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

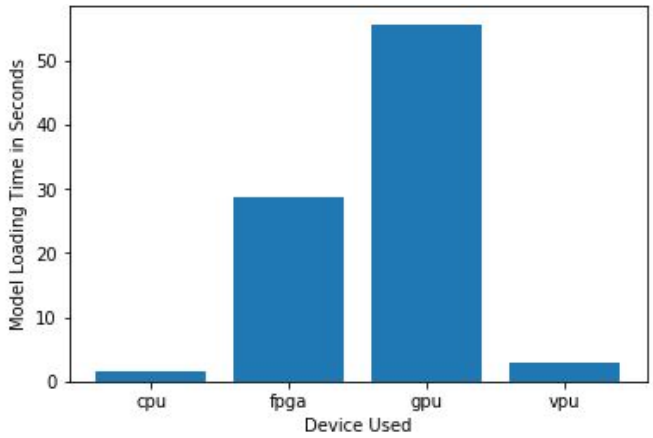
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The system would need to be able to run inference on the video stream very quickly	Once programmed with a suitable bitstream, FPGAs can execute neural networks with high performance and very little latency. The high performance comes from the ability to run many sections of the FPGA in parallel
The system would also need to be flexible so that it can be reprogrammed and optimized to quickly detect flaws in different chip designs.	Field-Programmable Gate Arrays (FPGAs) are chips designed with maximum flexibility, so that they can be reprogrammed as needed in the field
Naomi Semiconductors has plenty of revenue to install a quality system	FPGAs are expensive but the client will be able to afford them given that their revenue is plenty
They would ideally like it to last for at least 5-10 years.	FPGAs that use devices from Intel's Internet of Things Group have a guaranteed availability of 10 years, from start of production.

Queue Monitoring Requirements

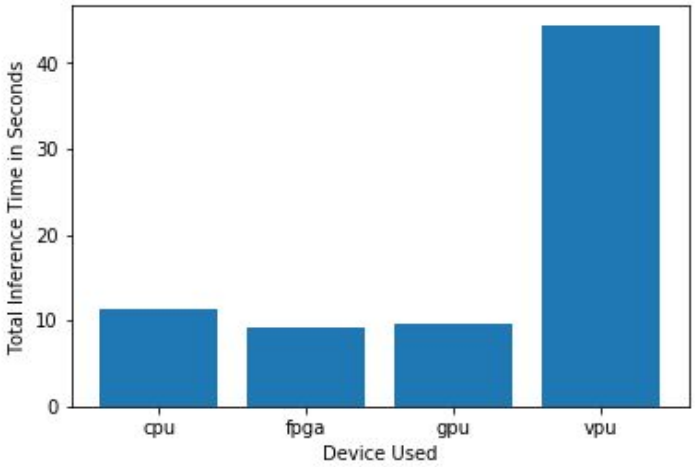
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

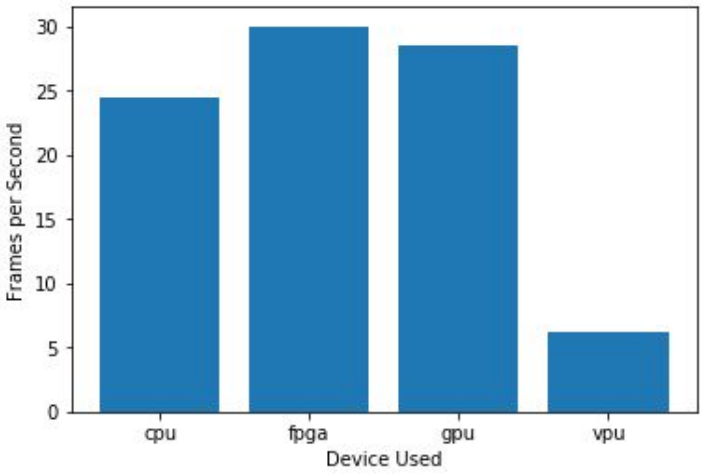
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
As the client wants to install a re-programmable system with quicker inference rates even if expensive, FPGAs will be the go-to choice of hardware. This is further corroborated by the graphs attached above. We can see that FPGAs provided the highest frames per second processing and lowest inference time. The client currently has cameras that records at 30-35 FPS and the above graph shows that FPGAs do have the same processing rate.

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)
IGPU

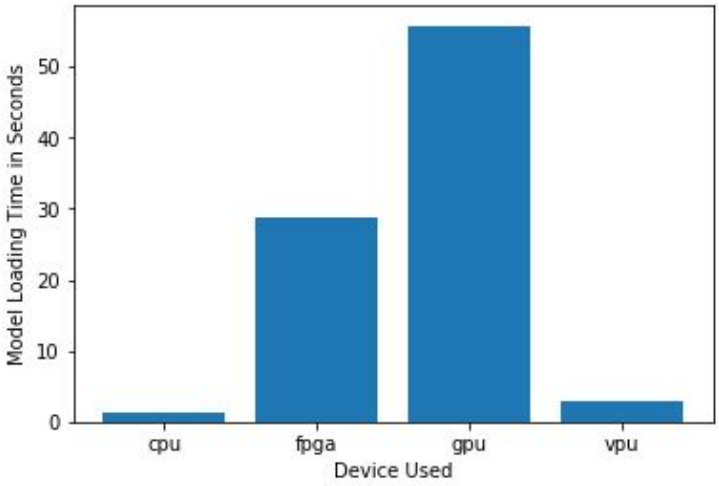
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client would like to save as much as possible on his electric bill.	Configurable Power Consumption - unused sections in a GPU can be powered down to reduce power consumption.
The client does not have much money to invest in additional hardware.	An integrated GPU (IGPU) is a GPU that is located on a processor alongside the CPU cores and shares memory with them. Most of the store's checkout counters already have a modern computer, each of which has an Intel i7 core processor. So there is no need to purchase additional hardware.

Queue Monitoring Requirements

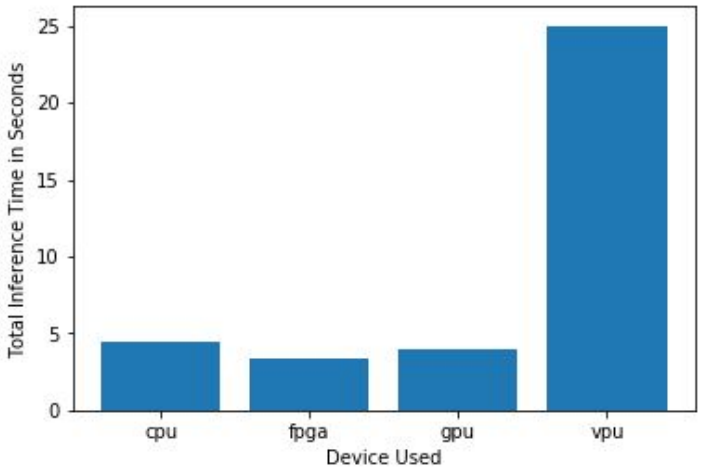
Maximum number of people in the queue	*2-5*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

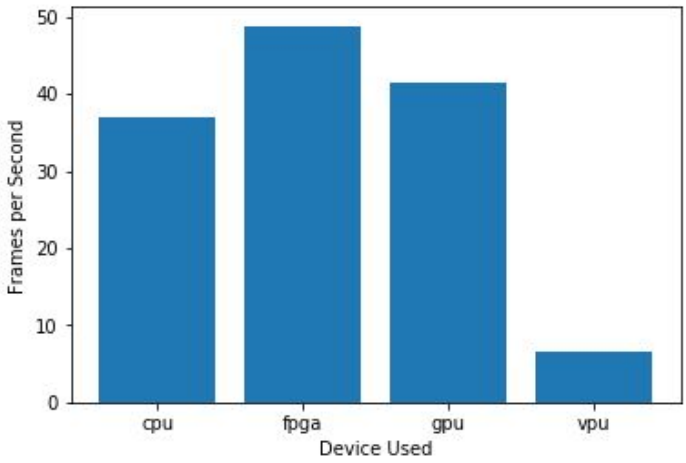
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
Although FPGAs are the quickest w.r.to inference times (which may help with faster checkouts at the counter thereby increasing the client's profits significantly), the client does not have the budget to afford an FPGA. Also, the client already has some i7 Intel processor borne computers at checkout which comes with an IGPU anyway. So, considering the fact that the client's budget is low and the power consumption should be as minimum as possible, IGPUs are the best choice of hardware.

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?**

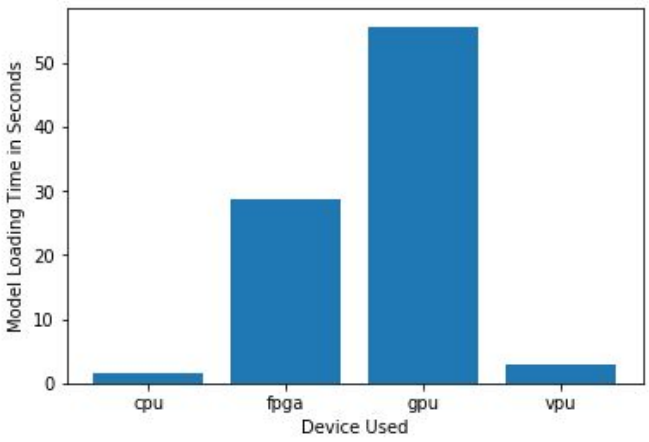
The client budget allows for a maximum of $300 per machine	VPU or NCS2 is an inexpensive option costing only about $70 to $100 and would fit in the price range.
The client would like to save as much as possible both on hardware and future power requirements	VPUs are small, low-cost, low-power devices that can dramatically improve the performance of a system without the need to upgrade the other hardware.
The CPUs in these machines are currently being used to process and view CCTV footage for security purposes and no significant additional processing power is available to run inference	The Neural Compute Stick 2 (NCS2) is a USB3.1 plug and play removable VPU for AI inferencing. The Myriad X has a very low power consumption of only 1-2 watts.

Queue Monitoring Requirements

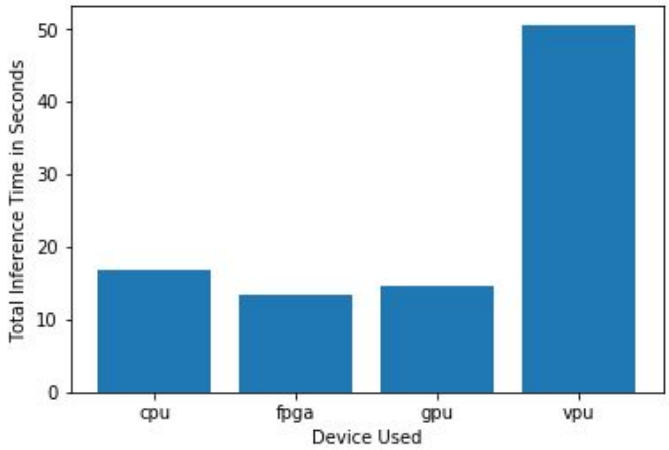
Maximum number of people in the queue	*7-15*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

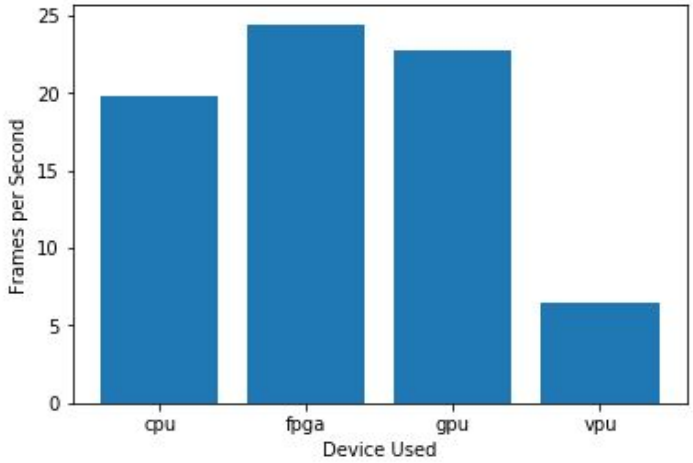
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
FPGAs can provide quickest inference times but it needs more investment. So although the client wants to quickly direct the crowd, there is a constraint on cost and power availability. This indicates that VPUs are the best option given that they consume the lowest power of all available choices at the lowest cost. However this does come with higher inference time and lowest FPS - a tradeoff the client must be willing to accept.