Choose the Right Hardware

The following is the Hardware Proposal Template for Project title “Smart Queue Monitoring System” in Intel Edge AI for IoT Developers Nanodegree Scholarship Program.

Please note:

Even though a choice between FP32, FP16, and FP11 were present, only FP16 was chosen and used. The reason for this is that FP16 is supported across all devices and with FP16 models, twice the number of calculations can be performed in a given time as compared to FP32. Furthermore, unlike FP11 and INT8 model types, FP16 represents a balancing point between lower precision and accuracy of the model. For these reasons, FP16 models were preferred.

Scenario 1: Manufacturing

Client Requirements and Potential Hardware Solution

Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
As per the requirement of the Client, Naomi Semiconductors, the most appropriate hardware at this point seems to be a **FPGA.**

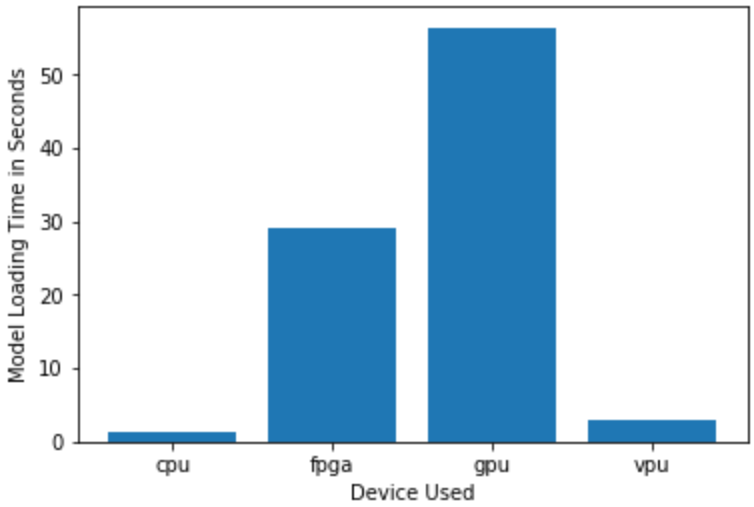
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client under discussion requires a system that can be easily repackaged/repurposed	FPGAs can be easily re-programmed for different solutions by loading Bitstream appropriate for that solution
The client needs the system to have a throughput for queue detection.	FPGAs can offer high throughput.
The client needs the system to have low latency for chip fault detection	FPGAs are suitable for low latency tasks as well. They have low inference time when coupled with CPU as backup.
The system should be long lasting (as per question and requirements, the system should last at least 5 year or more).	FPGAs have a longer lifespan compared to other alternatives

Queue Monitoring Requirements

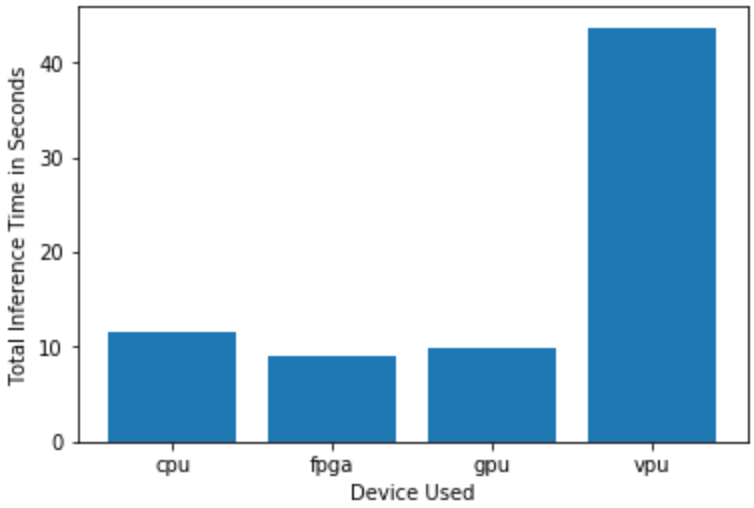
Maximum number of people in the queue	2
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

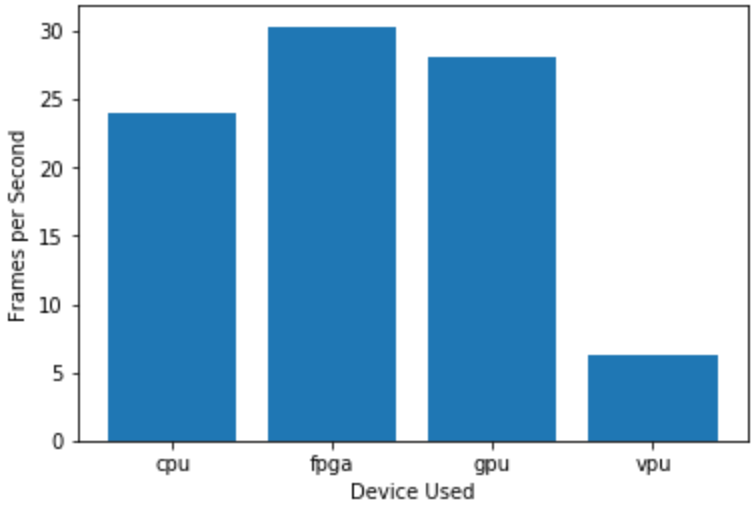
The 3 comparison graphs obtained after running the process are:-



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

The test results were surprising in the fact that out of all four hardware models - CPU, Integrated GPU, FPGA and VPU (Neural Compute Stick 2), the VPU performed poorly. The FPGA has consistent good performance across the used metrics as evident from the graphs. Thus, the Final Hardware Recommendation remains the same.

Write-up: Final Hardware Recommendation
FPGA

Scenario 2: Retail

Client Requirements and Potential Hardware Solution

Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
Mr.Lin of PriceRight Singapore outlet might consider going with a CPU + IGPU (MULTI) Solution.

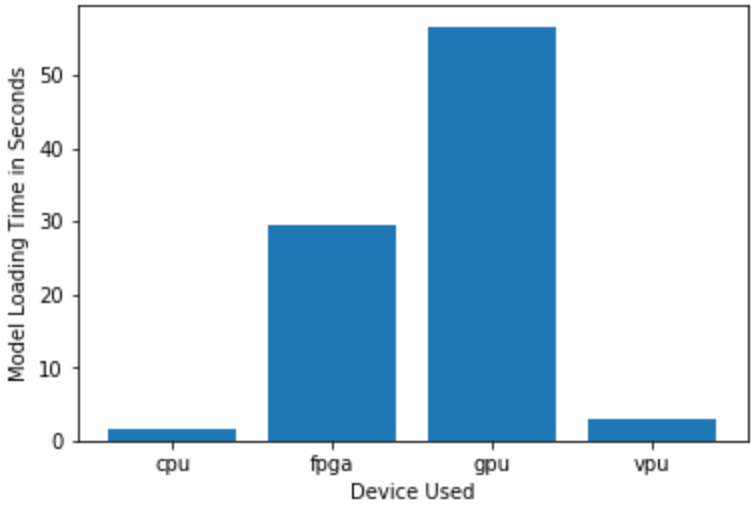
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The average wait time of a person in a queue is 230 seconds in normal while 350-400 seconds during rush hour.	A low latency device is required. However, the inference time does not need to be that low since average wait time in a normal scenario is > 1 minute.

Mr. Lin already has a i7 core processor computer running.	These are not used for computationally intensive tasks. As such the CPU can be leveraged.
Mr. Lin would like to invest as low as possible for the solution	Use of CPU would minimize any hardware related upgrade cost

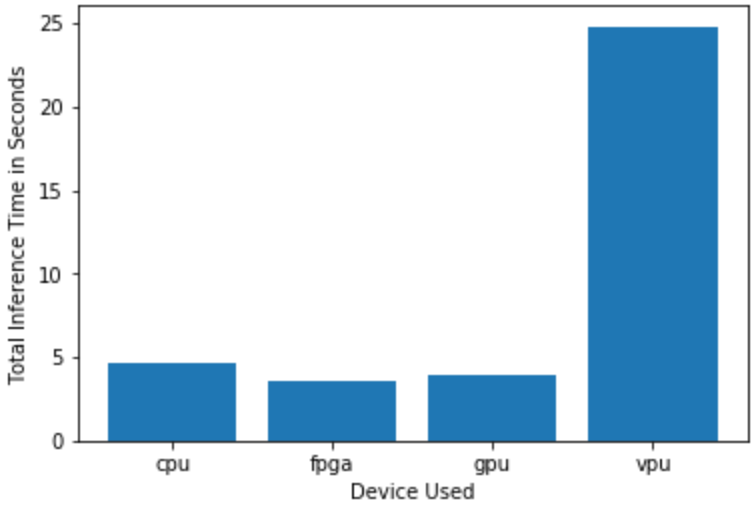
Queue Monitoring Requirements

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

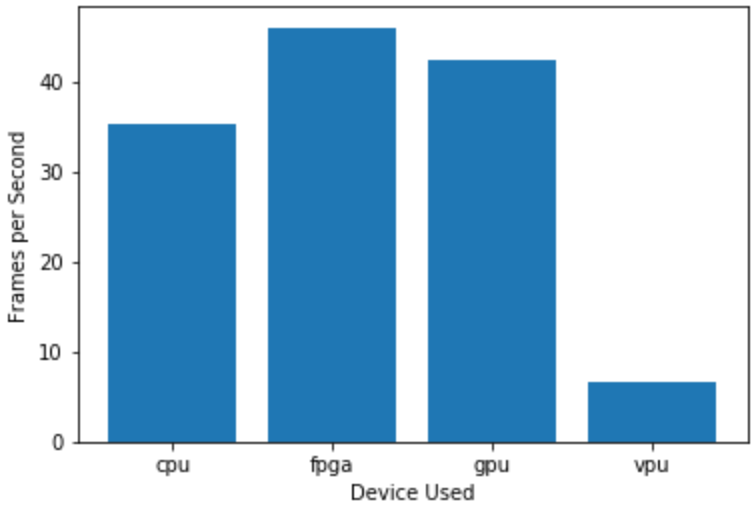
Test Results



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

From the above metrics, we can observe that:-

1. In Terms of Model Inference Time, the VPU is not faring very well. The IGPU (labelled as GPU) offers performance second to FPGA.
2. In terms of Model Load time, the IGPU takes the most time to load the model while the CPU loads the model the fastest. This does not affect inference in a real-time scenario where the model is already loaded and has been working beforehand.
3. GPU (the IGPU + CPU combo) offers second to FPGA FPS metrics outranking the Neural Compute Stick 2 (labelled VPU).

Thus, the initial recommendation of a combination of IGPU+CPU is the optimal solution here. It also fits well within the budget constraints of Mr. Lin

Write-up: Final Hardware Recommendation
IGPU + CPU

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 with a fallback on CPU (MULTI) seems to be a right choice at the first assessment of the scenario

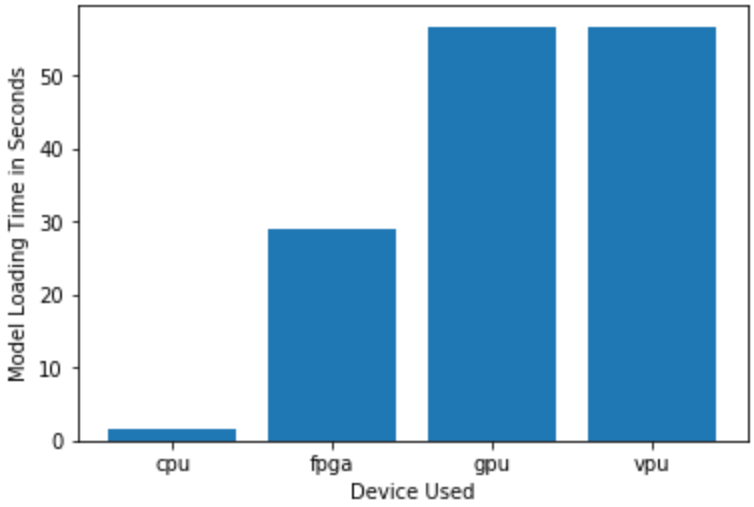
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
“In peak hours they currently have over 15 people on average in a single queue outside every door in the Metro Rail” & “On office hours there is a train every 2 mins” Thus, Low-latency is required.	VPUs like NCS 2 can accelerate inference time and provide lower latency.
“Ms. Leah's budget allows for a maximum of $300 per machine” Thus, budget.	NCS 2 is well within budget
“The CPUs in these machines are currently being used to process and view CCTV footage for security purposes and no significant additional processing power is available to run inference” No-extra resource to spare normally.	But using a VPU, the inference task can be offloaded to free up resources on the machines. The pre and post processings may also be offloaded to VPU. The CPU would act as a fallback device or be used when there is no more request space on the VPU.

Queue Monitoring Requirements

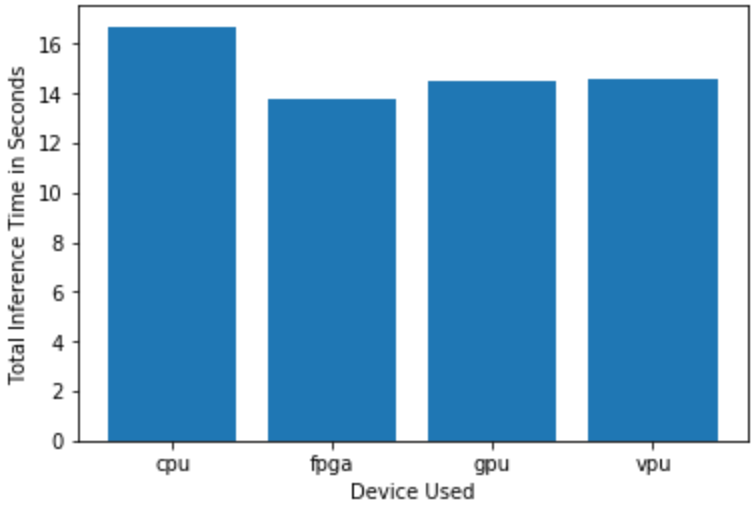
Maximum number of people in the queue	4
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

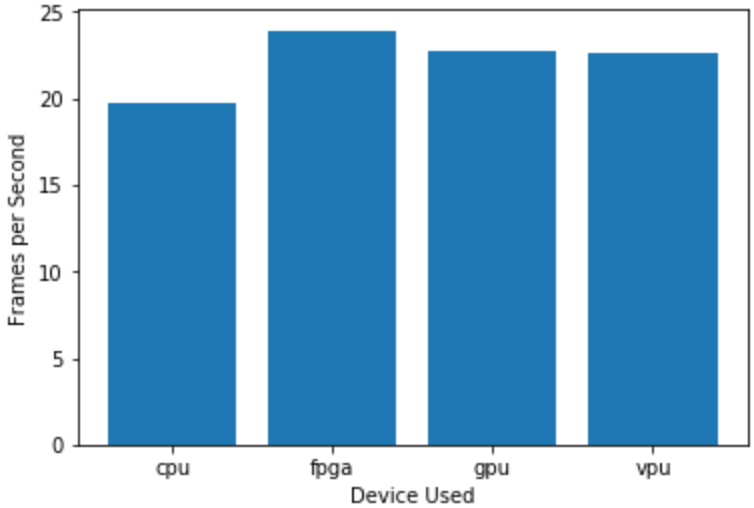
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

From the above graphs, it can be seen that almost all solutions offer results which are closely contested. The VPU keeps up the performance in a consistent manner across the metrics. Given that the CPU and Integrated GPUs cannot be used for inference in this case and FPGAs would be too costly a setup, VPUs (Intel Neural Compute Stick) form a better solution.

Write-up: Final Hardware Recommendation
VPU (Intel Neural Compute Stick)