Choose the Right Hardware

Proposal Template by Manasse Ngudia

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

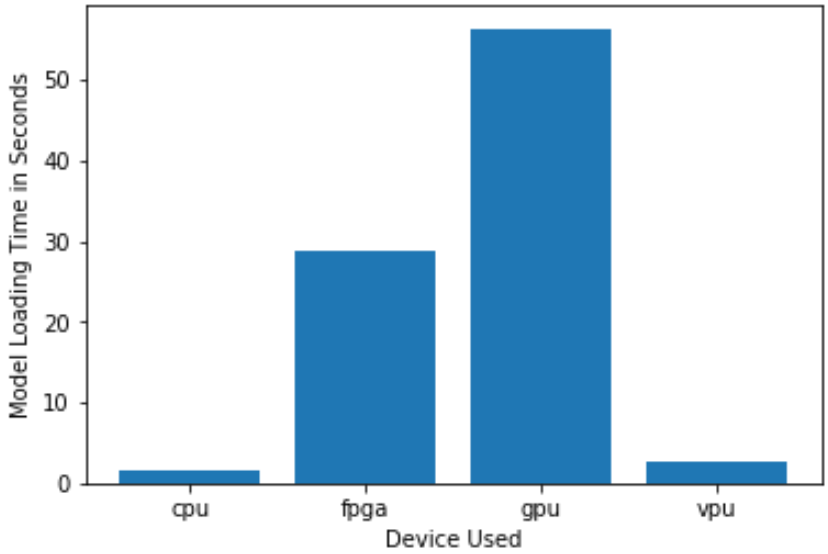
Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
The client needs a system that can last for at least 5-10 years	**Long Lifespan.** FPGAs have a long lifespan. For example, FPGAs that use devices from Intel's Internet of Things Group have a guaranteed availability of 10 years, from start of production.
The client needs a system to be flexible so that it can be reprogrammed and optimized	**Flexibility.** FPGAs are flexible in a few different ways: They are field-programmable; they can be reprogrammed to adapt to new, evolving, and custom networks . Various precision options (FP16, 11 and 9 bit) are supported—allowing developers a balance between speed and accuracy. 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.

Queue Monitoring Requirements

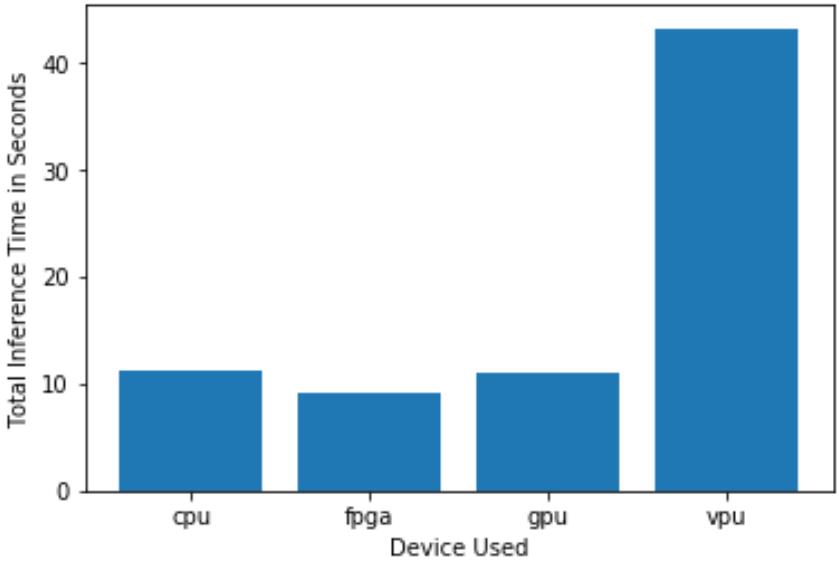
Maximum number of people in the queue	*5*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

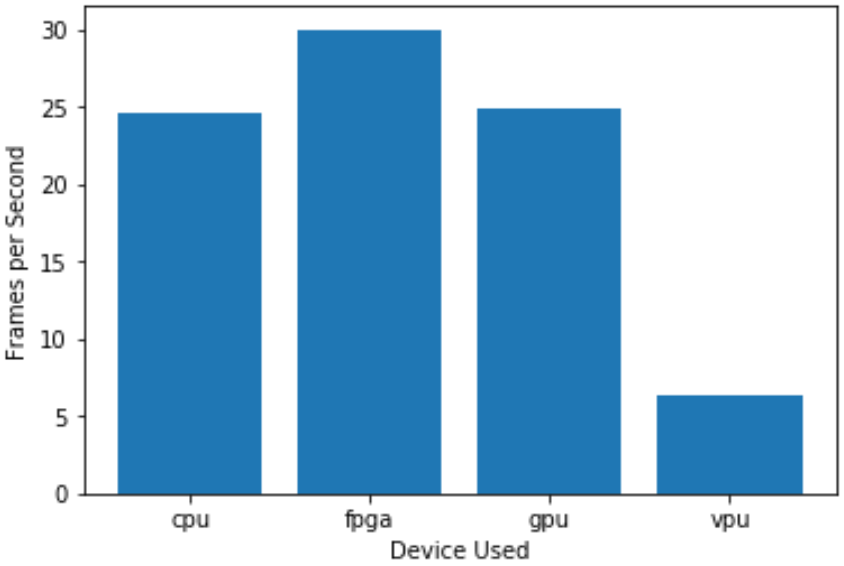
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 implement a quality system, which still represents a significant investment, and would ideally like it to last at least 5 to 10 years. FPGAs have a long lifespan. For example, FPGAs that use Intel's "Internet of Things" group has a 10-year guaranteed availability from the start of production.* ❖ *The client needs the system to be flexible so that it can be reprogrammed and optimized. FPGAs are field programmable; they can be reprogrammed to adapt to new, scalable and customized networks* ❖ *The client needs the system to be flexible so that it can be reprogrammed and optimized. FPGAs are field programmable; they can be reprogrammed to adapt to new, scalable and customized networks* ❖ *The customer wants a system that must be able to make inferences about the video stream very quickly. As we can see from the results of the above tests, FPGAs have the fastest inference compared to other devices, and meets the customer's requirements.* ❖ *The 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 production line. As shown by the results of the above tests, the FPGA indicates nearly 30 FPS compared to other devices and meets customer requirements* ❖ *Although CPUs and VPUs take much less time to load the model based on test results, these devices do not meet the customer's requirements. While the FPGAs take up to 30 seconds to load the model, they meet all customer requirements and be 100% on time, which means that they can operating continuously 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

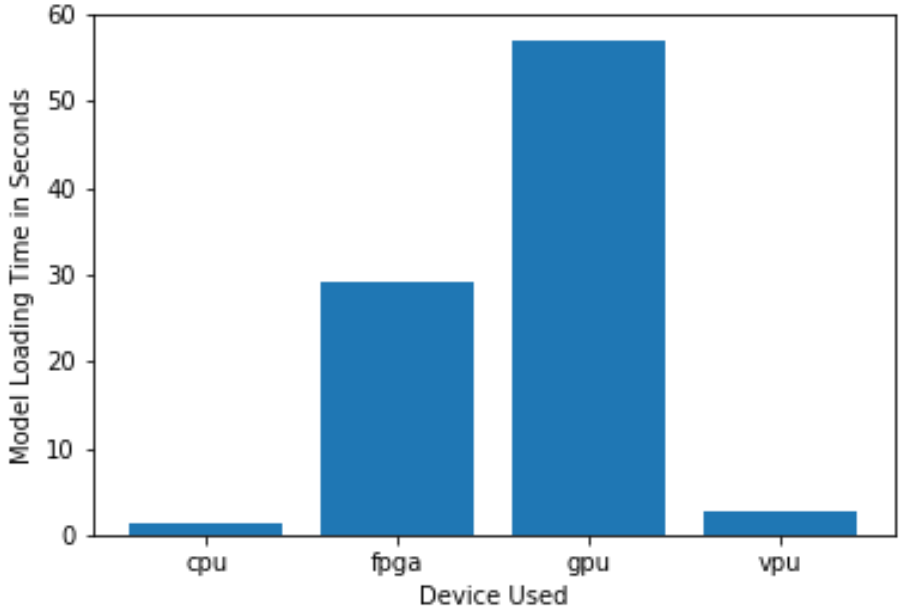
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client does not have much money to invest in additional hardware	An **integrated GPU (IGPU)** is a GPU that is located on a processor alongside the CPU cores and shares memory with them. The client already has modern computers equipped with an Intel i7 and contains GPUs.
The client would like to save as much as possible on his electric bill.	**Configurable Power Consumption.** 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.

Queue Monitoring Requirements

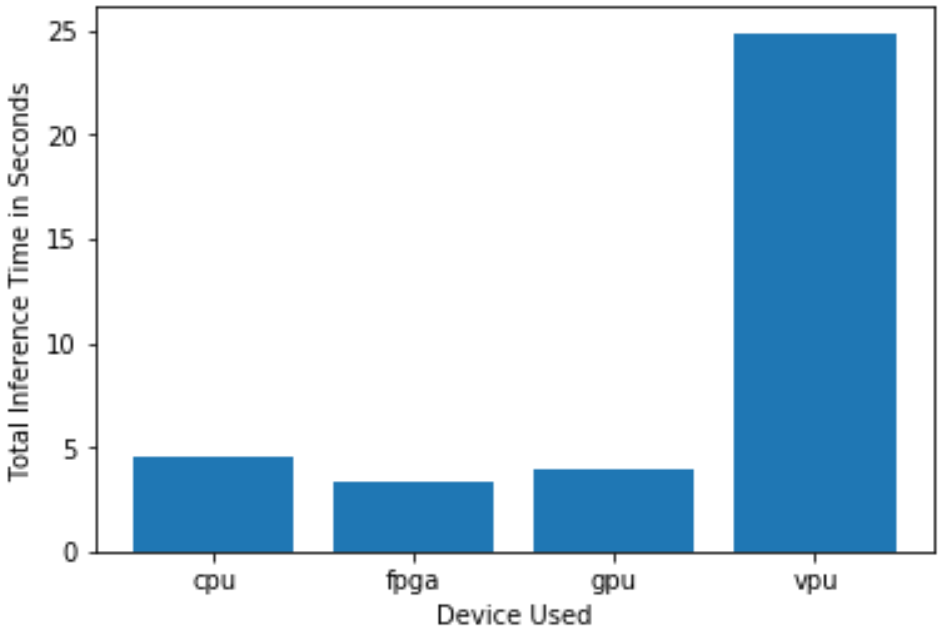
Maximum number of people in the queue	*During normal hours :2 and 5 during rush hours*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

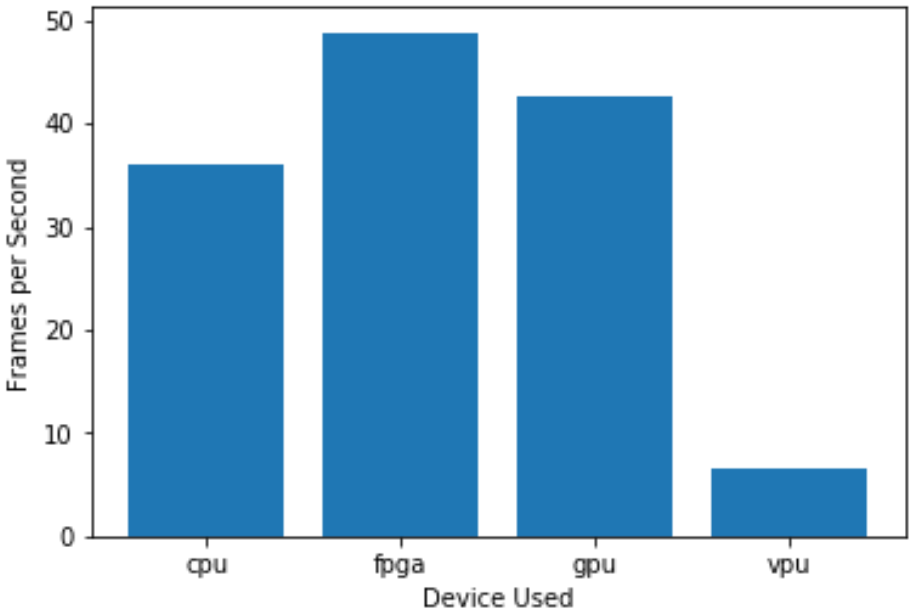
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 a lot of money to invest in additional equipment, which means that he cannot to afford to buy a VPU or FPGA. The client can use his modern computers equipped with an Intel i7 and contains IGPUs.* ❖ *The customer wants to save as much as possible on their electricity bill. A CPU for high performance requires more power to operate while in the IGPU the clock frequency for slicing and chopping can be controlled separately. This means that unused sections of a GPU can be turned off to reduce power consumption. Customer needs are therefore met by an IGPU.* ❖ *After testing the results, it can be said that the IGPU takes less inference time than the CPU and the more than the FPGA. But an FPGA does not meet the customer's requirements, so it is ignored.* ❖ *The IGPU also processes more frames per second [FPS] than the CPU, but takes longer to load. the model that the CPU, it does not meet the customer's needs. Therefore, a CPU is the hardware that meets the customer's requirements.*

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

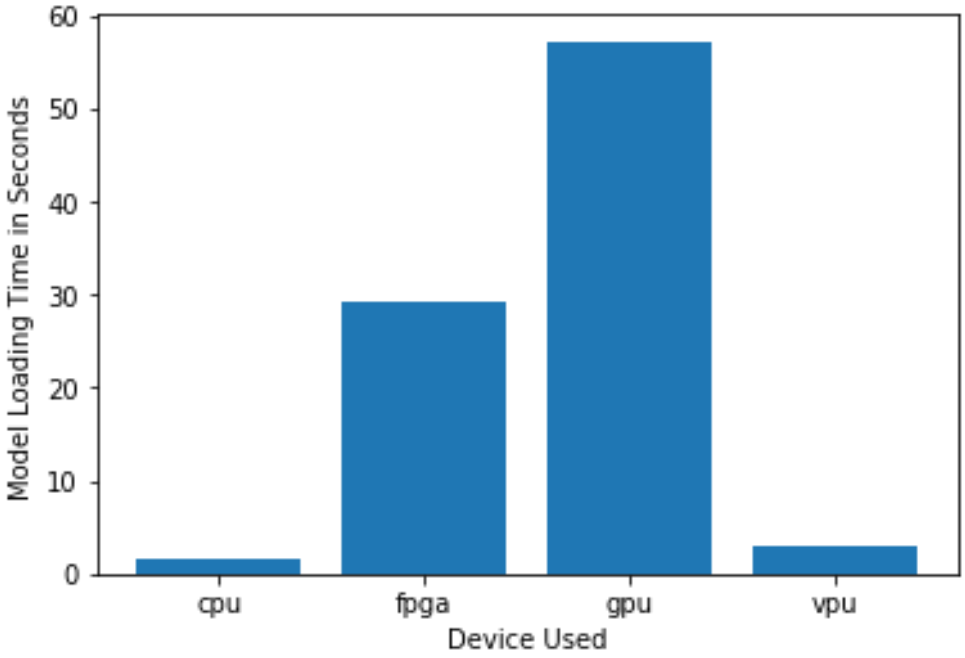
Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
The client's budget allows for a maximum of $300 per machine, and she would like to save as much as possible both on hardware and future power requirements.	VPUs orNCS2 are small, low-cost, low-power devices that can dramatically improve the performance of a system without the need to upgrade the other hardware. It's inexpensive compared to other AI accelerators which costs around 70-100$. NCS2 is meant to be a low-power device so that it can be easily deployed at the edge.
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	*The client's CPU requires more processing power and the NCS2 can be used to make inferences about the models because it requires very low processing power to operate inference.*

Queue Monitoring Requirements

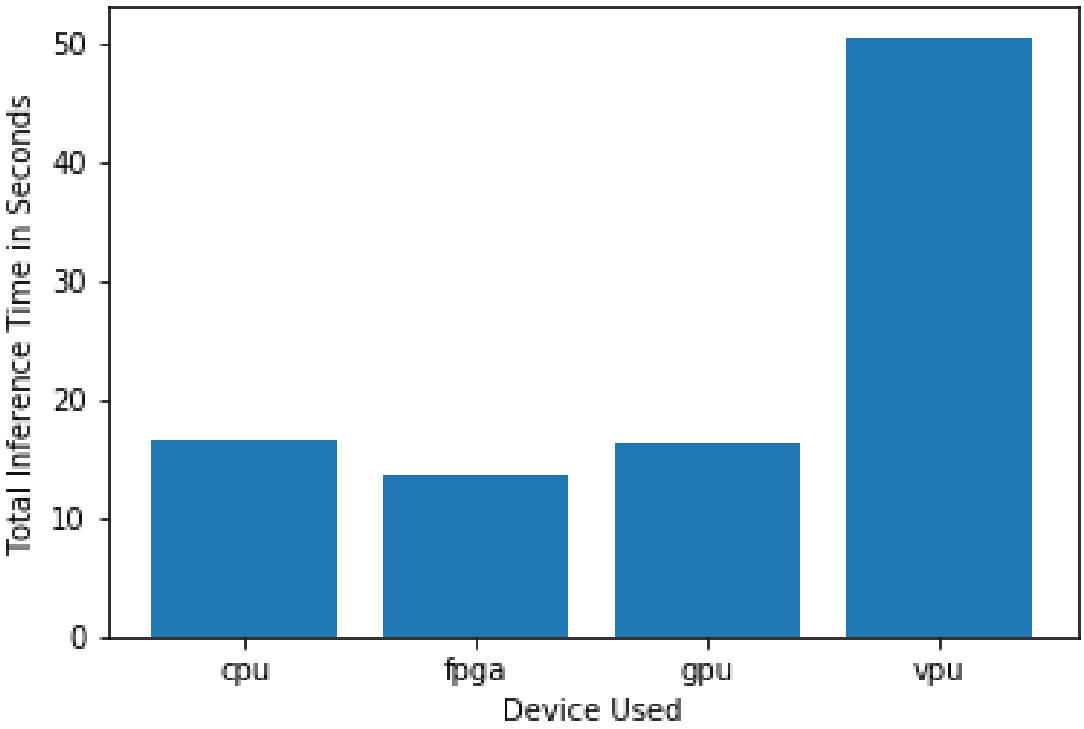
Maximum number of people in the queue	*During non-peak hours it 7 and 15 during the peak hours*
Model precision chosen (FP32, FP16, or Int8)	*FP16*

Test Results

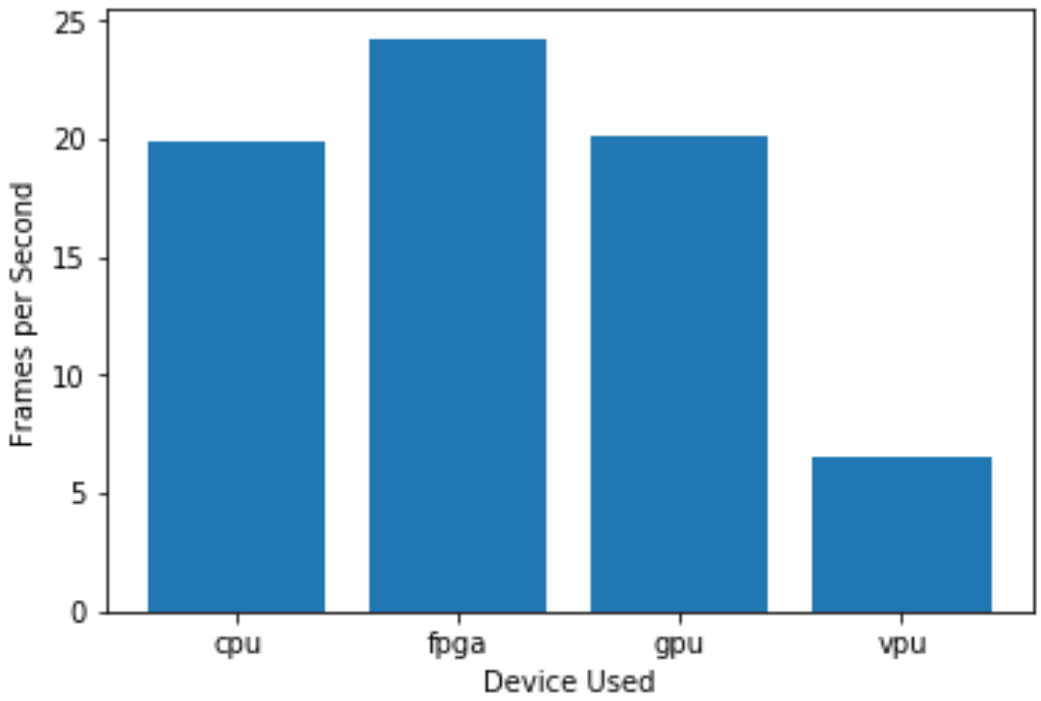
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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 to reach a maximum of $300 per machine and to save as much as possible at the same time on equipment and future energy needs. A VPU or NCS2 will be the necessary hardware used for the Edge AI systems.* ❖ *The client cannot use the FPGA because it is very expensive and costs more than $300 in budget, as the The CPU is currently used to process and view CCTV images for security purposes and has no Additional processing power is available to perform the inference, the NCS2 hardware meets the requirements.* ❖ *Concerning the test results, it can be said that the inference time of the VPU is significantly higher than that of the CPU, IGPU and FPGA. But this hardware like CPU, IGPU and FPGA does not meet the client's requirements since the FPGA is very expensive.* ❖ *The VPU reads fewer frames per second [FPS] than the CPU, IGPU and FPGA, which does not meet the customer's requirement. Therefore, The VPU or NCS2 is considered to be the required hardware which meets all the requirements of the client.* ❖ *The VPU model loading time is less than the FPGA and IGPU, but more than CPU. The CPU's does not meet the requirements since it takes a lot power to process and run the inference on models.*