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

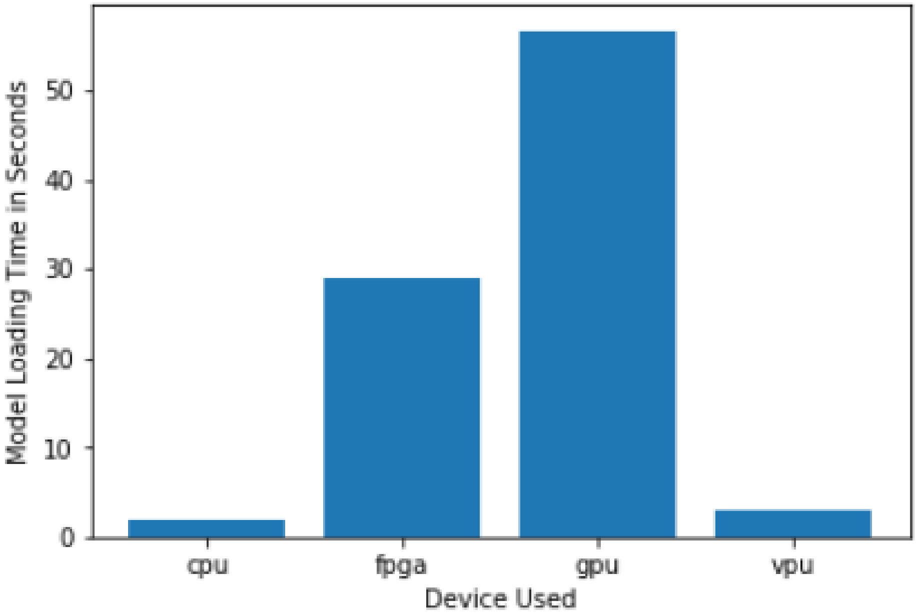
Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
The client requires a flexible so that it can be reprogrammed and optimized to quickly detect flaws in different chip designs.	FPGA can be reprogrammed to meet client needs.
The client requires very fast inference on the video stream.	FPGA can run multiple operations in parallel thus providing very fast inference.
The client requires the hardware to last for at least 5-10 years	FPGA can easily last for 10 years and is also in industrial environments.

Queue Monitoring Requirements

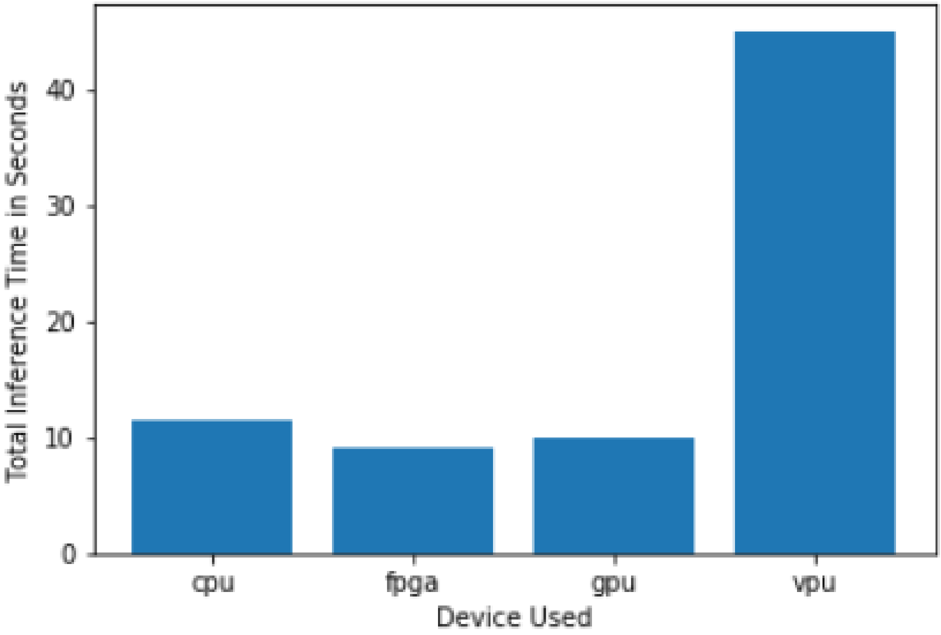
Maximum number of people in the queue	2
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

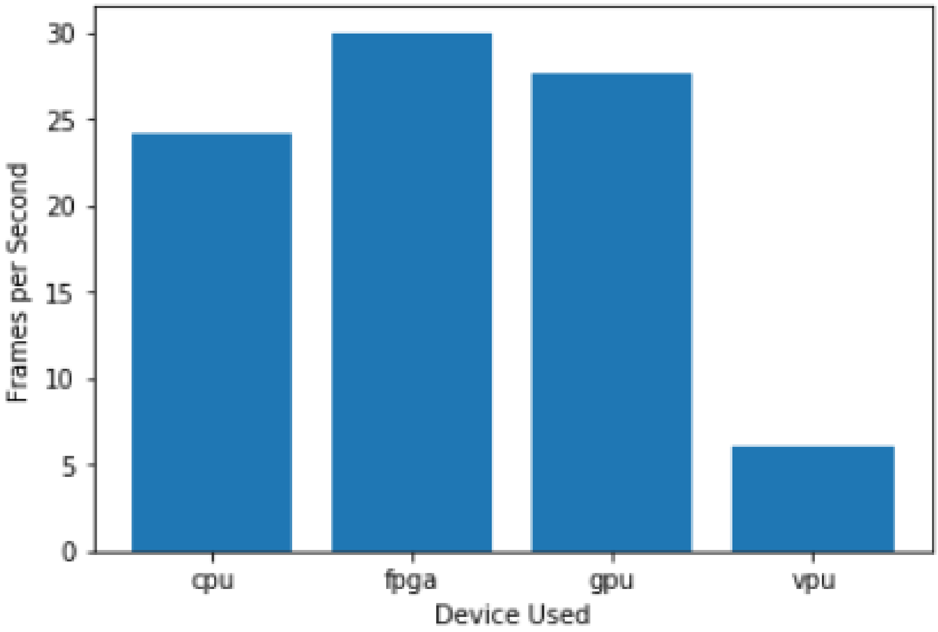
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 It can be reprogrammed easily which satisfies one of the client requirements. It is also the fastest among the other devices which is absolutely great for the client. However, it takes some time to load the model but the client has no requirement for that, so we good. FPGAs are also robust and can last for over 5 years.

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)
VPU

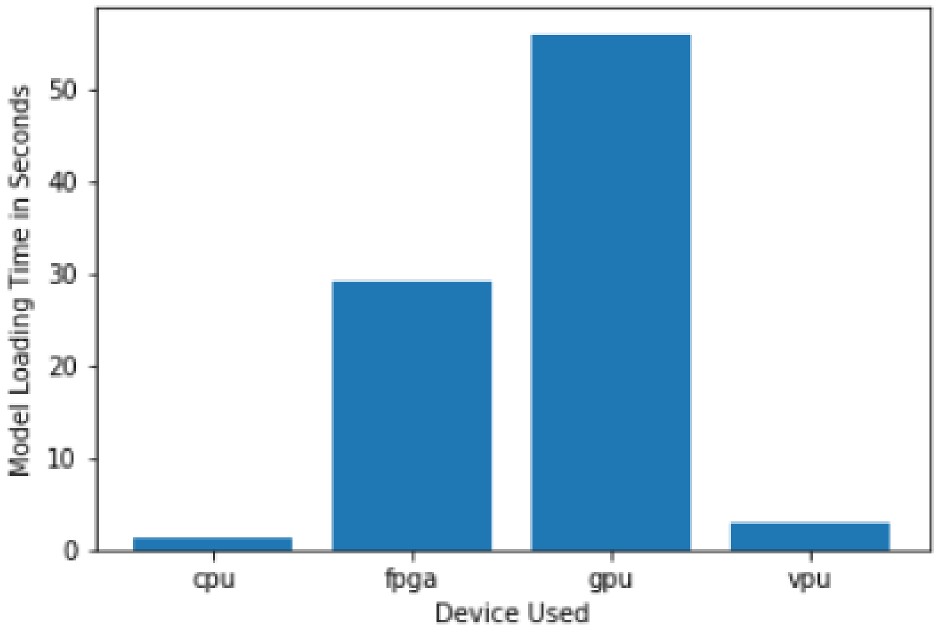
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client has low budget.	VPU would perfectly fit in the price range as it is less expensive than many other AI accelerators
The client requires low power consumption.	VPUs are low low-power devices. Although CPUs are available, but they consume much higher power.

Queue Monitoring Requirements

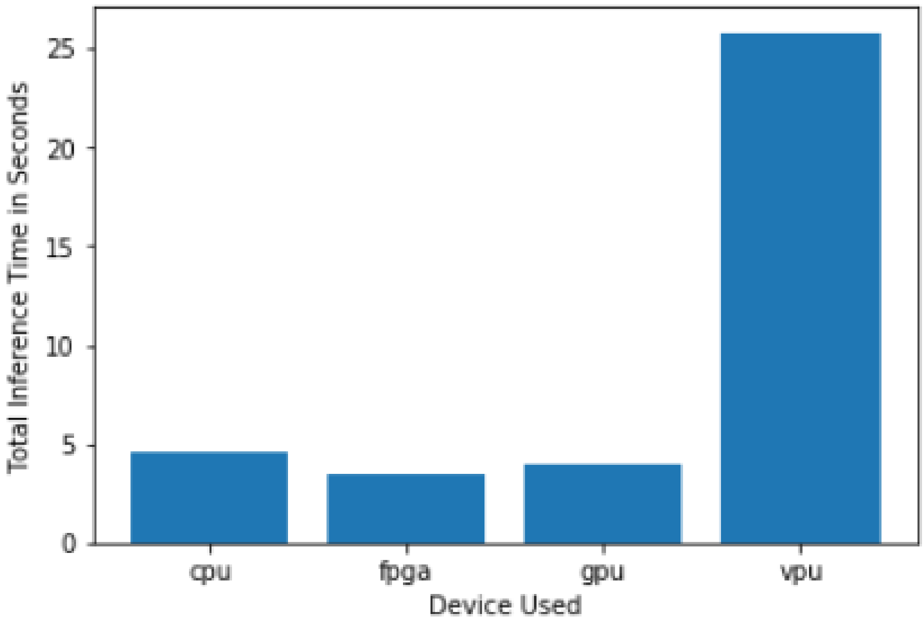
Maximum number of people in the queue	5
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

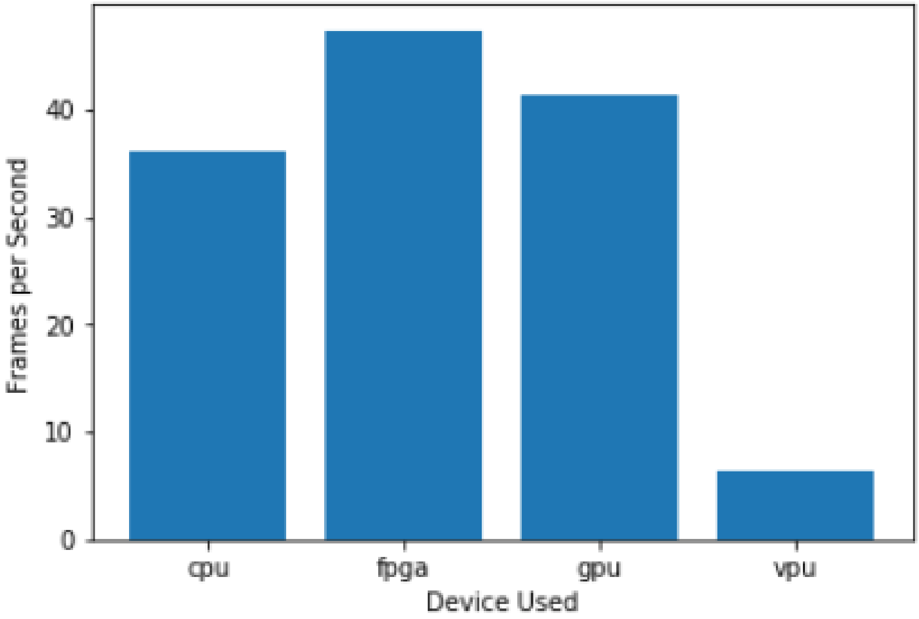
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
VPU It has the lowest power consumption among the other devices, and it doesn't cost much, so the client would be happy with it. However, it takes really long time to do inferencing on the video, hence low fps. VPU is fast at loading the model which the client might find helpful. The client might need to upgrade the device in the future when he gets more money. For now, VPU satisfies the client needs.

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)
CPU

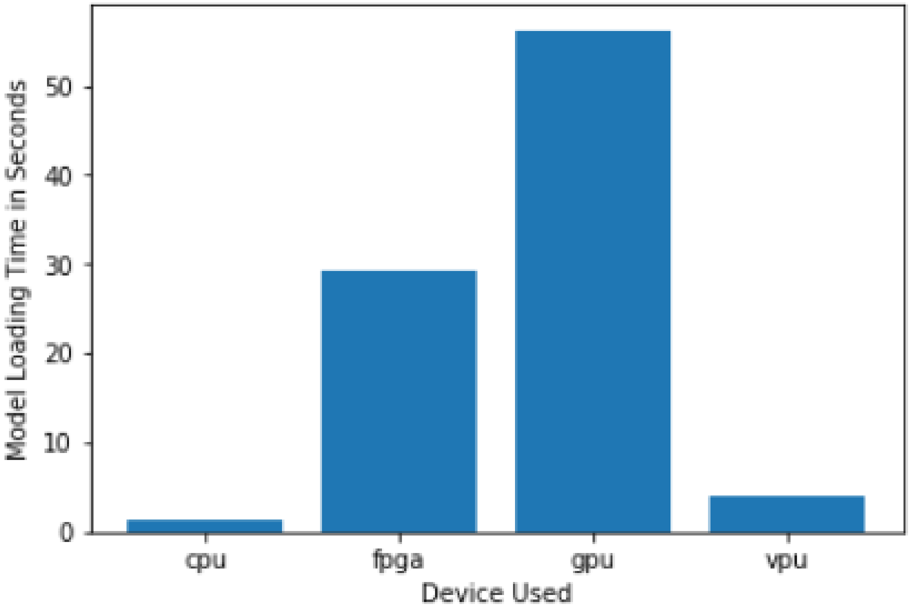
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client would like to save as much as possible both on hardware and future power requirements ($300 for each device).	CPUs can be of moderate price and give excellent performance.
The client would like to run the inference in real-time quickly.	CPUs can load the model very quickly and do the inference relatively fast.

Queue Monitoring Requirements

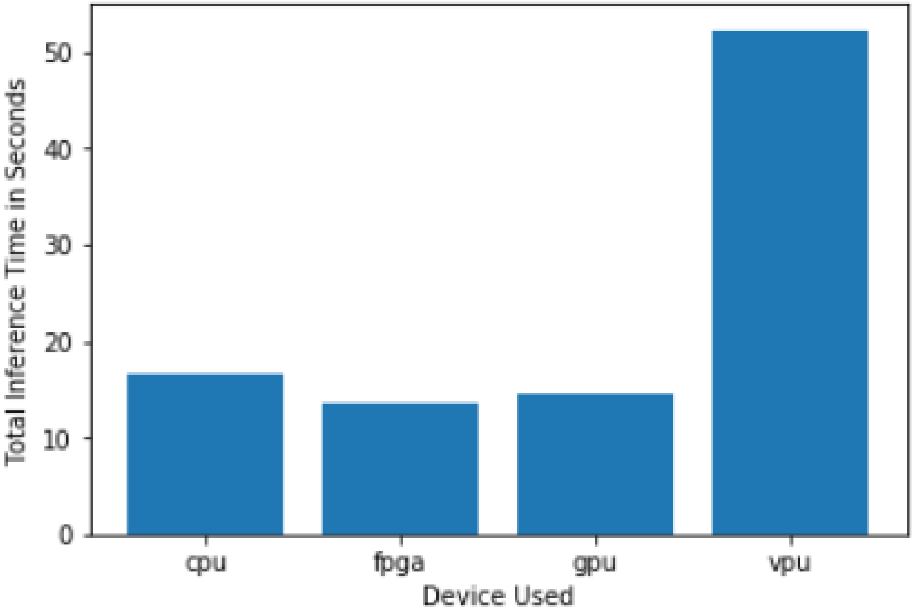
Maximum number of people in the queue	20
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

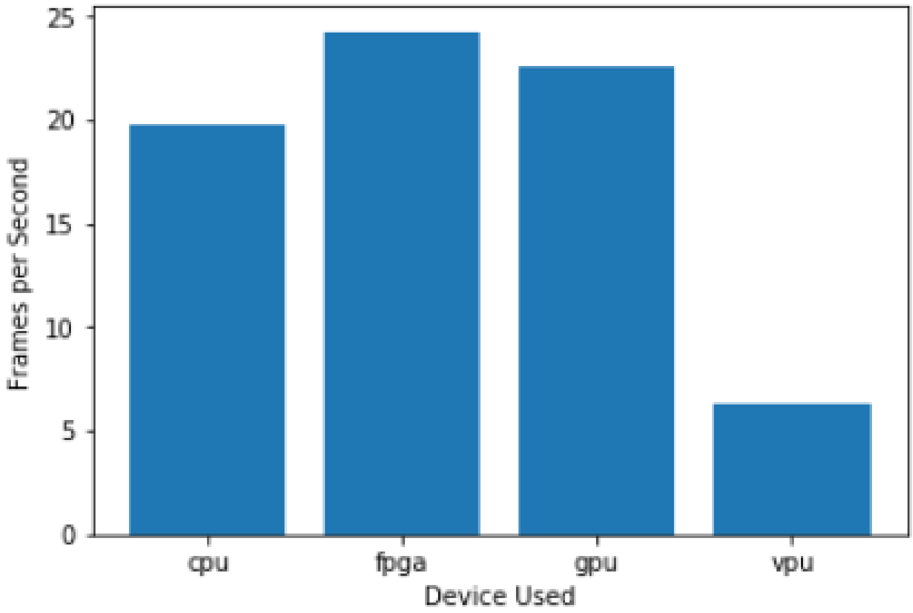
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
CPU or GPU Both devices can be used but it depends on the specifications and the price of each one. Both have approximately the same price which can be below 300$ as the client wants. As shown in the above figures, both can achieve similar, yet great performance. But CPU is way faster than GPU when it comes to model loading. VPU cannot be used here because it is slow and the client wants quick real time monitoring.