Smart Digital Junction

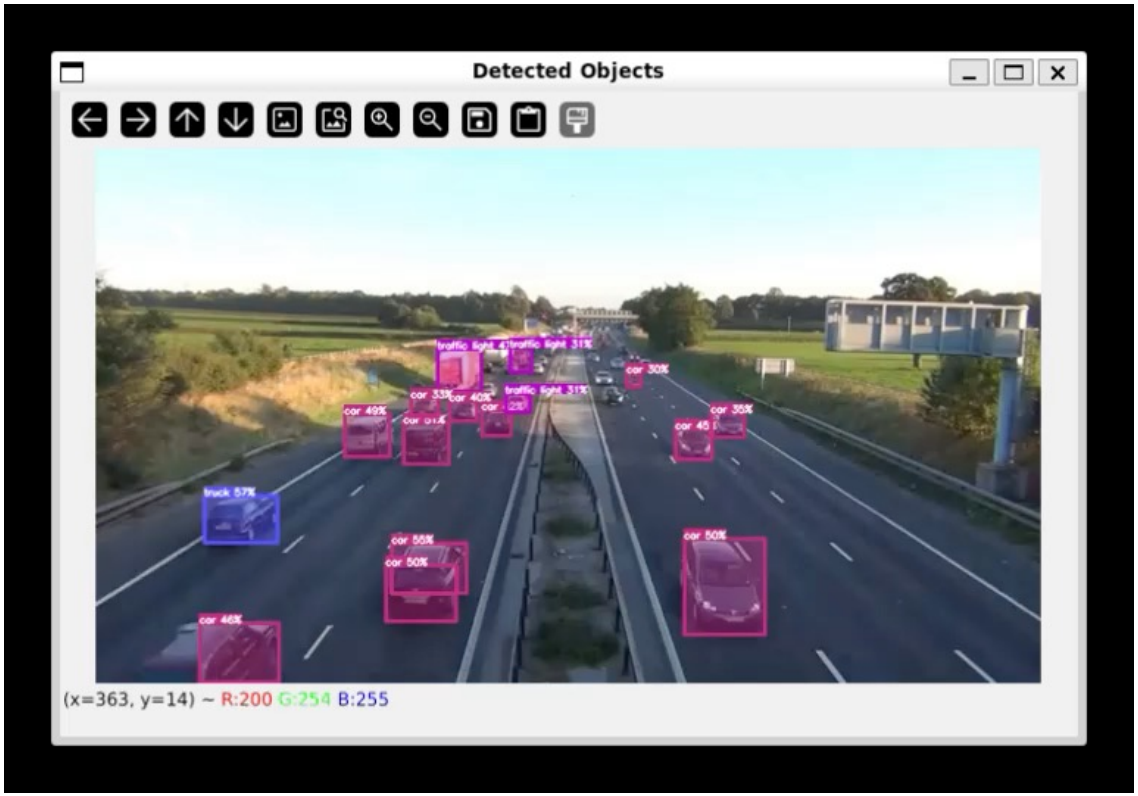
Traffic Applications using Hailo-8

Overview

Previously – *Can we use Hailo?”*

- Feasibility of using the HAILO-8 for real-time traffic monitoring
- Determined with basic benchmarks against Desktop PC + GPU environment

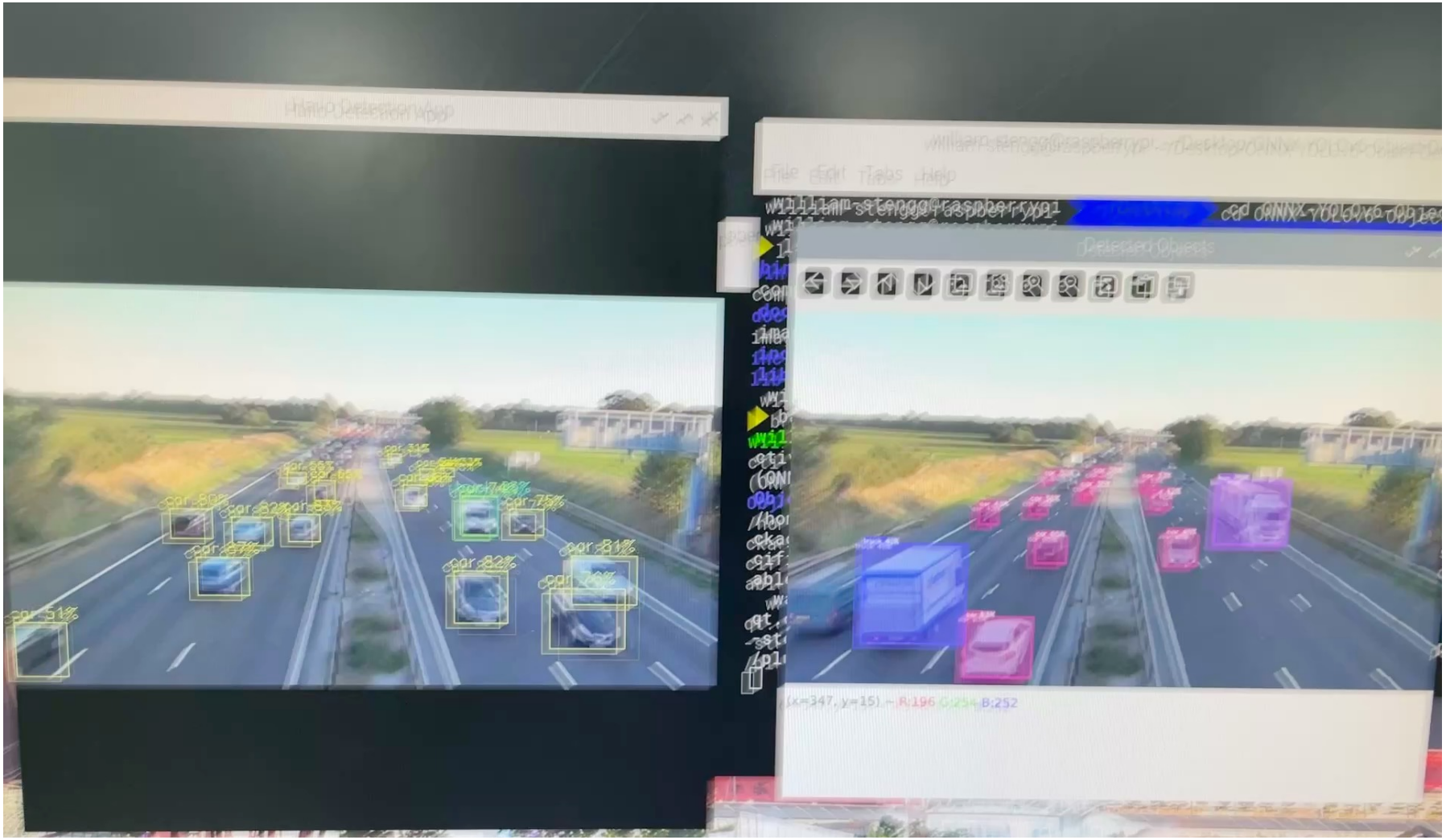
Desktop PC + GPU



- Time taken: 66.3 seconds
- 1.34x faster than real-time
- 74.6% of total duration

RPi + Hailo-8

RPi Only



Overview

Now – *Can we use Hailo for our own applications?*

- Evaluating practical applications for the Hailo-8 module
- **Task:** Simultaneous Object Detection, Classification + Tracking, Speed Estimation
- **Evaluation:** Cost, Power & Resources Consumption

Applications

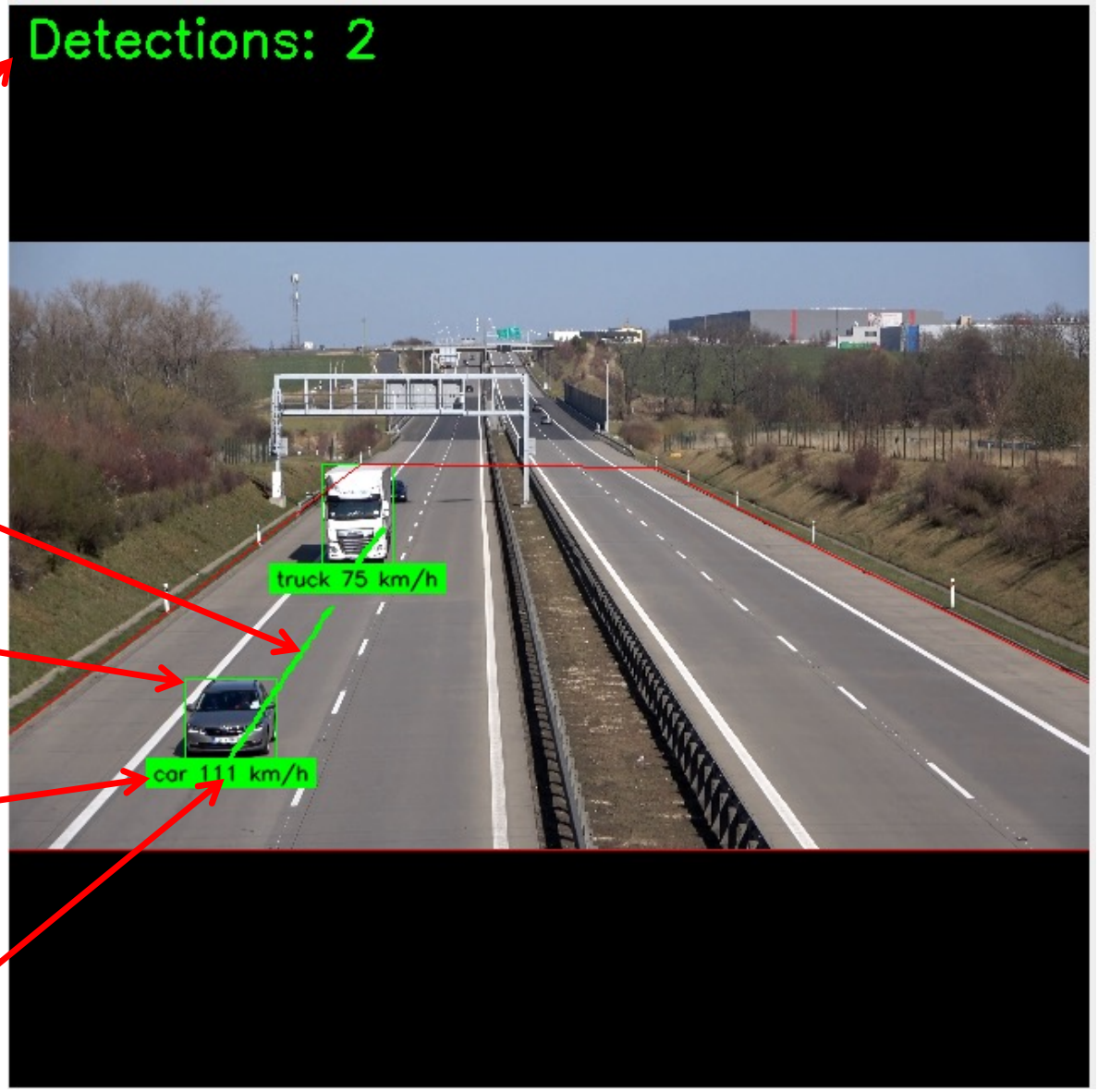
No. of Objects Detected

Path tracking

Object Detection

Classification

Speed Estimation



Speed Estimation with Yolov8m

Speed Estimation with Yolov8m

CPU Utilization & Power Consumption

```
Device ID                                       Utilization (%)         Architecture
-------------------------------------------------------------------------------------
0000:01:00.0                                    98.6                    HAILO8

Model                                           Utilization (%)         FPS           F
-------------------------------------------------------------------------------------
yolov8m                                         98.6                    29.0          5

Model                                           Stream
-------------------------------------------------------------------------------------
yolov8m                                         yolov8m/input_layer1
yolov8m                                         yolov8m/conv83_123
yolov8m                                         yolov8m/conv82_123
yolov8m                                         yolov8m/conv70_123
yolov8m                                         yolov8m/conv58_123
yolov8m                                         yolov8m/conv71_123
yolov8m                                         yolov8m/conv57_123
```

- Average Power: ~2.5W
- CPU (Hailo) Utilization:
  - 90-99% (Yolov8l)
  - 70-90% (Yolov8m)
  - 10-30% (Yolov6n)
- Cost: ~$300

```
william-stengg@raspberrypi ~ hailortcli measure-power
Executing on device: 0000:01:00.0
[HailoRT] [warning] Using the overcurrent protection dvm for power measurement will disable the overcurrent protection.
If only taking one measurement, the protection will resume automatically.
If doing continuous measurement, to enable overcurrent protection again you have to stop the power measurement on this dvm.
Current power consumption (W): 2.54011
```

Accuracy

- Hailo8 is able to support up to Yolov8l model
- Dense traffic analysis with ~100% NPU utilization
- Conventionally, Yolov6n used for real-time applications. However, Hailo can run a much larger Yolov8l for real-time applications.

Model	mAP@0.5	mAP@0.5:0.95	Inference Speed (FPS)	Model Size
YOLOv8l	~55%	~40%	20-30 FPS	~120 MB
YOLOv8m	~50%	~35%	30-40 FPS	~90 MB
YOLOv6n	~48%	~33%	40-60 FPS	~45 MB

Comparison b/w GPUs and HAILO-8

Running 24/7/365 inference with YOLOv8m (at 99% uptime)
** Using SP Electricity Tariff of 31.72 cents/kWh [incl. GST]*

Device	Inference Speed (FPS)	Cost ($)	Power (W)	Operating Cost* ($ per year)
RTX 2080	20-60 FPS	$800 + $1000	~215W	$591.44
Jetson Orin	200-250 FPS	$1999	~60W	$165.05
HAILO-8	30-60 FPS	**$300**	~5W	**$13.75**

- All platform can meet 20fps (minimum for real-time appln.)
- All platform can run YOLOv8m model (very accurate for real-time appln.)
- HAILO = **97.7%** operating cost reduction compared to RTX GPU
- HAILO = **91.6%** operating cost reduction compared to Jetson

Conclusion

- HAILO-8 module performs very well for real-time traffic monitoring
- Computationally intensive applications requiring conventional Desktop PC + GPU environments can be run on HAILO-8
- Much lower cost and energy consumption with **same accuracy**

Conclusion

- No significant trade-offs for running inference
- How?
  - Hailo uses a special compiler that converts YOLO models to HEF (Hailo Executable File)
  - This conversion process is computationally intensive, but only needed once
    - ^From previous slide costs about $2 (~7 kWh) for 1 model conversion
- Do conversion **once**, run inference **many times** at low cost
- Accuracy can be improved as needed by compiling a bigger model
  - Hailo-8 can support up to YOLOv8l (2x larger than v8m)
  - Hailo-15 theoretically up to YOLOv8x (4x larger than v8m)

Other Applications

- License Plate Recognition