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

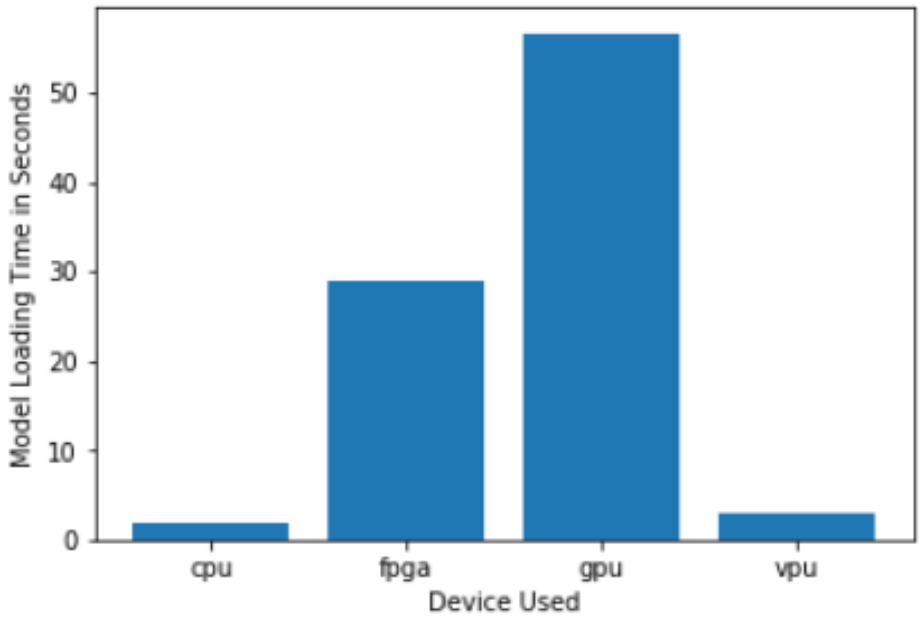
Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
The client requires a flexible so that it can be reprogrammed and optimized to quickly detect flaws in different chip designs.	FPGA can be reprogrammed to meet client needs.
The client requires very fast 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

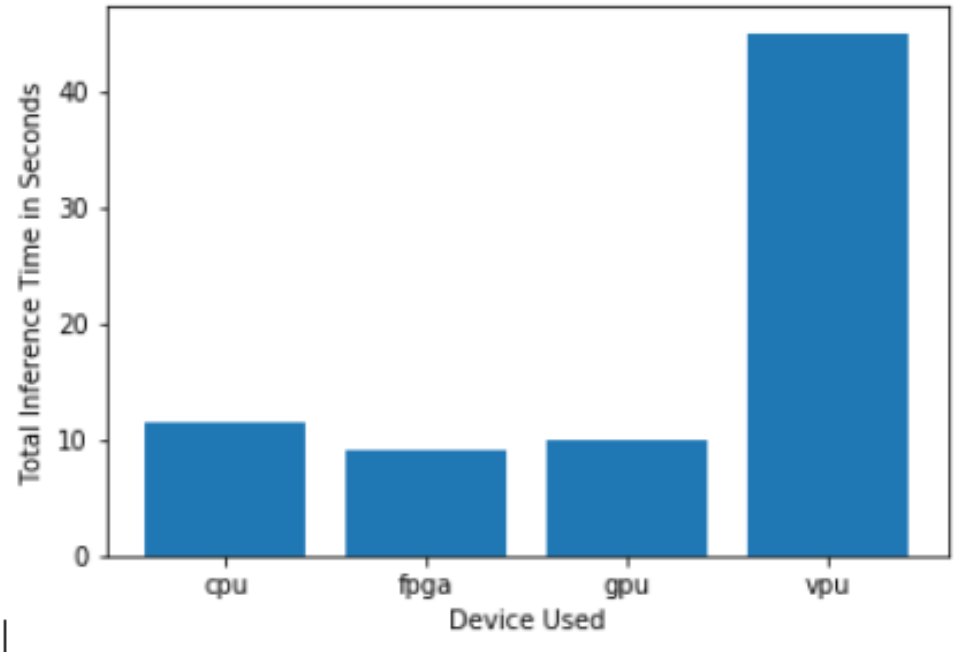
Maximum number of people in the queue	2
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

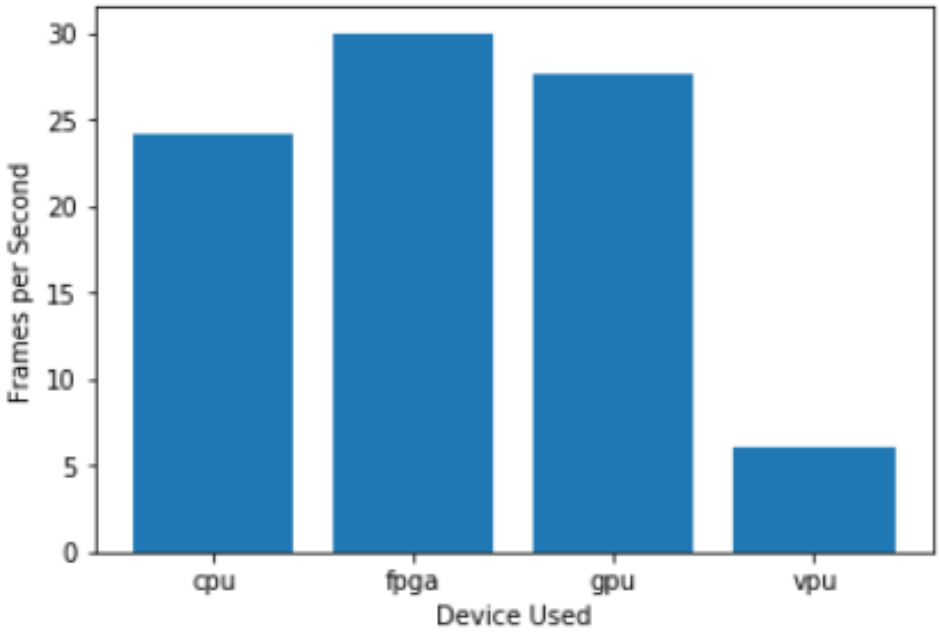
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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)
pre-existing CPUs

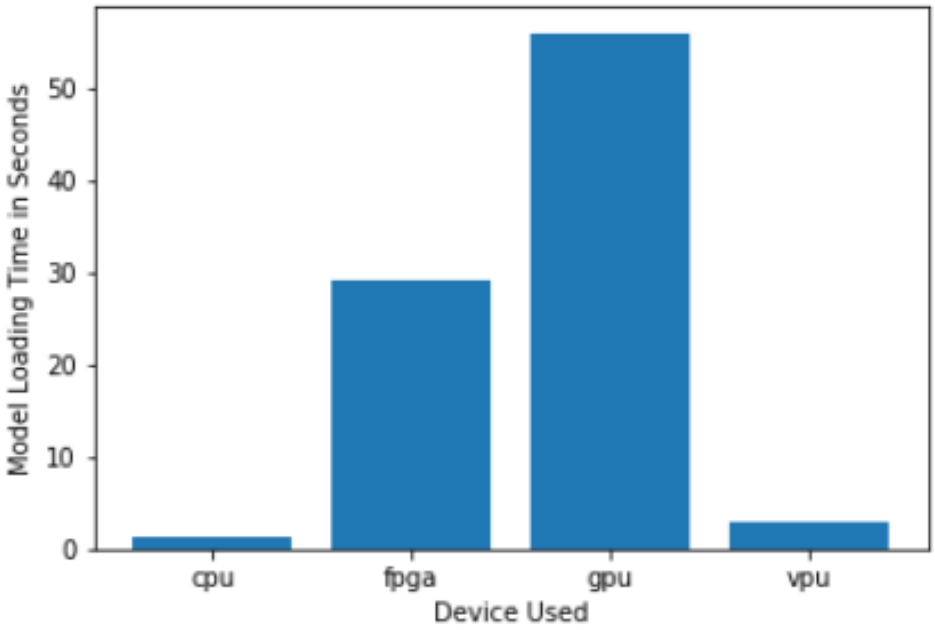
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client has low budget.	pre-existing CPUs would be perfect fit for the client as no extra cost is required.
The client wants to keep power consumption where it is	CPUs are already consuming power so the power consumption would not increase.

Queue Monitoring Requirements

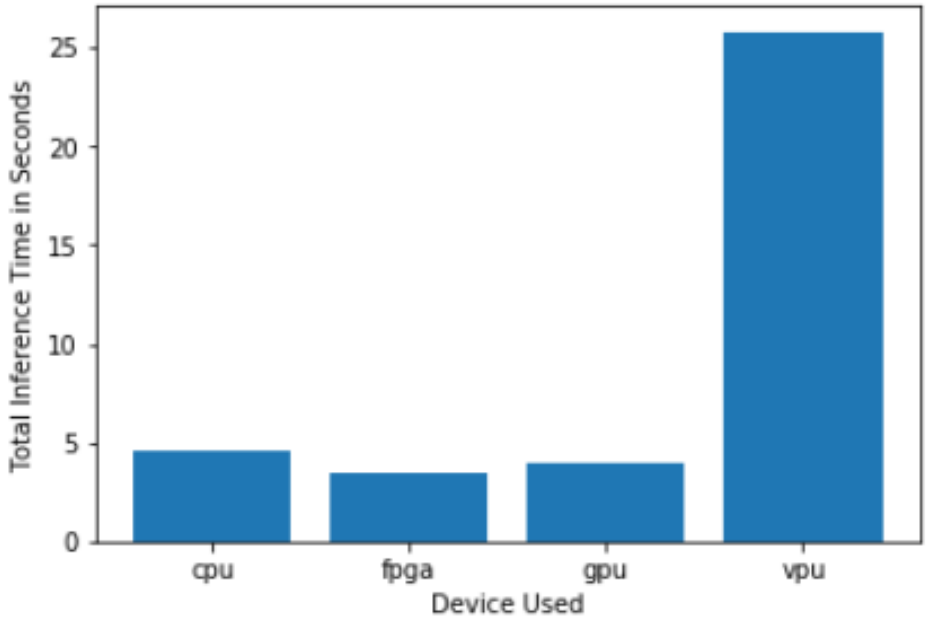
Maximum number of people in the queue	5
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

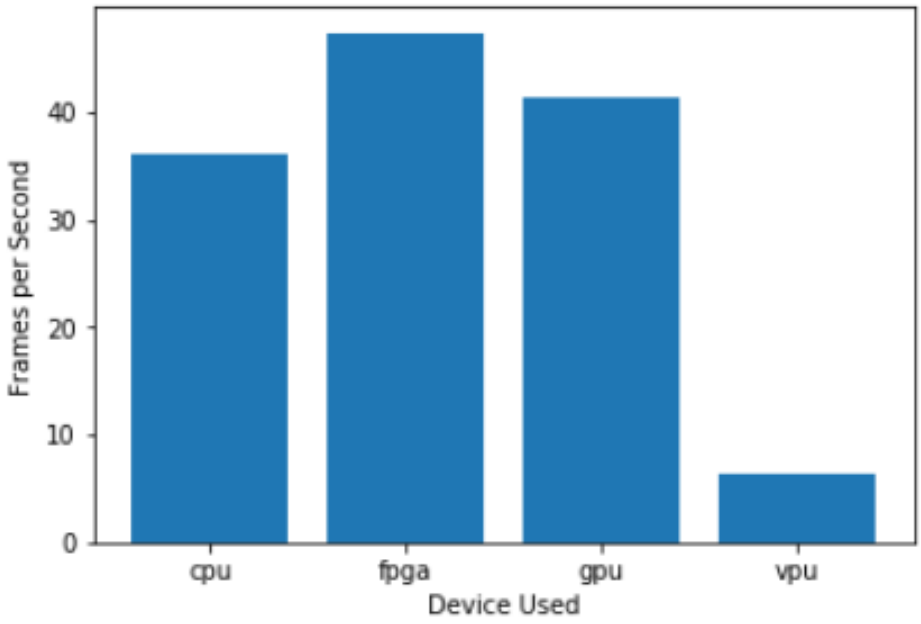
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
pre-existing CPUs Using the pre-existing CPUs would maintain the current power consumption since the client wants to keep power consumption where it is. Adding an accelerator, like an FPGA or VPU, does not seem to be necessary; the pre-existing CPUs can likely handle the inference task since it is not especially demanding.

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

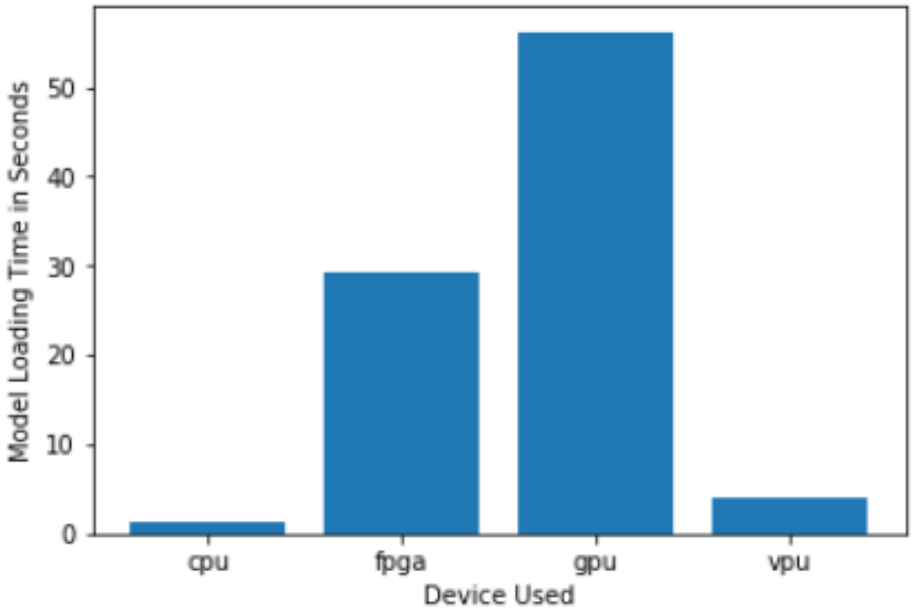
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client would like to save as much as possible both on hardware and future power requirements ($300 for each device).	VPUs are cheap in price and give excellent performance.
The client would like to run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

Queue Monitoring Requirements

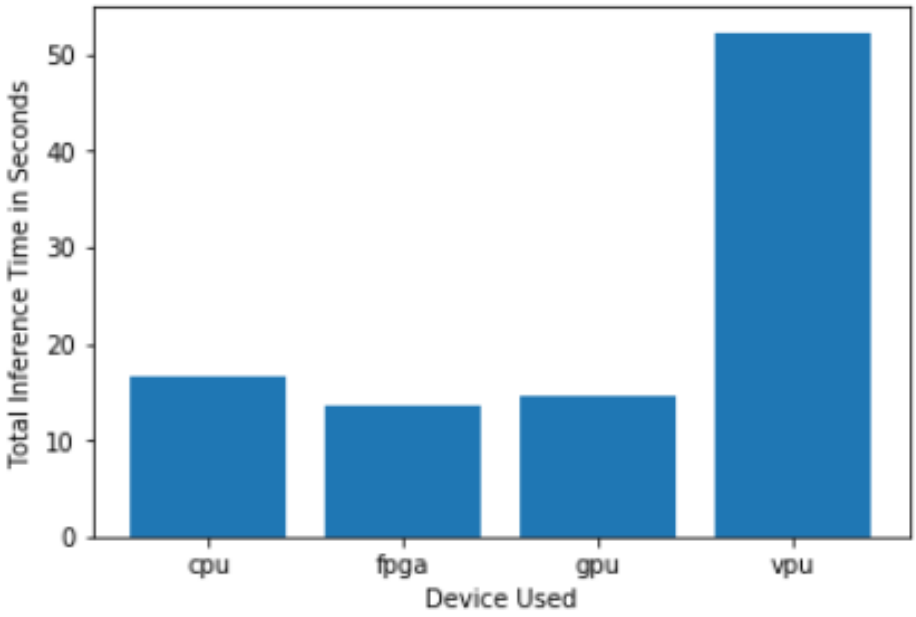
Maximum number of people in the queue	20
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

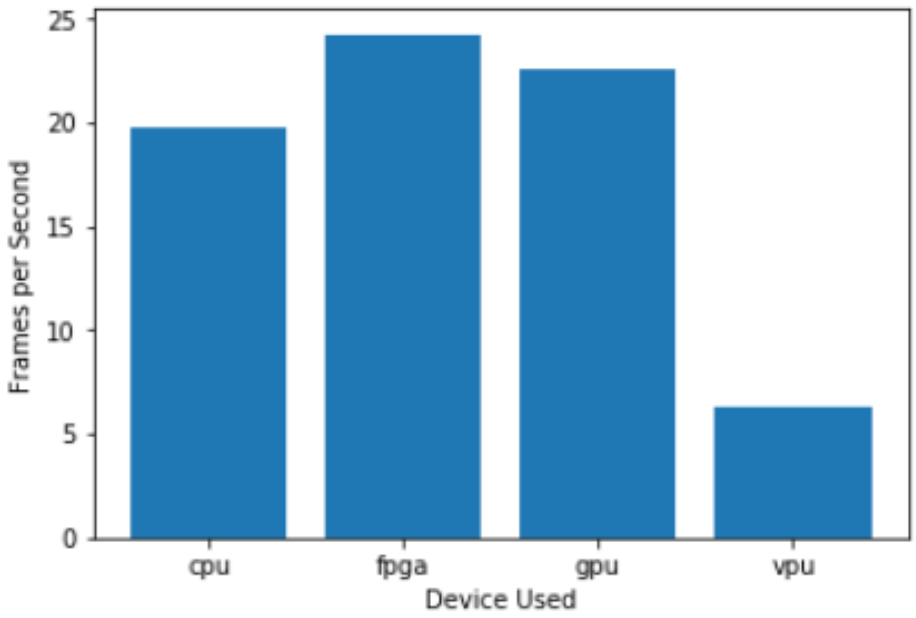
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
VPU The NCS2 is a USB plug and play device and could be easily added to an All-In-One PC since the pre-existing systems are closed, so installing new internal hardware like CPU or GPU is not an option. The client needs a high-performance device so that it can do real-time inference and quickly direct the passenger, but the pre-existing CPUs are already near capacity and do not have additional processing power to run inference (Using the current CPU is not an option, but adding VPU would provide the needed acceleration.)