MXNetOnACL

Performance Report

2017-10-11

OPEN AI LAB

Reversion Record

Date	Rev	Change Description	Author
2017-9-22	0.2.0	Initial version	
2017-10-11	0.3.0	Test on ACL v17.09	

catalog

1 Purpose

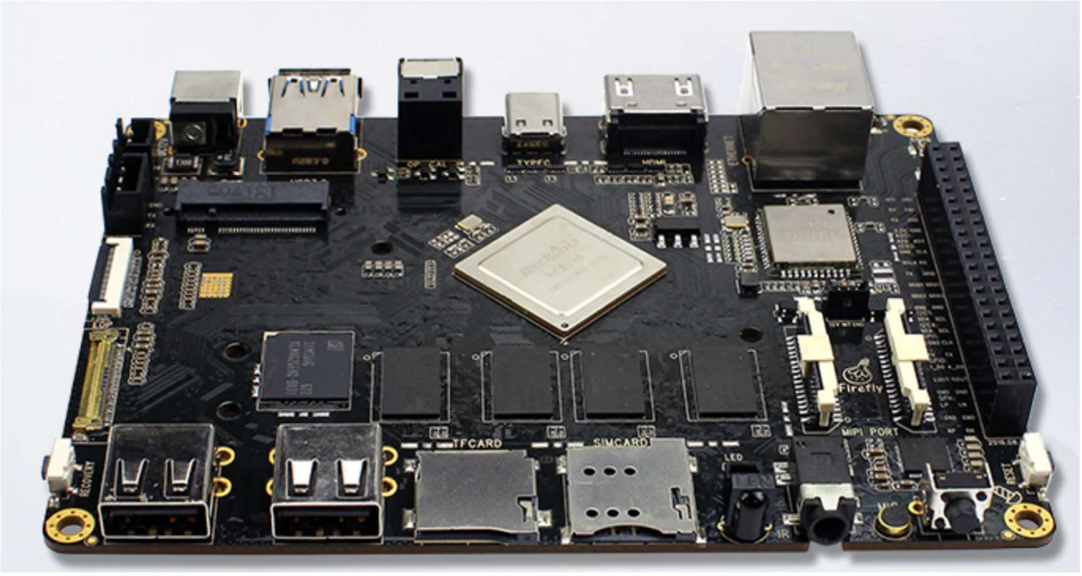
This Report is tested on RK3399 platform and the Arm Compute Library is version 17.09. The report includes both CPU data and GPU data. We collected the data on AlexNet, GoogLeNet, SqueezeNet and MobileNet. Note that the CPU data is on a single A72 core.There is no performance improvement for mixed mode on MXNetOnACL while on the CaffeOnACL the mixed mode can improve performance 1.9X 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.

We almost can't get any performance improvement by mixed mode.

	Original MXNet (ms)	Mixed Mode (ms)	Performance Gain
AlexNet	518	502	1.03X
GoogleNet	562	482	1.17X
SquezzeNet	116	137	0.85X
MobileNet	242	309	0.78X

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.

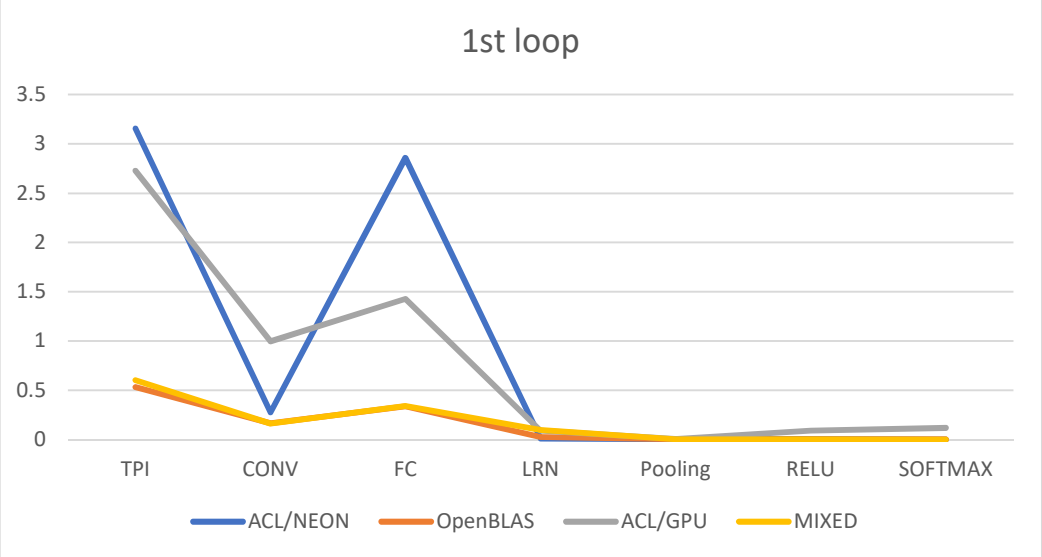
The details see user manual section “Use Cases”.

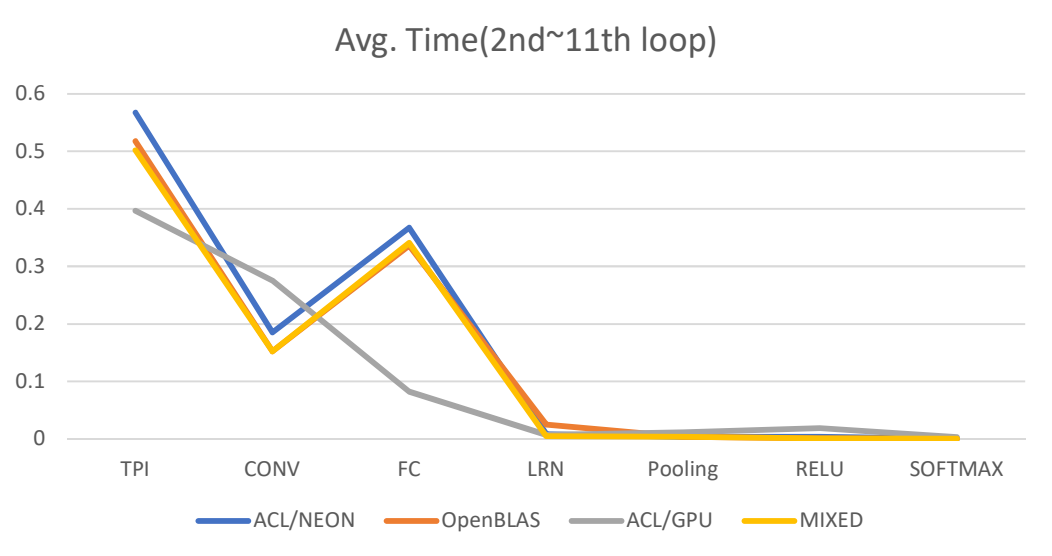
4.1 AlexNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	3.1555	0.2686	2.5913	0.1848	0.1074
OpenBLAS	0.5316	0	0	0	0
ACL/GPU	2.7274	0.4949	0.4104	1.6452	0.1907
MIXED	0.6026	0.0012	0.0034	0.0891	0.0016
Avg. Time					
ACL/NEON	0.5674	0.0017	0.4632	0	0.0037
OpenBLAS	0.5177	0	0	0	0
ACL/GPU	0.3967	0.0089	0.1839	0	0.0178
MIXED	0.5018	0.0010	0.0033	0	0.0015

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	3.1555	0.2759	2.8584	0.0095	0.0037	0.0077	0.0002

OpenBLAS	0.5316	0.1638	0.3364	0.0250	0.0038	0.0011	0.0015
ACL/GPU	2.7274	0.9961	1.4278	0.0904	0.0054	0.0887	0.1190
MIXED	0.6026	0.1624	0.3404	0.0948	0.0038	0.0011	0.0001
Avg. Time							
ACL/NEON	0.5674	0.1850	0.3674	0.0079	0.0034	0.0036	0.0001
OpenBLAS	0.5177	0.1518	0.3366	0.0247	0.0034	0.0011	0.0001
ACL/GPU	0.3967	0.2749	0.0821	0.0069	0.0109	0.0189	0.0030
MIXED	0.5018	0.1519	0.3405	0.0049	0.0034	0.0011	0.0001

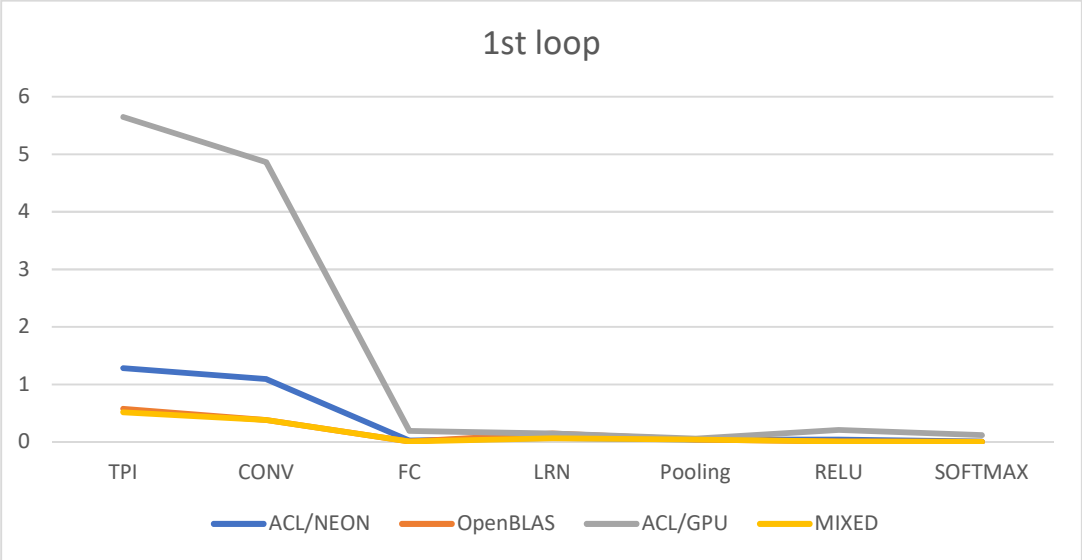


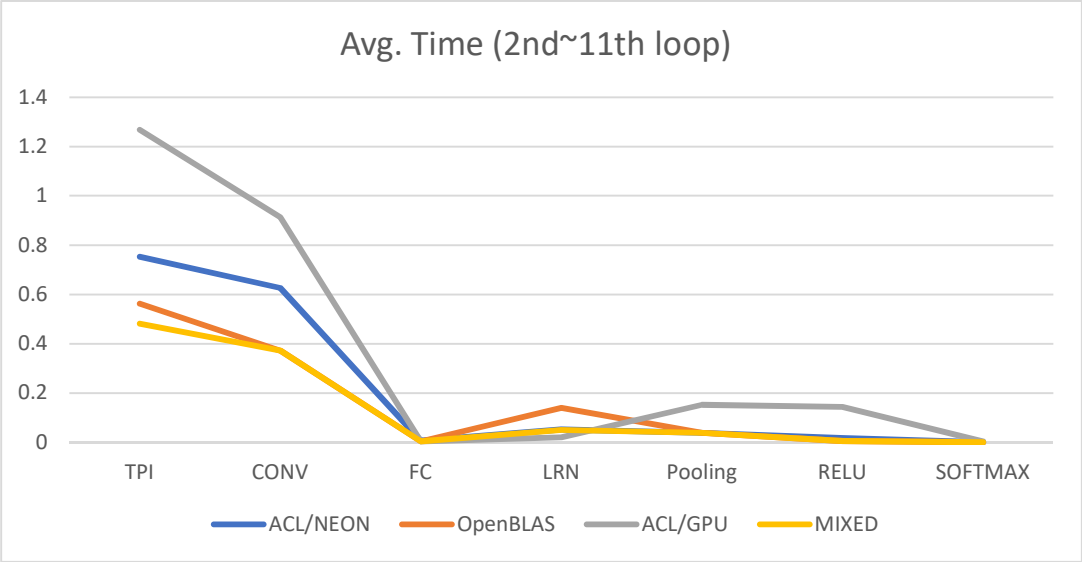


4.2 GoogleNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	1.2801	0.2339	0.7411	0.2271	0.1823
OpenBLAS	0.5749	0	0	0	0
ACL/GPU	5.6480	0.3564	1.0167	4.1540	0.2998
MIXED	0.5125	0.0177	0.0540	0.0028	0.0130
Avg. Time					
ACL/NEON	0.7524	0.0205	0.6728	0	0.0374
OpenBLAS	0.5624	0	0	0	0
ACL/GPU	1.2682	0.1098	0.9022	0	0.2035
MIXED	0.4819	0.0021	0.0539	0	0.0071

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	1.2801	1.0897	0.0208	0.0633	0.0394	0.0406	0.0014
OpenBLAS	0.5749	0.3800	0.0044	0.1425	0.0394	0.0056	0.0001
ACL/GPU	5.6480	4.8611	0.1874	0.1411	0.0559	0.2078	0.1140
MIXED	0.5125	0.3805	0.0046	0.0588	0.0394	0.0054	0.0001
Avg. Time							
ACL/NEON	0.7524	0.6266	0.0063	0.0514	0.0382	0.0180	0.0001
OpenBLAS	0.5624	0.3724	0.0044	0.1394	0.0381	0.0055	0.0001
ACL/GPU	1.2682	0.9127	0.0034	0.0202	0.1522	0.1429	0.0030
MIXED	0.4819	0.3724	0.0045	0.0491	0.0382	0.0054	0.0001



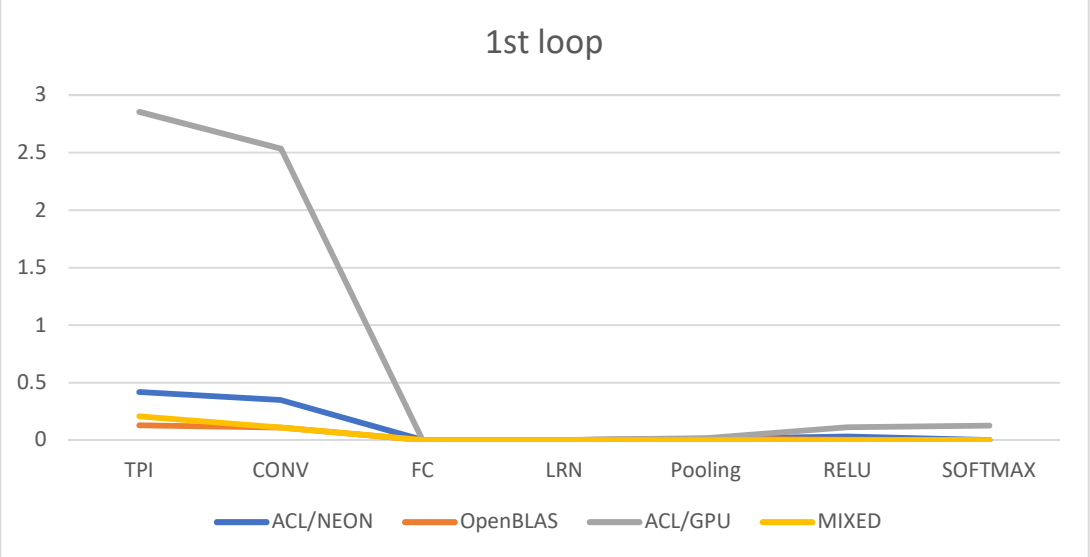


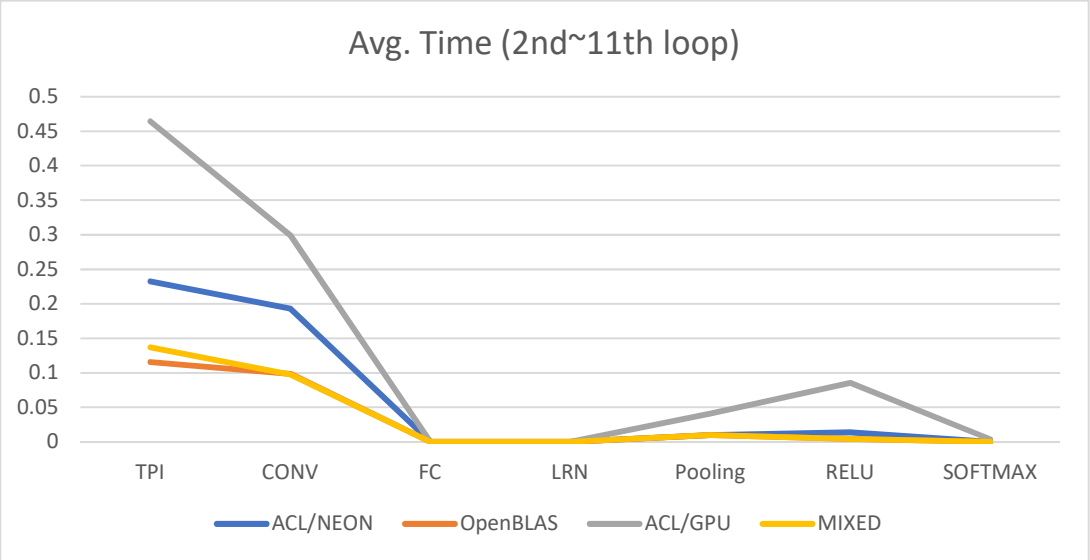
4.3 SqueezeNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	0.4169	0.0878	0.2122	0.0852	0.0670
OpenBLAS	0.1276	0	0	0	0
ACL/GPU	2.8538	0.1170	0.3093	2.3741	0.1189
MIXED	0.2071	0.0125	0.0160	0.0468	0.0149
Avg. Time					
ACL/NEON	0.2324	0.0084	0.1991	0	0.0219

OpenBLAS	0.1155	0	0	0	0
ACL/GPU	0.4646	0.0468	0.3084	0	0.1105
MIXED	0.1369	0	0.0191	0	0.0059

	TPI	CONV	FC	LRN	Pooling	RELU	SOFTMAX
1st							
ACL/NEON	0.4169	0.3472	0	0	0.0103	0.0293	0.0003
OpenBLAS	0.1276	0.1086	0	0	0.0104	0.0043	0.0001
ACL/GPU	2.8538	2.5341	0	0	0.0173	0.1110	0.1248
MIXED	0.2071	0.1097	0	0	0.0010	0.0004	0.0000
Avg. Time							
ACL/NEON	0.2324	0.1927	0	0	0.0096	0.0138	0.0001
OpenBLAS	0.1155	0.0986	0	0	0.0096	0.0043	0.0001
ACL/GPU	0.4646	0.2989	0	0	0.0408	0.0855	0.0035
MIXED	0.1369	0.0976	0	0	0.0096	0.0042	0.0001



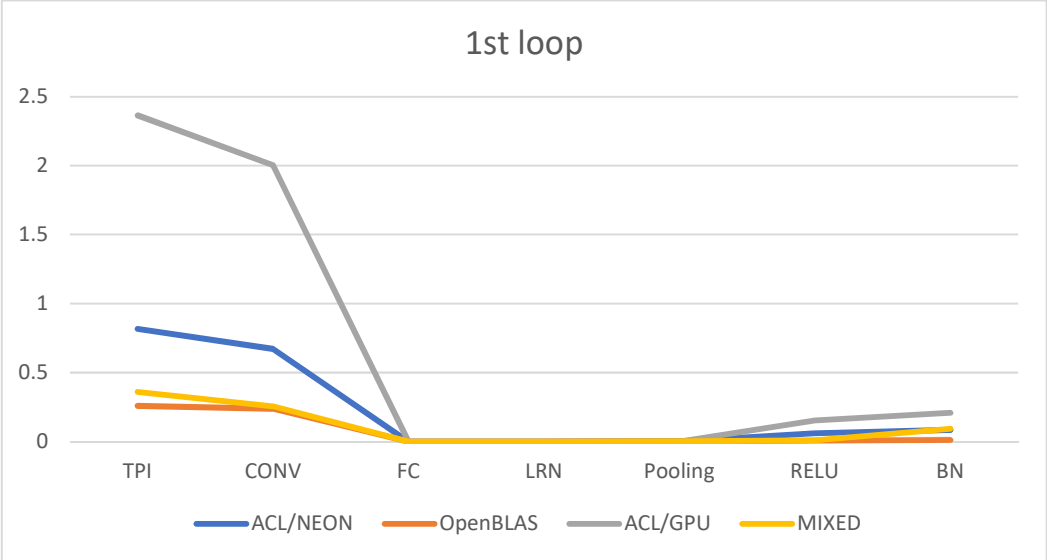


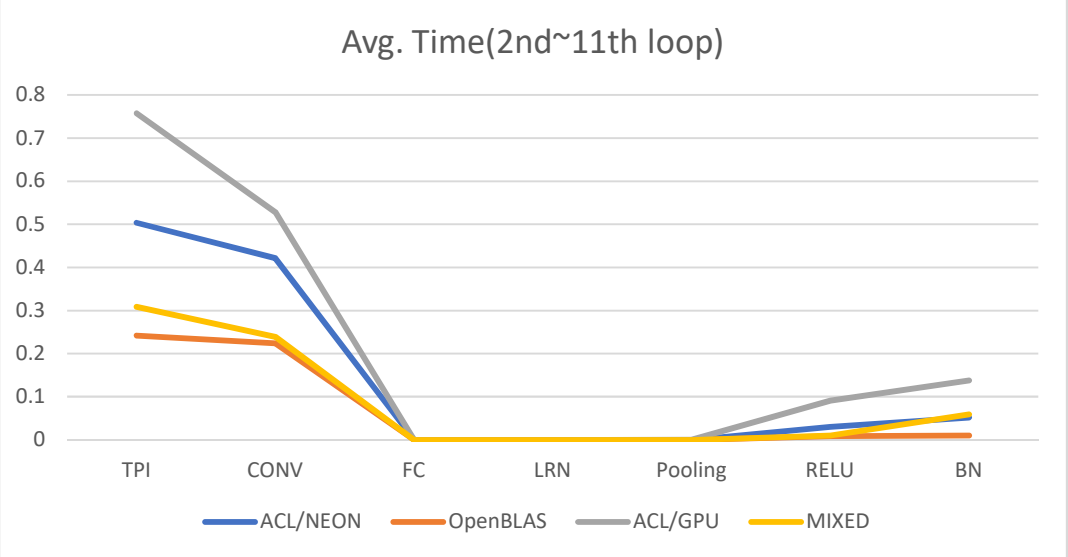
4.4 MobileNet

	TPI	Allocate	Run	Config	Copy
1st					
ACL/NEON	0.8166	0.2372	0.3783	0.0720	0.1883
OpenBLAS	0.2592	0	0	0	0
ACL/GPU	2.3644	0.3252	0.4289	1.4343	0.2696
MIXED	0.3601	0.0462	0.0288	0.0004	0.0307
Avg. Time					
ACL/NEON	0.5036	0.0311	0.3581	0.0005	0.0587
OpenBLAS	0.2419	0	0	0	0
ACL/GPU	0.7573	0.0835	0.4073	0.0245	0.1590
MIXED	0.3088	0.0151	0.0284	0	0.0293

	TPI	CONV	FC	LRN	Pooling	RELU	BN
1st							
ACL/NEON	0.8166	0.6711	0	0	0.0001	0.0596	0.0858
OpenBLAS	0.2592	0.2381	0	0	0.0001	0.0084	0.0126
ACL/GPU	2.3644	2.0040	0	0	0.0005	0.1516	0.2083
MIXED	0.3601	0.2562	0	0	0.0001	0.0100	0.0938
Avg. Time							

ACL/NEON	0.5036	0.4214	0	0	0.0001	0.0300	0.0521
OpenBLAS	0.2419	0.2236	0	0	0.0001	0.0084	0.0097
ACL/GPU	0.7573	0.5280	0	0	0.0003	0.0909	0.1381
MIXED	0.3088	0.2394	0	0	0.0001	0.0103	0.0591





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	2.2368	0.9531	0.9470
1xA72	0.5674	0.5177	0.5018
2xA72	0.3890	0.4778	0.4192
4xA53	0.7766	0.6638	0.6406
2xA72+4xA53*	0.4125	0.4816	0.4606

GoogleNet

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	1.9349	1.4855	1.3214
1xA72	0.7524	0.5624	0.4819
2xA72	0.4681	0.4243	0.3350
4xA53	1.3416	0.7624	0.6376
2xA72+4xA53*	1.1123	0.4429	0.4238

SqueezeNet.

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	0.5199	0.3068	0.3532
1xA72	0.2324	0.1155	0.1369
2xA72	0.1526	0.0784	0.0926
4xA53	0.4244	0.1549	0.2009
2xA72+4xA53*	0.4539	0.0858	0.0995

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

	ACL/NEON(s)	OpenBLAS(s)	MIXED(s)
1xA53	1.1847	0.6413	0.7436
1xA72	0.5036	0.2419	0.3088
2xA72	0.3607	0.1871	0.2284
4xA53	0.8798	0.4106	0.5012
2xA72+4xA53*	0.8396	0.1900	0.4739

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	0.9470	1.3214	0.3532	0.7436
1xA72	0.5018	0.4819	0.1369	0.3088
2xA72	0.4192	0.3350	0.0926	0.2284
4xA53	0.6406	0.6376	0.2009	0.5012
2xA72+4xA53	0.4606	0.4238	0.0995	0.4739

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/GPU	0.0821	0.0034	0	0
FC/ACL/NEON	0.3674	0.0063	0	0
FC/OpenBLAS	0.3366	0.0044	0	0