Hardware Proposal

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

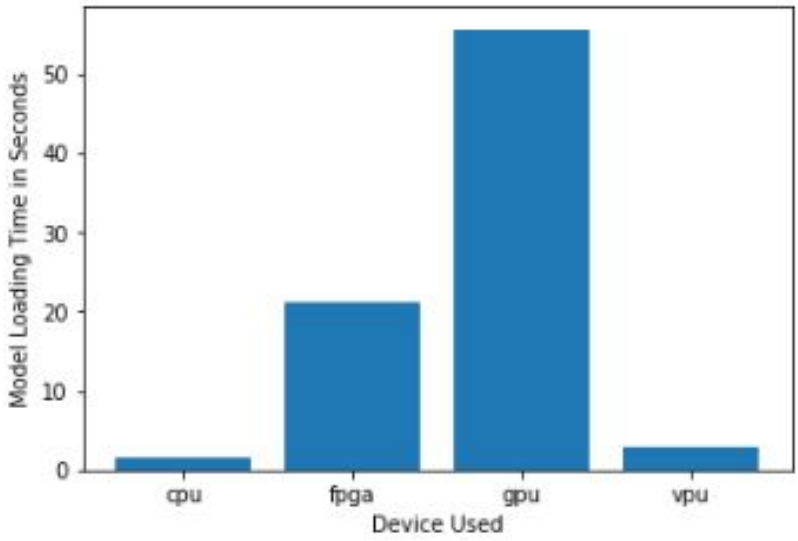
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
need to be flexible so that it can be reprogrammed	FPGAs are flexible and reprogrammable
Workers alternate shifts to keep the floor running 24 hours a day so that packaging continues nonstop	FPGAs can run continuously 24 hours a day, 7 days a week, 365 days a year
would ideally like it to last for at least 5-10 years	FPGAs have a long lifespan

Queue Monitoring Requirements

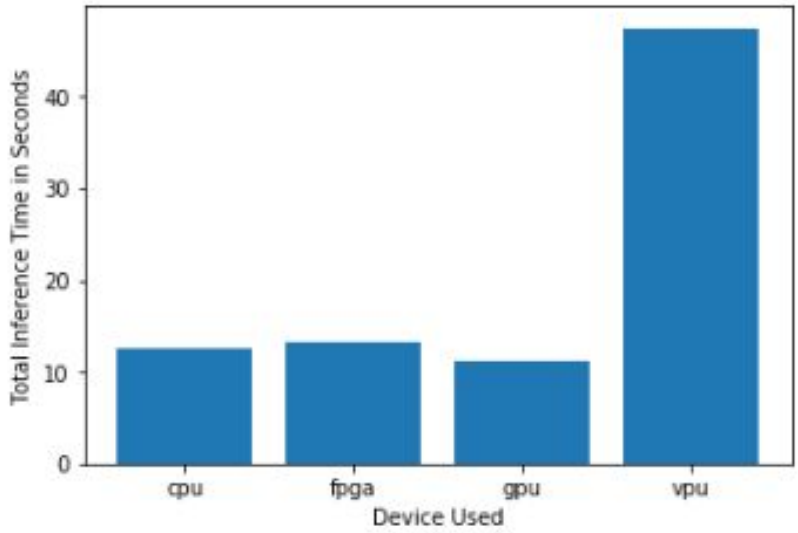
Maximum number of people in the queue	Unknown (about 10 from the video)
Model precision chosen (FP32, FP16, or Int8)	FP16 is preferable for FPGAs

Test Results

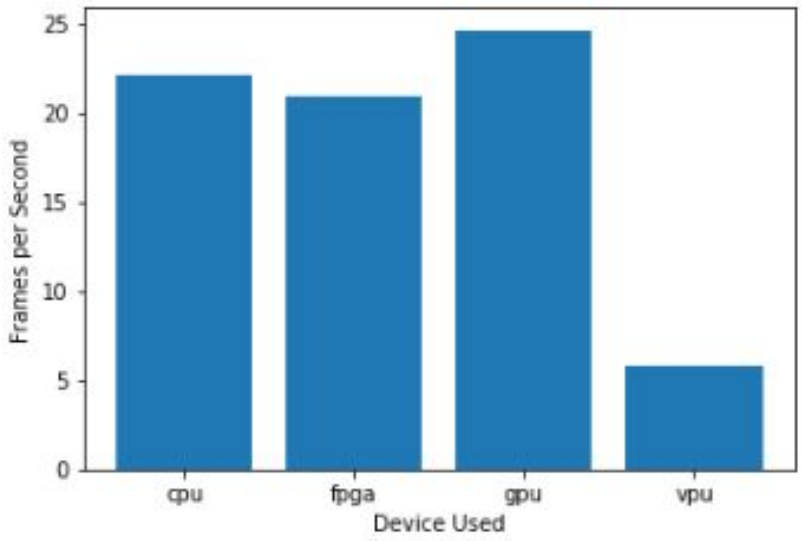
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
The GPU has the biggest model loading time, followed by FPGA, VPU and CPU respectively. This doesn't change through all 3 scenarios. For the Manufacturing scenario, the VPU has the lowest frames per second (biggest inference time), followed by FPGA, CPU and GPU respectively. Since the client's requirements are that (1) it needs to be flexible so that it can be reprogrammed, (2) it needs to run 24h a day nonstop and (3) it needs to last for at least 5-10 years, the FPGA is the optimal hardware 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

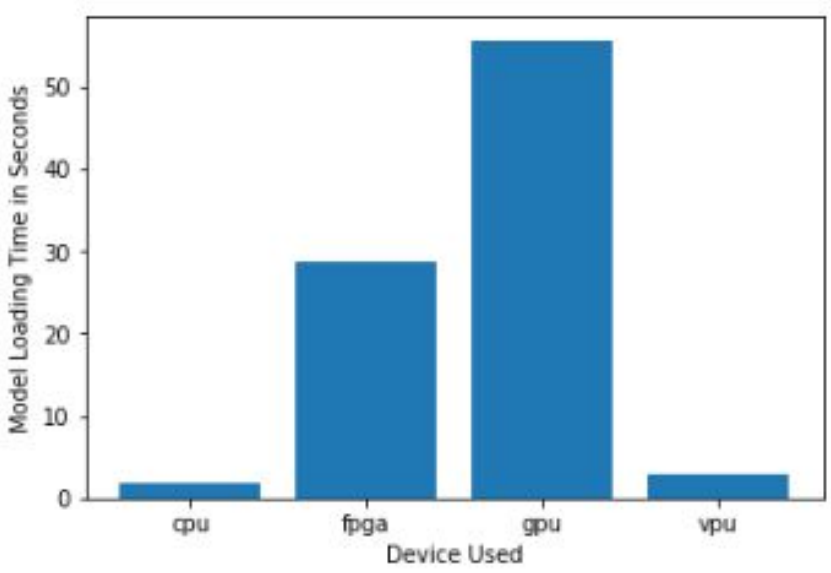
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
not much money to invest in additional hardware	Most of the store's checkout counters already have a modern computer, each of which has an Intel i7 core processor, so no need to buy new hardware
save as much as possible on his electric bill	CPUs require less power consumption than GPUs

Queue Monitoring Requirements

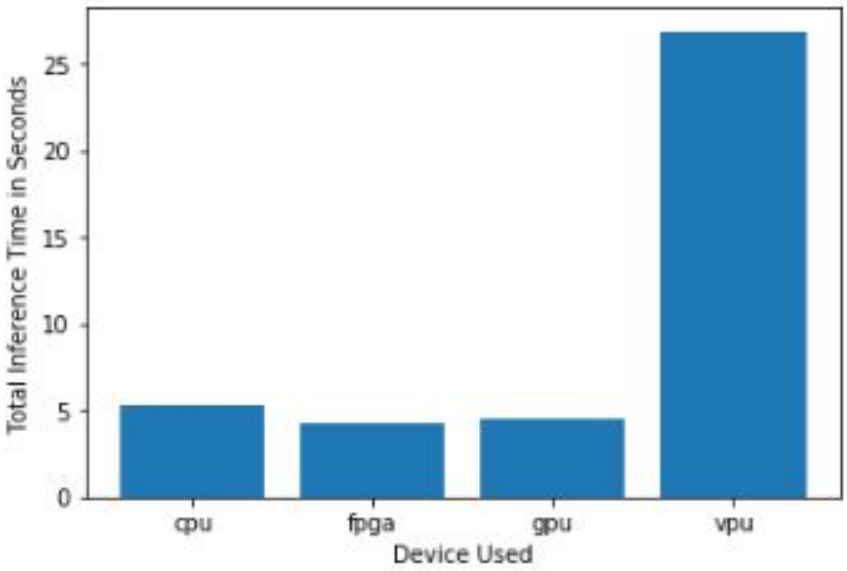
Maximum number of people in the queue	5
Model precision chosen (FP32, FP16, or Int8)	FP32 is preferred for CPUs

Test Results

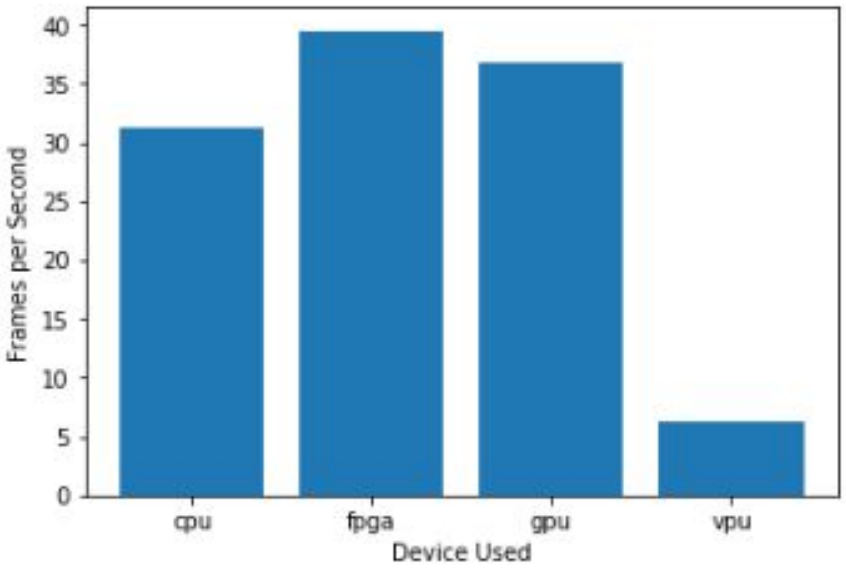
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 GPU has the biggest model loading time, followed by FPGA, VPU and CPU respectively. This doesn't change through all 3 scenarios. For the Retail scenario, the VPU has the lowest frames per second (biggest inference time), followed by CPU, GPU and FPGA respectively. Since the client's requirements are that (1) there is not much money to invest in additional hardware and they already have modern computers with

i7 core processors and (2) they want to save as much as possible on his electric bill, the CPU is the optimal hardware for this scenario.

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

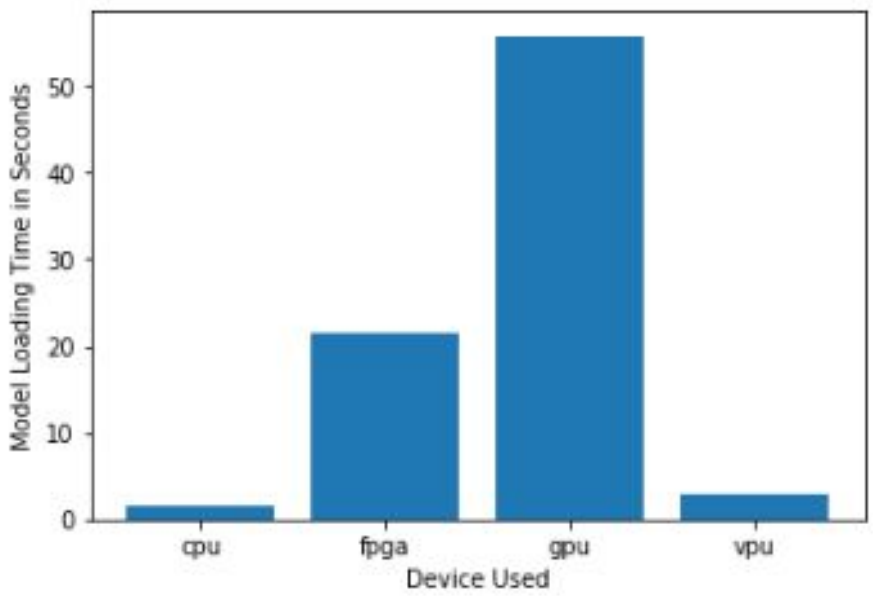
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
no significant additional processing power is available to run inference	VPU or NCS2 has a power consumption of only ~1W
budget allows for a maximum of $300 per machine	VPU or NCS2 has a MSRP of only $69 USD (July 14, 2019)

Queue Monitoring Requirements

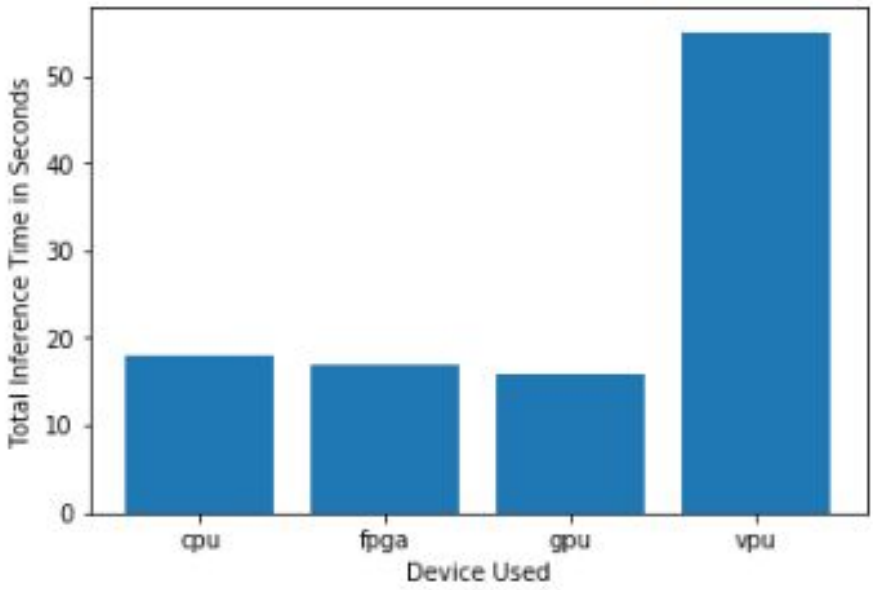
Maximum number of people in the queue	15
Model precision chosen (FP32, FP16, or Int8)	FP16 is required for VPUs

Test Results

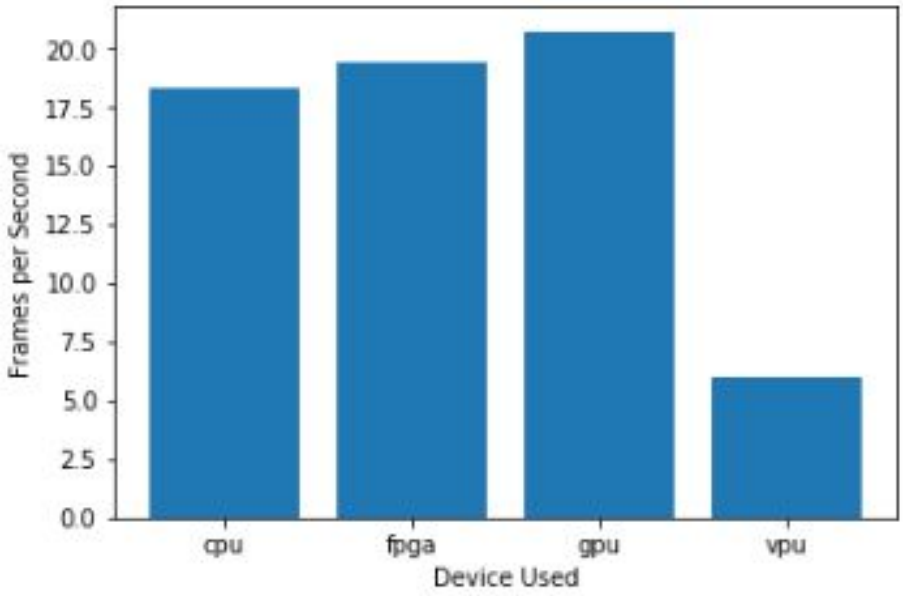
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 GPU has the biggest model loading time, followed by FPGA, VPU and CPU respectively. This doesn't change through all 3 scenarios. For the Transportation scenario, the VPU has the lowest frames per second (biggest inference time), followed by CPU, FPGA and GPU respectively. Since the clients requirements are that (1) there is no significant additional processing power available to run inference and that (2) the budget allows for a maximum of $300 per machine, the VPU is the optimal hardware for this scenario.