Choose the Right Hardware

# 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*** *would be most suitable for Manufacturing client mainly because of its durability (lasts for 10 years), ability to handle multiple instructions in parallel and re-programmable capabilities.* |

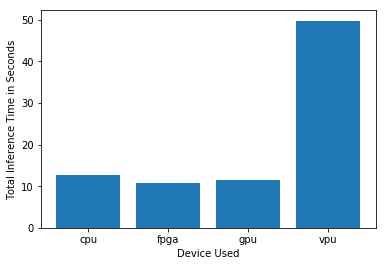
|  |  |
| --- | --- |
| **Requirement Observed**  **(Include at least two.)** | **How does the chosen hardware meet this requirement?** |
| The client has vision cameras installed across multiple belts. Each camera records 30-35 FPS and image processing needs to be 5 tasks per sec. | The key requirement here is parallel processing from multiple feeds from a large network of cameras. Inference speed requirement is low so no IGPU is needed.  FPGA can support large network with parameters as large as 2 million parameters. |
| The client needs the hardware to last long at least 5-10 years | FPGA’s life span is 10 years from the year of production. |
| The client would like to repurpose the system to monitor faulty chips being packaged to improve quality. | Client needs a programmable hardware so FPGA would be best for this scenario as they are flexible and reprogrammable to adapt to new,evolving networks |
| Budget does not seem to be concern here | FPGA although expensive should meet the client’s requirements. |

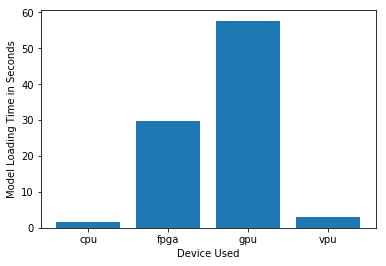
## Queue Monitoring Requirements

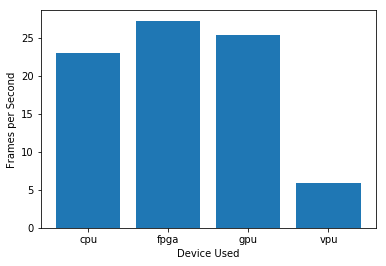
|  |  |
| --- | --- |
| **Maximum number of people in the queue** | There is no specific number so I would expect 20-100 people on the floor. |
| **Model precision chosen (FP32, FP16, or Int8)** | To get the correct headcount, during a shift change, is most important so I would say FP16. FP32 would be more appropriate , FPGA that we have supports only 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).

**

**

**

## 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** |
| Client needs a hardware that they can re-purpose and re-programme to monitor Faulty packaging, they have large enough budget and they need a hardware that can last for 10 years. **FPGA** would be most suitable for Manufacturing client . Based on the above graphs, FPGA would be most suitable because of Lowest inference time (10 per sec) and highest FPS. The model load time is pretty high,higher that CPU and VPU however since this is not going to be done frequently FPGA should be considered. |

# 

# 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.

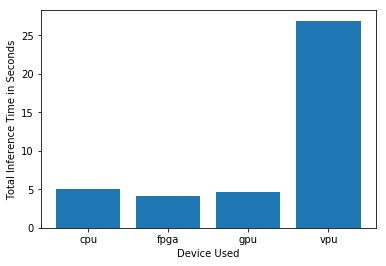
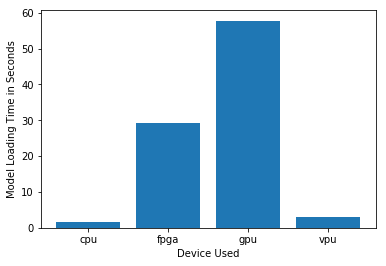
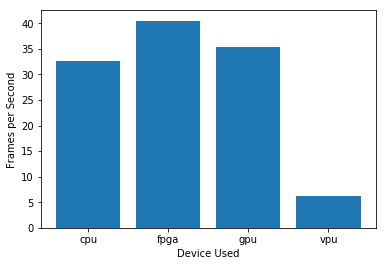
|  |
| --- |
| **Which hardware might be most appropriate for this scenario?**  **(CPU / IGPU / VPU / FPGA)** |
| ***VPU*** *would be most suitable for Retail client due to low cost and low power consumption needs.* |

|  |  |
| --- | --- |
| **Requirement Observed**  **(Include at least two.)** | **How does the chosen hardware meet this requirement?** |
| The client requires a low cost low power device to be connected to their CPU. | VPU or NCS2 would fit in price range as they cost less than $100 an consume power less than 50W |
| The client already has underutilized Intel i7 core processor | CPU can handle additional computation if needed. So would recommend using CPU with VPU. |

## 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).**

## 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** |
| ***VPU*** *consumes least power and cost least compared to other hardwares. However VPU takes maximum time to run inferences compared to CPU,GPU and FPGA and processes less frames compared to other hardwares. To improve performance the VPU could offload some of its computation tasks to CPU since the client has CPUs that re not completely utilized.* |

# 

# 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.

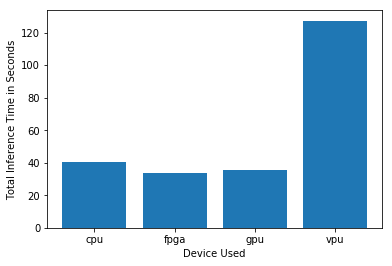
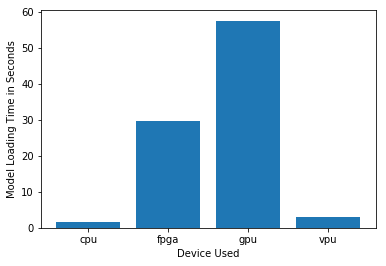
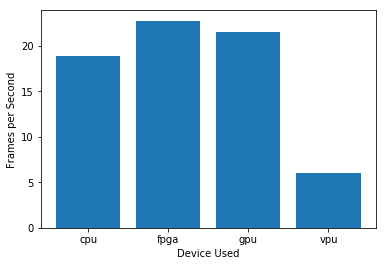
|  |
| --- |
| **Which hardware might be most appropriate for this scenario?**  **(CPU / IGPU / VPU / FPGA)** |
| *External GPU cards* |

|  |  |
| --- | --- |
| **Requirement Observed**  **(Include at least two.)** | **How does the chosen hardware meet this requirement?** |
| The client CPU machines are utilized completely. | VPU,NSC2 will not work if we need to fallback on CPU for additional computation. We need a GPU or FPGA |
| Client has multiple CPUs and needs additional hardware for them under a budget of $300 per machine. | *FPGA will be too expensive so dedicated external GPU card for each machine is best choice* |
| There are no limitations to power consumption and the additional hardware is just needed to accomplish one task headcount in queue. | *No requirement for re-programming so external GPU card would be best.* |
| *The client needs hardware with low inference time, every train arrives in 2 minutes during peak hours. So people need to be notified quickly if less congested queues are available.* | *GPUs are best for analyzing video frames and running inference quickly in matter of milliseconds.* |

## Queue Monitoring Requirements

|  |  |
| --- | --- |
| **Maximum number of people in the queue** | *15* |
| **Model precision chosen (FP32, FP16, or Int8)** | *FP32,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).**

## 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** |
| ***External GPUs*** *are perfect for this client as it is within their budget and can be fitted to existing machines to run inference. Client does not have the budget to buy FPGA or have the need for a system that lasts long or that could be reprogrammed. They cannot use VPU because they need to process large number of frames and run inference quickly. Based on the above graphs we can conclude that GPU’s FPS and Inference time is pretty close to FPGA and hence would be a perfect fit to client requirements. Model loading does not happen frequently so High Model Load time for GPU might not cause much impact.* |

# 