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

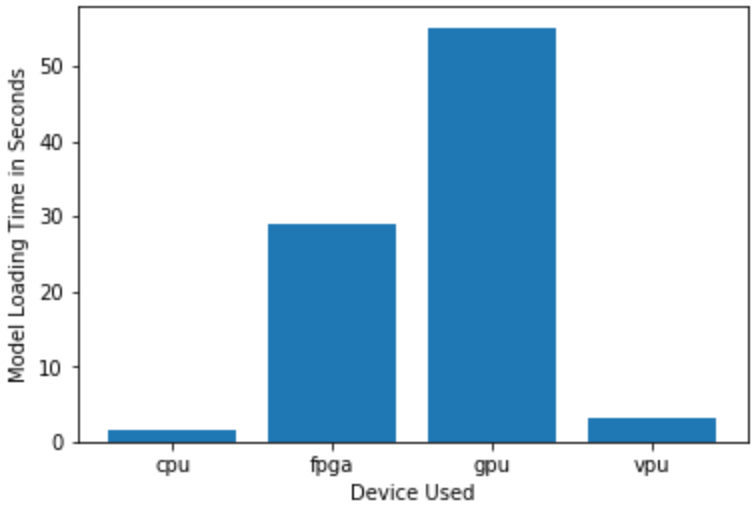
Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
The client requires the system to be able to run inference on the video stream very quickly.	*We can program an FPGA to act as an AI accelerator so that it performs well when running inference.*
The system would also need to be flexible so that it can be reprogrammed and optimized to quickly detect flaws in different chip designs.	*FPGAs are field-programmable. 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.*
The client would ideally like the system to last for at least 5-10 years.	*FPGAs have a long lifespan. For example, FPGAs that use devices from Intel's Internet of Things Group has guaranteed availability of 10 years, from the start of production.*
The client wants a system to monitor the number of people in the factory line and would like the image processing task to be completed five times per second.	*Once programmed with a suitable bitstream, FPGAs can execute neural networks with high performance and very little latency. They also support large networks making them useful in deep learning.*

Queue Monitoring Requirements

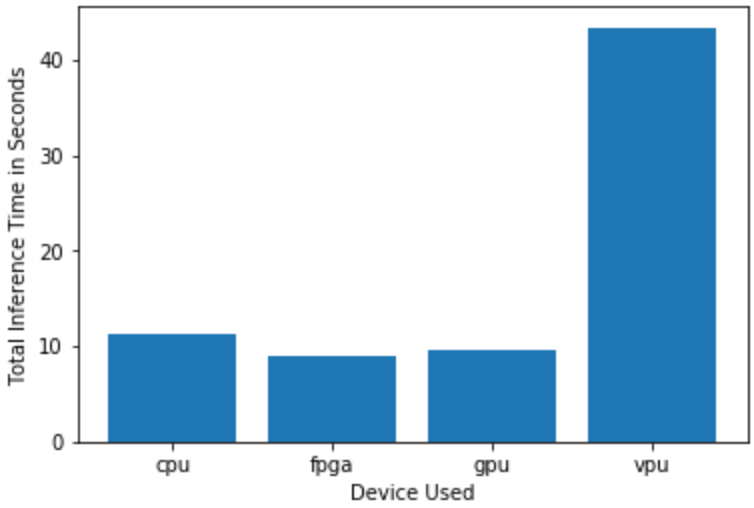
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

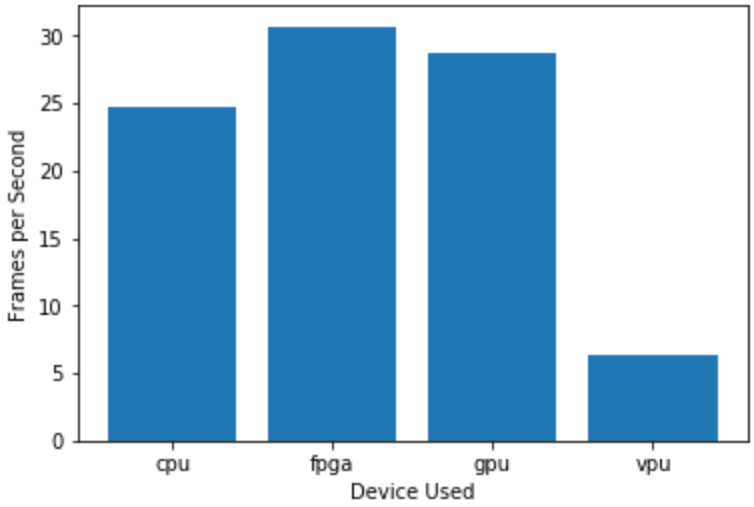
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
The primary requirements requested by the client are: *The first issue regarding productivity during shift transitions* - *Monitor number of people in the factory line* - *Image processing task to be performed 5 times per second* *The second issue regarding identifying flawed chips before packaging and shipping* - *The system needs to run inference on video stream quickly* - *The system needs to flexible to be reprogrammed according to the new designs* - *Lifespan of at least 5 - 10 years* *Field-Programmable Gate Arrays (FPGAs) are chips designed with maximum flexibility so that they can be reprogrammed as needed in the field. FPGA can be reprogrammed to act as an AI accelerator so that it performs well when running inference. Once programmed with a suitable bitstream, FPGAs can execute neural networks with high performance and very little latency. FPGAs can be deployed in harsh environments like factory floors and still perform optimally. FPGAs also have a long lifespan. In fact, FPGAs that use devices from Intel's Internet of Things Group has guaranteed availability of 10 years, from production.* *It can be observed from the test results that FPGA has shown the best performance. It meets the requirement of quick inference time with consuming less than 10 seconds. It also has 30-35 FPS which is the highest among the compared hardware types. Hence, FPGA can be considered as the best possible choice from the four hardware types for this scenario of manufacturing.*

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

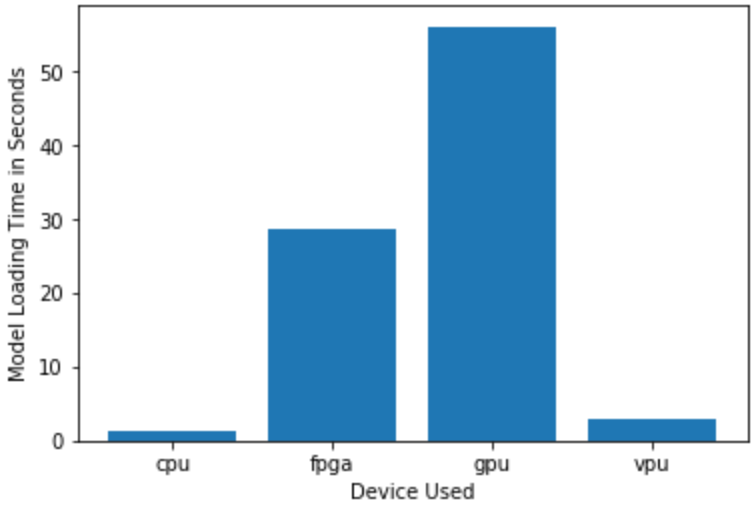
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
Most counters already have a modern computer with an Intel i7 core processor.	*IGPU is located on a processor alongside the CPU cores and shares memory with them.*
Limited budget resulting in no additional hardware requirement.	*No need for adding dedicated external GPU.*
Less power consumption	*Unused sections in a GPU can be powered down to reduce power consumption.*

Queue Monitoring Requirements

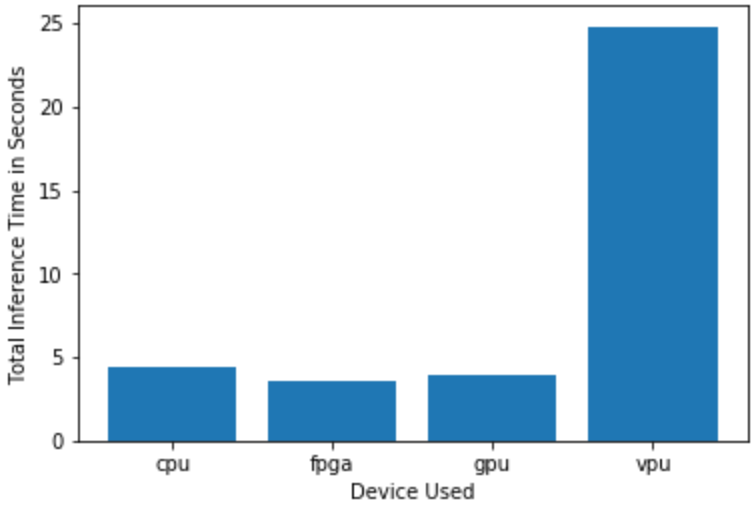
Maximum number of people in the queue	*3*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

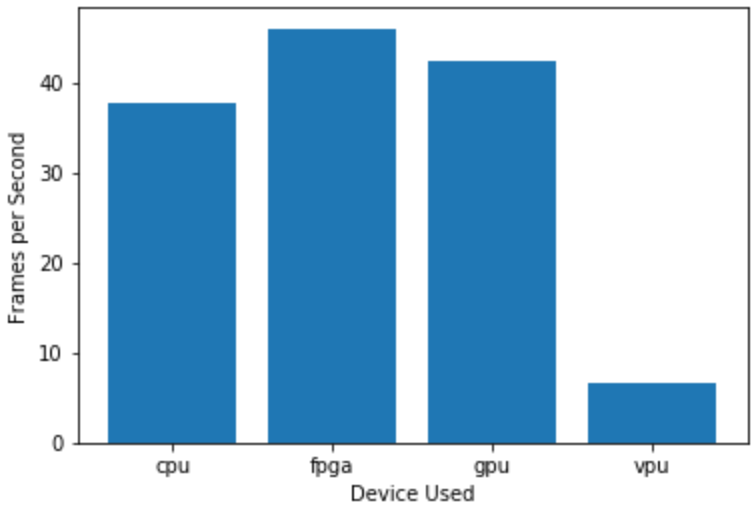
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
The client addresses the problem of congestion at the checkout counter by requesting for a system to direct people to less-congested queues in the store using Edge AI. It is stated that the store is already equipped with modern computers with Intel i7 core processor. Another important requirement of the client is to minimize the installation of additional hardware and power consumption as much as possible hence, employing a restricted budget. *In order to avoid additional hardware and minimize expenditure, we have to eliminate the usage of FPGA and VPU as they required $100 - $1000 additional cost. With the availability of Intel i7 core processor using Integrated GP (IGPU) is the best choice over CPU. Also compared to CPU, IGPU has fairly less inference time and high FPS of 40 - 45.*

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

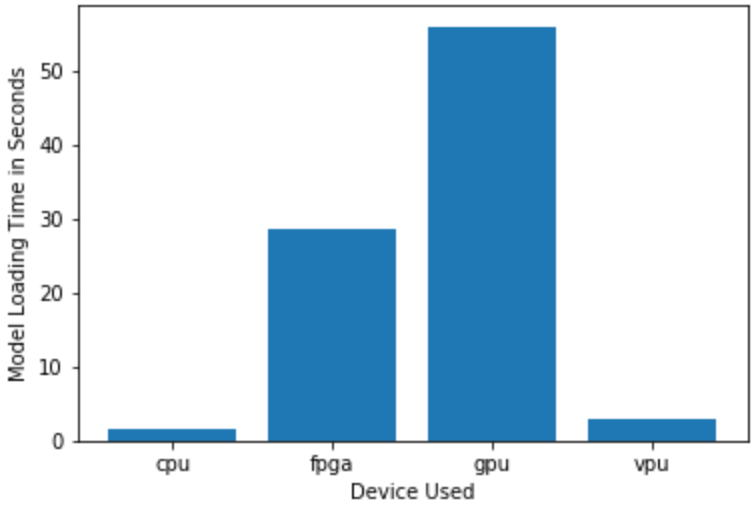
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client's budget is up to $300 per machine.	*NCS2 is a plug and play removable VPU for AI inference with the cost ranging from $70 - $100*
The client has no additional processing power available to run inference.	*Myriad X VPU has two on-chip CPUs to run the host inference and on-chip coordination between NCE, vector processor, imaging accelerators.*
The client requires less power consumption.	*VPUs are low-power devices with 1-2 watts of power consumption.*

Queue Monitoring Requirements

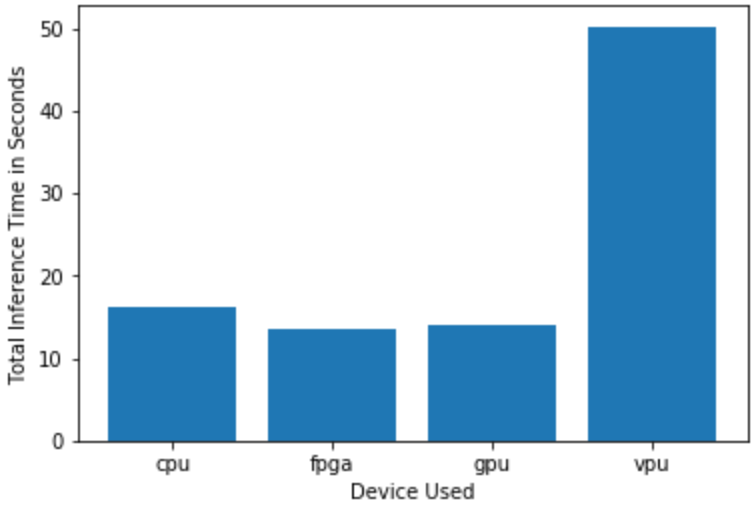
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

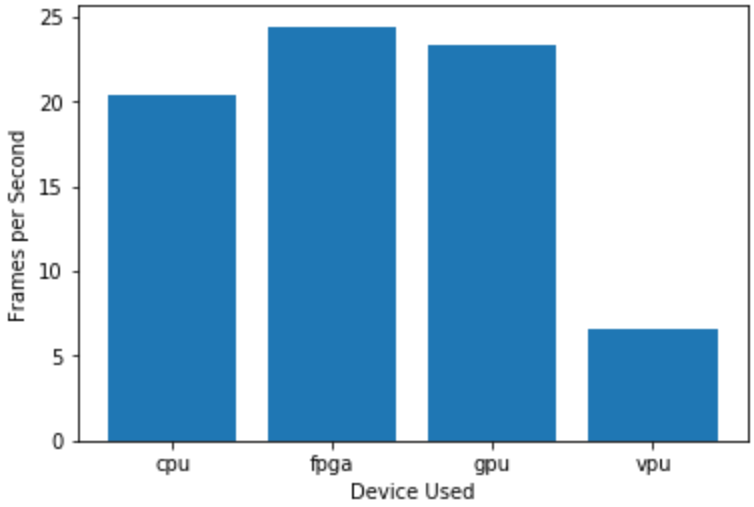
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
The client would like to automate the process of directing the Delhi Metro passengers to less congested areas during peak time using an Edge AI system. The system is required to monitor the queues in real-time and quickly direct the crowd in the proper manner. *There is no significant additional processing power available to run inference as all the CPUs in the machines are being used to process and view CCTV footage for security purposes.* *The client's budget allows for a maximum of $300 per machine.* *VPUs are accelerators that are specialized for AI tasks related to computer vision – such as CNNs and image processing. They are small, low-cost, low-power devices. They contain hardware accelerator optimized for running deep learning neural networks at low power without any loss in accuracy.* *The Myriad X has very low power consumption of 1-2 watts. Its 2.5Mb of on-chip memory reduces latency and power consumption.* *NCS2 is a USB3.1 plug and play removable VPU for AI inferencing. It using Myriad X VPU and cost from $70 - $100 meeting the client's budget.* *Hence, the VPU hardware type is considered as the best choice for Transportation scenario.*