# ECE5545/CS5775 A1

Naman Makkar

TOTAL POINTS

## 7.5 / 10

QUESTION 1

### *1* Submit your report 7.5 / 10

**- 0 pts** Correct

**- 0.25 pts** Use of log with flops-vs-latency is unjustified/unnecessary

**- 0.5 pts** Points are plotted above the roofline. Possible explanations are provided but the points are in some cases 2x higher than peak throughput which indicates the presence of a bug.

**- 0.25 pts** You attempted to justify the incorrect forward:backward ratio, but you did not root-cause the issue.

**- 0.5 pts** Plots are hard to read in more than one place

**- 2 pts** missing question 5

**- 2 pts** Rooflines are horizontal lines (incorrect)

**- 3 pts** Missing writeup, commentary and multiple plots are not visible

**- 0.5 pts** Bar charts and pie charts are hard to read

**- 0.5 pts** incorrect gpu roofline - mem BW looks wrong

**- 0.25 pts** Commentary is very brief or missing for some questions

**- 0.25 pts** noisy measurement

✓ **- 0.5 pts** *FLOPs ratio missing in Q5*

**- 0.75 pts** Pie charts in Q5 (mostly) missing or incorrect

✓ **- 0.5 pts** *rooflines are plotted as 2 separate lines*

✓ **- 0.5 pts** *Q5 pie charts missing*

**- 0.5 pts** forward:backward flops numbers way off in some cases

**- 0.5 pts** Rooflines are all in one plot making it hard to read them

**- 0.5 pts** GPU latency measurments are incorrect (should use events)

**- 0.5 pts** incorrect explanation

**- 1 pts** FLOPS/Latency ratios missing in Q5

**- 1 pts** Unlabeled axes in plots

**- 1 pts** Models are vertical lines on roofline plot where they are supposed to also include actual performance.

**- 2 pts** Rooflines are just points.

**- 1 pts** Incorrect roofline

**- 1 pts** op intensity needs to intersect with roofline

**- 0.5 pts** Points are plotted above the roofline. Explanation is not provided, indicates a bug.

✓ **- 1 pts** *Analysis of experiments insufficient.*

**- 0.25 pts** Correlation Matrix seems wrong

**- 0.25 pts** Models not distinguished in plots

**- 0.5 pts** Q4 Plot Missing/incorrect

**- 2 pts** Section 3 Missing

**- 2 pts** Section 4 Missing

**- 0.25 pts** Minor enhancements to plots possible. unclear in some places

# Assignment 1 (USING 1 LATE DAY)

**Naman Makkar (nbm49)**

February 2023

## 1 Chip Analysis

### 1.1 GPUs

#### 1.1.1 NVIDIA TESLA T4

The NVIDIA TESLA T4 gives a **peak performance of 8.1 TFLOPs/second** for FP32 single precision while it gives a peak performance of 65 TFLOPs/second for mixed precision and a peak performance of 130 TOPs/second and 260 TOPs/second for INT8 and INT4 precision respectively.
The NVIDIA Tesla T4 has a **memory bandwidth of 320 GB/s**. It makes use of the GDDR6 memory.

#### 1.1.2 NVIDIA TESLA V100 for PCIe Based Servers

The NVIDIA TESLA V100-PCIe gives a **peak performance of 14 TFLOPs/second** for FP32 single precision, 112 TFLOPs/second. It provides a **memory bandwidth of 900 GB/s**. It makes use of the HBM2(High Bandwidth Memory 2) memory.

#### 1.1.3 NVIDIA TESLA P100 for PCIe Based Servers

The NVIDIA Tesla P100 provides a **peak performance of 9.3 TFLOPS** for FP32 single precision, 18.7 TFLOPS for half precision with a memory bandwidth of **732 GB/s** with the help of the HBM2 memory.

#### 1.1.4 NVIDIA A100 80GB PCIe

The NVIDIA A100 provides a **peak performance of 19.5 TFLOPS for FP32 single precision**. With a **memory bandwidth of 1935 GB/s**.

#### 1.1.5 NVIDIA H100 PCIe

The NVIDIA H100 provides a **peak performance of 51 TFLOPS for FP32 single precision** with a **memory bandwidth of 2TB/s**.

## 1.2 CPUs

### 1.2.1 2x Intel Xeon E5-2699 v4 in dual socket configuration

Base clock speed of 2.2GHz and a maximum turbo frequency of 3.6 GHz. It has 22 cores each which provides a peak performance of **Number of cores * Clock Speed*2(single precision FP)*2(ADD/MUL)*2(number of chips)= 22*2.2*2*2*2 = 387.2 GFLOPS**. Whereas the memory bandwidth of the 2 chips can be calculated as **Memory frequency*4(number of channels)*2(number of bytes transferred by channel per cycle)*2(number of chips) = 2200*4*2*2 = 17.6 GB/s**.

### 1.2.2 2x Intel Xeon E5-2697 v2 in dual socket configuration

Clock speed of 2.7 GHz without turbo mode with 12 cores each. Theoretical peak performance can be calculated as **Number of cores * Clock Speed*8(single precision FP)*2(ADD/MUL)*2(number of chips)=12*2.7*8*2*2= 1036.8 GFLOPS**. It has a memory frequency of 1866 GHz. The memory bandwidth can be calculated as **Memory frequency*4(number of channels)*8(number of bytes transferred by channel per cycle)*2(number of chips) = 1866*4*8*2 = 119.424 GB/s**

### 1.2.3 Intel Xeon Phi 7120A coprocessor

This has a **theoretical peak performance of 2420 GFLOPS** and a **memory bandwidth of 352 GB/s**.

## 1.3 ASIC

### 1.3.1 TPUv4

TPUv4 provides a **peak performance of 275 TFLOPS for BFLOAT16 and INT8 precision** with a **memory bandwidth of 1200 GB/s**.

## 1.4 SoC

### 1.4.1 Apple A14 Bionic

Apple's A14 Bionic chip has a **maximum memory bandwidth of 5.3375 GB/s** with a **peak performance of 1000 GFLOPS for FP32 single precision, 2000 GFLOPS for FP16 Half Precision**.

# 2 DNN Compute and Memory Analysis

## 2.1 10 Models

The 10 models chosen are - EfficientNetB4, ResNet18, ShuffleNetv2_x2.0, SqueezeNet_1.1, EfficientNetB0, EficientNetB1, EfficientNetB2, MobileNetV3

### 2.2 Model FLOPs and Memory Footprint

Input size chosen for the tests is **(3, 224, 224)** for different batch sizes.
MACs have been calculated using the **thop** library, and FLOPs have been calculated as MACs * 2
The FLOPs of each model can be seen in **Figure 1**
The Memory footprint of the models has been calculated with the help of the **torchsummary** library and the Parameter size has been acquired in **MB**

### 2.3 Overlay Operational Intensity

The Operational Intensity has been calculated as **Number of FLOPs/ Total number of bytes accessed**. This was carried out with the help of the torchsummary library which provided the memory of the Input size, in addition to the parameter size of each model as well as the forward pass size.
Looking at Figure 3, the memory bandwidth and peak performance has been plotted for the **Google Colab CPU (Intel Xeon @ 2.20 MHz)** and the **Google Colab GPU (NVIDIA Tesla T4)**. Additionally, the operational intensities of the model have been plotted as vertical lines. The point at which the operational intensities intersect the roofline provides us with the peak performance of the model for that particular architecture

### 2.4 Memory Bound vs Compute Bound

Looking at **Figure 3**, it can be noted that the operational intensities are to the right of the intersection between the memory bandwidth of the CPU and the peak performance of the CPU. This indicates that all of the workloads are compute bound in the case of the CPU. While, looking at the GPU roofline, we can notice that all of the workloads are memory bound.

## 3 DNN Performance Benchmarking

### 3.1 Plotting Inference Latency vs FLOPs

The latency has been calculated for both CPU and GPU for batch sizes 1, 64, 256 for input tensor (3, 224, 224). **Figure 4** contains the plot for FLOPs vs Inference Latency for CPU, whereas **Figure 5** shows the plot for FLOPs vs Inference Latency for GPU.

### 3.2 Plotting Parameters vs Inference Latency

In **Figure 6** the Parameters vs Inference Latency has been plotted for batch sizes 1, 64, 256 for input tensor (3, 224, 224) on the CPU while in **Figure 7** the same has been done for the GPU.

### 3.3 Rank Correlation Coefficients between CPU Latency, GPU Latency, FLOPs, Parameters

The plot of the Spearman correlation coefficients can be observed in **Figure 8**

### 3.4 FLOPs vs Parameters

FLOPs provide an estimate of the number of mathematical operations required for carrying out a forward or a backward pass with the model, however FLOPs do not take into account the hardware architecture or the overhead involved in the forward or backward pass of the model. FLOPs do not take into account the memory footprint of the model and cannot be utilised to estimate the appropriate input size or batch size during training.
Parameters provide us with a way of correctly estimating the compute requirements for the model. However, the number of parameters of the model do not give us any intuition into how the performance of the model is calculated.
FLOPs are more useful when we need to calculate the model performance in OPs/sec, while Parameters are more useful when calculating the compute requirements and the memory footprint of the model.

### 3.5 Plotting Throughput for CPU and GPU

Since **Throughput = Inferences/second** the throughput has been calculated as **Batch Size/Latency**

The graphs of Parameters vs Throughput have been plotted for CPU and GPU in **Figure 9 and 10** respectively.

## 4 Hardware Utilization and Peak Performance

### 4.1 Plotting Performance on Roofline Plot

Plots have been provided in **Figure 11** and **Figure 12**.

### 4.2 Comments on Performance

It can be observed from the Roofline plots that the models are compute bound on the CPU and memory bound on the GPU. The models that are run on input tensors with batch size = 1 showed the highest model performance in Ops/second and the ones with batch size = 256 show the lowest model performance for both the CPU and GPU.

## 5 Inference vs training

Since we are utilising deep Convolutional Neural Networks, it can be assumed that the FLOPs ratio for backward to forward pass is **2:1**

The latency ratios for forward to backward pass can be visualised in **Figure 13** for each model for a batch size of 1 on CPU, for input tensor (3, 224, 224).

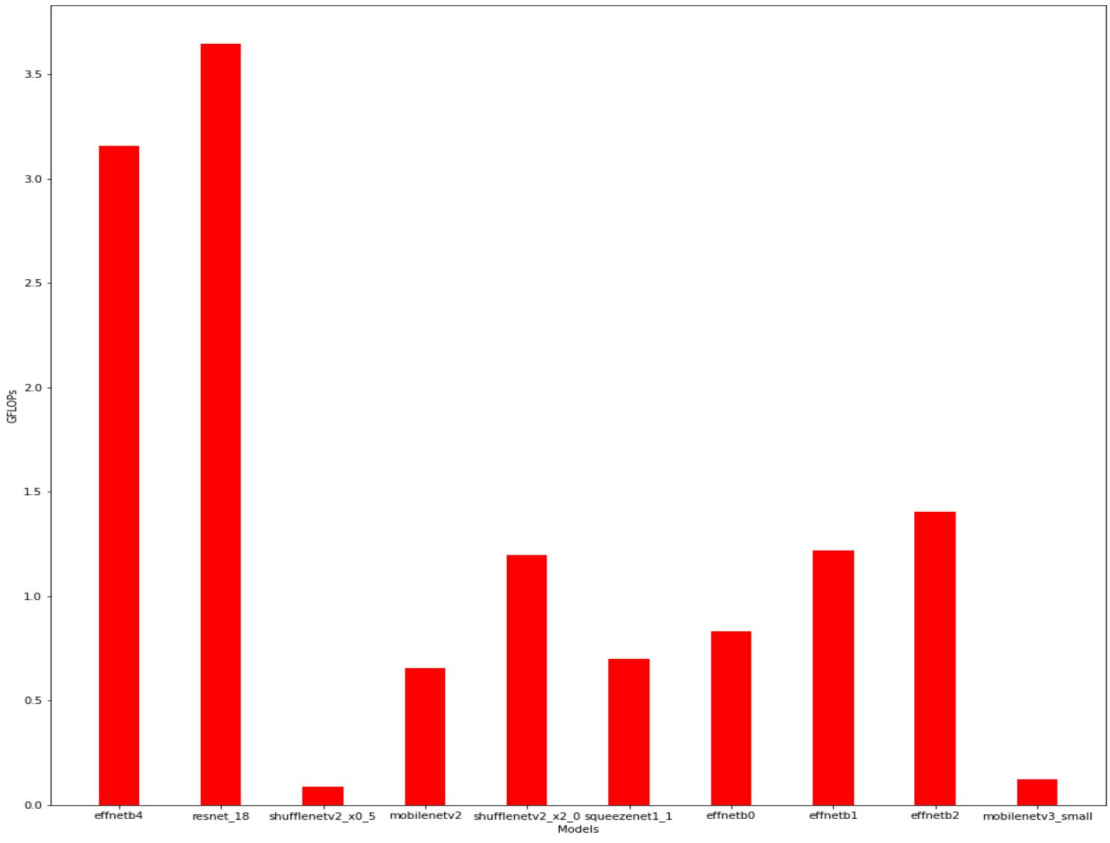


Figure 1: The FLOPs of each model shown in a bar chart

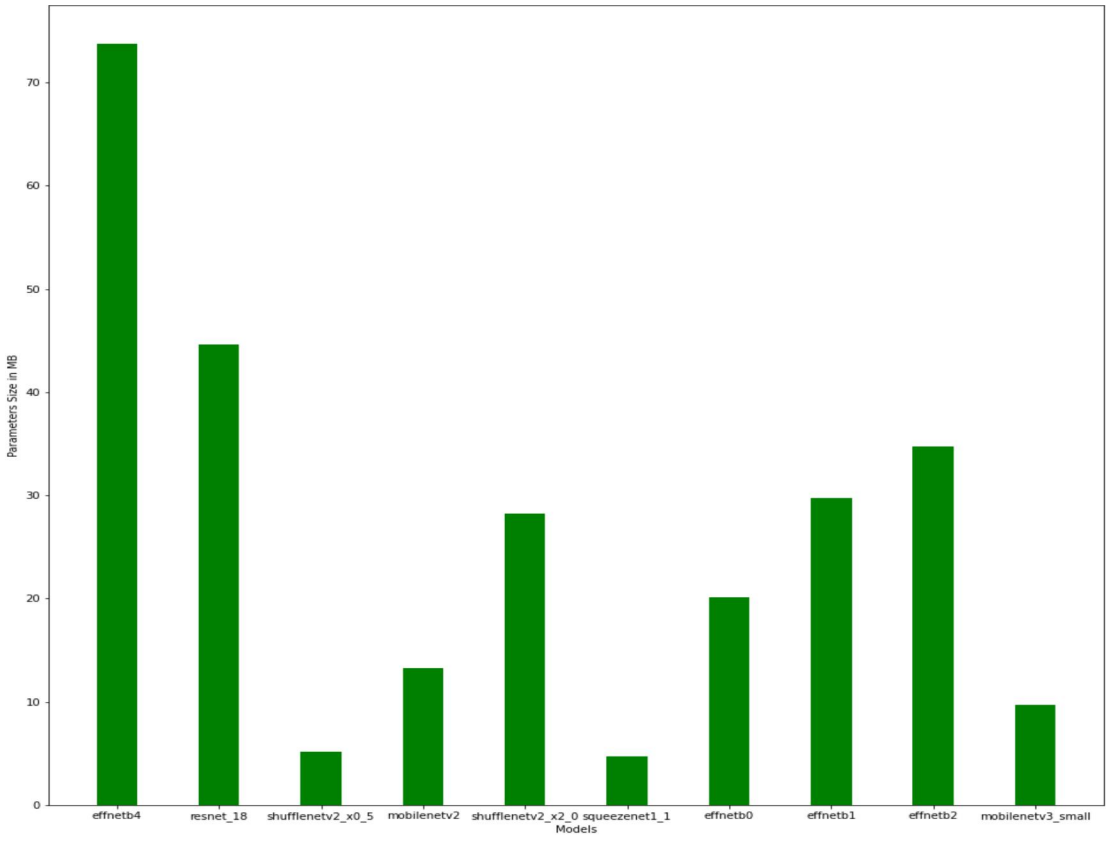


Figure 2: The Parameter size of each model shown in a bar chart

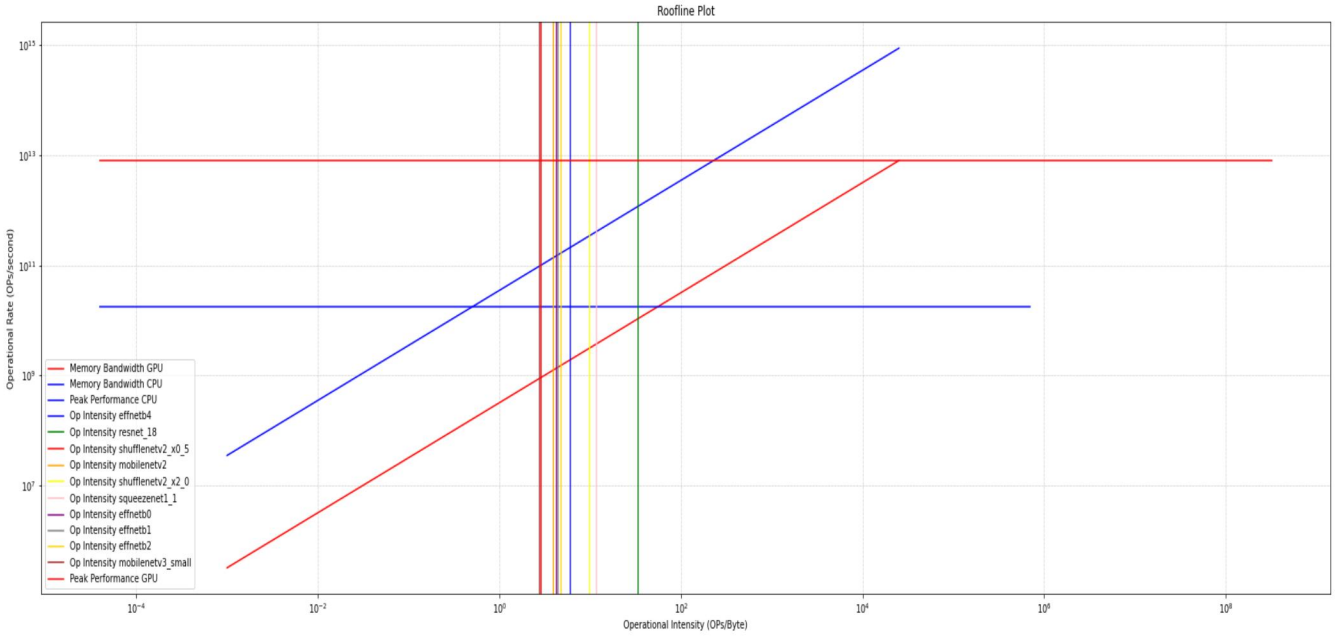


Figure 3: The Operational Intensity of each model plotted as a vertical line, with the peak performance and memory bandwidth of both CPU and GPU plotted

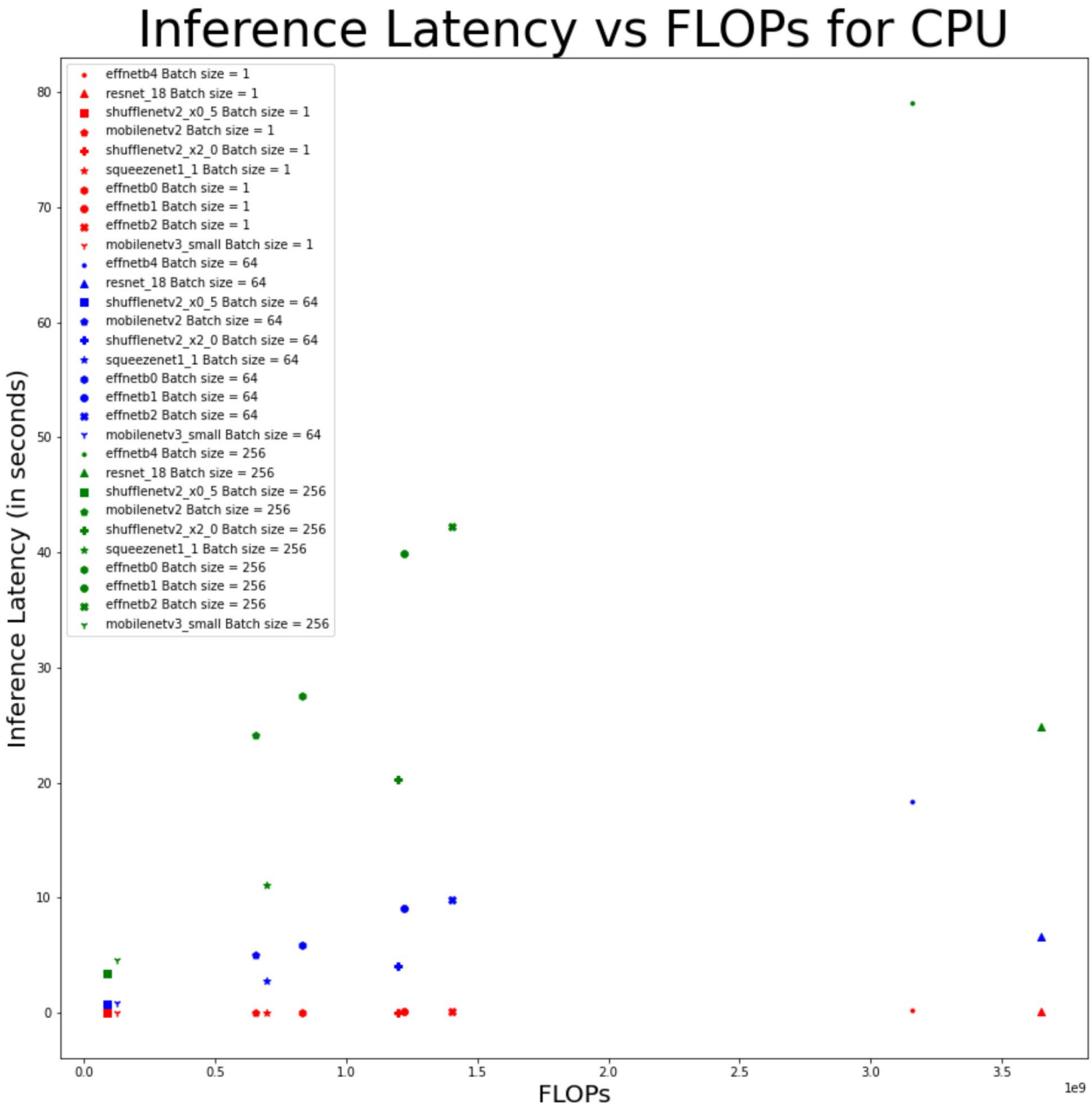




Figure 4: The FLOPs vs Inference latency of each model has been plotted for batch sizes 1, 64 and 256 for input tensor (3, 224, 224). This plot only shows inference latency on **CPU**

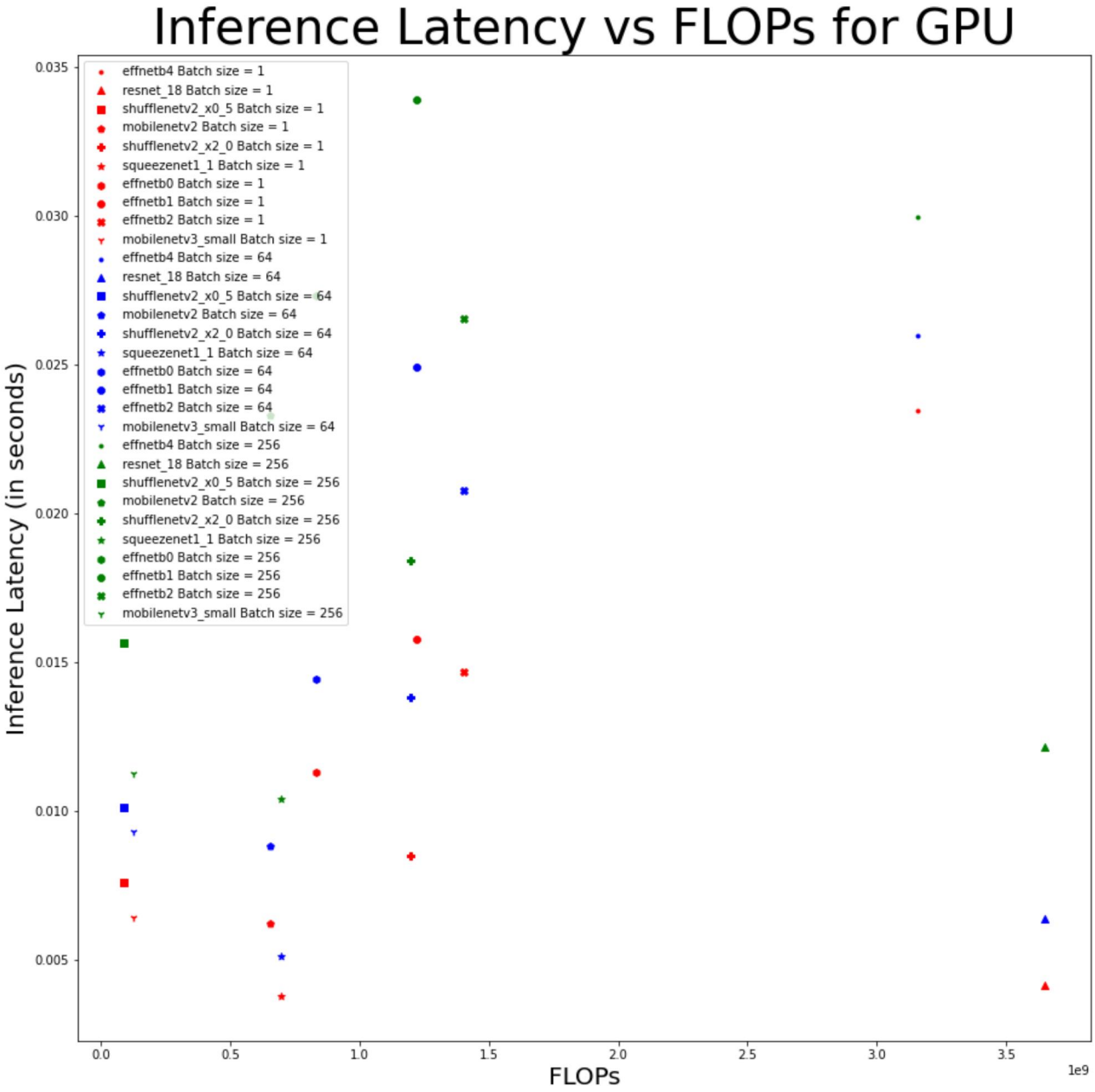




Figure 5: The FLOPs vs Inference latency of each model has been plotted for batch sizes 1, 64 and 256 for input tensor (3, 224, 224). This plot shows inference latency on **GPU**

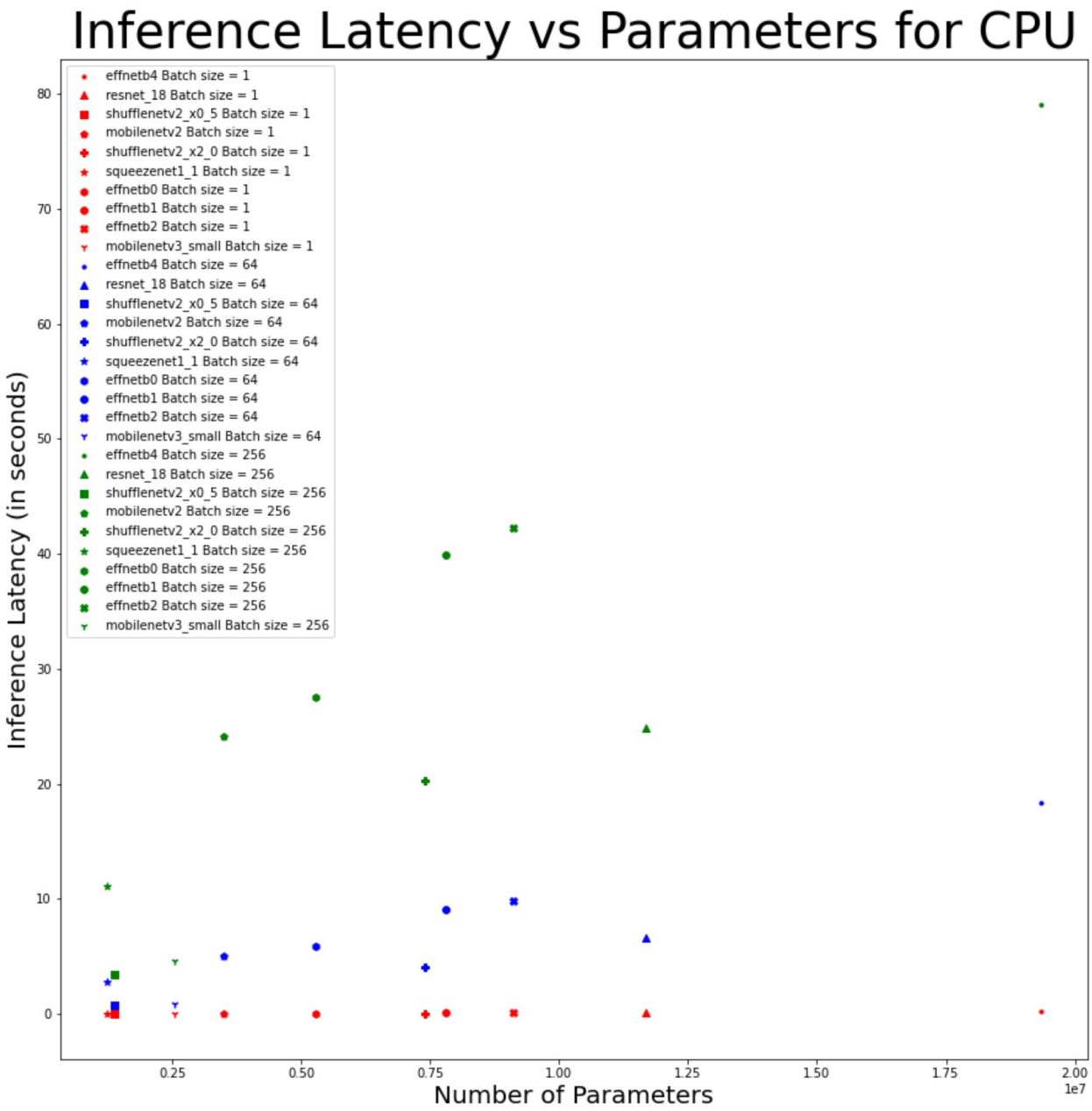




Figure 6: The **Parameters vs Inference latency** of each model has been plotted for batch sizes 1, 64 and 256 for input tensor (3, 224, 224). This plot only shows inference latency on **CPU**

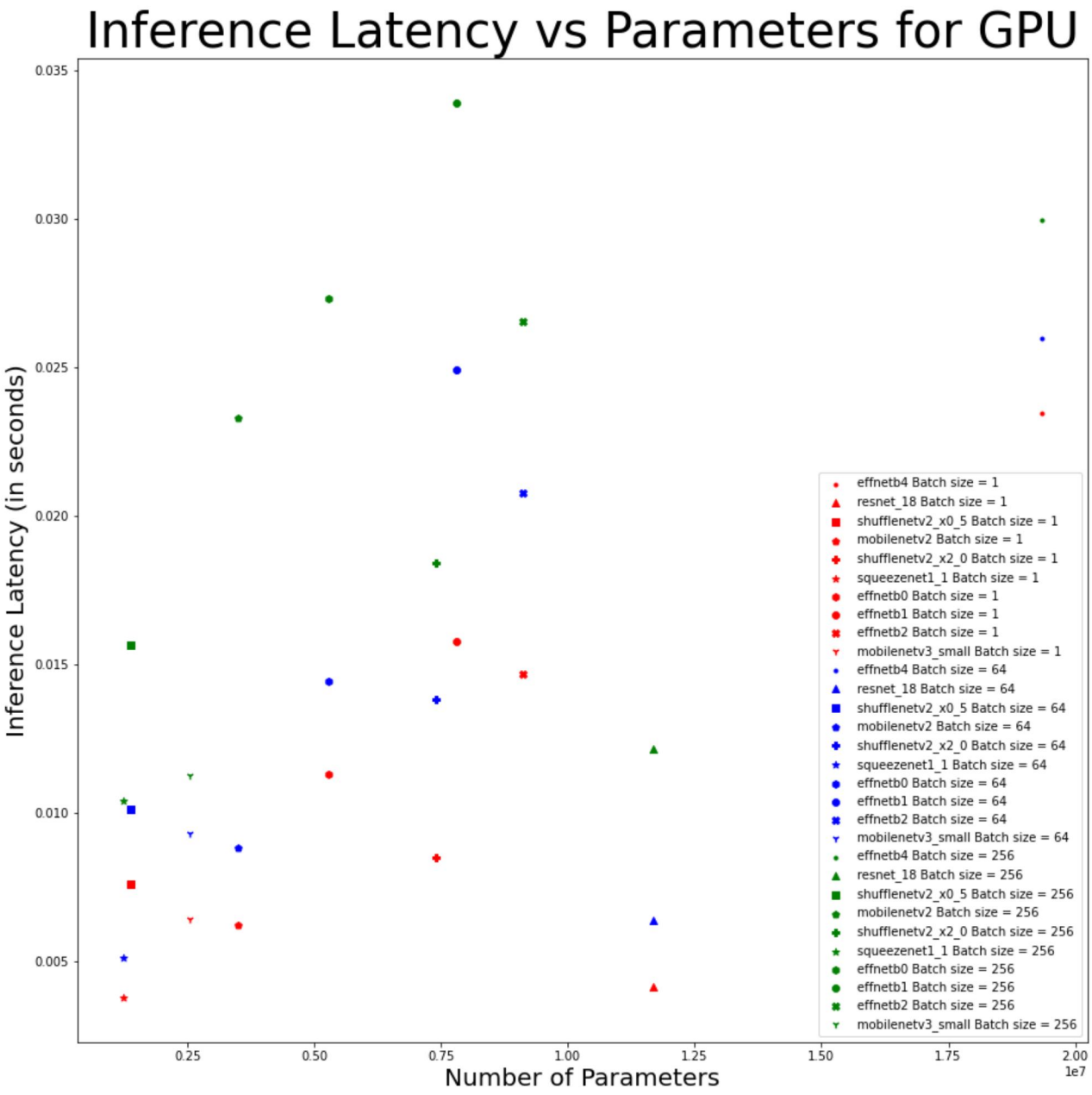




Figure 7: The **Parameters vs Inference latency** of each model has been plotted for batch sizes 1, 64 and 256 for input tensor (3, 224, 224). This plot shows inference latency on **GPU**

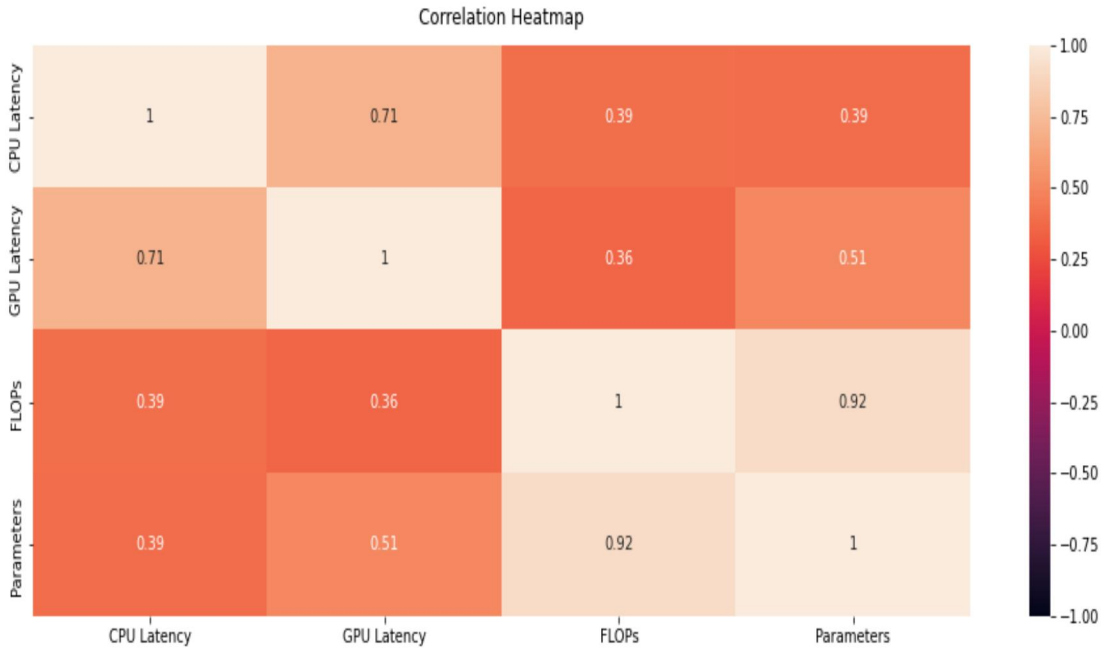


Figure 8: The Spearman's correlation coefficients have been plotted as a 2D matrix for CPU Latency, GPU Latency, FLOPs and Parameters

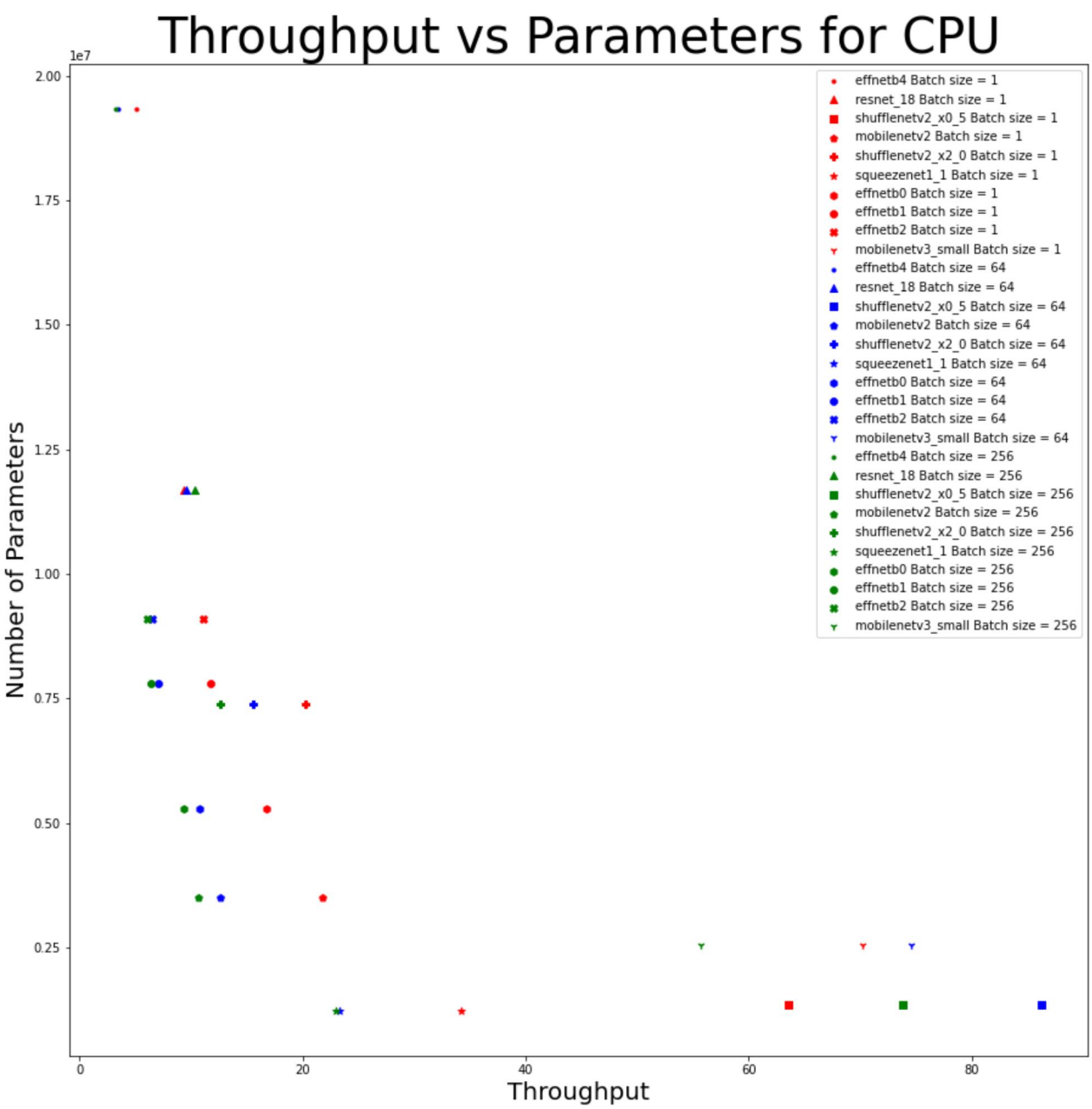




Figure 9: The **Parameters vs Throughput** of each model has been plotted for batch sizes 1, 64 and 256 for input tensor (3, 224, 224). This plot only shows inference latency on **CPU**

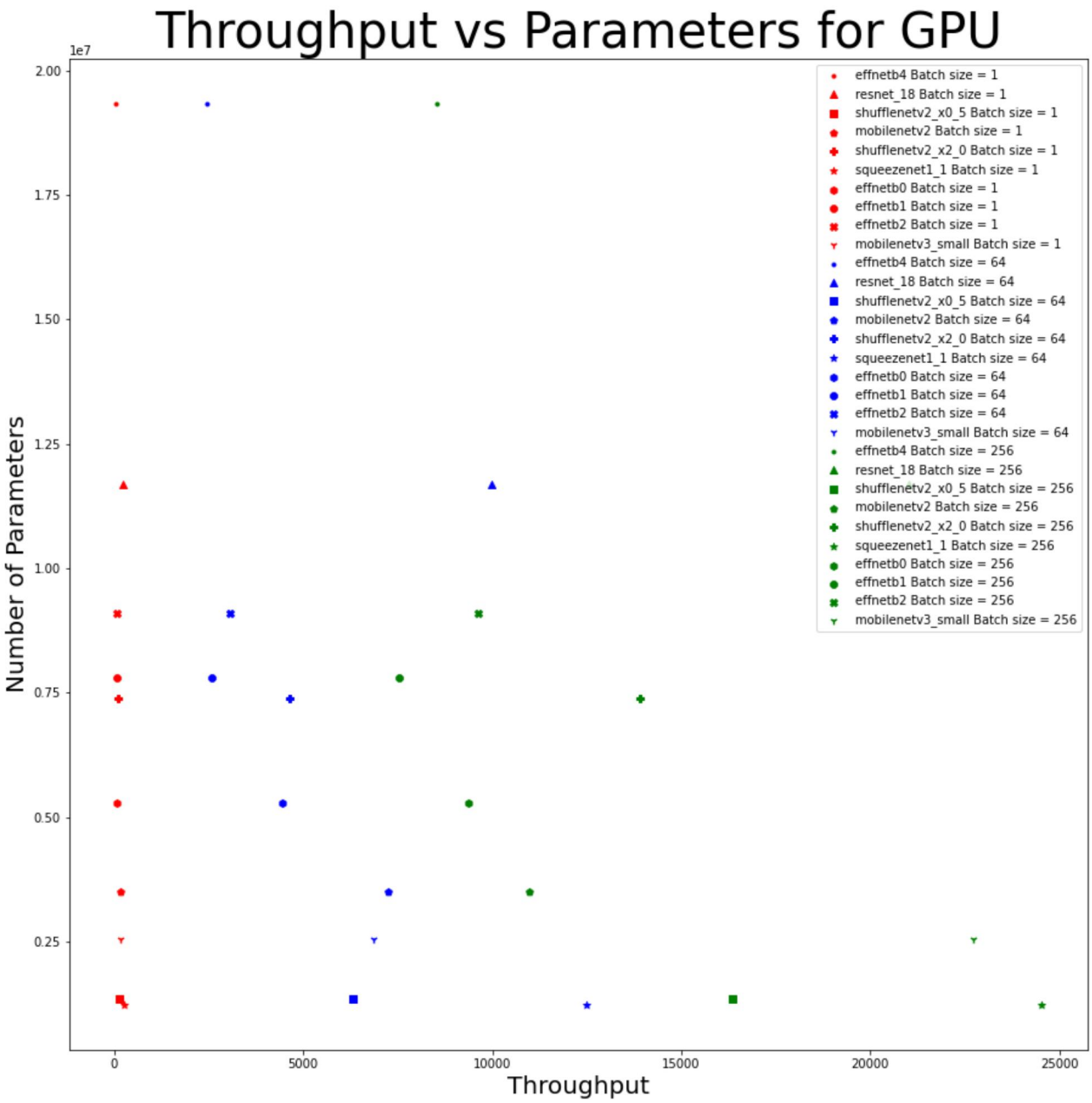




Figure 10: The **Parameters vs Throughput** of each model has been plotted for batch sizes 1, 64 and 256 for input tensor (3, 224, 224). This plot only shows inference latency on **GPU**

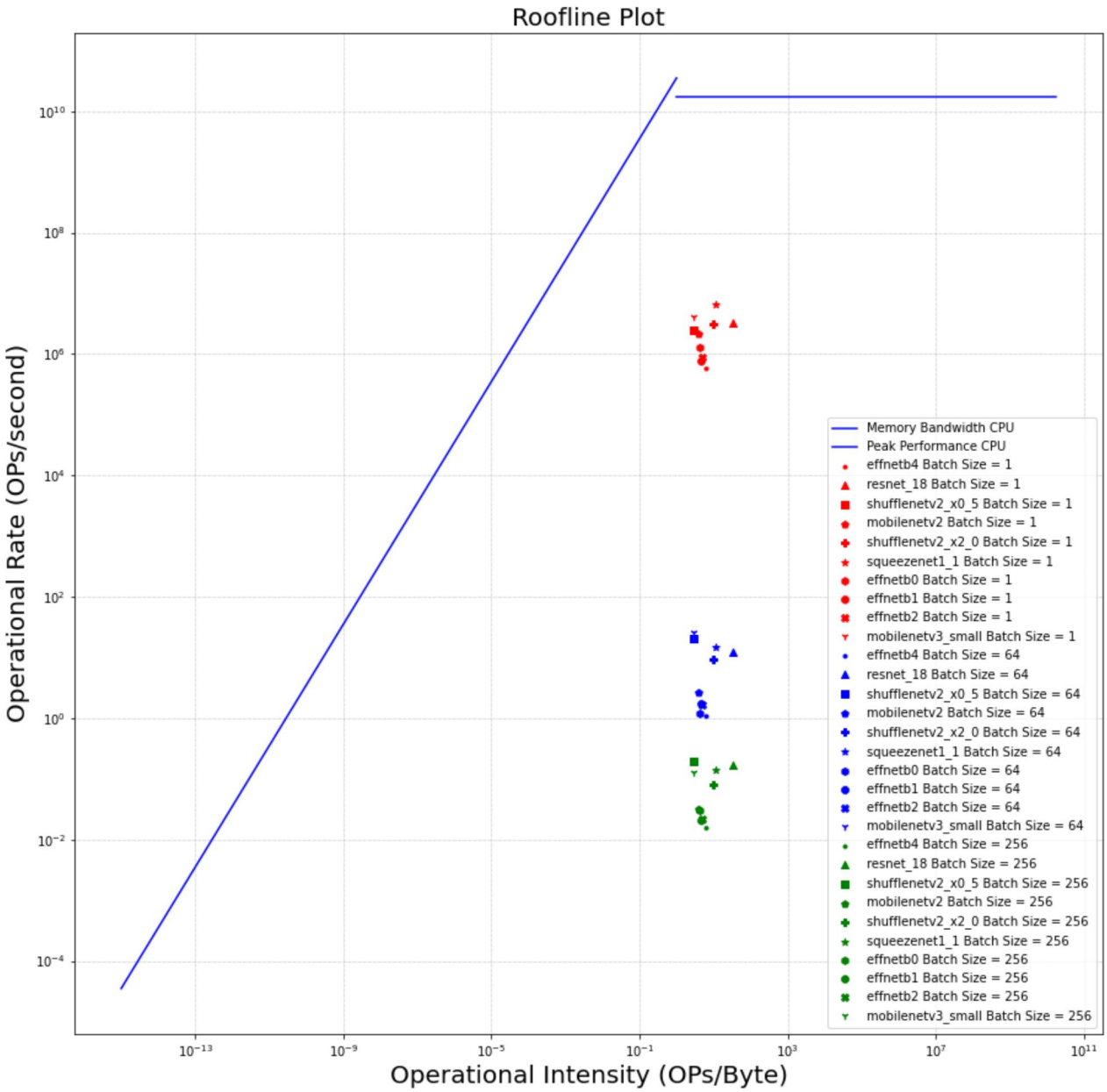




Figure 11: The roofline plot with model performance plotted alongside operational intensity for the CPU.

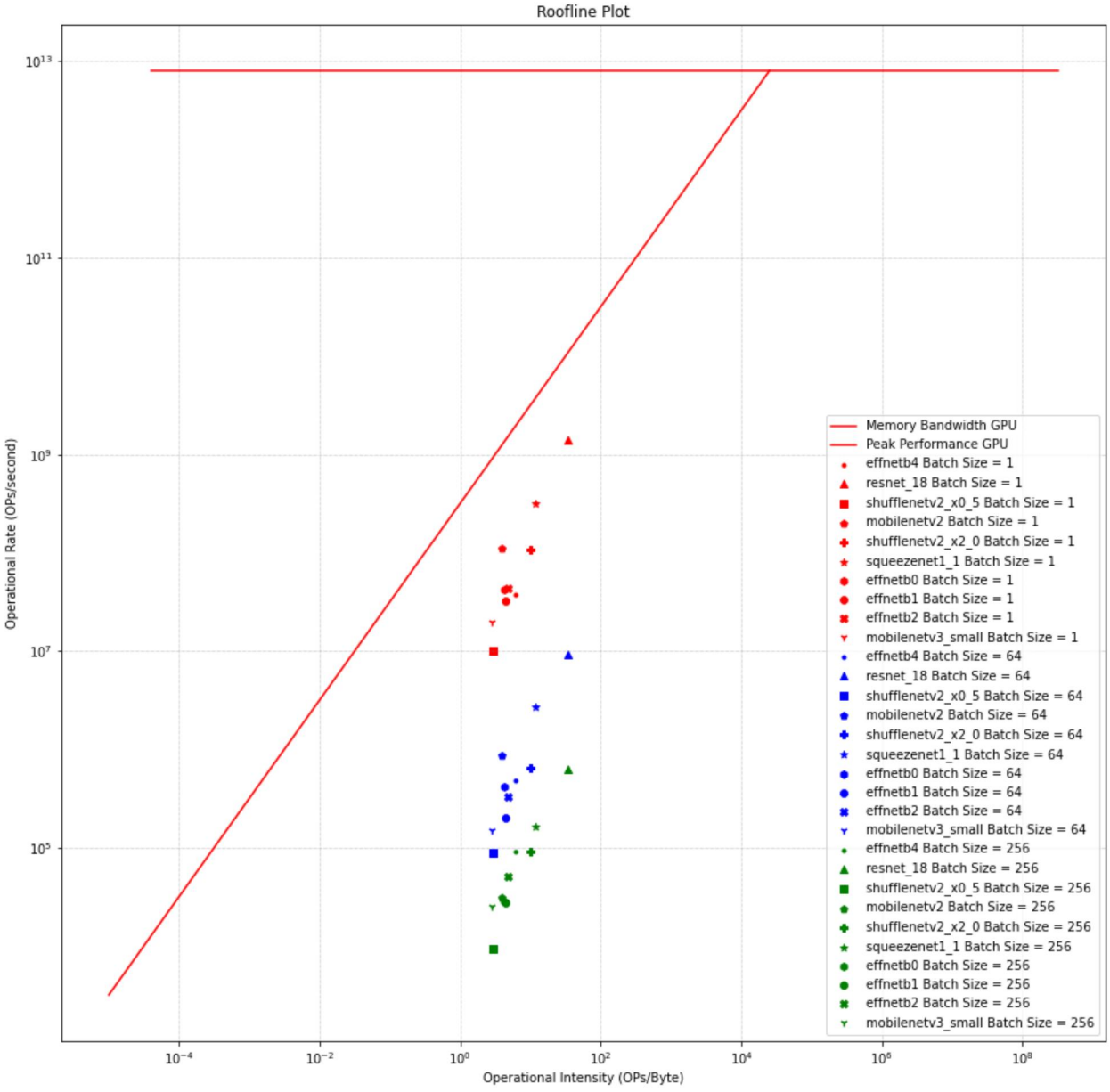




Figure 12: The roofline plot with model performance plotted alongside operational intensity for the GPU.

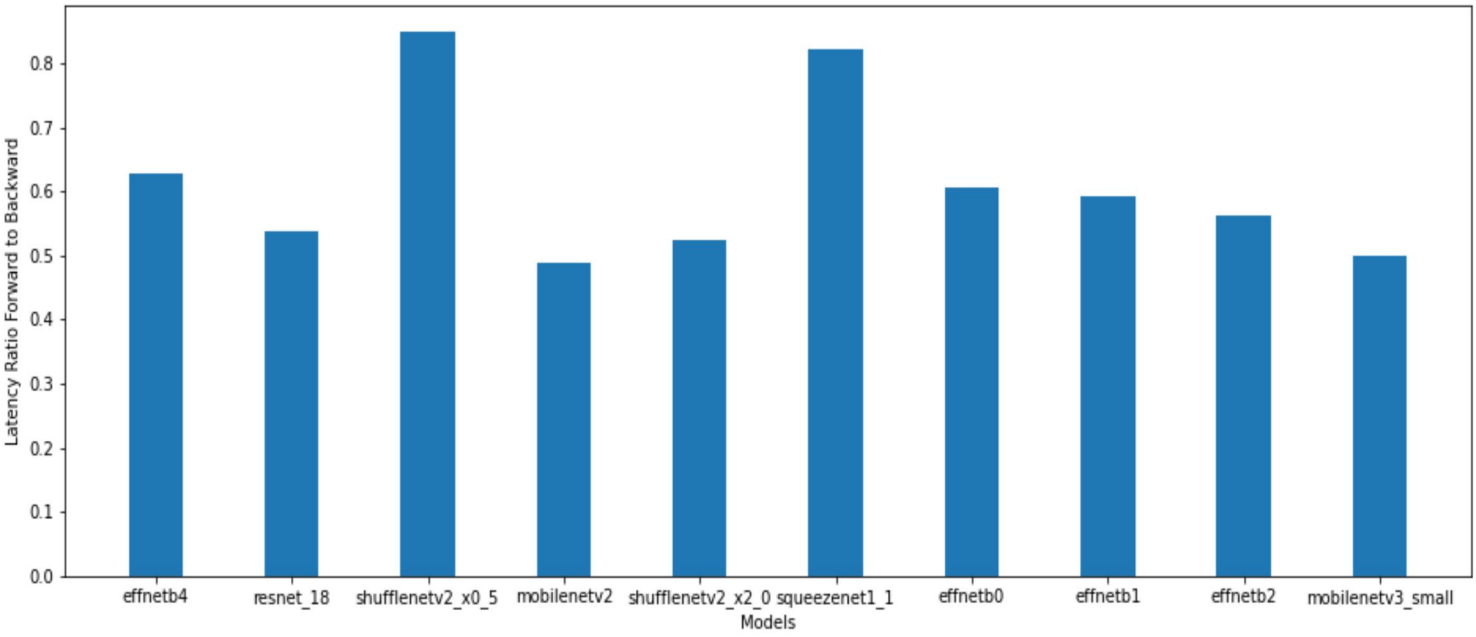


Figure 13: This bar chart shows the Latency ratio for forward pass vs backward pass for each model on CPU for a batch size of 1, input tensor of (3, 224, 224)

## 1 Submit your report 7.5 / 10

**- 0 pts** Correct

**- 0.25 pts** Use of log with flops-vs-latency is unjustified/unnecessary

**- 0.5 pts** Points are plotted above the roofline. Possible explanations are provided but the points are in some cases 2x higher than peak throughput which indicates the presence of a bug.

**- 0.25 pts** You attempted to justify the incorrect forward:backward ratio, but you did not root-cause the issue.

**- 0.5 pts** Plots are hard to read in more than one place

**- 2 pts** missing question 5

**- 2 pts** Rooflines are horizontal lines (incorrect)

**- 3 pts** Missing writeup, commentary and multiple plots are not visible

**- 0.5 pts** Bar charts and pie charts are hard to read

**- 0.5 pts** incorrect gpu roofline - mem BW looks wrong

**- 0.25 pts** Commentary is very brief or missing for some questions

**- 0.25 pts** noisy measurement

✓ **- 0.5 pts** *FLOPs ratio missing in Q5*

**- 0.75 pts** Pie charts in Q5 (mostly) missing or incorrect

✓ **- 0.5 pts** *rooflines are plotted as 2 separate lines*

✓ **- 0.5 pts** *Q5 pie charts missing*

**- 0.5 pts** forward:backward flops numbers way off in some cases

**- 0.5 pts** Rooflines are all in one plot making it hard to read them

**- 0.5 pts** GPU latency measurments are incorrect (should use events)

**- 0.5 pts** incorrect explanation

**- 1 pts** FLOPS/Latency ratios missing in Q5

**- 1 pts** Unlabeled axes in plots

**- 1 pts** Models are vertical lines on roofline plot where they are supposed to also include actual performance.

**- 2 pts** Rooflines are just points.

**- 1 pts** Incorrect roofline

**- 1 pts** op intensity needs to intersect with roofline

**- 0.5 pts** Points are plotted above the roofline. Explanation is not provided, indicates a bug.

✓ **- 1 pts** *Analysis of experiments insufficient.*

**- 0.25 pts** Correlation Matrix seems wrong

**- 0.25 pts** Models not distinguished in plots

**- 0.5 pts** Q4 Plot Missing/incorrect

**- 2 pts** Section 3 Missing

**- 2 pts** Section 4 Missing

**- 0.25 pts** Minor enhancements to plots possible. unclear in some places