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

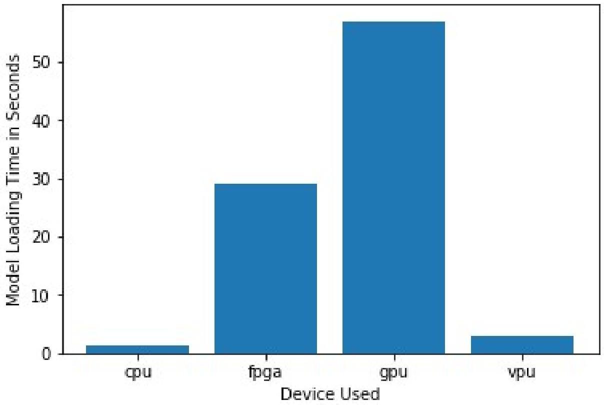
Requirement Observed **(Include at least two.)**	**How does the chosen hardware meet this requirement?**
Monitor the number of people in the factory line. The factory already has vision camera installed at every belt and Mr. Vishwas would like the image processing task to be completed **five times** per second.	The Intel Vision Accelerator Design with Intel Arria 10 FPGA can support more than 20 channels of video inputs, along with vision use cases such as facial detection and recognition.
A significant percentage of the semiconductor chips being packaged for shipping have flaws. These are not detected until the chips are used by clients. If these flaws could be detected prior to packaging, this would save money and improve the company's reputation.	FPGAs are field-programmable; they can be reprogrammed to adapt to new, evolving, and custom networks Various precision options (FP16, 11 and 9 bit) are supported—allowing developers a balance between speed and accuracy. The bitstreams being used can be updated without changing the hardware. This allows you to improve the performance of your system without replacing the FPGA. FPGAs can also support large networks, with a capacity to handle networks that have more than 2 million parameters.
Naomi Semiconductors has plenty of revenue to install a quality system and they would ideally like it to last for at least 5-10 years.	FPGAs have long life-span of upto 10 years.

Queue Monitoring Requirements

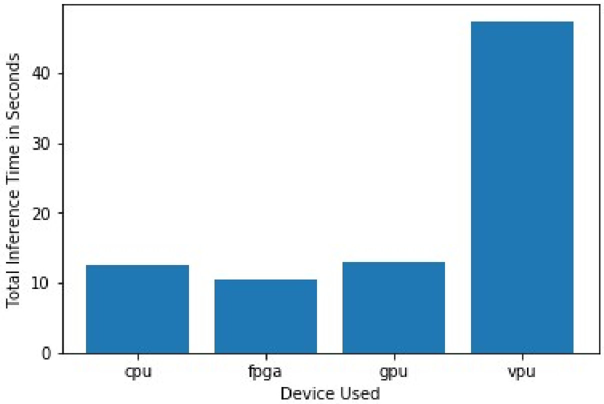
Maximum number of people in the queue	9
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

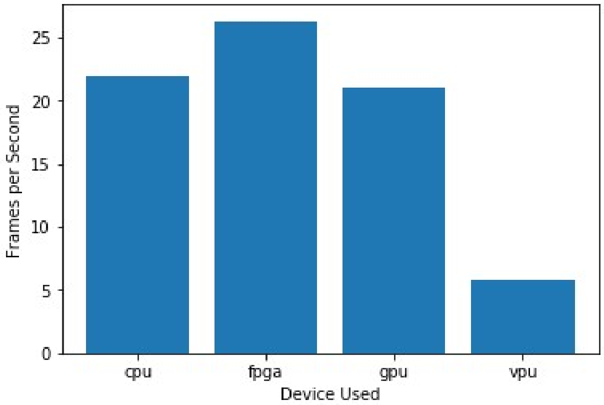
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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. *Field-Programmable Gate Arrays (FPGAs) are chips designed with maximum flexibility, so that they can be reprogrammed as needed in the field (i.e., after manufacturing and deployment).* *Because they are re-programmable, FPGA is good for prototyping and low-volume production.* *In this scenario, CPUs have also faired good in the test results. But, the client would like to invest in a product that can be re-programmed and which has a long life span.* *FPGA satisfies the client's requirements. Hence, the final recommendation is FPGA.*

Device	**Inference Time**	**FPS**	**Load Time**
CPU	12.6	21.90476	1.403547
GPU	13.1	21.0687	57.004
VPU	47.4	5.82278	3.0307
FPGA	10.5	26.2857	29.2405

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?**
Mr. Lin does not have much money to invest in additional hardware.	Most of the store's checkout counters already have a modern computer – Intel i7 core processor. So, Mr. Lin does not need to invest in a new hardware.
Mr. Lin would like to save as much as possible on his	Since we are not replacing any existing hardware or adding

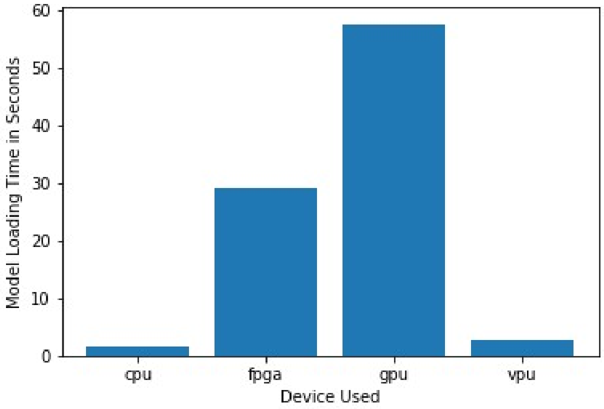
electric bill.	any additional hardware, Mr. Lin will not spend any more money on electricity. We need to redirect the people from a crowded queue to a less congested queue. This can be achieved by developing a smart queuing system.

Queue Monitoring Requirements

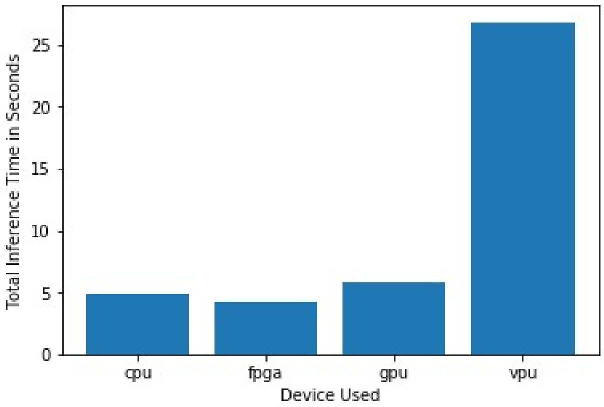
Maximum number of people in the queue	2 to 5
Model precision chosen (FP32, FP16, or Int8)	FP32

Test Results

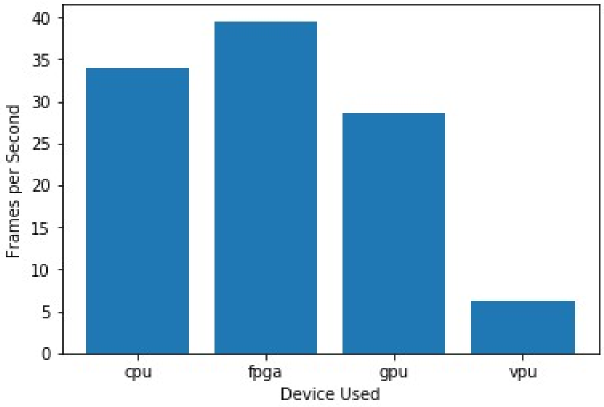
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
CPU. *Mr. Lin already has modern computer (Intel i7 processor) installed on most of the counters. These processors are only used to carry out minimal tasks, We will be using the existing hardware and adding a smart queuing system to it.* *In this scenario, FPGAs have performed best in the test results. But FPGAs are very expensive. GPU has also performed fairly well in the test results. Most CPUs come with integrated GPUs. It is not mentioned if the existing computers that Mr. Lin has, do have integrated GPUs. If they do, some of the process can be off-loaded to the GPU.* *Even if the existing computer does not have IGPUs, the existing CPUs suffice the requirements for a smart queuing system, without adding any additional budget towards hardware or electricity bill. CPU has a quick load time and can process about 34 frames per second. This will help in quickly redirecting the customers to a less congested lane.*

Device	Inference Time	FPS	Load Time
CPU	4.9	33.87755	1.53244
GPU	5.8	28.62069	57.5293
VPU	26.8	6.19403	2.833195
FPGA	4.2	39.52381	29.00681

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)
Neural Compute Stick 2 (NCS2)

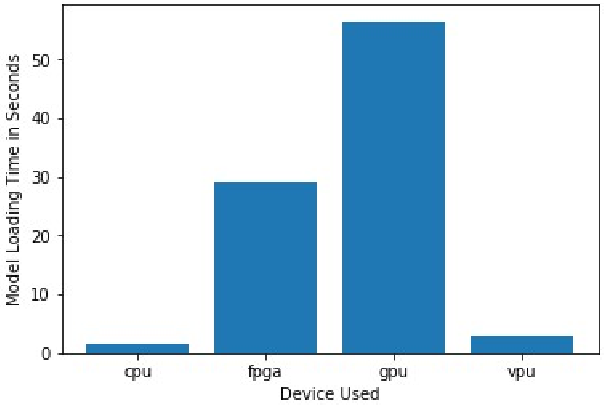
Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
Ms. Leah's budget allows for a maximum of $300 per machine	NCS2 is an inexpensive option, typically costing around $70 to $100 and would fit in the price range.
The CPUs in existing machines are currently being used to process and view CCTV footage for security purposes. **No significant additional processing power is available to run inference.**	(NCS2) is a USB3.1 plug and play removable VPU for AI inferencing and is a low power device.

Queue Monitoring Requirements

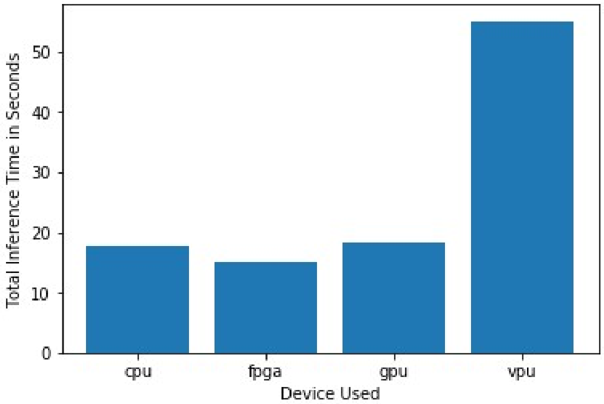
Maximum number of people in the queue	8
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

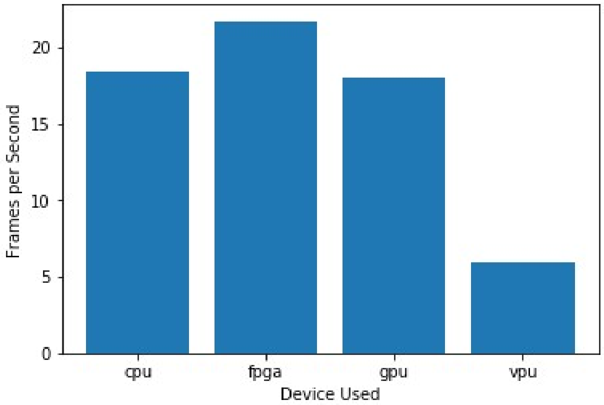
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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

Neural Compute Stick 2 (NCS2)

The CPUs in the current system are being used to view CCTV footage for security purposes and no significant additional processing power is available to run inference. Ms. Leah's budget is limited to $300/- per machine. NCS2 or VPUs are specialized for image processing. VPUs have specialized accelerators for image processing. VPUs have very low power consumption and are inexpensive, costing around $70 to $100. NCS2 also allow multiple inferences to run in parallel, They can be easily deployed at the edge. One drawback of NCS2 is that it has high inference time.

FPGA would be a good alternative since they are flexible in that they can be reprogrammed to adapt to new networks and have a long life span of upto 10 years. In this scenario, FPGAs have a low inference time and can process about 22 frames per second. Although VPUs and FPGAs can both be used as AI accelerators, FPGAs cost a lot more (mentioned in lesson 5 – FPGA specifications; in Choosing the Right Hardware).

Considering Ms. Leah's budget and power consumption options, the recommended hardware is NCS2.

Device	Inference Time	FPS	Load Time
CPU	17.9	18.4357	1.48066
GPU	18.3	18.03279	56.45765
VPU	55.1	5.989111	2.813685
FPGA	15.2	21.71053	29.11557