MXNetOnACL

Performance Report

2018-01-26

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
2018-01-26	0.3.0	Test on ACL v17.12	Huifang

catalog

1 Purpose

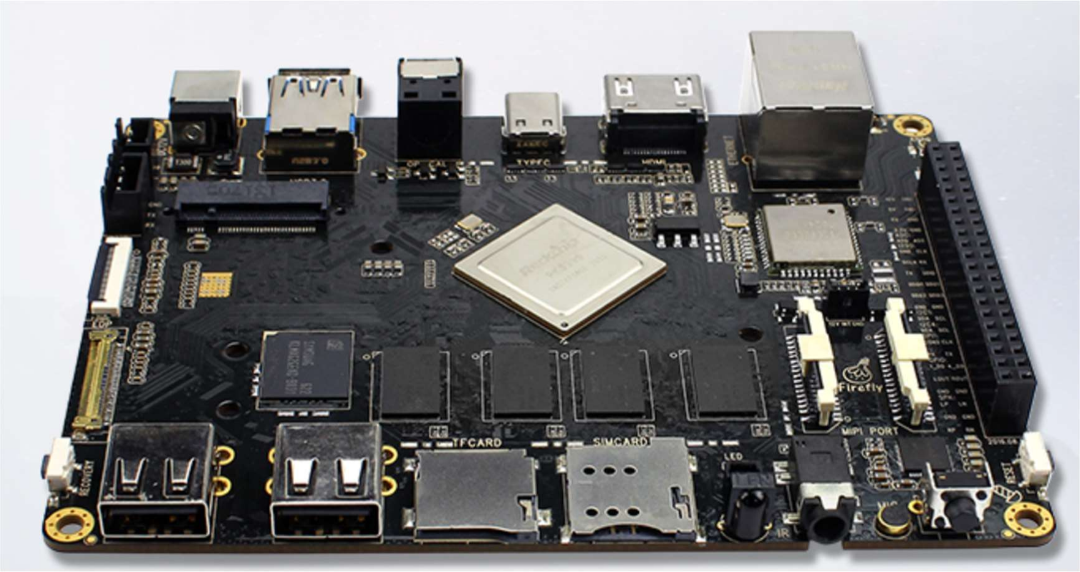
This Report is tested on RK3399 platform and the Arm Compute Library is version 17.12. The report includes both CPU data and GPU data. We collected the data on AlexNet, GoogLeNet, SqueezeNet and MobileNet. Note that the CPU data is on a single A72 core. There is no performance improvement for mixed mode on MXNetOnACL while on the CaffeOnACL the mixed mode can improve performance 1.8X for the best case. The reason is to be determined, but a potential reason is that Caffe matrix data is stored as row by row and MXNet's is column by column.

2 Test Environment

Hardware SoC: Rockchip RK3399

- GPU: Mali T864 (800MHz)
- 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



3 Original MXNet has better Performance

ACL layers CONV, CONV, FC, LR, Pooling, RELU, SOFTMAX are worse than OpenBLAS on CPU, only FC on GPU has better performance. This is different with CaffeOnACL. The reason is to be determined, but potential reason is that Caffe matrix data is stored as row by row and MXNet’s is column by column.

For the total time spent per inference, achieved about 1.12X performance in the best case.

	Original MXNet (ms)	Mixed Mode (ms)	Performance Gain
AlexNet	0.5763	0.5214	1.11X
GoogleNet	0.5700	0.5093	1.12X
SquezzeNet	0.1159	0.1360	0.85X
MobileNet	0.2425	0.2948	0.82X

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:

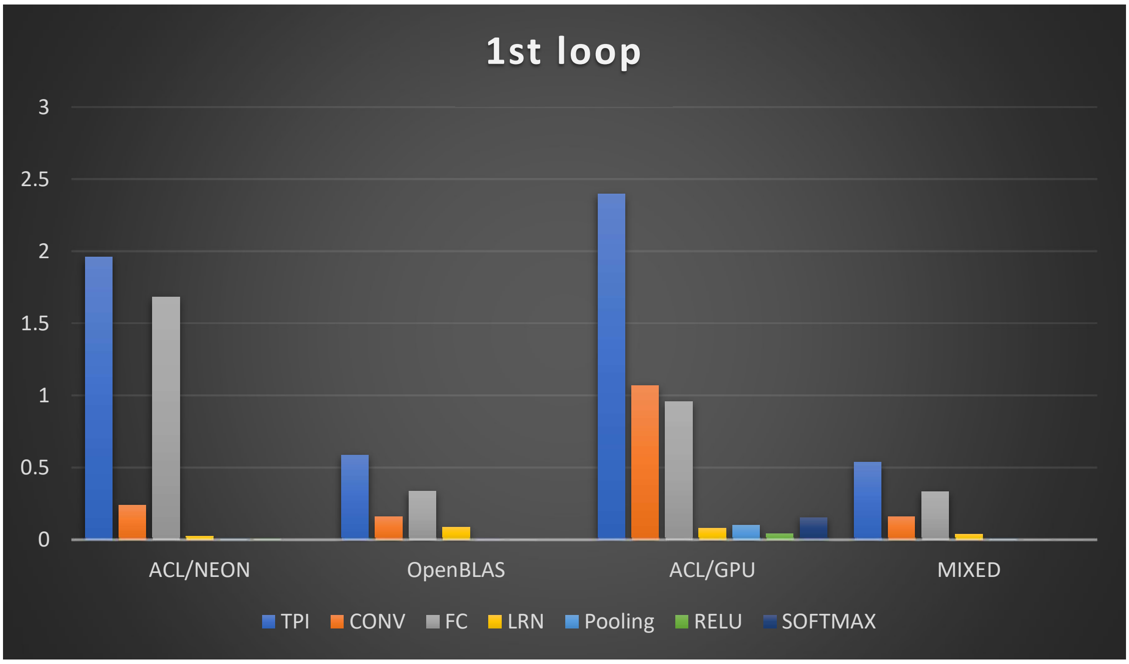
- ✧ TPI: The total time for per inference
- ✧ Avg. Time: tested total 10 times from 2nd to 11th and calculated the average time.
- ✧ The unit of all the data columns in tests below is second.

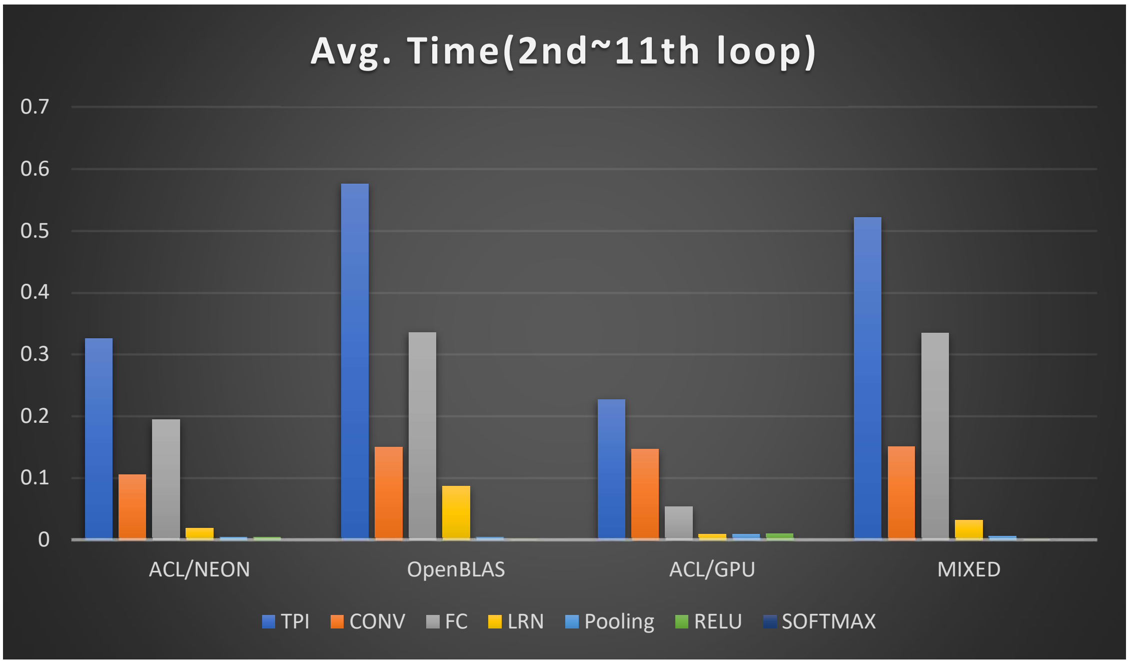
The details see user manual section “Use Cases”.

4.1 AlexNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	1.9609	0.1726	1.4442	0.2151	0.1272
OpenBLAS	0.5857	0	0	0	0
ACL/GPU	2.3976	0.1675	0.0635	1.3837	0.7808
MIXED	0.5369	0.0033	0.0316	0.0015	0.0054
Avg. Time					
ACL/NEON	0.3249	0	0.3159	0	0.0084
OpenBLAS	0.5763	0	0	0	0
ACL/GPU	0.2267	0	0.0119	0	0.2138
MIXED	0.5214	0	0.0309	0	0.0049

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	1.9609	0.2407	1.6823	0.0237	0.0062	0.0077	0.0002
OpenBLAS	0.5857	0.1573	0.3358	0.0874	0.0034	0.0016	0.0001
ACL/GPU	2.3976	1.0694	0.9582	0.0772	0.0995	0.0414	0.1518
MIXED	0.5369	0.1605	0.3328	0.0357	0.0067	0.0011	0.0001
Avg. Time							
ACL/NEON	0.3249	0.1045	0.1942	0.0186	0.0039	0.0036	0.0001
OpenBLAS	0.5763	0.1500	0.3356	0.0862	0.0034	0.0011	0.0001
ACL/GPU	0.2267	0.1466	0.0530	0.0081	0.0088	0.0098	0.0005
MIXED	0.5214	0.1502	0.3341	0.0310	0.0050	0.0011	0.0001



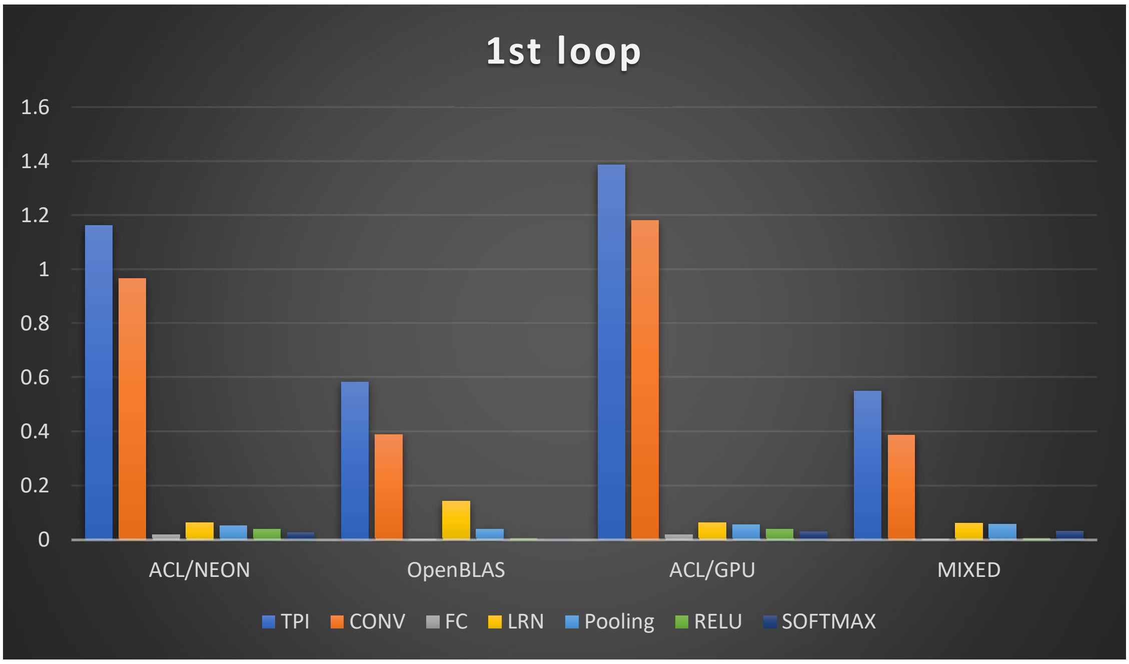


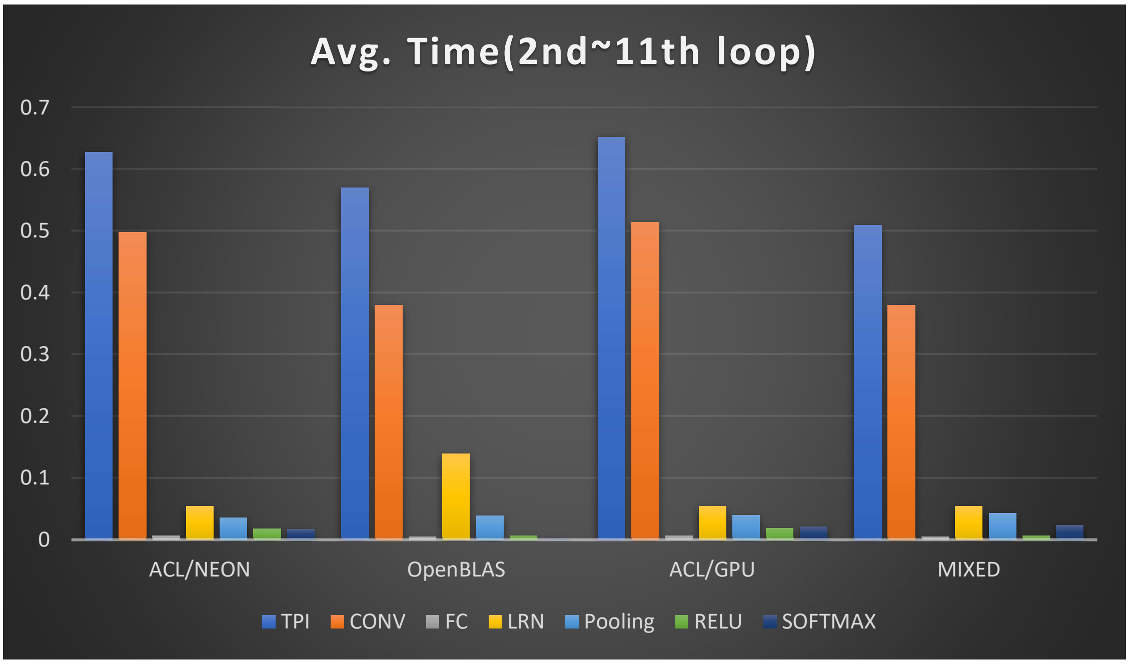
4.2 GoogleNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	1.1634	0.0848	0.6364	0.2223	0.2110
OpenBLAS	0.5832	0	0	0	0
ACL/GPU	1.3875	0.0848	0.6398	0.2221	0.4317
MIXED	0.5497	0.0247	0.0856	0.0035	0.0357
Avg. Time					
ACL/NEON	0.6273	0	0.5690	0	0.0548
OpenBLAS	0.5700	0	0	0	0
ACL/GPU	0.6524	0	0.5720	0	0.0767
MIXED	0.5093	0	0.0852	0	0.0327

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	1.1634	0.9662	0.0193	0.0629	0.0510	0.0389	0.0250
OpenBLAS	0.5832	0.3884	0.0040	0.1422	0.0392	0.0065	0.0027
ACL/GPU	1.3875	1.1812	0.0192	0.0627	0.0554	0.0392	0.0294
MIXED	0.5497	0.3867	0.0047	0.0620	0.0581	0.0065	0.0316
Avg. Time							
ACL/NEON	0.6273	0.4984	0.0061	0.0536	0.0352	0.0176	0.0165

OpenBLAS	0.5700	0.3795	0.0045	0.1392	0.0381	0.0064	0.0023
ACL/GPU	0.6524	0.5144	0.0061	0.0537	0.0397	0.0178	0.0206
MIXED	0.5093	0.3793	0.0045	0.0535	0.0424	0.0065	0.0231

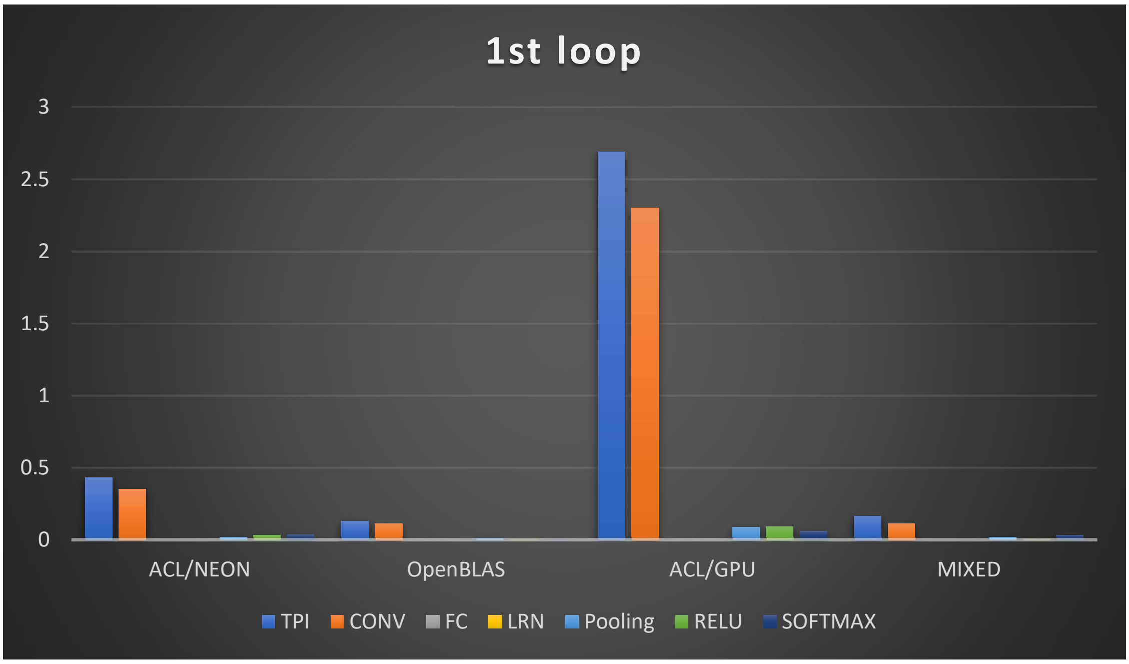


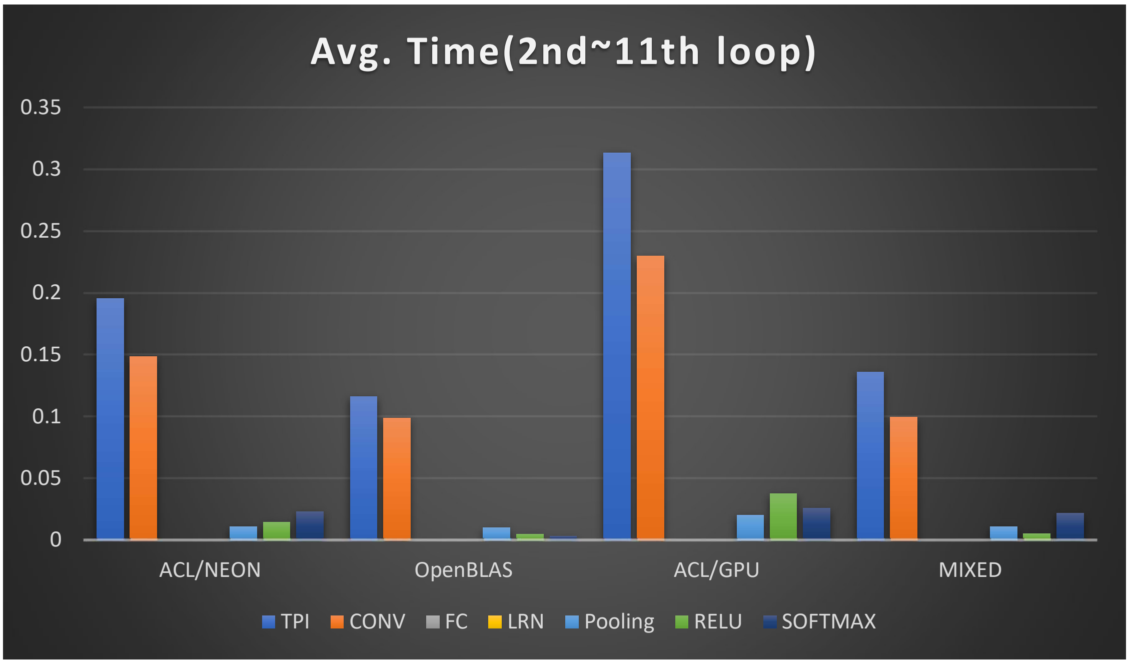


4.3 SqueezeNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	0.4317	0.0432	0.1719	0.0875	0.1246
OpenBLAS	0.1277	0	0	0	0
ACL/GPU	2.6887	0.0332	0.0456	2.2527	0.3517
MIXED	0.1621	0.0129	0.0190	0.0005	0.0144
Avg. Time					
ACL/NEON	0.1955	0	0.1583	0	0.0352
OpenBLAS	0.1159	0	0	0	0
ACL/GPU	0.3135	0	0.0282	0	0.2824
MIXED	0.1360	0	0.0186	0	0.0128

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	0.4317	0.3520	0	0	0.0162	0.0303	0.0330
OpenBLAS	0.1277	0.1082	0	0	0.0104	0.0049	0.0041
ACL/GPU	2.6887	2.3031	0	0	0.0858	0.0886	0.0586
MIXED	0.1621	0.1093	0	0	0.0165	0.0045	0.0317
Avg. Time							
ACL/NEON	0.1955	0.1483	0	0	0.0104	0.0141	0.0227
OpenBLAS	0.1159	0.0987	0	0	0.0097	0.0046	0.0029
ACL/GPU	0.3135	0.2296	0	0	0.0200	0.0375	0.0256
MIXED	0.1360	0.0991	0	0	0.0105	0.0048	0.0216



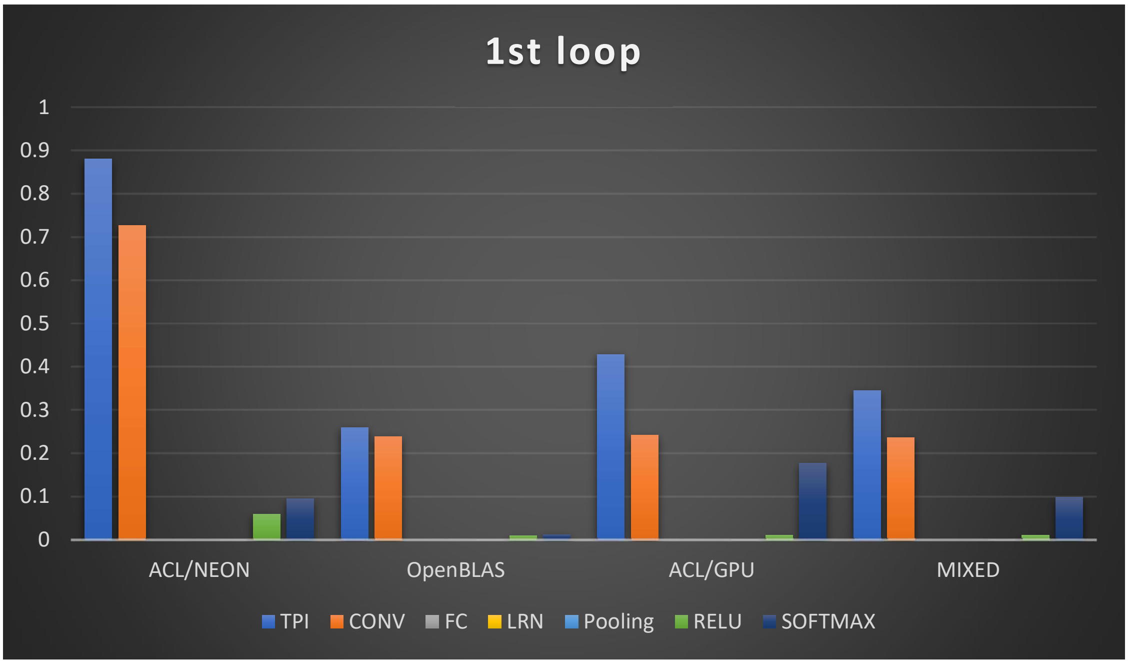


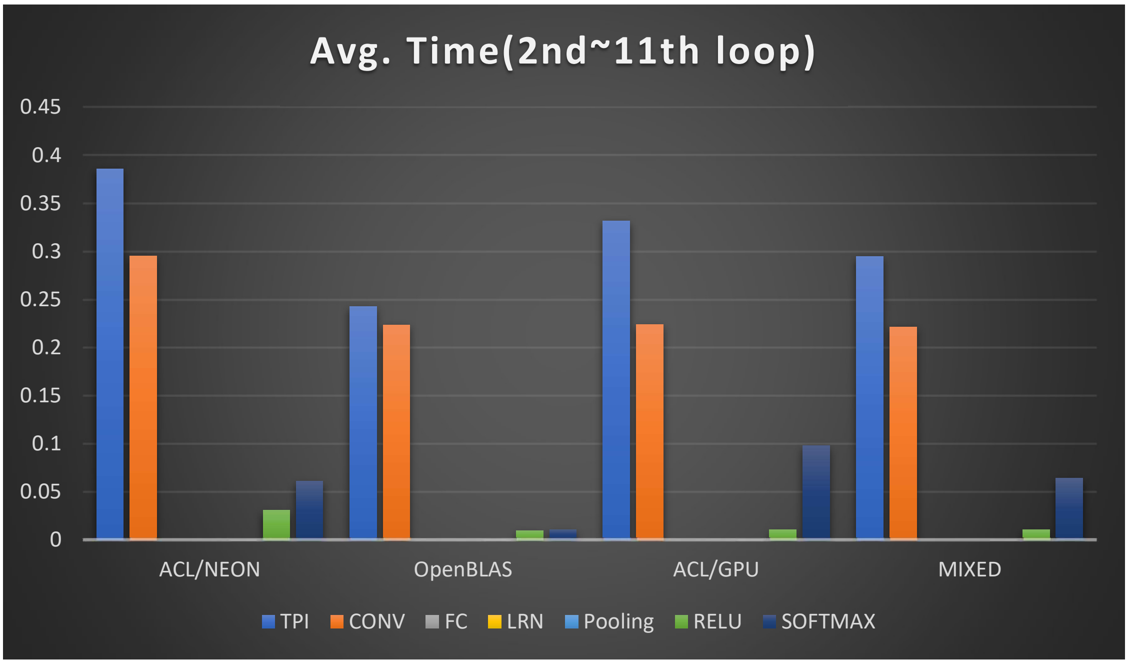
4.4 MobileNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	0.8803	0.0786	0.2476	0.0725	0.3852

OpenBLAS	0.2598	0	0	0	0
ACL/GPU	0.4285	0.0097	0.0119	0.0684	0.0829
MIXED	0.3443	0.0283	0.0268	0.0006	0.0392
Avg. Time					
ACL/NEON	0.3858	0.0003	0.2236	0.0005	0.0775
OpenBLAS	0.2425	0	0	0	0
ACL/GPU	0.3316	0	0.0173	0	0.0793
MIXED	0.2948	0	0.0261	0	0.0365

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	0.8803	0.7260	0	0	0.0001	0.0597	0.0945
OpenBLAS	0.2598	0.2388	0	0	0.0001	0.0090	0.0119
ACL/GPU	0.4285	0.2418	0	0	0.0001	0.0102	0.1763
MIXED	0.3443	0.2364	0	0	0.0001	0.0102	0.0976
Avg. Time							
ACL/NEON	0.3858	0.2949	0	0	0.0001	0.0302	0.0606
OpenBLAS	0.2425	0.2235	0	0	0.0001	0.0088	0.0101
ACL/GPU	0.3316	0.2238	0	0	0.0001	0.0098	0.0978
MIXED	0.2948	0.2212	0	0	0.0001	0.0099	0.0636





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

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	0.3296	0.5723	0.9249
1xA72	0.3249	0.5763	0.5214
2xA72	0.3244	0.5119	0.4495
4xA53	0.3237	1.8043	0.6156
2xA72+4xA53	0.3226	2.4070	0.4735

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

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	1.2134	1.3857	1.3033
1xA72	0.6273	0.5700	0.5093
2xA72	0.4364	0.4245	0.3594
4xA53	0.6019	0.7345	0.6632

2xA72+4xA53	0.3600	0.4597	0.5770

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

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	0.3852	0.3078	0.3625
1xA72	0.1955	0.1159	0.1360
2xA72	0.1567	0.0793	0.1017
4xA53	0.2446	0.1542	0.2112
2xA72+4xA53	0.1347	0.0887	0.1107

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

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	0.8061	0.6418	0.7363
1xA72	0.3858	0.2425	0.2948
2xA72	0.3161	0.1886	0.2421
4xA53	0.5732	0.4085	0.5013
2xA72+4xA53	0.2943	0.1923	0.2507

5.2 The TPI In Mixed mode

The TPI data for different CPU cores in mixed mode:

	AlexNet(s)	GoogleNet(s)	MobileNet(s)	SqueezeNet(s)
1xA53	0.9249	1.3033	0.7363	0.3625
1xA72	0.5214	0.5093	0.2948	0.1360
2xA72	0.4495	0.3594	0.2421	0.1017
4xA53	0.6156	0.6632	0.5013	0.2112
2xA72+4xA53	0.4735	0.5770	0.2507	0.1107

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.

	AlexNet(s)	GoogleNet(s)	SquezzeNet(s)	MobileNet(s)
FC/ACL/NEON	0.1942	0.0061	0	0
FC/OpenBLAS	0.3356	0.0045	0	0
FC/ACL/GPU	0.0530	0.0061	0	0