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

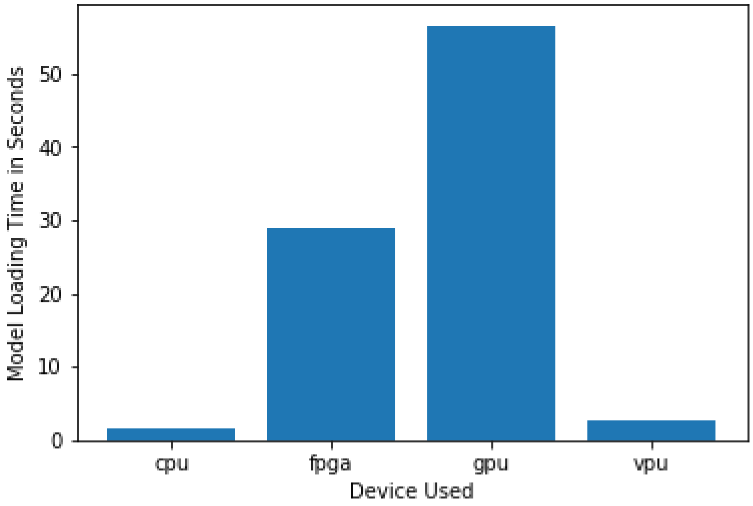
Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
Example requirement: The client requires a tiny device to be connected to their CPU—and their budget is only about $100 for each device.	*Example explanation:* VPU or NCS2 is only about 27.40 mm in size and would fit in the price range.
The company is growing and last year had a revenue of 2 million dollars	*The FPGA is an expensive device that fits the needs of a big budget.*
The company wants the solution to be reprogrammable.	*The FPGA is reprogrammable and can be repurposed for both use cases the client has mentioned.]*

Queue Monitoring Requirements

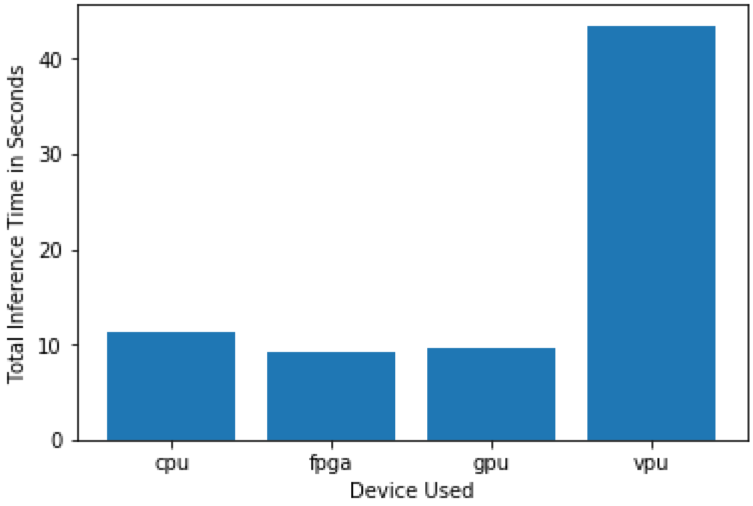
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP32 for CPU, FP16 for the rest*

Test Results

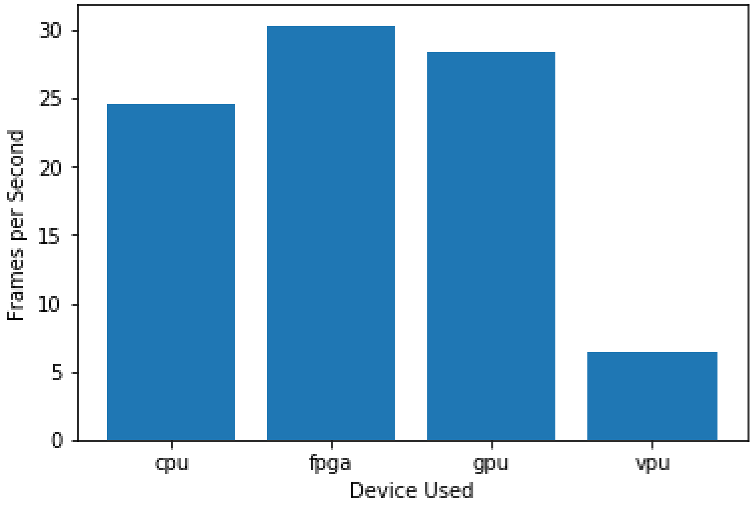
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
From the above results we can see that the FPGA shows the best performance in FPS (satisfies the required 30-35 FPS by the customer) and inference time. Also, the model load time is somewhat average. Given the big budget, the need for high FPS, the need to perform inference 5 times per second, and the fact that they need a reprogrammable solution, we conclude that the FPGA is indeed the best choice for them.

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?**
Example requirement: The client requires a tiny device to be connected to their CPU—and their budget is only about $100 for each device.	*Example explanation:* VPU or NCS2 is only about 27.40 mm in size and would fit in the price range.
Mr. Lin has limited budget.	*The ability to use the CPUs his computers already have would keep his investment budget low.*
Mr. Lin's CPUs are not utilized to their full potential at the moment.	*The resources that are not currently utilized by the existing CPUs can be used for the new tasks of inference.*

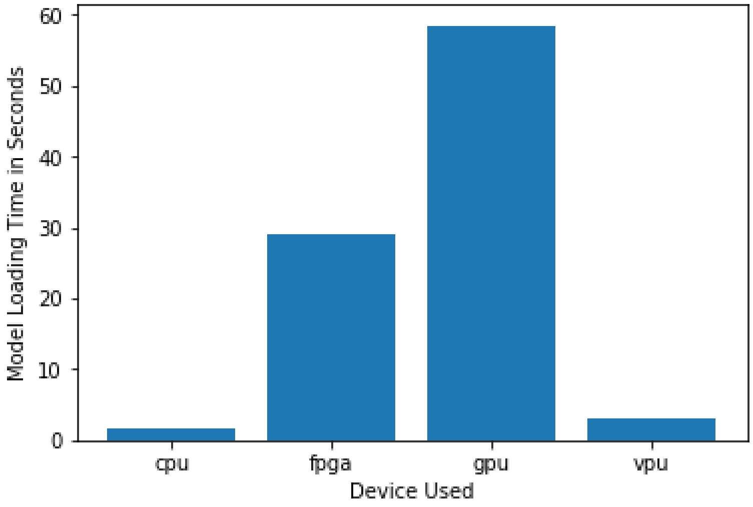
Queue Monitoring Requirements

Maximum number of people in the queue	*3*

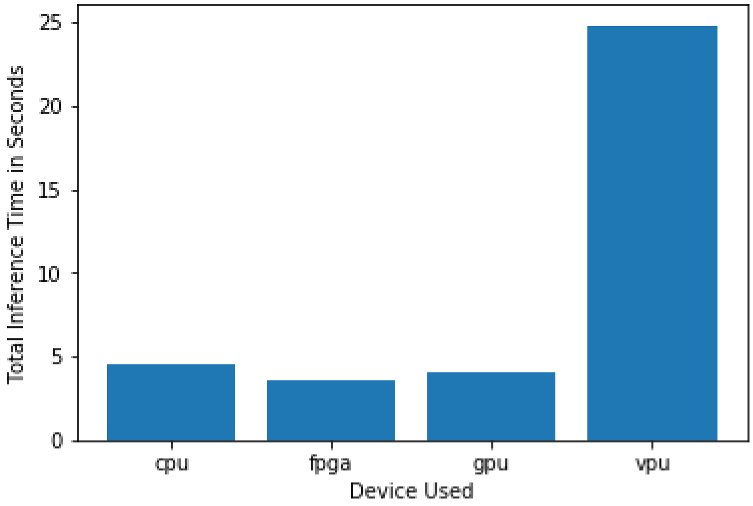
Model precision chosen (FP32, FP16, or Int8)	*FP32 for CPU, FP16 for rest*

Test Results

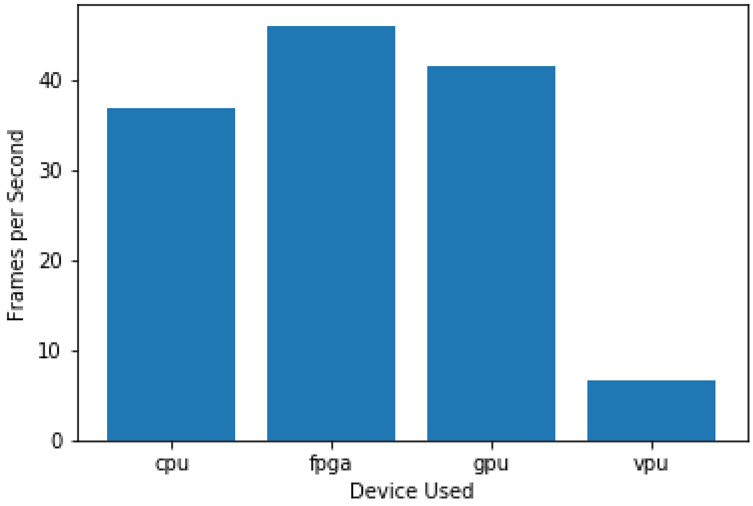
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 CPU we suggested exhibits very low inference time, comparable to gpu and fpga. The frames per second are also pretty high, close to fpga and gpu as well. Finally, the model load time for CPU is the lowest among all devices. To sum up, we can safely conclude that the CPU is the right choice for Mr. Lin.

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?**
Example requirement:	*Example explanation:*

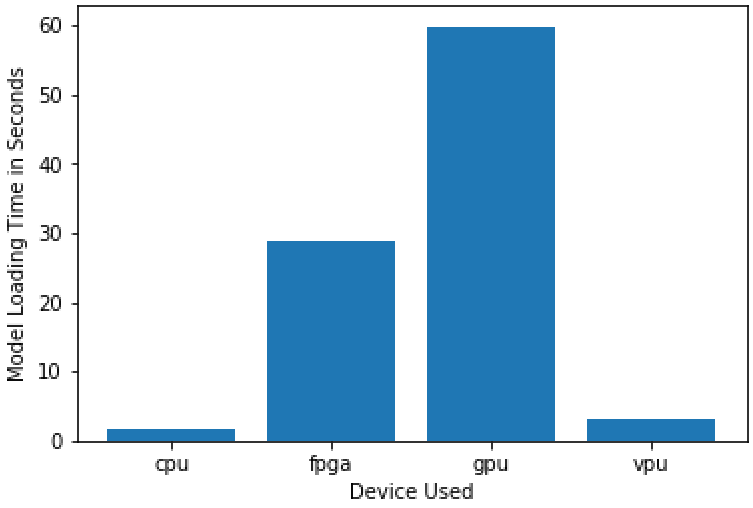
The client requires a tiny device to be connected to their CPU—and their budget is only about $100 for each device.	VPU or NCS2 is only about 27.40 mm in size and would fit in the price range.
The CPUs of the all in one PCs are already working to the maximum.	*The VPU (for example the NCS2) will be able to easily expand the capacity of the all in one PCs for inference.*
The budget is limited to 300 dollars per machine.	*The Neural Compute Stick 2 is an affordable extension to the all in one pcs for this budget.*

Queue Monitoring Requirements

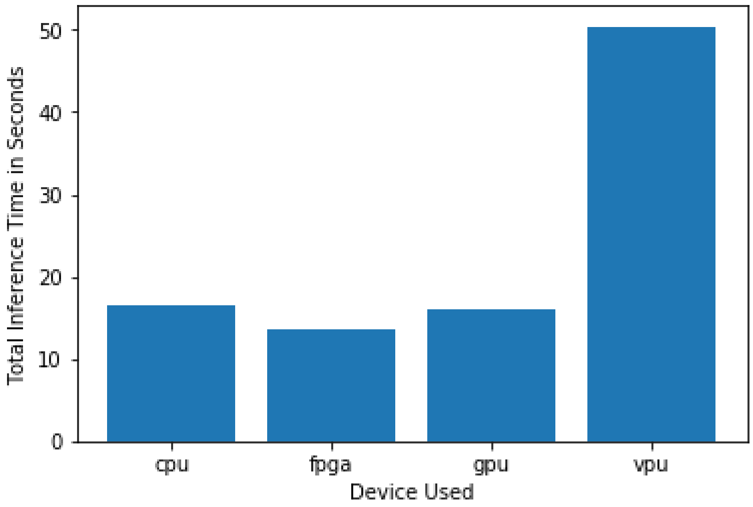
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP32 for CPU, FP16 for the rest*

Test Results

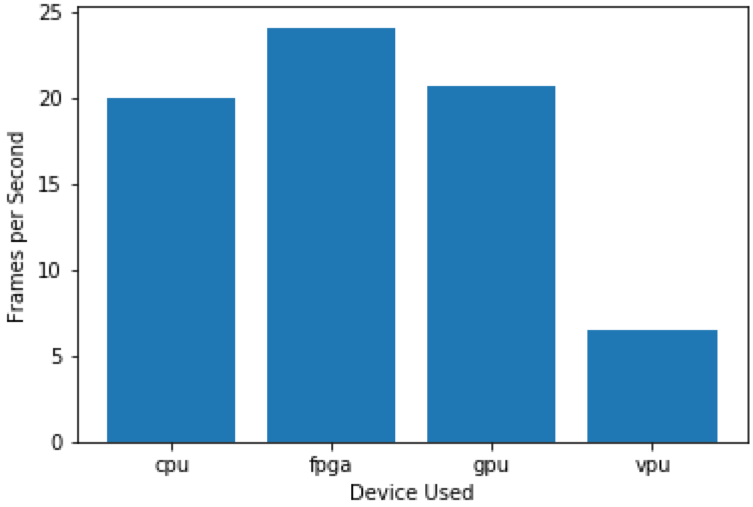
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 inference time for the VPU is higher than the other devices. Also, the FPS are lower than the other devices as well. Finally, the model load time is the lowest among all devices. We can show the customer the limitations of the device and the potential they may have for faster inference with another device. However, given their tight budget and overloaded CPUs we still believe the VPU (e.g. the Neural Compute Stick 2) is the best option for them.