# Choose the Right Hardware

This project is one of the requirements for the successful completion of the Intel's Edge AI for IoT Developers Nanodegree Scholarship Program with Udacity. It consists of hardware proposals for 3 different Smart Queue Monitoring System scenarios: Manufacturing, Retail, and Transportation.

## Scenario 1: Manufacturing

### Client Requirements and Potential Hardware Solution

| **Which hardware might be most appropriate for this scenario?<br>(CPU / IGPU / VPU / FPGA)** |
|---|
| FPGA |

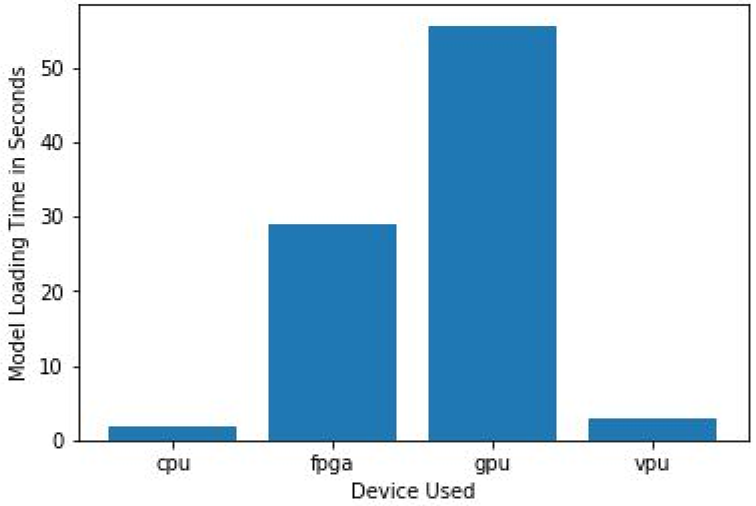
| **Requirement Observed** | **How does the chosen hardware meet this requirement?** |
|---|---|
| The system should have a low latency and be able to run inference quickly to detect flawed chips. | FPGA offers high performance and low latency. |
| The company wants to venture in an existing system: Intel Pentium 4/3000. | The FPGA hardware can be used together with a CPU and even results in better performance. |
| Mr. Vishwas claims that new designs would be created regularly, as such, the system needs to be flexible. | FPGAs are designed for maximum flexibility, that is, they can be re-programmable by loading the appropriate bitstream. |

### Queue Monitoring Requirements

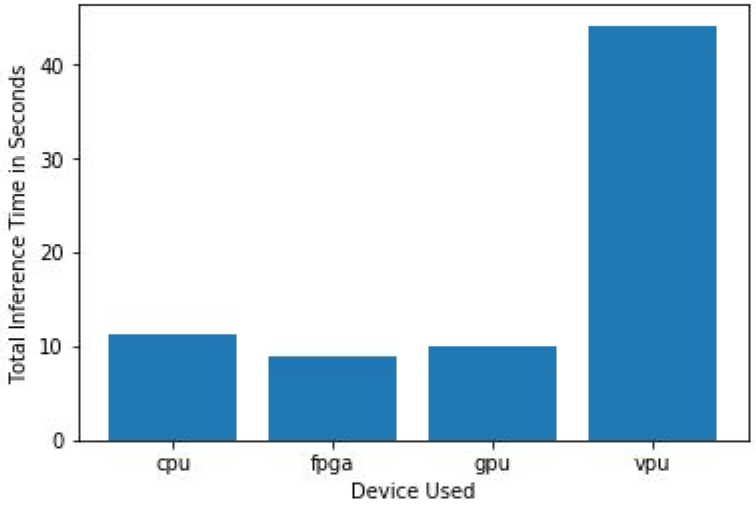
| | |
|---|---|
| **Maximum number of people in the queue** | 2 |
| **Model precision chosen (FP32, FP16, or Int8)** | FP16 |

### Test Results

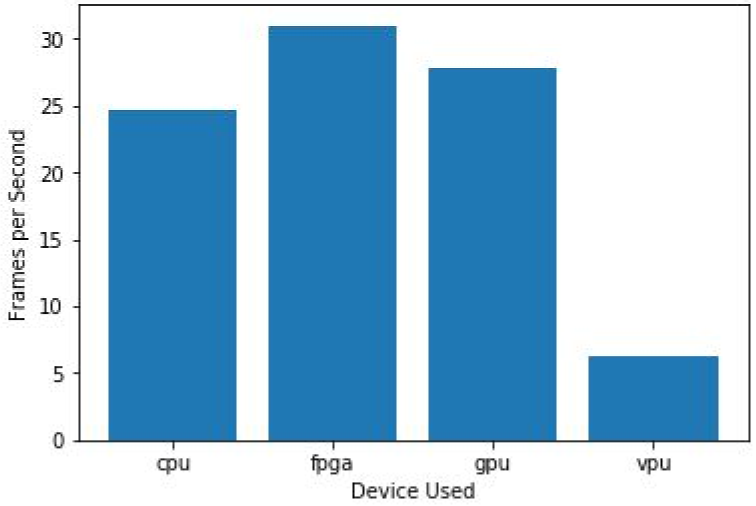
Performance of each device would be compared based on three key metrics of an Edge AI system: Model Loading Time, Inference Time, and Frames Per Second. Below are images from the analysis:



***Model Load Time***



***Inference Time***



***FPS***

## Final Hardware Recommendation

| Write-up: Final Hardware Recommendation |
|---|
| BEST CHOICE: FPGA<br>Naomi Semiconductors is a manufacturing company known for its industrial-grade standard in producing semiconductor chips. Though the company made a good revenue in the previous year, it hopes to achieve maximum revenue with an already existing system - an Intel Pentium 4/3000. This implies that the company does not intend to invest in purchasing a new CPU/iGPU. Mr. Vishwas, the company's representative also reports that the system should be able to fit into new designs that would be created subsequently.<br>Test results show that the FPGA hardware which was the proposed hardware pre-test, performs relatively okay, and best among all the options. It results in the least inference timing which improves the system's performance in detecting failed chips. Finally, the FPGA as a re-programmable system would support any new design by the company in the future. |

# Scenario 2: Retail

## Client Requirements and Potential Hardware Solution

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

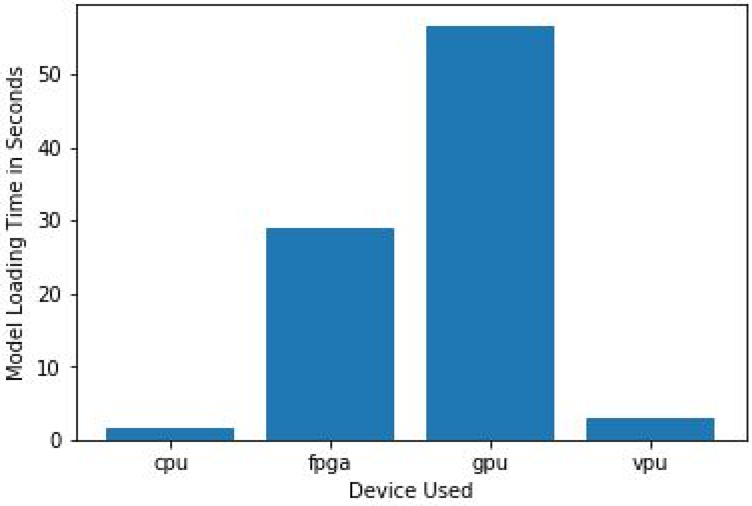
| Requirement Observed | How does the chosen hardware meet this requirement? |
|---|---|
| The average wait time at checkout during weekdays and weekends is 230 secs and about 350 - 400 secs respectively. | The IGPU offers reduced latency and improved performance. |
| The client hopes to use an already existing system - intel core i7 processor - that is currently subjected to minimal computational tasks. | The system has multiple CPUs/cores that could be used. |
| Mr. Lin further posits that the company has no money to invest in additional hardware. | The Integrated GPU typical of every Intel processor could be leveraged in the need of extra performance. |

## Queue Monitoring Requirements

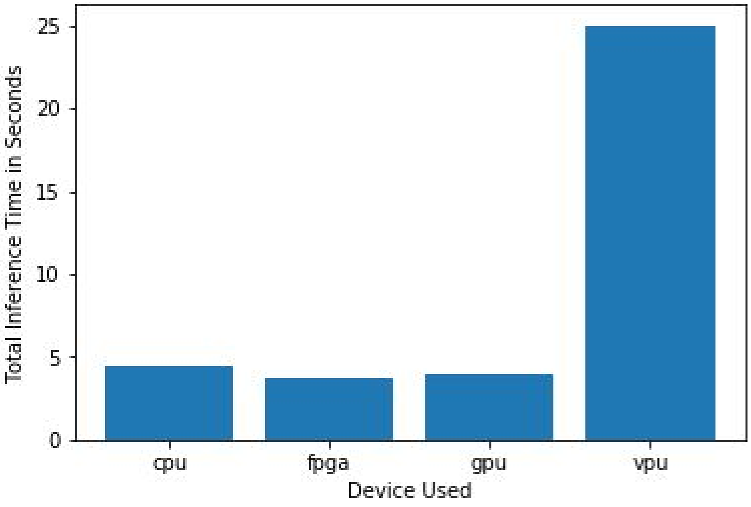
| | |
|---|---|
| **Maximum number of people in the queue** | 2 |
| **Model precision chosen (FP32, FP16, or Int8)** | FP16 |

## Test Results

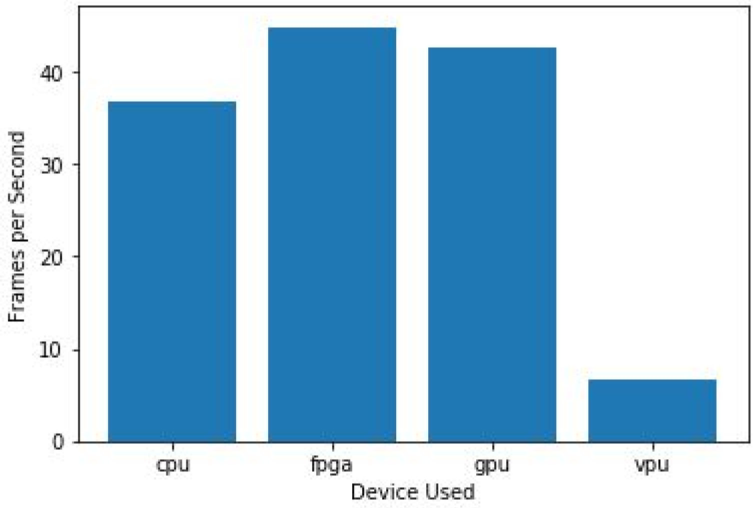
Performance of each device would be compared based on three key metrics of an Edge AI system: Model Loading Time, Inference Time, and Frames Per Second. Below are images from the analysis:
(NOTE: The 'gpu' shown in the plot represents 'IGPU' hardware.)



***Model Load Time***



***Inference Time***



***FPS***

## Final Hardware Recommendation

| **Write-up: Final Hardware Recommendation** |
|---|
| BEST CHOICE: IGPU<br><br>PriceRight Singapore is a small retail store with regular customers. The store manager, Mr. Lin, reports that when there is congestion at checkout, profit obtained is lower than expected. This implies the need for a system that redirects customers to less-congested queues. In addition to this, the company does not intend to invest in additional hardware. Thus, the need to maximize the already existing system.<br><br>Unarguably, the FPGA hardware appears to be the best choice from the test result with its attractive performance in all three metrics. But, the no-additional hardware constraint rules it out for contention. Fortunately, the proposed hardware pre-test, IGPU, comes second best. Its low inference timing and high FPS implies image processing and inference would be done in quick time. This inevitably reduces congestion. The towering model loading time of the IGPU does not affect inference, because the model must have been loaded and working beforehand. |

# Scenario 3: Transportation

## Client Requirements and Potential Hardware Solution

| **Which hardware might be most appropriate for this scenario?<br>(CPU / IGPU / VPU / FPGA)** |
|---|
| VPU |

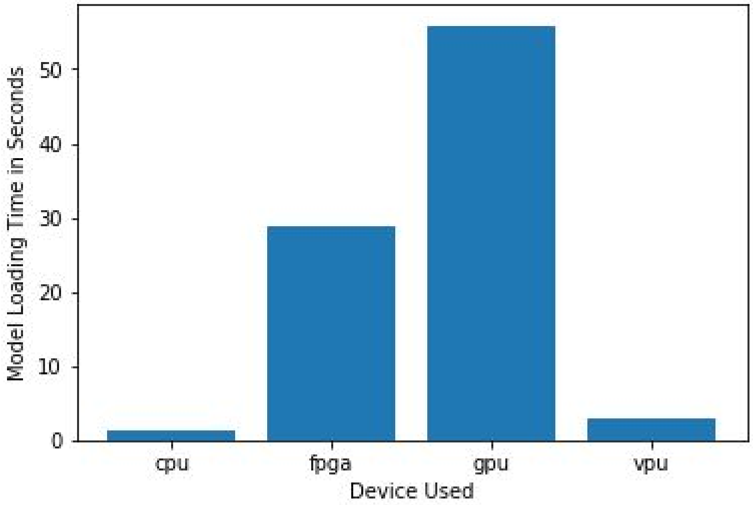
| Requirement Observed | How does the chosen hardware meet this requirement? |
|---|---|
| In peak hours, over 15 people on average in a single queue. | The VPU device can accelerate inference time substantially and offers low latency. |
| The client explains that they are trying to save as much cost as possible. | The VPU (NCS2) is one of the cheapest AI accelerators that can be purchased. |
| Max of $300 is budgeted per machine. | An NCS2 cost just about $100 or less. |
| Ms. Leah confirms that no significant processing power is available to run inference. | The VPU device is known to require low power and is still quite efficient. |

## Queue Monitoring Requirements

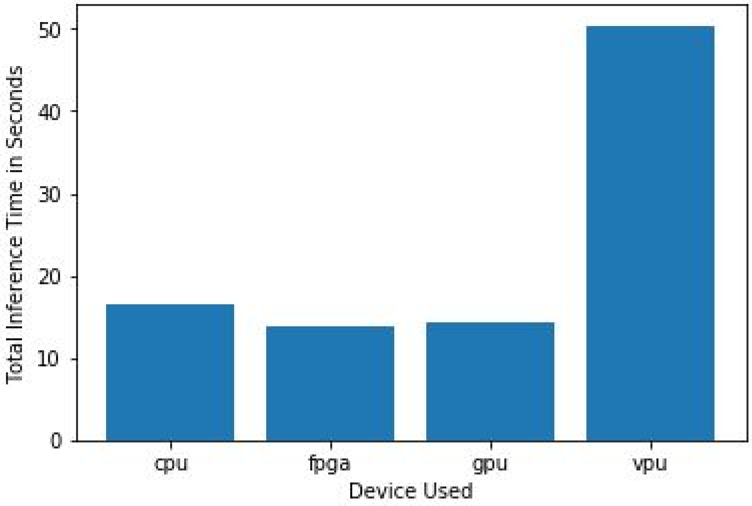
| | |
|---|---|
| **Maximum number of people in the queue** | 4 |
| **Model precision chosen (FP32, FP16, or Int8)** | FP16 |

## Test Results

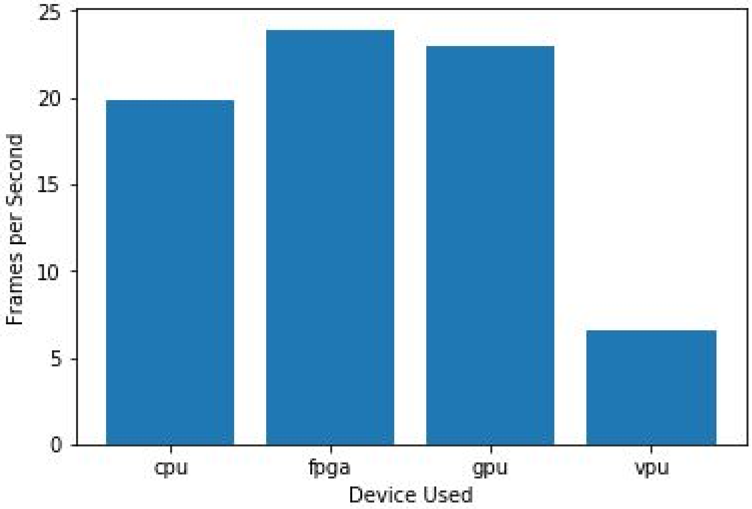
Performance of each device would be compared based on three key metrics of an Edge AI system: Model Loading Time, Inference Time, and Frames Per Second. Below are images from the analysis:
(NOTE: The VPU used in analysis is the Intel's Neural Compute Stick 2)



***Model Load Time***



***Inference Time***



***FPS***

## Final Hardware Recommendation

| **Write-up: Final Hardware Recommendation** |
|---|
| BEST CHOICE: VPU<br><br>Based on the three metrics, the FPGA performs best. However, the available budget does not make it feasible.<br>Conclusively, due to cost and power constraints, the VPU hardware is recommended. The high inference time is as a result of the low power availability. Already, 7 CCTV cameras are connected to all-in-one PCs. Again, only a device like the VPU is acceptable. |