

# Computer Organization

Lab13 Cache(1), Load Program on CPU

Pipeline tips(optional)





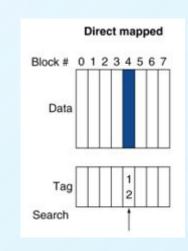
- > Cache (1)
  - Direct Mapped Cache(implementation)
    - > Components, Read, Write

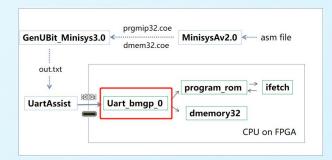
# > Load program on CPU

➤ Getting the updated coe files through **Uart** port (2 tools, Modification on CPU)

# Pipeline related tips

- Register( storage, stall, flush)
- > Forwarding
- > Determine the time of a cycle

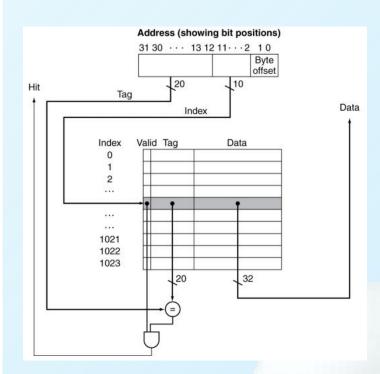






# **Direct Mapped Cache**

- Direct mapped cache
  - > ONE data in MEMORY is mapped to ONLY ONE location in CACHE.
  - Location determined by the ADDRESS:
    - > The **lower bits of address** define the place of the unit(aka **block**) in the cache.
    - > The **higher bits of address** are called **tag** which would be **stored in the cache unit**(aka cache block).

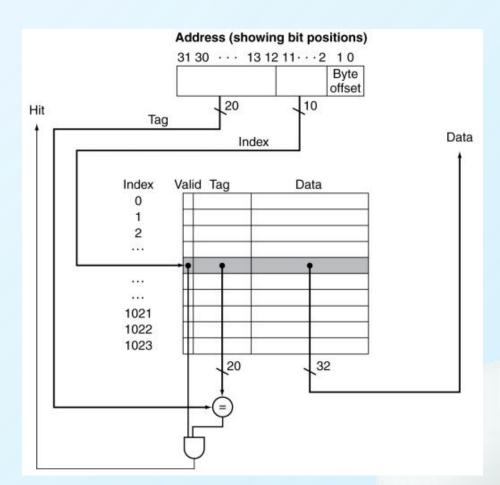




# Direct Mapped Cache continued

#### Componets in Cache

- Valid(1bit)
  - ➤ Initial value is 1′ b0
  - > Turns to 1' b1 while the cache unit has been written which means there has been data in the cache unit
- > **Tag:** Higer bits of the address
- > Data: Store the data which has been cached

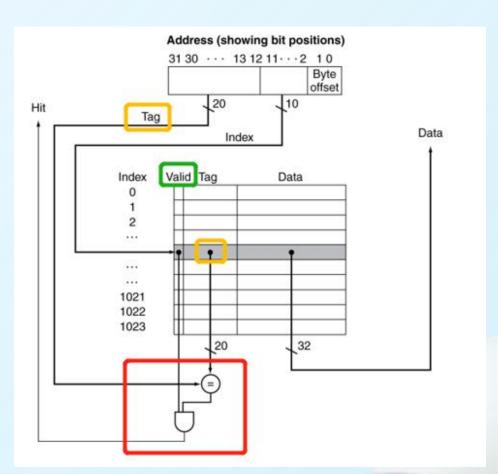




# Direct Mapped Cache continued

#### Cache Hit vs Cache Miss

- > Hit
  - > Find the cache unit based on the 'Index' of 'Address'
  - > Find the 'valid' and 'Tag' parts from the cache unit
  - ➤ If 'valid' is 1 and the 'Tag' is the same as the tag part(higer bits) of 'Address', then it is cache hit.
- Miss: If not hit, then miss
- hit ratio = hits / accesses

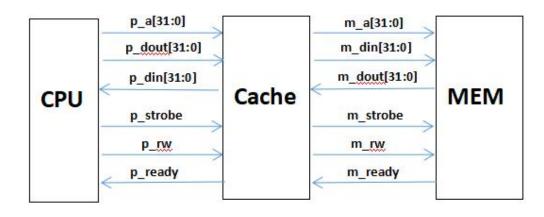




# Implement Direct Map Cache(ports)

```
module cache
#(parameter A WIDTH=32, parameter C INDEX=10, parameter D WIDTH=32)
(clk,resetn,
p a, p dout, p din, p strobe, p rw, p ready,
m a, m dout, m din, m strobe, m rw, m ready);
    input clk, resetn;
     input [A_WIDTH-1:0] p_a; //address of memory tobe accessed
    input [D_WIDTH-1:0] p_dout; //the data from cpu
     output [D WIDTH-1:0] p din; //the data to cpu
    input p_strobe; // 1 means to do the reading or writing
    input p_rw; // 0:read, 1:write
     output p ready; // tell cpu, outside of cpu is ready
                               //address of memory tobe accessed
     output [A WIDTH-1:0] m a;
     input [D_WIDTH-1:0] m_dout; //the data from memory
    output [D_WIDTH-1:0] m_din; //the data to memory output m_strobe; //1 means to do the reading or writing
     output m rw;
                       //0:read, 1:write
                                   //memory is ready
     input m ready;
```

**Q.** What's the relationship between A\_WIDTH, **C\_INDEX** and D\_WIDTH?



#### TIPS:

parameter in verilog makes the design more **flexiable**, which is highly recommended.



# Implement Direct Map Cache(componets)

#### **Build the cache**, complete the following statement

```
// d_valie is a piece of memory stored the valid info for every block reg d_valid [ 0 : complete code here p7-1];

// T_WIDTH is the width of 'Tag' localparam T_WIDTH = complete code here p7-2;

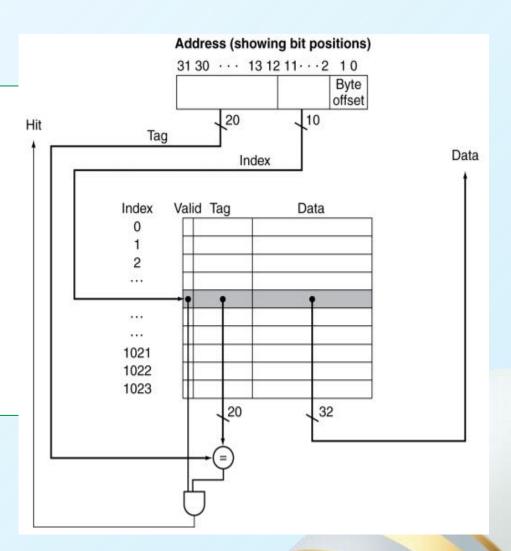
//d_tags is a piece of memory stored the tag info for every block reg [T_WIDTH-1: 0] d_tags [0 : complete code here p7-1];

//d_data is a piece of memory stored the data for every block reg [D_WIDTH-1:0] d_data [0 : complete code here p7-1];
```

A WIDTH: the width of 'Address', here its value is 32.

C INDEX: the width of 'Index', here its value is 10.

T\_WIDTH: used as the width of 'Tag'. 'Byte offset': 2 bit-width





### Implement Direct Map Cache (cache hit vs cache miss)

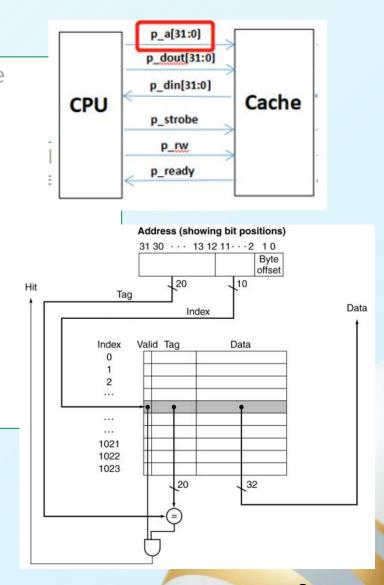
#### Complete the following code to implement cache hit

```
// d valie is a piece of memory stored the valid info for every block in cache
// d tags is a piece of memory stored the tag info for every block in cache
// d data is a piece of memory stored the data for every block in cache
wire [C INDEX-1:0] index = p_a[C_INDEX+1:2];
wire [T WIDTH-1:0] tag = p_a[ A_WIDTH-1 : C_INDEX+2];
wire valid = d valid[index];
wire [T WIDTH-1:0] tagout = d tags[index];
wire [D WIDTH-1:0] c dout = d data[index];
//cache control
wire cache hit = complete the code here p8
wire cache miss = ~cache hit;
```

A\_WIDTH: the width of 'Address', here its value is 32.

C\_INDEX: the width of 'Index', here its value is 10.

T WIDTH: used as the width of 'Tag'. 'Byte offset': 2 bit-width

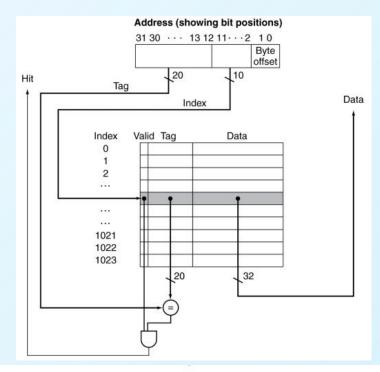


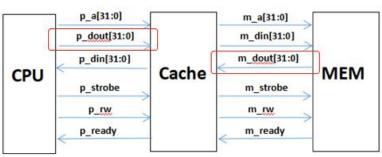


# Implement Direct Map Cache(cache write)

Complete the following code to implement cache write

```
wire c write = p rw | cache miss & m ready;
always @ (posedge clk, negedge resetn)
     if( resetn == complete code here p9-1 ) begin
          integer i;
          for (i=0; i<(complete code here p9-2); i=i+1)
               d valid[i] <=1'b0;
     end
     else if(c write==1'b1)
          d valid[index] <=1'b1;
always @ (posedge clk)
     if(c write==1'b1) d tags[index] <= tag;
wire sel in = p rw;
wire [D WIDTH-1:0] c din = complete code here p9-3;
always @ (posedge clk)
     if(c_write==1'b1) d_data[index] <= c din;</pre>
```







# Implement Direct Map Cache(memory write)

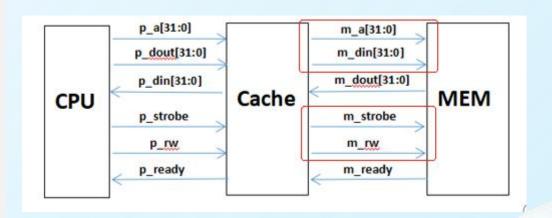
Complete the following code to implement memory write (write\_through)

```
// write data (m_din) to the memory unit specified by m_a
    assign m_a = p_a;

assign m_din = p_dout;

// p_strobe (1 means to do the reading or writing, 0 means else)
// m_strobe (1 means to do the reading or writing, 0 mean else)
// p_rw, m_rw (0:read, 1:write)

assign m_strobe = p_strobe & (p_rw | cache_miss);
assign m_rw = complete code here p10;
```





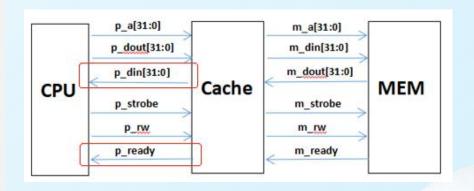
# Implement Direct Map Cache(CPU read)

#### Complete the following code to implement CPU read

```
//read data to CPU
// if cache hit, read c_dout from cache to p_din, else read m_dout to p_in
// wire [D_WIDTH-1:0] c_dout = d_data[index];
    wire sel_out = cache_hit;

assign p_din = complete code here p11 ? c_dout: m_dout;

assign p_ready = ~p_rw & cache_hit | (cache_miss | p_rw) & m_ready;
```





- > 1. Complete the code of the Directly Mapped Cache(page 7-11)
- > 2. Build a testbench to verify the function of the Directly Mapped Cache.



# How to make CPU work on a new program?

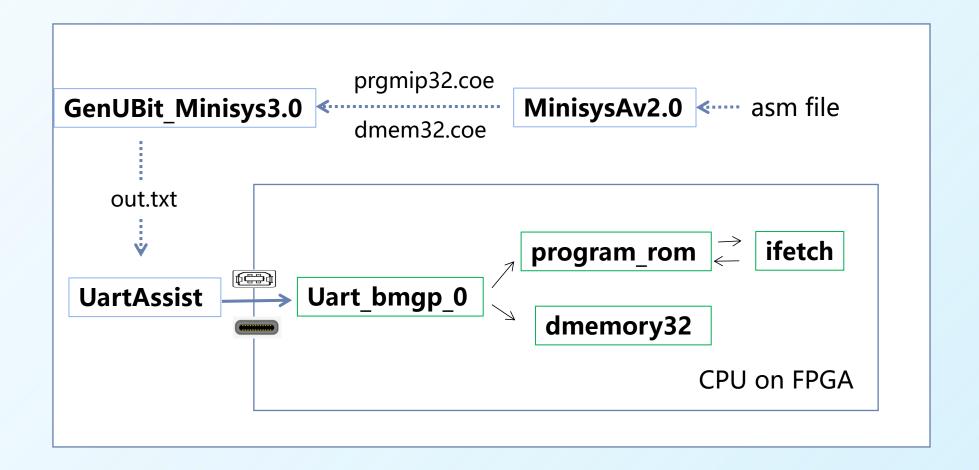
- How to make CPU work on a new program?
  - new program -> new machine code and initial data (new coe(s))
  - > Solution1:
    - > update the **ProgramRom** and **DataRAM** of CPU with new coe(s) -> re-generate bitstream of updatad CPU -> re-program the FPGA chip by updated bitstream file

#### > Solution2:

- While CPU work on Communicate mode: CPU get the new coe(s) by uart port, then rewrite its 'PrgramROM' and DataRAM
- While CPU work on Normal mode: work on the updated program
- 2 tools(GenUBit\_Minisys3.0, UartAssist) and some modifications on the CPU are needed for solution2.



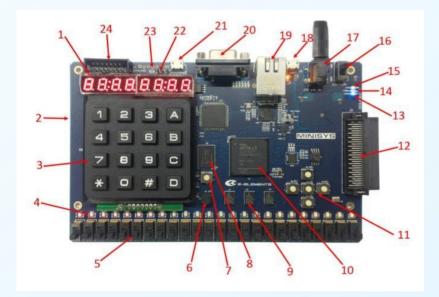
# Solution2 on load program on CPU



"GenUBit\_Minisys3.0", "UartAssist" and "UARTCoe\_v3.0" could be found in the "uart\_tools.rar" of "labs/lab13\_uart" on BlackBoard site



# Uart Interface on Minisys Board and EGO1 Board





**port 18**(as shown on the left hand) is the USB to UART interface.

For **Minisys board(new vesrion**, the type of USB\_Jtag interface is **typeC**), USB\_Jtag and USB to UART interface share the same port.



For **EGO1 board**, USB\_Jtag and USB(typeC) to UART interface share the same port.

The handbook of Minisys board and EGO1 board could be found in the "Handbook\_of\_Minisys\_EGO" of "labs" on BlackBoard site



# Changes on Single Cycle CPU

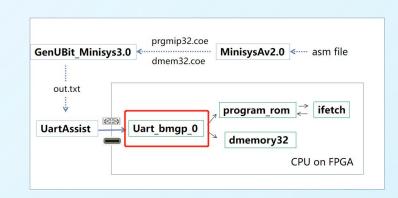
- 1. Two working modes on the CPU
  - > Normal mode vs Uart Communication mode
- 2. A new module(Uart\_bmgp\_0) which works as Uart interface
- 3. A new clock for uart communication
- 4. Changes
  - **>** 3-1) **CPU top:**

New module, new ports, new internal connection and new logic

> 3-2) Changes on **Data-meomroy** 

Working mode: Normal mode vs Uart Communication mode

- > 3-3) Changes on IFetch
  - > Change IP core "prgrom" from ROM to RAM
  - > Working mode: Normal mode vs Uart Communication mode
  - > Separate "prgrom" from IFetch (optional)

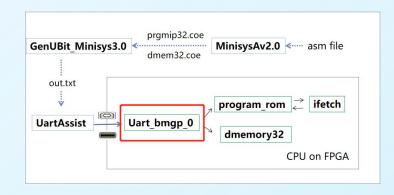




### Add an IP core Which Processes Uart Data

#### Step1: Add the **IP core** to IP catalog(User Repository) of vivado.

- The Communication between this IP core and Uart port:
  - **Receive** data from Uart port and forward to data-memory and instruction-memory
  - Send data back to uart port to info that all the data has been received.

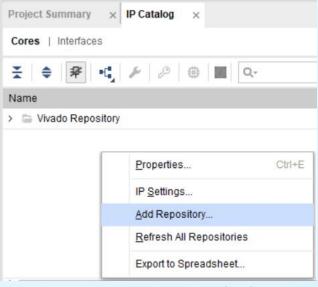


Step1-1: download the ziped IPcore file, then unzip it



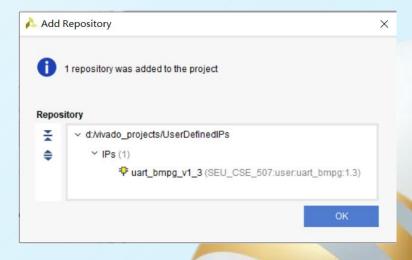
The IP core could be found in the "labs/lab13\_uart" on BlackBoard site

Step1-2: open "IP Catalog" pane, right click on the blank space and select "Add Repository"



CS214 wangw6@sustech.edu.cn

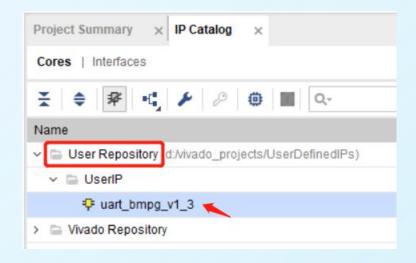
Step1-3: in "Add Repository" pane select the diretory where the upzipped IPcore is placed. vivado would detect the IPcore and pop up the following prompt window, click "OK".

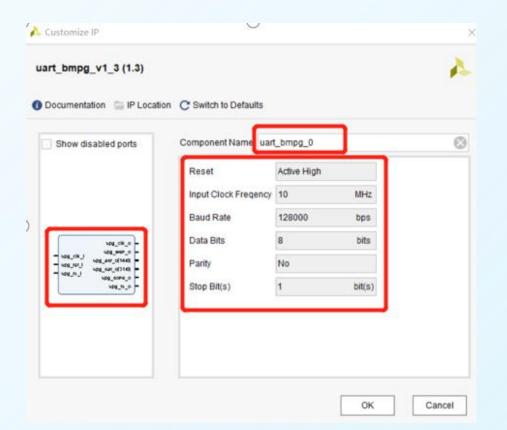




# Add an IP core Which Processes Uart Data continued

**Step2:** Add the IP core(uart\_bmpg\_0) from IP catalog into vivado project: in "IP catalog" pane, the IPcore(uart\_bmpg\_v1\_3) could be found in the "User Repository", click it to add it to your vivado project.

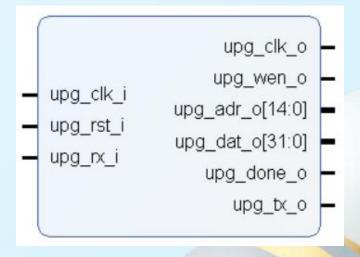




#### **NOTE:**

Don't change the settings of this IP core.

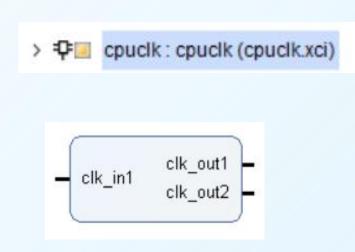
While using the IPcore, its name, features on uart communication and ports are important!

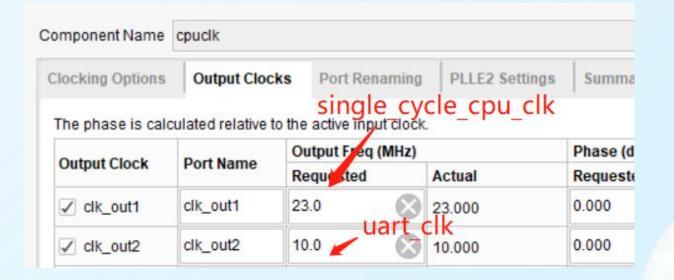




# Add a New Clock For The New IP core

- Reset the "cpuclk" IP core to make a new clock
  - Add a new clk\_out (clk\_out2) whose frequence is 10 Mhz for the IP core(uart\_bmpg\_0) which is used for Uart communication(on last page)







# Changes on CPU Top Module

```
module CPU_TOP(
    input fpga_rst, //Active High
    input fpga_clk,
    input[23:0] switch2N4,
    output[23:0] led2N4,

// UART Programmer Pinouts
// start Uart communicate at high level
    input start_pg, // Active High
    input rx, // receive data by UART
    output tx // send data by UART
);
```

```
out.txt

program_rom ifetch

UartAssist

uart_bmgp_0

dmemory32

CPU on FPGA
```

```
// For Minisys, the package_pin relationship in the constraints file set_property -dict {IOSTANDARD LVCMOS33 PACKAGE_PIN Y19} [get_ports rx] set_property -dict {IOSTANDARD LVCMOS33 PACKAGE_PIN V18} [get_ports tx] // For EGO1, the package_pin relationship in the constraints file set_property -dict {IOSTANDARD LVCMOS33 PACKAGE_PIN T4} [get_ports rx] set property -dict {IOSTANDARD LVCMOS33 PACKAGE_PIN N5} [get_ports tx]
```

Here the usage of "fpga\_rst" and "start\_pg" are only one type of implements, not the request.

The Y19(UART\_RX) and V18(UART\_TX) are the USB-UART pins of the FPGA chip(Artix7 fgg484) on Minisys Board.

The T4(UART\_RX) and N5(UART\_TX) are the USB-UART pins of the FPGA chip(Artix7 csg324) on EGO1 Board



### Changes on CPU Top Module continued

```
module CPU_TOP(
    input fpga_rst, //Active High
    input fpga_clk,
    input[23:0] switch2N4,
    output[23:0] led2N4,

// UART Programmer Pinouts
// start Uart communicate at high level
    input start_pg, // Active High
    input rx, // receive data by UART
    output tx // send data by UART
);
```

```
// UART Programmer Pinouts
wire upg_clk, upg_clk_o;
wire upg_wen_o; //Uart write out enable
wire upg_done_o; //Uart rx data have done

//data to which memory unit of program_rom/dmemory32
wire [14:0] upg_adr_o;

//data to program_rom or dmemory32
wire [31:0] upg_dat_o;
```

Q1. How many types of working mode on the CPU which support uart communication to download the program and the data? How to identify different types of working mode?

Q2. What's the relationship between the working mode and the "fpga\_rst" and "start\_pg"?

TIPS: Here the usage of "fpga\_rst" and "start\_pg" are only one type of implements, not the request.



# **Changes on Demeory32**

```
module dmemory32 (
              ram clk i,
                                 // from CPU top
    input
    input
              ram wen i,
                                 // from Controller
                                // from alu result of ALU
    input [13:0]
                   ram adr i,
                                // from read data 2 of Decoder
    input [31:0]
                  ram dat i,
                                 // the data read from data-ram
                   ram dat o,
    output [31:0]
    // UART Programmer Pinouts
                                // UPG reset (Active High)
    input
                   upg rst i,
                   upg_clk_i, // UPG ram_clk i (10MHz)
    input
                                 // UPG write enable
                   upg wen i,
    input
                   upg adr i,
                               // UPG write address
    input [13:0]
                   upg_dat_i, // UPG write data
    input [31:0]
                   upg done i // 1 if programming is finished
    input
```

Q. While "**kickOff**" is 1'b1, what's the working mode of the CPU? How about while "**kickOff**" 1'b0?

```
ram_clk_i
ram_wen_i
ram_adr_i[13:0]
ram_dat_i[31:0]
upg_rst_i
upg_clk_i
upg_wen_i
upg_adr_i[13:0]
upg_dat_i[31:0]
upg_done_i

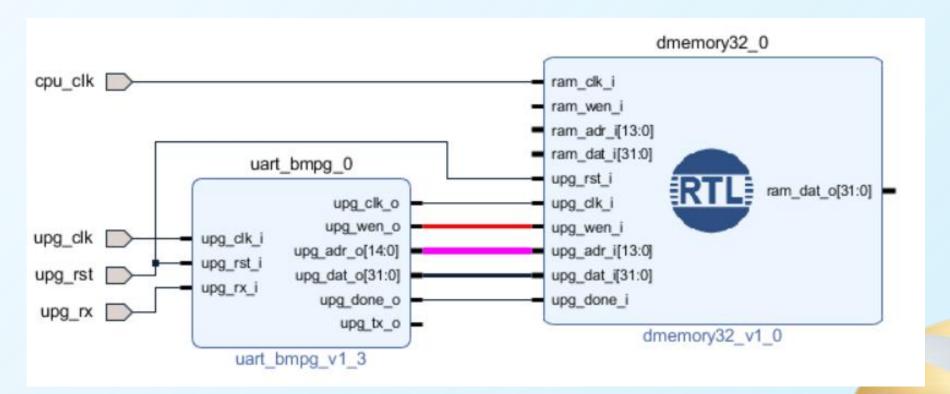
dmemory32_v1_0
```

```
wire ram clk = !ram clk i;
/* CPU work on normal mode when kickOff is 1.
CPU work on Uart communicate mode when kickOff is 0.*/
wire kickOff = upg_rst_i | (~upg_rst_i & upg_done_i);
ram ram (
     .clka (kickOff ?
                      ram clk
                                  : upg clk i),
     .wea (kickOff?
                       ram wen i : upg wen i),
     .addra (kickOff ?
                      ram adr i
                                 : upg adr i),
     .dina (kickOff ?
                       ram dat i : upg dat i),
     .douta (ram dat o)
```



### Changes on Demeory32 continued

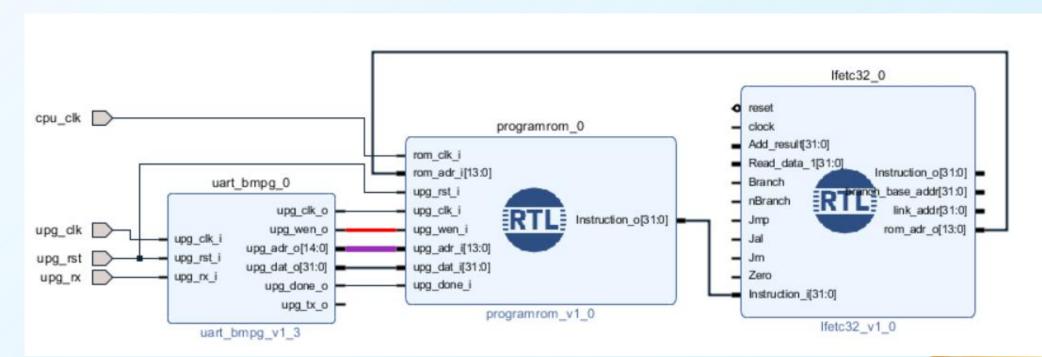
- upg\_wen\_i ( uart write enable on Dmemory32) :
  - determined by: upg\_wen\_o(from uart\_bmpg\_0) & upg\_adr\_o[14] (from uart\_bmpg\_0)
- upg\_adr\_i[13:0] (uart write address on Dmemory32):
  - connect with: upg\_adr\_o[13:0] (from uart\_bmpg\_0)





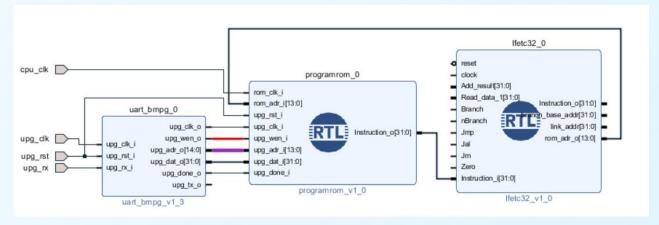
### **Changes on IFetch**

- Separae IP core( "programrom\_0" ) which stores the Instruction from IFeth(optional)
- Change prgrom from ROM to RAM(writabel while work on Uart communication mode, read only while work on normal mode)
- upg\_wen\_i ( uart write enable on "programrom" ), determined by: upg\_wen\_o(from uart\_bmpg\_0) & (!upg\_adr\_o[14]) (from uart\_bmpg\_0)
- > upg\_adr\_i[13:0] (uart write address on "programrom"), connect with upg\_adr\_o[13:0] (from uart\_bmpg\_0)





### Changes on IFetch continued



```
module programrom (
    // Program ROM Pinouts
                  rom clk i,
                             // ROM clock
    input
    input[13:0] rom adr i,
                            // From IFetch
           [31:0] Instruction o, // To IFetch
    output
    // UART Programmer Pinouts
                  upg rst i,
                              // UPG reset (Active High)
    input
                 upg_clk_i, // UPG clock (10MHz)
    input
                 upg_wen i, // UPG write enable
    input
                 upg_adr_i, // UPG write address
    input[13:0]
                  upg_dat_i, // UPG write data
    input[31:0]
                  upg done i
                              // 1 if program finished
    input
```

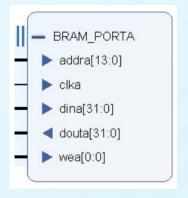


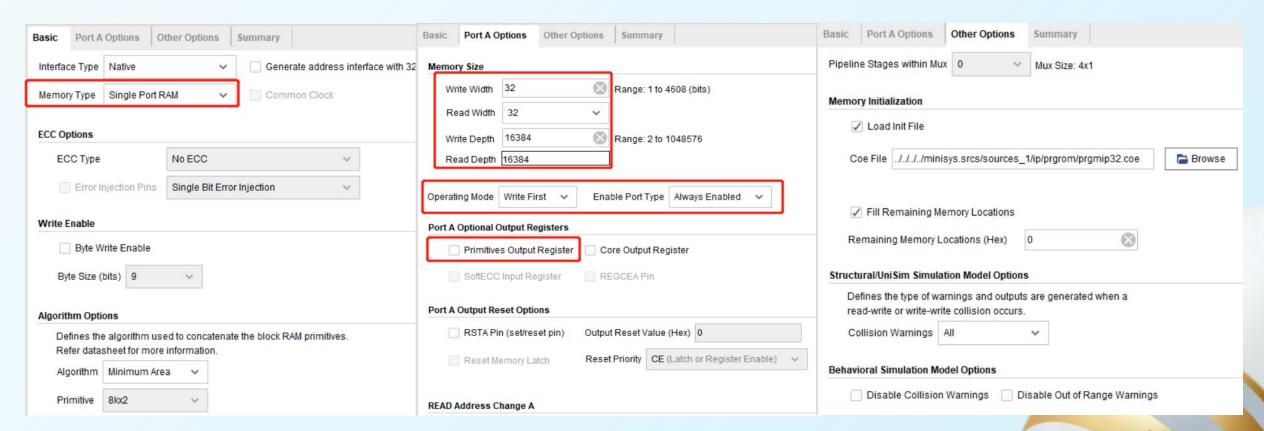
### Changes on IFetch continued

#### Make a **new** programrom(which is a **RAM** memory):

TIPS about the "programrom":

- > While on CPU communication mode, "programrom" is writable.
- While on CPU normal mode, "programrom" is readOnly.

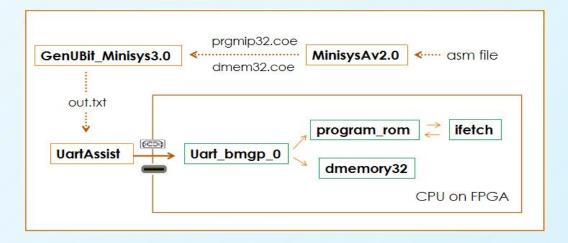


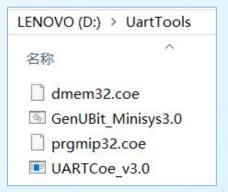




### **Tools(1): Generate the Data For Uart Port**

- Step1: Using "MinisysAv2.0" to assemble the asm file and generate the coe files(prgmip32.coe and dmem32.coe)
- Step2: Using "GenUBit\_Minisys3.0" to merge the coe files(prgmip32.coe and dmem32.coe) into one file "out.txt"
  - Tips on Step2:
  - put "prgmips32.coe" and "dmem32.coe" into the same directory with "UARTCoe\_v3.0" and "GenUBit\_Minisys3.0", or you will need to make some modification on GenUBit Minisys3.0





d:\UartTools>GenUBit\_Minisys3.0.bat 2 files are read successfully Hexadecimal file(s) detected. Done.

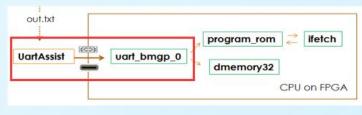
"GenUBit\_Minisys3.0", "UartAssist" and "UARTCoe\_v3.0" could be found in the "uart\_tools.rar" of "labs/lab13\_uart" on BlackBoard site

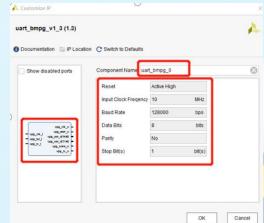


# Tools(2): Using "UartAssist"

- > Step 1: Connect the Computer which runs "UartAssist" with Minisysboard on which your designed CPU has already been programed on its FPGA chip.
- > Step 2: Double click on "UartAssist" to open it
- > Step 3: Set the items in "串口设置" as the settings of screen snap on the right hand, then click on "打开"
  - ➤ **NOTICE:** "串口号" could be an **serial port** other than "COM4", which **is up to your Computer**. The port which you choose here and then click on "打开" hasn't report error is the right port.







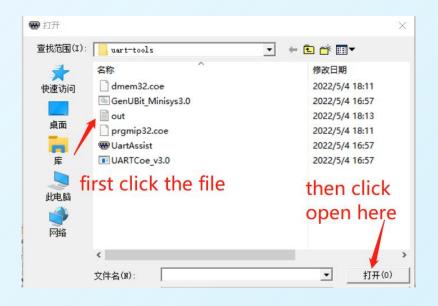


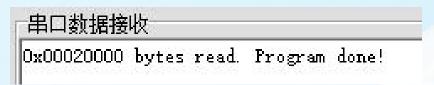
# Tips: Using "UartAssist" continued

- Step 4: Make sure the CPU on FPGA works on uart communication mode.
- > Step 5: Set the items in "发送区设置" as the screen snap on the right hand. Click on "启用文件数据源", find the data file which is to be transformed by uart port to FPGA chip.
- Step 6: Wait until a notice info "Program done!" has appeared in the "串口数据接收" window as the screen snap on the right hand

(The "Program done!" here means the Uart communication done).









### **Practice**

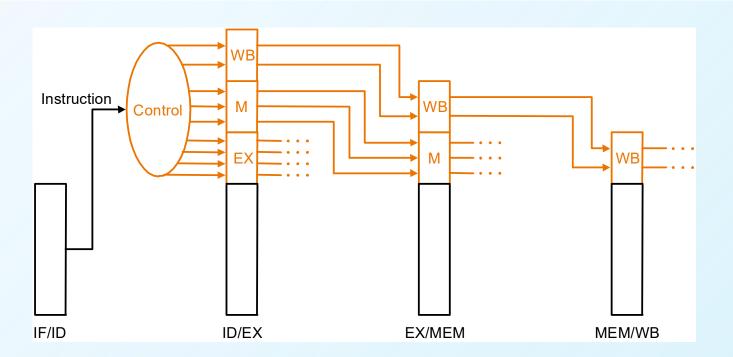
- > 1. Modify the Data-memory module, do the unit test on the updated Data-memory.
- > 2. Modify the IFetch, do the unit test on the updated IFetch.
- > 3. Update the CPU with uart communication implemented.
- ➤ 4. Do the test: Programe FPGA chip only once, make the CPU run the different programs by using uart communication to update the instructions and data of new program.

It is strongly recommended to conduct unit test first, and then conduct integration test after passing the unit test.



# Pipeline related Tips: register storage

- The process of instruction in CPU are divded to 5 stage: IF(update PC,instruction fetch), ID(instruction decoder: generate controll signals and prepare the data), EX(instruction execution), MEM(access Data memroy), WB(write back).
- Pipeline level registers are request to store and forward the data, address, and control information related to instructions to the module in the next processing stage.



#### NOTE:

The data, address, and control information of the same instruction should be synchronously transmitted to the target module.

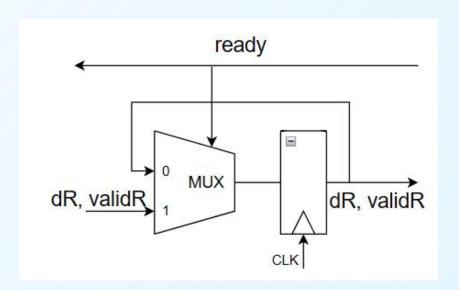
Tip: Here "IF/ID" is the pipeline level registers, which is between the stage IF and ID,Other situations are similar.



# Pipeline related Tips: register read/write

valid/ ready: It is a commonly used communication mode, where "valid" indicates that the data tobe sent is ok, "ready" indicates that the receiver is ready, and only when both signals are enable indicates that the transmitted data signal can be effectively received.

The **valid/ ready** could be used to implement the **stall** to handle the harzad in pipeline.



```
assign upstream_ready = downstream_ready;

wire stall = ~upstream_ready;

data_nxt = rst?0: ( stall? dowstream_data : upstream_data);

valid_nxt = rst?0: ( stall? dowstream_valid : upstream_valid);

DFF #(bits) dataR(clk, data_nxt, dowstream_data);

DFF #(1) validR(clk, valid_nxt, dowstream_valid);

//DFF here is a parameterized D-flipflop, its1st port is clock, 2nd port is the D data, the 3rd port is the output Q. The bitwidth of
```

D and Q is specified by the parameter.

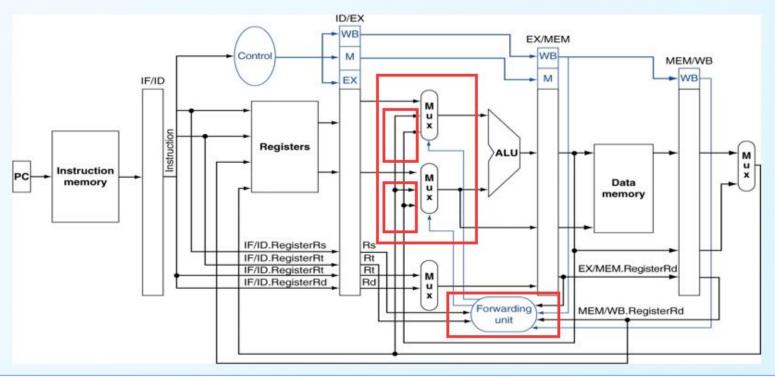
NOTES: The implementation is for reference only, not request.



### Pipeline related Tips: forward

Use result when it is computed, do not wait for it to be stored in a register, requires extra connections in the

datapath.



#### Control signals for forwarding

EX hazard

if (EX/MEM.RegWrite && (EX/MEM.RegRd!=0) &&(EX/MEM.RegRd==ID/EX.RegRs)) ForwardA=10 // the ALU operand is forwarded from the prior ALU reslut.

#### MEM hazard

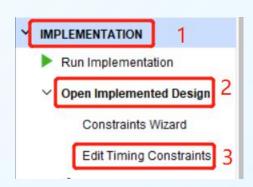
if (MEM/WB.RegWrite && (MEM/WB.RegRd!=0) && (MEM/WB.RegRd==ID/EX.RegRs)) ForwardA=01 // the ALU operand is forward from Data Memory or an earlier ALU result.



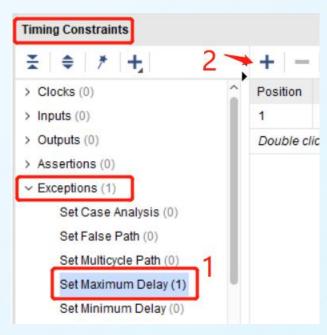
# Pipeline Tips: determine the Clock cycle(1)

1) Adding Simple Timing Constraints 2) Running Implementation based on the Timing Constraints 3) Checking if there is an exception caused by the Timing Constraints, No exception means the maximum delay for the circuit is ok, otherwise a longer delay is needed.

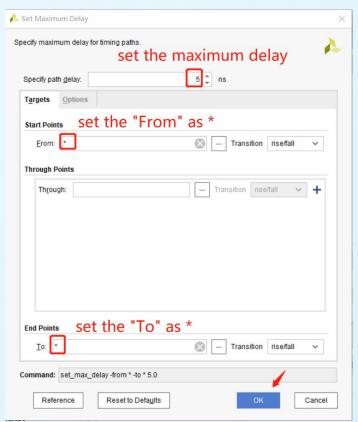
Adding Simple Timing Constraints could be done by add a description "set\_max\_delay -from \* -to \* 5.000" to the constraint file(here 8 is the maximum delay) or by GUI (follow the following steps):



Step1: Click the "**Edit Timing Constraints**" of "Open Implementated Design" to open "Timing Constraints" pane.



Step2: Click "**Set Maximum Delay**" of "Exceptions" in "Timing Constraints".



Step3: Set the "path delay", "From", "To" in "Set Maximum Dalay" pane

Step4:Save the constraint settings.

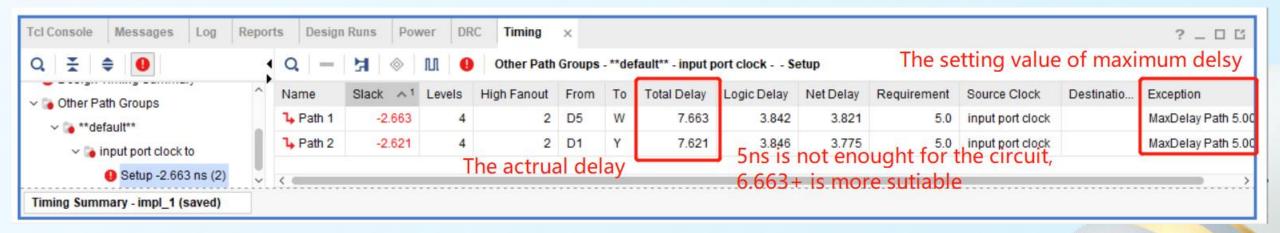


# Pipeline Tips: determine the Clock cycle(2)

Run the implementation again to check if the maximum delay set in the constraint is suitable for the circuit.

A new pane named "Timing" would be invoke at the bottom of the vivado after the implementation finished.

If the maximum delay set in the constraint file is not suitable, an implementation error would be invoked to tell more information.





# Pipeline Tips: determine the Clock cycle(3)

Here the maximum delay is adjust to 8ns (edit the discription in the constraints file directly.

Re running implementation to check the delay setting.

After implementation, it is found there is NO implementation error about Timing, 8ns is suitable for the circuit.



