Choose the Right Hardware

Proposal Template

Scenario 1: Manufacturing

Client Requirements and Potential Hardware Solution

Look through the scenario and find any relevant client requirements. Then, suggest a potential hardware type and explain how this hardware would satisfy each of the requirements.

Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
FPGA

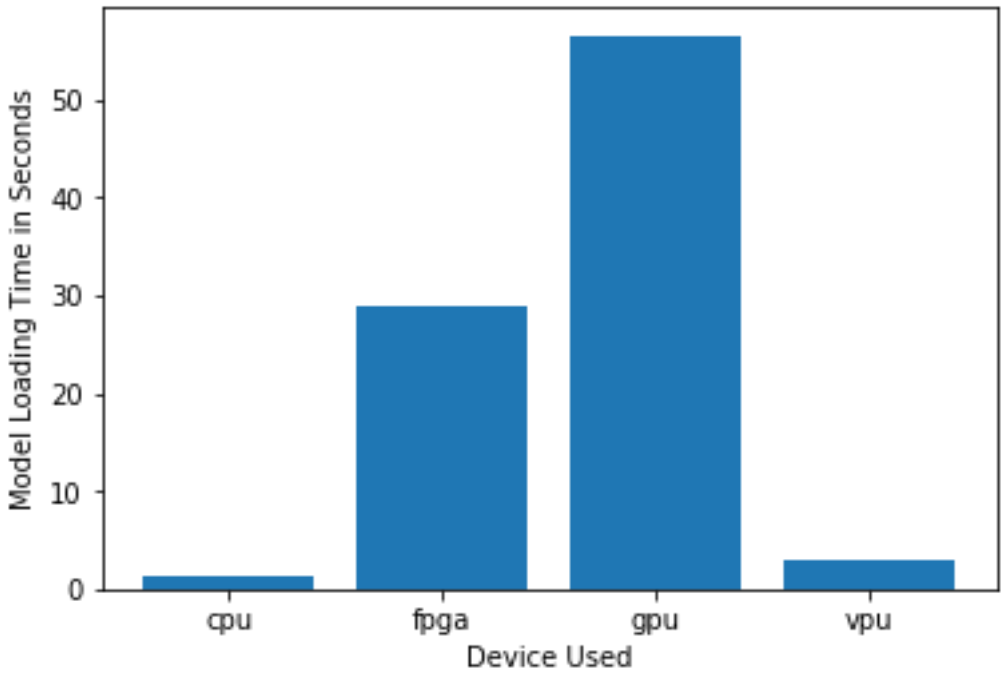
Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
The client requires good quality system.	*According to documentation, the quality of FPGA is better than other devices.*
Lifetime of system should be 5-10 years	*Intel guarantees that the FPGA will be available for next 10 years.*
System should be flexible for future reprogramming and optimization.	*We can reprogram FPGA.*

Queue Monitoring Requirements

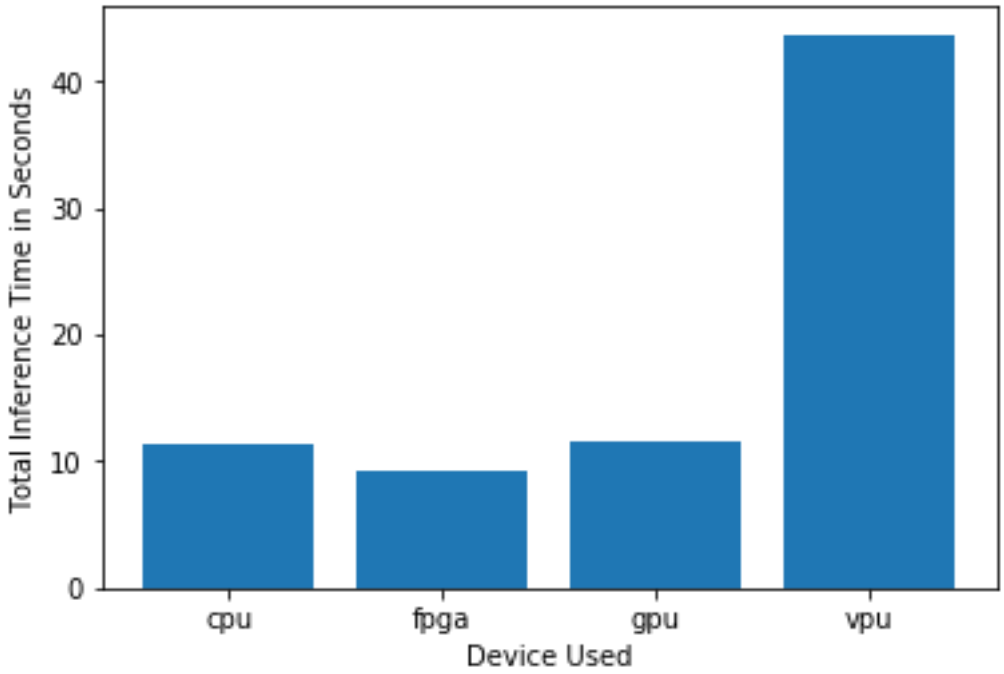
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

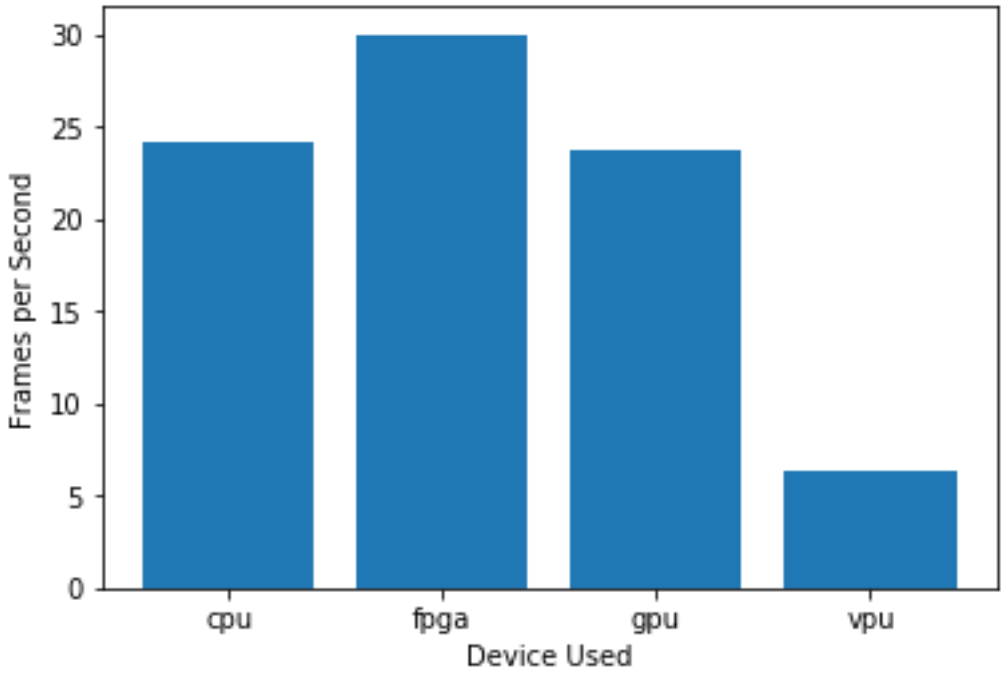
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

Now synthesize your points from above and provide a brief write-up describing why the chosen hardware is the best choice for this scenario. Be sure to discuss the client's requirements, the test results, and how these relate to one another (e.g., perhaps one of the devices performed better than the rest, but does not meet one of the client's requirements).

Write-up: Final Hardware Recommendation
• *The client is looking for quality system, also wants to invest for next 5-10 years. As we know FPGAs have long lifetime. IIT's group have guaranteed the availability of FPGA for next 10 years.* • *The client wants the system to be reprogrammed and optimize in future. Field-programmable FPGA's are best suitable for this requirement as they can be reprogrammed for evolving and custom networks.* • *Client's video streaming requirements are that the system should be able to run inference very quickly. As shown in result, FPGAs inference is faster than other devices which is according to client's requirements.* • *CPU and VPU took 0-5 seconds to load model according to test results but these hardware doesn't meets client requirements. GPU took 60 seconds which is too much. While FPGAs took 30 seconds to load model and meet all other requirements.*

Scenario 2: Retail

Client Requirements and Potential Hardware Solution

Look through the scenario and find any relevant client requirements. Then, suggest a potential hardware type and explain how this hardware would satisfy each of the requirements.

Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
IGPU

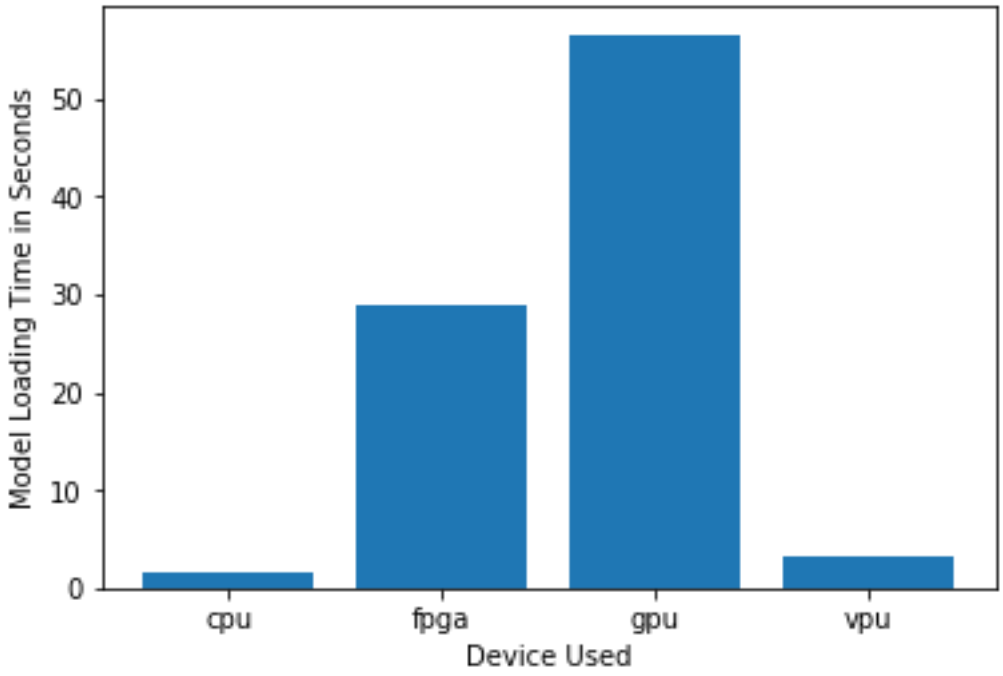
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
Client is not interested in investing additional amount for hardware.	*No other amount is required for graphical processing if we use IGPU, as it comes with integrated Graphical processor with CPU.*
Power consumption means a lot for client, client wants to reduce as much as he can for electricity bill.	*IGPU has capability to minimize the resources to be use, while if we use CPU only it will require more power for high performance.*

Queue Monitoring Requirements

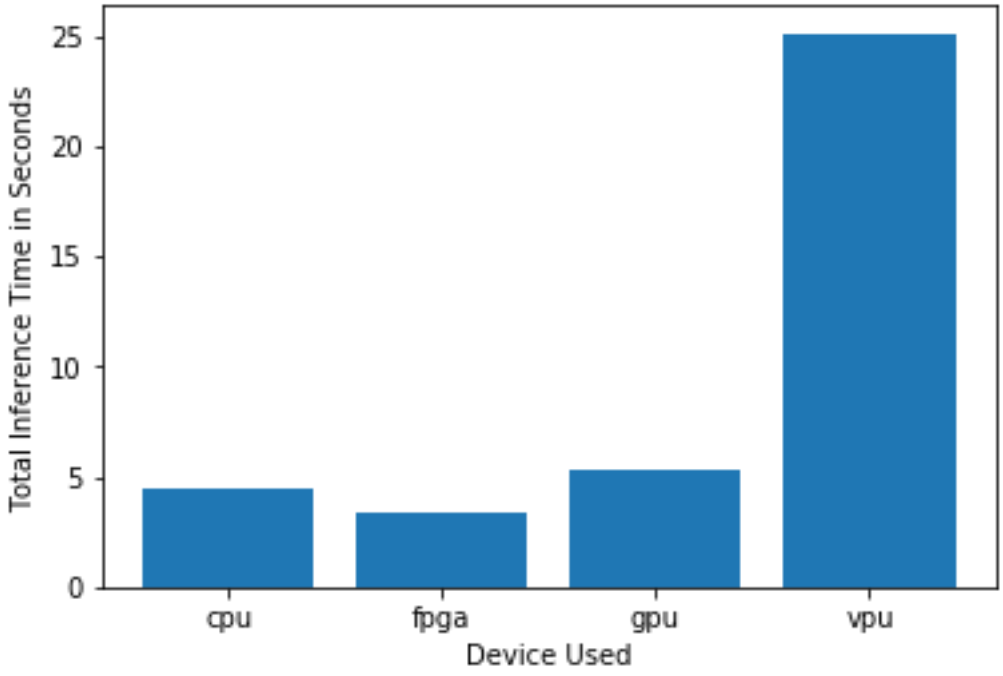
Maximum number of people in the queue	*2 (normal hours) - 5 (rush hours)*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

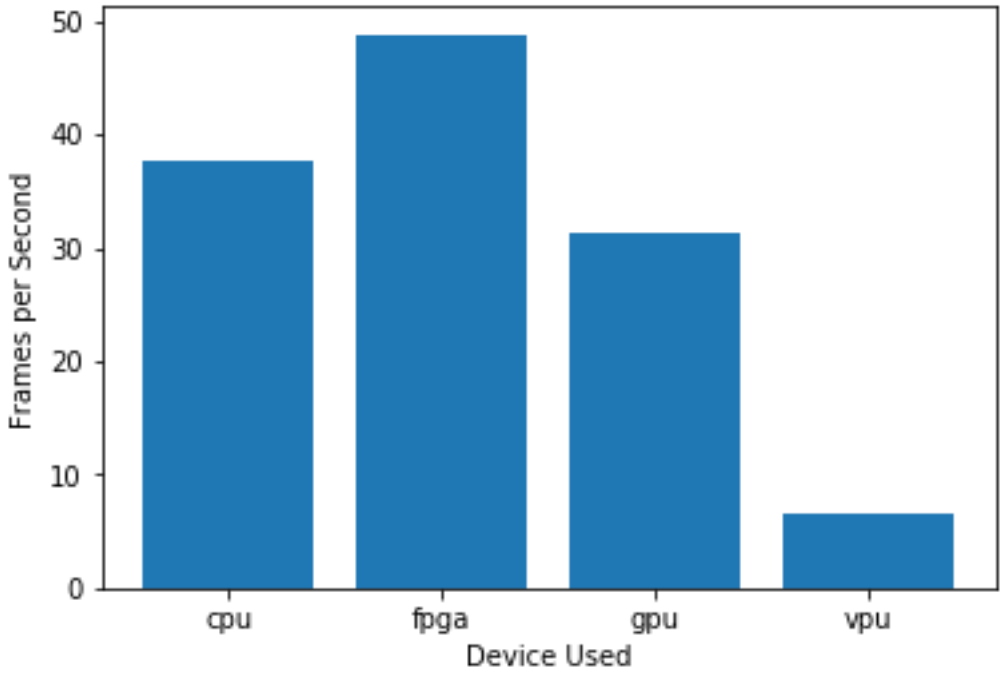
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

Now synthesize your points from above and provide a brief write-up describing why the chosen hardware is the best choice for this scenario. Be sure to discuss the client's requirements, the test results, and how these relate to one another (e.g., perhaps one of the devices performed better than the rest, but does not meet one of the client's requirements).

Write-up: Final Hardware Recommendation
• *Client is not interested in investing additional amount for hardware, which clearly shows that client cannot afford to purchase a VPU or FPGA. So, the best options are CPU or IGPU.* • *Power consumption means a lot for client, client wants to reduce as much as he can for electricity bill. CPU required more power for high performance. While the IGPU is capable to minimize the power consumptions because the clock rate for slice and un-slice are treated separately. By this we satisfy client's requirement.* • *Testing results shows that the IGPU takes less Inference time as compared to CPU and more than FPGA. But an FPGA does not fulfill the client's requirement, so we ignored it.* • *The IGPU process more frames as compared to CPU. But it takes more time to load the model. IGPU is the right hardware according to our client's requirement.*

Scenario 3: Transportation

Client Requirements and Potential Hardware Solution

Look through the scenario and find any relevant client requirements. Then, suggest a potential hardware type and explain how this hardware would satisfy each of the requirements.

Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
VPU

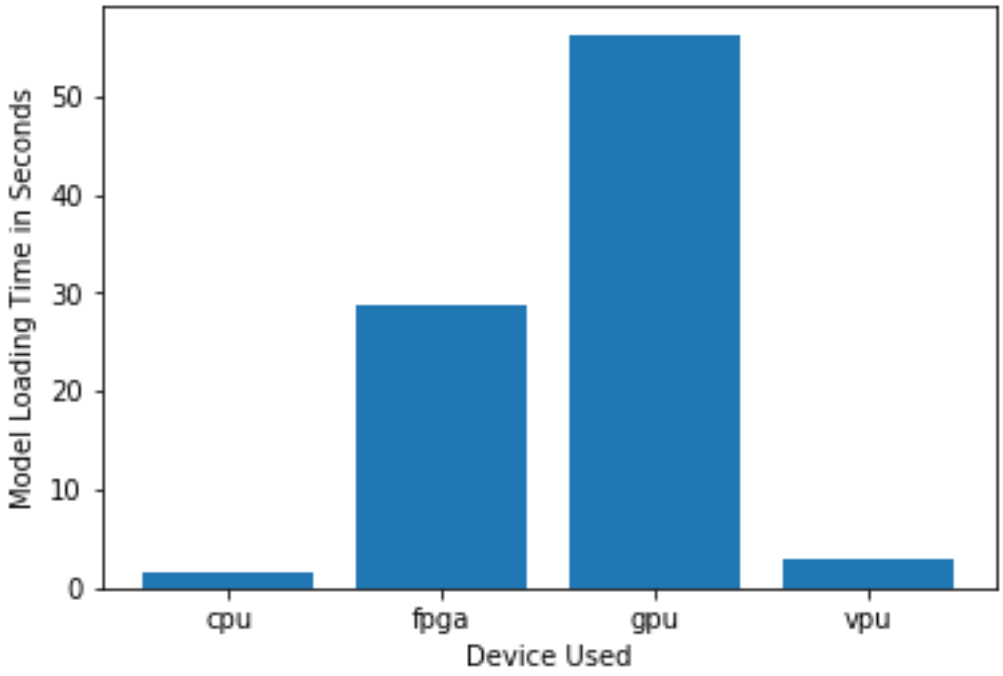
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
Client's budget for each machine is up-to $300, and client wants to save as much as we can possibly on hardware and future power consumption.	*If we look at price range, the VPU or NCS2 is acceptable as compared to other AI accelerators which cost 70-100$. NCS2 is easily deployed at the edge which is also low-power device.*
Currently the client is using CPU to process and visualize CCTV footage for security purpose, but not enough amount of processing power is available to run inference.	*The client's CPU required more processing power and the NCS2 can be used to run inference on the models because it requires low processing power to run inference.*

Queue Monitoring Requirements

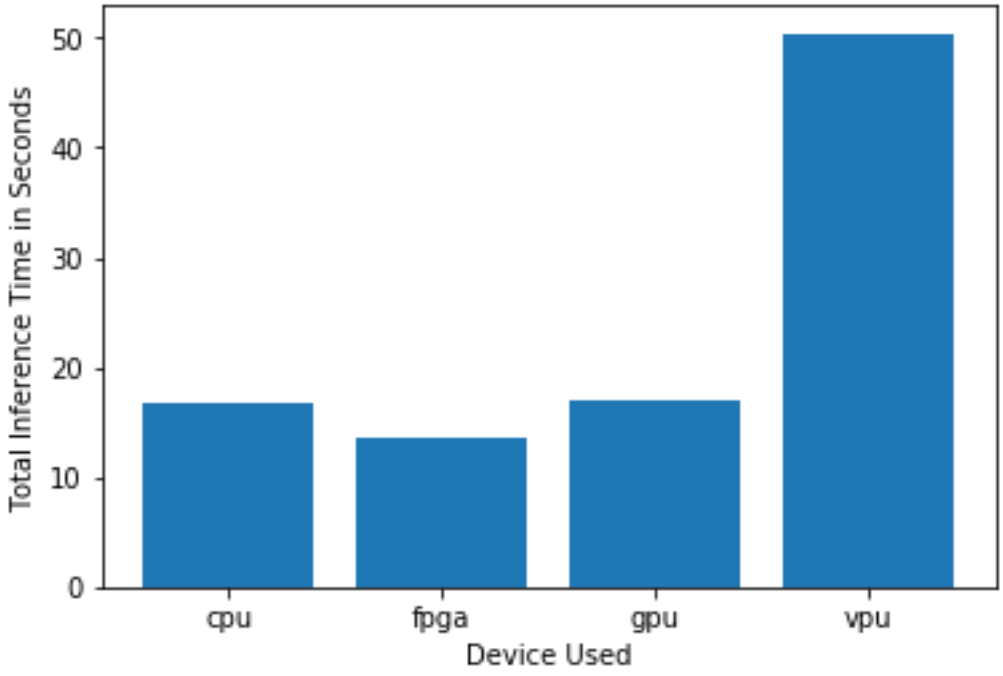
Maximum number of people in the queue	*6(non-peak hours) – 15(peak hours)*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

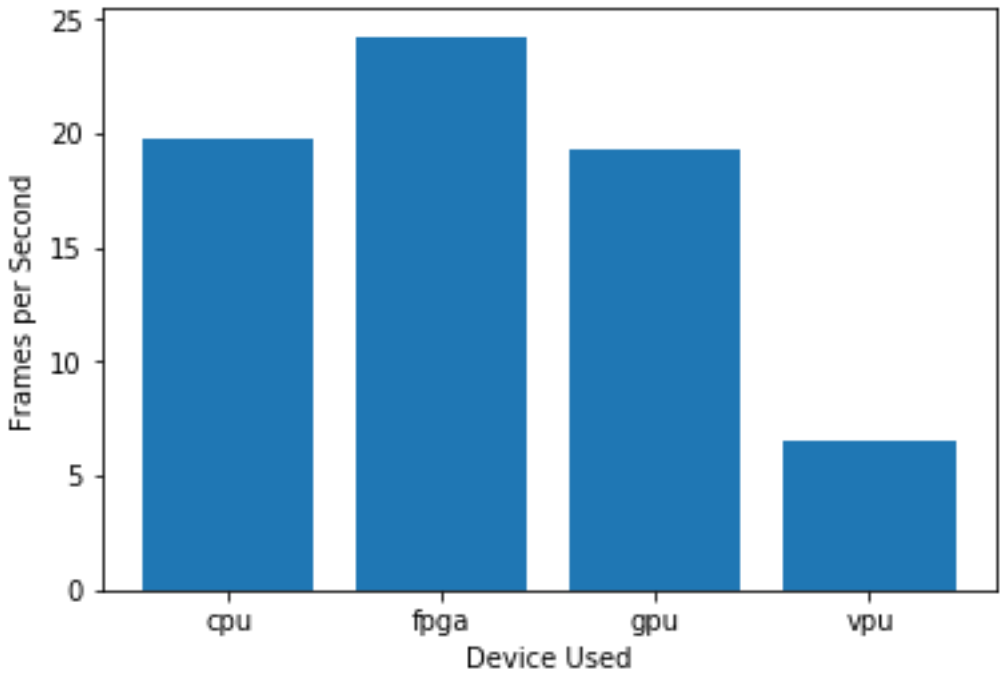
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

Now synthesize your points from above and provide a brief write-up describing why the chosen hardware is the best choice for this scenario. Be sure to discuss the client's requirements, the test results, and how these relate to one another (e.g., perhaps one of the devices performed better than the rest, but does not meet one of the client's requirements).

Write-up: Final Hardware Recommendation
• *According to client's budget, he cannot bear the FPGA which costs more than 300$. As client mentioned that the CPU's are in use for processing and visualization of CCTV footage and security purpose there is no processing power available to tun inference, so the NCS2 hardware meets the client's requirement.* • *As shown in results the VPU's inference time is higher as compared to CPU, IGPU and FPGA. But none of them meets the client's requirements.* • *FPGA's price is very high which is out of client's budget.* • *The VPU's model loading time is less than FPGA and IGPU, but more than CPU. As client already using CPU for some purpose, which is not capable to run inference on model.* • *The VPU or NCS2 is recommended according to the client's requirements.*