Choose the Right Hardware

Scenario 1: Manufacturing

Mr. Vishwas is the VP of Engineering at Naomi Semiconductors, a manufacturer known for its industrial-grade standard in producing semiconductor chips. Recently, the company has been venturing into Intel Pentium 4/3000 chip production—and they want to maximize their revenue in this venture. Their other chips in the last year have earned them two million dollars alone. With such good revenue, their expansion into the Intel Pentium 4/3000 industry is an obvious next step.

There are several steps involved in the chip manufacturing process:

- Step 1: Produce a silicon ingot
- Step 2: Create blank wafers
- Step 3: Use these wafers to reproduce a patterned wafer
- Step 4: Create and test dies
- Step 5: Assemble bond dies into packages
- Step 6: Test packaged dies
- Step 7: Ship dies to customers

Mr. Vishwas explains that there have been several roadblocks in this pipeline. The entire process should take around 6 to 8 weeks—but currently, it is taking 10 to 12 weeks. This is reducing their revenue by 30%. Mr. Vishwas has noticed that Step 7 (shipping to customers) seems to be taking the most time. 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.

Workers alternate shifts to keep the floor running 24 hours a day so that packaging continues nonstop, but Mr. Vishwas has noticed a slow-down in production during the shift transition periods. Between shifts, he has observed a 70% dip in the production rate of packaged containers.

To help understand and address these issues, Mr. Vishwas 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. Mr. Vishwas would like the image processing task to be completed five times per second.

Once this productivity problem has been addressed, Mr. Vishwas would like to be able to repurpose the system to address a second issue. The second issue Mr. Vishwas has 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 Naomi Semiconductors 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.

Client Requirements and Potential Hardware Solution

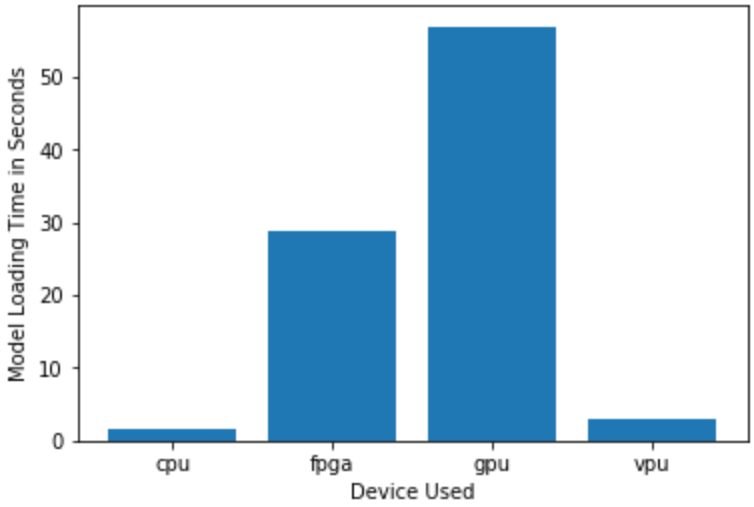
Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
FPGA

Requirement Observed (Include at least two.)	**How does the chosen hardware meet this requirement?**
The client requires the system to be flexible. They want to be able to reprogram it for multiple chip design. The new designs are created regularly.	FPGA is reprogrammable and good for low-volume production.
The client requires the system to run inference on the video stream very quickly.	FPGA is high performance and low latency.
The client has plenty of revenue to install a quality system, but they'd like it to last for at least 5-10 years.	FPGA is expensive, but it has a long lifespan.

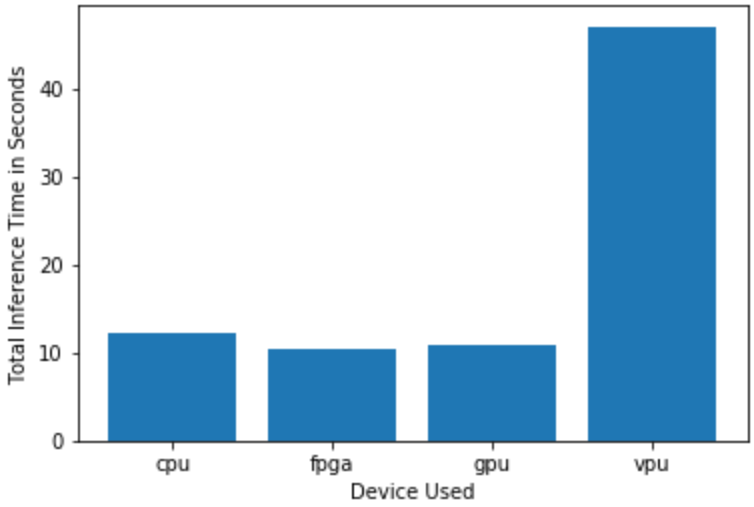
Queue Monitoring Requirements

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

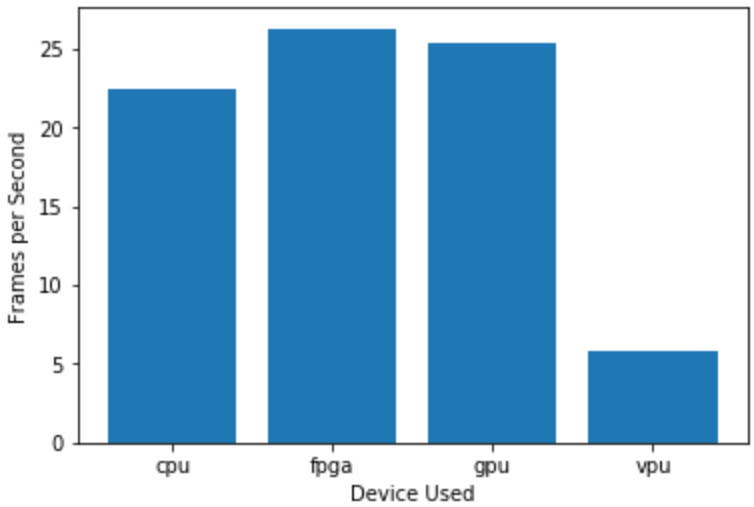
Test Results



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

Write-up: Final Hardware Recommendation
FPGA is the best choice for this scenario. The client requires it to run inference very quickly. Among all devices, FPGA has the shortest inference time. The client also requires it to complete the image processing task five times per second. FPGA can process 26 FPS, the highest among the devices. Additionally, as required, FPGA has a long lifespan and is reprogrammable.

Scenario 2: Retail

PriceRight Singapore has one of its smaller outlets in the tiny neighborhood of Dover. Mr. Lin is the store manager—and like any good store manager, he wants to use Edge AI to help maximize his profit this year.

Most of the customers are regulars at the store. Mr. Lin 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 Mr. Lin has seen wavering sales. When wait times are short and checkout happens smoothly, he sees a jump in his revenue from 6 to 20%. However, if there is congestion at the checkout counter, his profits only go up to 4-5%.

Mr. Lin believes this problem can be easily solved by directing people to less-congested queues in the store, and he is 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.

Mr. Lin 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. Mr. Lin also believes in giving fair employment and good wages. He pays his staff with proper salaries, along with substantial bonuses twice a year. As a result, Mr. Lin does not have much money to invest in additional hardware, and also would like to save as much as possible on his electric bill.

Client Requirements and Potential Hardware Solution

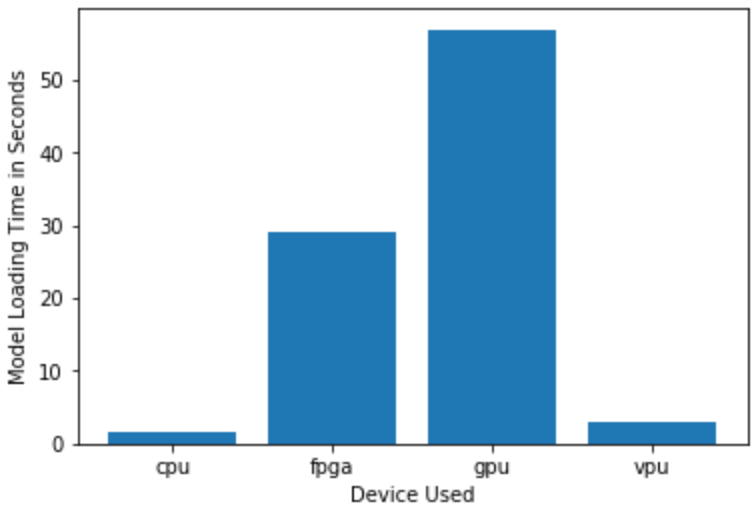
Which hardware might be most appropriate for this scenario? (CPU / IGPU / VPU / FPGA)
CPU

Requirement Observed (Include at least two.)	How does the chosen hardware meet this requirement?
Power requirement. The client requires to save as much as possible on the electric bill.	CPU can handle the required FPS - which doesn't need to be very fast for counting people at the counter - and less FPS implies less power consumption.
The client uses the computers to carry out some minimal tasks.	CPU has the resources to carry on edge AI tasks.
Budget. Most of the store's checkout counters already have a modern computer with Intel i7 core processor. The client doesn't have much money to invest in additional hardware.	CPU doesn't require the client to spend extra money.

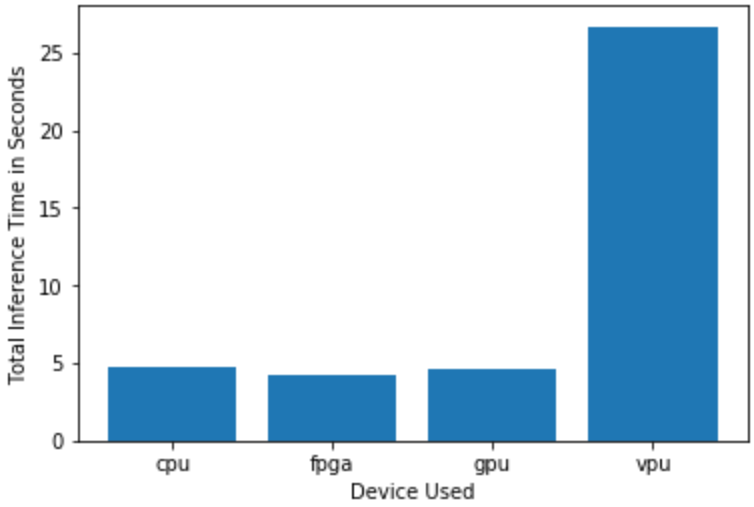
Queue Monitoring Requirements

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

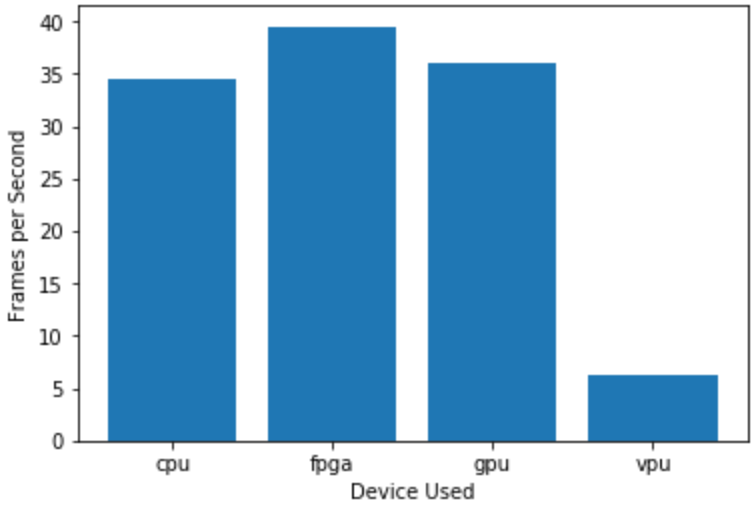
Test Results



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

Write-up: Final Hardware Recommendation
CPU is the best choice for this scenario. The client doesn't want to spend money on hardware. This left us with CPU and GPU. The client requires low power consumption, and lower FPS consumes less power. The CPU's FPS is lower, while running inference at around the same speed as the GPU.

Scenario 3: Transportation

Ms. Leah is the Innovation head for Delhi Metro Rail Services. Delhi Metro is an urban passenger transportation system connecting Ghaziabad, Faridabad, Gurgaon, Noida, Bahadurgarh, and Ballabhgarh in

the National Capital Region of India. Delhi Metro makes 2,700 trips every day and is one of the busiest metros in India.

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. Ms. Leah 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.

Client Requirements and Potential Hardware Solution

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 would like to save as much as possible on future power requirements.	VPU or NCS2 has a very low power consumption of only 1-2 watts.
Budget. The client has a budget of a maximum of $300. She wants to monitor with 7 cameras at the same time.	VPU or NCS2 is inexpensive. It costs around $70 to $100 per stick. NCS2 can run 4 inferences per stick, and 7 cameras need 2 sticks - we only need to spend $200.
The CPUs are being used to process and view CCTV footage and no significant processing power is available to run inference.	VPU is an AI accelerator. The pre-existing CPU will not be doing any calculation.

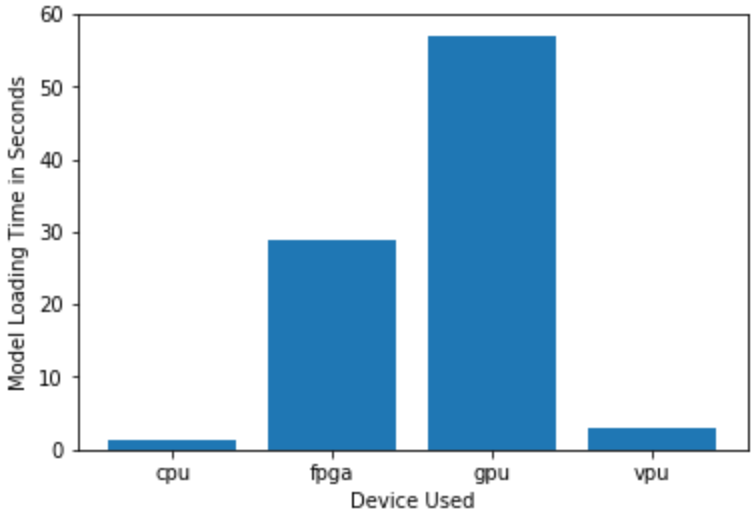
Queue Monitoring Requirements

Maximum number of people in the queue	7

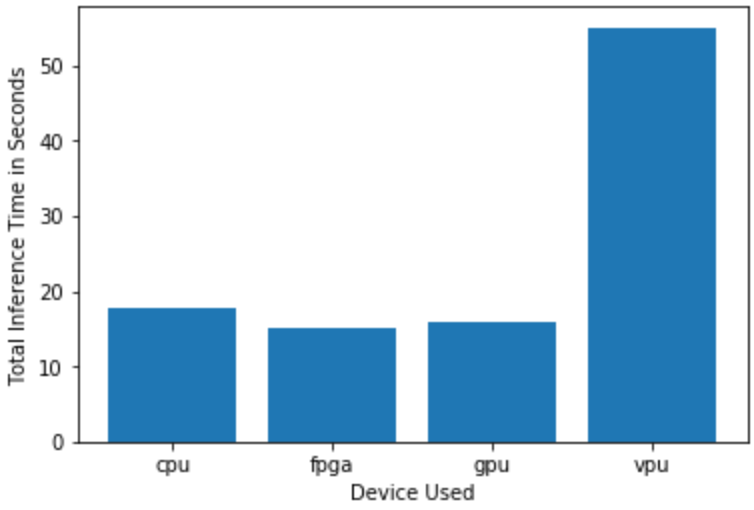
Model precision chosen (FP32, FP16, or Int8)	FP16

Test Results

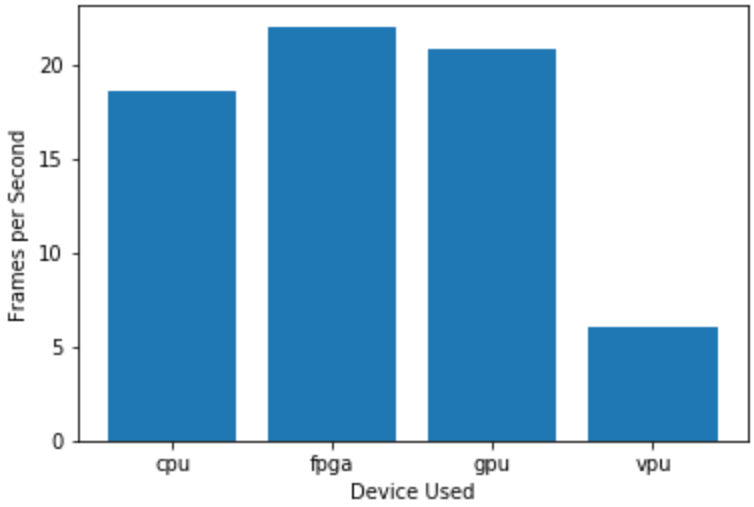
After you've tested your application on all four hardware types (CPU, IGPU, VPU, and FPGA), copy the matplotlib output showing the comparison into the spaces below. You should have three graphs (for model load time, inference time, and FPS).



Model Load Time



Inference Time



FPS

Final Hardware Recommendation

Write-up: Final Hardware Recommendation
VPU is the best choice for this scenario. The client has a \$300 budget. Her PCs have no additional process power available to run inference. It's either to upgrade the processor (and possibly the motherboard) or add AI accelerators like VPU. This left us with VPU as the only option. Although VPU takes longer to run inference, the client doesn't require it to run fast. On top of this, the client requires to save as much power as possible. VPU has the lowest FPS, which implies a very low power consumption.