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)**
CPU/FPGA

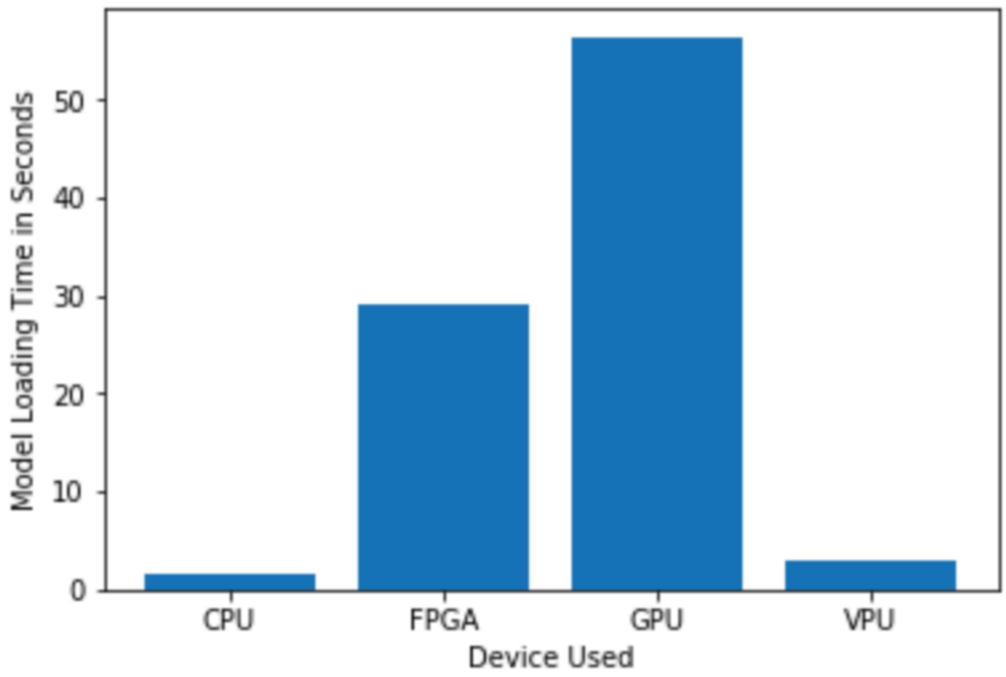
Requirement Observed **(Include at least two.)**	**How does the chosen hardware meet this requirement?**
The floor needs to be running 24 hours a day and the client requires the image processing task to be completed five times per second.	*FPGAs are designed to have 100% on-time performance, meaning they can be continuously running 24 hours a day, 7 days a week, 365 days a year.*
To be able to detect chip flaws without slowing down the packaging process and to run inference on the video stream very quickly.	*FPGAs have **High performance, low latency.***
The system would also need to be flexible so that it can be reprogrammed and optimized to quickly detect flaws in different chip designs, and they would ideally like it to last for at least 5-10 years.	*FPGA is flexible and the lifetime is also long and about 10 years.*

Queue Monitoring Requirements

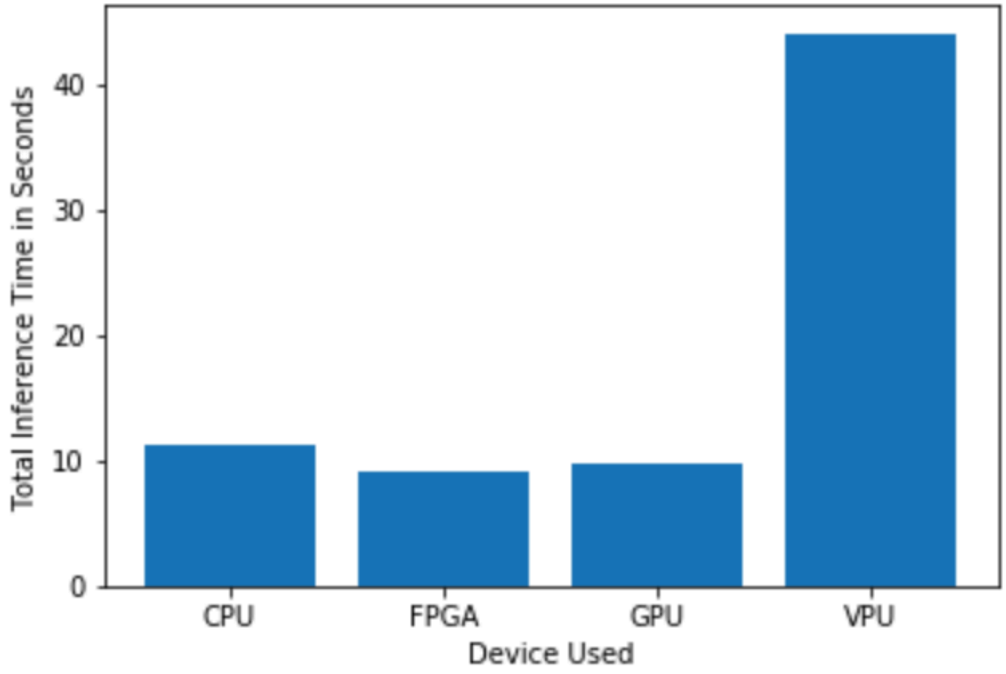
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP32*

Test Results

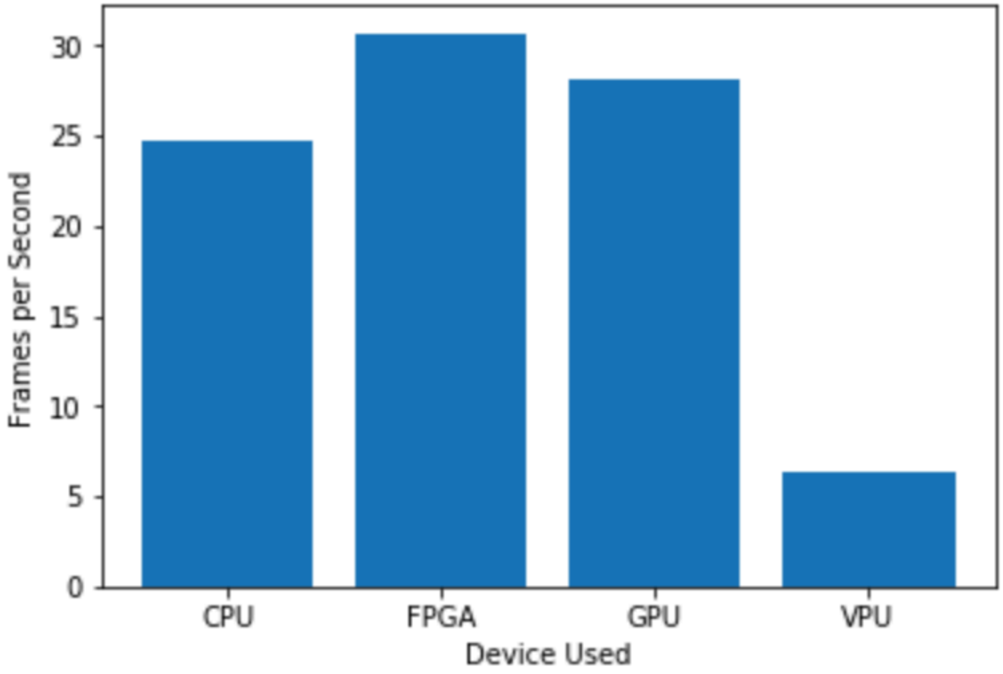
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
FPGA_time ~30 + 9s (Load time + Inference time) *GPU_time~55 + 10s* *CPU_time~11 + 2s* *VPU_time~45 + 3s* *Comparing the timing calculated from the calculations, it seems that CPU has the lowest timing and is fastest. However, considering the client's requirements FPGA would be the best option for the following reasons:* • *Their lifetime is about 10 years which is what the client wants.* • *It is flexible and has 100% on-time performance and can be used 24/7/365.* • *They have high performance and low latency. The only downside is the load time. But that is only one time. Once the model is loaded then inference can be performed quickly.*

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)
CPU/VPU

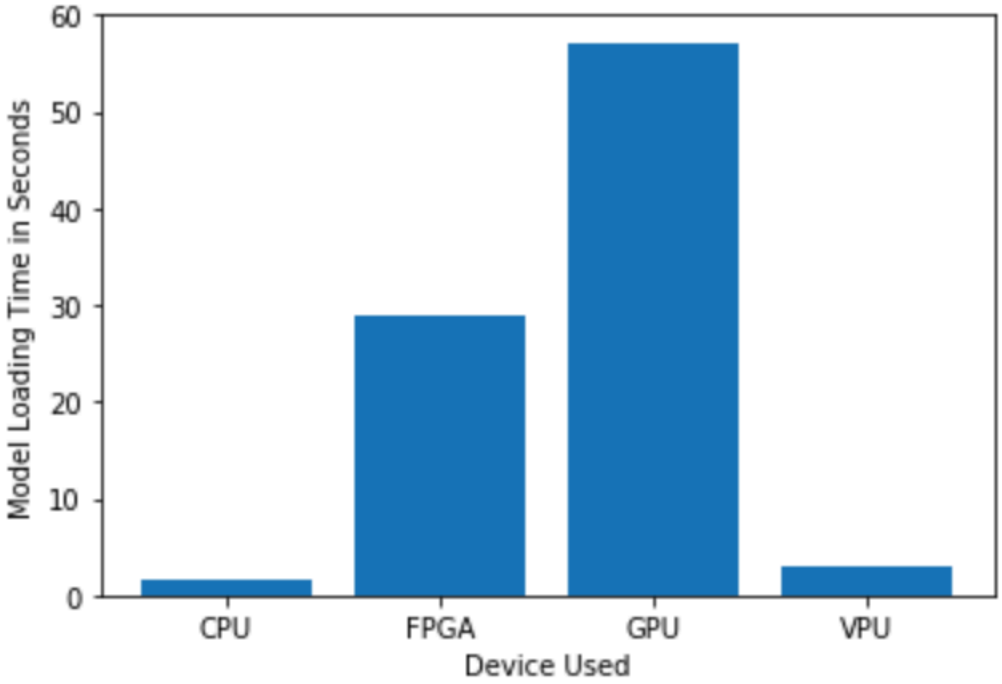
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
Client does not have much money to invest in additional hardware	*VPUs are very cheap and cost about 70-100$.*
Client would like to save as much as possible on his electric bill.	*VPU has a very low power consumption of only 1-2 watts.*
Do rapid computations to guide people into less congested queues.	*VPUs have specialized accelerators for image processing.*

Queue Monitoring Requirements

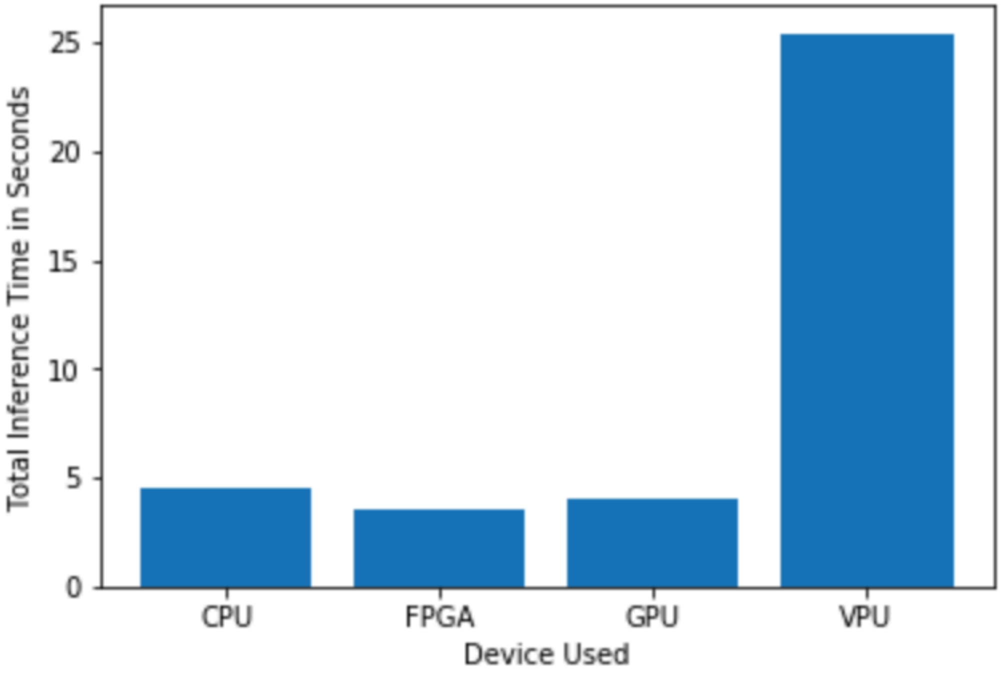
Maximum number of people in the queue	*1*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

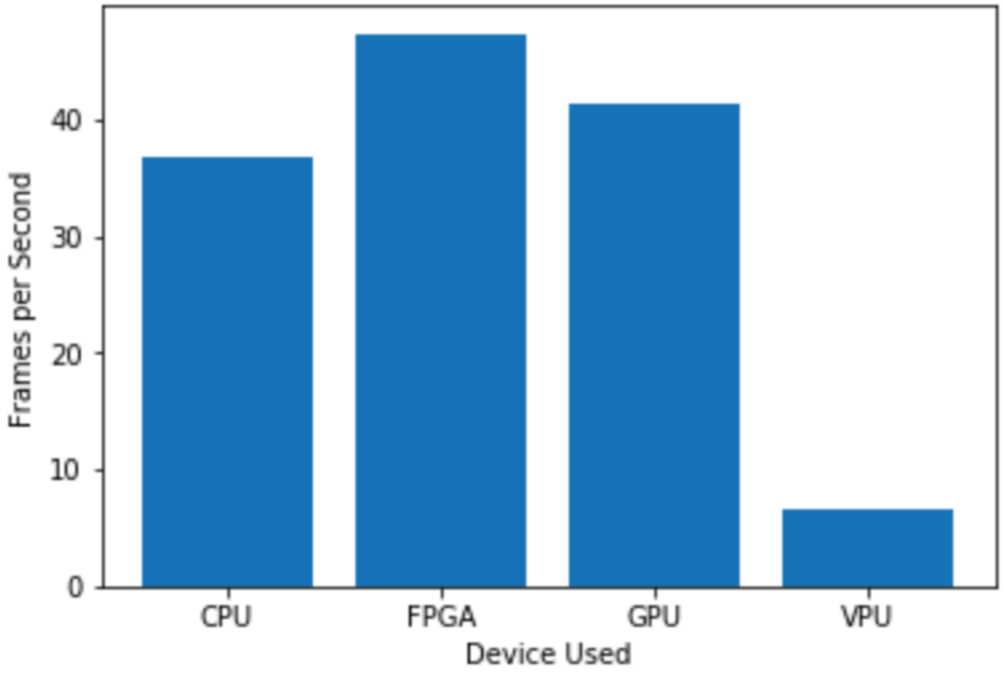
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
FPGA_time ~29 + 3s (Load time + Inference time) *GPU_time~57 + 4s* *CPU_time~5 + 2s* *VPU_time~25 + 3s* *Comparing the timing calculated from the calculations, it seems that CPU has the lowest timing and is fastest. However, considering the client's requirements VPU would be the best option for the following reasons:* • *It has low power consumption and would allow the client to save on the electricity bill.* • *They are very cheap and satisfy the financial limitations of the client.* • *They have specialized accelerators for image processing. The only downside is the load time. But that is only one time. Once the model is loaded then inference can be performed pretty quickly and the inference time is comparable with CPU as well.*

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)**
Intel Atom Processor/IGPU

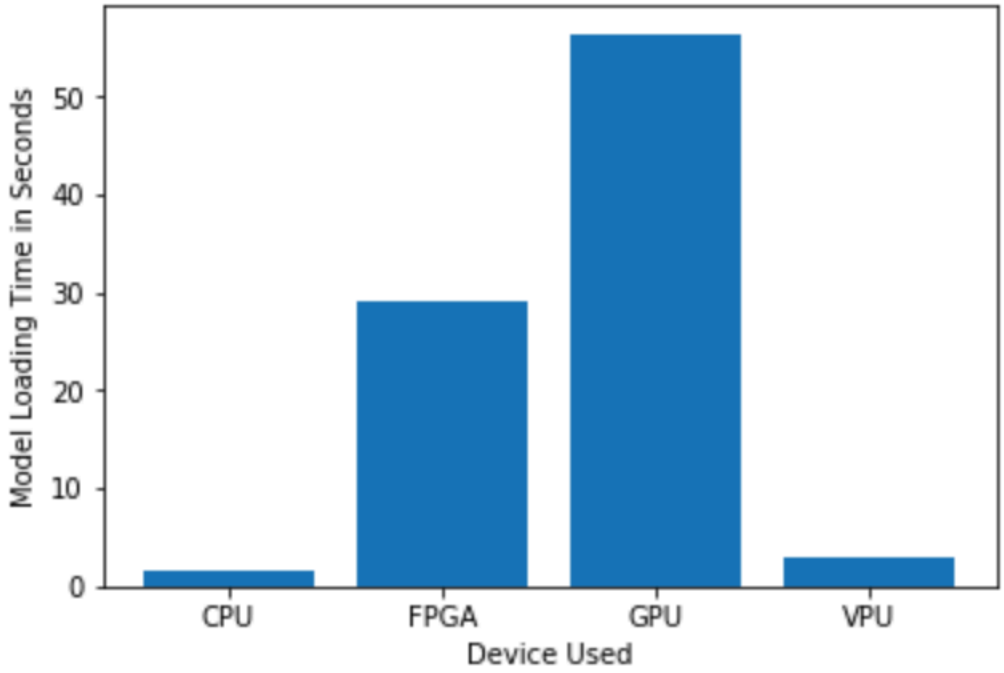
Requirement Observed **(Include at least two.)**	**How does the chosen hardware meet this requirement?**
The client's budget allows for a maximum of $300 per machine.	*It is on the same chip as CPU and only costs 57$*
The client would like to save as much as possible both on hardware and future power requirements.	*IGPUs have **Configurable Power Consumption.** The clock rate for the slice and unslice can be controlled separately. This means that unused sections in a GPU can be powered down to reduce power consumption.*
There are lots of images and workload to be processed by each pc.	*IGPUs have better performance with larger batch size. IGPUs generally can handle a much larger number of processes at once.*

Queue Monitoring Requirements

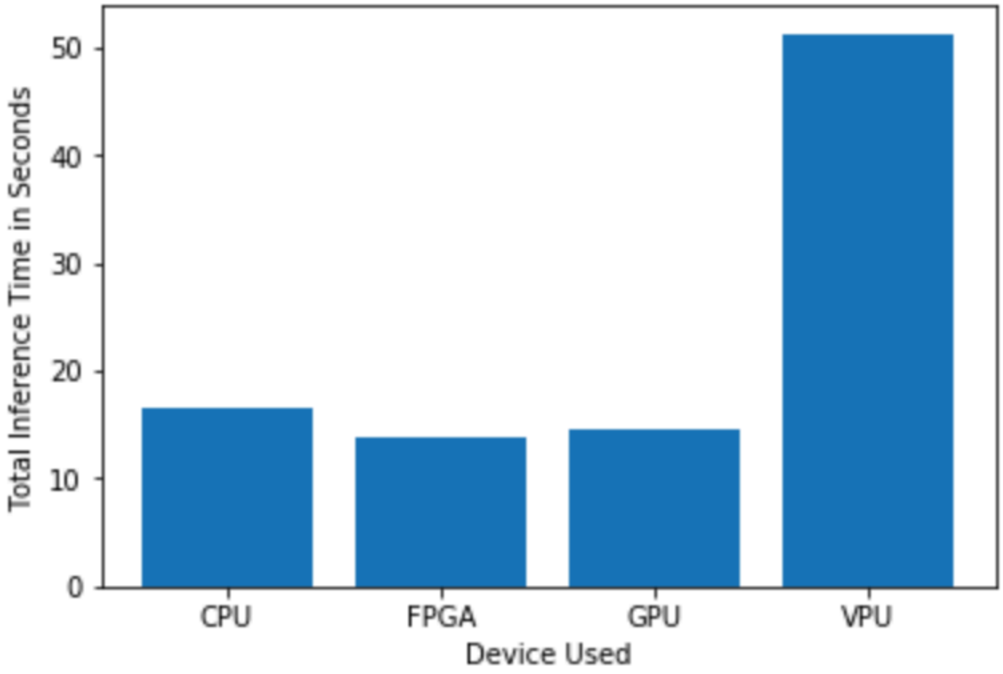
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

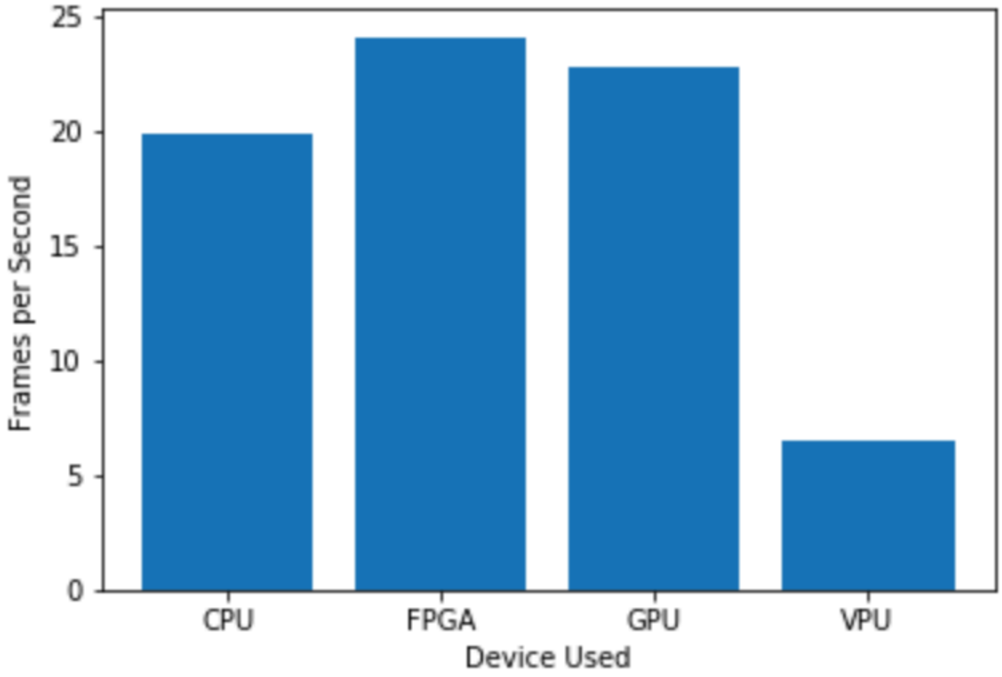
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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

FPGA_time ~30 + 13s (Load time + Inference time)
GPU_time~55 + 15s
CPU_time~2 + 17s
VPU_time~3 + 50s
Comparing the timing calculated from the calculations, it seems that CPU, GPU, and FPGA have comparable timings. However, to make the final choice we need to consider client's requirements as well:

- *There are lots of images and workload to be processed by each pc. IGPUs are suitable for batch processing and taking advantage of this we can even reduce the inference time more.*
- *IGPUs have Configurable Power Consumption and allow for power saving.*
- *Considering the client's budget, they would be the best option. Because FPGAs are much more expensive, and CPU would not be able to achieve as much performance as IGPU for large number of batches.*

Therefore, IGPUs would be the best option.