Choosing the Right Hardware

INTEL® EDGE AI FOR IOT DEVELOPERS NANODEGREE

SWASTIK NATH

Scenario 1: Manufacturing

Client Requirements and Potential Hardware Solution

In accordance with the manufacturing scenario an industrial semiconductor manufacturer which follows a 7-step process to build the semiconductors chips. As according to the Engineering team there are several constraints in this complex manufacturing process and typically it should take about 6 to 8 weeks but as the problem it is currently taking 10-12 weeks leading to a reduction of revenue by 30%.

The problematic scenario was identified in the process of Shipping to Customers. This part of the process involves the manual labor of packaging the chips into boxes. There is one particular shop floor—which has two industrial belts—that has shown slower production than the rest.

To help understand and address these issues, the client wants a system to monitor the number of people in the factory line. The factory has a vision camera installed at every belt. Each camera records 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. The client would like the image processing task to be completed five times per second.

Once this productivity problem has been addressed, they would like to be able to repurpose the system to address a second issue. The second issue encountered is that a significant percentage of the semiconductor chips being packaged for shipping have flaws. These are not detected until the chips are used by clients. If these flaws could be detected prior to packaging, this would save money and improve the company's reputation.

To be able to detect chip flaws without slowing down the packaging process, the system would need to be able to run inference on the video stream very quickly. Additionally, because there are multiple chip designs—and new designs are created regularly—the system would also need to be flexible so that it can be reprogrammed and optimized to quickly detect flaws in different chip designs.

While the client has plenty of revenue 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.

Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
*As the scenario explains the client has plenty of resources to spend for a great performing machine, we might suggest the client to opt for an **FPGA coupled with a CPU** for this scenario with HETERO plugin enabled.*

Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
The device will be installed in a factory plane which will be receiving input streams from each belt with 35 fps and the image processing needs to be done five times per second.	*The FPGAs will surely be able to perform better in this scenario, as it will be able to process the input streams faster than any other option.*

The system needs to accelerate the slowed down shipping process by monitoring the number of people involved.	*As we can see from the figures 1, 2, and 3 from below we can see that the overall latency of FPGA in terms of inference time is considerably low and the frames per second is also higher than other options. In terms of the Model loading time, the model needs to be loaded only once, so it is safe to overlook that aspect. As the inference will be happening 5 times per second, the inference time is the most important factor we should be monitoring.*
The system needs to be flexible enough to adapt to the new changes to address other issues.	*When it comes to flexibility, Field Programmable Gate Arrays (FPGAs) is what we need to opt to. In terms of re-programmability, only FPGAs offer the highest level of customization.*
The system needs to be resilient enough to be operational for 5-10 years in a harsh environment like factory plane.	*The Intel FPGA's are supported for up to 10 years from the date of initial release and can work under harsh environments.*

Queue Monitoring Requirements

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

Test Results

We test our application on all four hardware types (CPU, IGPU, VPU, and FPGA), corresponding visualizations showing the comparison are provided below. We have three graphs (for model load time, inference time, and FPS).

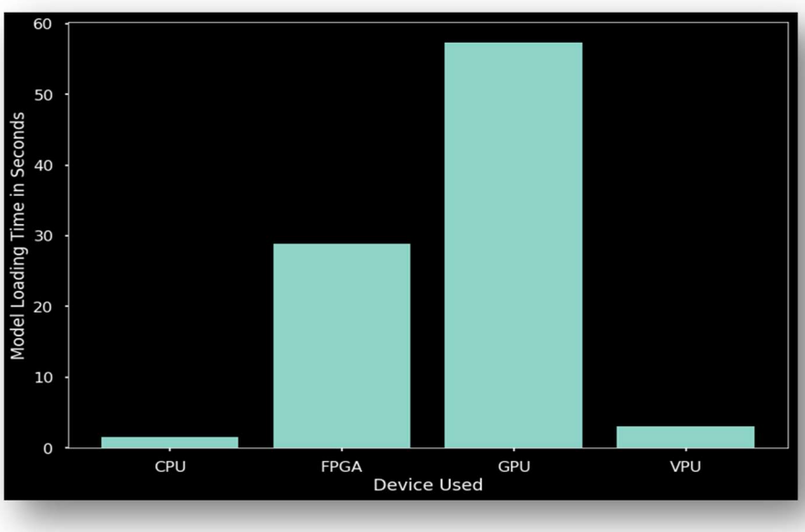

Figure 1: Model Loading Time in Manufacturing Scenario.

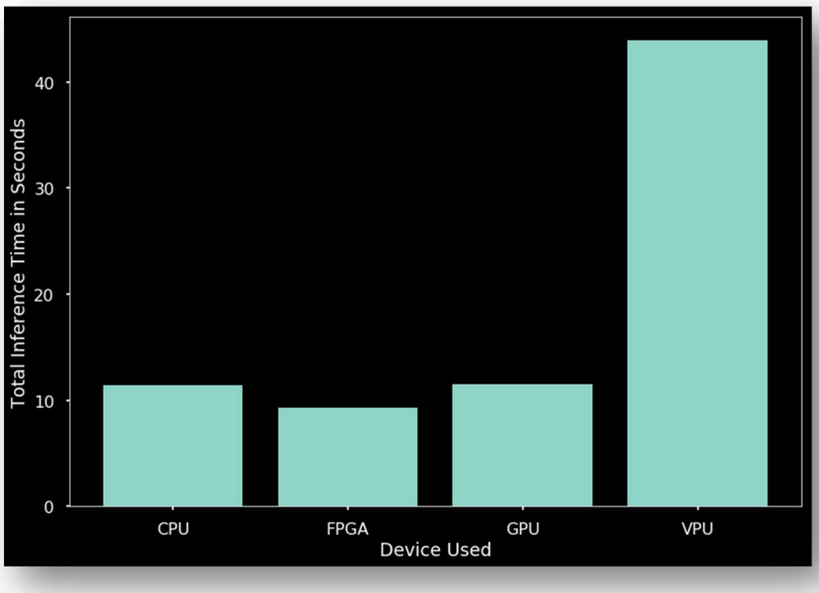

Figure 2: Inference Time in Seconds for Manufacturing scenario.

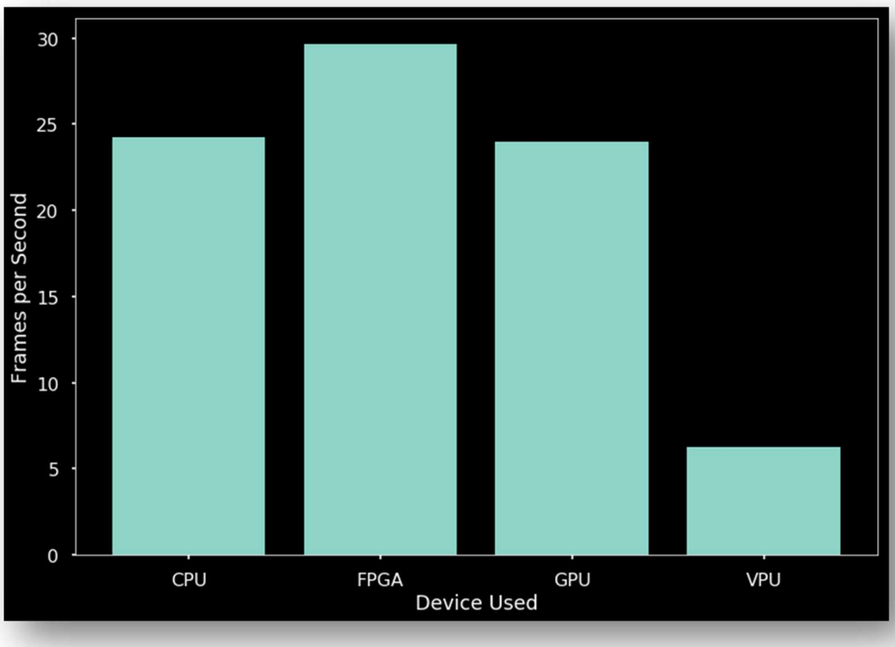

Figure 3: FPS for Manufacturing Scenario.

Final Hardware Recommendation

As we had anticipated, the FPGA coupled with CPU using HETERO plugin does great in terms of overall Inference Time, Frames per Second with the best performance in chart. But in case of Model Load time it takes a few extra seconds, however as per the criteria of the client the inference will be done on frames acquired from multiple cameras five times per second and the model will be loaded only once. So, for the sake of performance and reduction in latency we choose to stick with our preliminary hardware recommendation.

Final Hardware Recommendation
*For the manufacturing scenario we recommend our client to use the **FPGA coupled with CPUs with HETERO plugin** as because it provides faster inference time, higher frames per second and greater flexibility to be reprogrammed.*

Scenario 2: Retail

Client Requirements and Potential Hardware Solution

A Retail outlet chain would like to seek help of Edge AI to help maximize their yearly profits and customer satisfaction. Most of the customers are regulars at the store. Client has seen an average of about 200 people in the store during weekdays. On the weekends, this increases to between 500 and 1000. The maximum number of people visit the store during the holidays. Most customers spend 30-50 mins in the store during a single visit. Out of this, they have an average wait time of 230 seconds at the checkout counters. But on the weekends, the wait time can increase substantially. The average time spent is 40 mins at the store and 350-400 seconds at the checkout line. 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).

It is during rush hours that client has seen wavering sales. When wait times are short and checkout happens smoothly, he sees a jump in their revenue from 6 to 20%. However, if there is congestion at the checkout counter, his profits only go up to 4-5%.

The client believes this problem can be easily solved by directing people to less-congested queues in the store, and they are interested in using an Edge AI system to do so.

Most of the store's checkout counters already have a modern computer, each of which has an Intel i7 core processor. Currently these processors are only used to carry out some minimal tasks that are not computationally expensive. The client employs close to 300 employees, including staff that work in transportation, on the store floor, and at the checkout counter. Although the store's annual sales are $7 million in food alone, the net profit is only about 1.1% of this. The client also believes in giving fair employment and good wages. They pay their staff with proper salaries, along with substantial bonuses twice a year. As a result, the client does not have much money to invest in additional hardware, and also would like to save as much as possible on his electric bill.

Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
*As the client already in possession of powerful **CPUs** we might suggest to stick with the CPUs to perform the inference as it would drastically reduce the costs such as extra hardware, power bills.*

Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
The system should be able to smartly recognize congestions in queue to better maintain customer satisfaction and minimize the time spent by the customer in a queue by directing them to another queue.	*The Intel Core i7 processors are very powerful to perform inference over the accumulated video frames from around the store. As we can see from the figure 4, 5, and 6 the model load time and inference time are significantly lower in this case with significantly higher FPS.*
The system should be able to direct the customers to queues to minimize their waiting time without a delay.	*With CPUs the inference time is significantly lower and the Frames per second is significantly higher thereby leading to minimal delay in directing the customers.*
The system must be able to keep track of large number of people in every video frame per camera for a large amount of time.	*In order to conform to the requirement criteria, the inference time must be low and FPS must be higher to be able to provide real-time decisions.*

The client does not have much resources to spend on the system and would also like to reduce overall cost like power bills etc.	*Deploying with CPUs will help us using the client's underutilized pre-installed hardware without incurring extra hardware or power cost.*

Queue Monitoring Requirements

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

Test Results

We test our application on all four hardware types (CPU, IGPU, VPU, and FPGA), corresponding visualizations showing the comparison are provided below. We have three graphs (for model load time, inference time, and FPS).

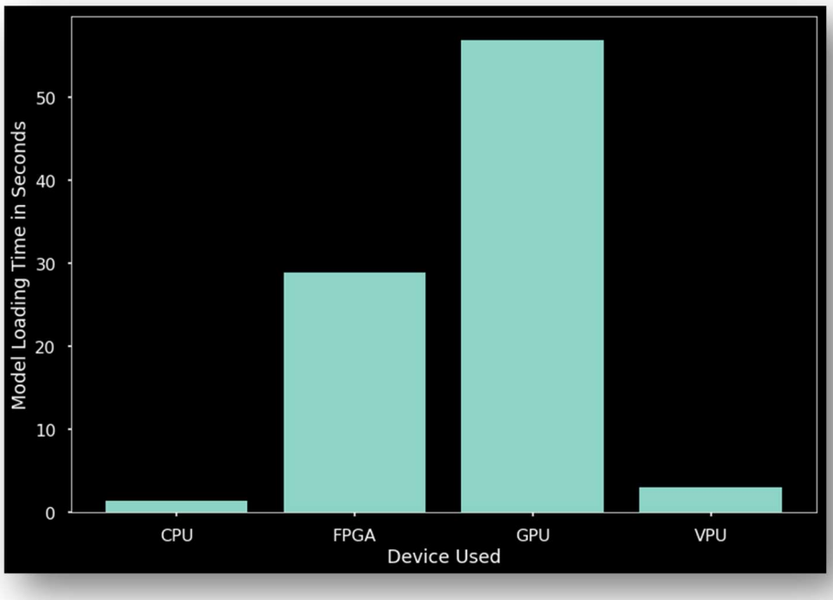

Figure 4: Model Loading Time in Retail Scenario

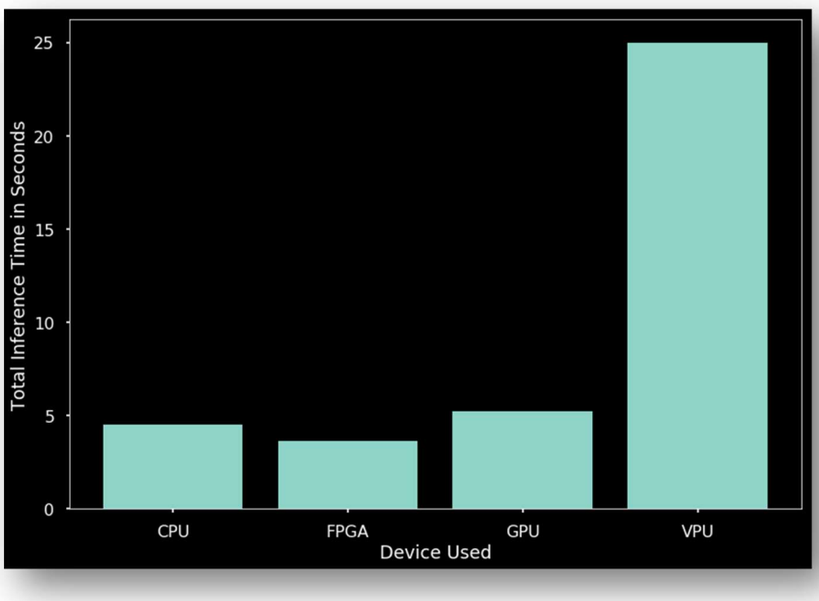

Figure 5: Inference Time in Seconds for Retail scenario.

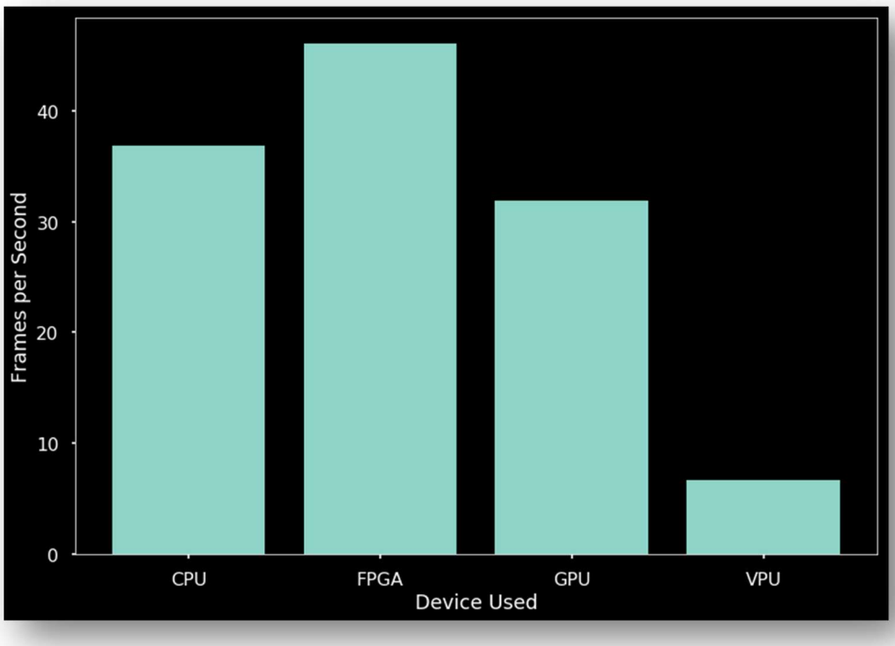

Figure 6: FPS for Retail Scenario.

Final Hardware Recommendation

The FPGAs coupled with CPUs with HETERO plugin actually outperforms the CPUs in terms of lower inference time and higher frames per second. The model is loaded only once across the inference pipeline, so FPGAs were actually a better choice than CPUs. But the cost of FPGA is much higher than the pre-installed CPUs and will incur high power usage. So, the client requirements will not be conformed with FPGAs due to higher price and higher power usage. So, we will stick to CPUs in this scenario.

Final Hardware Recommendation
*We will be using the **CPUs** in this scenario as they are already preinstalled and will not further incur extra charges while providing better utilization of the hardware, lower model load time, inference time and significantly higher frame rates.*

Scenario 3: Transportation

Client Requirements and Potential Hardware Solution

A large-scale underground railway urban passenger transportation service makes around 2700 trips every day and is one of the busiest transportation providers of a nation. During peak hours, some areas of the platform get highly congested, while other areas remain relatively open. In some cases, passengers trying to board in the more congested areas are unable to get on, even though there is space on the train.

Currently, this congestion is handled manually by door operators, who help direct passengers to less congested areas during peak time. The client would like to automate this using an Edge AI system that would monitor the queues in real-time and quickly direct the crowd in the right manner.

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.

They monitor the entire situation with 7 CCTV cameras on the platform. These are connected to closed All-In-One PCs that are located in a nearby security booth. 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. 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.

Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
*In accordance with the client requirements, to save additional cost on hardware and power requirements we might suggest to use the **Integrated GPUs (IGPU)** of the currently installed CPUs.*

Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
The deployed system should be agile enough to perform inference so that large number of people can be directed to form proper queue in real time to avoid congestion.	*From the figures 7,8, 9 we observe that the inference time using IGPU is quite low and the FPS is quite high thereby leading to the deployment real-time. Although the model loading time is high, we can safely overlook this in this scenario because the model is loaded only once in the entire pipeline.*
The system should be able to perform inference on video frames accumulated from multiple sources.	*We can see from the figures that the lower inference time and higher frames per second makes the system real-time.*
The deployed system should be able not to incur additional hardware charges and power usage.	*By using the unutilized Integrated GPUs of the pre-installed systems, we will be saving the client's requirements of reducing power usage and additional hardware charges.*

Queue Monitoring Requirements

Maximum number of people in the queue	8
Model precision chosen (FP32, FP16, or Int8)	*FP32*

Test Results

We test our application on all four hardware types (CPU, IGPU, VPU, and FPGA), corresponding visualizations showing the comparison are provided below. We have three graphs (for model load time, inference time, and FPS).

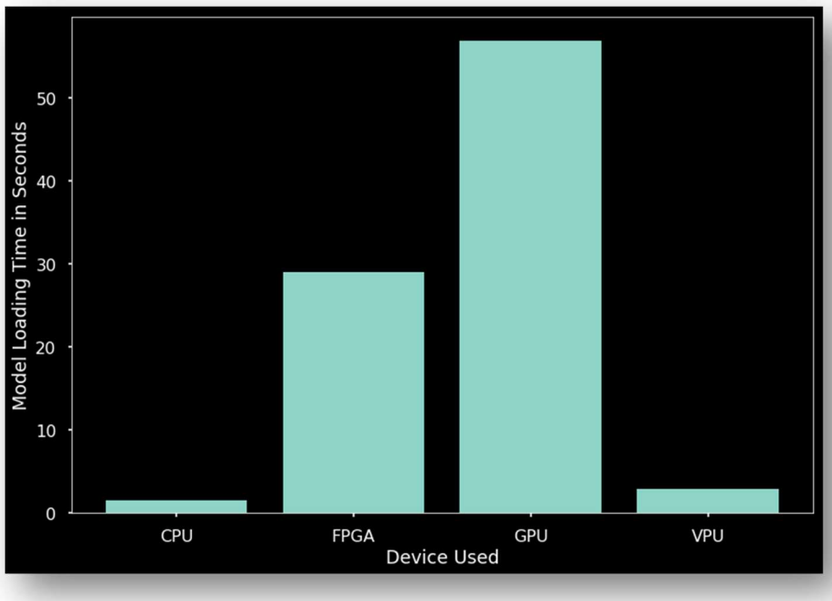

Figure 7: Model Loading Time in Transportation Scenario

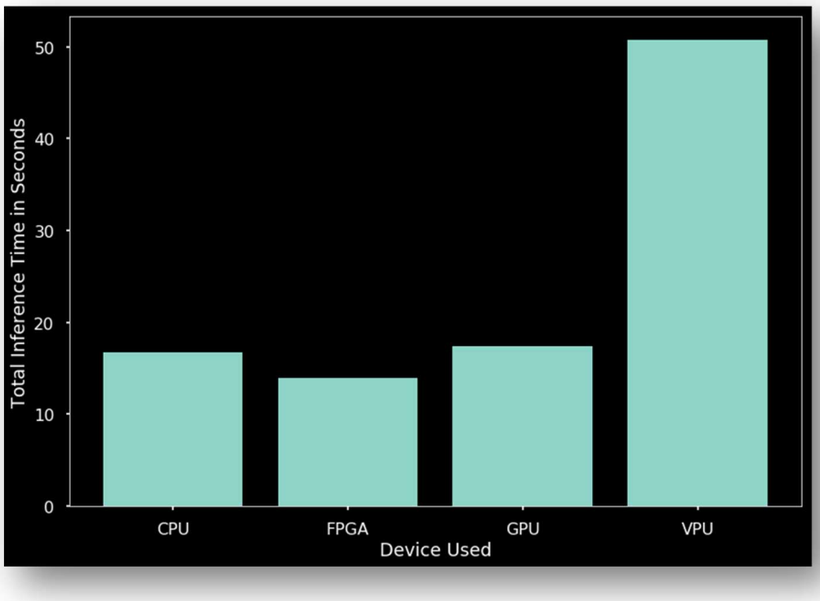

Figure 8: Inference Time in Seconds for Transportation scenario.

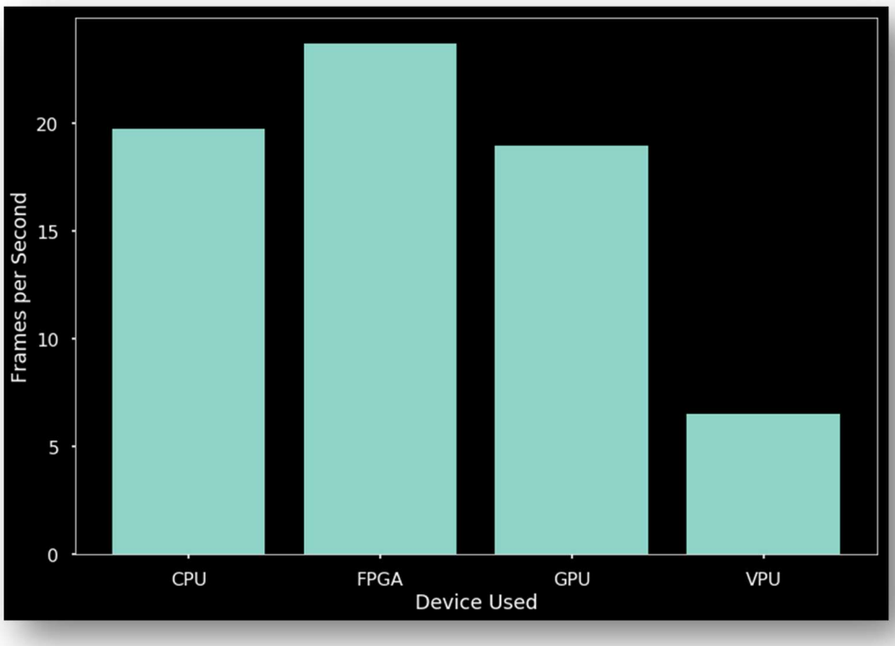

Figure 6: FPS for Transportation Scenario.

Final Hardware Recommendation

The FPGAs coupled with CPUs with HETERO plugin actually outperforms the IGPUs in terms of lower inference time and higher frames per second. The model is loaded only once across the inference pipeline, so FPGAs were actually a better choice than IGPUs. But the cost of FPGA is much higher than the pre-installed IGPUs and will incur high power usage. So, the client requirements will not be conformed with FPGAs due to higher price and higher power usage. So, we will stick to IGPUs in this scenario.

Write-up: Final Hardware Recommendation
*We will be using the **IGPUs** in this scenario as they are already preinstalled and will not further incur extra charges while providing better utilization of the hardware, lower model load time, inference time and significantly higher frame rates.*