Choose the Right Hardware

Proposal Template by Olukolatimi David

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

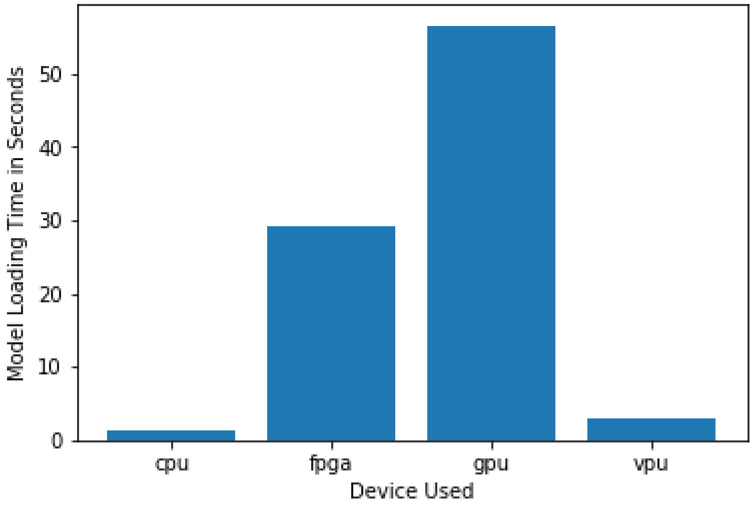
Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
The client requires a system that can be reprogrammed to solve different tasks.	***Flexibility -*** *FPGA(Field Programmable Gate Array), from the name, it is an hardware that can be reprogrammed It supports various precision options (FP16, FP11, FP9), thus it allows developers to balance between speed and accuracy. The bitstream used in configuring an FPGA can be changed without changing the hardware.*
The client requires an hardware that would last at-least 5 to 10 years.	**Long Lifespan -** 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

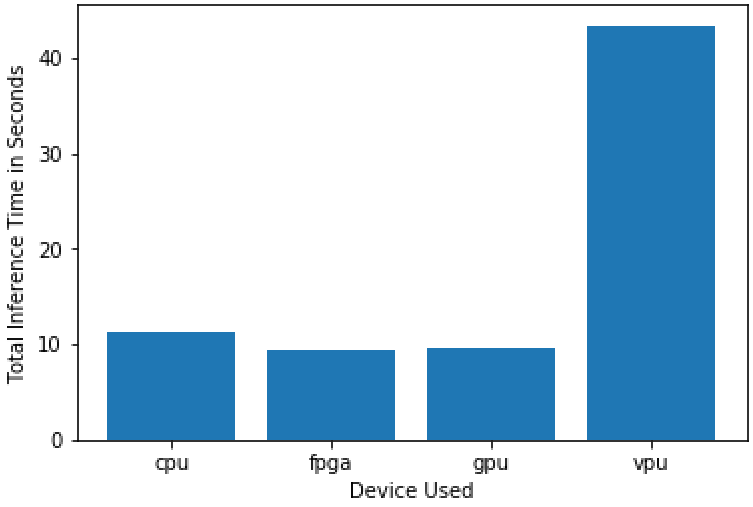
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

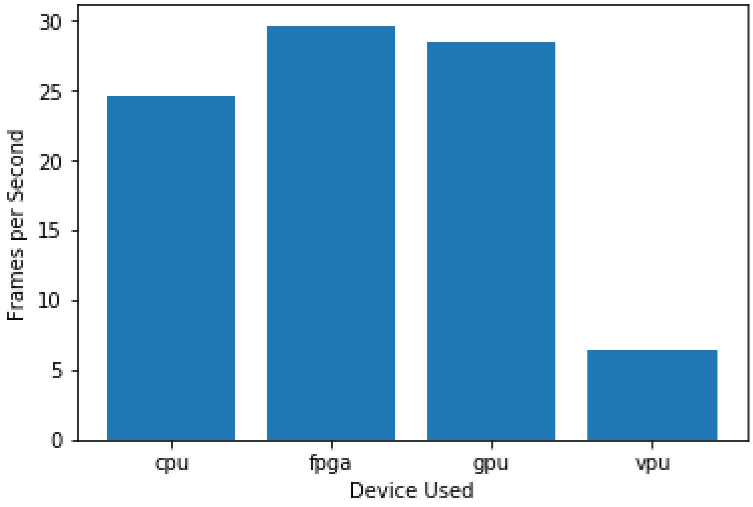
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 wants to implement a quality system which is a significant investment and would really like the system to last for at least 5 to 10 years. FPGAs have a long lifespan like how* FPGAs that use devices from Intel’s Internet of Things Group have a guaranteed availability of 10 years, from start of production.

- ✓ The client wants a flexible system. The customer requires a system that can be easily changed and still be very well optimized. FPGAs are field programmable, and can be reprogrammed easily to adjust to new configurations.

- ✓ The client wants a system that can make inference on the video very quickly. From the inference_time graph, we can see that the FPGA runs inference fastest compared to the oher devices and hence meets the customer’s requirement.

- ✓ Despite that the model loading time on the FPGA is higher than the model load time for CPU and VPU, the FPGA meets the client’s requirement and we can trade-off the model load time. The FPGA can be online 100%, i.e it can be operating continuously for 24hours a day.

- ✓ The client’s camera records video at 30-35 Frames per second(FPS) and the FPGA reads at about 30 Frames per seconds.

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

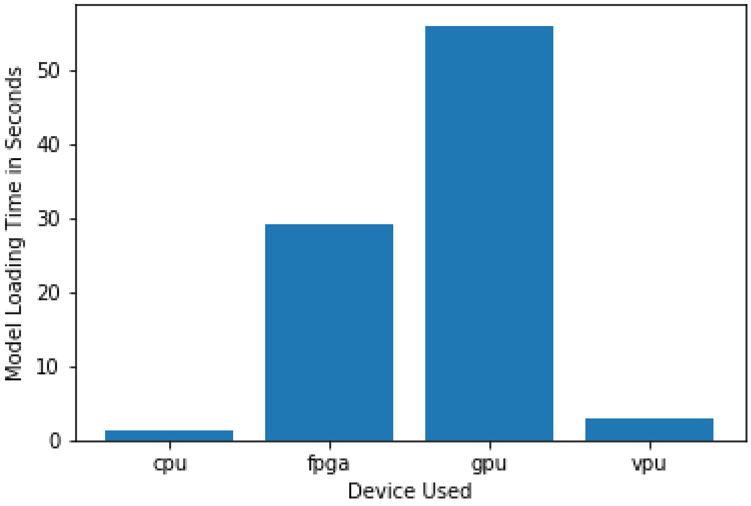
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client has already computers with intel i7 processors not being used for any computational expensive tasks.	*An IGPU is a GPU that is located on a processor alongside the CPU cores and they both share memory. The client already has modern computers equipped with intel i7 and contains GPUs.*
The client does not have much money to invest in new hardwares.	*Since there's already modern computers that have intel i7, the client won't need to purchase new VPUs or even FPGA, since we're trying to minimize cost and still be efficient.*
The client wants to save as much as possible on his electric bill.	***Power Consumption -** Since the clock rate of the slice and unslice can be controlled separately, the unused section of the GPU can be powered down to reduce power consumption.*

Queue Monitoring Requirements

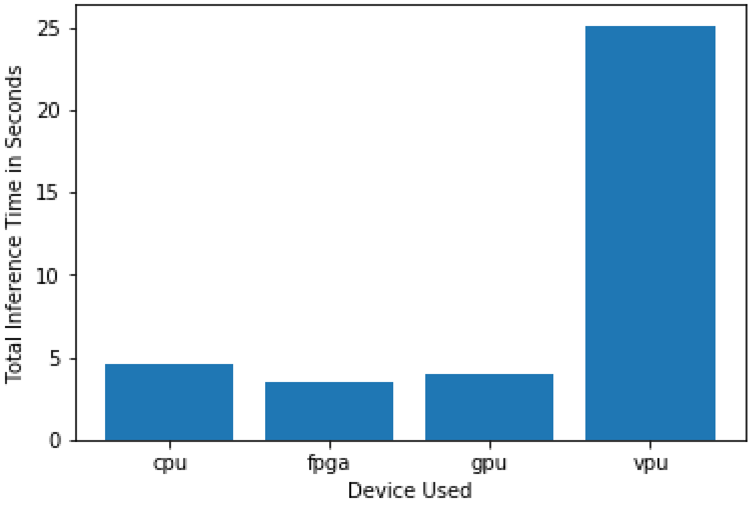
Maximum number of people in the queue	*During rush hours : 5*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

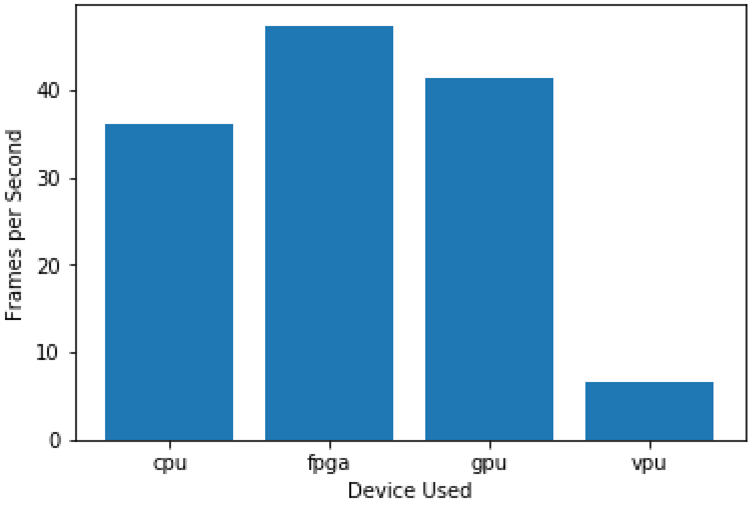
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 does not have a lot of money to invest in additional equipment, meaning cannot afford to buy a VPU or FPGA. The client can use the modern computers equipped with intel i7 and contain IGPU.* ✓ *The customer wants to save as much money as possible on electric bill. A CPU for high performance would require a lot of power, while in IGPU, the clock rate of the slicing can be controlled separately. This means the unused sections of the GPU can be turned off to reduce power consumption.* ✓ *The IGPU takes less time to run inference on video than the CPU and VPU, however not FPGA, but considering the cost of an FPGA, the IGPU is what meets the client's needs.* ✓ *The IGPU also processes more Frames per second than the CPU.* ✓ *Although the model load time of the IGPU is about 50 seconds more than that of the CPU, the model load time can be a trade-off and then consider the inference time, frames per second and power.*

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 budgets at most $300 per machine.	NCS2 is only about $100 and would fit in the price range.
The client wants an edge AI system that would monitor the queues in real-time and quickly direct the crowd in the right manner.	VPU or NCS2 can run inference very fast because the vector processors in a VPU can break up a complex instruction and then execute many tasks in a parallel manner.
The CPUs in the client's machine are currently being used to process and view CCTV footage for security purposes and no additional processing power is available to run inference.	*The client's PC requires more processing power, the NCS2 is a low powered device and it can be used to make inference on models.*

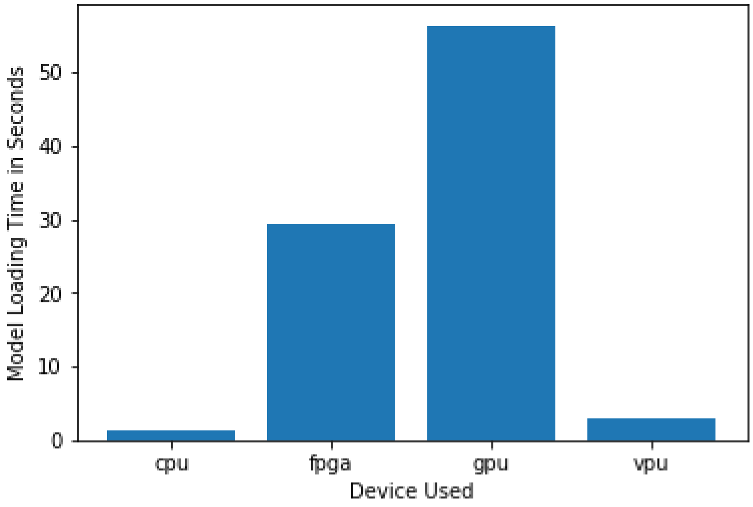
Queue Monitoring Requirements

Maximum number of people in the queue	*During Peak-Hours : 15*

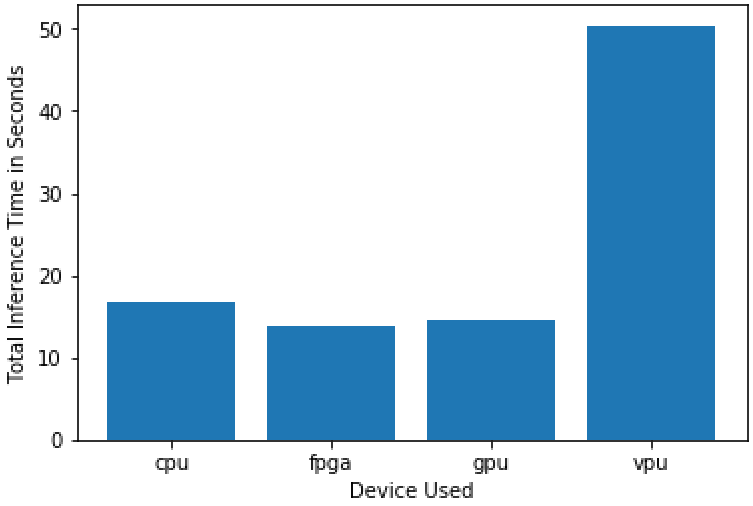
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

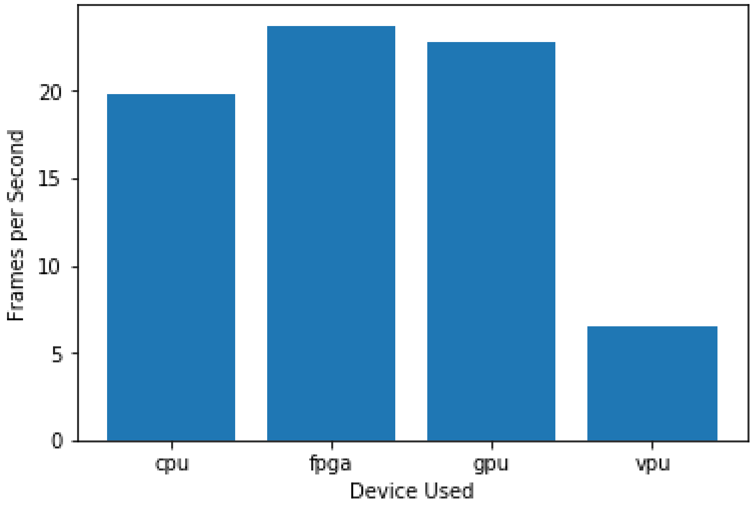
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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

- *The client has a budget of at most $300 per machine, and she would like to save as much as possible both on hardware and future power requirements. A VPU or NCS2 is the necessary hardware for the edge AI system.*
- *The client cannot use FPGA because it costs way more than the $300 budget.*
- *The client cannot use CPU because they are used to process and view CCTV footage for security purposes and no significant additional processing power is available to run inference.*
- *Although the VPU reads fewer frames than the CPU, FPGA, and IGPU, the VPU is what still meets the client's needs.*
- *The model load time for the VPU is faster than the FPGA and IGPU but slightly slower than CPU.*
- *The Inference time for the VPU is significantly higher than that of the CPu, IGPU and FPGA , but these hardwares don't meet the client needs, FPGA is too expensive, CPU is already being used extensively.*

UDACITY

UDACITY