Choose the Right Hardware

Proposal Template by David Francisco Dávila Ortega

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)
[TODO: Type your answer here] I believe that Naomi Semiconductors requirements would be met with a FPGA.

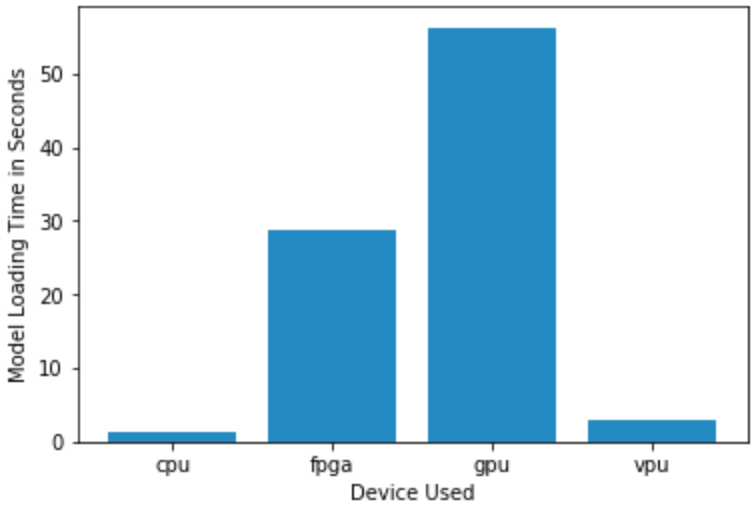
Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
Example requirement: The client requires a tiny device to be connected to their CPU—and their budget is only about $100 for each device.	*Example explanation:* VPU or NCS2 is only about 27.40 mm in size and would fit in the price range.
[TODO: Type your answer here] The client want to invest for equipment that “lasts for at least 5-10 years”.	*[TODO: Type your answer here]* FPGAs have a long lifespan - FPGAs that use devices from Intel’s Internet of Things Group have a guaranteed availability of 10 years, from start of production.
[TODO: Type your answer here] “Workers alternate shifts to keep the floor running 24 hours a day”	*[TODO: Type your answer here]* *FPGAs are designed to have 100% on-time performance, meaning they can be continuously running 24 hours a day, 7 days a week, 365 days a year. They are also able to function over a wide range of temperatures, from 0° C to 60° C. This means that FPGAs can be deployed in harsh environments like factory floors and still perform optimally.*
[TODO: Type your answer here] “Naomi Semiconductors has plenty of revenue to install a quality system”	*[TODO: Type your answer here]* *Given all the features that* FPGAs present they are quite pricey, surpassing the 1000 USD barrier/each.

Queue Monitoring Requirements

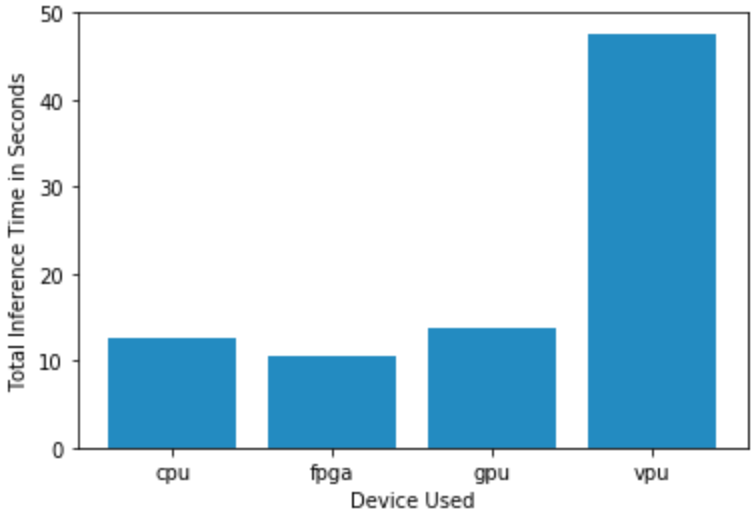
Maximum number of people in the queue	*[TODO: Type your answer here]* Mr. Vishwas would like the image processing task to be completed five times per second.
Model precision chosen (FP32, FP16, or Int8)	*[TODO: Type your answer here—choose from]* FP16

Test Results

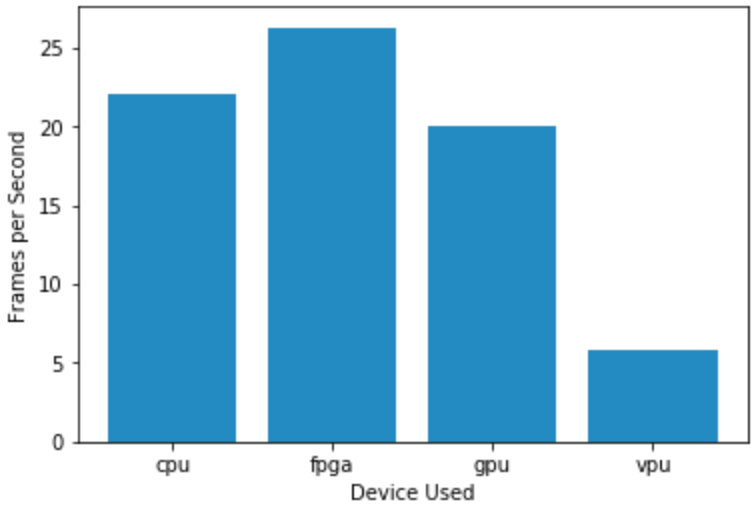
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
[TODO: Type your answer here] By looking at the results I conclude that the best choice for this scenario would be the **use of a FPGA** since it shows the slowest inference time and processes more than 25 fps, maybe the time to load the model is not so low (28 s) but taking into account all the client requirements previously discussed, a FPGA is most likely the best fit. Here I used the Heterogeneous Plugin to run inference on both the FPGA and the CPU as a fallback device.

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)
[TODO: Type your answer here] I believe that PriceRight Singapore requirements would be met with a VPU.

Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
Example requirement: The client requires a tiny device to be connected to their CPU—and their budget is only about $100 for each device.	*Example explanation:* VPU or NCS2 is only about 27.40 mm in size and would fit in the price range.
[TODO: Type your answer here] *"Mr. Lin would like to save as much as possible on his electric bill".*	*[TODO: Type your answer here]* *VPUs show* Low power consumption - e.g. The processor in the NCS2 (the Myriad X) has a very low power consumption of only 1-2 watts.
[TODO: Type your answer here] The store counts with Intel i7 core processor PCs only used for minimal tasks that are not computationally expensive.	*[TODO: Type your answer here]* The NCS2 is meant to be a low-power device so that it can be easily deployed at the edge; however, one drawback of this is that it cannot process as many frames per second (FPS) as some other devices.
[TODO: Type your answer here] *"Mr. Lin does not have much money to invest in additional hardware"*	*[TODO: Type your answer here]* Compared to other AI accelerators, the NCS2 is an inexpensive option, typically costing around $70 to $100.

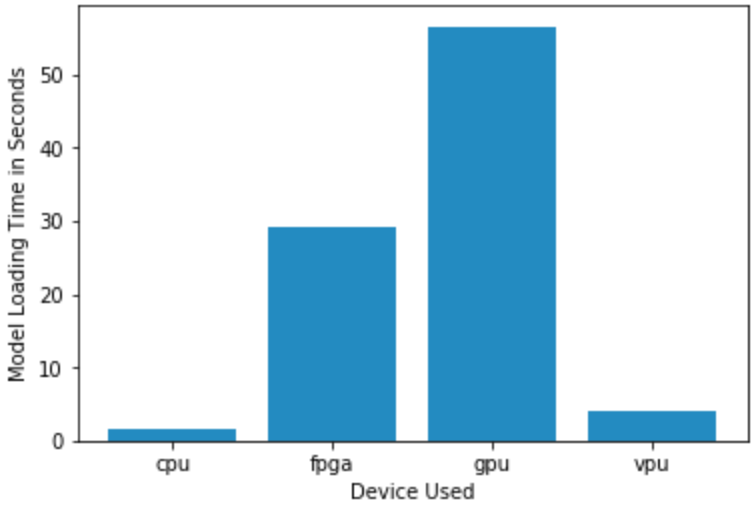
Queue Monitoring Requirements

Maximum number of people in the queue	*[TODO: Type your answer here]* The total number of people in the checkout queue ranges from an average of 2 per queue (during normal daily hours) to 5 per queue (during rush hours).

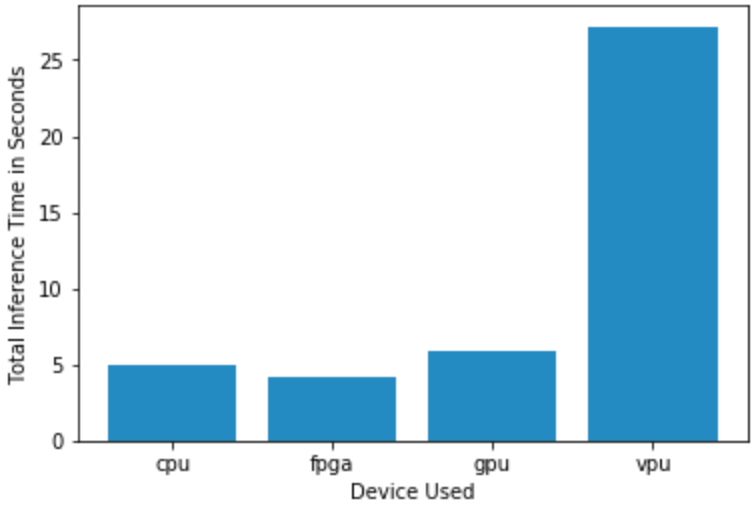
	The total number of people in the checkout queue ranges from an average of 2 per queue (during normal daily hours) to 5 per queue (during rush hours).
Model precision chosen (FP32, FP16, or Int8)	*[TODO: Type your answer here—choose from]* FP16 only

Test Results

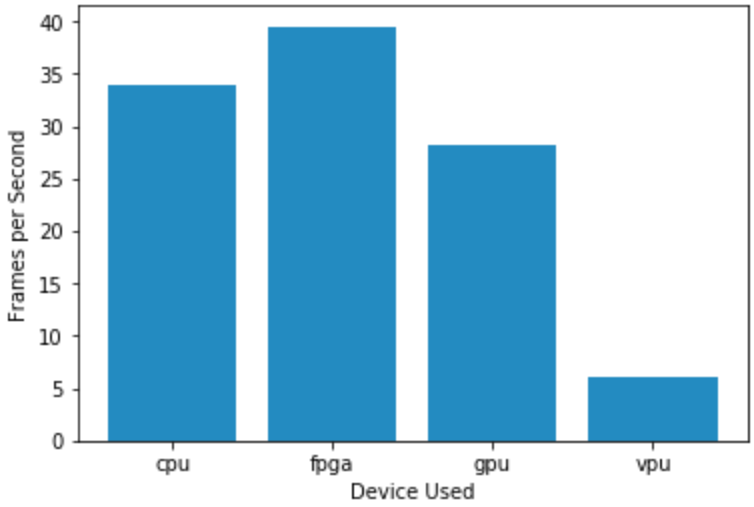
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
[TODO: Type your answer here] By looking at the results I conclude that the best choice for this scenario would be the **use of a CPU** since it shows the slowest inference time and processes 34 fps, besides the time to load the model is very low (1-2 s). Considering this and all the client requirements previously discussed, a CPU is most likely the best fit. Here I used the Multi-Device Plugin while performing inference with the VPU since the client states that they count with modern computers, each of which has an Intel i7 core processor that currently are only used to carry out some minimal tasks that are not computationally expensive. I also used the Heterogeneous Plugin to run inference on both the FPGA and the CPU as a fallback device.

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)
[TODO: Type your answer here] I believe that Delhi Metro Rail Services requirements would be met with an IGPU.

Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
Example requirement: The client requires a tiny device to be connected to their CPU—and their budget is only about $100 for each device.	*Example explanation:* VPU or NCS2 is only about 27.40 mm in size and would fit in the price range.
[TODO: Type your answer here] 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.	*[TODO: Type your answer here]* The IGPU can share memory with the All-In-One PCs from the Company. GPUs generally can handle a much larger number of processes at once compared to a CPU). Thus, if our data is divided into batches, an IGPU can process multiple batches simultaneously, which can sometimes give a significant boost in performance.
[TODO: Type your answer here] Ms. Leah'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.	*[TODO: Type your answer here]* IGPUs show configurable power consumption - On an 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.
[TODO: Type your answer here]	*[TODO: Type your answer here]*

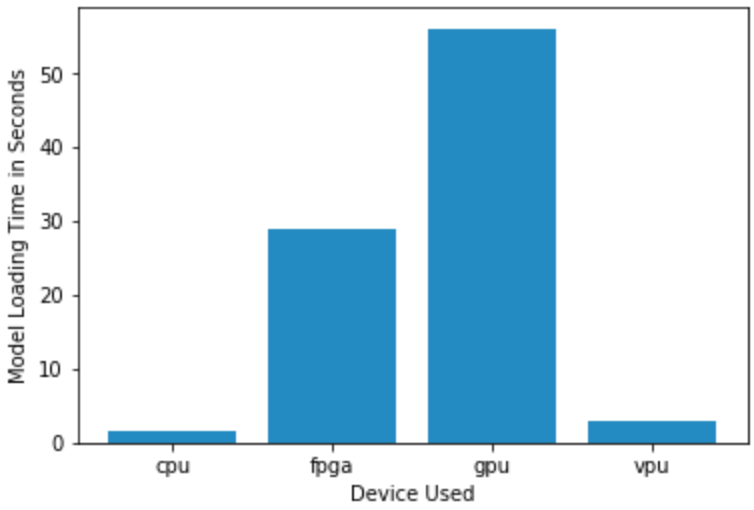
Queue Monitoring Requirements

Maximum number of people in the queue	*[TODO: Type your answer here]* In peak hours they currently have over 15 people on average in a single queue outside every door in the Metro Rail. But during non-peak hours, the number of people reduces to 7 people in a single queue. On office hours there is a train every

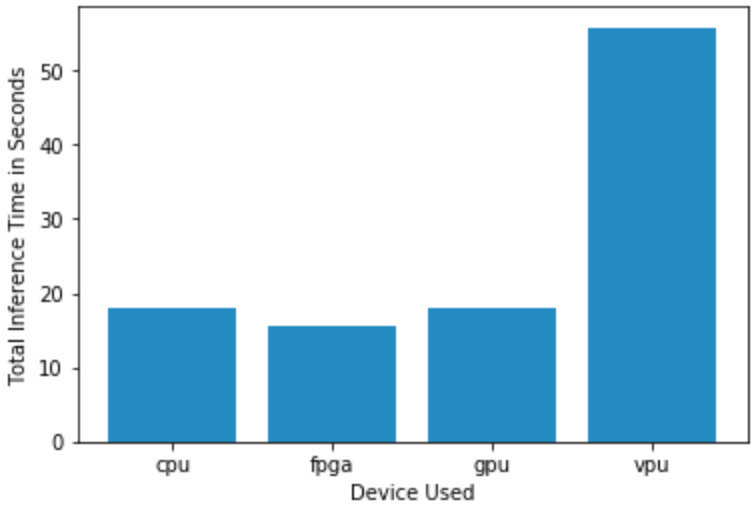
	2 mins. However, on the weekends the time increases to up to 5 mins since some of their drivers work only 5 days a week.
Model precision chosen (FP32, FP16, or Int8)	*[TODO: Type your answer here—choose from]* FP16

Test Results

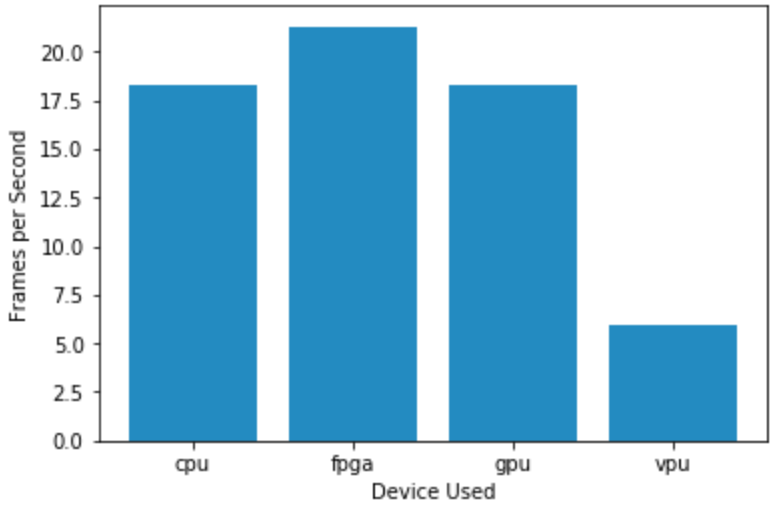
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



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
[TODO: Type your answer here] Here I used the Heterogeneous Plugin to run inference on both the FPGA and the CPU as a fallback device. By looking at the results I conclude that the best choice for this scenario would be the **use of an IGPU** since it shows a slow inference time (~20 s) and processes 18 fps, perhaps the only drawback would be regarding the time taken to load the model (< 50 s). Considering this and all the client requirements previously discussed, a IGPU is most likely the best fit.