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

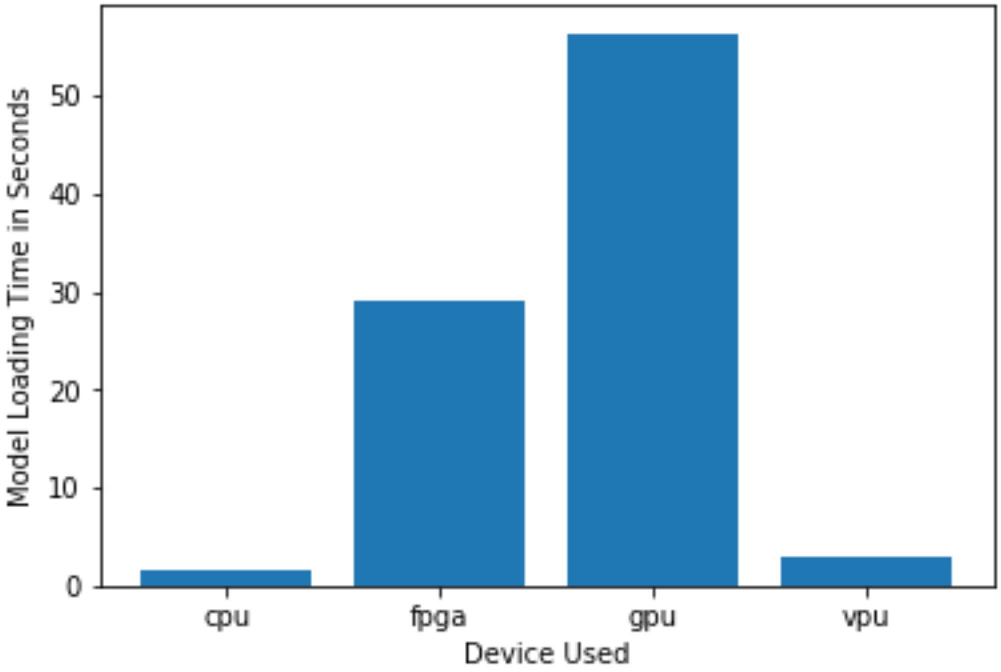
Requirement Observed **(Include at least two.)**	**How does the chosen hardware meet this requirement?**
The system installed by Naomi Semiconductors would need to be flexible so that it can be reprogrammed and optimized to quickly detect flaws in different chip designs.	FPGAs can be reprogrammed to adapt to new, evolving, and custom networks.
Naomi Semiconductors would ideally like a device to last for at least 5-10 years.	FPGAs have a long lifespan. For example, FPGAs that use devices from Intel's Internet of Things Group have a guaranteed availability of 10 years, from start of production.
Naomi Semiconductors has plenty of revenue to install a quality system and requires a system to run inference very quickly on a video stream	Although they are more expensive than other options such as VPUs and CPUs, once programmed with a suitable bitstream, FPGAs can execute neural networks with high performance and very little latency.
The Naomi Semiconductors floor runs 24 hours a day so that packaging continues nonstop	FPGAs are designed to have 100% on-time performance, meaning they can be continuously running 24 hours a day, 7 days a week, 365 days a year.

Queue Monitoring Requirements

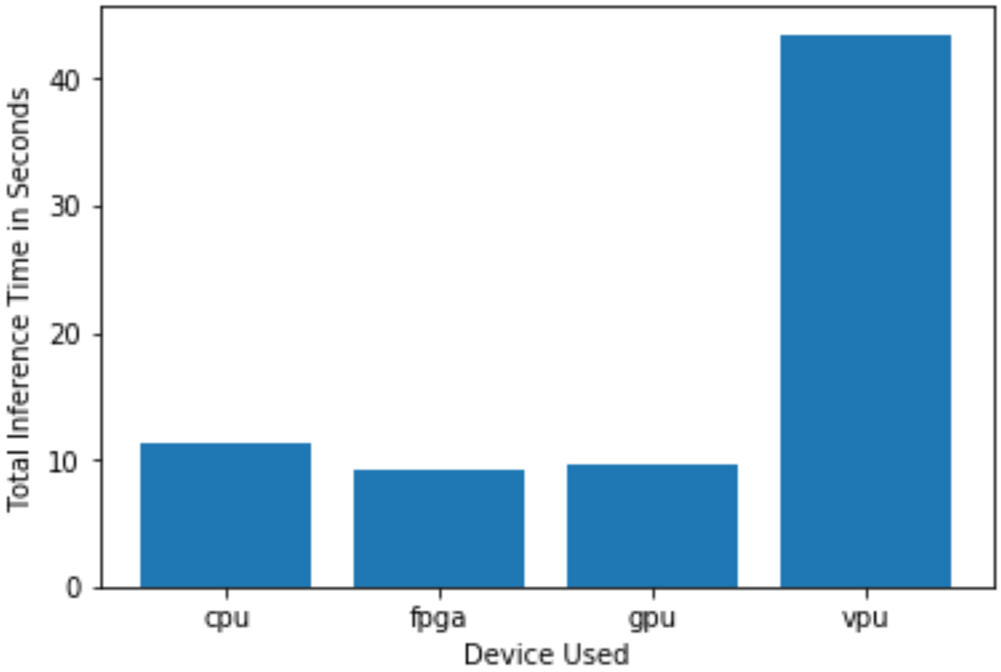
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

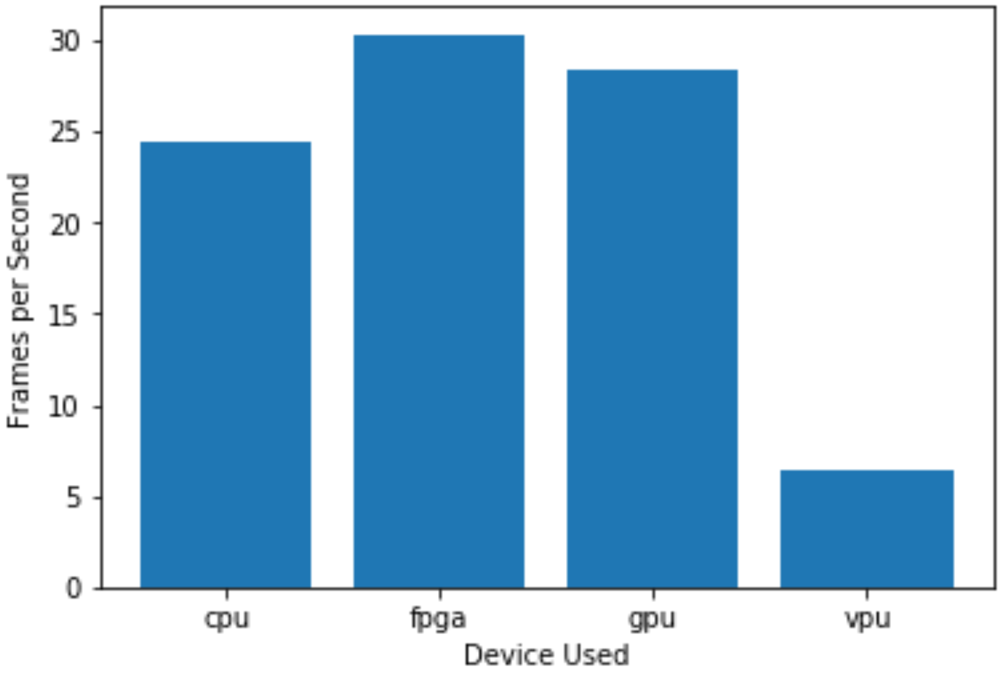
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
As expected the FPGA looks to be the best option in this scenario. Although the model load time is higher than both the CPU and VPU. The model is going to be run on a factory floor 24/7 365 days a year so this overhead is likely only going to have to be handled once. FPGAs are designed to handle this kind of continuous workload and have a guaranteed lifespan of 10 years which was a requirement of Naomi Semiconductors. Naomi Semiconductors also requires a system to run inference very quickly on a video stream and be able to handle camera feed of 30 – 35 frames per second. As seen in the above graphs FPGAs have the quickest inference time of the 4 pieces Hardware tested and they also had the best FPS of 30 frames per second. Another requirement of Naomi Semiconductors is that the chosen device needs to be flexible so that it can be reprogrammed. This is a core feature of an FPGA and therefore meets this requirement. Lastly Naomi Semiconductors has plenty of revenue to install a quality system so it can afford to purchase the FPGAs which are the best fit for their 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

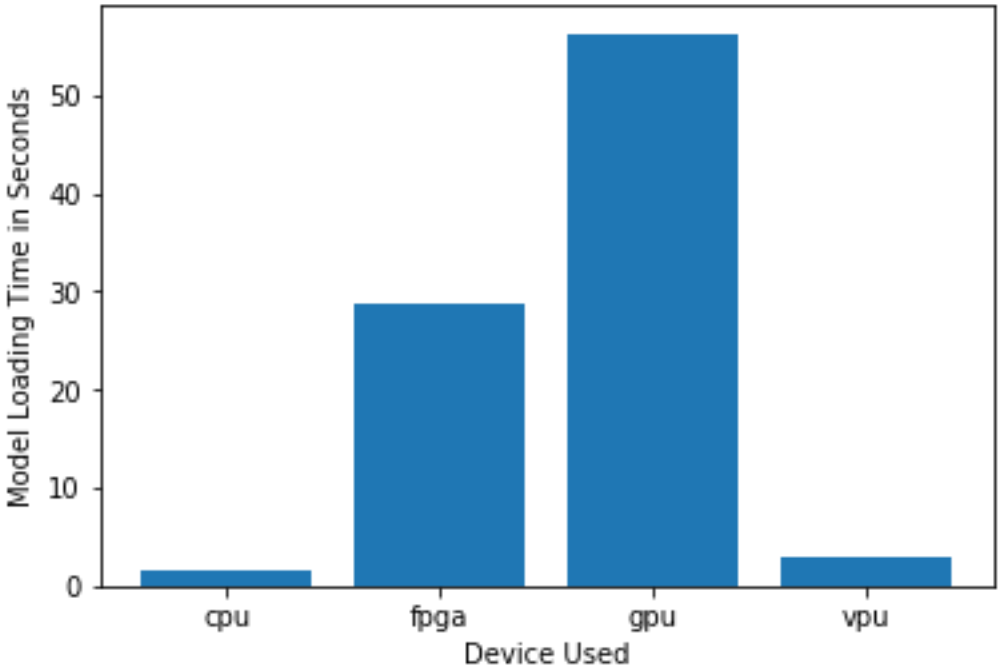
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
Does not have much money to invest in additional hardware	The client has Intel i7 core processors that are only used to carry out some minimal tasks that are not computationally expensive. By using these to handle inference we can save the client money on buying new hardware.
Wants to save as much as possible on his electric bill	IGPUs have configurable power consumption. The clock rate for the slice and unslice of IGPUs can be controlled separately. This means that unused sections in a GPU can be powered down to reduce power consumption.

Queue Monitoring Requirements

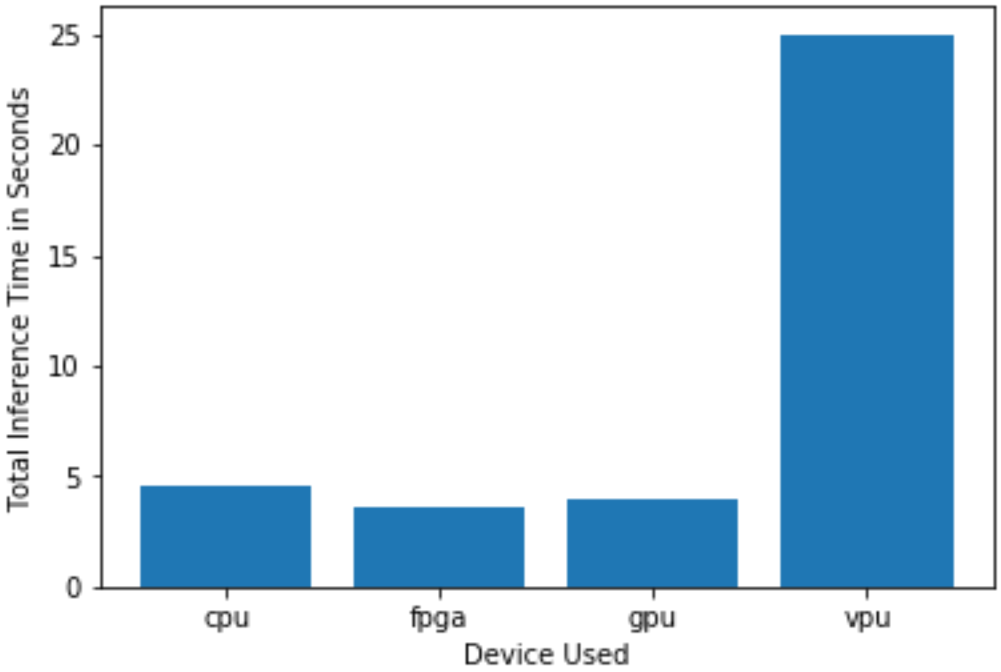
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

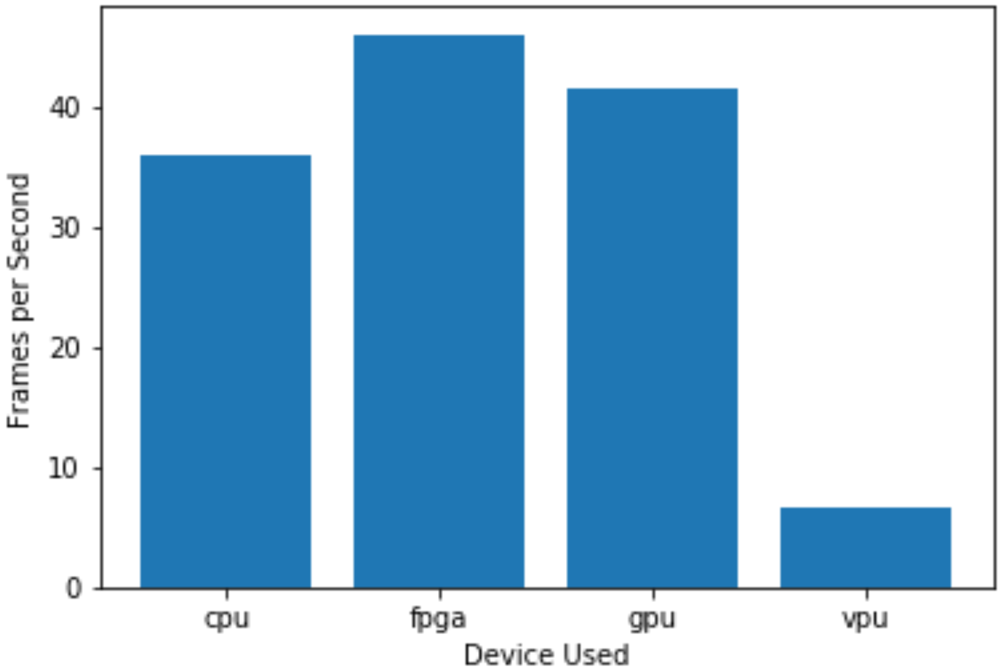
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
Despite having the slowest model load time the IGPU has one of the fastest total inference times and also one of the best frames per second rates. We can also utilise the clients existing Intel i7 core processors, meaning we won't have to buy any additional hardware to use an IGPU for inference. Lastly IGPUs have configurable power consumption which allows unused parts of the IGPU to be powered down reducing power consumption. These characteristics meet all of the clients requirements and for that reason I believe using an IGPU is the best option in this scenario.

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?
The budget allows for a maximum of $300 per machine	VPUs such as the Intel Neural Compute Stick have a price of between $70 - $100 per device which comes in well under the budget of $300 per machine.
Save as much as possible both on hardware and future power requirements.	VPUs are cheaper to buy than new CPUs, GPUs and FPGAs. They also are designed to have very low power consumption. This means that VPUs are the most cost efficient option when buying new hardware.
Inference must be handled alongside processes used to view CCTV footage for security purposes.	VPUs such as the Intel Neural Compute Stick can be easily added to existing hardware using a convenient USB3.1 plug and play interface. They will then handle the work load caused by inference.
Need to handle input from 7 CCTV cameras that are connected to closed All-In-One PCs	Each Intel Neural Compute Stick added can handle up to 4 inference requests at the one time. However Adding multiple sticks (or other Myriad X devices) will allow multiple inferences to run in parallel. So

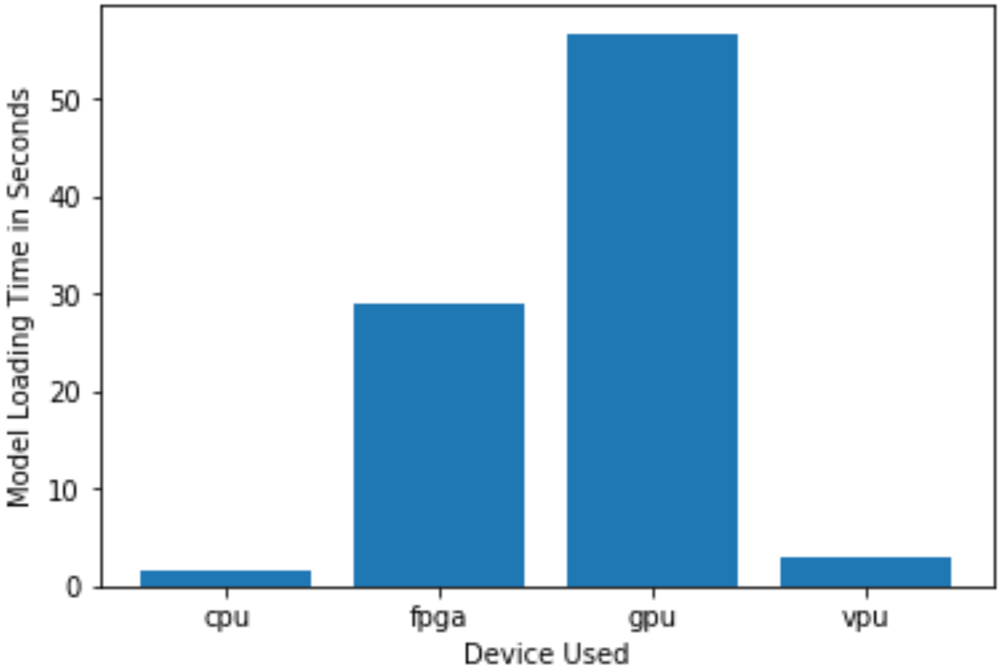
	adding multiple VPUs will allow us to handle the input from the 7 CCTV cameras at the one time.

Queue Monitoring Requirements

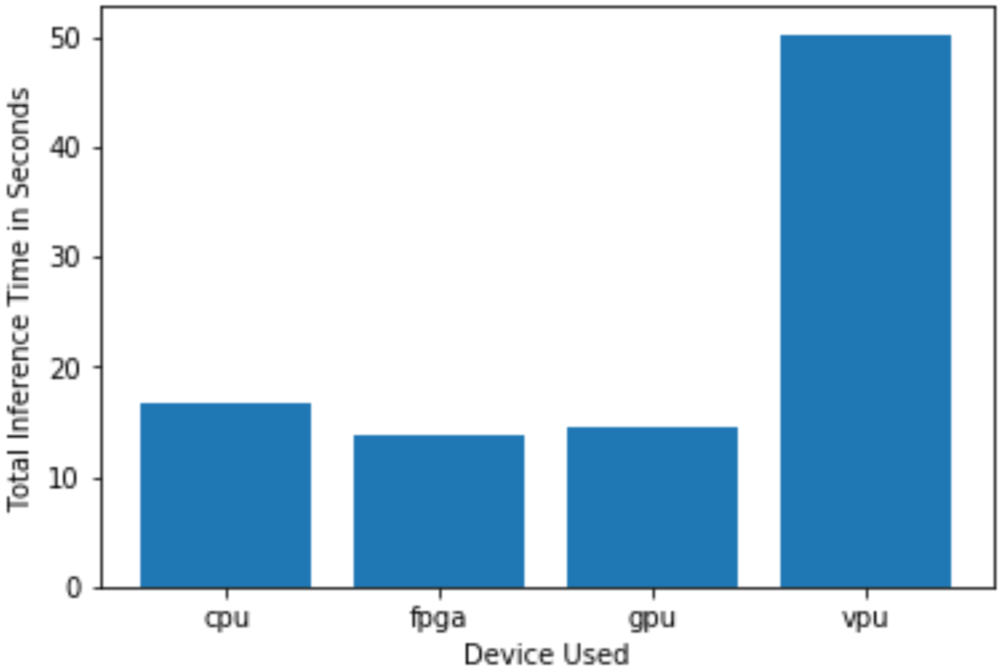
Maximum number of people in the queue	*15*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

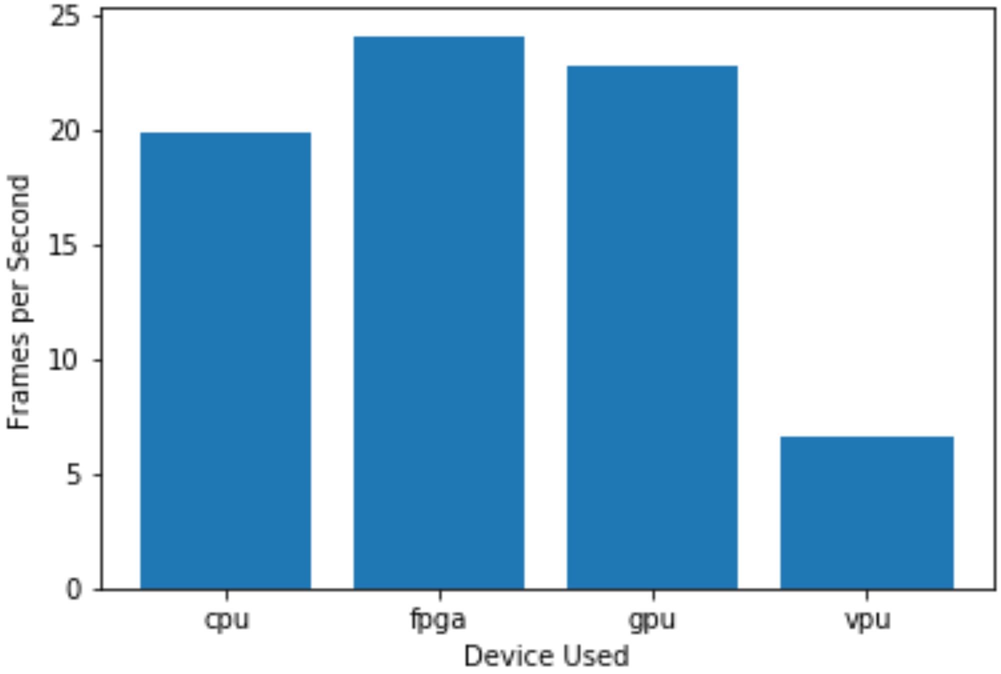
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
Although the VPU has significantly worse inference times and frames per second rate than the other devices I still think it is the best device for Delhi Metro Rail Services requirements. VPUs should still be able to handle identifying 7 – 15 people in a queue whilst also being a cost efficient solution. The clients priorities seem to be mainly cost related and the VPU is definitely the best option for this criteria. VPUs can easily be added to the clients existing hardware. They will perform best when handling multiple inference requests in parallel for the 7 CCTV cameras the client uses. They also have a low power consumption so the clients cost of running their systems won't drastically increase. For these reasons I think a VPU is the best the option in this scenario.