Choose the Right Hardware

Scenario 1: Manufacturing

Client Requirements and Potential Hardware Solution

Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
FPGA

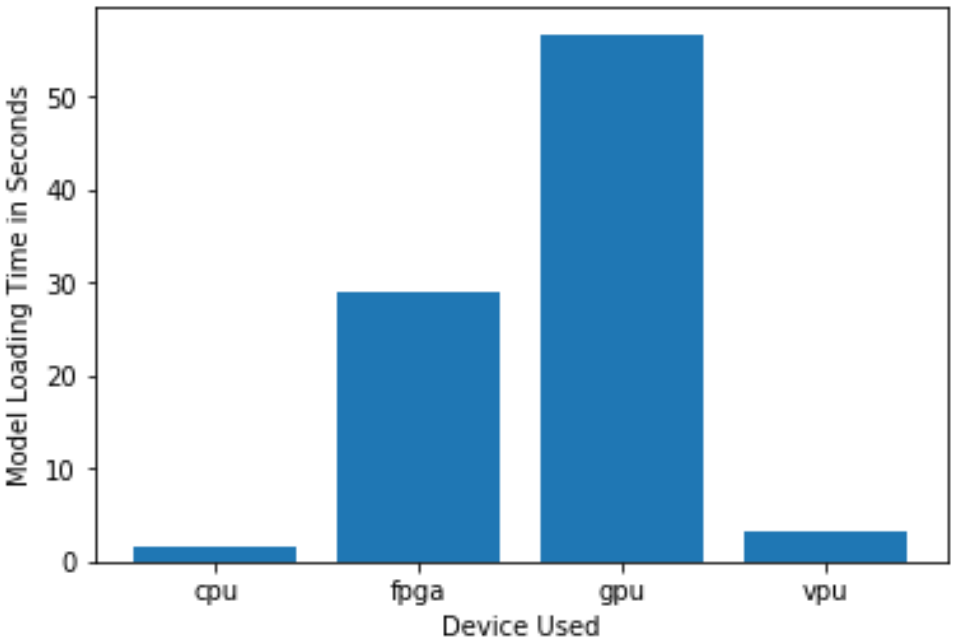
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client would like the image processing task to be completed five times per second.	We can take advantage of the massive parallelism of the FPGA to process the images.
The client would need the system to be able to run inference on the video stream very quickly.	We can program an FPGA to act as an AI accelerator for the particular model we are going to run, so that it performs well when running inference.
The client need the system to be flexible so that it can be reprogrammed and optimized to quickly detect flaws in different chip designs.	FPGA can be reprogrammed to optimize the performance for different functions as needed.
The client would 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 have a guaranteed availability of 10 years, from start of production.

Queue Monitoring Requirements

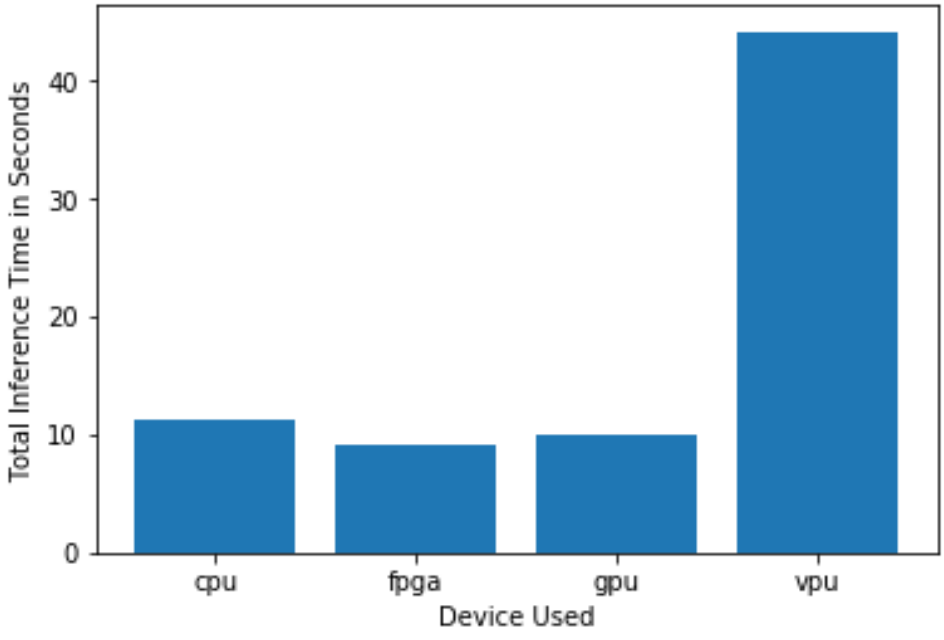
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*CPU → FP32 | VPU, GPU, FPGA → FP16*

Test Results

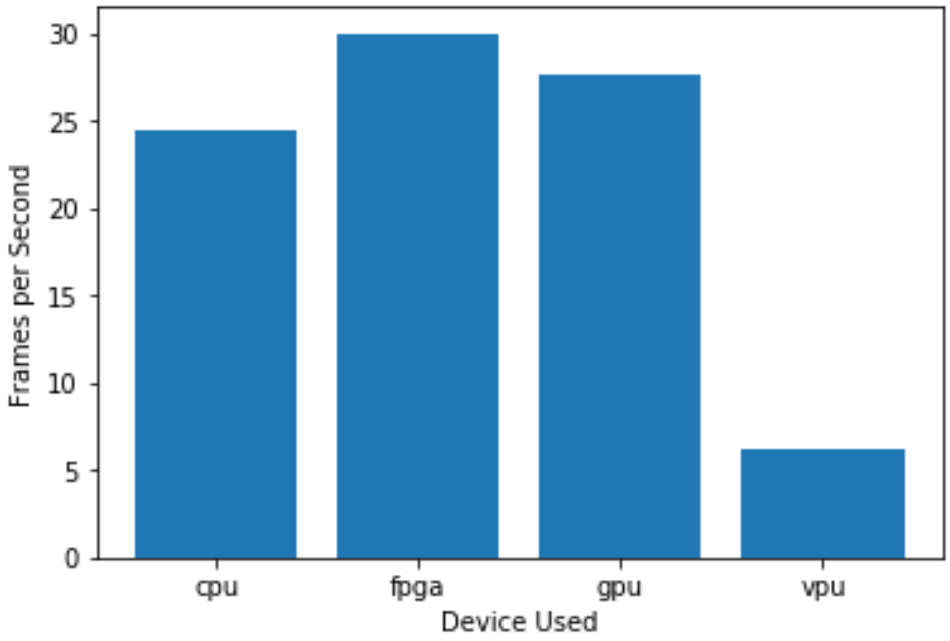
After testing the application on all four hardware types (CPU, IGPU, VPU, and FPGA), you can see below the comparison graphs (for model load time, inference time, and FPS).



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

Write-up: Final Hardware Recommendation
From the test result, we can see that FPGA have the best inference time and FPS and that is exactly what the client want. *The client also needs the system to be flexible so that it can be reprogrammed and would like it to last for at least 5-10 years, we can find these two requirements in FPGA characteristics. At the end, we can say that FPGA is the best choice.*

Scenario 2: Retail

Client Requirements and Potential Hardware Solution

Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
IGPU

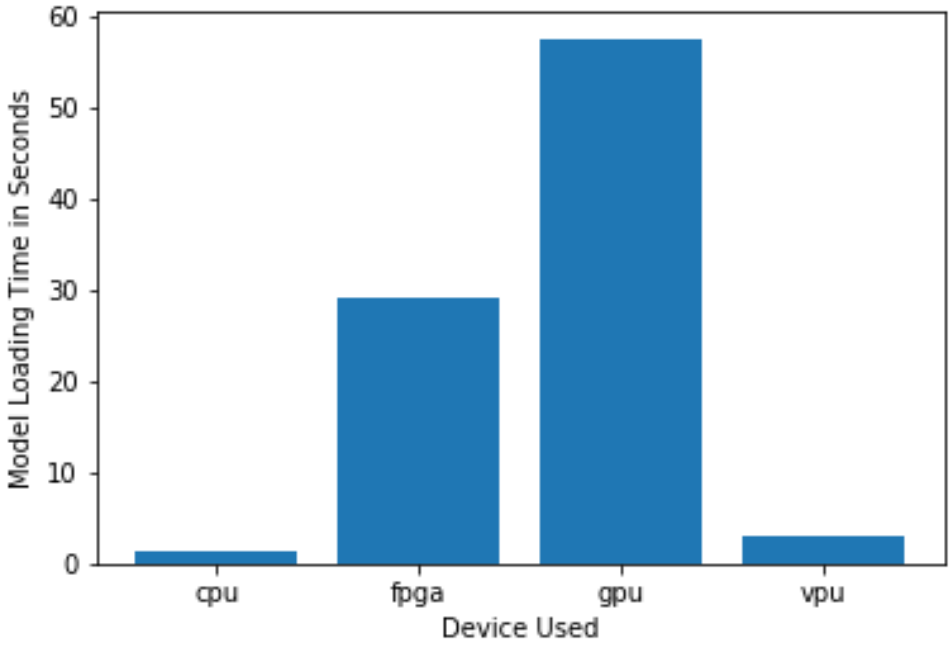
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client does not have much money to invest in additional hardware.	We will use the existing hardware "Intel i7 core processor" since it have an IGPU.
The client would like to save as much as possible on his electric bill.	We will not add any new device.

Queue Monitoring Requirements

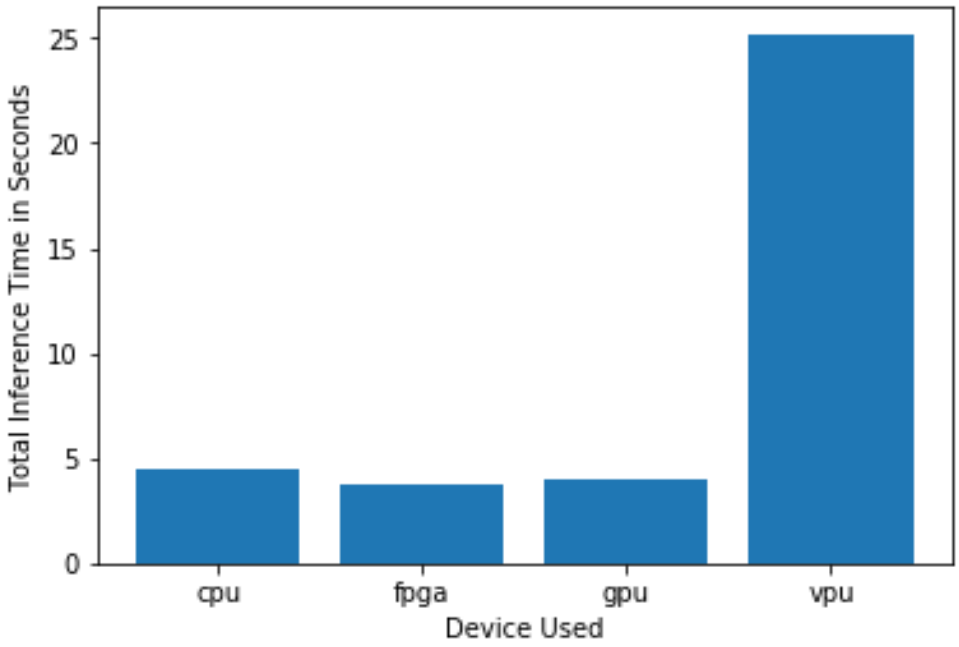
Maximum number of people in the queue	*2*
Model precision chosen (FP32, FP16, or Int8)	*CPU → FP32 | VPU, GPU, FPGA → FP16*

Test Results

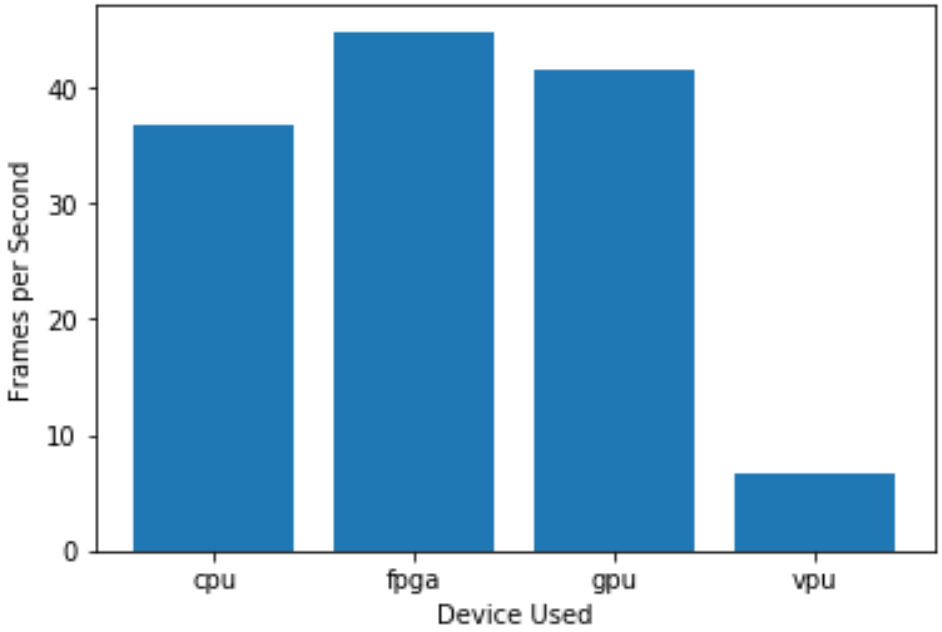
After testing the application on all four hardware types (CPU, IGPU, VPU, and FPGA), you can see below the comparison graphs (for model load time, inference time, and FPS).



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

Write-up: Final Hardware Recommendation
From the test result, we see that FPGA is the best choice but since the client does not have much money to invest in additional hardware, and want to save as much as possible on his electric bill. We will use the IGPU in the existing hardware "Intel i7 core processor".

Scenario 3: Transportation

Client Requirements and Potential Hardware Solution

Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
VPU

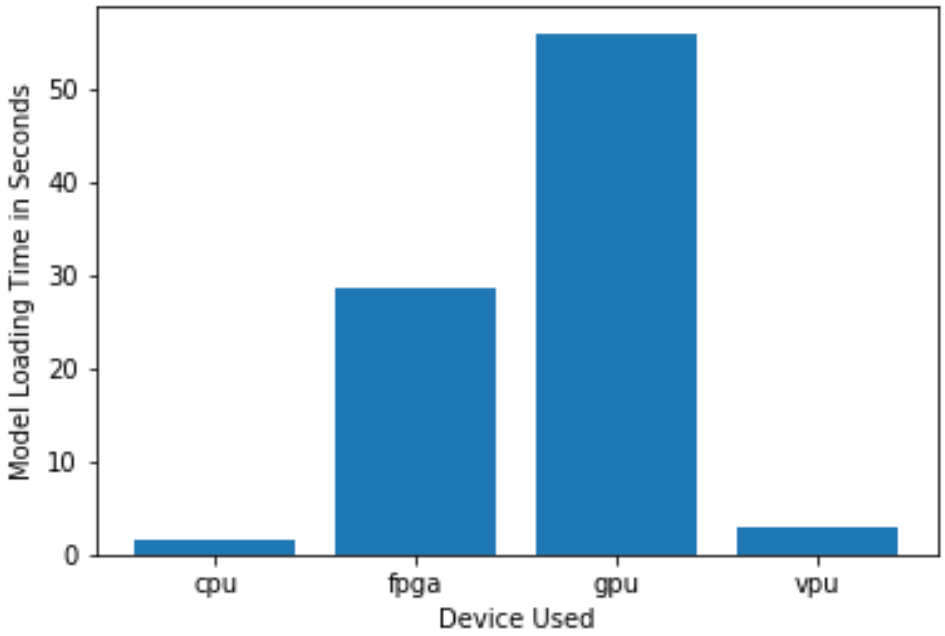
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The CPUs in client machines are currently being used to process and view CCTV footage for security purposes, no significant additional processing power is available to run inference, and the client budget allows for a maximum of $300 per machine.	VPU can be used to accelerate the performance of a pre-existing system and it is a low-cost device typically costing around $70 to $100.
The client would like to save as much as possible on future power requirements.	VPU is a low power device.

Queue Monitoring Requirements

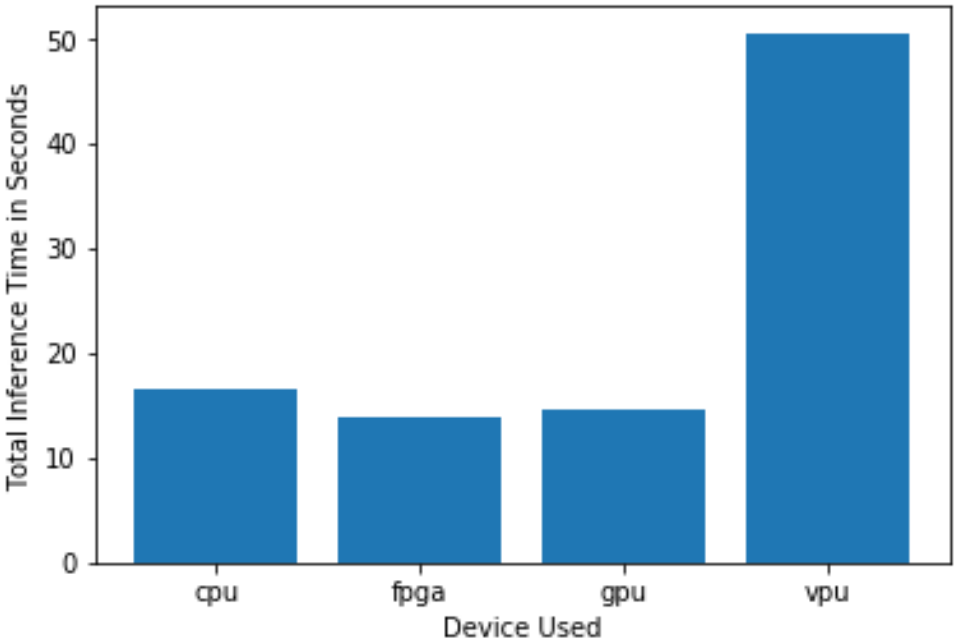
Maximum number of people in the queue	*6*
Model precision chosen (FP32, FP16, or Int8)	*CPU → FP32 | VPU, GPU, FPGA → FP16*

Test Results

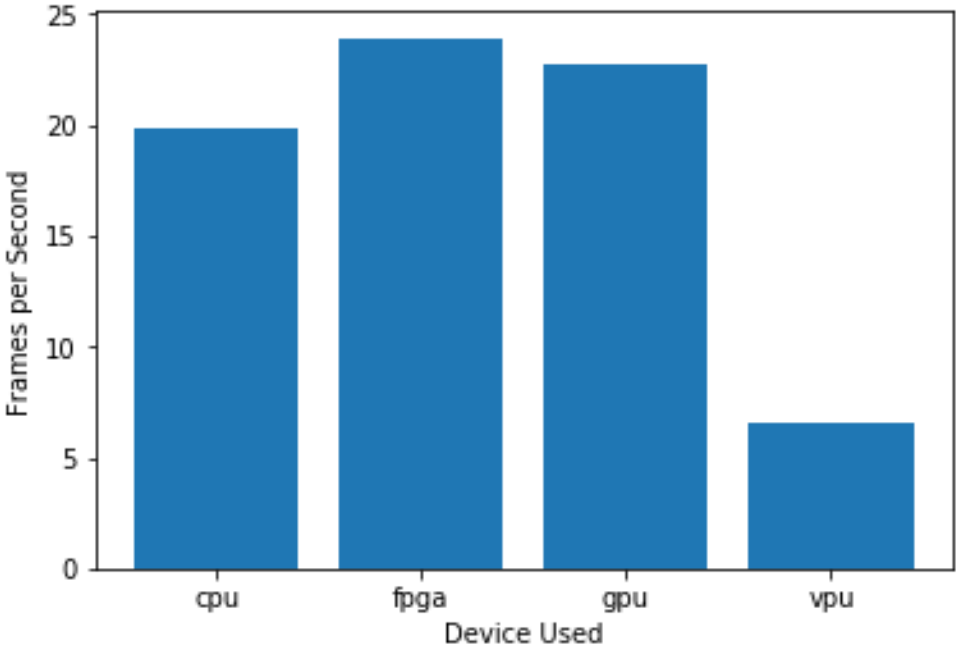
After testing the application on all four hardware types (CPU, IGPU, VPU, and FPGA), you can see below the comparison graphs (for model load time, inference time, and FPS).



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

Write-up: Final Hardware Recommendation
From the test result, we can see that FPGA , GPU and CPU can be a good choice in terms of inference time and FPS, but since the client would like to save as much as possible on hardware and future power requirements, and his machines have no significant additional processing power to run inference, We will use VPU since it is low-cost and low power device.