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

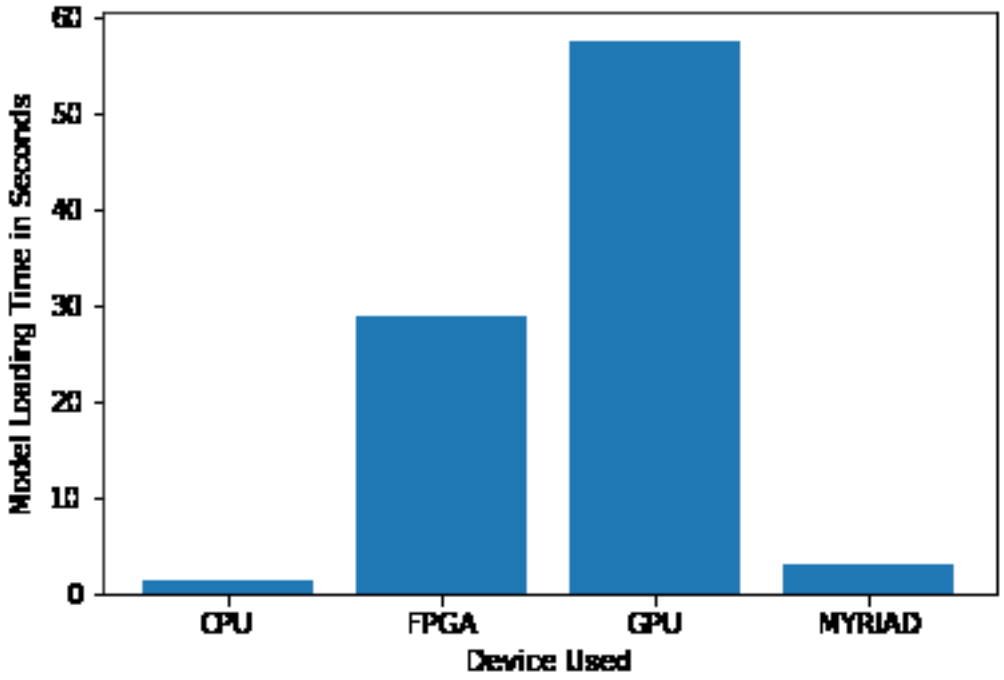
Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
Must run 24 hours a day because the factory floor runs 24 hours a day, every day.	*FPGAs are very robust and can run for 24-7-365.*
Must be reprogrammable and flexible. First it will be used to monitor the number of people in the factory line, then it will be used to view chip designs to find flaws. New designs are created regularly.	*FPGAs are field-programmable, which means they can be reprogrammed with different bitstreams by the customer.*
System must last at least 5 – 10 years	*FPGAs have a very long lifespan and are guaranteed to be available for at least 10 years.*
Inference must be done quickly in order to detect chip flaws without slowing down the packaging process.	FPGAs have very high performance and low latency.

Queue Monitoring Requirements

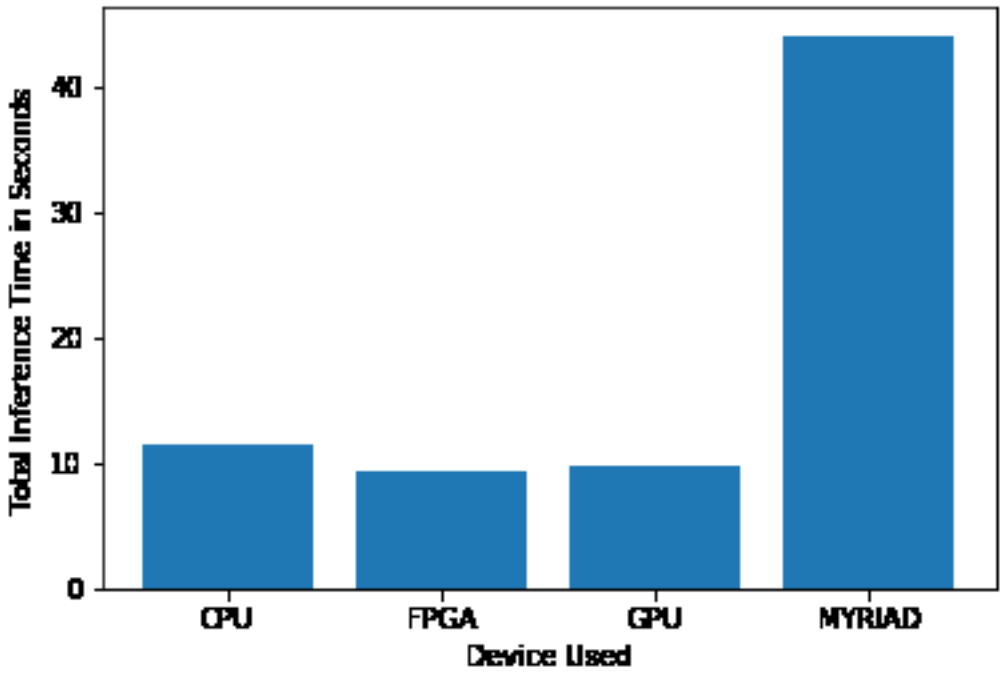
Maximum number of people in the queue	*2*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

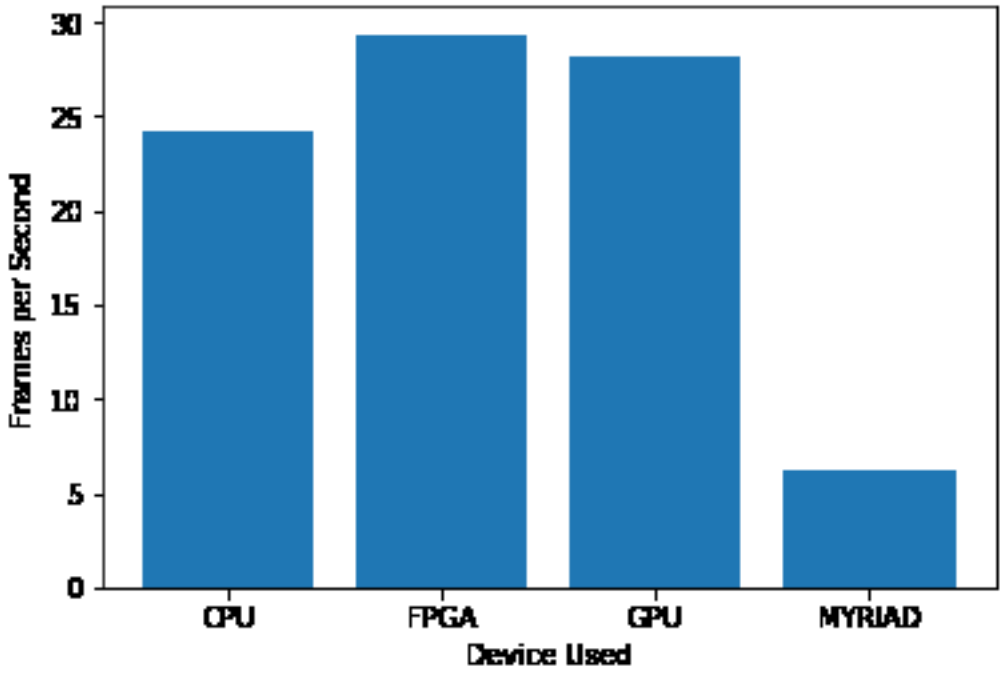
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
The client needs a high-performance device that can be reprogrammable and will last for at least 10 years. The FPGA fits the bill. As can be seen from the results above, the FPGA had the highest FPS and the lowest inference time which shows it's superior performance to all the other devices. The model load time, however, was the second lowest. A relatively slow model load time is not a concern because this is a one-time hit. After the model is loaded, the rest of the time is spent on actual inference. The FPGA can be reprogrammed to optimize it for the different workloads the client will throw at it.

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

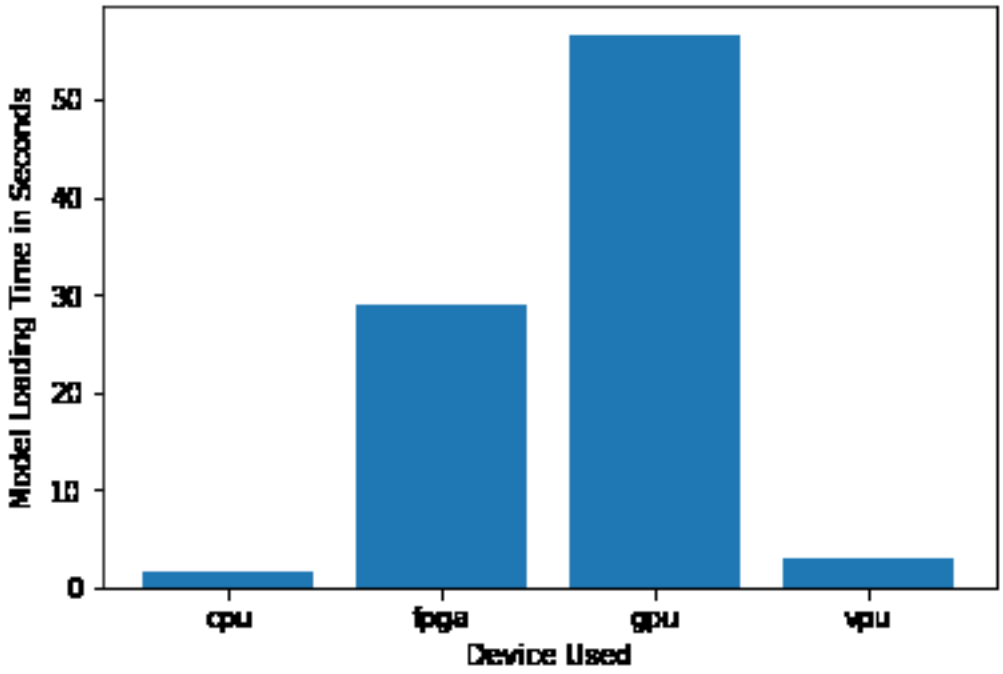
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
Currently installed systems have underutilized Intel Core i7 processors.	*Current PCs have CPUs that can be used for inferencing.*
No money for extra hardware.	*He has to use the hardware he already has, which are the CPUs.*
[TODO: Type your answer here]	*[TODO: Type your answer here]*
[TODO: Type your answer here]	*[TODO: Type your answer here]*

Queue Monitoring Requirements

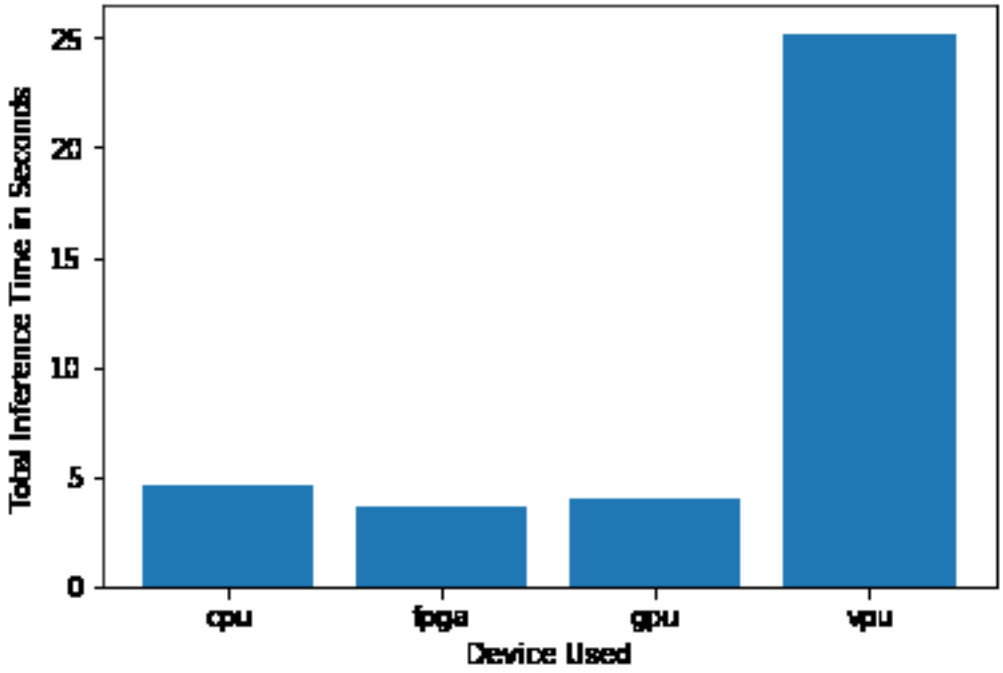
Maximum number of people in the queue	*2*
Model precision chosen (FP32, FP16, or Int8)	*FP32*

Test Results

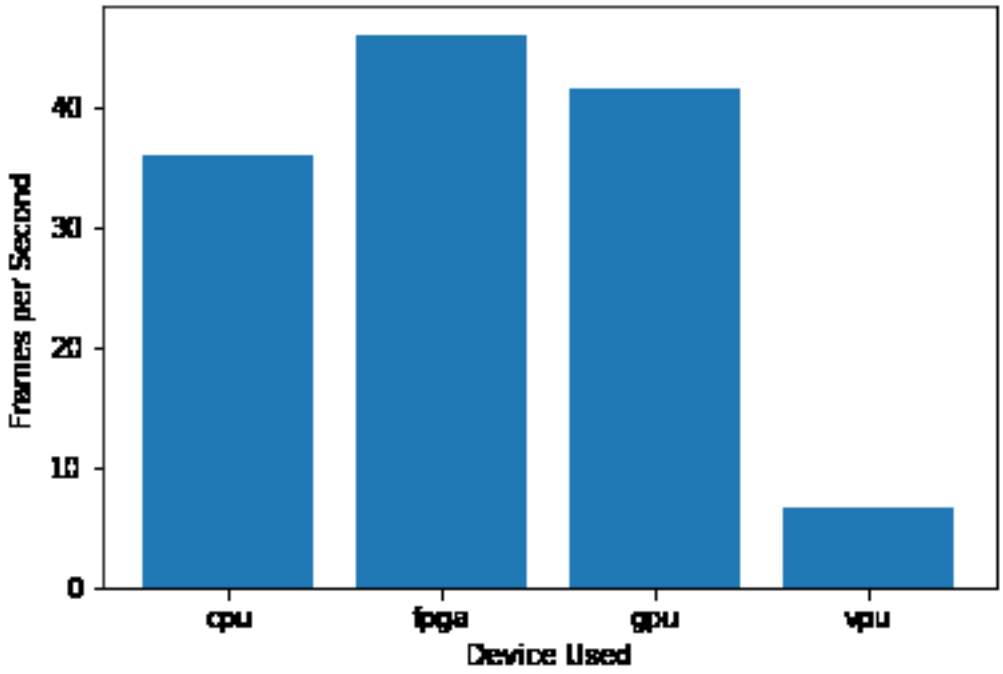
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
Mr. Lin does not have money to spend on extra hardware so he must use the hardware he already has. Additionally, his current processors are underutilized. Initially, I recommended the CPU device but the results show that the GPU may be the better choice. The GPU performed better than the CPU with a higher fps and a lower total inference time compared to the CPU. The model load time was much slower on the GPU. A relatively slow model load time is not a concern because this is a one-time hit. After the model is loaded, the rest of the time is spent on actual inference. The GPU is also already integrated into Mr. Lin's Intel Core i7 processors so by using the GPU he does not have to purchase new hardware and the CPU remains free to do its other tasks. Also, the GPU can be more power efficient than the CPU, which will help Mr. Lin save as much as possible on the electric bill. *The FPGA performed the best but it is too expensive and will require extra hardware purchases.*

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)
NCS2

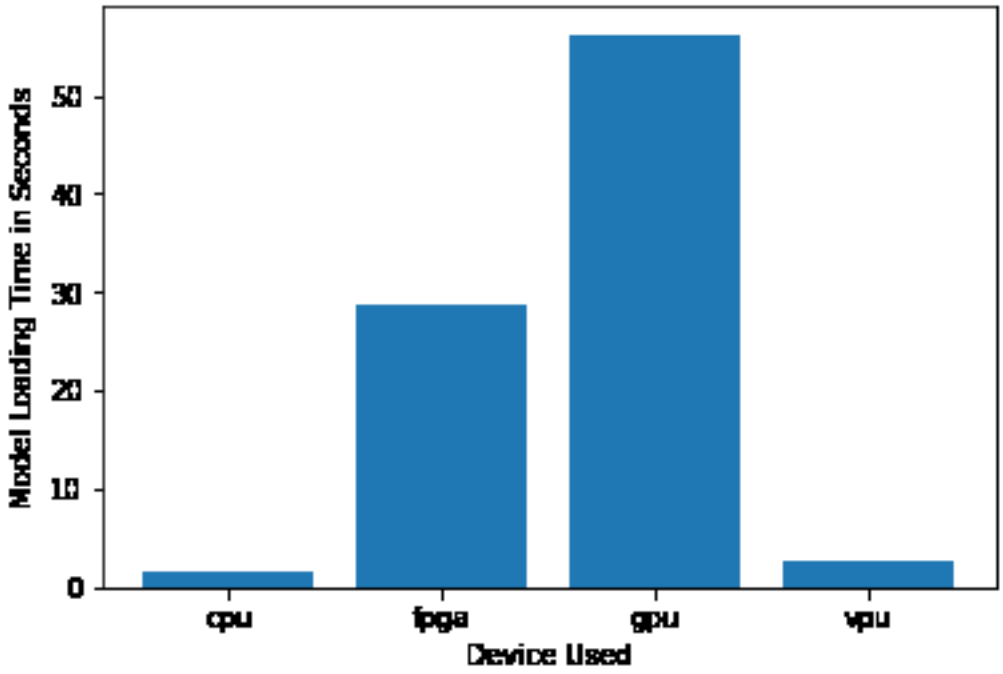
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
Current CPUs are being used for other purposes.	*Must use a separate device.*
Can spend up to $300 per machine for a new device.	*The Intel NCS2 retail price is $79.*
Client wants to save on power requirements.	*NCS2 is a low-power device with 1W TDP.*
[TODO: Type your answer here]	*[TODO: Type your answer here]*

Queue Monitoring Requirements

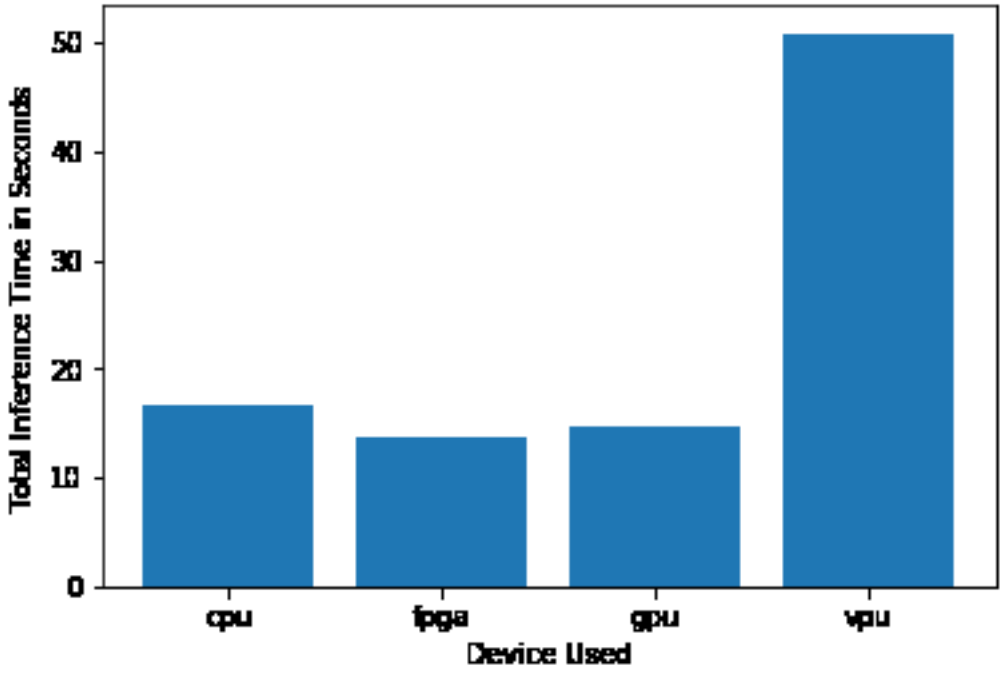
Maximum number of people in the queue	*15*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

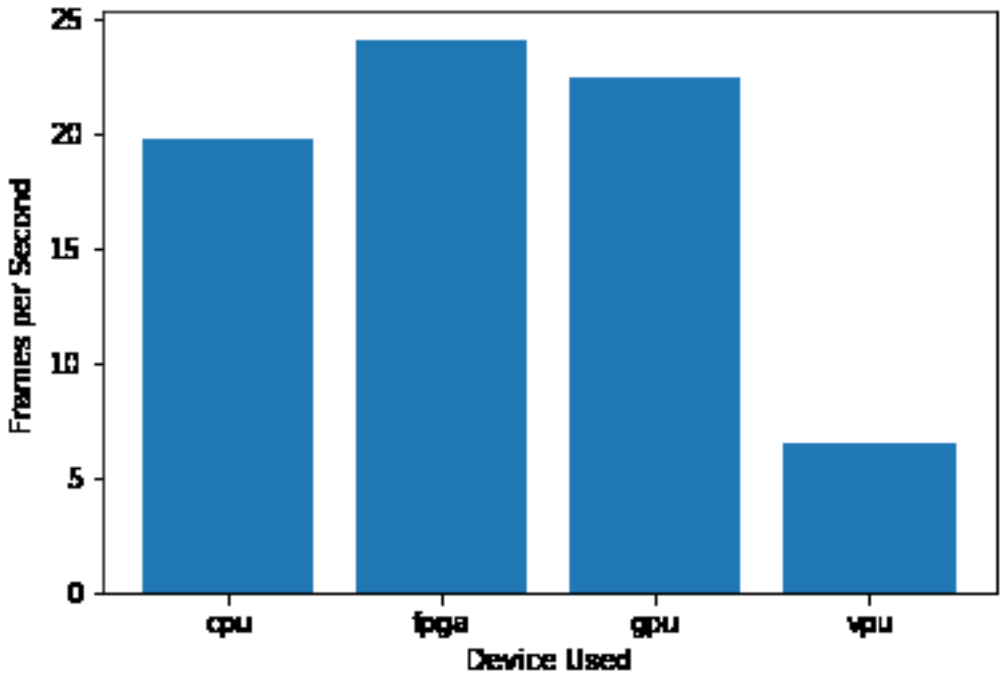
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
In this case, the client has a small budget, $300, for extra hardware purchases and already has AIO PCs. However, the PCs are already being used for video processing and they do not have enough bandwidth for additional processing. The current workload on the AIOs precludes the use of CPU and GPU. The solution must also have low power requirements. The only device that fits all of these criteria is the Intel Neural Compute Stick 2. They cost about $79, are low power, and they can easily be incorporated into the current systems by simply plugging them into the PCs. The performance of the NCS2 in this case is the worst of the bunch, however. It has the lowest FPS and the highest total inference time. The model load time is extremely low, which is good, but it's not as important as inference time. The CPU and GPU have much better performance but since they are already fully loaded they cannot be used for inference. So, the VPU will get the job done at the right price but performance may suffer. This is the tradeoff.