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

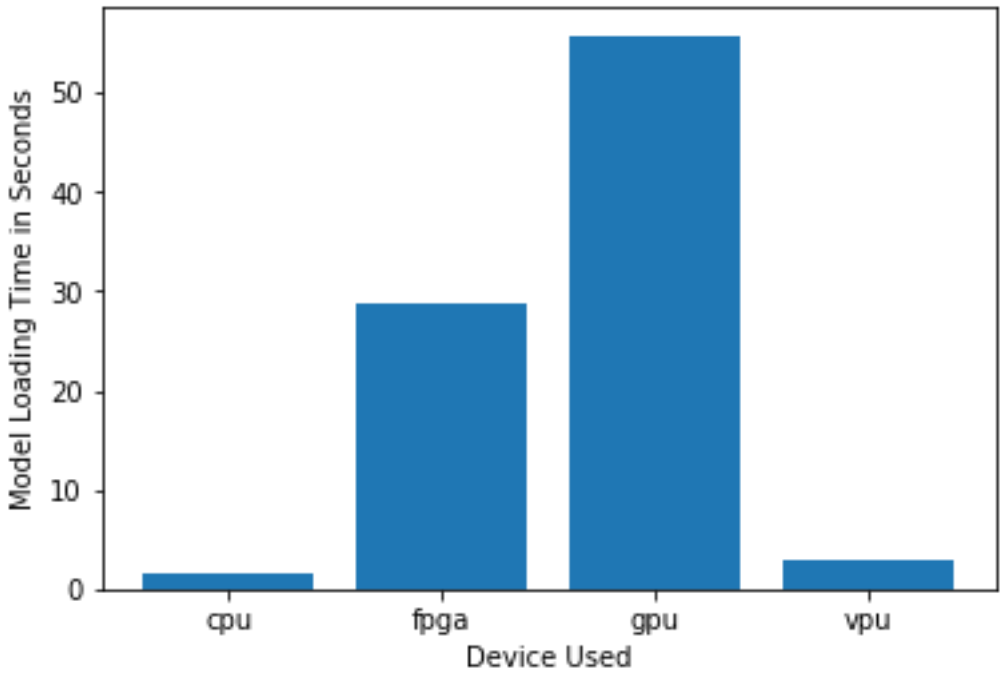
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
To detect chip flaws without slowing down the packaging process. For this purpose the system needs to be able to run inference on the video stream very quickly.	The FPGA can execute neural networks with high performance. The FPGA can execute neural networks with very little latency.
Because there are multiple chip designs—and new designs are created regularly—the system needs to be flexible so that it can be reprogrammed and optimized to quickly detect flaws in different chip designs.	FPGAs are field-programmable; they can be reprogrammed to adapt to different custom networks.
Naomi Semiconductors has plenty of revenue to install a quality system, this is still a significant investment and they would ideally like it to last for at least 5-10 years.	FPGAs have a long lifespan. FPGAS manufactured by some renowned companies like Intel have a guaranteed availability of 10 years, from start of production.
Workers alternate shifts to keep the floor running 24 hours a day so that packaging continues nonstop. Slow-down in production happens during the shift transition periods.	FPGAs are designed to have 100% on-time performance. They can be continuously running 24 hours a day, 7 days a week, 365 days a year.

Queue Monitoring Requirements

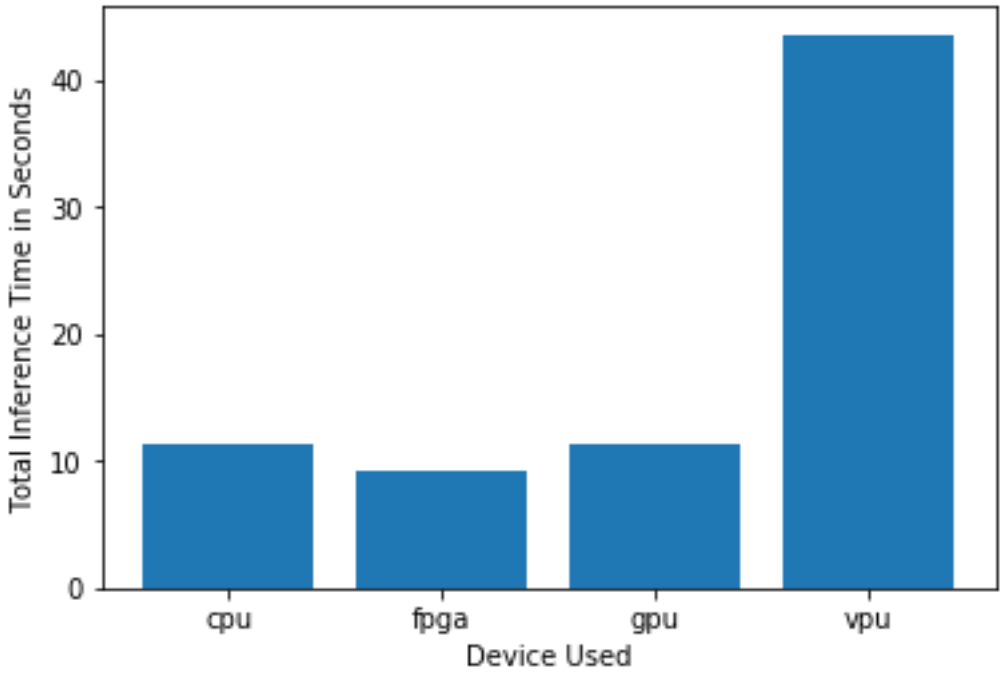
Maximum number of people in the queue	4 People.
Model precision chosen (FP32, FP16, or Int8)	FP16 (Compatible with specifications of FPGAs).

Test Results

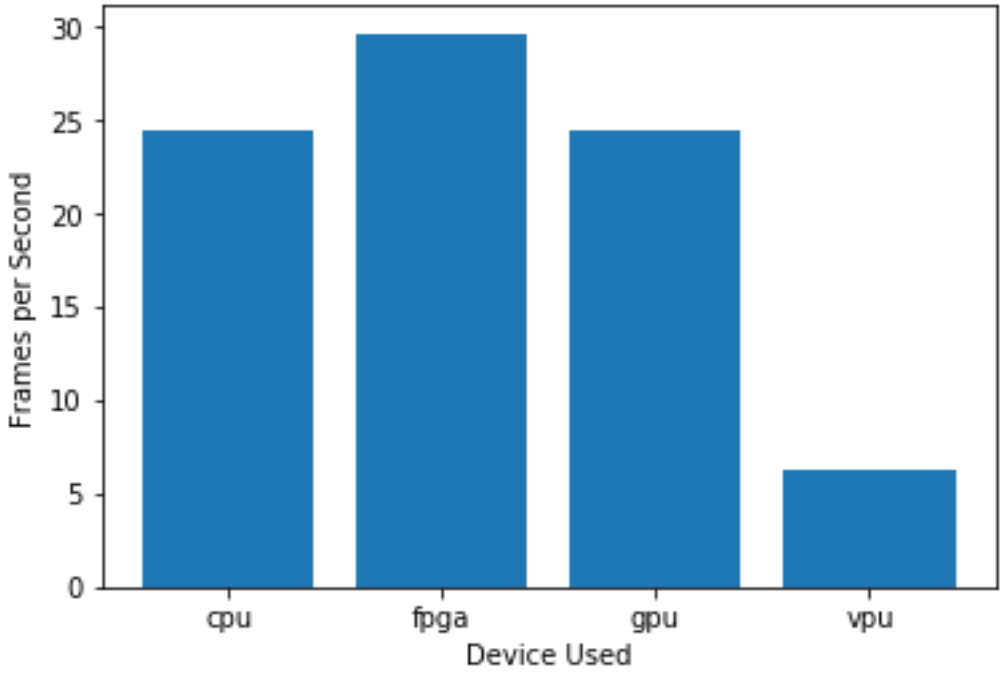
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 the first graph, the model loading time of the FPGA is located more or less in the middle of the model loading time of the GPU and the model loading time of the other devices (CPU and VPU). The model loading time is not a very critical factor in this scenario and FPGA performance is superior in inference time and frame per second processing. The total inference time of the FPGA is the lowest, which means that the device is the fastest of the four devices. This confirms that the FPGA is an excellent device for this scenario because high speed is a customer requirement. The FPGA has the capacity to process more frames per second than the other devices (approximately 30 FPS), and this processing capacity is compatible with the video stream recorded by the cameras. **The test results validate that FPGA is the best choice for this scenario.**

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?
The majority of the checkout counters at the store have a computer with an Intel i7 core processor. And these processors are being underused.	The processors of the modern computers that the client has in the checkout counters are in sufficient capacity to perform the required task because the

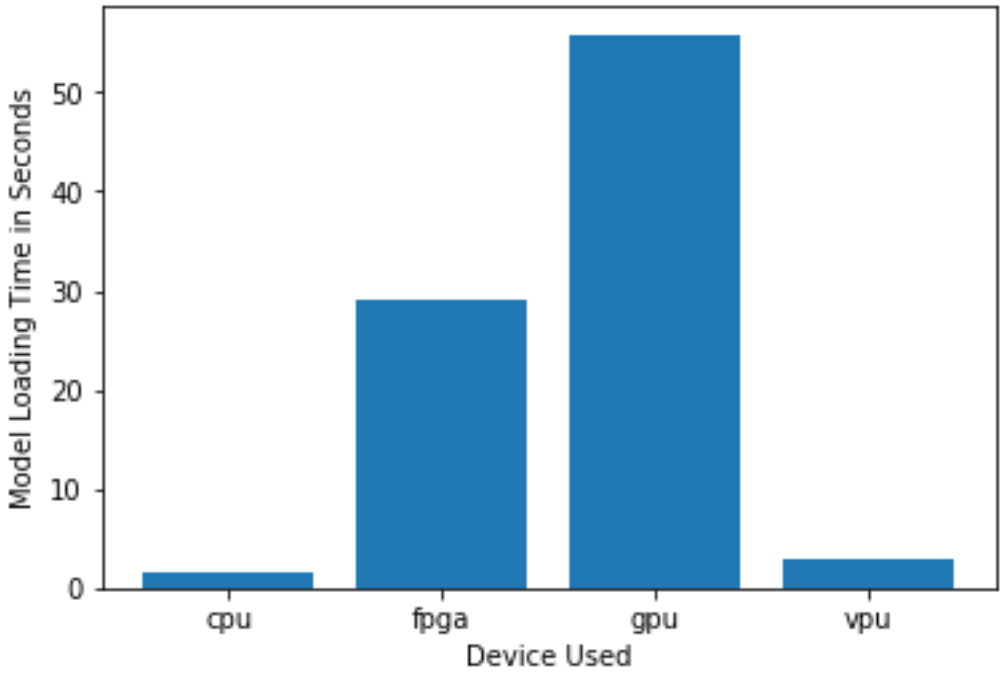
Average wait time spent at the checkout counters: Weekdays: 230 seconds. Weekends: 350 – 400 seconds.	latency and the speed of the inference do not require a computer with superior processing capabilities.
The client does not have much money to invest in additional hardware.	There is no need to purchase new hardware because the Intel i7 processors are capable of performing the required tasks.
The client would like to save as much as possible on his electric bill.	Since the customer already pays for electricity service for the equipment he already has, and the characteristics of the task to be executed by the processors do not require high performance, using the existing CPUs, electricity service does not have to be raised considerably.

Queue Monitoring Requirements

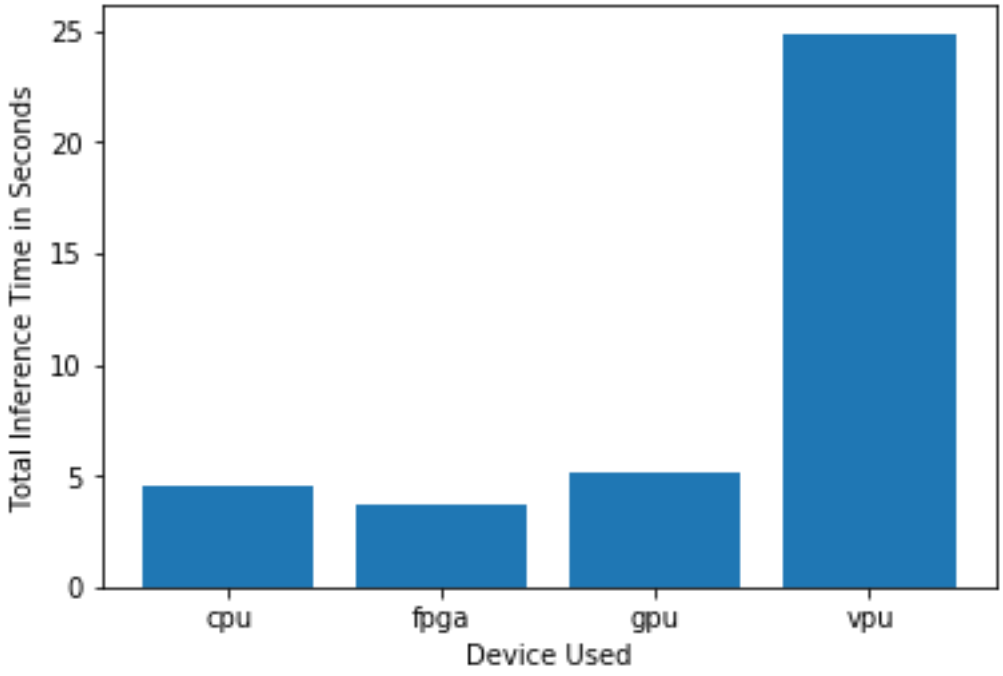
Maximum number of people in the queue	5 People.
Model precision chosen (FP32, FP16, or Int8)	FP32

Test Results

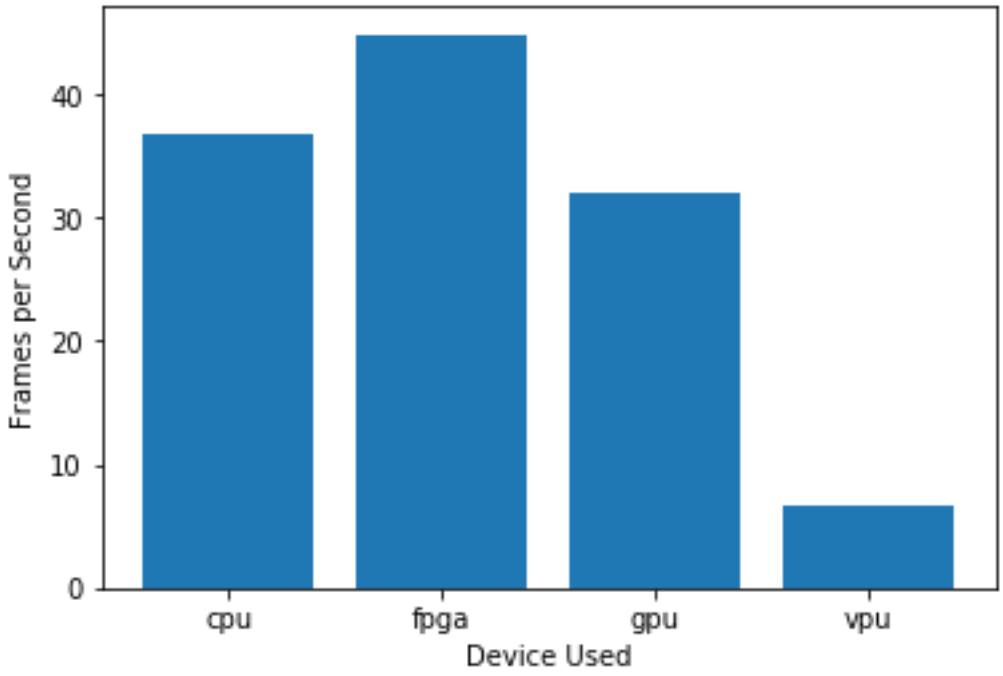
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
In the first graph it is possible to see that the model loading time of the CPU is the lowest among the 4 devices and this parameter is a good criterion for choosing the CPU, taking into account that the client does not want to invest a lot of money in additional hardware. The total inference time of the CPU and the FPGA are the lowest, which means these devices are the fastest to perform inference. Considering that high inference speed is not a critical factor in this project and that it is important for the customer to save money, there is no reason to buy a FPGA and the CPU is still the best option. the FPGA can process more frames per second but the difference with the CPU is not too much considering the characteristics of the project, so it is concluded that according to this performance parameter, the CPU still represents the best option. **The analysis of the test results and the evaluation of the requirements of the customer, lead to the conclusion that the best choice for this scenario is the CPU.**

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

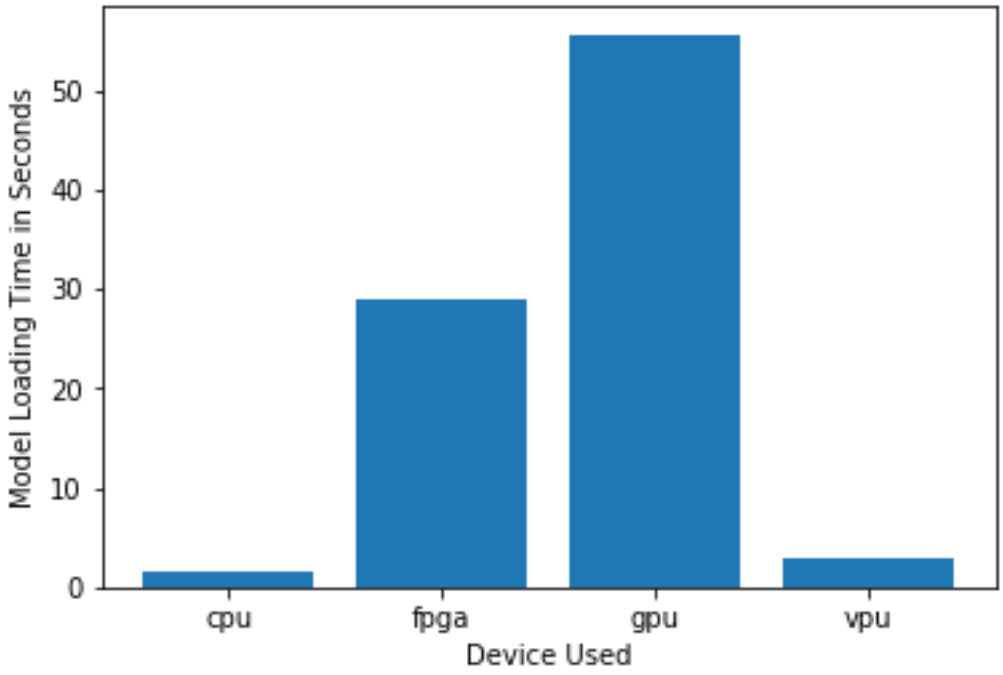
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The CPUs in the available machines are used to process and analyze footage CCTV and therefore there is not enough processing capacity available to make an inference.	A VPU accelerates the performance of the pre-existing CPU. The CPU doesn't need to be a powerful one, since it will not actually be doing any calculations.
The customer's budget is small (about $300 per machine).	The price of an Intel NCS2 is about $75.65 https://www.amazon.com/Intel-Neural-Compute-Stick-2/dp/B07KT6361R
The customer would like to save as much as possible both on hardware and future power requirements.	VPUs are small, low-cost, low-power devices that can dramatically improve the performance of a system without the need to upgrade the other hardware.

Queue Monitoring Requirements

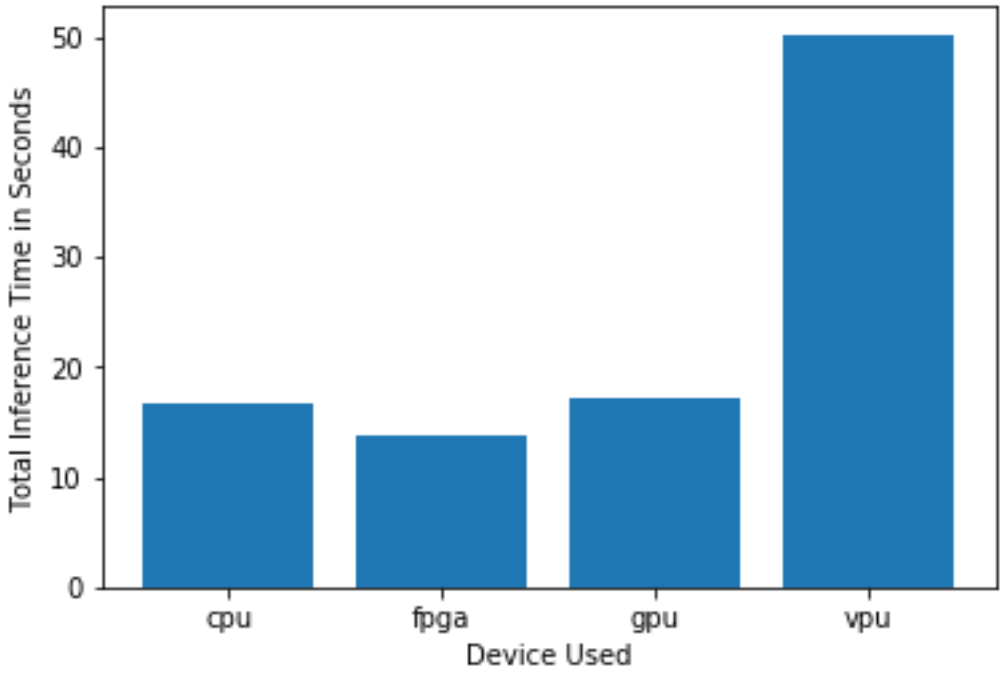
Maximum number of people in the queue	15 People.
Model precision chosen (FP32, FP16, or Int8)	FP16 Taking into consideration that the Intel NCS2 only supports FP16.

Test Results

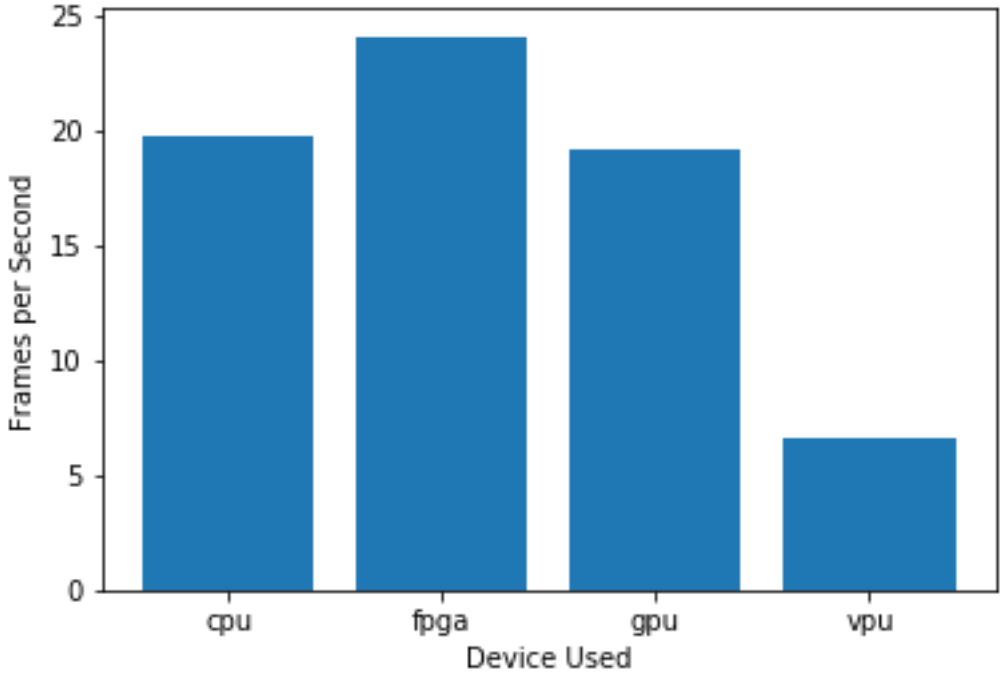
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 the first graph it is possible to see that the model loading time of the VPU is very short and this parameter is a good criterion for choosing the VPU, taking into account that the client has a narrow budget. The VPU takes a lot of time for inference (almost three times the time of the other devices), The time for inference is less than one minute and taking into consideration that in office hours there is a train every 2 minutes, the implementation of a system with a VPU, It could still work in an acceptable way, considering that the client does not have too much budget to invest in hardware. The VPU can process less frames per second than the other devices but frames processing is not a critical requirement for this project. **The analysis of the test results and the evaluation of the requirements of the customer, lead to the conclusion that the best choice for this scenario is the VPU.**