

## How to Prepare and Run NPU System Demo (CPU + FPGA)

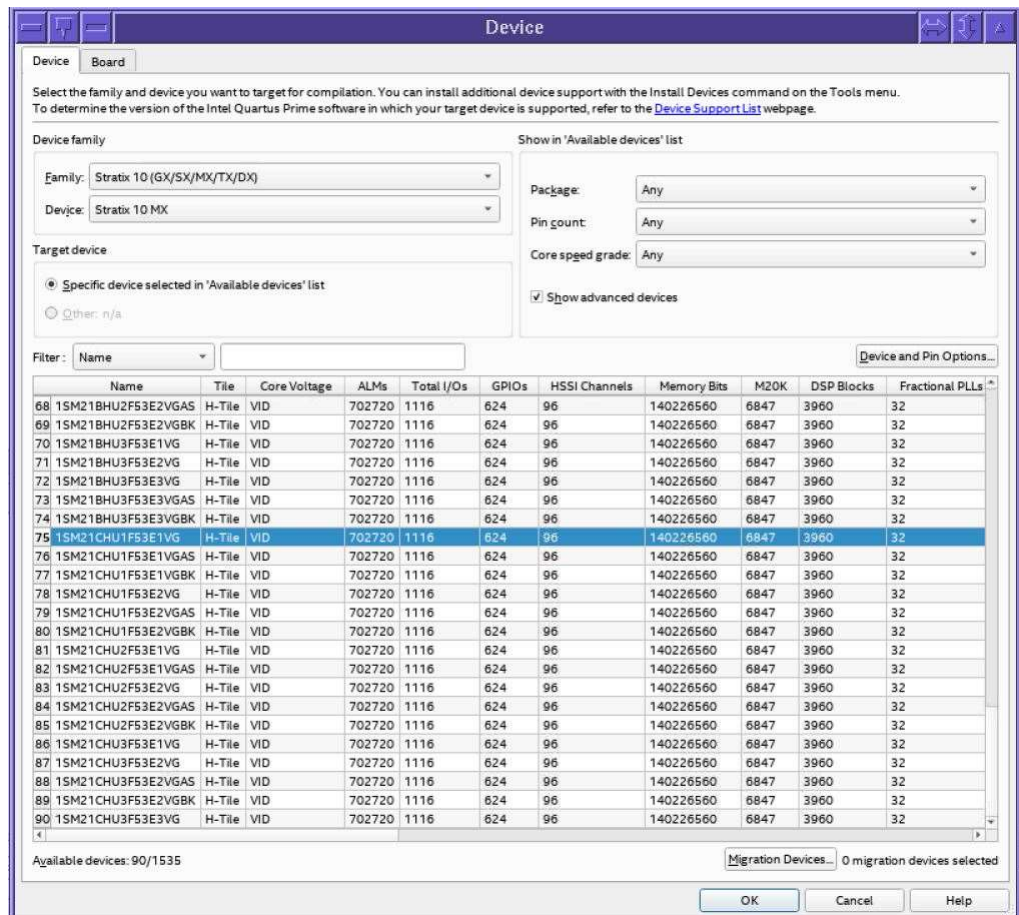
### 1. Start in the directory where the zip file has been unpacked

- 1) We call this directory the <project home> directory

### 2. PCIe EP example setup & Integration with NPU

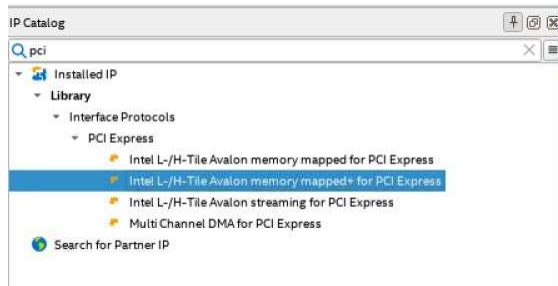
#### 2.1 PCIe EP example project setup

- 1) Go to <project home> directory
- 2) Create one new project with new project wizard & choose empty project
- 3) Choose S10 MX device with below item at first

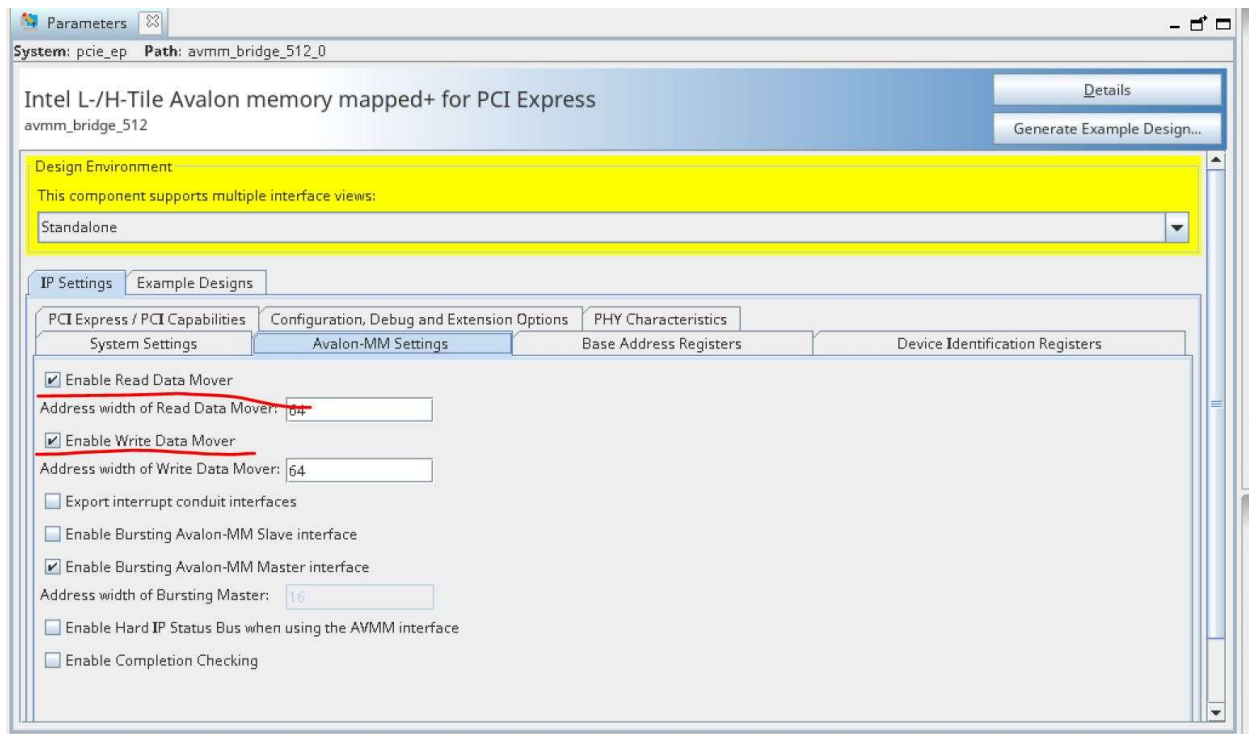


- 4) Directly click "finish"

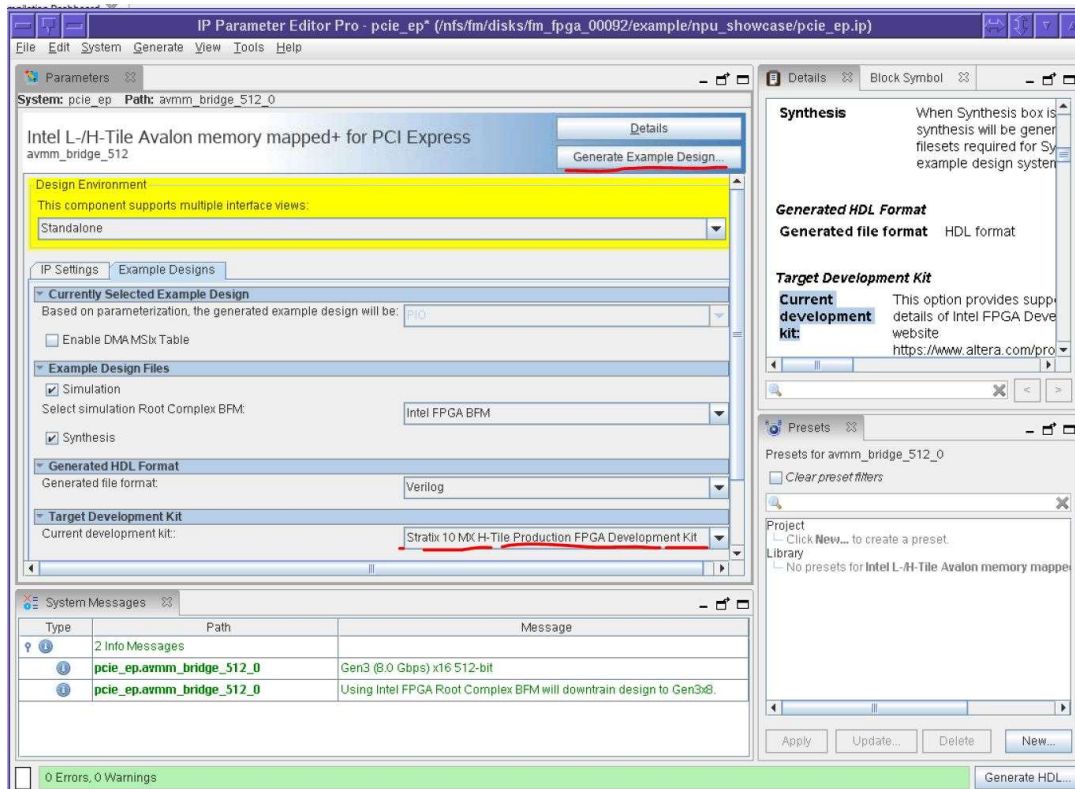
- 5) Search "pci" in IP Catalog & choose that map + as below, then click "add.." button.



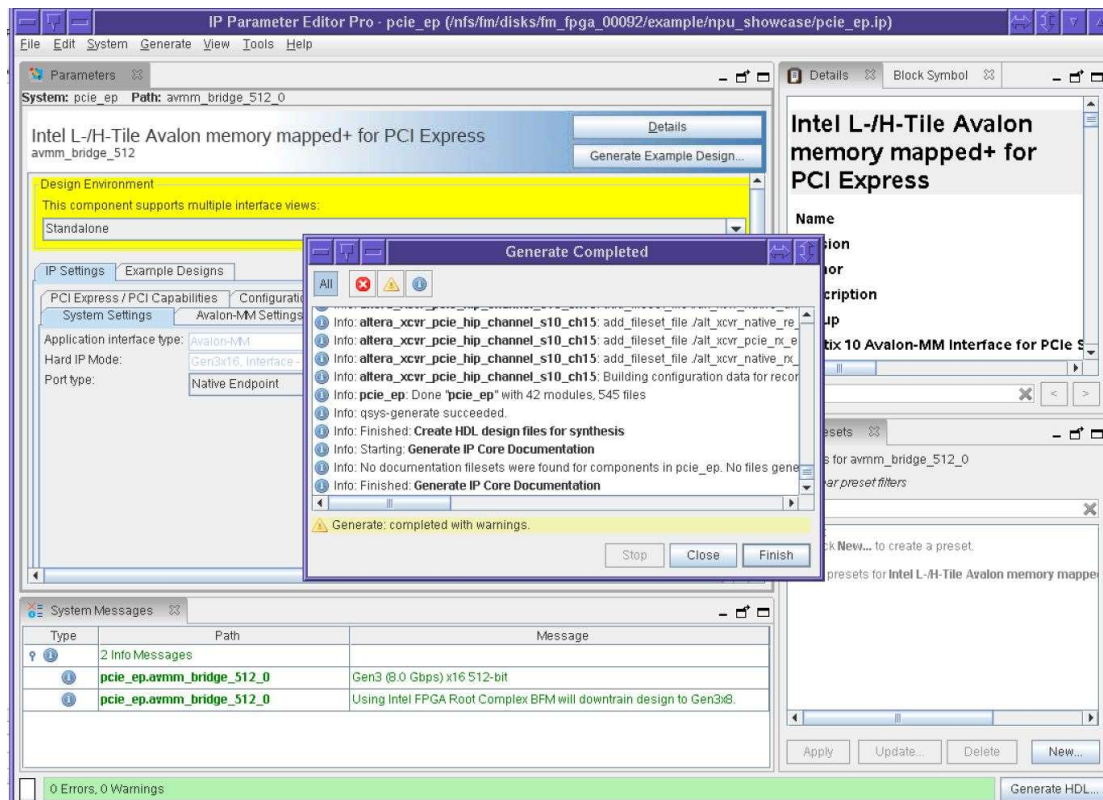
- 6) Create new PCIe endpoint IP
- 7) In PCIe EP config wizard, in “Avalon-MM settings” tab, make sure selection for “Enable Read Data Mover” & “Enable Write Data Mover”. (default selected)



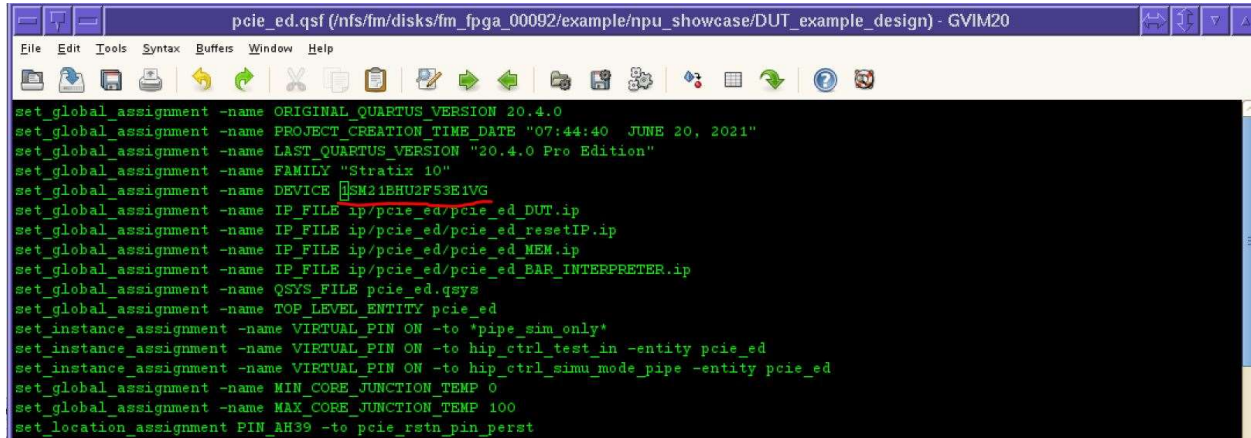
- 8) In “Example Design” tab, choose S10 MX development kit, then click “Generate Example Design” button.



9) Save & generate.

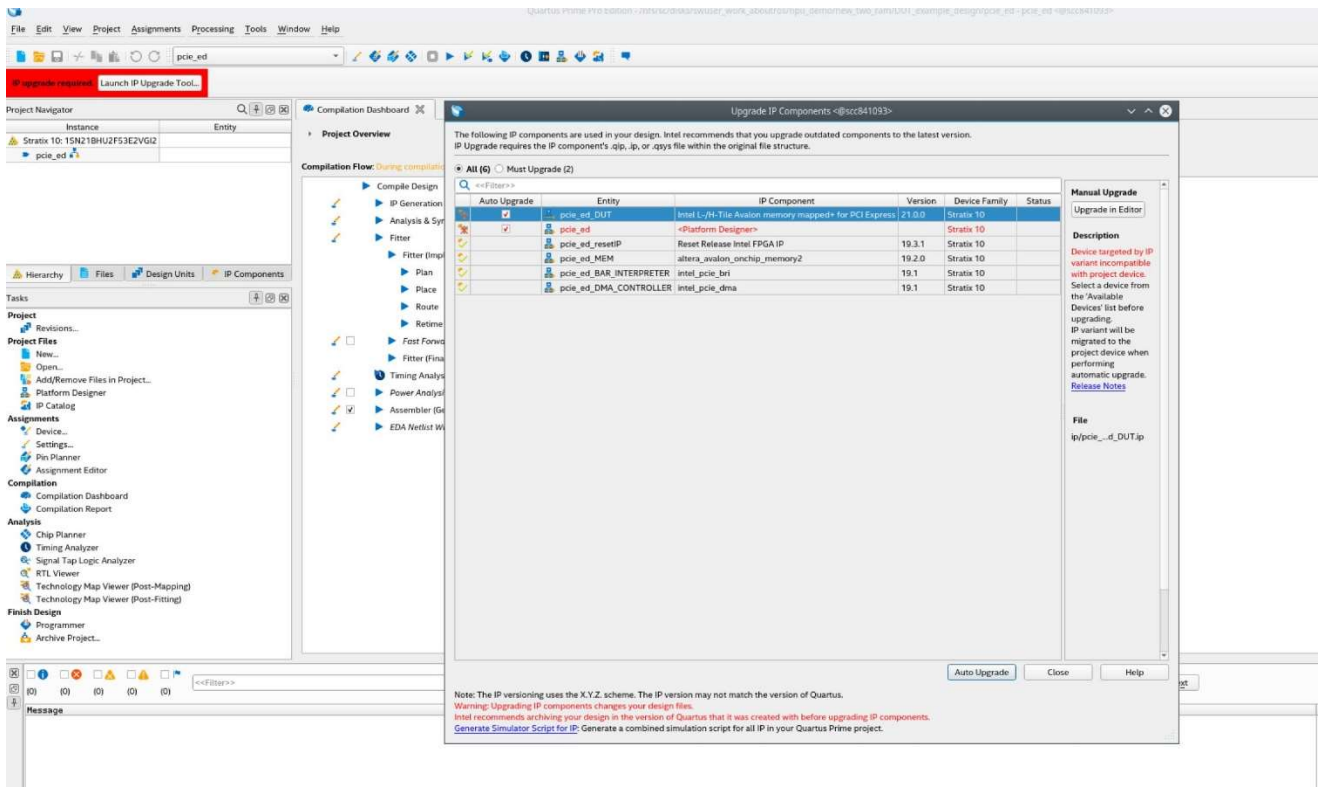


- 10) Since Stratix 10 NX can't be selected from Quartus GUI, you have to manually edit \*.qsf under your new PCIe example generate folder to update device name to "1SN21BHU2F53E2VGI2"



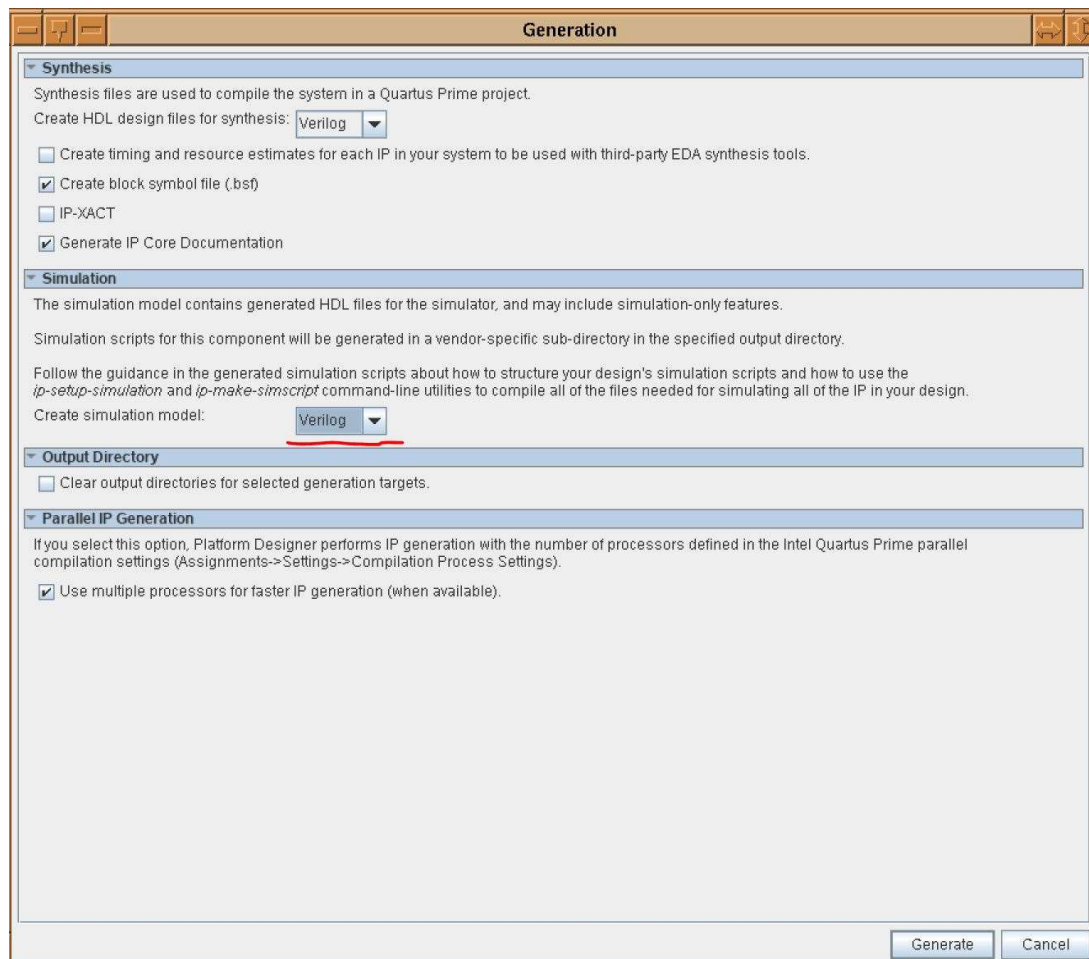
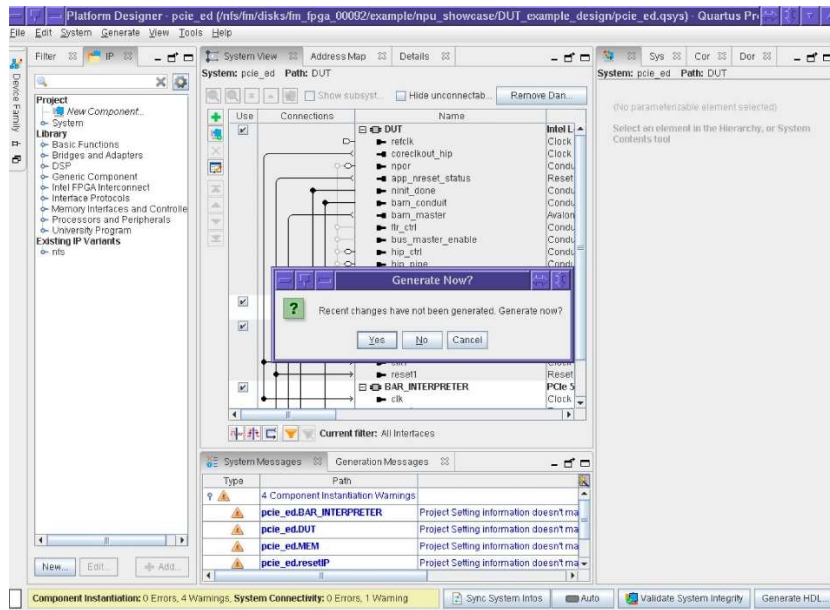
```
set_global_assignment -name ORIGINAL_QUARTUS_VERSION 20.4.0
set_global_assignment -name PROJECT_CREATION_TIME_DATE "07:44:40 JUNE 20, 2021"
set_global_assignment -name LAST_QUARTUS_VERSION "20.4.0 Pro Edition"
set_global_assignment -name FAMILY "Stratix 10"
set_global_assignment -name DEVICE 1SN21BHU2F53E1VG
set_global_assignment -name IP_FILE ip/pcie_ed/pcie_ed_DUT.ip
set_global_assignment -name IP_FILE ip/pcie_ed/pcie_ed_resetIP.ip
set_global_assignment -name IP_FILE ip/pcie_ed/pcie_ed_MEM.ip
set_global_assignment -name IP_FILE ip/pcie_ed/pcie_ed_BAR_INTERPRETER.ip
set_global_assignment -name QSYS_FILE pcie_ed.qsys
set_global_assignment -name TOP_LEVEL_ENTITY pcie_ed
set_instance_assignment -name VIRTUAL_PIN ON -to *pipe_sim_only*
set_instance_assignment -name VIRTUAL_PIN ON -to hip_ctrl_test_in -entity pcie_ed
set_instance_assignment -name VIRTUAL_PIN ON -to hip_ctrl_simu_mode_pipe -entity pcie_ed
set_global_assignment -name MIN_CORE_JUNCTION_TEMP 0
set_global_assignment -name MAX_CORE_JUNCTION_TEMP 100
set_location_assignment PIN_AH39 -to pcie_rstn_pin_perst
```

- 11) Re-load project (\*.qpf) under your new PCIe example generate folder, then you will see below notice with RED (you have to update IP for you manually change device name)

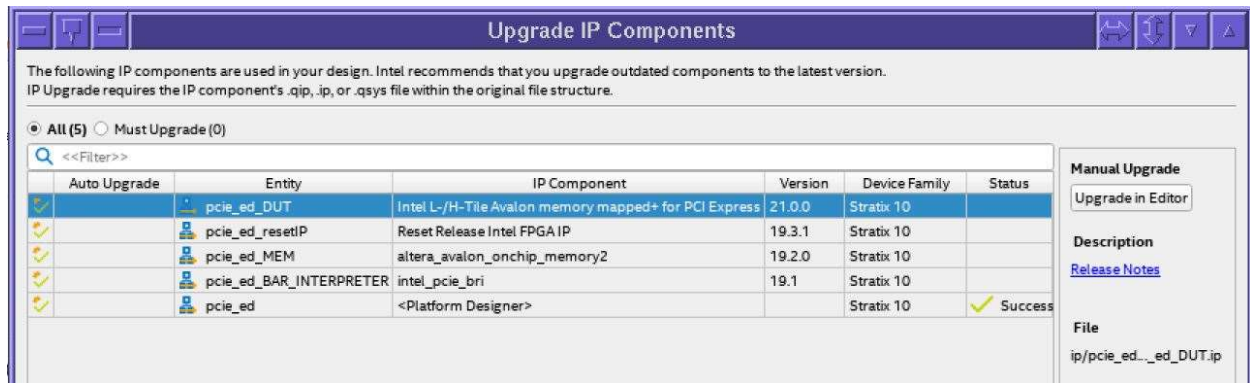




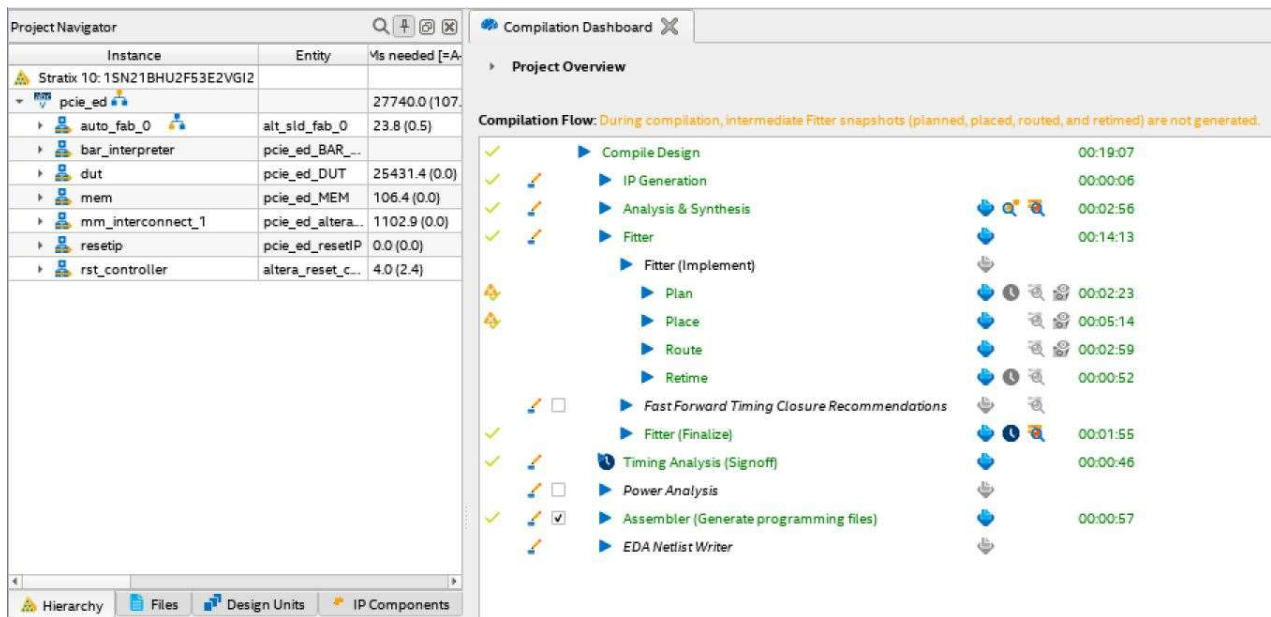
- 12) Click “Update in Editor” button, load platform designer, save → exit → generate.  
(Note, please remember to choose Verilog for simulation)



13) Update IP in red one by one to finally all IP updated & without highlighted with red.



14) Click “run” button to generate one PCIe EP image at first before integrating with NPU IP.



## 2.2 How to get current dual-RAM FPGA module (1Mbyte DMA address)

### 2.2.1 Extend DMA address (connected to DUT) from 32KB to 1MB

- Open your project qsys file, double click that “MEM” IP, in right side property window, you need to change “Total memory size” from 32768(32KB) to 1048576(1MB)

The image shows the Qsys System View and Parameters window for the **pcie\_ed MEM** component. The System View on the left displays a complex interconnection of components including **BAR\_INTERPRETER**, **DMA\_CONTROLLER**, and **MEM**. The Parameters window on the right shows the configuration for the **On-Chip Memory (RAM or ROM) Intel FPGA IP**.

**Parameters:**

- HDL entity name:** pcie\_ed\_MEM
- IP file:** ip/pcie\_ed/pcie\_ed\_MEM.ip
- Memory type:** RAM (Writable)
- Dual-port access:** ☒ (Single clock operation: ☐)
- Read During Write Mode:** DONT\_CARE
- Block type:** AUTO
- Size:**
  - Slave s1 Data width: 512
  - Total memory size: 1048576
- Read latency:**
  - Slave s1 Latency: 2
  - Slave s2 Latency: 1
- ROM/RAM Memory Protection:** Reset Request: Enabled
- ECC Parameter:** Extend the data width to support ECC bits: Disabled

**System Messages:**

Type	Path	Message
4 Component Instantiation Warnings		
Warning	pcie_ed.BAR_INTERPRETER	Project Setting information doesn't match IP Variant file. Double-click to open System Info tab.
Warning	pcie_ed.DMA_CONTROLLER	Project Setting information doesn't match IP Variant file. Double-click to open System Info tab.
Warning	pcie_ed.MEM	Project Setting information doesn't match IP Variant file. Double-click to open System Info tab.
Warning	pcie_ed.resetIP	Project Setting information doesn't match IP Variant file. Double-click to open System Info tab.
1 System Connectivity Warning		
Warning	pcie_ed.DUT	DUT.fir_ctrl must be exported, or connected to a matching conduit as it has unconnected inputs.

Activity: 7 Errors, 1 Warning

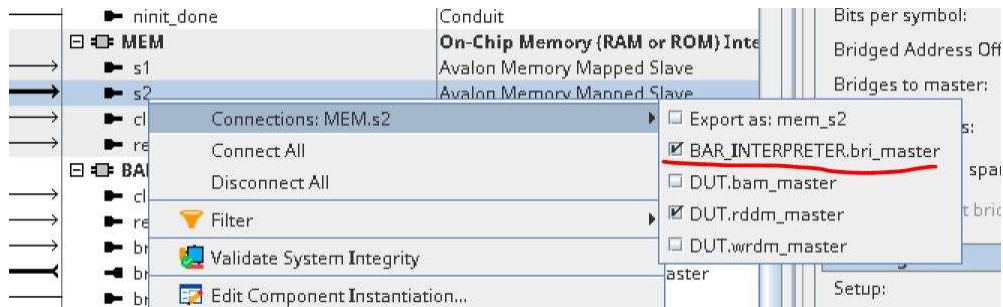
- b) In Qsys system view, on “mem.s1” & “mem.s2” interfaces, **dis-connect** their connection with **BAR\_INTERPRETER.bri\_master** (this is for we don’t need access DMA RAM content through PIO mode, we use small size DMA read to replace PIO polling, & big buffer size easy to get conflict on **BAR\_INTERPRETER** bus address).

The image shows the Qsys System View with a context menu open for the **Connections: MEM.s1** connection. The menu options are:

- Connect All
- Disconnect All
- Filter
- Validate System Integrity
- Edit Component Instantiation...
- Add...

The **Parameters** window on the right shows the configuration for the **avalon\_slave** component:

- Parameters:**
  - Address group:
  - Address units:
  - Associated clock:
  - Associated reset:
  - Bits per symbol:
  - Bridged Address Offset:
- Timing:**
  - Slave
  - Master
  - Setup:



- c) In “Address Map” window of Qsys, you need to unblock MEM.s1 & MEM.s2 memory map setting at first, then change base address from 0x1\_0000 to 0x10\_0000 to resolve address map conflict.

Slave	BAR_INTERPRETER.bri_master	DUT.wrdm_master	DUT.rddm_master	
BAR_INTERPRETER.b...				0x000
DMA_CONTROLLER,...	0x0000_0000 - 0x0000_0fff		0x0000_0000_0000_0000 -	
MEM.s1		0x0000_0000_0001_0000 -		
MEM.s2			0x0000_0000_0001_0000 -	

Slave	BAR_INTERPRETER.bri_master	DUT.wrdm_master	DUT.rddm_master	
BAR_INTERPRETER.b...				0x000
DMA_CONTROLLER,...	0x0000_0000 - 0x0000_0fff		0x0000_0000_0000_0000 -	
MEM.s1		0x0000_0000_0010_0000 -		
MEM.s2			0x0000_0000_0010_0000 -	

## 2.3 Apply NPU patch to the Quartus project

- Go to patch directory  
cd <project home>/patch
- Apply patch  
./setup.sh

## 3. Hardware platform setup & Run NPU demo

### 3.1 host PC selection

- X86 CPU & can support at least AVX256, better for AVX512 (can use lscpu in Linux to check CPU flags)
- Support Gen3 x16 PCIe slot

### 3.2 host OS selection

Suggest Ubuntu (> 18.0), or PCIe kernel driver maybe need minor change for kernel mode function support.

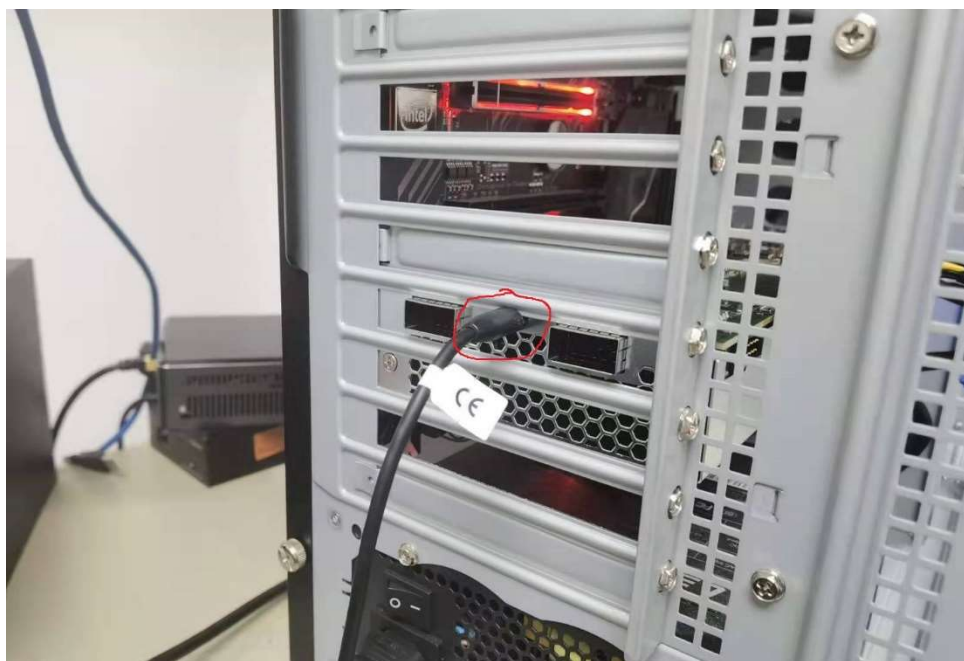
### 3.3 hardware setup

- Connect S10 NX card to one Gen3 x16 PCIe slot on host PC.
- Need to connect ATX power cable to power slot on S10 NX card.





- Connect USB cable between S10 NX card & your control PC (run Quartus programmer & SigTap tool)

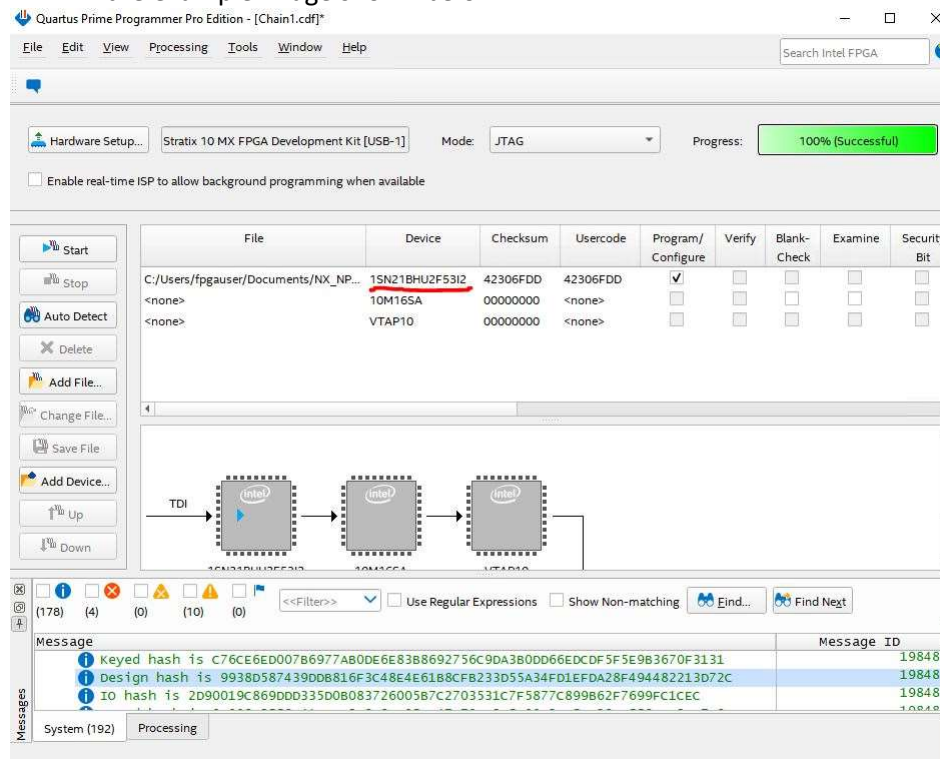


### 3.4 Control PC tool

please install Quartus program tool (not need full Quartus install, but please use 20.4 version)

### 3.5 How to run NPU on FPGA board

- Download pcie\_ed.sof generated in <project directory>/DUT\_example\_design/ to S10 NX (make sure image is using this device name)
  - Click "Auto Detect"
  - Click "Add file" to add the pcie\_ed.sof bitstream
  - Remove the redundant entry, Use "Up" button to make the sequence matching the example image shown below.



- You need to reboot host PC (Note: it is "reboot", can't be "power down")
- Copy <project home>/DUT\_example\_design/software folder to host PC
- Go to "kernel/linux". For the first time, compile the source files by typing "make".
- Load the driver by typing "sudo ./load"
- Type "lspci -d 1172: -vvv" to make sure you can find that PCIe device, BAR2 enabled & driver load

```
intel@flexpsthj-19:~/test_pcie/software/user/dma_test$ lspci -d 1172: -vvv
17:00.0 Unassigned class [ff00]: Altera Corporation Device 0000 (rev 01)
Control: I/O- Mem+ BusMaster+ SpecCycle- MemWINV- VGASnoop- ParErr- Stepping- SERR+ FastB2B- DisINTx-
Status: Cap+ 66MHz- UDF- FastB2B- ParErr- DEVSEL=fast >TAbort- <TAbort- <MAbort- >SERR- <PERR- INTx-
Latency: 0, Cache Line Size: 32 bytes
Interrupt: pin A routed to IRQ 53
NUMA node: 0
Region 0: Memory at b5e10000 (64-bit, prefetchable) [size=64K]
Region 2: Memory at b5e00000 (64-bit, prefetchable) [size=64K]
Capabilities: <access denied>
Kernel driver in use: intel_fpga_pcie_drv
Kernel modules: altera_cvp
```

- g) Go to "user/npu\_test". For the first time, compile the source files by typing "make".
- h) Copy benchmark data. E.g., to run mlp\_batch4104, type "cp ./mlp\_batch4104/\* ./"
- i) Type "./real\_npu\_test" to run the test
- j) Type 0 for automatic PCIe device find

```

*****
Intel FPGA PCIe Link Test
Version 2.0
0: Automatically select a device
1: Manually select a device
*****
> |
intel@flexstbj-i9:~/rui/software/user/npu_test$ ./real_npu_test

*****
Intel FPGA PCIe Link Test
Version 2.0
0: Automatically select a device
1: Manually select a device
*****
> 0
Opened a handle to BAR 0x2 of a device with BDF 0x6500
Allocate Kernel memory succesully!
Filling 280 MRFs with 40 elements x 88 words
MRF buffer size = 1835008 Bytes
Finished parsing MRF file!
Filling instruction memory with 48 elements x 34 words
Instruction buffer size = 262144 Bytes
Finished parsing Instructions file!
Filling input memory with 40 elements x 114912 words
Input buffer size = 7602176 Bytes
Finished parsing Inputs file!
Finished parsing Golden Outputs file!
Finished sending MRFs
Finished sending instructions
Latency is 3.563000 ms
Finished sending all inputs and receiving all results!
TEST PASSED!

```

### 3.6 [Optional] Sigtap debugging on FPGA

- new one \*.stp file with debug signals you want, & include it into FPGA image build
- Copy that \*.stp to control PC
- Use SigTap tool to open this \*.stp
- Select correct hardware & device in tool
- Set trigger condition in that "setup" tab.

Quartus Prime Signal Tap Logic Analyzer Pro Edition - [stp1.stp]

File Edit View Processing Tools Help

Search Intel FPGA

Instance Manager: Ready to acquire

Instance Status Enabled

auto\_sign... Not running ✓

JTAG Chain Configuration: JTAG ready

Hardware: Intel Stratix 10 MX FPGA Dev Setup...

Device: @1: 15N21BHJ[1|2|2|251| Scan Chain

Bridge index: None detected

SOF Manager:

Signal Configuration:

Clock: refclk\_clk

Data:

Schema length: 612

trigger: 2021/06/21 10:25:34 #0 Lock mode: Allow all changes

Type	Alias	Name	Data Enable	Trigger Enable	Trigger Conditions
Basic AND			1099	1099	1
pcie_ep_npu_shell_0	avs_s0_chipselect		✓	✓	0
pcie_ep_npu_shell_0	avs_s0_write		✓	✓	100h
...	...	...	...	...	...

Hierarchy Display: auto\_signaltap\_0

- Then kick off “capture” button, it will stop & show waveform when trigger condition hit.

Quartus Prime Signal Tap Logic Analyzer Pro Edition - [stp1.stp]

File Edit View Processing Tools Help

Search Intel FPGA

Instance Manager: Ready to acquire

Instance Status Enabled

auto\_sign... Not running ✓

JTAG Chain Configuration: JTAG ready

Hardware: Intel Stratix 10 MX FPGA Dev Setup...

Device: @1: 15N21BHJ[1|2|2|251| Scan Chain

Bridge index: None detected

SOF Manager:

log: Trig @ 2021/06/21 10:25:38 (0:0:4.1 elapsed)

click to insert time bar

Type	Alias	Name	Value
pcie_ep_npu_shell_0	avs_s0_write		100h
...	...	...	...

Hierarchy Display: auto\_signaltap\_0

pcie\_ep\_np...