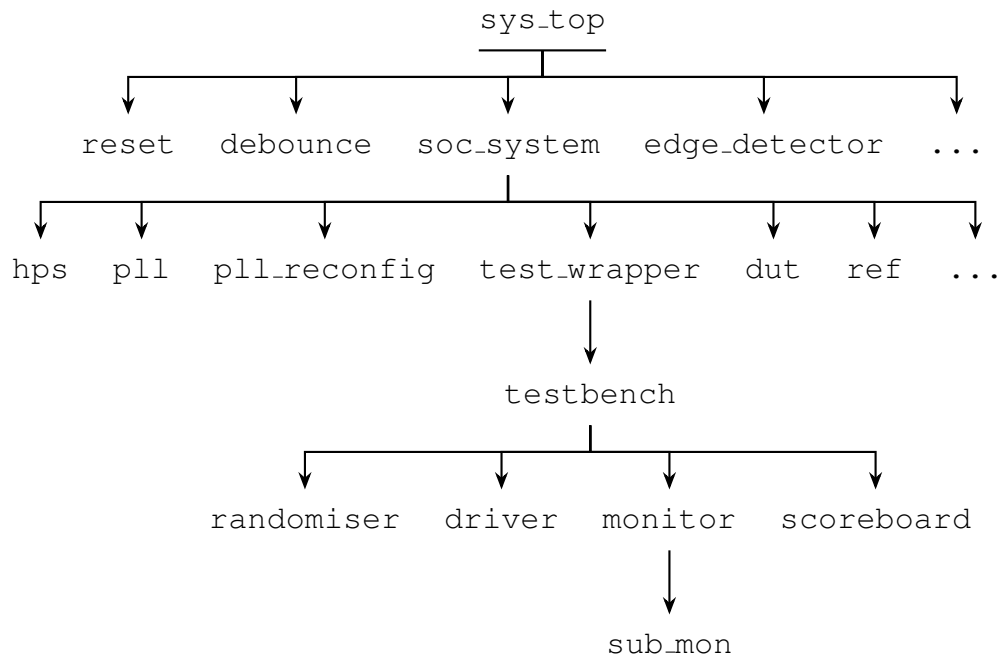


Figure 7.2: Hierarchy of the Full Hardware System

The main section of the testbench is instantiated in the wrapper as a separate module named `testbench`. This module instantiates and connects all main components of the testbench. While this module seems unnecessary, being another wrapper in a bigger wrapper, it does however, provide 2 advantages.

Firstly, it simplifies the development cycle of the framework. During the compilation of a Qsys system, all relevant files are copied into a folder and the system is rendered as a Verilog file that has its direct dependencies contained in that folder. While this is arguably a benefit in terms of dependency man-

agement, it makes the development more difficult since whenever something is changed in the interface between testbench modules, the entire `soc_system` needs to be recompiled to update the dependencies. If forgotten, it can cause confusions as the testbench could be still on the old iteration even though a new compilation at the `sys_top` level has been performed.

Secondly, it makes the simulation of the testbench more straight forward. Verifying the correctness of the Avalon slave in the wrapper is important, but there is no need to go through the interface protocol whenever a new input signal is desired in testbench simulations. Having the `testbench` module allows direct manipulation and examination of the signals in and out of the testbench, without worrying about the HPS side of things in simulations.

Chapter 8

Hardware Implementation

8.1 Randomiser

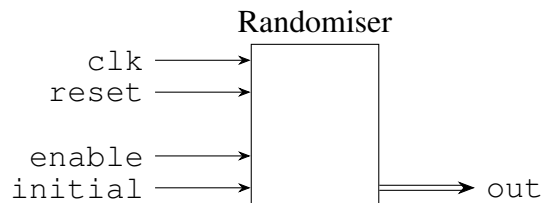


Figure 8.1: Randomiser Block Diagram

Implementing the randomiser is straight forward. A possible set of taps for a 32-bit Galois LFSR is [32, 30, 26, 25]. Referring back at Figure 5.2 on page 16, the logic is to XOR the bits left of the taps with bit 0, and simple right shift for all other bits. For driver to control the randomiser, an enable signal and an initial signal is added as input in addition to clock and reset. The `initial` signal seeds the LFSR.

8.2 Driver

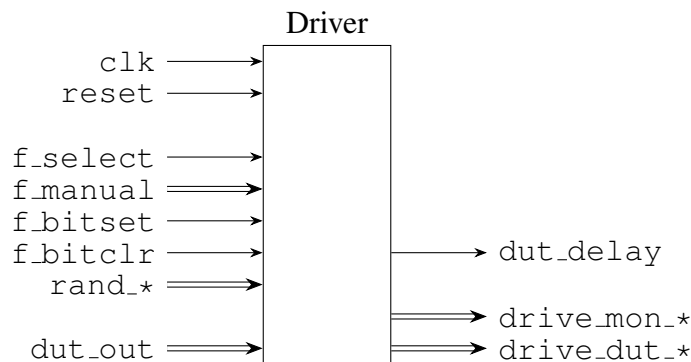


Figure 8.2: Driver Block Diagram

The filter select signal `f_select` selects the mode of operation of the driver. When it is set, the driver will read from `f_manual` and feed them to the output. Otherwise, the driver will take the output of the randomisers at `rand_*`, set and clearing specific bits according to `f_bitset` and `f_bitclr`.

The output is immediately sent to the DUT from the ports `drive_dut_*`. The output is also delayed for a number of cycles before being sent to the monitor from the ports `drive_mon_*`. This delay is known and thus can be configured by the user before compiling the testbench.

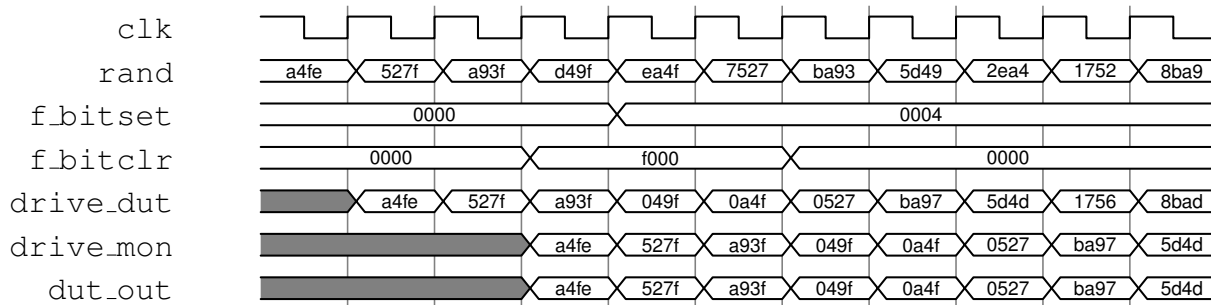


Figure 8.3: Driver Waveform

Figure 8.3 shows how the waveform of the implemented driver looks like when operating in auto mode. It should be stated that all waveforms in this report has been obtained from simulation, and are verified to be correct. For clarity, unnecessary signals and signal values are omitted.

In the example waveform, we assume the design is a simple 1 input 1 output module where the input is passed to the output after 2 cycles. The driver passed `rand` to `drive_dut` after a cycle. When `f_bitclr` is `f000`, the top 4 bits of the output are set to 0, and when `f_bitset` is `0004`, bit 2 of the output is set to 1. It should be noted, that if the same bit is set and cleared by the user, `f_bitclr` takes priority. This is an arbitrary choice, and is noted in the user guide.

The driver delays the output to monitor by 2 cycles, and we can see that this aligns it the output from the DUT.

8.3 Monitor

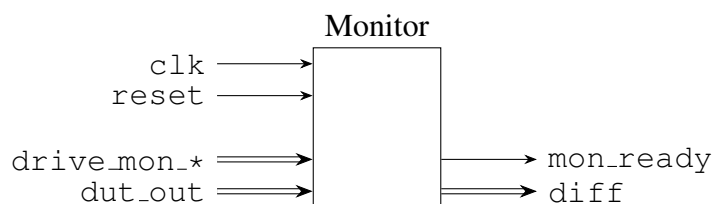


Figure 8.4: Monitor Block Diagram

The monitor takes DUT inputs from the driver, and distributes them to a few sub-monitors. Each sub-monitor containing a reference design then produces the correct results `mon_out` from the inputs with

a relaxed time budget. The monitor then checks the difference between the reference output, `mon_out`, and the DUT output, `dut_out` with XOR gates. This results in `diff`, where each bit set to 1 indicates a wrong bit in the DUT output. `mon_ready` will be set after the distributor has completed an entire round, where the first meaningful `diff` value becomes available.

As the number of sub-monitors and the width of the tested unit are parametrised, the design of the distributor was not straightforward. An one-hot counter is set up to determine the currently active sub-monitor. Since the sub-monitors are clocked `NUM_SUB_MON` times slower, an array of `NUM_SUB_MON` registers each of size `WIDTH` is also created for each input or output of the design. These serve as the interface between the sub-monitors and the rest of the design.

8.3.1 Sub-monitors

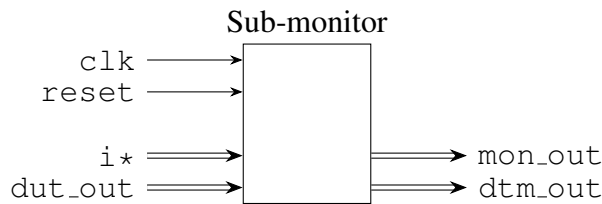


Figure 8.5: Sub-monitor Block Diagram

The current sub-monitor is a part of the monitor module in the architecture design that interfaces with the reference module. In addition to connecting with the reference inputs and outputs, the sub-monitors also handle the delay of the reference module. This is due to the highly parametrised nature of the monitor module, which made it rather complex and inflexible to the addition of more features.

As such, there is an extra signal of `dtm_out`, which has the same value as `dut_out`, but delayed by the number of cycles that the reference design needs to complete its operation, thus aligning with `mon_out`. The other signals are directly connected to the reference module.

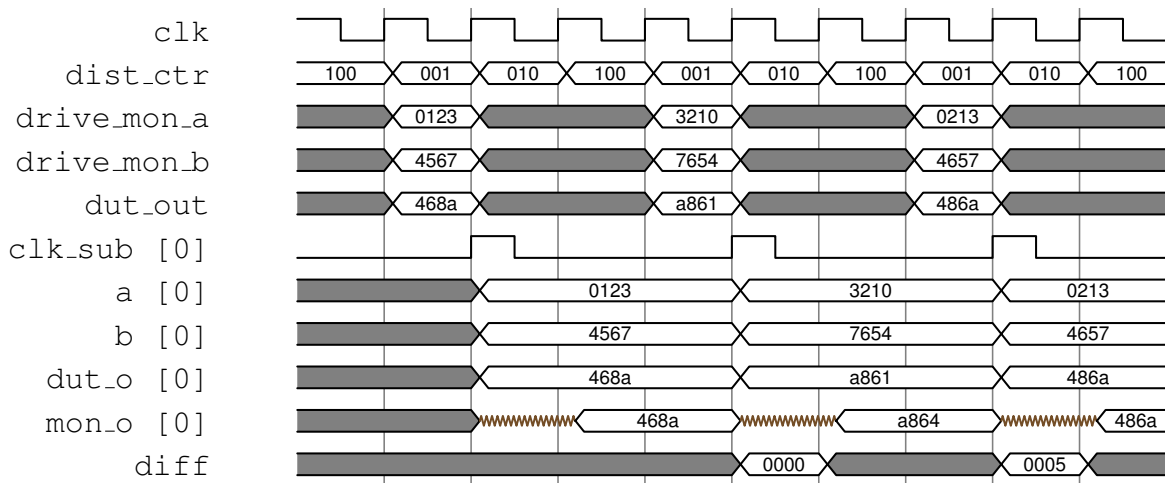


Figure 8.6: Monitor Waveform

Figure 8.6 shows the waveform of a monitor with `NUM_SUB_MON` as 3. The reference adder takes 1 cycle to complete, but it must be clocked at a frequency slower than that of the adder DUT. With 3 sub-monitors, the width of `dist_ctr` is 3, and its lowest bit corresponds to sub-monitor 0, which is shown in detail in this figure. `clk_sub` is the clock driving the sub-monitor, which is made by masking the DUT clock with a delayed version of `dist_ctr`. As it ticks, the I/O values are copied into the register arrays and held for 3 cycles. Within this time, the reference design completes its operation, fixing the value on `mon_o`. When the cycle of `dist_ctr` goes a full cycle and the `sub_clk` ticks again, this value is collected back and XOR'ed to form the final output of the monitor, `diff`. In the example of Figure 8.6, the second result was an error on the DUT, as it gave `a861` while the reference answer was `a864`. This means bit 0 and 2 were different, and `diff` is thus `0005`.

The other sub-monitors all work identically but each 1 DUT cycle later the last. This allows the reference design to run slower than the DUT, to still provide a constant stream of `diff` at the monitor output as designed.

8.4 Scoreboard

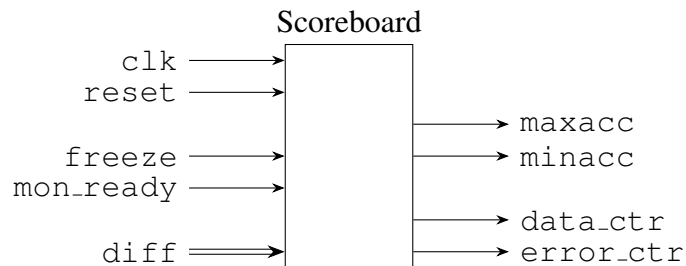


Figure 8.7: Scoreboard Block Diagram

This scoreboard tracks the number of valid test points going through with `data_ctr` and the number of errors within them with `error_ctr`. The external input `freeze` is exposed to the software to stop all counting in the scoreboard. As the current HPS-FPGA bridge set up only allows sequential reads to the FPGA registers, it is necessary to ensure the values do not change within a single set of read commands from the software.

Another implication of the current bridge set up is that there is no simple way of getting all the `diff` values out to the HPS for statistics calculations. This is a limitation of the current implementation, and will be discussed in the Further work section of this report. Therefore, the hardware will have to do some simple statistics.

Since we are interested in how the precision of the DUT degrades as the frequency increases, two signals are created to record the maximum and the minimum precision of the DUT output. Calculating the precision with the `diff` signal means counting the number of leading zeros (CLZ), as zeros indicate the correct bits. As the current implementation is limited with a maximum width of 32, the easiest way of doing CLZ fast is by padding zeros after the number to 32 bits, and then use a large lookup table with don't cares. This precision signal is named `acc`.

To keep track of the minimum precision, register `minacc` is first initialised to the maximum, and then for each smaller value observed, it will take on its smaller value. The comparison logic here is relatively expensive in this fast testbench design. As such, there is great incentive in the future to build a better communication method to allow offloading these operations to the HPS.

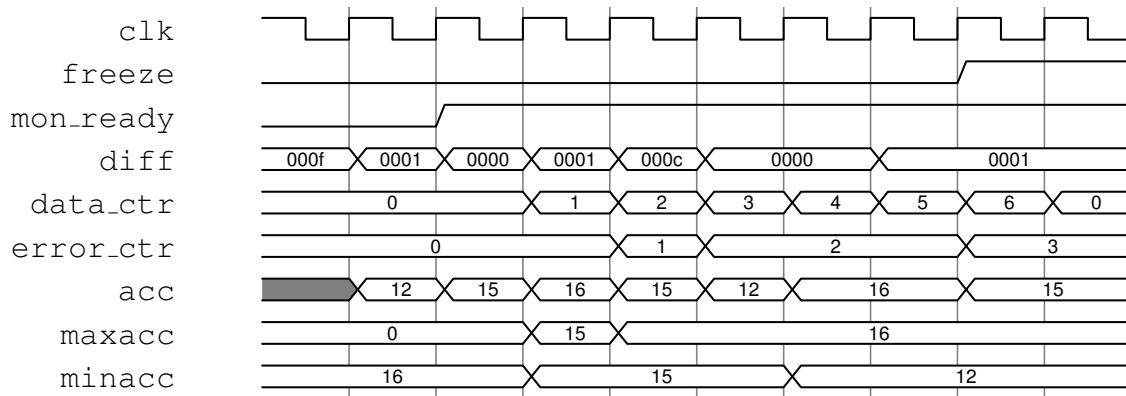


Figure 8.8: Scoreboard Waveform

Figure 8.8 shows a example waveform. The counters and the extrema trackers changes value only if we have `mon_ready && !freeze`.

8.5 Summary

Figure 8.9 provides a detailed look at how the individual hardware components are wired together to form the `testbench` module. The inputs and outputs of this module is contained within `test_wrapper`, which handles the AXI communications when coupled with the `hps` module. As discussed in the software implementation chapter, this allowed the software to access the testbench, completing the overall system implementation.

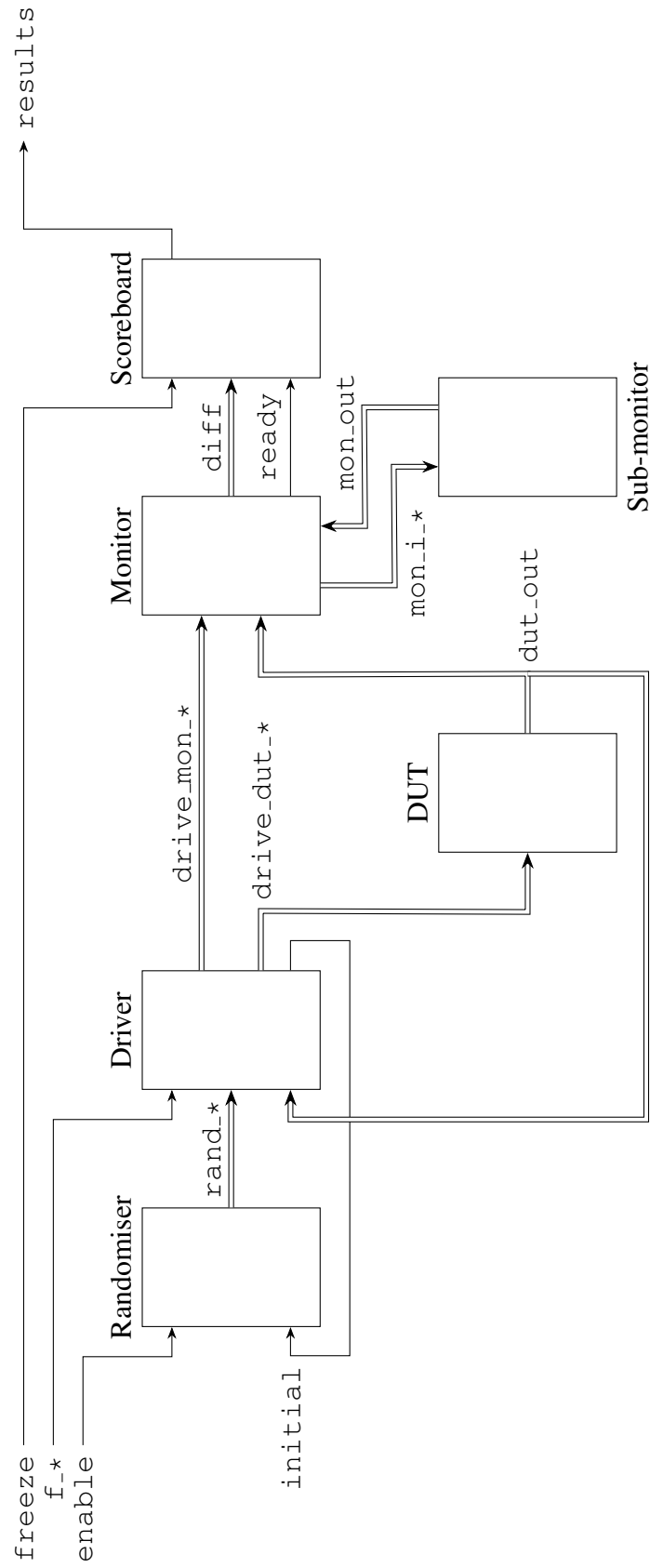


Figure 8.9: Block diagram of the implemented testbench

Chapter 9

Software Implementation

9.0.1 HPS Side

Need to take out this section and only keep the useful stuff The HPS runs Ubuntu and a Bash script has been written to load the bitstream onto the FPGA. Next, a program was written in Python to test the hardware design from the HPS. The interfaces are mapped onto the physical memory, thus they can be accessed by opening `/dev/mem`. Checking against the specifications [30], the lightweight master is at `0xFF20_0000`. Qsys allocates the memory spaces of modules relatively, so when it reports that the adder has been placed at `0x0010_0000`, it is physically at `0xFF30_0000`. The adder was designed to have its two inputs at `0x00` and `0x10` and its output at `0x20`, which were assigned by Qsys relatively to `0xFF30_0000`.

With the memory mapping understood, the program was designed to closely mirror this relative relationship between the modules using classes. For example, the adder defines its output at `0x20`, but its read and write methods are inherited from an AXI class that brings it to the correct physical address by adding the address of the bridge defined in it. This parallel between software and hardware should be helpful as the product gets more complex.

To verify what I have built and learnt was correct, 1000 add operations were executed separately with and without the hardware acceleration of the FPGA. The results were compared and confirmed that this training module worked. While called hardware acceleration, the FPGA actually had worse performance than the HPS in this testing case. The CPU is reasonably efficient in calculating additions, while calculating on the FPGA requires the adder I/O data going through the HPS-FPGA bridge. This incurs a significant overhead, thus slowing it down.

9.0.2 New Stuff Here

Most interactions through the HPS-FPGA bridge involves writing and reading different memory locations.

To set up a REPL, we need to define a list of expressions that the user is allowed to give to the program. (not tcl?)

Command	Explanation
reset	Resets the system and test results.
version	Prints the system version.
freq <speed>	Sets the clock speed to the specified value in MHz. Prints the actual frequency configured.
mode <m a>	Choose between <u>m</u> anual and <u>a</u> uto test mode.
manual <a b> <hex>	Give input in manual mode.
bitset <a b> <hex>	Force bits to be 1 in auto mode.
bitclr <a b> <hex>	Force bits to be 0 in auto mode.
run <time>	Runs the test for specified duration in ms. Prints the results at the end of the test.
exit	Exits the REPL.

Table 9.1: Commands accepted in test REPL

Chapter 10

Testing

10.1 Functional Correctness

10.1.1 Simulation

During implementation, the modules were simulated with ModelSim to verify their correctness. Initially we followed the data path of randomiser, driver, monitor, sub-monitor, and scoreboard, each module was attached and simulated. As the randomiser was a perfect source for setting up random tests for the following module, this process was relatively straightforward once the randomiser was ascertained to work correctly. Once the system design became stable, the testbench module can be simulated as a whole. Since RTL simulation is cheap, this is done after any design addition or change thereafter. As writing hardware can be a lot less intuitive than writing software, the simulations have prevented many errors from going on to the hardware.

10.1.2 FPGA Testing

After each major functional addition, the framework is tested as whole on the development board. This means that it is integrated with a test DUT, usually an adder, and the entire system is synthesised and programmed on the FPGA. A basic test script was written early on to for this system level testing. The FPGA testing serves to confirm the functional correctness shown by the software simulations.

However, the confirmatory nature of the FPGA tests stopped after the number of sub-monitors instantiated by the monitor was made configurable. The test results no longer agrees with simulation results. To discover the cause of this discrepancy, another debug method was suggested by my supervisor.

10.1.3 Post-fit Simulation

Previously, the simulation was done with modules files before synthesis. In order to be closer to hardware, we can synthesise and fit the design with first, then with the *EDA Netlist Writer*, a massive Verilog file can be generated containing the entire design. This can then be fed into ModelSim for a more accurate simulation.

With this, a signal was found to be off by a clock cycle so we fixed it in the design. However, this was not sufficient, as the FPGA test results still differ from that of the post-fit simulation. The next proposed debug method by my supervisor is to use *SignalTap*, which can probe signals inside of the FPGA, allowing us to see the actual waveforms in hardware. Being a much more time-consuming method, we decided to first take a closer look at the design code before committing to it. Among other issues, my supervisor was able to identify a clock which I have handled rather carelessly. After spending some time tidying up this clock, the system again behaved correctly as predicted by the simulation.

10.2 Maximum Frequency

As a important benchmark of the framework, the maximum frequency is closely monitored during the implementation process. This is done with *TimeQuest Timing Analyser*, which is a tool that can provide an estimate on the maximum frequency the testbench and run safely on. If this value is lower than what is required, the tool can also provides a list of the slowest offending signal paths for us to optimise on.

During FPGA tests, the physical maximum frequency the board can achieve is measured with a script that will run the same test repeatedly with increasing frequencies. These are useful as the software estimations are usually more conservative than what is possible on hardware.

10.3 Out-of-the-box Testing

10.3.1 Introduction

In an out-of-the-box(OOTB) test, the product is delivered to test users as a packaged box. Its unpacking process, in which the users sets up and uses the product, is then observed to study the intuitiveness of the design. This testing method is used in this project as one of the key determinant for the success of this project is how convenient it is for users to configure and make use of the framework.

For this project, we have made up a scenario where the users have designed an adder with a suspected error that they wish to ascertain. The user guide [Appendix A], the testbench, and a flawed adder design was provided to the test volunteers. An extra piece of logic was added to an adder design, so that if both inputs are odd, the adder will produce an incorrect output. The test volunteers are made aware of the fact that there might be an error relating to the parity (evenness) of the inputs.

Unfortunately, due to time constraints and the limited number of people with Quartus knowledge that I have access to, this test was only performed by me first and one volunteer after. Nevertheless, this process still proved to be helpful with many design flaws being identified, and a list of possible improvements being drawn out. The possible improvements will be discussed in the further work chapter of this report, while the identified flaws will be evaluated within this chapter.

10.3.2 Hardware Configuration

For the hardware portion, the test user were given adders of width 16 which is expected to work up to 3 times as fast as the reference. He was able to correctly modify the default values for `WIDTH` where is 32, and for `NUM_SUB_MON`, where the default is 2. However, there seemed to be an issue with Qsys where the widths of ports are not always updated when the module parameters are modified. As he had not much familiarity with Qsys, and was not expected to have any knowledge of the inner workings of the `test_wrapper` module, the test had to be interrupted. To fix this, we deleted and re-added both the `test_wrapper` and the `dut_adder` module, modifying their default values before their were placed into the *System Contents* window. A note was also added in the user guide.

After dealing with the minor obstacle, he was able to generate the HDL from Qsys, and compile the full design with Quartus without issues.

10.3.3 Using the Test Software

The uploading and the programming procedure was smooth, the volunteer followed the guide and successfully primed the development board for testing.

He then had to read the command list of the testing software and devise a plan to pinpoint the relationship between the parity of the inputs and the correctness of the design output. At this point, he suggested that while some of the commands were explained succinctly, the commands on manipulating the inputs were not explained in sufficient detail and left him somewhat confused. As such, I explained the details and improved the user guide accordingly.

While explaining, the volunteer quickly realised that he can use the `bitset` and `bitclr` commands to fix the evenness of the inputs. However, in doing so he typed in a series of instructions that the test software never had to deal with during initial testing. This revealed a bug in the software where the frequency of the PLL output is not well defined if the software issued a reset command after setting the frequency.

After a second interruption to fix this bug, the test was able to continue. With some experimenting in the REPL interface, the user was able to correctly conclude that the design gives incorrect values only when both inputs are odd. Test automation was also tested, in which the volunteer had no trouble getting the software to run with his command list file.

10.3.4 Testing Results

Despite the two intermissions, the OOTB was reportedly a smooth experience for the test volunteer. This showed that the framework is reasonably intuitive and user-friendly. However, the test also highlighted a few major places where the project can do better in.

Over the entire duration of the OOTB testing process, a significant amount of time was wasted fiddling with Quartus and Qsys GUIs. It was also the step most where the framework is the most prone to user mistakes. For example, a misclick in the Qsys *System Contents* window may only cause an error during the synthesis of the entire project, and for someone not familiar with Qsys, this error may take a very long time to be identified and fixed. As such, having a way to automate the hardware configuration process would be a great improvement to the usability of the product. Furthermore, a unified configuration system would be even better, as the user will not have to figure out what customisation options have to be done in the GUIs before compile, and what can still be manipulated during in the testing software. A method of accomplishing both will be proposed in the further works chapter of this report.

Another key improvement arising from the OOTB testing was on the readability of the user guide. It is difficult to produce a guide that will be suitable to readers in all skill levels. Therefore, while this one test has made it slightly better, feedback from more users and especially users with different levels of experience and knowledge is definitely still needed.

Chapter 11

Evaluation

11.1 Product Metrics

11.1.1 Robustness

The next few measures look at the performance of the final product. First, the maximum stress of which the testbench can provide without failing is a good metric. This can be quantitatively measured by the maximum data throughput across the DUT, and the maximum frequency that the DUT can be running where the testbench remains reliable. A robust testbench with a higher maximum frequency can reveal a wider picture in the performance of the DUT. This would hopefully allow more insights to be gained regarding the DUT, or it could mean that the testbench can be used for future designs that may be faster than the current one. As the main quantitative metric, this will be a vital indicator of the project's success.

11.1.2 Flexibility

The flexibility of the testbench is also vital to the product's performance. The testbench should be suitable for a range of DUTs. This will allow the testbench to be used for future experiments. The flexibility can be measured by the number of configurable parameters that it has, and the range of which these parameters can be adjusted to.

11.1.3 User-friendliness

The ease of use of the testbench can be another evaluation point. On the hardware side, the verification system can be packaged into a Qsys module. Given the DUT is also a module with an agreed interface, it can easily be connected in Qsys for testing. On the software side, a user-friendly interface could be built. A usable command line interface may be good enough, but a simple graphic interface could make the tests much more visual and informative.

The interface could also provide information on the failure in the DUT. A better testbench will provide more insightful details when the DUT fails. This would make debugging or evaluating the design much simpler. Along with the GUI, this project has many optional extensions that would be discussed further in section 6.2. After the main goal of the project being met, the number of optional functions implemented becomes a good measure of the progress of the project.

11.1.4 Note

A noteworthy point in evaluation is regarding the progress of the sister project. The purpose of the testbench is to verify and stress arithmetic designs. If these designs are not available near the end of this project, it would be difficult to empirically prove the capabilities of the testbench and its surrounding system. Since the project comprises a verification system, the results from the benchmarks should not be used for evaluation of this project. It is not impossible, as there are still substitutes for them. For functional purposes, standard off-the-shelf arithmetic modules could be used in-lieu. For other purposes, it is possible to have a model done before the actual design starts in the paired project. While this would allow this project to progress, it would be extra work for the other project. In all, it would be nice to have a solid arithmetic module completely to run in this testbench, but without one, the system can still be built and completed, albeit generating less useful data towards the overall aim of the project.

11.2 Project Metrics

11.2.1 Implementation Plan

One natural way of measuring the success of the project is to look at the actual progress and compare it to the implementation plan. It should be noted that no plan is perfect, so some deviation is allowed. However, if there is significant delay from the plan, there must be justifications given.

11.2.2 Fallbacks and Extensions

The fallbacks and extensions were detailed in the implementation plan. The progression of these extension tasks can be used as a point of evaluation to the project. This is already reflected in the product metrics, as 6.1.2 and 6.1.3 examines mainly the success of the extension tasks, while only 6.1.1 analyses the achievements of the core task. However, this does not mean that this project will be completely sealed if all extension tasks are complete before the final deadline, there is always more potential for future work. In the very unlikely case that the project progresses to such a state, new ideas such as dynamic voltage scaling will be drafted and evaluated.

Chapter 12

Conclusion

Chapter 13

Further Work

13.1 Verilog Preprocessor

13.1.1 Additional Customisation

1. LFSR with variable width
2. variable number of inputs/outputs to DUT

13.1.2 Unified Interface

1. Qsys
2. Quartus
3. Software

13.2 Automatic Delay Reconfiguration

Since finding out the delay of the DUT is useful for further simplifying the configuration process of the framework, a delay tester is built within the driver. It counts the number of cycles for the DUT to produce its output as `dut_delay`, but it still has a few limitations in its design. As such, using this to reconfigure the driver to monitor delay on the fly is not yet possible. How this might be achieved will be discussed in the Further Work chapter of this report.

The delay tester is built with a simple FSM.

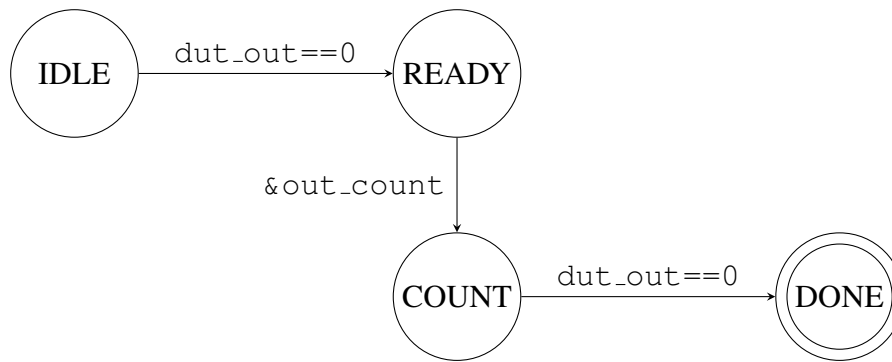


Figure 13.1: Delay Tester FSM

We first start a counter called `out_count`. Whenever this reaches all 1's, the driver output is set to some value with a known safe DUT output. In this example, the safe output value is 0, which means when the delay tester is active, no other output from the driver can result in a DUT output of 0.

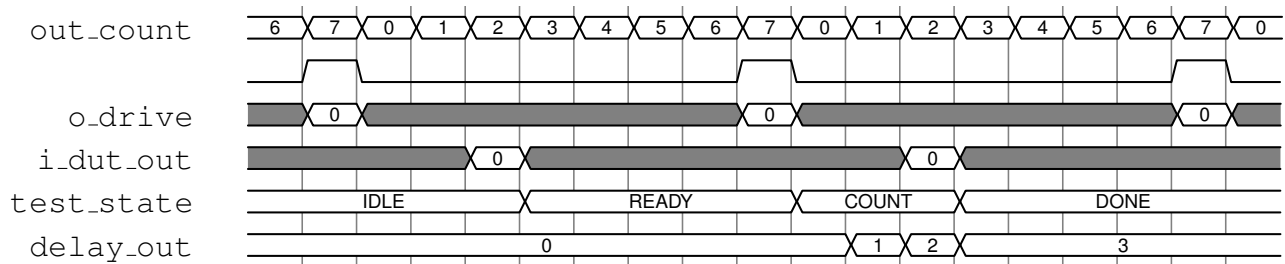


Figure 13.2: 3-bit Delay Tester Waveform

The FSM starts in state IDLE. When the first 0 output is detected from the DUT, the FSM enters the READY state. It now knows that it can enter the COUNT state when the `out_count` becomes all 1's again, triggering the next safe test input. The FSM leaves the COUNT state for the DONE state when the safe output of 0 is detected. The delay tester process is now complete and `out_count` can be deactivated.

With this, the DUT's delay in clock cycles is the same as the number of cycles that the FSM stayed in state COUNT. The delay counter `delay_out` increments itself every cycle if the FSM is in that state. When the FSM enters the DONE state, the value of the delay counter is the delay of the DUT. With a 3-bit counter as shown in the timing diagram, it can measure this delay for up to 8 clock cycles. Longer delays can be measured by extending the width of `out_count`.

13.2.1 Reference Delay

Currently reference is sub-monitor `-i` which needs to pass through `dut_out` as well. To have a pure module, need this delay control.

Appendix A

User Guide

A.1 Introduction

This project provides an extensible framework for at-speed evaluation of arithmetic hardware. It is currently only implemented and tested with the following environment:

Item	Version
Hardware	Cyclone V SX SoC development board
HPS System	Ubuntu 16.04.6 LTS (GNU/Linux 3.10.31 armv7l)
HPS Python	2.7.12
Quartus Prime	16.0.0.211

Table A.1: Tested Environment

To download this implementation, use

```
git clone https://github.com/MerelyLogical/arithmetic-testbench.git
```

It includes both hardware and software files.

A.2 Configuring Hardware

The hardware needs to be configured to fit the design before compiled and uploaded onto the development board. This guide provides the steps to this configuration with Quartus' GUI in a beginner-friendly way. It should be noted that this configuration can also be done by editing the `.qsys` file directly. The summary of this section is to first integrate the user's design into the Qsys system as a component. Then the system is compiled into synthesisable HDL. The whole system is then synthesised fitted, and then assembled into an `.sof` file. In order to upload this to the FPGA through the HPS later, it is converted to an `.rbf`, but this is optional if a different method of programming the FPGA is desired.

1. Open Quartus 16.0
2. File → Open Project... → `cy5-systest.qpf`.
3. Tools → Qsys
4. File → Open... → `soc_system.qsys`. (The open dialogue may pop-up automatically on start up of Qsys.)

5. In the IP Catalog window, click New..., which will bring out the component editor. This turns your designed components into Qsys components, which will then be integrated into the Qsys system.
6. Enter basic information of your design in the *Component* tab of the component editor.
7. In the *Files* tab, add your design files, select the top-level module and click on *Analyze Synthesis Files*.
8. Module parameters and be set in the *Parameters* tab.
9. The *Signals & Interfaces* tab should have detected all ports of your design. First ensure that the clock and the reset inputs are detected and categorised in the interface list correctly. Then to allow connection to the rest of the testbench, move the two inputs and the one output of your design into a *Conduit* interface. Associate the clock and the reset signal to the conduit. Then name the two inputs a and b, and name the output as out.
10. Finish... → Yes, Save → Yes, save before refresh.
11. The created component should now show up in the *IP Catalog*.
12. Add the component to the system, enter the desired parameter values before clicking on *Finish*.
13. Connect the signals of your design using the *System Contents* window of Qsys. The clock should be connected to the `outclk0` signal of the component `pll_dut`. The reset signal should be connected to the `clk_reset` signal of the component `clk_ref_50M`. The conduit should be connected to the opposing conduit in the `test_wrapper` module.
14. Click on the `test_wrapper` module and configure its parameters where necessary. `NUM_SUB_MON` determines how many sub-monitors are spawned to run in parallel. `WIDTH` should match the width of your design's I/O width.
15. Fix any outstanding errors in the *Messages* window if preset.
16. Click on Generate HDL...
17. Ensure *Create HDL design files for synthesis* is not on none.
18. Generate the HDL files.
19. Click on Finish to close Qsys.
20. Quartus should prompt you to add the `.qip` file and the `.sip` file. Follow them by going to Project → Add/Remove Files to Project...
21. After making sure that the two Qsys files are included, start the compilation by going to Processing → Start Compilation.
22. By the end of the compilation, the assembler should have produced `output_files/cy5-systest.sof`. We need this in a different format, so go to File → Convert Programming Files...
23. Select `.rbf` as the output type.
24. Click on the only entry in the input files table, and click on Add File... to add the produced `.sof` file.
25. The hardware setup is completed after obtaining the `.rbf` file.

A.3 Setting up the Board

By the end of this sections, we should have the FPGA programmed and the software uploaded to the HPS, ready to start running the tests. In this guide we will be using rsync to upload the `/scripts` folder and the `.rbf` file into the board, and programming the FPGA from the HPS, but other methods are equally as valid.

1. Copy the `.rbf` file into `/scripts` and upload it with `rsync`.

```
rsync -rtuv ./ root@ee-cy5soc3.ee.ic.ac.uk:/home/____/scripts
```

2. Connect to the board. (This can be done by using PuTTY on Windows or a `ssh` command on Linux.)

```
ssh root@ee-cy5soc3.ee.ic.ac.uk
```

3. Program the FPGA.

```
cd /home/____/scripts
./program_fpga.sh output_file.rbf
```

A.4 Running Tests

```
python -O ./run_test.py [file]
```

If no file is provided, the software will enter a REPL, where the commands available are as follows. If a file is provided, the software will read each line and execute the commands with the same syntax requirements, except for `exit` which will not work in a file.

Command	Explanation
<code>reset</code>	Resets the system and test results.
<code>version</code>	Prints the system version.
<code>freq <speed></code>	Sets the clock speed to the specified value in MHz. Prints the actual frequency configured.
<code>mode <m a></code>	Choose between <u>m</u> anual and <u>a</u> uto test mode.
<code>manual <a b> <hex></code>	Give input in manual mode.
<code>bitset <a b> <hex></code>	Force bits to be 1 in auto mode.
<code>bitclr <a b> <hex></code>	Force bits to be 0 in auto mode.
<code>run <time></code>	Runs the test for specified duration in ms. Prints the results at the end of the test.
<code>exit</code>	Exits the REPL.

Table A.2: Commands accepted in test REPL

Notes:

- Arguments in angle brackets are required. Arguments in square brackets are optional. Vertical bars separate all possible options.
- Frequency configuration is done by a PLL which has limited granularities. As such the actual frequency may differ from the desired frequency.

- Argument `<a|b>` is used to select which input this command will apply to.
- `<hex>` needs to be in the format `^(0x)?[0-9a-fA-F]{1,8}$`. In words, it takes 1 to 8 digits of case-insensitive hexadecimal digits. No other characters including space or underscore are allowed, base prefix `0x` is unnecessary but allowed.
- Under the current hardware environment, a safe range for `<speed>` in MHz is from 50 to 400.
- When the same bit is set and cleared, clear always take priority.

Bibliography

- [1] I. Ahmed, S. Zhao, J. Meijers, O. Trescases and V. Betz, “Automatic BRAM Testing for Robust Dynamic Voltage Scaling for FPGAs”, *Int. Conf. on Field-Programmable Logic and Applications*, 2018.
- [2] A. Amin, W. Shinwari, “High-Radix Multiplier-Dividers: Theory, Design, and Hardware”, *IEEE Trans. Comput.*, vol. 1, no.8, 2008.
- [3] R.P. Brent, “A Regular Layout for Parallel Adders”, *IEEE Trans. Comput.*, vol. C-31, pp. 260-264, 1982.
- [4] B. Catanzaro, and B. Nelson, “Higher Radix Floating-Point Representations for FPGA-Based Arithmetic”, *Proceedings of the 51st Annual Design Automation Conference*, 2005.
- [5] L. Chen, F. Lombardi, P. Montuschi, J. Han and W. Liu, “Design of Approximate High-Radix Dividers by Inexact Binary Signed-Digit Addition”, *Proceedings of the on Great Lakes Symposium on VLSI*, 2017.
- [6] R. Duncan, “A Survey of Parallel Computer Architectures”, *Computer*, vol. 23, pp. 5-16, 1990.
- [7] J.W. Duran, “An Evaluation of Random Testing”, *IEEE Trans. on Software Engineering*, vol. SE-10, no. 4, pp. 438-444, 1984.
- [8] M.D. Ercegovic, “On-line Arithmetic: An Overview”, *28th Annual Technical Symposium*, pp. 86-93, International Society for Optics and Photonics, 1984.
- [9] M.D. Ercegovic, and T. Lang, “Digital Arithmetic”, Morgan Kaufmann, 2003.
- [10] S. Hazwani, et al, “Randomness Analysis of Pseudo Random Noise Generator Using 24-bits LFSR”, *Fifth Int. Conf. on Intelligent Systems, Modelling and Simulation*, 2014.
- [11] P. Kornerup, “Reviewing High-Radix Signed-Digit Adders”, *IEEE Trans. Comput.*, vol.64, no. 5, pp. 1502-1505, 2015.
- [12] H. Li, J.J. Davis, J. Wickerson and G.A. Constantinides, “ARCHITECT: Arbitrary-precision Constant-hardware Iterative Compute”, *Int. Conf. on Field-Programmable Technology*, 2017.
- [13] T. Lynch, and M.J. Schulte, “A High Radix On-line Arithmetic for Credible and Accurate Computing”, *Journal of Universal Computer Science*, vol. 1, no. 7, pp. 439-453, 1995.
- [14] T. Lynch, and M.J. Schulte, “Software for High Radix On-line Arithmetic”, *Reliable Computing*, vol. 2, no. 2, pp. 133-138, 1996.

- [15] G. Marsaglia, “Xorshift RNGs”, *Journal of Statistical Software*, 2003.
- [16] H.R. Srinivas, and K.K. Parhi, “High-Speed VLSI Arithmetic Processor Architectures Using Hybrid Number Representation”, *J. of VLSI Sign. Process.*, vol. 4. pp. 177-198, 1992.
- [17] K. Shi, D. Boland, and G.A. Constantinides, “Accuracy-Performance Tradeoffs on an FPGA through Overclocking”, *Proc. Int. Symp. Field-Programmable Custom Computing Machines*, pp. 29-36, 2013.
- [18] K. Shi, D. Boland, E. Stott, S. Bayliss, and G.A. Constantinides, “Datapath Synthesis for Overclocking: Online Arithmetic for Latency-Accuracy Trade-offs”, *Proceedings of the 13th Symposium on Field-Programmable Custom Computing Machines*, pp. 1-6, ACM, 2014.
- [19] O. Šćekić “FPGA Comparative Analysis”, *University of Belgrade*, 2005.
- [20] A.F. Tenca, and M.D. Ercegovac, “Design of high-radix digit-slices for on-line computations”, 2007.
- [21] K.S. Trivedi, and M.D. Ercegovac, “On-line Algorithms for Division and Multiplication”, *IEEE Trans. Comput.*, vol. C-26, no. 7, pp. 667-680, 1977.
- [22] P. Whyte, “Design and Implementation of High-radix Arithmetic Systems Based on the SDNR/RNS Data Representation” *Edith Cowan University*, 1997.
- [23] Y. Zhao, J. Wickerson, and G.A. Constantinides, “An Efficient Implementation of Online Arithmetic”, *Int. Conf. on Field-Programmable Technology*, 2016.
- [24] Accellera Systems Initiative, “Universal Verification Methodology 1.2 User’s Guide”, 2015.
- [25] Altera Corporation, “Cyclone V SoC Development Board Reference Manual”, 2015.
- [26] Altera Corporation, “Memory System Design”, *Embedded Design Handbook*, 2010.
- [27] Altera Corporation, “Introduction to Altmemory IP”, *External Memory Interface Handbook: Reference Material*, vol. 3, 2012.
- [28] Altera Corporation, “Phase-Locked Loop Basics, PLL,”.
- [29] Altera Corporation, “Creating Qsys Components”, 2018.
- [30] Altera Corporation, “Cyclone V Hard Processor System Technical Reference Manual”, 2018.
- [31] Imperial College “An Ethics Code”, *Imperial College Research Ethics Committee*, 2013.
- [32] Intel Corporation, “Cyclone V SoC Development Kit and Intel SoC FPGA Embedded Development Suite”.
- [33] Intel Corporation, “Introduction to Intel FPGA IP Cores”, 2018.
- [34] Intel Corporation, “Avalon Interface Specifications”, 2018.
- [35] RocketBoards.org, “GSRD 14.1 User manual”, 2015.
- [36] Xilinx, Inc, “Zynq-7000 All Programmable SoC”, 2018.
- [37] Xilinx, Inc, “ZedBoard (Zynq Evaluation and Development) Hardware User’s Guide”, 2012.