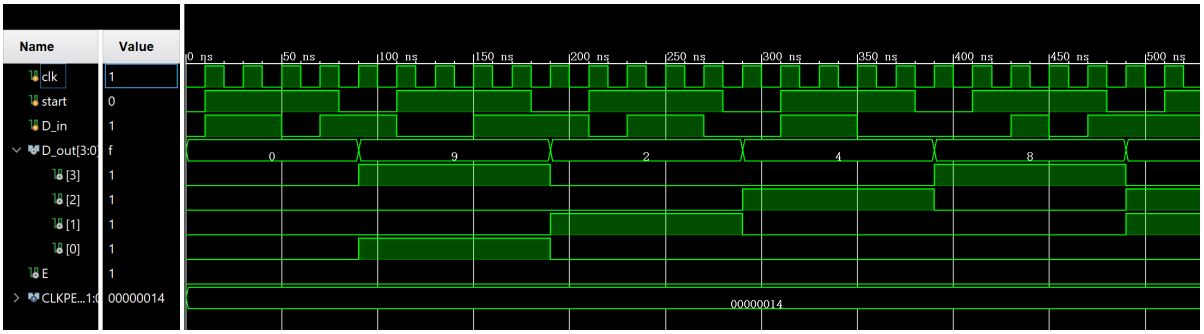
Report Lab 2

In the first two part of this report I will focus on the timing report. In the FSM part of the report I will focus on the trade-off between PPA.

2.1 Gray to BCD Decoder

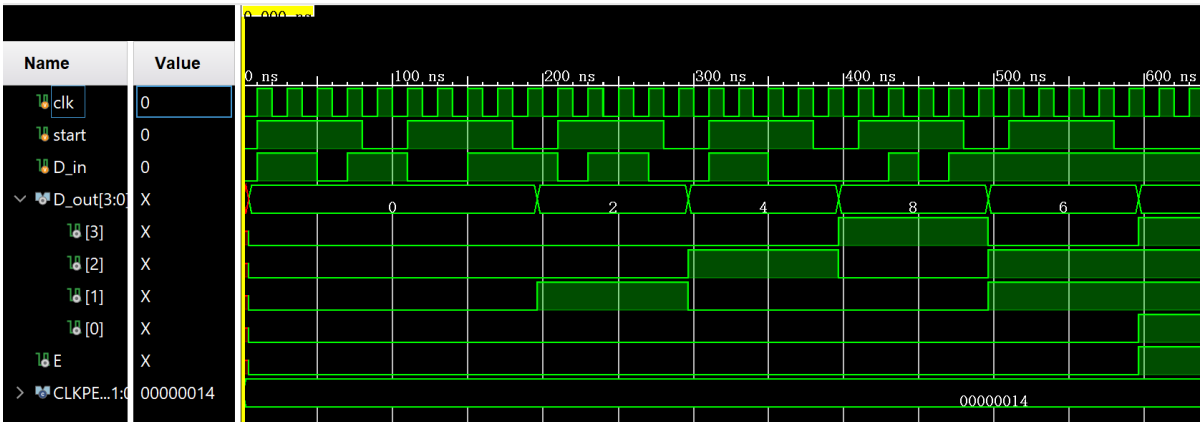
Step 1: Behavior Simulation

The waveform of behavior simulation is shown as below:

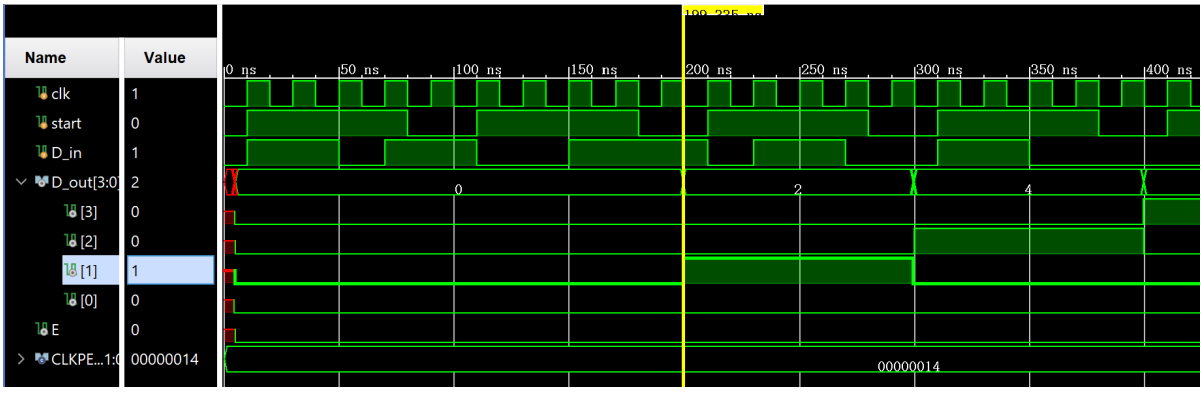


After checking all functions work properly.

To make comparison more clear, the post-synthesis timing simulation and post-implementation timing simulation are also shown in here. The delay of the post-implementation timing simulation is longer than post-synthesis timing simulation due to the P&R.



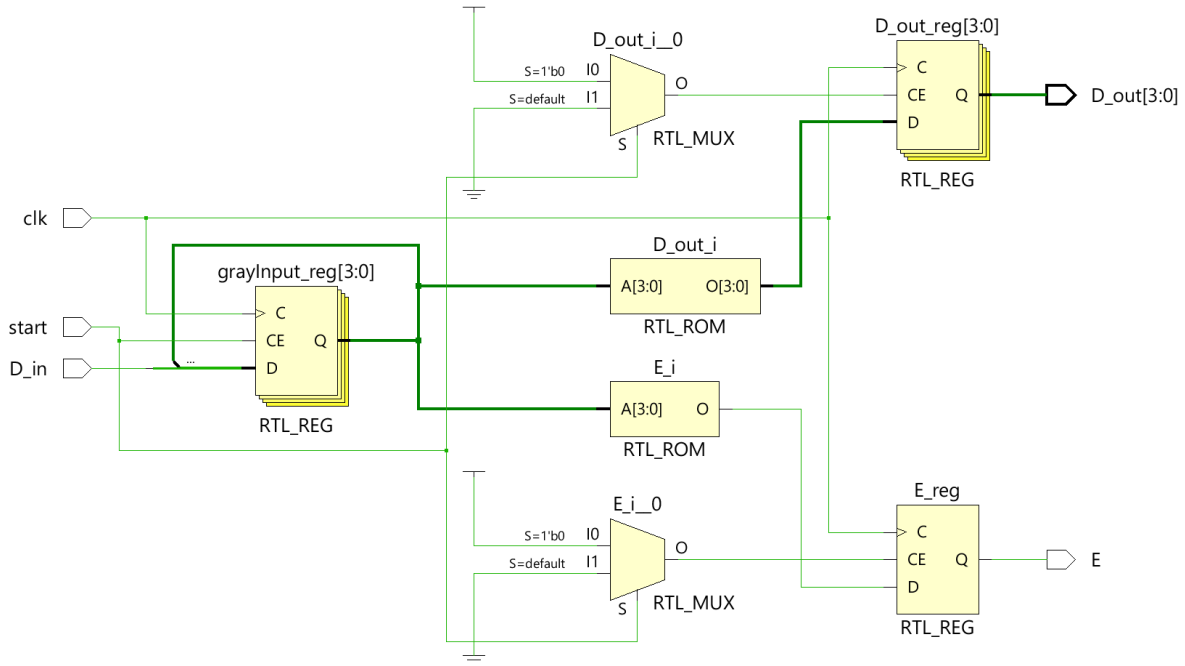
post-synthesis timing simulation



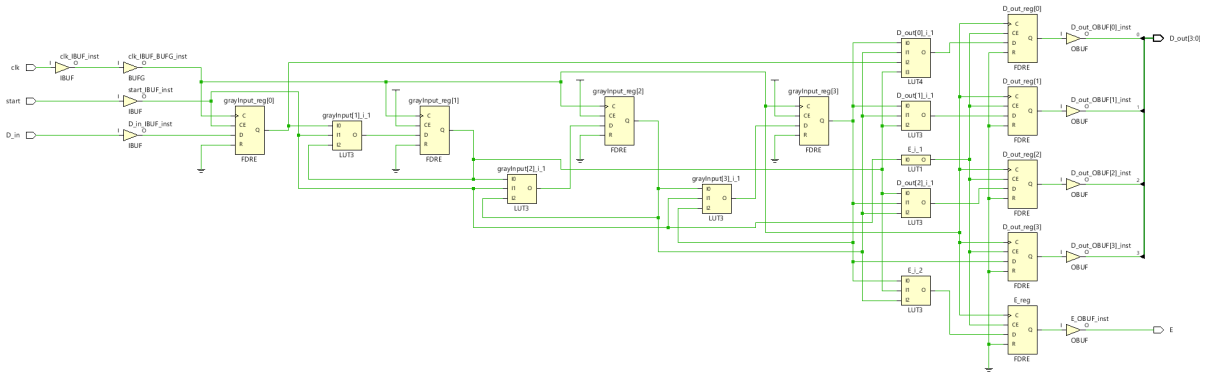
post-implementation timing simulation

Step 2: Synthesis

Then start elaborating to convert RTL description into schematic:



Then one step further, using synthesis to make this netlist more realistic:

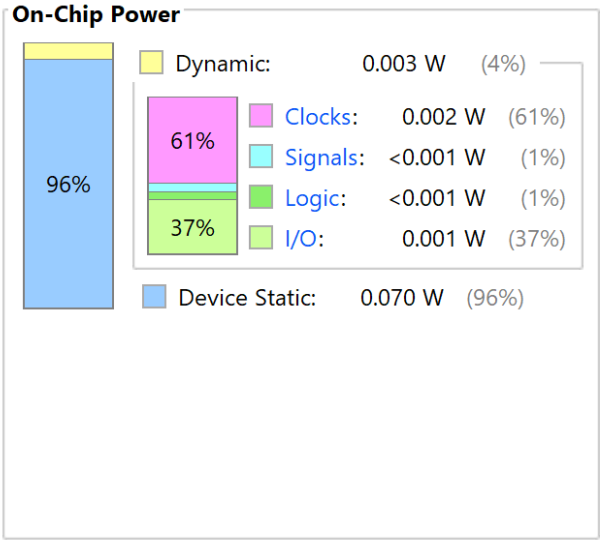


And check the power report of this synthesis result:

Power estimation from Synthesized netlist. Activity derived from constraints files, simulation files or vectorless analysis. Note: these early estimates can change after implementation.

Total On-Chip Power:	**0.073 W**
Design Power Budget:	**Not Specified**
Power Budget Margin:	**N/A**
Junction Temperature:	**25.4°C**
Thermal Margin:	59.6°C (11.9 W)
Effective ϑJA:	5.0°C/W
Power supplied to off-chip devices:	0 W
Confidence level:	Medium

Launch Power Constraint Advisor to find and fix invalid switching activity



Compared with the power derived in Lab 1, thought the I/O is still the highest power consumption part, the ratio has dropped from around 97% to 90% due to the fact that the area of this circuit is larger than the simple CC_11 or top level. However, since the area of this decoder could still be small, the I/O still consumes most of the power.

So far everything we did is the same as Lab 1, **however, this is no longer a purely combinatorial circuit, it has clock signal**. So it is also worth deriving its timing report. The most important parameter is **Worst Negative Slack (WNS)**

$$WNS \equiv T_{constraint} - T_{actual}$$

where the constraint time is also known as worst delay. It is the delay under worst cases, any delays in the circuit should **NEVER** exceed this value otherwise the input or the signal won't be captured by some components correctly. If

1. WNS < 0: constraints are too optimistic (should be thus modified)
2. WNS > 0: GOOD
3. WNS >>>>0: having margin, can relax constraint time a little bit (make it smaller) to improve the operating frequency.

In this Lab, first set constraint time to 10ns, which means that the frequency is 100MHz, the WNS is as follow (WNS may be **too large**):

Design Timing Summary

Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	8.299 ns	Worst Hold Slack (WHS):	0.148 ns	Worst Pulse Width Slack (WPWS):	4.500 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	0.000 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0	Number of Failing Endpoints:	0
Total Number of Endpoints:	8	Total Number of Endpoints:	8	Total Number of Endpoints:	10

All user specified timing constraints are met.

which is 8.299ns >> 0, so we can make it smaller to get better performance (higher frequency). So set it to 4ns which means that the frequency is 250MHz, the WNS is as follow, which is more **reasonable**:

Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	2.299 ns	Worst Hold Slack (WHS):	0.148 ns	Worst Pulse Width Slack (WPWS):	1.500 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	0.000 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0	Number of Failing Endpoints:	0
Total Number of Endpoints:	8	Total Number of Endpoints:	8	Total Number of Endpoints:	10

All user specified timing constraints are met.

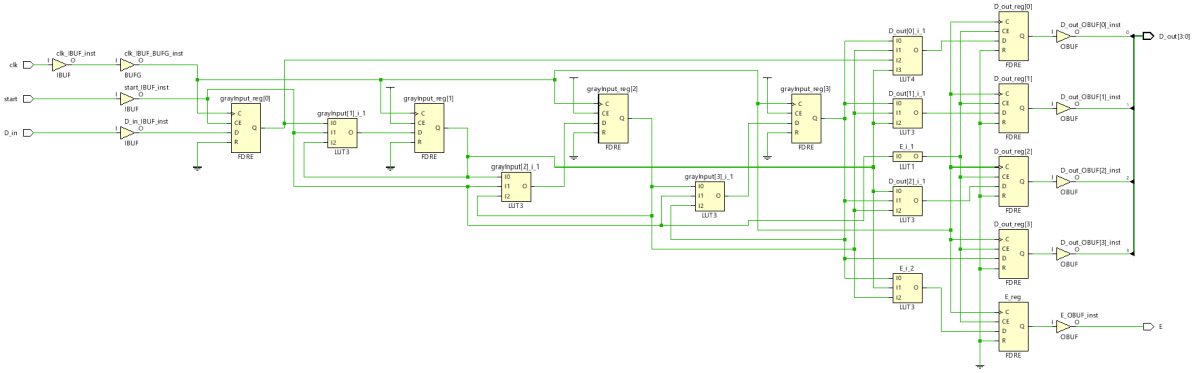
if further increase the frequency to 500MHz (2ns), the WPWS will become negative, which is **bad**:

Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	0.299 ns	Worst Hold Slack (WHS):	0.148 ns	Worst Pulse Width Slack (WPWS):	-0.155 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	-0.155 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0	Number of Failing Endpoints:	1
Total Number of Endpoints:	8	Total Number of Endpoints:	8	Total Number of Endpoints:	10

Timing constraints are not met.

Step 3: Implementation

In this step, P&R is done. Each component generated and converted in synthesis step are assign properly to the corresponding position on FPGA and then wired. The schematic is as below:



Then have a look at power report, which is quite different from the one derived in synthesis:

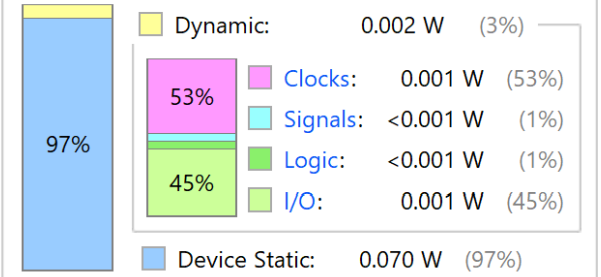
Summary

Power analysis from Implemented netlist. Activity derived from constraints files, simulation files or vectorless analysis.

Total On-Chip Power:	**0.072 W**
Design Power Budget:	**Not Specified**
Power Budget Margin:	**N/A**
Junction Temperature:	**25.4°C**
Thermal Margin:	59.6°C (11.9 W)
Effective ϑJA:	5.0°C/W
Power supplied to off-chip devices:	0 W
Confidence level:	Medium

Launch Power Constraint Advisor to find and fix invalid switching activity

On-Chip Power



We can see that in here the static power take 97% of the total power consumption, this is because in implementation step, each component is assigned to the real and physical x, y position. And this circuit is larger than the one designed in Lab 1, so some components might be far away from each other and then wired, which means that in this real case implementation the actual schematic will be quite different from the one in synthesis (since this one consider the real delay on real wire, while synthesis the position is more like an ideal case).

Then do the timing report to see if the WNS is good. The period still remain as 4ns, we can see in here the WNS even improves compared with the one derived from synthesis step. This might because during this step, Vivado is still trying to optimize the circuit, in here I suppose Vivado succeed in optimization...

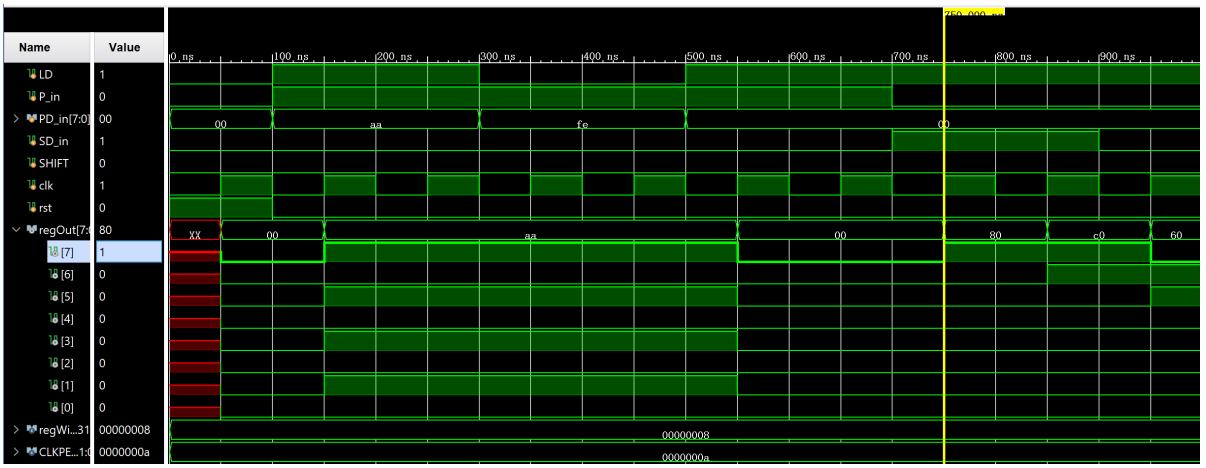
Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	2.428 ns	Worst Hold Slack (WHS):	0.188 ns	Worst Pulse Width Slack (WPWS):	1.500 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	0.000 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0	Number of Failing Endpoints:	0
Total Number of Endpoints:	8	Total Number of Endpoints:	8	Total Number of Endpoints:	10

All user specified timing constraints are met.

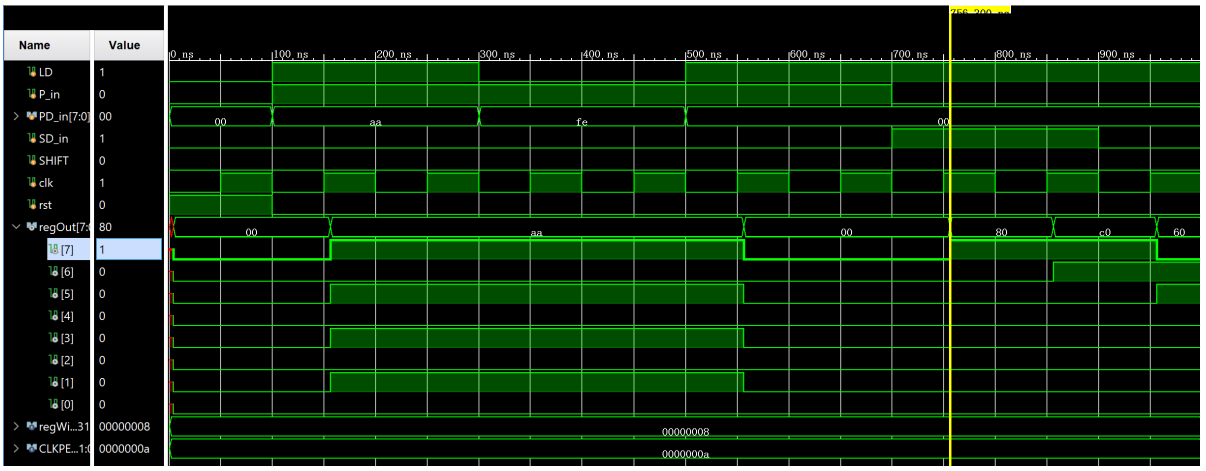
2.2 Shift Register

Step 1: Behavior Simulation

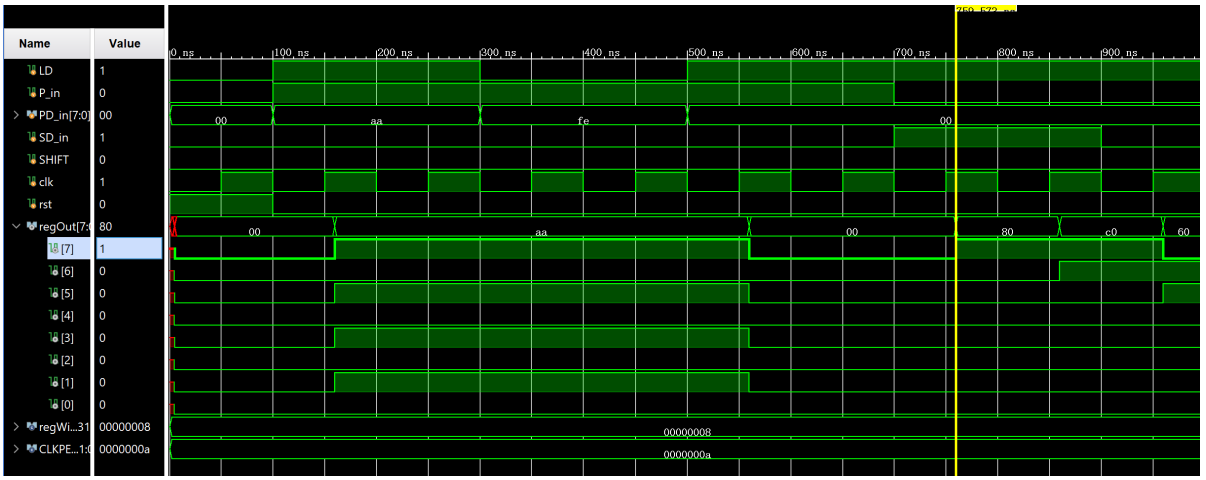
The result of behavior simulation is shown as below:



As what we did in 2.1, the timing simulation of post-synthesis and post-implementation are shown below for better comparison:



post-synthesis timing simulation

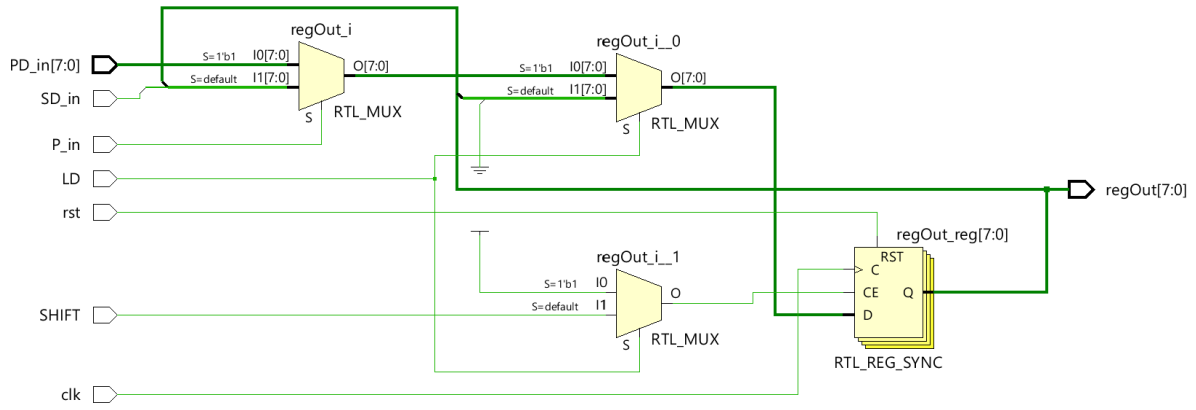


post-implementation timing simulation

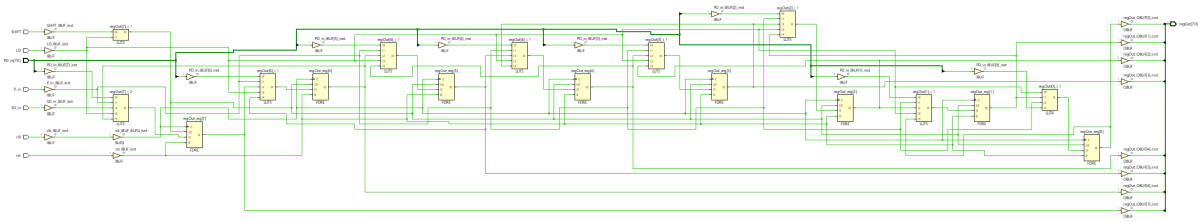
The delay from behavior to post-implementation increases as expected due to the P&R. From 0ns in behavior simulation to 200ns in post-synthesis simulation and ends up at 572ns in post-implementation simulation.

Step 2: Synthesis

Schematic in elaborating step:



Schematic in synthesis step:



with power report:

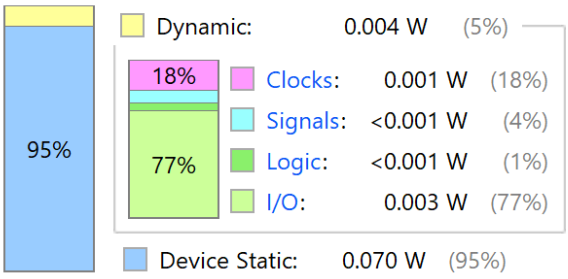
Summary

Power estimation from Synthesized netlist. Activity derived from constraints files, simulation files or vectorless analysis. Note: these early estimates can change after implementation.

Total On-Chip Power:	**0.074 W**
Design Power Budget:	**Not Specified**
Power Budget Margin:	**N/A**
Junction Temperature:	**25.4°C**
Thermal Margin:	59.6°C (11.9 W)
Effective ϑJA:	5.0°C/W
Power supplied to off-chip devices:	0 W
Confidence level:	Low

Launch Power Constraint Advisor to find and fix invalid switching activity

On-Chip Power



Now let's focus on the timing report for this design and compare it with the decoder:

Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	2.339 ns	Worst Hold Slack (WHS):	0.139 ns	Worst Pulse Width Slack (WPWS):	0.500 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	0.000 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0	Number of Failing Endpoints:	0
Total Number of Endpoints:	8	Total Number of Endpoints:	8	Total Number of Endpoints:	9

All user specified timing constraints are met.

Here the period, or let's say the target time is set to 4ns as well. However the WNS in this step is 2.339ns which is larger than 2.299ns in decoder case. This is not what I expected since the **critical path** of this circuit might be longer than the decoder, which is the opposite. Therefore I guess the reasons could be as below:

1. Actually the critical path for the shift register is shorter than decoder, which leads to a smaller actual time and thus bigger WNS.
2. The EDA does some better optimization here.

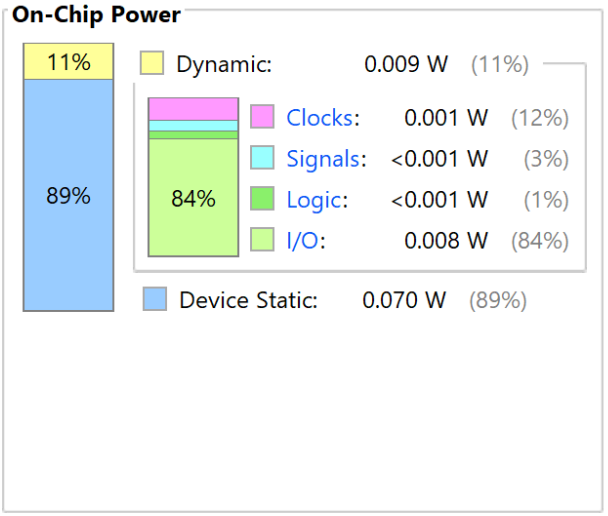
Step 3: Implementation

The power report of implementation is shown as below. The dynamic power consumption is higher than synthesis one.

Power analysis from Implemented netlist. Activity derived from constraints files, simulation files or vectorless analysis.

Total On-Chip Power:	**0.079 W**
Design Power Budget:	**Not Specified**
Power Budget Margin:	**N/A**
Junction Temperature:	**25.4°C**
Thermal Margin:	59.6°C (11.9 W)
Effective ϑJA:	5.0°C/W
Power supplied to off-chip devices:	0 W
Confidence level:	Low

Launch Power Constraint Advisor to find and fix invalid switching activity



And the timing report:

Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	2.697 ns	Worst Hold Slack (WHS):	0.241 ns	Worst Pulse Width Slack (WPWS):	0.500 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	0.000 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0	Number of Failing Endpoints:	0
Total Number of Endpoints:	8	Total Number of Endpoints:	8	Total Number of Endpoints:	9

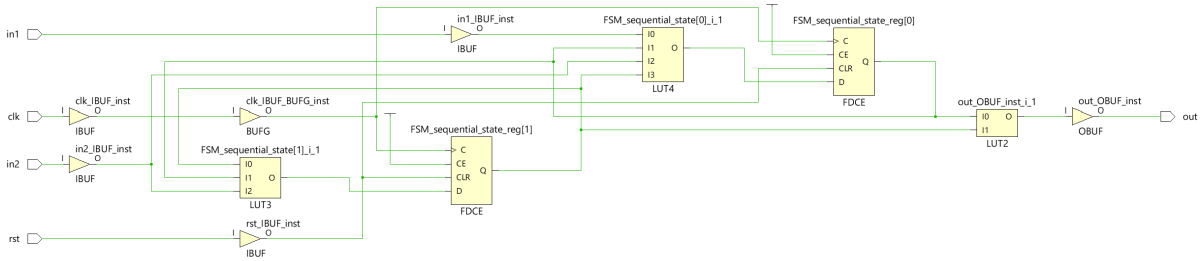
All user specified timing constraints are met.

(To be honest I do not see how is it possible that after P&R the WNS can get better, since I suppose after P&R the wire should be longer thus the critical path and actual time should be longer... leading to a worse WNS. Maybe the power of EDA optimization)

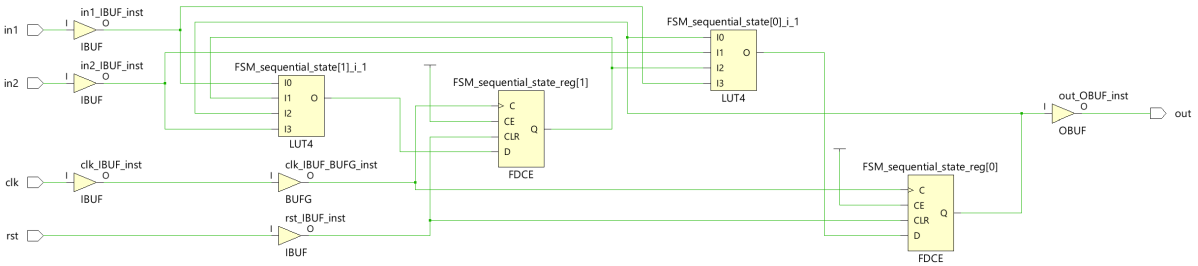
2.3 FSM

1. Comparison between schematic

The schematic of one hot FSM:



The schematic of minimal coding FSM:



We can see that even if the RTL of two FSM are totally the same except the way of coding state, the schematic derived is quite different.

For one hot FSM: 4 states are 0001, 0010, 0100, 1000, respectively.

For min code FMS: 4 states are 00, 01, 10, 11, respectively.

When designing FSM, one thing needs to pay attention to is that before synthesis step, you have to make sure that the sensitive list with one "always" block should be the same otherwise there will be errors and fail to synthesis. For example if you write "always @ (posedge clk, rst)", then you include 2 different sensitive parameters into the list, which will have errors when doing synthesis.

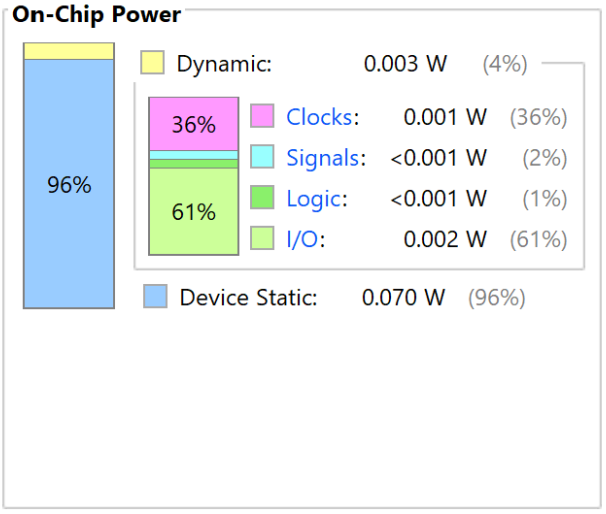
2. Comparison between power

The power report of one hot FSM:

Power analysis from Implemented netlist. Activity derived from constraints files, simulation files or vectorless analysis.

Total On-Chip Power:	**0.073 W**
Design Power Budget:	**Not Specified**
Power Budget Margin:	**N/A**
Junction Temperature:	**25.4°C**
Thermal Margin:	59.6°C (11.9 W)
Effective ϑJA:	5.0°C/W
Power supplied to off-chip devices:	0 W
Confidence level:	Medium

Launch Power Constraint Advisor to find and fix invalid switching activity

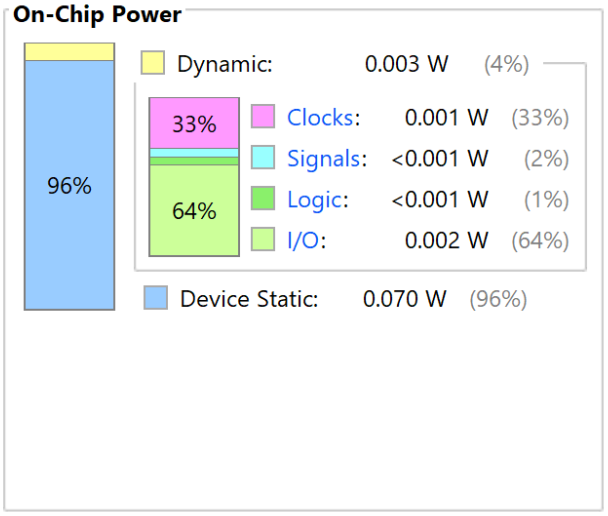


The power report of minimal coding FSM:

Power analysis from Implemented netlist. Activity derived from constraints files, simulation files or vectorless analysis.

Total On-Chip Power:	**0.073 W**
Design Power Budget:	**Not Specified**
Power Budget Margin:	**N/A**
Junction Temperature:	**25.4°C**
Thermal Margin:	59.6°C (11.9 W)
Effective ϑJA:	5.0°C/W
Power supplied to off-chip devices:	0 W
Confidence level:	Medium

Launch Power Constraint Advisor to find and fix invalid switching activity



We can see that the total on-chip power for minimal coding is 0.073W which is almost the same as one hot FSM.

3. Comparison between area

The area report of one hot FSM:

Site Type	Used	Fixed	Available	Util%
Slice LUTs*	2	0	20800	<0.01
LUT as Logic	2	0	20800	<0.01
LUT as Memory	0	0	9600	0.00
Slice Registers	2	0	41600	<0.01
Register as Flip Flop	2	0	41600	<0.01
Register as Latch	0	0	41600	0.00
F7 Muxes	0	0	16300	0.00
F8 Muxes	0	0	8150	0.00

The area report of minimal coding FSM:

Site Type	Used	Fixed	Available	Util%
Slice LUTs*	1	0	20800	<0.01
LUT as Logic	1	0	20800	<0.01
LUT as Memory	0	0	9600	0.00
Slice Registers	2	0	41600	<0.01
Register as Flip Flop	2	0	41600	<0.01
Register as Latch	0	0	41600	0.00
F7 Muxes	0	0	16300	0.00
F8 Muxes	0	0	8150	0.00

We can see that the number of FFs is the same in both minimal code FSM and one hot FSM. However the LUT used in minimal FSM is less than one hot FSM, which leads to a less area consumption. Since this is a simple and small FSM with only 4 states, the difference is not that significant. When the size of the FSM becomes larger and state increases, the area consumption for one hot FSM could be significantly higher than minimal code FSM.

4. Comparison between timing

The timing report of one hot FSM:

Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	2.832 ns	Worst Hold Slack (WHS):	0.260 ns	Worst Pulse Width Slack (WPWS):	1.500 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	0.000 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0	Number of Failing Endpoints:	0
Total Number of Endpoints:	2	Total Number of Endpoints:	2	Total Number of Endpoints:	3

All user specified timing constraints are met.

The timing report of minimal coding FSM:

Setup		Hold		Pulse Width	
Worst Negative Slack (WNS):	2.427 ns	Worst Hold Slack (WHS):	0.261 ns	Worst Pulse Width Slack (WPWS):	1.500 ns
Total Negative Slack (TNS):	0.000 ns	Total Hold Slack (THS):	0.000 ns	Total Pulse Width Negative Slack (TPWS):	0.000 ns
Number of Failing Endpoints:	0	Number of Failing Endpoints:	0	Number of Failing Endpoints:	0
Total Number of Endpoints:	2	Total Number of Endpoints:	2	Total Number of Endpoints:	3

All user specified timing constraints are met.

Both reports are derived under time period is 4ns (frequency is 250MHz). We can see that the WNS of minimal code FSM is larger than one hot FSM, which means that there are more space to improve the operating frequency for one hot FSM than minimal code FSM. In other word, when put both performance to the limit, the performance of minimal code FSM will be worse than one hot FSM. This can be understood, since the bit flip rate of one hot FSM is smaller then minimal code FSM, which means that the speed of transfer between states can be faster in one hot FSM, leading to a better performance.

5. Trade-off between PPA

From the discussion before, we can know that:

FSM type	State bit width	Power consumption	Performance (max. frequency)	Area
One Hot FSM	Longer (4-bit)	Same (total on-chip power = 0.073W)	Higher (WNS = 2.832ns @ 250MHz)	Larger (LUTs = FFs = 2)
Minimal Code FSM	Shorter (2-bit)	Same (total on-chip power = 0.073W)	Lower (WNS = 2.427ns @ 250MHz)	Smaller (LUTs = 1, FFs = 2)

Therefore:

- if one cares about area, then use minimal code FSM.
- if one cares about performance, then use one hot FSM