# Caffe-HRT

Performance Report

2018-02-09

OPEN AI LAB

# Reversion Record

| Date | Rev | Change Description | Author |
|---|---|---|---|
| 2017-9-22 | 0.1.0 | Initial version | Joey |
| 2017-10-11 | 0.2.0 | Test on ACL v17.09 | Joey |
| 2017-11-28 | 0.3.0 | Test on ACL v17.10 | Huifang |
| 2018-01-25 | 0.5.0 | Test on ACL v17.12 | Huifang |

# catalog

# 1 Purpose

This Report is tested on RK3399 platform and the Arm Compute Library is version 17.12. The report includes both CPU data and GPU data. We collected the data on AlexNet, GoogLeNet, SqueezNet, MobileNet, ResNet18, ResNet34, ResNet50.Note that the CPU data is on a single A72 core. And we found the mixed mode can improve performance 2.8X for the best case.

# 2 Test Environment

Hardware SoC:  firefly

http://www.t-firefly.com/product/rk3399.html

- GPU: Mali T864 (800MHz)
- RAM: 2G
- CPU: Dual-core Cortex-A72 up to 2.0GHz (real frequency is 1.8GHz); Quad-core Cortex-A53 up to 1.5GHz (real frequency is 1.4GHz)

Operating System: Ubuntu 16.04

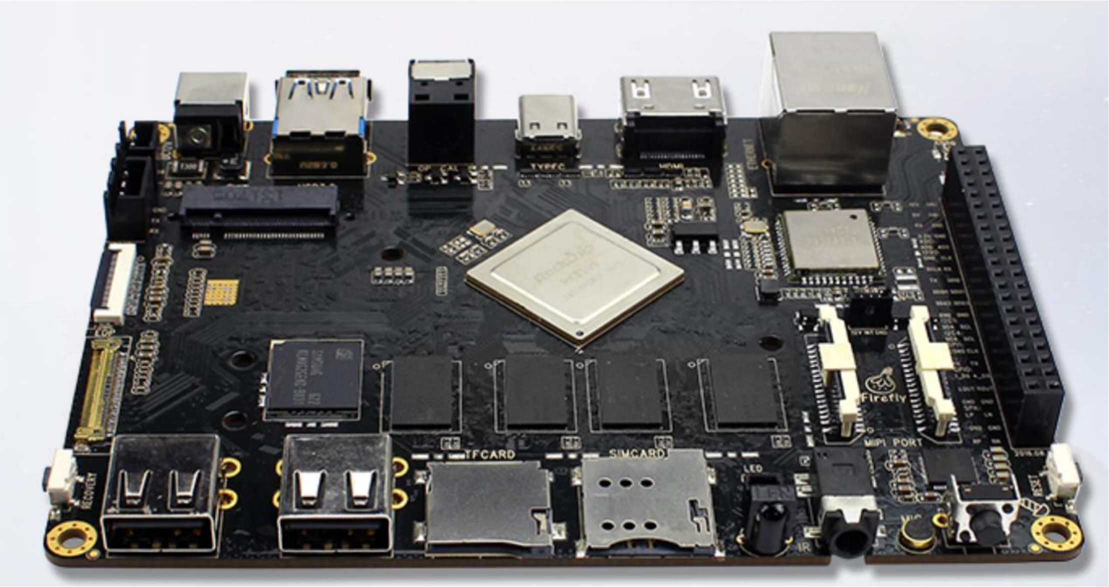

Figure 1 firefly board

# 3 Performance Improvement Achievement

The ACL_NEON's LRN and POOLING are better, and ACL_CL(GPU) has the better performances on large FC while OpenBLAS has better on CONV. It's possible to gain better performance on mixing the calculation on different component, for example, using OpenBLAS layers (SoftMax, RELU, FC, CONV) and ACL_NEON layers (LRN, Pooling) in neural network.

After we mixed the layers calculation on OpenBLAS and ACL, it's very easy to mix the layers calculation by exporting environment variable BYPASSACL, details in User Guide 5.2.

For the total time spent per inference, we have achieved about 2.81X performance in the best case.

Table 1 Performance comparation

| | Original Caffe (ms) | Caffe-HRT (ms) | Performance Gain (ms) |
|---|---|---|---|
| AlexNet | 932.70 | 534.80 | 1.74 |
| GoogleNet | 1387.70 | 494.00 | 2.81 |
| SquezzeNet | 144.30 | 144.30 | 1.00 |
| MobileNet | 305.00 | 292.80 | 1.04 |
| ResNet18 | 509.20 | 492.60 | 1.03 |
| ResNet34 | 1040.00 | 1024.30 | 1.02 |
| ResNet50 | 1095.30 | 1089.10 | 1.01 |

# 4 Performance

For GPU, the OpenCL driver need compile CL kernel for the first time running, but after 2nd time, the CL kernel may not be compiled. This will impact performance. Here we list the 1st data separately. We tested total 10 times from 2nd to 11th and calculated the average time. The data in the below tables are in the unit of second.

The items (TPI, Allocate, Run, Config, Copy, FC, CONV, LRN, Pooling, RELU, SOFTMAX) in the below tables:

- TPI: The total time for per inference
- Avg. Time: tested total 10 times from 2nd to 11th and calculated the average time.
- The unit of all the data columns in tests below is second.

The details see user manual section “Use Cases”.

## 4.1 AlexNet

Table 2 AlexNet Performance for configuration

| | TPI (s) | Allocate (s) | Run (s) | Config (s) | Copy (s) |
|---|---|---|---|---|---|
| 1st | | | | | |
| ACL/NEON | 3.2593 | 0.1744 | 2.7276 | 0.2189 | 0.1352 |
| OpenBLAS | 0.9527 | | | | |
| ACL/GPU | 2.4794 | 0.1822 | 0.0648 | 1.5011 | 0.7269 |
| MIXED | 0.5609 | 0.0046 | 0.0328 | 0.0013 | 0.0057 |
| Dynamic | 2.3379 | 0.1654 | 0.0616 | 1.3311 | 0.7747 |
| Avg. Time | | | | | |
| ACL/NEON | 0.5908 | | 0.5811 | | 0.0090 |
| OpenBLAS | 0.9327 | | | | |
| ACL/GPU | 0.1542 | | 0.0127 | | 0.1406 |
| MIXED | 0.5348 | | 0.0319 | | 0.0045 |
| Dynamic | 0.2033 | | 0.0096 | | 0.1916 |

Table 3 AlexNet performance for each Layer

| | TPI (s) | CONV (s) | FC (s) | LRN (s) | Pooling (s) | RELU (s) | SOFTMAX (s) |
|---|---|---|---|---|---|---|---|
| 1st | | | | | | | |
| ACL/NEON | 3.2593 | 0.3279 | 2.8777 | 0.0382 | 0.0072 | 0.0081 | 0.0003 |
| OpenBLAS | 0.9527 | 0.1703 | 0.3423 | 0.4292 | 0.0093 | 0.0014 | 0.0002 |
| ACL/GPU | 2.4794 | 1.1812 | 0.9399 | 0.0827 | 0.0770 | 0.0456 | 0.1529 |
| MIXED | 0.5609 | 0.1702 | 0.3436 | 0.0384 | 0.0070 | 0.0015 | 0.0002 |
| Dynamic | 2.3379 | 1.0402 | 0.9898 | 0.0792 | 0.0748 | 0.0014 | 0.1525 |
| Avg. Time | | | | | | | |
| ACL/NEON | 0.5908 | 0.1769 | 0.3734 | 0.0320 | 0.0046 | 0.0039 | 0.0001 |
| OpenBLAS | 0.9327 | 0.1556 | 0.3421 | 0.4253 | 0.0082 | 0.0015 | 0.0001 |
| ACL/GPU | 0.1542 | 0.0913 | 0.0374 | 0.0070 | 0.0081 | 0.0098 | 0.0005 |
| MIXED | 0.5348 | 0.1542 | 0.3425 | 0.0320 | 0.0046 | 0.0015 | 0.0001 |
| Dynamic | 0.2033 | 0.1339 | 0.0508 | 0.0081 | 0.0085 | 0.0014 | 0.0006 |

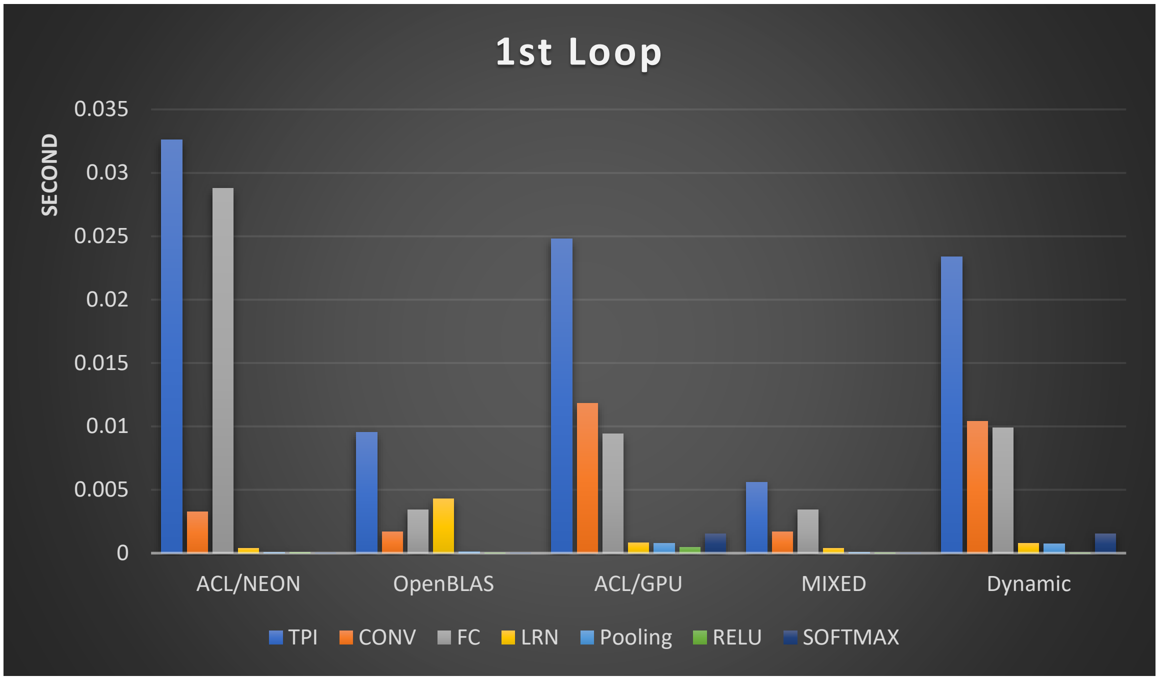

Figure 2 1st loop for AlexNet

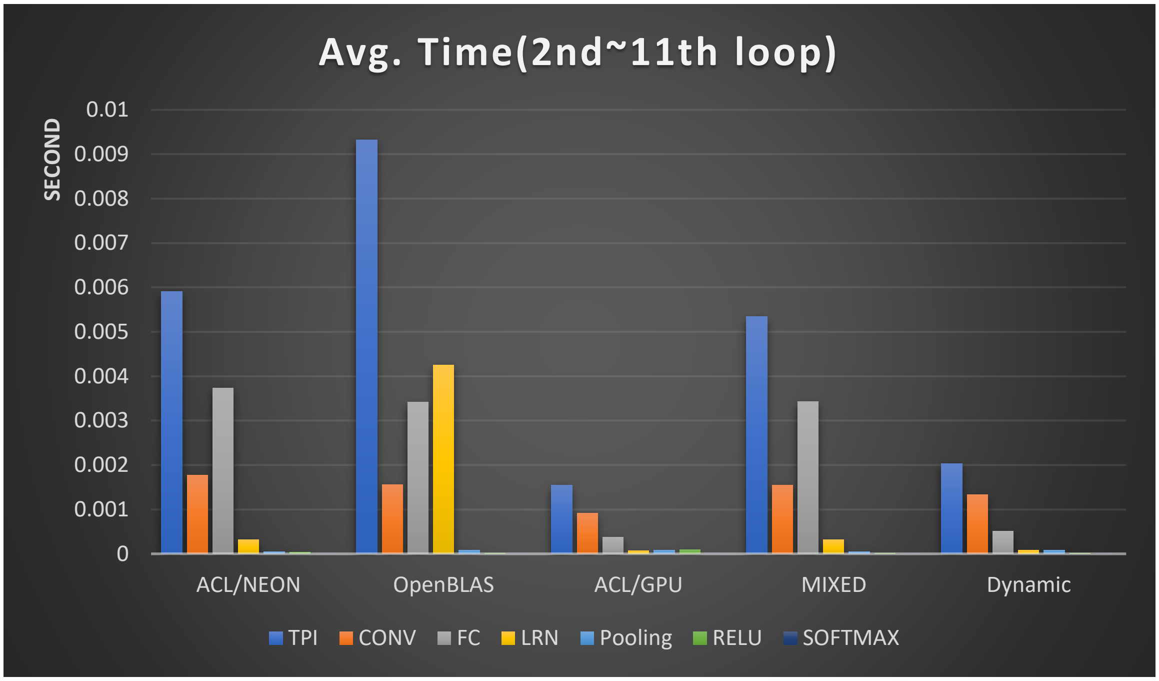

Figure 3 Avg. Time(2nd~11th loop) for AlexNet

## 4.2 GoogleNet

Table 4 GoogleNet Performance for configuration

| | TPI (s) | Allocate (s) | Run (s) | Config (s) | Copy (s) |
|---|---|---|---|---|---|
| 1st | | | | | |
| ACL/NEON | 1.1912 | 0.0850 | 0.6307 | 0.2243 | 0.2253 |
| OpenBLAS | 1.4333 | | | | |
| ACL/GPU | 4.9562 | 0.1154 | 0.1244 | 3.5634 | 1.1277 |
| MIXED | 0.5651 | 0.0249 | 0.0829 | 0.0033 | 0.0314 |
| Dynamic | 2.4819 | 0.0284 | 0.0331 | 1.7637 | 0.3478 |
| Avg. Time | | | | | |
| ACL/NEON | 0.6268 | | 0.5631 | | 0.0586 |
| OpenBLAS | 1.3877 | | | | |
| ACL/GPU | 0.6005 | | 0.0786 | | 0.5136 |
| MIXED | 0.4940 | | 0.0824 | | 0.0230 |
| Dynamic | 0.5724 | | 0.0190 | | 0.2719 |

Table 5 GoogleNet Performance for each layer

| | TPI (s) | CONV (s) | FC (s) | LRN (s) | Pooling (s) | RELU (s) | SOFTMAX (s) |
|---|---|---|---|---|---|---|---|
| 1st | | | | | | | |
| ACL/NEON | 1.1912 | 0.9765 | 0.0195 | 0.0635 | 0.0625 | 0.0385 | 0.0003 |
| OpenBLAS | 1.4333 | 0.4054 | 0.0043 | 0.9230 | 0.0882 | 0.0071 | 0.0002 |
| ACL/GPU | 4.9562 | 4.1194 | 0.1495 | 0.0841 | 0.2336 | 0.1447 | 0.1536 |
| MIXED | 0.5651 | 0.4051 | 0.0049 | 0.0629 | 0.0563 | 0.0072 | 0.0002 |
| Dynamic | 2.4819 | 1.8614 | 0.1487 | 0.0838 | 0.2233 | 0.0068 | 0.1531 |
| Avg. Time | | | | | | | |
| ACL/NEON | 0.6268 | 0.4938 | 0.0061 | 0.0542 | 0.0371 | 0.0175 | 0.0001 |
| OpenBLAS | 1.3877 | 0.3742 | 0.0045 | 0.9175 | 0.0818 | 0.0069 | 0.0001 |
| ACL/GPU | 0.6005 | 0.4611 | 0.0014 | 0.0113 | 0.0520 | 0.0515 | 0.0005 |
| MIXED | 0.4940 | 0.3752 | 0.0048 | 0.0537 | 0.0363 | 0.0068 | 0.0001 |
| Dynamic | 0.5724 | 0.4818 | 0.0023 | 0.0141 | 0.0647 | 0.0068 | 0.0006 |

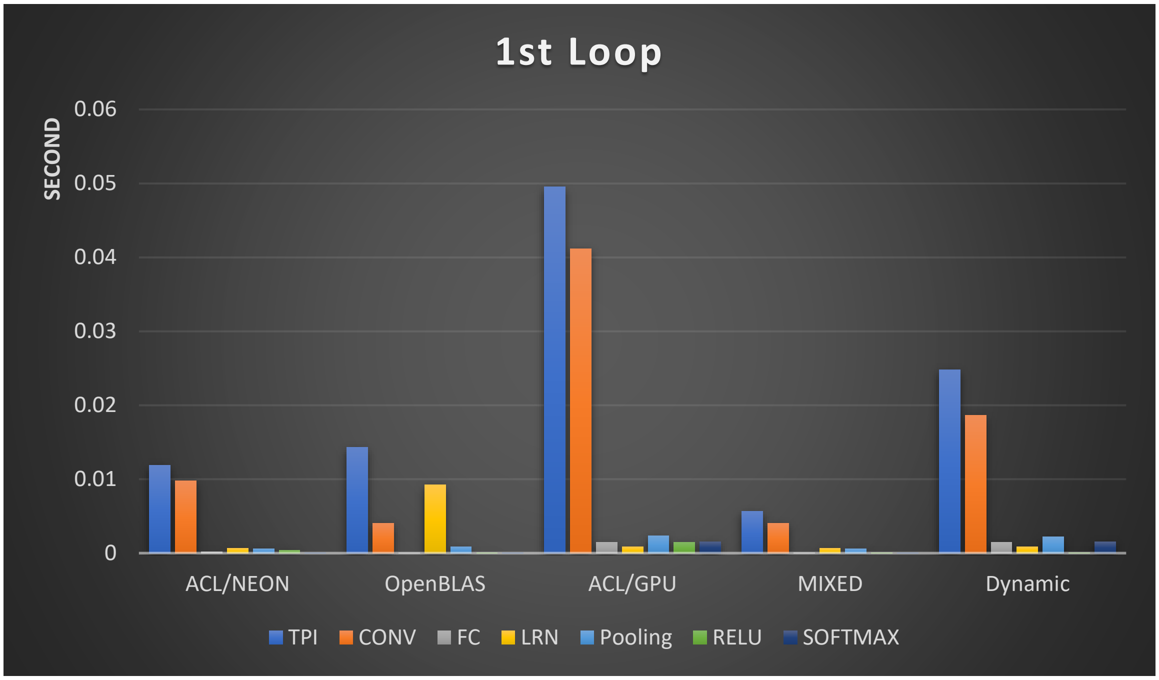

Figure 4 1st Loop for GoogleNet

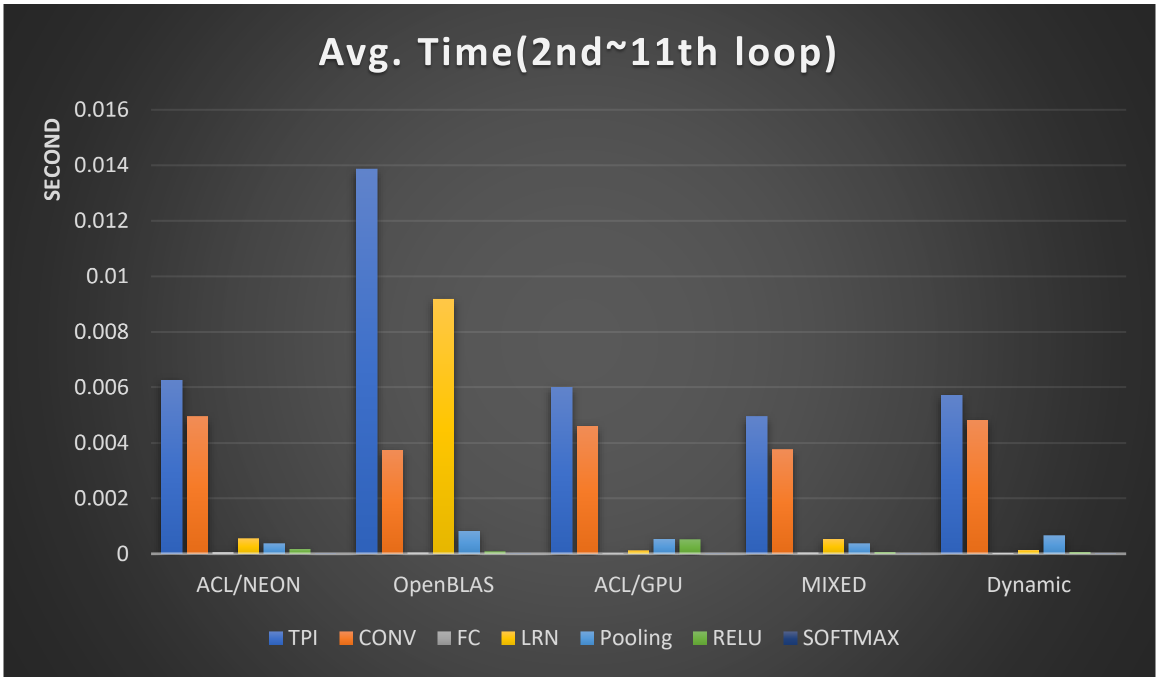

Figure 5 Avg. Time(2nd~11th loop) for GoogleNet

## 4.3 SqueezeNet

Table 6 SqueezeNet Performance for configuration

| | TPI (s) | Allocate (s) | Run (s) | Config (s) | Copy (s) |
|---|---|---|---|---|---|
| 1st | | | | | |
| ACL/NEON | 0.3977 | 0.0456 | 0.1723 | 0.0909 | 0.0799 |
| OpenBLAS | 0.1671 | | | | |
| ACL/GPU | 2.7463 | 0.0391 | 0.0448 | 2.2530 | 0.3988 |
| MIXED | 0.1887 | 0.0139 | 0.0193 | 0.0006 | 0.0181 |
| Dynamic | 0.4191 | 0.0017 | 0.0039 | 0.2581 | 0.0154 |
| Avg. Time | | | | | |
| ACL/NEON | 0.1950 | | 0.1602 | | 0.0320 |
| OpenBLAS | 0.1443 | | | | |
| ACL/GPU | 0.3377 | | 0.0301 | | 0.3037 |
| MIXED | 0.1503 | | 0.0190 | | 0.0127 |
| Dynamic | 0.1407 | | 0.0039 | | 0.0166 |

Table 7 SqueezeNet Performance for each layer

| | TPI (s) | CONV (s) | FC (s) | LRN (s) | Pooling (s) | RELU (s) | SOFTMAX (s) |
|---|---|---|---|---|---|---|---|
| 1st | | | | | | | |
| ACL/NEON | 0.3977 | 0.3117 | | | 0.0197 | 0.0320 | 0.0003 |
| OpenBLAS | 0.1671 | 0.1283 | | | 0.0260 | 0.0061 | 0.0002 |
| ACL/GPU | 2.7463 | 2.3229 | | | 0.1135 | 0.0954 | 0.1519 |
| MIXED | 0.1887 | 0.1283 | | | 0.0190 | 0.0059 | 0.0002 |
| Dynamic | 0.4191 | 0.1262 | | | 0.1314 | 0.0058 | 0.1493 |
| Avg. Time | | | | | | | |
| ACL/NEON | 0.1950 | 0.1478 | | | 0.0114 | 0.0151 | 0.0001 |
| OpenBLAS | 0.1443 | 0.1108 | | | 0.0245 | 0.0058 | 0.0001 |
| ACL/GPU | 0.3377 | 0.2540 | | | 0.0185 | 0.0385 | 0.0009 |
| MIXED | 0.1503 | 0.1114 | | | 0.0117 | 0.0058 | 0.0001 |
| Dynamic | 0.1407 | 0.1103 | | | 0.0203 | 0.0059 | 0.0013 |

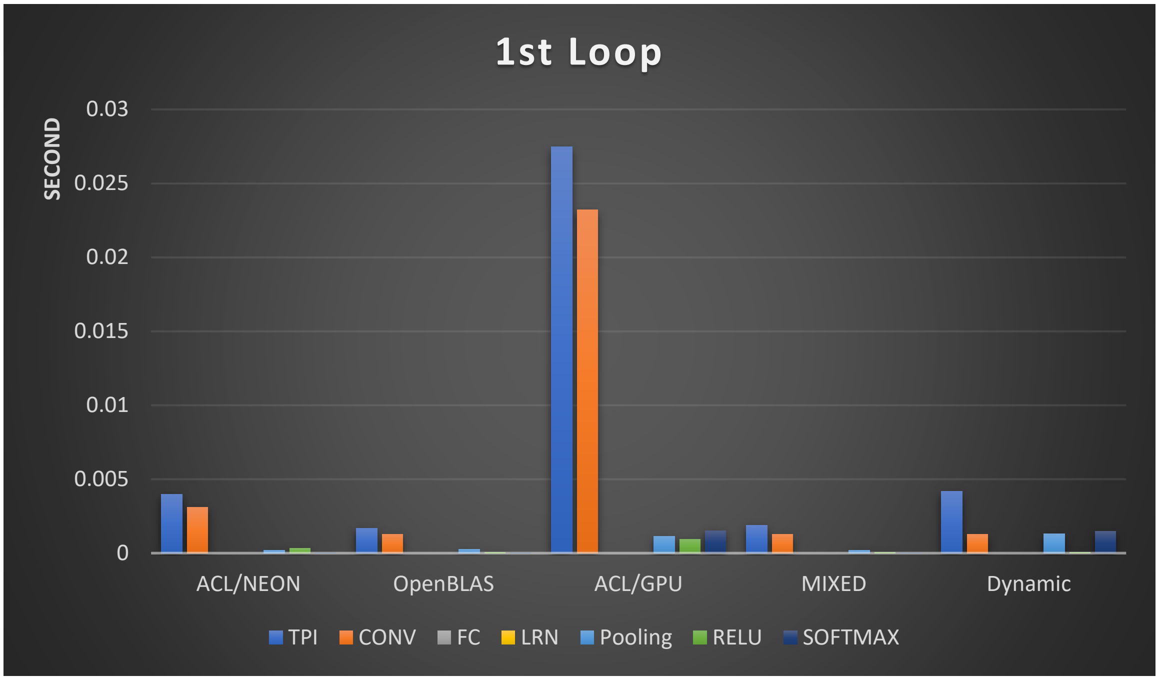

Figure 6 1st Loop for SqueezeNet

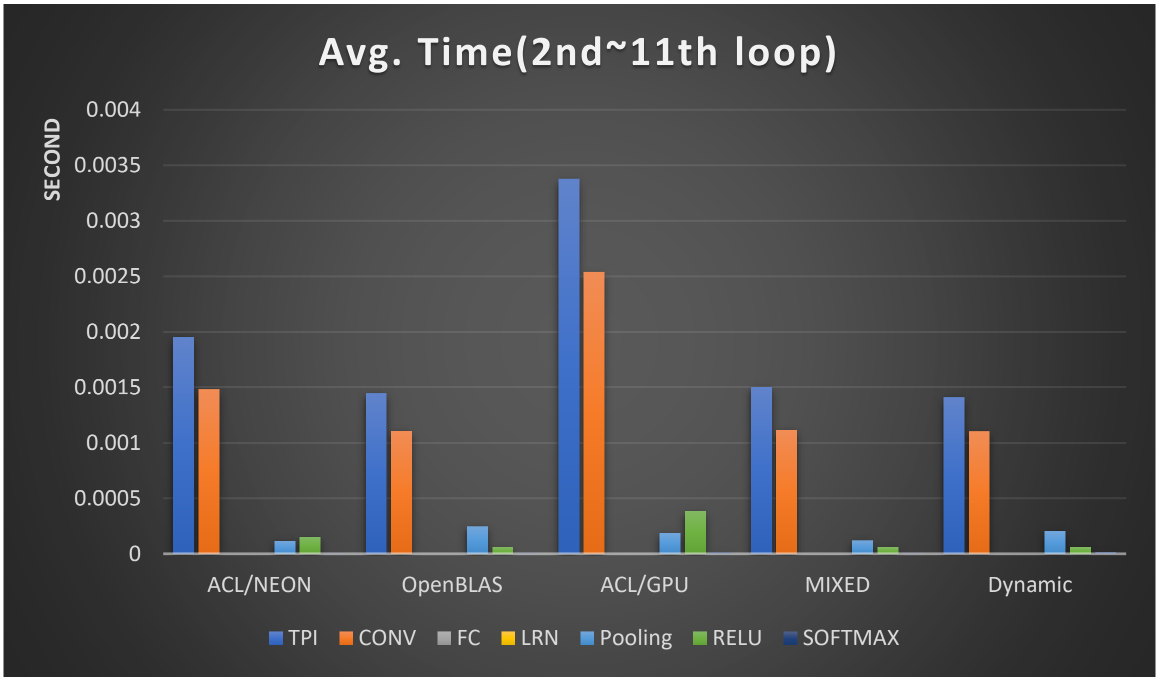

Figure 7 Avg. Time(2nd~11th loop) for SqueezeNet

## 4.4 MobileNet

Table 8 MobileNet Performance for configuration

| | TPI (s) | Allocate (s) | Run (s) | Config (s) | Copy (s) |
|---|---|---|---|---|---|
| 1st | | | | | |
| ACL/NEON | 0.7769 | 0.0823 | 0.2627 | 0.0788 | 0.2096 |
| OpenBLAS | 0.3874 | | | | |
| ACL/GPU | 2.3577 | 0.0758 | 0.0407 | 1.4363 | 0.6599 |
| MIXED | 0.3911 | 0.0291 | 0.0257 | 0.0006 | 0.0314 |
| Dynamic | 0.3818 | | | | |
| Avg. Time | | | | | |
| ACL/NEON | 0.3717 | | 0.2229 | | 0.0590 |
| OpenBLAS | 0.3050 | | | | |
| ACL/GPU | 0.5565 | | 0.0399 | | 0.4244 |
| MIXED | 0.2928 | | 0.0244 | | 0.0300 |
| Dynamic | 0.2957 | | | | |

Table 9 MobileNet Performance for each layer

| | TPI (s) | CONV (s) | FC (s) | LRN (s) | Pooling (s) | RELU (s) | SOFTMAX (s) |
|---|---|---|---|---|---|---|---|
| 1st | | | | | | | |
| ACL/NEON | 0.7769 | 0.6268 | | | 0.0006 | 0.0619 | |
| OpenBLAS | 0.3874 | 0.2897 | | | 0.0005 | 0.0111 | |
| ACL/GPU | 2.3577 | 2.0609 | | | 0.0006 | 0.1270 | |
| MIXED | 0.3911 | 0.2897 | | | 0.0006 | 0.0112 | |
| Dynamic | 0.3818 | 0.2850 | | | 0.0006 | 0.0108 | |
| Avg. Time | | | | | | | |
| ACL/NEON | 0.3717 | 0.2836 | | | 0.0005 | 0.0324 | |
| OpenBLAS | 0.3050 | 0.2271 | | | 0.0005 | 0.0108 | |
| ACL/GPU | 0.5565 | 0.3976 | | | 0.0005 | 0.0655 | |
| MIXED | 0.2928 | 0.2260 | | | 0.0005 | 0.0108 | |
| Dynamic | 0.2957 | 0.2216 | | | 0.0005 | 0.0108 | |

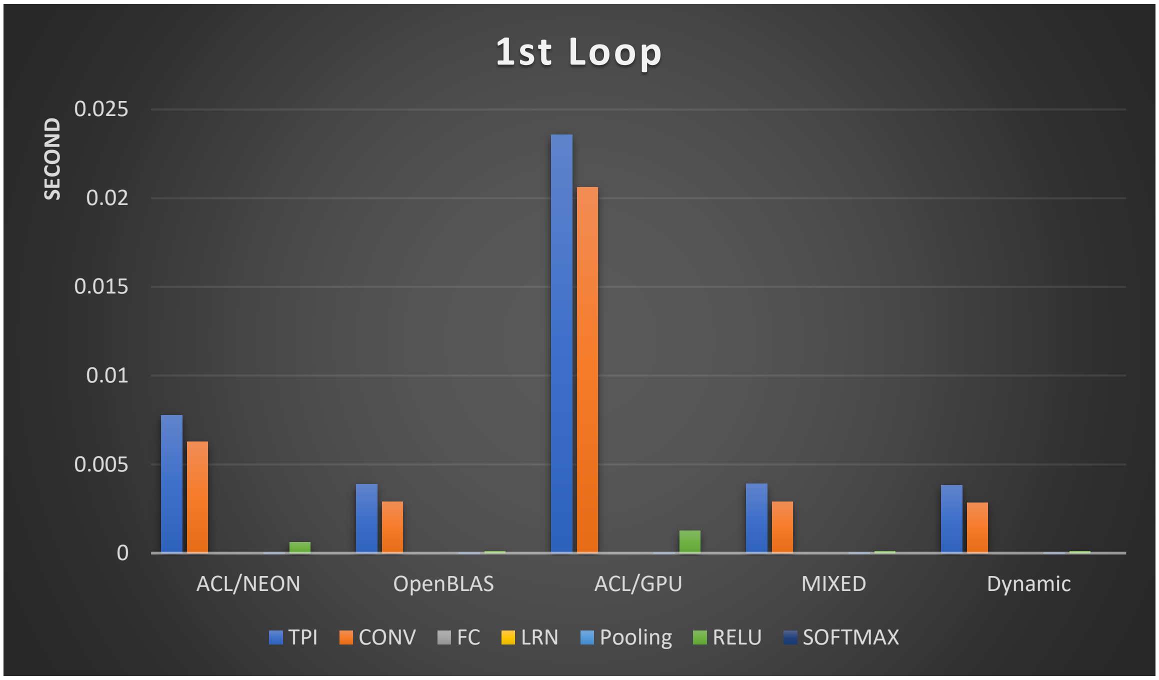

Figure 8 1st Loop for MobileNet

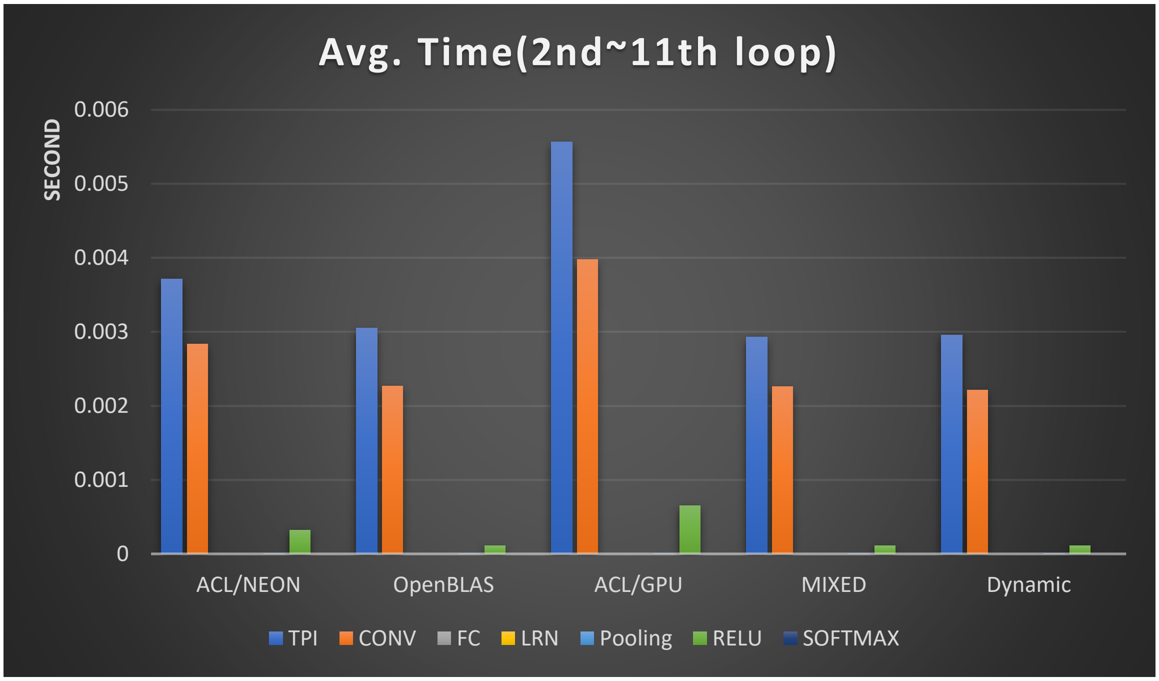

Figure 9 Avg. Time(2nd~11th loop) for MobileNet

## 4.5 ResNet18

Table 10 ResNet18 Performance for configuration

| | TPI (s) | Allocate (s) | Run (s) | Config (s) | Copy (s) |
|---|---|---|---|---|---|
| 1st | | | | | |
| ACL/NEON | 1.1815 | 0.0768 | 0.6987 | 0.2222 | 0.1755 |
| OpenBLAS | 0.5649 | | | | |
| ACL/GPU | 1.1847 | 0.0766 | 0.7012 | 0.2228 | 0.1759 |
| MIXED | 0.5684 | 0.0197 | 0.0218 | 0.0023 | 0.0182 |
| Dynamic | 2.1104 | 0.0294 | 0.0236 | 0.9794 | 0.7553 |
| Avg. Time | | | | | |
| ACL/NEON | 0.6183 | | 0.5783 | | 0.0375 |
| OpenBLAS | 0.5092 | | | | |
| ACL/GPU | 0.6197 | | 0.5796 | | 0.0376 |
| MIXED | 0.4926 | | 0.0179 | | 0.0166 |
| Dynamic | 0.5937 | | 0.0059 | | 0.3074 |

Table 11 ResNet18 Performance for each layer

| | TPI (s) | CONV (s) | FC (s) | LRN (s) | Pooling (s) | RELU (s) | SOFTMAX (s) |
|---|---|---|---|---|---|---|---|
| 1st | | | | | | | |
| ACL/NEON | 1.1815 | 1.0910 | 0.0100 | | 0.0090 | 0.0270 | 0.0002 |
| OpenBLAS | 0.5649 | 0.4943 | 0.0022 | | 0.0150 | 0.0049 | 0.0002 |
| ACL/GPU | 1.1847 | 1.0940 | 0.0101 | | 0.0091 | 0.0271 | 0.0003 |
| MIXED | 0.5684 | 0.4987 | 0.0101 | | 0.0094 | 0.0050 | 0.0002 |
| Dynamic | 2.1104 | 1.6923 | 0.1416 | | 0.0735 | 0.0048 | 0.1519 |
| Avg. Time | | | | | | | |
| ACL/NEON | 0.6183 | 0.5679 | 0.0029 | | 0.0059 | 0.0138 | 0.0001 |
| OpenBLAS | 0.5092 | 0.4497 | 0.0021 | | 0.0144 | 0.0049 | 0.0001 |
| ACL/GPU | 0.6197 | 0.5692 | 0.0029 | | 0.0059 | 0.0139 | 0.0001 |
| MIXED | 0.4926 | 0.4523 | 0.0026 | | 0.0060 | 0.0048 | 0.0001 |
| Dynamic | 0.5937 | 0.5394 | 0.0016 | | 0.0095 | 0.0048 | 0.0006 |

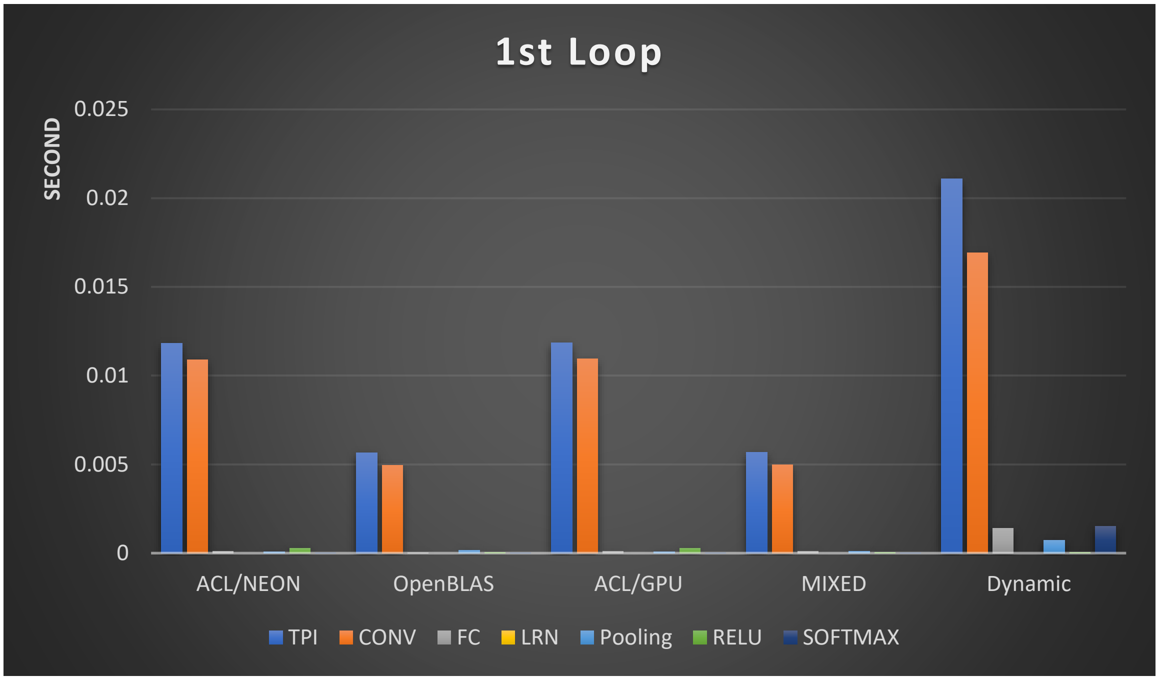

Figure 10 1st Loop for ResNet18

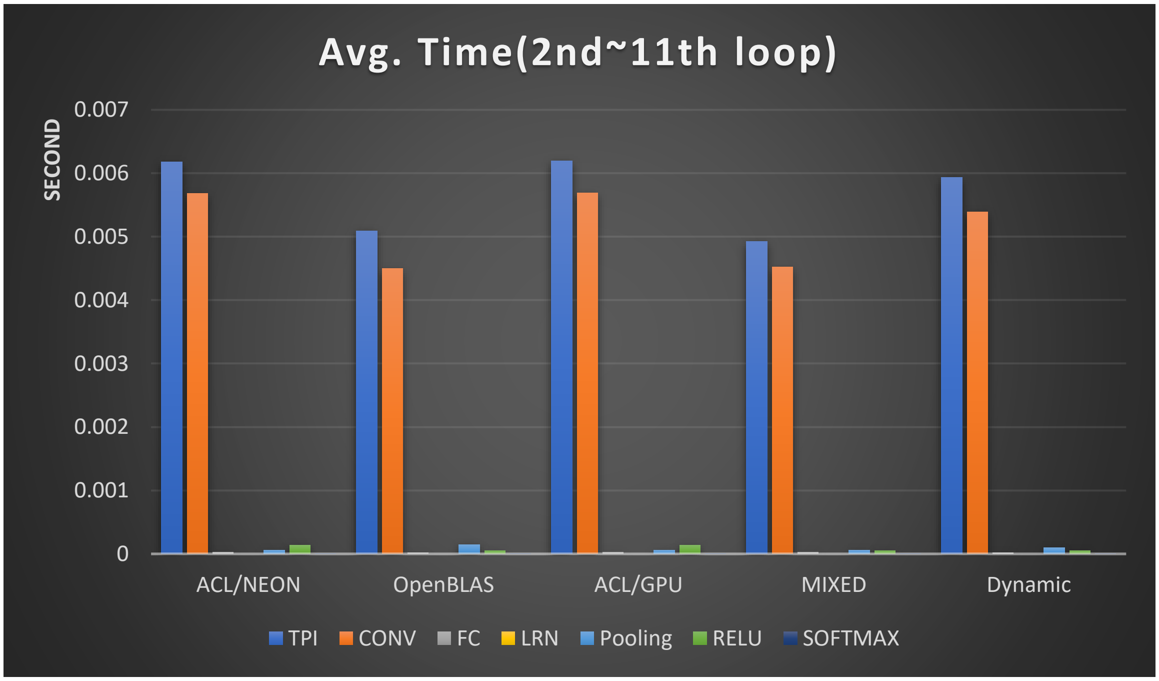

Figure 11 Avg. Time(2nd~11th loop) for ResNet18

## 4.6 ResNet34

Table 12 ResNet34 Performance for configuration

| | TPI (s) | Allocate (s) | Run (s) | Config (s) | Copy (s) |
|---|---|---|---|---|---|
| 1st | | | | | |
| ACL/NEON | 2.1758 | 0.1026 | 1.2860 | 0.3670 | 0.2691 |
| OpenBLAS | 1.1373 | | | | |
| ACL/GPU | 2.2497 | 0.1023 | 1.2864 | 0.3659 | 0.3324 |
| MIXED | 1.1341 | 0.0043 | 0.0094 | 0.0019 | 0.0042 |
| Dynamic | 3.3660 | 0.0624 | 0.0483 | 1.0911 | 1.5508 |
| Avg. Time | | | | | |
| ACL/NEON | 1.2212 | | 1.0666 | | 0.0314 |
| OpenBLAS | 1.0400 | | | | |
| ACL/GPU | 1.2230 | | 1.0651 | | 0.0326 |
| MIXED | 1.0243 | | 0.0057 | | 0.0027 |
| Dynamic | 1.1603 | | 0.0118 | | 0.6130 |

Table 13 ResNet34 Performance for each layer

| | TPI (s) | CONV (s) | FC (s) | LRN (s) | Pooling (s) | RELU (s) | SOFTMAX (s) |
|---|---|---|---|---|---|---|---|
| 1st | | | | | | | |
| ACL/NEON | 2.1758 | 1.9753 | 0.0100 | | 0.0089 | 0.0408 | 0.0003 |
| OpenBLAS | 1.1373 | 0.9692 | 0.0023 | | 0.0154 | 0.0075 | 0.0001 |
| ACL/GPU | 2.2497 | 2.0376 | 0.0102 | | 0.0090 | 0.0408 | 0.0002 |
| MIXED | 1.1341 | 0.9630 | 0.0101 | | 0.0103 | 0.0075 | 0.0002 |
| Dynamic | 3.3660 | 2.8446 | 0.1420 | | 0.0751 | 0.0074 | 0.1529 |
| Avg. Time | | | | | | | |
| ACL/NEON | 1.2212 | 1.0716 | 0.0030 | | 0.0058 | 0.0204 | 0.0001 |
| OpenBLAS | 1.0400 | 0.8942 | 0.0022 | | 0.0147 | 0.0076 | 0.0001 |
| ACL/GPU | 1.2230 | 1.0713 | 0.0030 | | 0.0058 | 0.0205 | 0.0001 |
| MIXED | 1.0243 | 0.8868 | 0.0026 | | 0.0061 | 0.0074 | 0.0001 |
| Dynamic | 1.1603 | 1.0178 | 0.0016 | | 0.0096 | 0.0074 | 0.0006 |

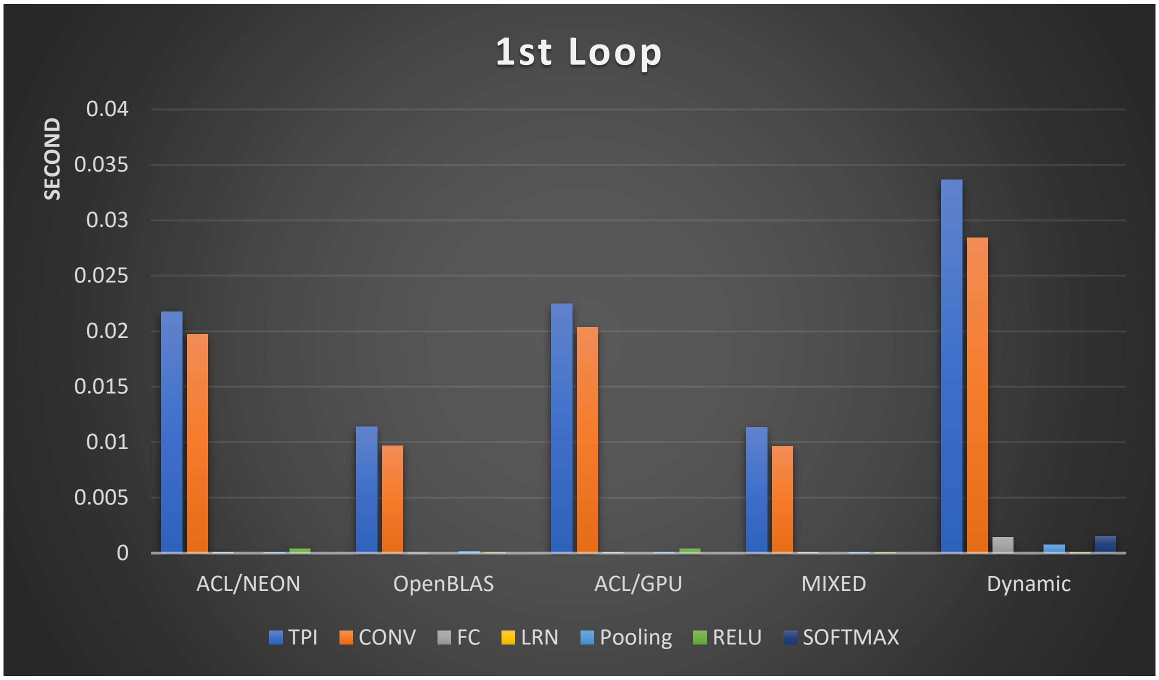

Figure 12 1st Loop for ResNet34

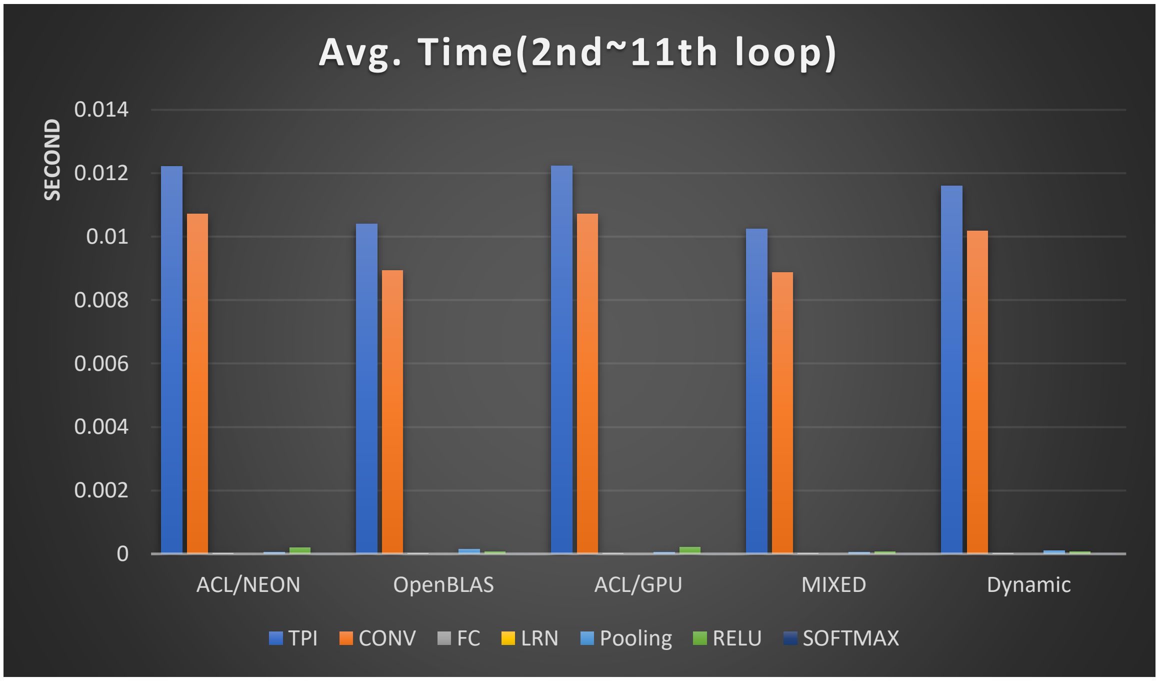

Figure 13 Avg. Time (2nd~11th loop) for ResNet34

## 4.7 ResNet50

Table 14 ResNet50 Performance for configuration

| | TPI<br>(s) | Allocate<br>(s) | Run<br>(s) | Config<br>(s) | Copy<br>(s) |
|---|---|---|---|---|---|
| 1st | | | | | |
| ACL/NEON | 3.1285 | 0.2472 | 1.5828 | 0.4180 | 0.8583 |
| OpenBLAS | 1.1975 | | | | |
| ACL/GPU | 3.1052 | 0.2407 | 1.5751 | 0.4076 | 0.8597 |
| MIXED | 1.2510 | 0.0694 | 0.0930 | 0.0071 | 0.0666 |
| Dynamic | 2.8057 | 0.0368 | 0.0281 | 0.9983 | 0.8684 |
| Avg. Time | | | | | |
| ACL/NEON | 1.4490 | | 1.2924 | | 0.1495 |
| OpenBLAS | 1.0953 | | | | |
| ACL/GPU | 1.4514 | | 1.2940 | | 0.1501 |
| MIXED | 1.0891 | | 0.0782 | | 0.0628 |
| Dynamic | 1.2521 | | 0.0071 | | 0.4667 |

Table 15 ResNet50 Performance for each layer

| | TPI<br>(s) | CONV<br>(s) | FC<br>(s) | LRN<br>(s) | Pooling<br>(s) | RELU<br>(s) | SOFTMAX<br>(s) |
|---|---|---|---|---|---|---|---|
| 1st | | | | | | | |
| ACL/NEON | 3.1285 | 2.7772 | 0.0419 | | 0.0097 | 0.1088 | 0.0002 |
| OpenBLAS | 1.1975 | 0.9830 | 0.0095 | | 0.0152 | 0.0190 | 0.0001 |
| ACL/GPU | 3.1052 | 2.7546 | 0.0431 | | 0.0098 | 0.1080 | 0.0003 |
| MIXED | 1.2510 | 0.9888 | 0.0416 | | 0.0103 | 0.0190 | 0.0002 |
| Dynamic | 2.8057 | 2.2338 | 0.1604 | | 0.0757 | 0.0187 | 0.1514 |
| Avg. Time | | | | | | | |
| ACL/NEON | 1.4490 | 1.2501 | 0.0120 | | 0.0066 | 0.0551 | 0.0001 |
| OpenBLAS | 1.0953 | 0.9224 | 0.0096 | | 0.0146 | 0.0191 | 0.0001 |
| ACL/GPU | 1.4514 | 1.2518 | 0.0120 | | 0.0067 | 0.0551 | 0.0001 |
| MIXED | 1.0891 | 0.9258 | 0.0123 | | 0.0070 | 0.0193 | 0.0001 |
| Dynamic | 1.2521 | 1.0935 | 0.0043 | | 0.0109 | 0.0188 | 0.0007 |

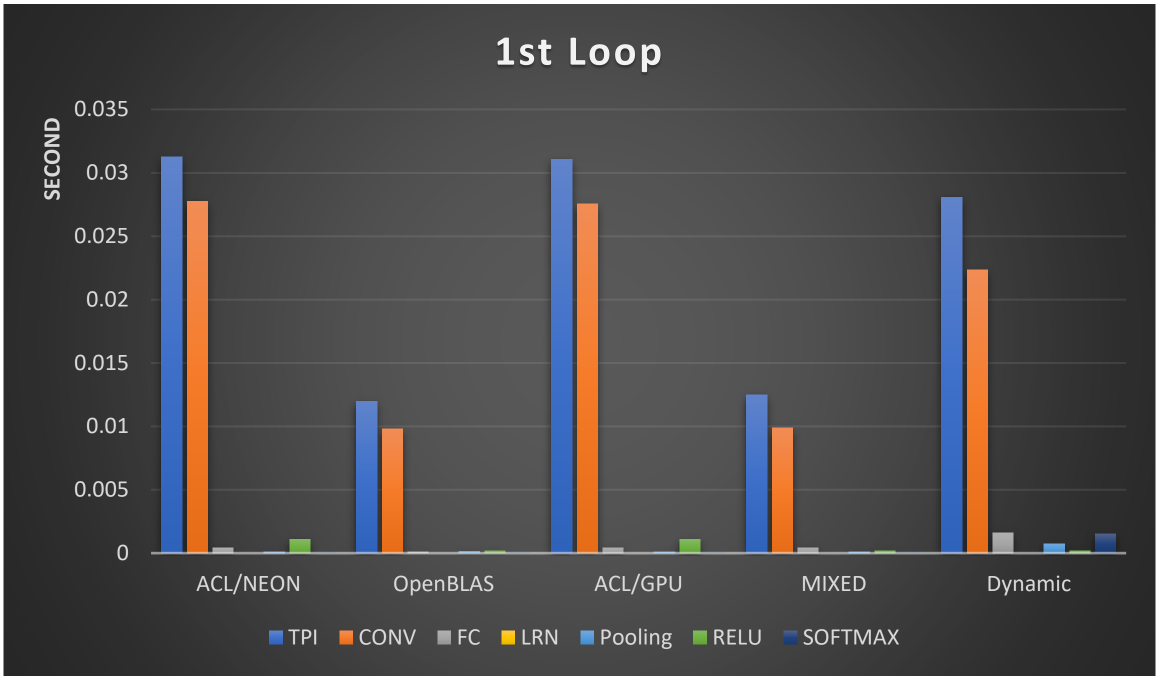

Figure 14 1st Loop for ResNet50

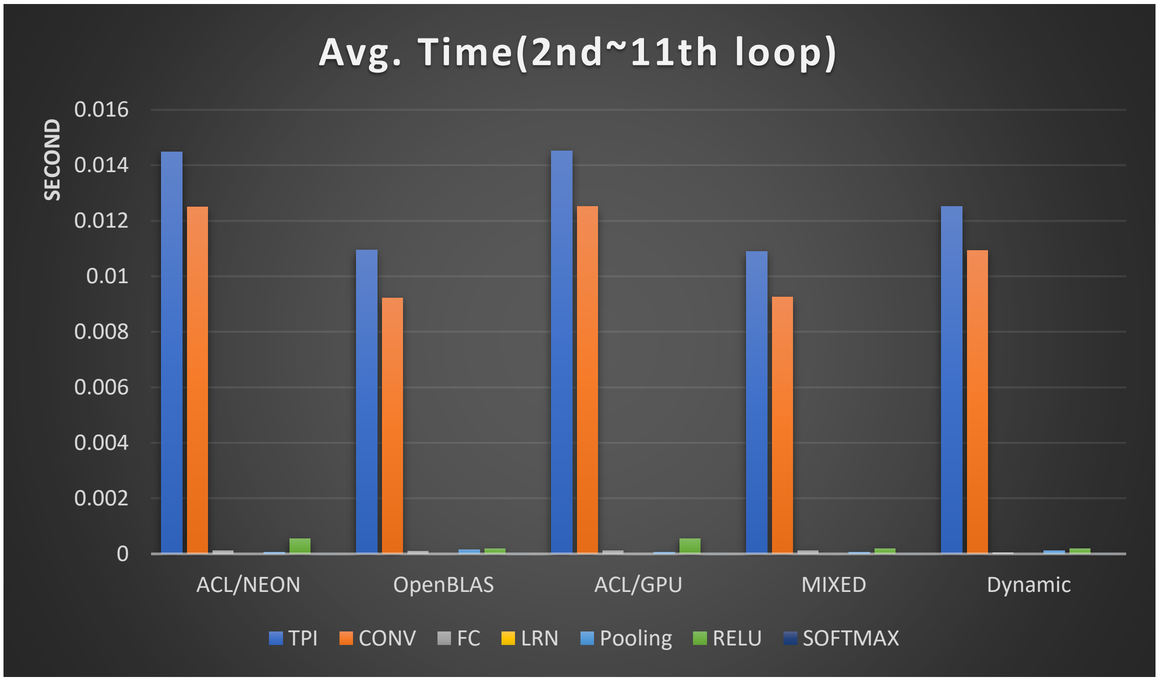

Figure 15 Avg. Time(2nd~11th loop) for ResNet50

# 5 Performance On Different Cores

The TPI is not very stable, it’s in wide fluctuation. The data in the tables is lower limit of the range.

## 5.1 The TPI Data For ACL/NEON, OpenBLAS And Mixed Mode

AlexNet TPI data for ACL/NEON, OpenBLAS and mixed mode

Table 16 AlexNet TPI data

| | ACL/NEON (s) | OpenBLAS (s) | MIXED (s) | Dynamic (s) |
|---|---|---|---|---|
| 1xA53 | 1.9461 | 1.8653 | 0.9466 | 0.2606 |
| 1xA72 | 0.5908 | 0.9327 | 0.5348 | 0.2033 |
| 2xA72 | 0.3256 | 0.8736 | 0.4534 | 0.1731 |
| 4xA53 | 0.5987 | 1.6108 | 0.6669 | 0.3181 |
| 2xA72+4xA53* | 0.4166 | 0.8930 | 0.6584 | 0.2493 |

GoogleNet TPI data for ACL/NEON, OpenBLAS and mixed mode

Table 17 GoogleNet TPI data

| | ACL/NEON (s) | OpenBLAS (s) | MIXED (s) | Dynamic (s) |
|---|---|---|---|---|
| 1xA53 | 1.2299 | 3.3694 | 1.2992 | 1.0878 |
| 1xA72 | 0.6268 | 1.3877 | 0.4940 | 0.5724 |
| 2xA72 | 0.4039 | 1.2322 | 0.3459 | 0.4550 |
| 4xA53 | 0.8351 | 2.7240 | 0.6335 | 0.6078 |
| 2xA72+4xA53* | 0.7288 | 1.7410 | 0.3653 | 0.5355 |

MobileNet TPI data for ACL/NEON, OpenBLAS and mixed mode

Table 18 MobileNet TPI data

| | ACL/NEON (s) | OpenBLAS (s) | MIXED (s) | Dynamic (s) |
|---|---|---|---|---|
| 1xA53 | 0.7996 | 0.8226 | 0.7645 | 1.0878 |
| 1xA72 | 0.3717 | 0.3050 | 0.2928 | 0.5724 |
| 2xA72 | 0.3662 | 0.2325 | 0.2419 | 0.4550 |
| 4xA53 | 0.6415 | 0.5526 | 0.5204 | 0.6078 |
| 2xA72+4xA53* | 0.3526 | 0.2388 | 0.2451 | 0.5355 |

SqueezeNet TPI data for ACL/NEON, OpenBLAS and mixed mode.

Table 19 SqueezeNet TPI data

| | ACL/NEON (s) | OpenBLAS (s) | MIXED (s) | Dynamic (s) |
|---|---|---|---|---|
| 1xA53 | 0.4142 | 0.3690 | 0.4016 | 0.3394 |
| 1xA72 | 0.1950 | 0.1443 | 0.1503 | 0.1407 |
| 2xA72 | 0.1389 | 0.1004 | 0.1080 | 0.0979 |
| 4xA53 | 0.3253 | 0.1942 | 0.2235 | 0.1645 |
| 2xA72+4xA53* | 0.2133 | 0.1048 | 0.2060 | 0.0980 |

ResNet18 TPI data for ACL/NEON, OpenBLAS and mixed mode.

Table 20 ResNet18 TPI data

| | ACL/NEON (s) | OpenBLAS (s) | MIXED (s) | Dynamic (s) |
|---|---|---|---|---|
| 1xA53 | 1.1766 | 1.3175 | 1.2812 | 1.0980 |
| 1xA72 | 0.6183 | 0.5092 | 0.4926 | 0.5937 |
| 2xA72 | 0.7495 | 0.9831 | 0.7672 | 0.4562 |
| 4xA53 | 0.7411 | 0.5978 | 0.5561 | 0.6561 |
| 2xA72+4xA53* | 0.6469 | 0.3839 | 0.3678 | 0.6013 |

ResNet34 TPI data for ACL/NEON, OpenBLAS and mixed mode.

Table 21 ResNet34 TPI data

| | ACL/NEON (s) | OpenBLAS (s) | MIXED (s) | Dynamic (s) |
|---|---|---|---|---|
| 1xA53 | 2.4465 | 2.8463 | 2.7018 | 2.2180 |
| 1xA72 | 1.2212 | 1.0400 | 1.0243 | 1.1603 |
| 2xA72 | 1.6986 | 1.5154 | 0.7037 | 0.8280 |
| 4xA53 | 1.7216 | 1.2313 | 1.2174 | 1.4044 |
| 2xA72+4xA53* | 1.3393 | 0.7840 | 0.7745 | 1.3369 |

ResNet50 TPI data for ACL/NEON, OpenBLAS and mixed mode.

Table 22 ResNet50 TPI data

| | ACL/NEON (s) | OpenBLAS (s) | MIXED (s) | Dynamic (s) |
|---|---|---|---|---|
| 1xA53 | 1.4386 | 2.9243 | 2.9795 | 2.6174 |
| 1xA72 | 1.4490 | 1.0953 | 1.0891 | 1.2521 |
| 2xA72 | 1.2705 | 0.7483 | 0.7406 | 0.9774 |
| 4xA53 | 2.0314 | 1.3296 | 1.2971 | 1.4878 |
| 2xA72+4xA53* | 1.5213 | 0.7921 | 1.0883 | 1.3830 |

## 5.2 The TPI In Mixed mode

The TPI data for different CPU cores in mixed mode:

Table 23 The TPI data for different

| | AlexNet (s) | GoogleNet (s) | MobileNet (s) | Squeeze Net(s) | ResNet 18(s) | ResNet 34(s) | Reset 50(s) |
|---|---|---|---|---|---|---|---|
| 1xA53 | 0.2606 | 1.0878 | 0.8048 | 0.3394 | 1.0980 | 2.2180 | 2.6174 |
| 1xA72 | 0.2033 | 0.5724 | 0.2957 | 0.1407 | 0.5937 | 1.1603 | 1.2521 |
| 2xA72 | 0.1731 | 0.4550 | 0.2311 | 0.0979 | 0.4562 | 0.8280 | 0.9774 |
| 4xA53 | 0.3181 | 0.6078 | 0.5532 | 0.1645 | 0.6561 | 1.4044 | 1.4878 |
| 2xA72+4xA53 | 0.2493 | 0.5355 | 0.2382 | 0.0980 | 0.6013 | 1.3369 | 1.3830 |

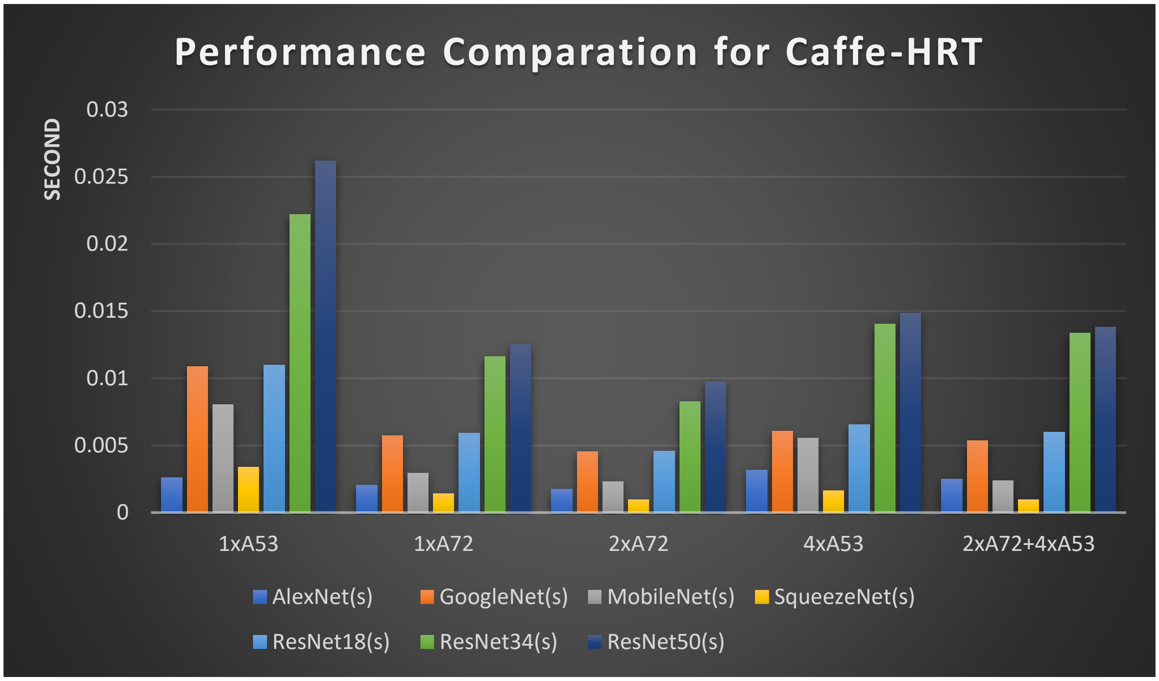

Figure 16 Performance Comparation

# 6 Conclusion

From the above test cases, we can deduce that: the performances of large FC are better under ACL_CL(GPU) than under NEON and OpenBLAS.

Table 24 Performance of FC layer for different models

| | Alex Net(s) | Google Net(s) | Squeeze Net(s) | Mobile Net(s) | ResNet18 Net(s) | ResNet34 Net(s) | ResNet50 Net(s) |
|---|---|---|---|---|---|---|---|
| FC/ACL/NEON | 0.1942 | 0.0061 | 0 | 0 | 0.0100 | 0.0030 | 0.0120 |
| FC/OpenBLAS | 0.3356 | 0.0045 | 0 | 0 | 0.0022 | 0.0022 | 0.0096 |
| FC/ACL/GPU | 0.0530 | 0.0061 | 0 | 0 | 0.0101 | 0.0030 | 0.0120 |