CaffeOnACL

Performance Report

2017-10-20

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	
2017-10-20	0.5.0	Test on ACL v17.10	

Catalog

1 Purpose

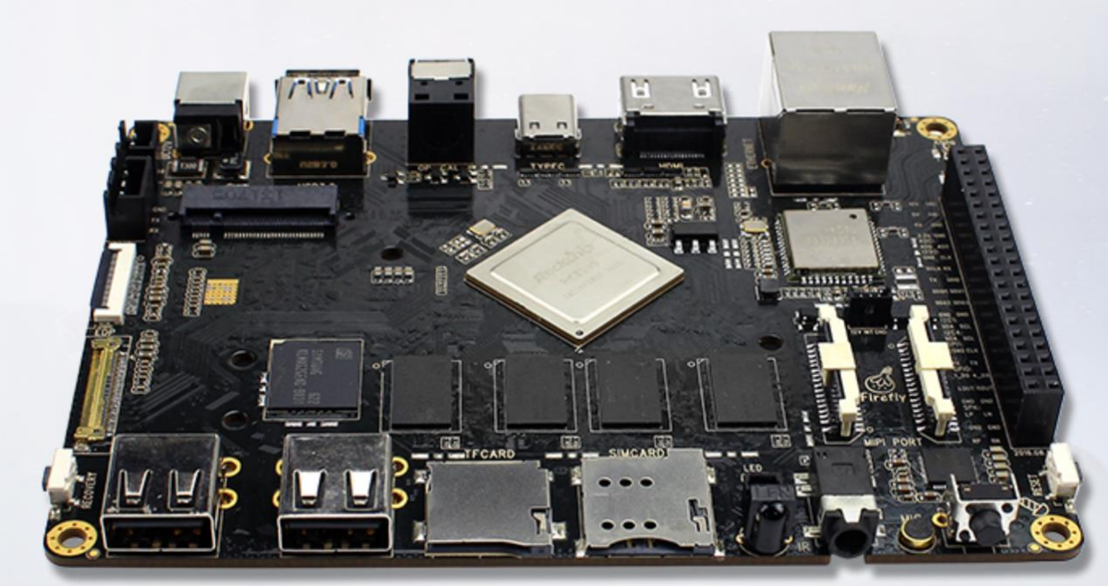
This Report is tested on RK3399 platform and the Arm Compute Library is version 17.10. The report includes both CPU data and GPU data. We collected the data on AlexNet, GoogLeNet, SqueezeNet and MobileNet. And we found the mixed mode can improve performance 2.92X for the best case.

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 2.92X performance in best case.

	Original Caffe(s)	Mixed Mode(s)	Performance Gain
AlexNet	0.8572	0.4949	1.73X
GoogleNet	1.2566	0.4303	2.92X
SquezzeNet	0.1329	0.1209	1.10X
MobileNet	0.2815	0.2649	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”.

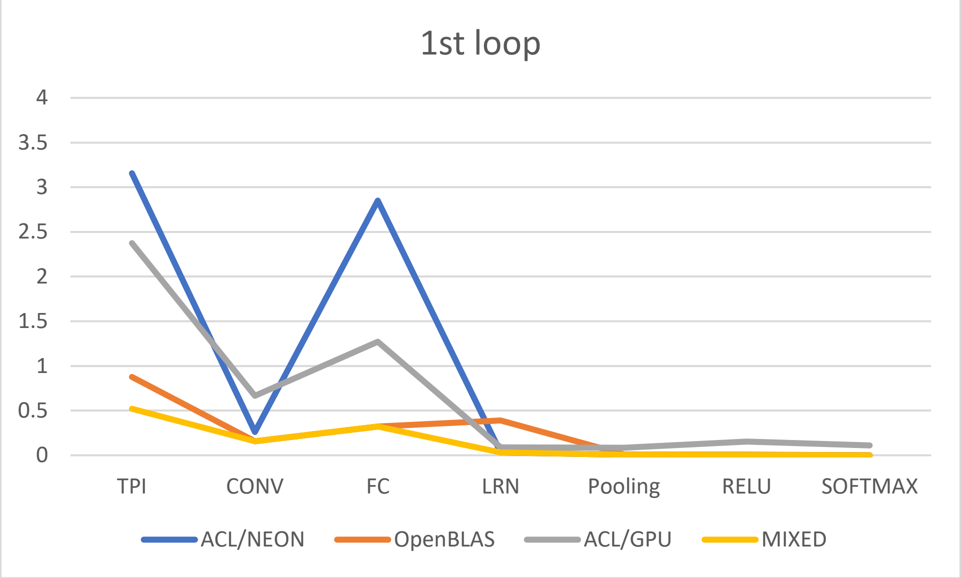
Note that the CPU data of this section is on a single A72 core.

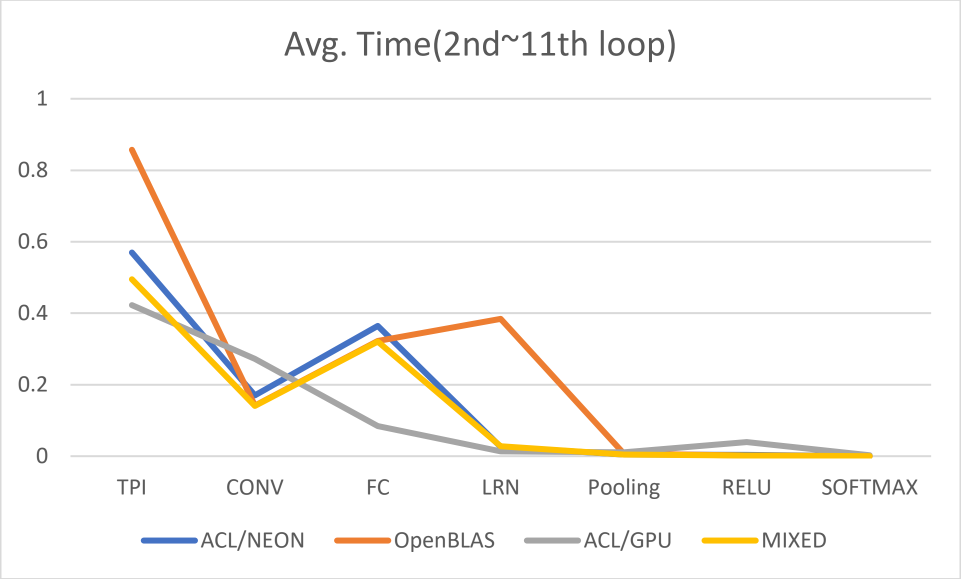
4.1 AlexNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	3.1564	0.1608	2.6011	0.1788	0.1154
OpenBLAS	0.8768	0.0000	0.0000	0.0000	0.0000
ACL/GPU	2.3744	0.1616	0.4528	1.3512	0.1840
MIXED	0.5206	0.0037	0.0282	0.0012	0.0053
Avg. Time					
ACL/NEON	0.5698	0.0000	0.4734	0.0000	0.0070
OpenBLAS	0.8572	0.0000	0.0000	0.0000	0.0000
ACL/GPU	0.4226	0.0000	0.1891	0.0000	0.0273
MIXED	0.4949	0.0000	0.0273	0.0000	0.0043

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	3.1564	0.2608	2.8492	0.0321	0.0065	0.0076	0.0002
OpenBLAS	0.8768	0.1557	0.3227	0.3881	0.0086	0.0015	0.0001
ACL/GPU	2.3744	0.6665	1.2700	0.0893	0.0847	0.1544	0.1094
MIXED	0.5206	0.1567	0.3226	0.0329	0.0067	0.0015	0.0001
Avg. Time							
ACL/NEON	0.5698	0.1705	0.3649	0.0264	0.0042	0.0037	0.0001
OpenBLAS	0.8572	0.1412	0.3221	0.3846	0.0076	0.0015	0.0001
ACL/GPU	0.4226	0.2719	0.0842	0.0134	0.0112	0.0392	0.0026

MIXED	0.4949	0.1411	0.3203	0.0275	0.0044	0.0015	0.0001



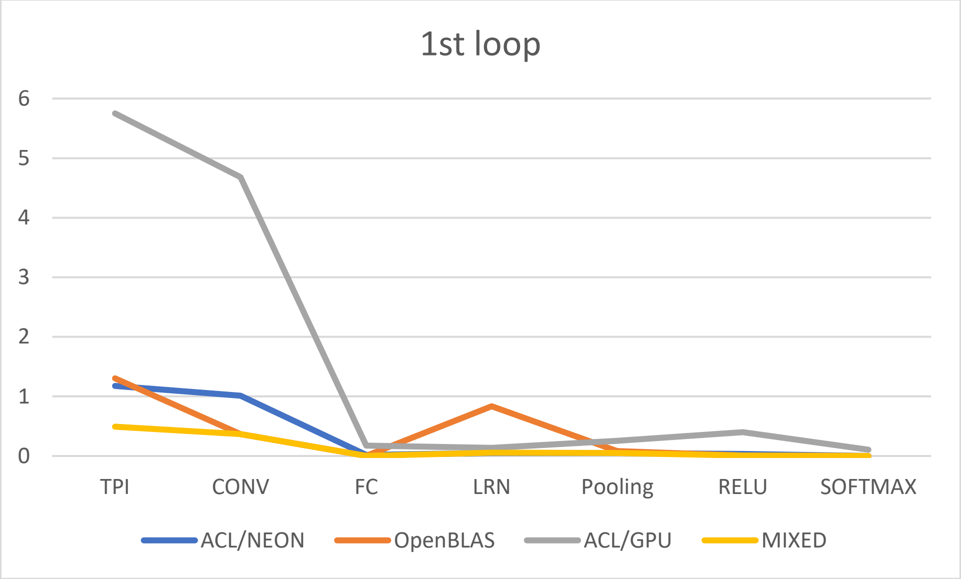


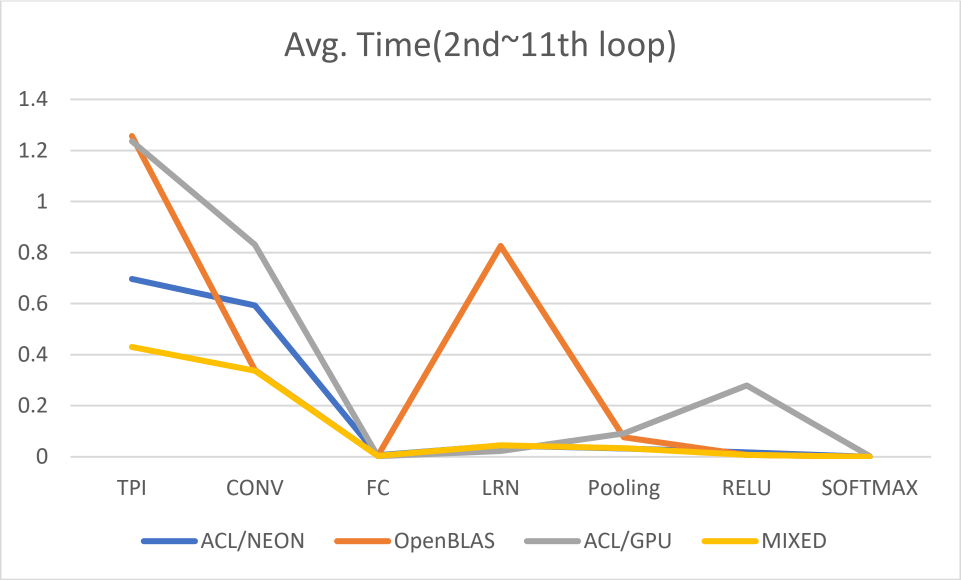
4.2 GoogleNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	1.1761	0.0734	0.7087	0.2112	0.1606
OpenBLAS	1.3035	0	0	0	0
ACL/GPU	5.7508	0.1004	1.0252	4.0865	0.3053
MIXED	0.4920	0.0166	0.0639	0.0024	0.0209
Avg. Time					
ACL/NEON	0.6966	0	0.6458	0	0.0432

OpenBLAS	1.2566	0	0	0	0
ACL/GPU	1.2360	0	0.8343	0	0.2439
MIXED	0.4303	0	0.0622	0	0.0149

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	1.1761	1.0117	0.0187	0.0530	0.0509	0.0365	0.0002
OpenBLAS	1.3035	0.3699	0.0047	0.8325	0.0841	0.0072	0.0002
ACL/GPU	5.7508	4.6793	0.1722	0.1346	0.2537	0.3999	0.1056
MIXED	0.4920	0.3681	0.0041	0.0541	0.0532	0.0072	0.0002
Avg. Time							
ACL/NEON	0.6966	0.5931	0.0061	0.0445	0.0332	0.0171	0.0001
OpenBLAS	1.2566	0.3402	0.0042	0.8263	0.0764	0.0070	0.0001
ACL/GPU	1.2360	0.8311	0.0031	0.0228	0.0919	0.2802	0.0029
MIXED	0.4303	0.3378	0.0042	0.0446	0.0336	0.0070	0.0001

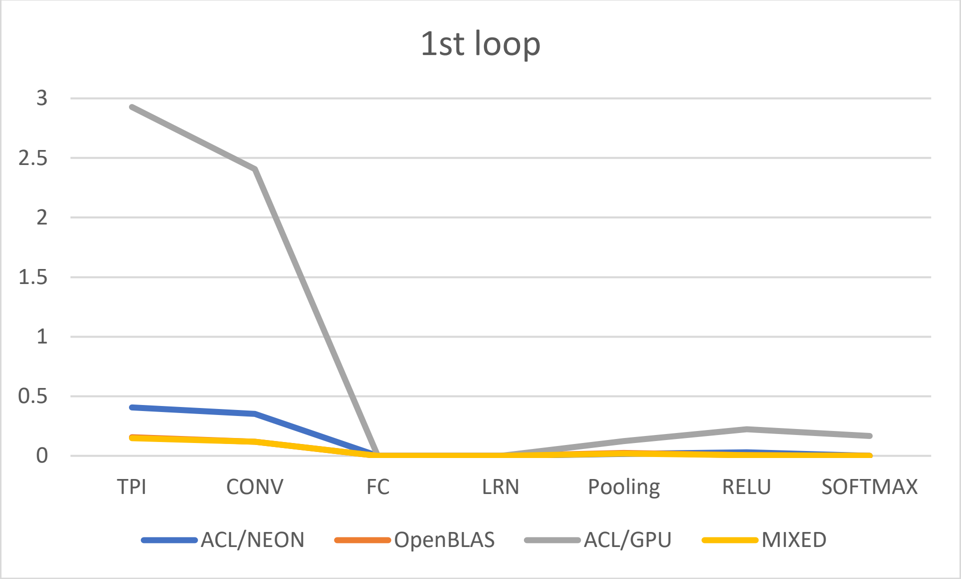


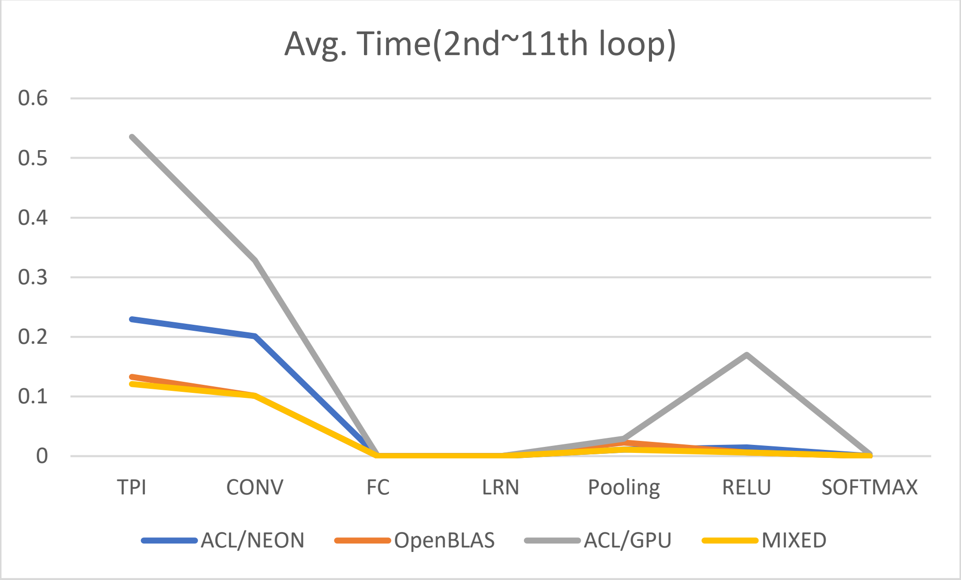


4.3 SqueezeNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	0.4061	0.0351	0.2115	0.0857	0.0605
OpenBLAS	0.1564	0	0	0	0
ACL/GPU	2.9276	0.0282	0.3271	2.3379	0.1043
MIXED	0.1480	0.0046	0.0066	0.0001	0.0054
Avg. Time					
ACL/NEON	0.2295	0	0.2016	0	0.0226
OpenBLAS	0.1329	0	0	0	0
ACL/GPU	0.5355	0	0.3281	0	0.1101
MIXED	0.1209	0	0.0059	0	0.0042

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	0.4061	0.3515	0	0	0.0175	0.0307	0.0002
OpenBLAS	0.1564	0.1187	0	0	0.0253	0.0061	0.0002
ACL/GPU	2.9276	2.4076	0	0	0.1243	0.2220	0.1674
MIXED	0.1480	0.1174	0	0	0.0182	0.0062	0.0002
Avg. Time							
ACL/NEON	0.2295	0.2010	0	0	0.0109	0.0147	0.0001
OpenBLAS	0.1329	0.1013	0	0	0.0226	0.0060	0.0001
ACL/GPU	0.5355	0.3285	0	0	0.0294	0.1697	0.0034
MIXED	0.1209	0.1010	0	0	0.0109	0.0060	0.0001

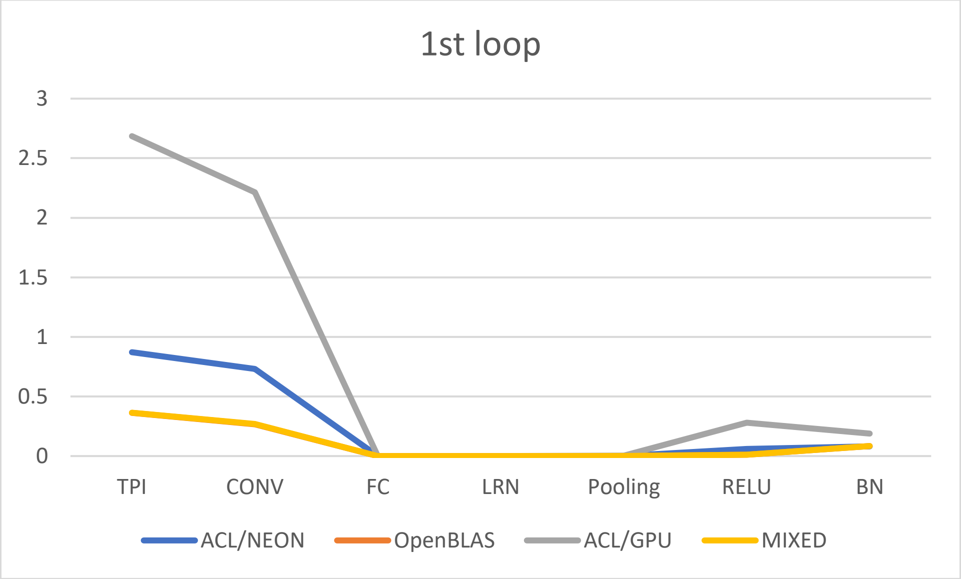


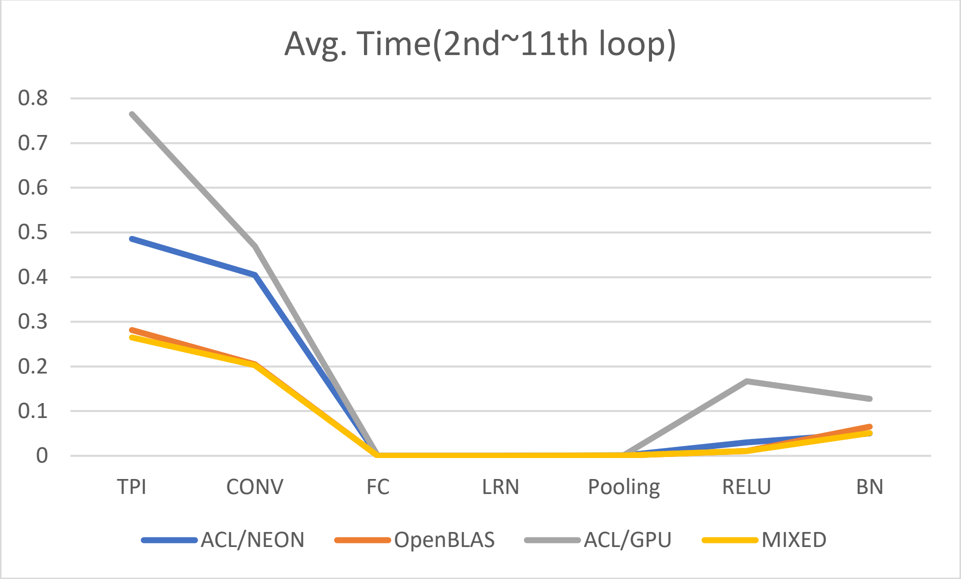


4.4 MobileNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	0.8716	0.0785	0.3901	0.0734	0.1929
OpenBLAS	0.3627	0	0	0	0
ACL/GPU	2.6849	0.0800	0.4513	1.5839	0.2768
MIXED	0.3637	0.0277	0.0248	0.0004	0.0280
Avg. Time					
ACL/NEON	0.4856	0	0.3516	0	0.0538
OpenBLAS	0.2815	0	0	0	0
ACL/GPU	0.7648	0	0.4042	0	0.1338
MIXED	0.2649	0	0.0235	0	0.0266

	TPI	CONV	FC	LRN	Pooling	RELU	BN
1st							
ACL/NEON	0.8716	0.7308	0	0	0.0005	0.0592	0.0811
OpenBLAS	0.3627	0.2664	0	0	0.0005	0.0113	0.0845
ACL/GPU	2.6849	2.2138	0	0	0.0023	0.2792	0.1895
MIXED	0.3637	0.2682	0	0	0.0004	0.0113	0.0837
Avg. Time							
ACL/NEON	0.4856	0.4046	0	0	0.0005	0.0304	0.0501
OpenBLAS	0.2815	0.2046	0	0	0.0004	0.0111	0.0654
ACL/GPU	0.7648	0.4691	0	0	0.0012	0.1667	0.1278
MIXED	0.2649	0.2024	0	0	0.0004	0.0111	0.0509





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	2.0606	1.7571	0.8954
1xA72	0.5691	0.8558	0.4963
2xA72	0.3801	0.8078	0.4266
4xA53	0.7521	1.5168	0.5995
2xA72+4xA53*	0.427	0.8607	0.4664

GoogleNet

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	1.7122	3.1558	1.1838
1xA72	0.6997	1.2602	0.4337
2xA72	0.4257	1.117	0.3061
4xA53	0.7341	2.5363	0.5642
2xA72+4xA53*	0.4907	1.2027	0.3356

SqueezeNet.

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	0.4748	0.3466	0.3244
1xA72	0.2295	0.1338	0.1237
2xA72	0.1478	0.097	0.085
4xA53	0.2655	0.1811	0.158
2xA72+4xA53*	0.1574	0.0995	0.0887

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

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	1.2002	0.8516	0.8011
1xA72	0.5234	0.3192	0.3073
2xA72	0.4456	0.2578	0.2626
4xA53	0.8617	0.5998	0.572
2xA72+4xA53*	0.4806	0.2776	0.2868

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	0.8954	1.1838	0.3244	0.8011
1xA72	0.4963	0.4337	0.1237	0.3073
2xA72	0.4266	0.3061	0.085	0.2626
4xA53	0.5995	0.5642	0.158	0.572
2xA72+4xA53	0.4664	0.3356	0.0887	0.2868

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)	SquezzeNet(s)	MobileNet(s)
LRN/ACL	0.0264	0.0445	0	0
LRN/OpenBLAS	0.0319	0.8263	0	0
FC/ACL/GPU	0.0842	0.0031	0	0
FC/ACL/NEON	0.3649	0.0061	0	0
FC/OpenBLAS	0.3501	0.0042	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.