Choose the Right Hardware

Kesha K. Kaneria

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)
FGPA is most appropriate for this scenario.

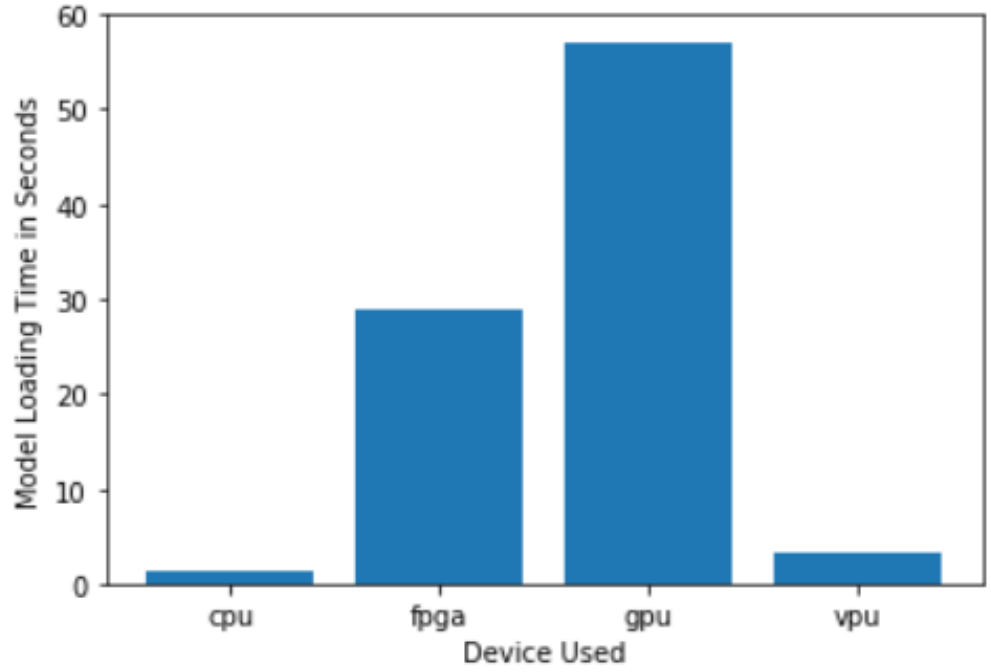
Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
The client wants to install a quality system, which can last for atleast 5-10 years.	*FPGA's have a long lifespan. Ex: FPGAs that use devices from Intel's IOT Group have a guaranteed availability of 10 years, from start of production.*
Should be flexible enough so that it can be optimized and reprogrammed also if needed.	*FPGA'S are field-programmable; they can be reprogrammed to adapt to new, evolving, and custom networks.*

Queue Monitoring Requirements

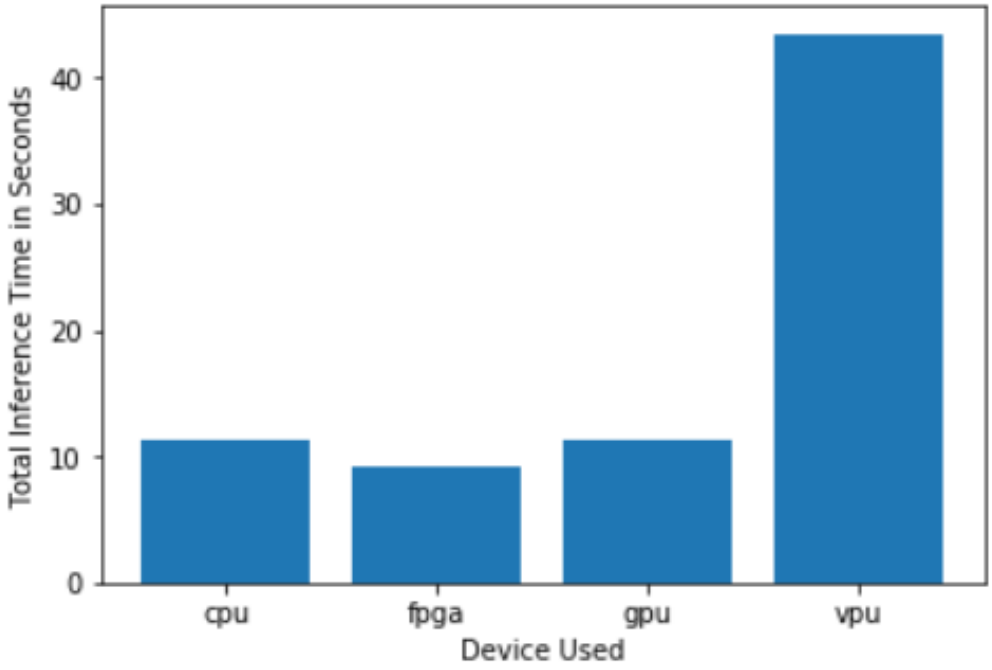
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

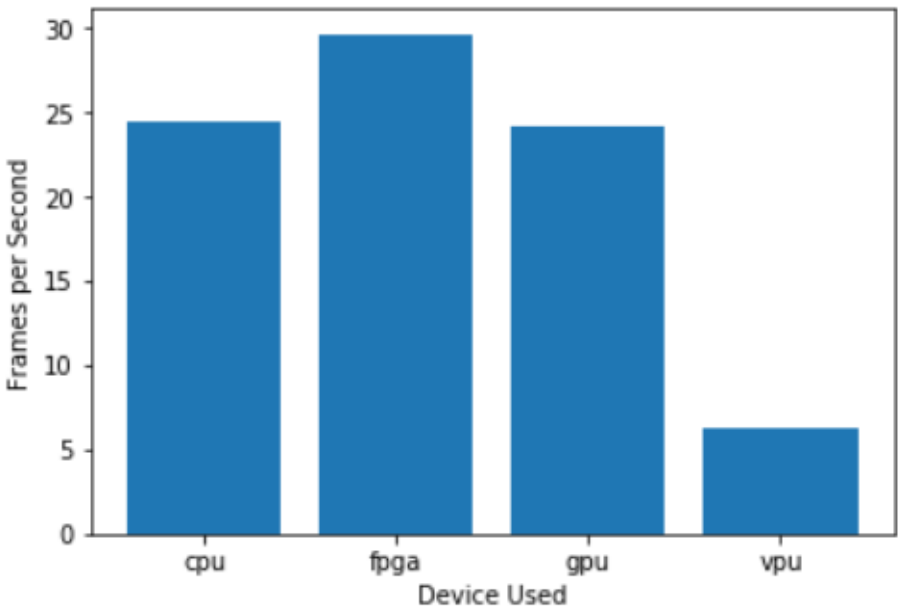
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
• *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. The client wants to install a quality system, this is still a significant investment and they would ideally like it to last for at least 5-10 years.* • *The client needs the system to be flexible so that it can be reprogrammed and optimized. FPGA'S are field-programmable; they can be reprogrammed to adapt to new, evolving, and custom networks* • *As we can see from above test results, FPGA's has the Faster Inference when compared to other devices and meets the client's requirements. The client wants a system would need to be able to run inference on the video stream very quickly.* • *Results shows, FPGA's reads nearly 30 FPS when compared to other devices and meets the client's requirements. Client's cameras record video at 30-35 FPS (Frames per Second) and this video stream can be used to monitor the number of people in the factory line.* • *Although CPU's and VPU's take very less time to load the model by given test results but, it may happen that these hardware devices do not meet the client's requirements. While FPGA's take nearly upto 30 sec to load the model they meet all the requirements needed by the client and have 100% on-time performance, meaning they can be continuously running 24 hours a day, 7 days a week, 365 days a year.*

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 is most appropriate for this scenario

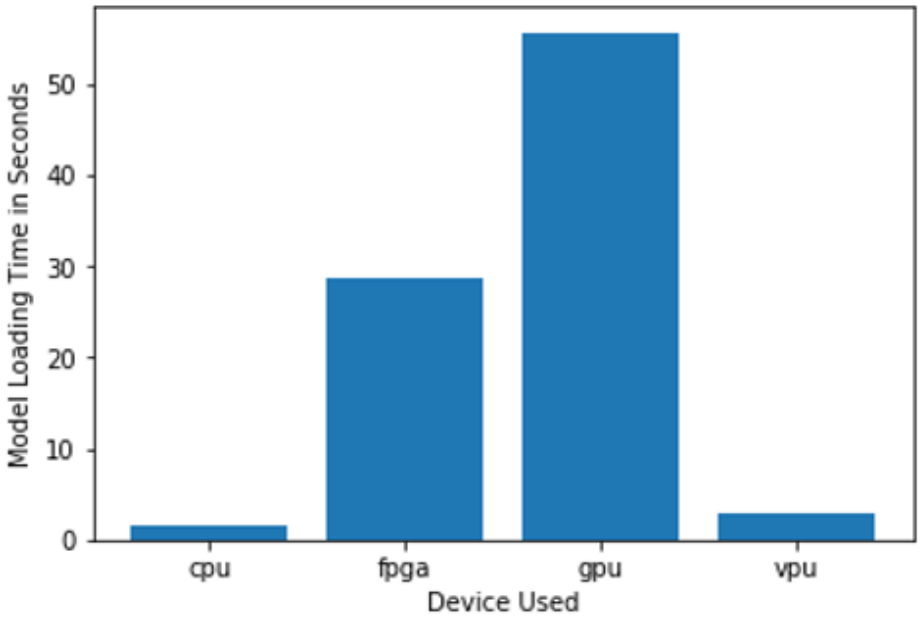
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client would like to save as much as possible on his electric bill.	*A CPU for its high performance requires more power to run whereas, In IGPU the clock rate for the slice and unslice can be controlled separately. This means that unused sections in a GPU can be powered down to reduce power consumption.*
The client does not have much money to invest in additional hardware.	The integrated IGPU comes along with the processor which has a GPU along with the other CPU core.

Queue Monitoring Requirements

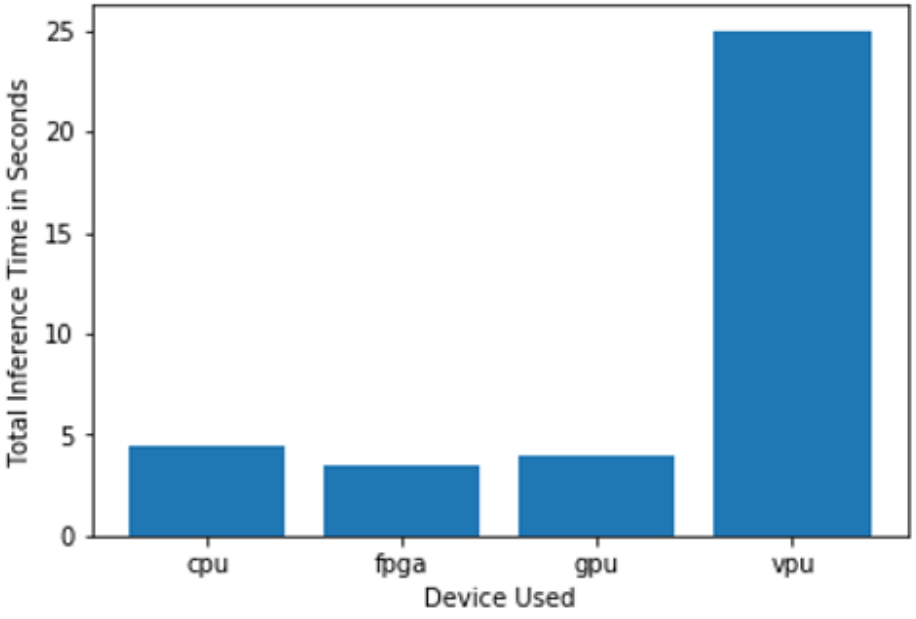
Maximum number of people in the queue	*2 [during normal hours] to 5 [during rush hours]*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

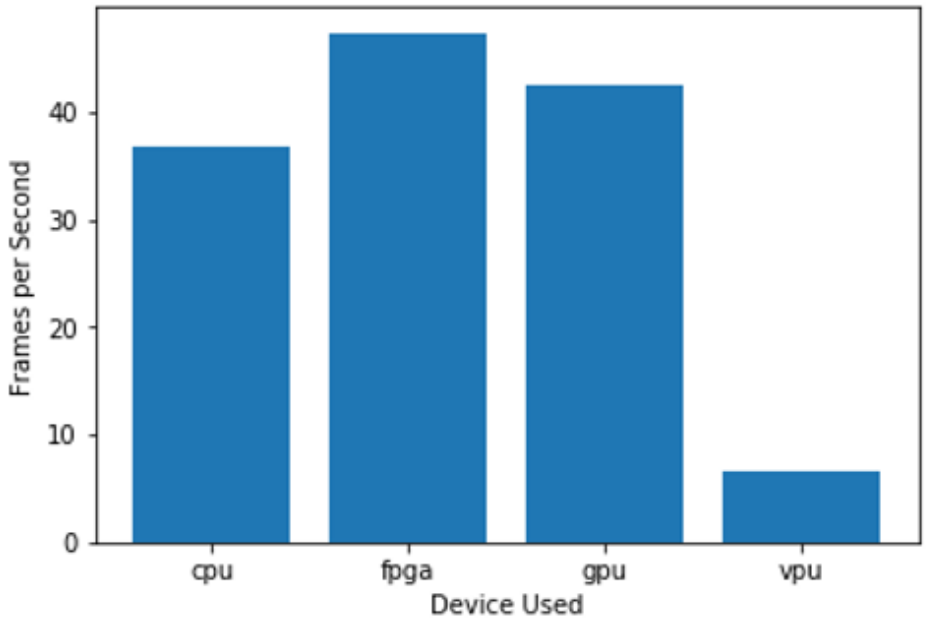
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 much money to invest in additional hardware, which means the client cannot afford to buy a VPU or an FPGA. So, the client must rely on the CPU or IGPU.* • *The client would like to save as much as possible on his electric bill. A CPU for its high performance requires more power to run whereas, In IGPU the clock rate for the slice and unslice can be controlled separately. This means that unused sections in a GPU can be powered down to reduce power consumption. Therefore, the client requirements are satisfied by an IGPU.* • *After testing the results, we can say that the IGPU takes less Inference time when compared to CPU and more than FPGA. But an FPGA does not meet the requirements of the client so, it is ignored.* • *The IGPU also process more Frames per second [FPS] when compared to CPU, but takes more time to load the model than the CPU, It does not satisfy the client's needs. Hence, An IGPU is the hardware that meets the requirements of the client.*

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 is most appropriate for this scenario

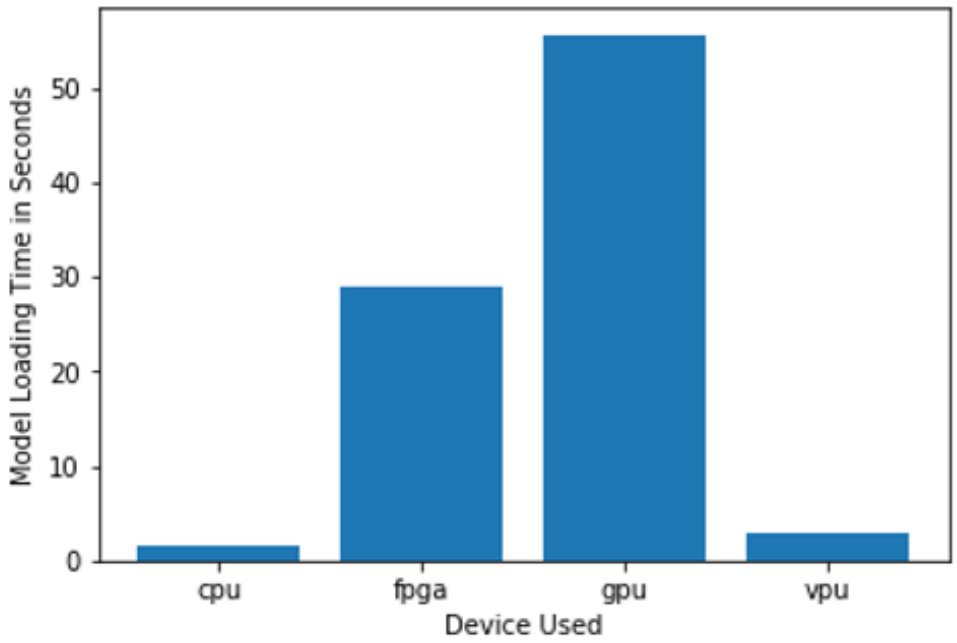
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The CPUs in client's machines are currently being used to process and view CCTV footage for security purposes and no significant additional processing power is available to run inference	*As the client's CPU takes more processing power and the NCS2 can be used to run inference on the models as it requires very low processing power to run inference.*
The client's budget allows for a maximum of $300 per machine, and she would like to save as much as possible both on hardware and future power requirements.	*The VPU or NCS2 is inexpensive compared to other AI accelerators which costs around 70-100$. NCS2 is meant to be a low-power device so that it can be easily deployed at the edge.*

Queue Monitoring Requirements

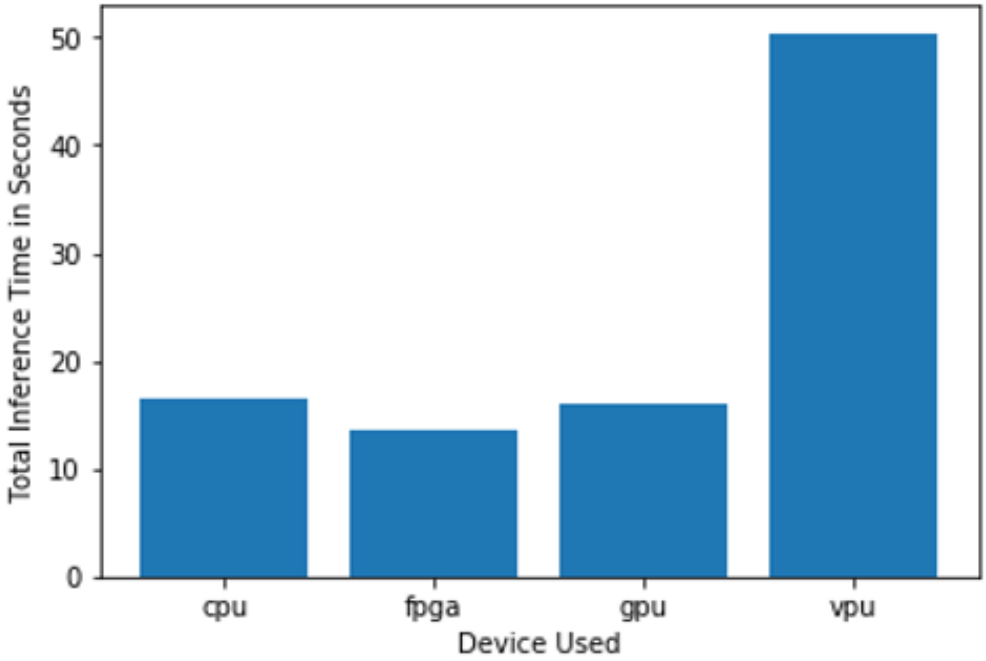
Maximum number of people in the queue	*7 [during non-peak hours] – 15 [during peak hours]*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

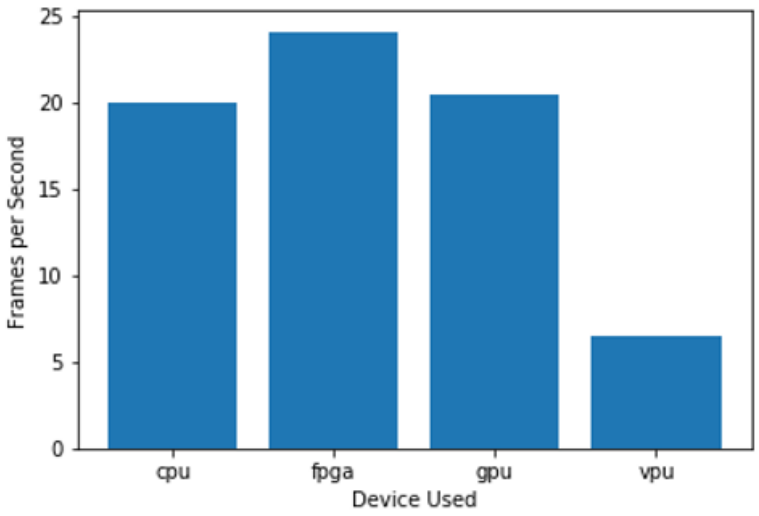
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
• *As the client's budget allows for a maxinmum of 300$ per machine and to save as much as possible both on hardware and future power requirements. A VPU or NCS2 will be the required hardware used for Edge AI systems.* • *The client cannot use FPGA because it is very expensive costing more than 300$ budget, As the client's CPU's currently being used to process and view CCTV footage for security purposes and no significant additional processing power is available to run inference, the NCS2 hardware meets the requirements.* • *After the test results, we can say that VPU's Inference Time is significantly higher when compared to CPU, IGPU and FPGA. But these hardwares [i.e., CPU, IGPU and FPGA] does not meet the client's requirements as FPGA is very expensive.* • *The VPU model loading time is less than the FPGA and IGPU, but more than CPU. The CPU's does not meet the requirements because it takes a lot power to process and run the inference on models.* • *The VPU reads less number of frames per second [FPS] than CPU,IGPU and FPGA which does not meet client's requirement. Therefore, The VPU or NCS2 is considered to be the required hardware which meets all the requirements of the client.*