CaffeOnACL

Performance Report

2017-10-11

OPEN AI LAB

Revision Record

Date	Rev	Change Description	Author
2017-9-22	0.3.0	Initial version	
2017-10-11	0.4.0	Test on ACL v17.09	

catalog

1 Purpose

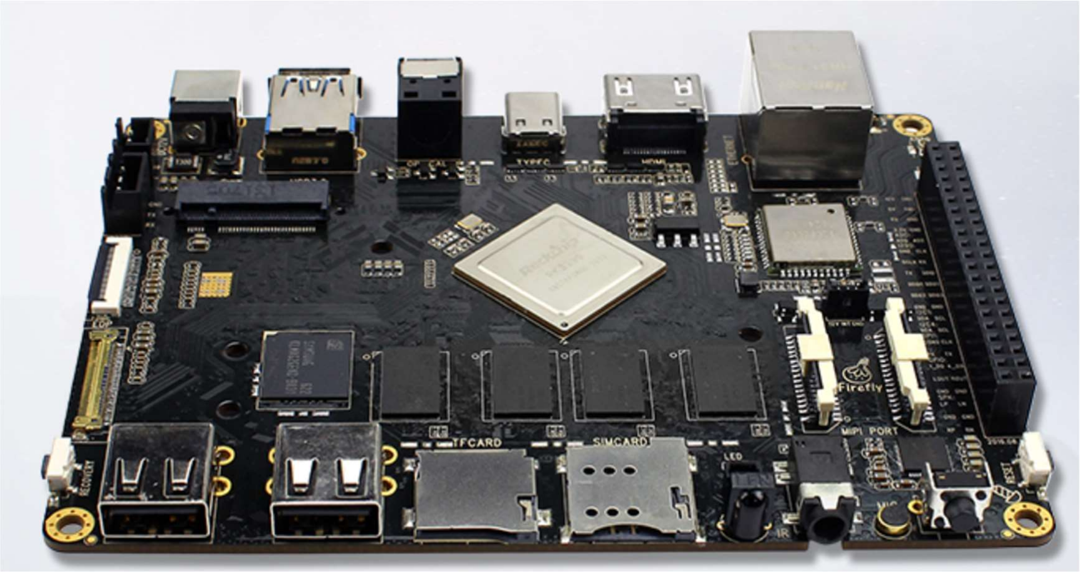
This Report is tested on RK3399 platform and the Arm Compute Library is version 17.09. The report includes both CPU data and GPU data. We collected the data on AlexNet, GoogLeNet, SqueezeNet and MobileNet. And we found the mixed mode can improve performance 1.90X for the best case.

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 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 comment, 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. We have achieved about 1.90X performance in best case.

	Original Caffe(ms)	Mixed Mode(ms)	Performance Gain
AlexNet	552	542	1.02X
GoogleNet	1403	737	1.90X
SquezzeNet	147	160	0.92X
MobileNet	304	287	1.06X

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”.

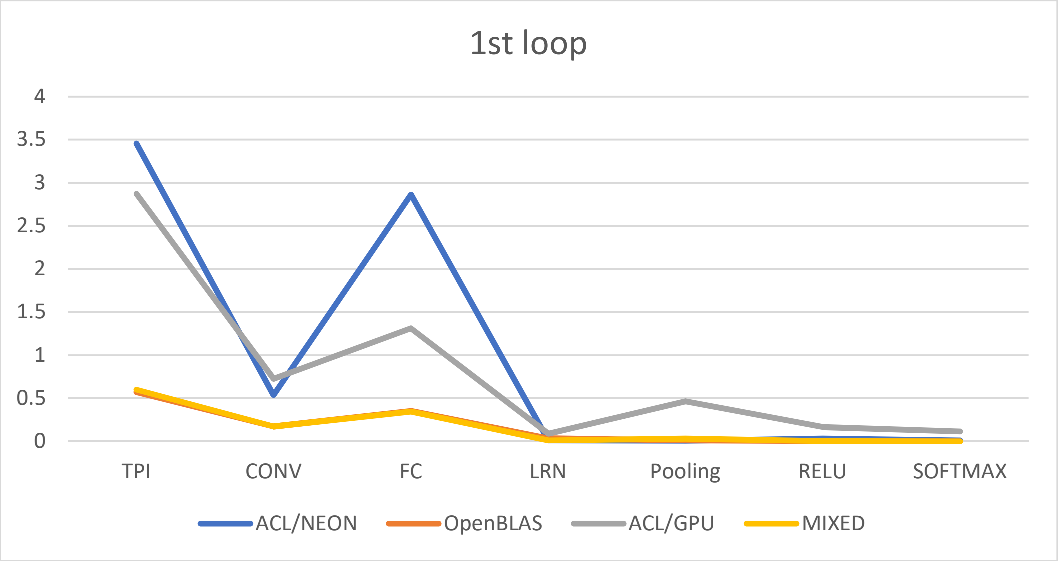
Note that the CPU data of this section is on a single A72 core.

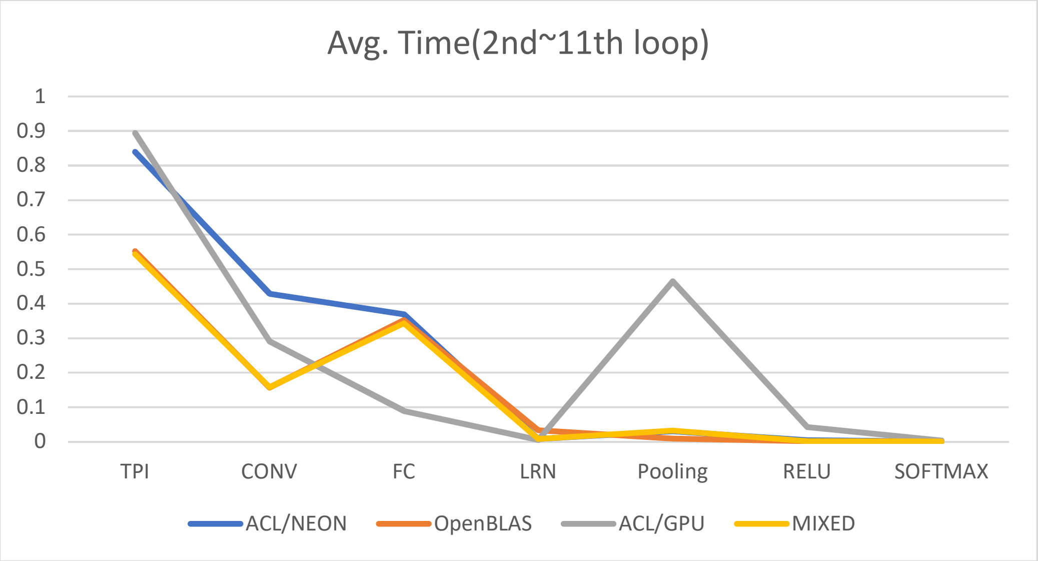
4.1 AlexNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	3.4552	0.1631	2.6166	0.1929	0.1067
OpenBLAS	0.5698	0	0	0	0
ACL/GPU	2.8717	0.1626	0.6913	1.4608	0.3092
MIXED	0.5997	0.0003	0.0339	0.0004	0.0028
Avg. Time					
ACL/NEON	0.8396	0	0.4889	0	0.0059
OpenBLAS	0.5522	0	0	0	0
ACL/GPU	0.8943	0	0.489	0	0.1839
MIXED	0.5428	0	0.0338	0	0.0027

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	3.4552	0.5390	2.8621	0.0109	0.0069	0.0326	0.0091
OpenBLAS	0.5698	0.1718	0.3523	0.0339	0.0102	0.0014	0.0002
ACL/GPU	2.8717	0.7244	1.3121	0.09	0.4659	0.1655	0.1138
MIXED	0.5997	0.1737	0.3451	0.0088	0.033	0.0015	0.0002
Avg. Time							
ACL/NEON	0.8396	0.4283	0.3692	0.0076	0.0307	0.0038	0.0006
OpenBLAS	0.5522	0.1567	0.3516	0.0333	0.0091	0.0015	0.0001
ACL/GPU	0.8943	0.2902	0.089	0.005	0.4646	0.0422	0.0034

MIXED	0.5428	0.1580	0.3431	0.0077	0.0324	0.0015	0.0001



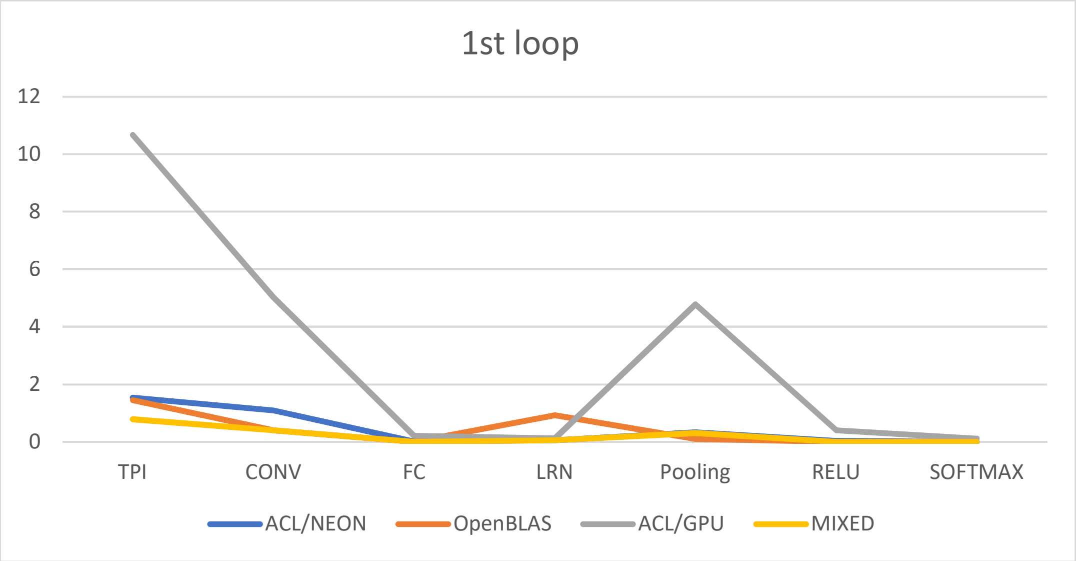


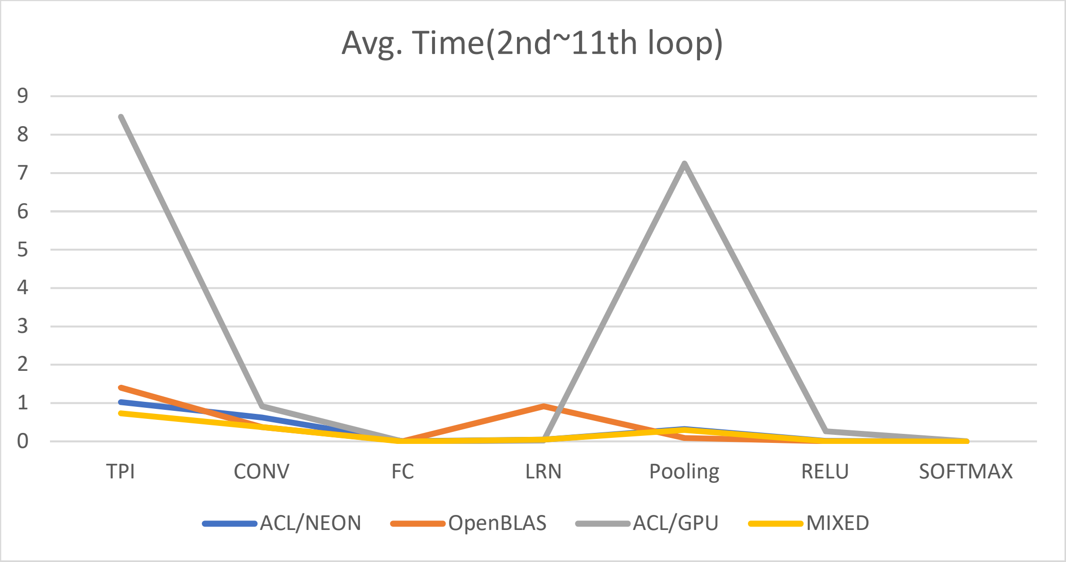
4.2 GoogleNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	1.5351	0.0641	0.9956	0.2177	0.2001
OpenBLAS	1.4486	0	0	0	0
ACL/GPU	10.6703	0.081	3.9034	4.427	1.904
MIXED	0.7867	0.0044	0.2941	0.0025	0.0280
Avg. Time					
ACL/NEON	1.0228	0	0.9287	0	0.0516

OpenBLAS	1.4031	0	0	0	0
ACL/GPU	8.4652	0	4.9121	0	2.6528
MIXED	0.737	0	0.0602	0	0.0185

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	1.5351	1.0858	0.0189	0.058	0.3296	0.0373	0.0003
OpenBLAS	1.4486	0.4093	0.0048	0.9225	0.0997	0.0070	0.0002
ACL/GPU	10.6703	5.0201	0.2127	0.1228	4.7843	0.41	0.114
MIXED	0.7867	0.4034	0.0046	0.0578	0.3088	0.007	0.0002
Avg. Time							
ACL/NEON	1.0228	0.6259	0.0058	0.0488	0.3226	0.0171	0.0006
OpenBLAS	1.4031	0.3776	0.0047	0.9182	0.0932	0.0068	0.0001
ACL/GPU	8.4652	0.9182	0.0029	0.0224	7.2459	0.2702	0.0026
MIXED	0.737	0.372	0.0044	0.0486	0.3029	0.0068	0.0001

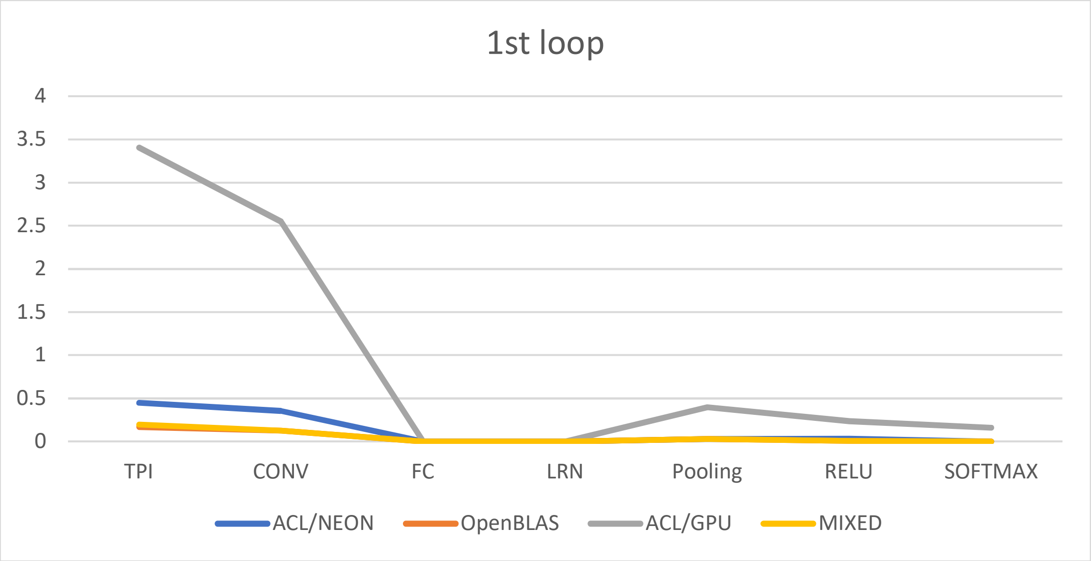


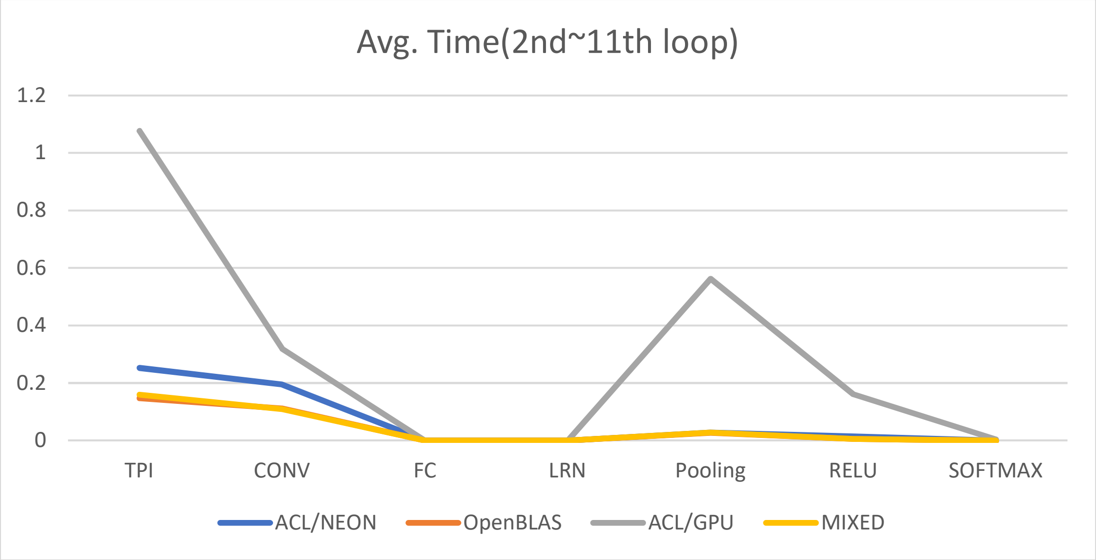


4.3 SqueezeNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	0.4473	0.0398	0.2335	0.089	0.0742
OpenBLAS	0.16833	0	0	0	0
ACL/GPU	3.4057	0.0344	0.5077	2.4941	0.2298
MIXED	0.1964	0.0084	0.0332	0.0003	0.0171
Avg. Time					
ACL/NEON	0.2526	0	0.2208	0	0.0262
OpenBLAS	0.1476	0	0	0	0
ACL/GPU	1.0769	0	0.656	0	0.3173
MIXED	0.1601	0	0.0337	0	0.0075

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	0.4473	0.354	0	0	0.0289	0.0314	0.0002
OpenBLAS	0.1683	0.1272	0	0	0.0286	0.0058	0.00002
ACL/GPU	3.4057	2.5469	0	0	0.3963	0.2354	0.1593
MIXED	0.1964	0.1259	0	0	0.0298	0.0062	0.0002
Avg. Time							
ACL/NEON	0.2526	0.1947	0	0	0.0276	0.015	0.0001
OpenBLAS	0.1476	0.1114	0	0	0.0271	0.0059	0.0001
ACL/GPU	1.0769	0.3191	0	0	0.5624	0.1617	0.0038
MIXED	0.1601	0.109	0	0	0.0285	0.0058	0.0001

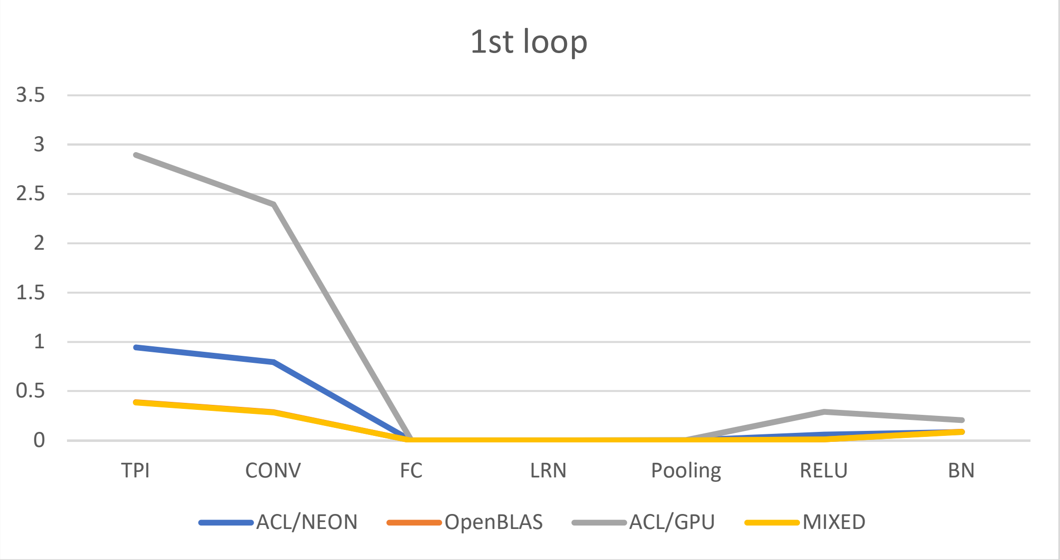


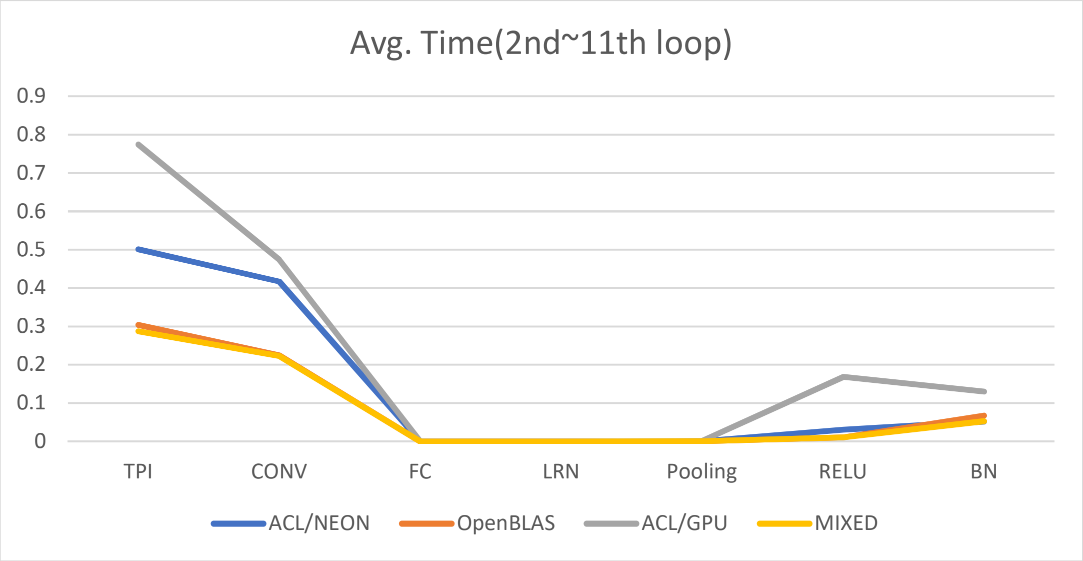


4.4 MobileNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	0.9437	0.0849	0.4044	0.0803	0.2218
OpenBLAS	0.389	0	0	0	0
ACL/GPU	2.8949	0.0861	0.4554	1.7	0.3279
MIXED	0.3836	0.0286	0.0251	0.0004	0.0289
Avg. Time					
ACL/NEON	0.5011	0	0.3556	0	0.056
OpenBLAS	0.3039	0	0	0	0
ACL/GPU	0.7744	0	0.4048	0	0.1361
MIXED	0.2874	0	0.0242	0	0.028

	TPI	CONV	FC	LRN	Pooling	RELU	BN
1st							
ACL/NEON	0.9437	0.7938	0	0	0.0007	0.0615	0.0877
OpenBLAS	0.389	0.2876	0	0	0.0007	0.011	0.0897
ACL/GPU	2.8949	2.3936	0	0	0.0032	0.2906	0.2075
MIXED	0.3836	0.2861	0	0	0.0008	0.0111	0.0856
Avg. Time							
ACL/NEON	0.5011	0.4174	0	0	0.0007	0.0306	0.0524
OpenBLAS	0.3039	0.2243	0	0	0.0007	0.0107	0.0679
ACL/GPU	0.7744	0.4742	0	0	0.0016	0.1684	0.1302
MIXED	0.2874	0.2227	0	0	0.0007	0.0108	0.0532





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

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	2.1262	0.9589	0.9537
1xA72	0.5698	0.5522	0.5428
2xA72	0.4128	0.4667	0.4648
4xA53	0.7843	0.6864	0.6708
2xA72+4xA53*	0.4626	0.6326	0.4908

GoogleNet

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	2.2941	3.4872	1.7562
1xA72	1.0228	1.4031	0.7076
2xA72	0.66	1.2518	0.5798
4xA53	1.7562	2.7282	1.056
2xA72+4xA53*	2.0334	1.3092	0.6077

SqueezeNet.

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	0.5963	0.3780	0.4152
1xA72	0.2526	0.1476	0.1601
2xA72	0.1669	0.1035	0.1142
4xA53	0.4664	0.196	0.2487
2xA72+4xA53*	0.6232	0.107	0.1175

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

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	1.2873	0.9061	1.2542
1xA72	0.5011	0.3039	0.5419
2xA72	0.4434	0.2676	0.4622
4xA53	0.9505	0.6338	0.9418
2xA72+4xA53*	0.6667	0.2657	0.5017

5.2 The TPI In Mixed mode

The TPI data for different CPU cores in mixed mode:

	AlexNet(s)	GoogleNet(s)	SqueezeNet(s)	MobileNet(s)
1xA53	0.9537	1.7562	0.4152	1.2542
1xA72	0.5428	0.7076	0.1601	0.5419
2xA72	0.4648	0.5798	0.1142	0.4622
4xA53	0.6708	1.056	0.2487	0.9418
2xA72+4xA53	0.4908	0.6077	0.1175	0.5017

6 Conclusion

From the above test cases, we can deduce that :

- the performances of LRN are better under ACL_NEON than under OpenBLAS
- the performances of large FC are better under ACL_CL(GPU) than under NEON and OpenBLAS

	AlexNet(s)	GoogleNet(s)	SquzzeNet(s)	MobileNet(s)
LRN/ACL	0.0076	0.0488	0	0
LRN/OpenBLAS	0.0333	0.9182	0	0
FC/ACL/GPU	0.089	0.0029	0	0
FC/ACL/NEON	0.3692	0.0058	0	0
FC/OpenBLAS	0.3516	0.0047	0	0

However, for different cases, you may see different result for different layers by using ACL or OpenBLAS. Therefore, for applications, you can select best solution by combining ACL and OpenBLAS together.