# Memory II, FSMs, Pipelining

6.205 Fall 2024

## Memory Example!

• In the first part of week 05 you're going to be displaying popcat on your FPGA over video.



- We need to store popcat
- Popcat is a 256X256 24 bit color image.
- How to encode with memories?

#### Popcat

- We could build a *24bit-wide* memory that has 256x256 (65,536) entries (deep) in it (one for each pixel)
- Math: 256x256x24 = 1,572,864 bits
- That's greater than >50% of our memory on the S50 FPGA.



## **Strategy**

- Images are large and take up lots of memory
- Want to save space, and store image using less memory
- Many images don't express every one of the 2<sup>24</sup> "true" colors.
- Why waste the space storing an unused possibility?
- So pick the *N* most popular and only display them:
	- You can store them using  $ceil(log_2(N))$  bits (save space)
	- Then use a color table to look up what full color (24 bit value) that corresponds to!

#### Color Lookup Table

- So use two memories:
	- 8-wide memory with 65536 entries for each pixel encoding one of 256 colors
	- 24-wide memory with 256 entries entries encoding those colors



#### Storage Savings



256 X 256 image @ 24 bits per pixel is: 256 X 256 X 24 bits = 1.572 Mbits (196.6 kBytes)

24 bit – 16M colors



8 bit – 256 colors

256 X 256 image with 8 bit color map: 256 X 256 X 8 bits + 256 X 24 = 0.530432 Mbits (66.304 kBytes)

## Keep Going As Needed



 $24 \text{ bit} - 16 \text{ M}$  colors  $8 \text{ bit} - 256 \text{ colors}$  6 bit – 16 colors



4 bit – 16 colors





 $2 \text{ bit} - 4 \text{ colors}$  1 bit – 2 colors

10/3/24 https://fpga.mit.edu/6205/F24 7





## Additional Tricks Can be Played

• Dithering, in particular can help with this problem, but we'll go into that in a future week.

## What about Latency?

- Yes What about latency.
- These things we just built are memories!
- Memory of any scale usually has latency involved with it
- Xilinx BRAM has how many cycles latency on a read?
- 2 cycles
- Lab/Week 5 will investigate

# FSM Modularity

#### Toward FSM Modularity

• Consider the following abstract FSM:



- Suppose that each set of states  $a_x...a_x$  is a "sub-FSM" that produces exactly the same outputs.
- Can we simplify the FSM by removing equivalent states? *No! The outputs may be the same, but the next-state transitions are not.*
- This situation closely resembles a procedure call or function call in software...how can we apply this concept to FSMs?

Acknowledgements: Rex Min

### The Major/Minor FSM Abstraction



- Subtasks are encapsulated in minor FSMs with common reset and clock
- Simple communication abstraction:
	- START: tells the minor FSM to begin operation (the call)
	- BUSY: tells the major FSM whether the minor is done (the return)
- The major/minor abstraction is great for...
	- Modular designs (*always* a good thing)
	- Tasks that occur often but in different contexts
	- Tasks that require a variable/unknown period of time
	- Event-driven systems

#### Inside the Major FSM



#### Inside the Minor FSM



10/3/24 https://fpga.mit.edu/6205/F24 14

#### Optimizing the Minor FSM

Good idea: de-assert BUSY one cycle early



*So make sure you if you do this, that last state always happens and always happens for one cycle* 

#### A Four-FSM Example



#### Four-FSM Sample Waveform



### Alternative to Busy Signals

- As an alternative to busy signals sometimes just having a single-cycle "valid" signals is sufficient.
- If the downstream systems involved are stateful enough to be able to keep track of various system's this can work
- Or you can do both. Depends on your design

#### Alternative to Busy Signals (Single-cycle asserts)



## A Divider

- The Divider from when we first talked about FSMs is an example of an system which might be a minor FSM in part of a larger major's algorithm
- Many things need division, but it would suck to have to rewrite it repeatedly.
- We want you to get practice with that in Week 5's lab



## Center of Mass Calculation in Lab05

- You will write a center-of-mass calculator that is best thought of as an FSM.
- For each frame of video:
	- Sum the x location and y location of every active pixel you come across
	- Keep track of how many pixels you've encountered
	- At end of frame (or beginning of next one, divide the two sums by the number of active pixels
	- This will give an average X,Y
- Division takes time!
- Need to create a major/minor FSM system

#### Lab 05 Center of Mass



- C.O.M. will be in a "data collection state" during the active portion of a video frame
- When the frame's active part is done, it needs to calculate the average x,y position of the "hot" pixels it has observed.
- To divide, the C.O.M. module hands the values it needs divided off to two separate dividers.
- C.O.M. waits on them monitoring their BUSY signals
- They can do division separately (in parallel)
- When done, they report back to the C.O.M with their result
- C.O.M. reports to outside world its calculation

# Pipelining

How to make sure signals are balanced going through a sequence of operations.

## Performance Metrics



- Latency (L):
	- time between arrival of new input and generation of corresponding output.
	- For purely combinational circuits this is just  $t_{\text{PD}}$ .
- Throughput (T):
	- Rate at which new outputs appear.
	- For purely combinational circuits this is just  $1/t_{\text{PD}}$  or  $1/L$ .

#### Performance of Combinational For combinational logic: Logic  $L = t_{PD}$  $X$   $F$   $H$   $P(X)$  $T = 1/t_{PD}$ We can't get the answer faster, but are H we making effective use of our hardware at all times? G X  $F(X)$ **XXXXXX**  $G(X)$ XXXXXXXXXXX  $P(X)$ XXXXXXXXXXXXXXXXXXXXXX

F & G are "idle", just holding their outputs stable while H performs its computation

### Retiming: A useful transform

*Propagation delays indicated by numbers:*



#### Retiming: A useful transform



*Assuming ideal registers: i.e.,*  $t_{PD} = 0$ ,  $t_{SETUP} = 0$ 

## Pipeline Diagrams



The results associated with a particular set of input data moves diagonally through the diagram, progressing through one pipeline stage each clock cycle.

## Pipeline Conventions

- a K-Stage Pipeline ("K-pipeline") is an acyclic circuit having exactly K registers on every path from an input to an output.
- a COMBINATIONAL CIRCUIT is thus a 0-stage pipeline.

## Pipeline Conventions

- CONVENTION:
	- Every pipeline stage, hence every K-Stage pipeline, has a register on its OUTPUT (not on its input).
- ALWAYS:
	- The CLOCK common to all registers must have a period sufficient to cover propagation over combinational paths PLUS (input) register  $t_{\text{PD}}$  PLUS (output) register  $t_{\text{SFTUP}}$ .

 $t_{\text{PD},\text{reg1}} + t_{\text{PD},\text{logic}} + t_{\text{SETUP},\text{reg2}} \leq t_{\text{CLK}}$ 

The LATENCY of a K-pipeline is K times the period of the clock common to all registers.

The THROUGHPUT of a K-pipeline is the frequency of the clock.

## ILL-formed Pipeline



For what value of K is the following circuit a K-Pipeline? \_

Problem:

Successive inputs get mixed: e.g.,  $B(A(X_{i+1}), Y_i)$ . This happened because some paths from inputs to outputs have 2 registers, and some have only 1!

This CAN'T HAPPEN on a well-formed K pipeline!

10/3/24 https://fpga.mit.edu/6205/F24 31

## Pipelining

*Let's say we want t<sub>clk</sub> to be 8ns* 



## Another Thing to Pipeline

*Let's say we want t<sub>clk</sub> to be 8ns* 



#### Step 1:

Add a register on the output.

#### Step 2:

From register. Draw a contour backwards that includes as much of the circuit that will fit inside required period. Add registers

Repeat until satisfied with T. Look for redundant registers

#### STRATEGY:

Focus your attention on placing pipelining registers around the slowest circuit elements (BOTTLENECKS).



*Assuming this interfaces with other modules that have registered outputs the input will chain will be ok (<= 8ns)*

## Another Pipeline Example





OBSERVATIONS:

- 1-pipeline improves neither L or T.
- T improved by breaking long combinational paths, allowing faster clock.
- Too many stages cost L, don't improve T.
- Back-to-back registers are often required to keep pipeline well-formed.
- *Better throughput here means we can run at higher clock rate*


# How often should you be adding FlipFlops in your FPGA?

- This comes with experience and getting to know your system.
- Most of what you want to do really is some form of math.
- So knowing how much math you can do in a clock cycle is useful

# The Complexity of Math Operations

Let's look at some basic math circuits:

 $+, -, *, /$ 

# "Full Adder" building block



# Adder: a circuit that does addition

Here's an example of binary addition as one might do it by "hand" :



If we build a circuit that implements one column:



we can quickly build a circuit to add two 4-bit numbers…



#### Subtraction:  $A-B = A + (-B)$

Using 2's complement representation:  $-B = \gamma B + 1$ 

So let's build an arithmetic unit that does both addition and subtraction. Operation selected by control input:



 $\sim$  = bit-wise complement

Ð

B

# "Full Adder" building block



Can also rewrite the carry path as:  $c_{out} = (a \& c_{in}) | (b \& c_{in}) | (a \& b)$ 

# Speed:  $t_{\text{PD}}$  of Ripple-carry Adder

 $C_O = AB + AC_I + BC_I$ 



Worst-case path: carry propagation from LSB to MSB, e.g., when adding 11…111 to 00…001.

$$
t_{\text{PD}} = (N-1)^*(t_{\text{PD,OR}} + t_{\text{PD,AND}}) + t_{\text{PD,XOR}} \approx \Theta(N)
$$
\nCl to CO

\nCl<sub>N-1</sub> to S<sub>N-1</sub>

\ndunder of the operator  $t_{\text{PQ,AND}}$ 

\nuence of the operator  $t_{\text{PQ,ND}}$ 

 $\Theta(N)$  is read "order N" : means that the latency of our adder grows at worst in proportion to the number of bits in the operands.

# The Carry Path Becomes Limiting

• Solution is the Carry-Look-ahead Adder:

$$
c_1 = G_0 + P_0.c_0
$$
  
\n
$$
c_2 = G_1 + P_1.G_0 + P_1.P_0.c_0
$$
  
\n
$$
c_3 = G_2 + P_2.G_1 + P_2.P_1.G_0 + P_2.P_1.P_0.c_0
$$
  
\n
$$
c_4 = G_3 + P_3.G_2 + P_3.P_2.G_1 + P_3.P_2.P_1.G_0 + P_3.P_2.P_1.P_0.c_0
$$

Can do some factoring/redesign and cutdown on  $t_{\text{pd}}$  of the carry path

$$
G_i = a_i.b_i
$$
  

$$
P_i = a_i \oplus b_i
$$



https://www.ece.uvic.ca/~fayez/courses/ceng465/lab\_465/project1/adders.pdf

10/3/24 https://fpga.mit.edu/6205/F24 44

#### Logic Slices Can Add/Subtract

• Can synthesize the addition of two 4 bit numbers with fast carry

$$
A_3A_2A_1A_0
$$
  
+
$$
B_3B_2B_1B_0
$$



#### Add/Subtract on the FPGA

- + and can be done combinationally very quickly:
	- 32 bit add can be done in a clock cycle (<10 ns) pretty easily
	- Several smaller adds (A+B+C+D) can be done in clock cycle as well (10 ns)
- CLBs (the generic function generators, of which we have a lot) are capable of being chained together to allow large adds.

# Slices can stack to give more bits

 $A_{11}A_{10}A_9A_8A_7A_6A_5A_4A_3A_2A_1A_0$  $+B_{11}B_{10}B_9B_8B_7B_6B_5B_4B_3B_2B_1B_0$ 



#### + or - in Verilog

- Generally + or on its own will get synthesized using logic slices unless specified
- Very large additions or subtractions may start to take too long!\*
- But doing a couple 32 bit adds in a 10 ns cycle should be possible…

*\*too is really with respect to a clock. If you're running on a 10 MHz clock, then things are different!* 

# But also the stuff around it matters too!

- Keep track of the stuff before and after your math.
- If you have a ton of if/else/ifs…or if you have a super-deeply nested if/if/if/ chain, all that stuff requires logic too.

#### Also Case Statements are Good

• If/elses and even parallel if's as shown on the previous page get encoded as priority logic



Fig 1

https://www.kevnugent.com/2020/10/22/verilog-blogpost\_002/

#### Also Case Statements are Good

• If logic can be structured without priority, then do it! Can yield simpler underlying logic.



# The stuff around it matters too!

- Keep track of the stuff before and after your math.
- If you have a ton of if/else/ifs…or if you have a super-deeply nested if/if/if/ chain, all that stuff requires logic too.
- **Also think about the stuff being used to calculate the if/else stuff.**



# Multiplication on the FPGA

- Multiplication can be done on the FPGA on 2's complement numbers
- Takes more time:
	- Depending on size of operands may/may not be doable in one clock cycle
- Where possible try to get away with bit shifts and adds.

# Multiplications with shifts

- <<1 is multiply by 2
- >>1 is divide by 2
- Can do a lot with this if get creative

```
\mathbf 1logic [7:0] x;
\overline{2}3 logic [7:0] y; //want this to be seven times X
   assign y = (x<<2) + (x<<1) + x;
```
• Vivado can be pretty good at figuring these things out for you, but largely only for constants.

#### Generic Digital Multiplication



 $a_1b_0+a_0b_1$   $a_0b_0$   $\leftarrow$  Product

*In base 2 multiplication these are all very simple calculations done with XOR*

*Some really cool factoring can be done to make the overall propagation delay of a multiplier relatively short, though there's a lot of logic in it\**

\*Lecture on Multiplier architectures: https://inst.eecs.berkeley.edu/~eecs151/sp18/files/Lecture21.pdf

10/3/24 https://fpga.mit.edu/6205/F24 56

# DSP Blocks

- Add-then-multiply is a common operation chain in many things, particularly Digital Signal Processing
- FPGA has dedicated hardware multiplier modules called DSP48 blocks on it
	- 150 of them on Urbana FPGA
	- Capable of single-cycle multiplies
- Can get inferred from using \* in your Verilog that isn't a power of 2:
	- x\*y, for example, will likely will result in DSP getting used
	- May take a full clock cycle so would need to budget tiing accordingly
- Can infer multiple for larger multiplies

# DSP48 Slice (High Level)



Figure 1-1: Basic DSP48E1 Slice Functionality

*Much of the benefit/speed of this module comes from the hardwired internal routing, keeping it very fast. This device is not as generalized as a LUT/logic cell. It can only do a subset of math operations*

https://www.xilinx.com/support/documentation/user\_guides/ug479\_7Series\_DSP48E1.pdf

# How much multiply can do?

- At 100 MHz on these boards, I'd aim for one 32 bit multiply per clock cycle (it'll use several DSP blocks to achieve that)
- Anything more is pushing it
- If you run out of DSP blocks, it'll revert to using the generic logic…and this will become a harder problem to satisfy
- Smaller multiply-adds you can maybe get away with in one clock cycle.

# Division

- The outlier in the  $+ * /$  set...
- Division is a significantly harder math operation to do compared to multiplication
- Where possible try to avoid
- Try to divide by powers of 2 (use right shift)!
- If you can't avoid we must do it. (week 05)

#### Lab 05 Center of Mass



- C.O.M. will be in a "data collection state" during the active portion of a video frame
- When the frame's active part is done, it needs to calculate the average x,y position of the "hot" pixels it has observed.
- To divide, the C.O.M. module hands the values it needs divided off to two separate dividers.
- C.O.M. waits on them monitoring their BUSY signals
- They can do division separately (in parallel)
- When done, they report back to the C.O.M with their result
- C.O.M. reports to outside world its calculation

# One "Bad" Attempt at Division

- In previous lecture looked at \*what\* this actually builds
- We can ask Vivado to synthesize division logic for us, and it actually will do it.
- This code constrains the act of division to having to exist between two clock edges.:

```
module top_level(
      input wire clk_100mhz, //clock @ 100 mhz
      input wire [15:0] sw, //switches
      input wire btnc, //btnc (used for reset)
      input wire btnu, //btnc (used for reset)
      input wire btnl, //btnc (used for reset)
      output logic [15:0] led //just here for the funs
      );
   logic old_btnl;
   logic old_btnu;
   logic old_btnc;
   logic [15:0] quotient;
   logic [15:0] dividend;
   logic [15:0] divisor;
  assign led = quotient;always ff @(posedge clk 100mhz)begin
   old\_btnl \leq btni:
    old btnu \leq btnu:
    old btnc \leq btnc;
   end
   always_ff @(posedge clk_100mhz)begin
    if (btnu & ~old btnu)begin
       quotient<= dividend/divisor; //divide
     end
     if (btnc & ~old_btnc)begin
       dividend <= sw; //divide //load dividend
     end
    if (btnl \& ~old_btnl)begin
       divisor <= sw; //divide //load divisor
```

```
 end
endmodule
```
end

#### Circuit Built:



### Build the Stupid Divider

*Violates timing!*





# Now Do same Thing With 32 bits:





*\*See lecture code for full implementation and build. (divider0)*

10/3/24 https://fpga.mit.edu/6205/F24 65

#### Build the Stupider Divider

Phase 20 Post Router Timing INFO: [Route 35-20] Post Routing Timing Summary | WNS=-72.004| TNS=-1004.354| WHS=0.227 | THS=0.000 | Phase 20 Post Router Timing | Checksum: 1d10fc4d8

#### 2. Slice Logic Distribution







*\*See lecture code for full implementation and build. (divider0)*

10/3/24 https://fpga.mit.edu/6205/F24 67

#### So conclusions

- +, -, \* can be done in a clock cycle with exceptions
- Watch out for flow-control logic…that can start to stack up
- / will never happen in one clock cycle. Accept that.
- Similar other things like square root, cosine, etc…those need clock cycles…or if you absolutely need those in one/two cycles, you do a lookup table of pre-computed values (takes huge amounts of memory)

# Back on pipelining…

- If we have time… A white blip
- In lab 05, early on you may see an artifact on popcat



# (Two registers coming from delay in memory access/read)

- Monitor drawing based on vsync, hsync, blank,
- But what image rom is giving it is 5 clock cycles behind
- At start of Death Star nothing in the "pipeline" yet







#### How to Fix?

• Delay the other signals so everybody is the same



*Turn the whole thing into a 5-stage pipeline!*
# Pipelining

- Pipeline in Verilog!
- Make sure other things are protected too!

```
logic hs_pip[4:0];
logic vs_pip[4:0];
logic b_pip[4:0];
always_ff@(posedge clk_in)begin
  hs\_pip[0] \leq hsync_in;vs\_pip[0] \leq vsync_in;b pip[0] \leq blank in;
  for (int i=1; i<5; i = i+1) begin
    hs_pip[i] <= hs_pip[i-1];
    vs\_pip[i] \leq vs\_pip[i-1];
    b pip[i] \leq b pip[i-1];
   end
end
assign hsync out = hs pip[4];
\textsf{assign} vsync_out = vs_pip[4];
assign blank out = b pip[4];
```
### Final Project Ideas

- Things with video and/or related topics are very "relevant" to FPGAs
- You have to move and process very large amounts of data with demanding timing.
- This is something software often cannot on its own.

### Live Pong



#### Glow Trails



### DigiEyes



## Final Project Info released by tomorrow

- Start Teaming!
- Teams of 2 or 3!