

Created on
Last modified

22/09/2017
24/01/2019

Author
Version
Document type

Almotive
1.4
Measurement Report

aiWare2 Benchmark Results

Measurement Report

CONFIDENTIAL

Created on	22/09/2017	Author	Almotive
Last modified	24/01/2019	Version	1.4
		Document type	Measurement Report

Contents

1	Introduction	3
2	Environment.....	3
2.1	HW and SW environment.....	3
2.2	Used neural networks.....	4
2.3	Methodology.....	4
2.3.1	GPU.....	5
2.3.2	aiWare	5
3	Benchmark	5
3.1	General considerations	5
3.2	Results.....	7
3.2.1	GOOGLENET_BASE	7
3.2.2	GOOGLENET_SEG	7
3.2.3	VGG16_BASE	7
3.2.4	VGG16_SEG.....	8
3.2.5	INCEPTION_V4_BASE.....	8
3.2.6	INCEPTION_V4_SEG.....	9
3.2.7	RESNET_50_BASE	9
3.2.8	RESNET_50_SEG	9
3.2.9	INCEPTION_RESNET_V2_BASE	10
3.2.10	INCEPTION_RESNET_V2_SEG	10
3.2.11	MOBILENET_BASE	10
3.2.12	MOBILENET_SEG	11
4	Conclusion.....	12
5	Appendix A – References	13
6	Appendix B – Terms and abbreviations.....	14
7	Appendix C – Document history	15

List of tables

Table 1. Specifications of used hardware	3
Table 2. Benchmarking results	11

Created on
Last modified

22/09/2017
24/01/2019

Author
Version
Document type

Almotive
1.4
Measurement Report

1 INTRODUCTION

As part of its product portfolio, Almotive offers a unique, application-independent and general NN accelerator IP, aiWare, which is scalable from embedded solutions up to data centers.

aiWare offers a dedicated AI computation core. It is independent from the type of the used neural network. Underpinned by its unique and scalable architecture, aiWare can process a wide range of image input of quality beyond 8K. aiWare also offers fixed and variable precision based on the application need (compilation time). aiWare is accompanied by a complex SDK that can make the necessary transformations to run a network trained in floating point domain on a fixed point aiWare variant. Even in this case—due to Almotive’s proprietary, hardware-level scaling technology—the quality degradation remains minimal.

This document contains the results of the benchmark measurements that Almotive carried out on generic NVIDIA GeForce GTX 1080 Ti GPU and on FPGA-based aiWare 2.0 evaluation kit.

2 ENVIRONMENT

2.1 HW AND SW ENVIRONMENT

Measurements were carried out in the following HW setup:

- GPU: NVIDIA GeForce GTX 1080 Ti
<https://www.geforce.co.uk/hardware/10series/geforce-gtx-1080-ti/>
- aiWare hardware IP 2.0 deployed on Nallatech 510T Compute Acceleration Card,
<http://www.nallatech.com/store/fpga-accelerated-computing/pcie-accelerator-cards/nallatech-510t-fpga-computing-acceleration-card/>
- CPU: Intel Xeon E3-1275 v5 3.6GHz
- Memory size: 64 GiB

The following table lists detailed specifications of used HW:

	GTX 1080 Ti	FPGA	
GMAC capacity	5669,888	GMAC capacity	194,56
GPU GFLOP capacity (GFLOP/s)	11339,776	FPGA GOP capacity (GOP/s)	389,12
CUDA cores	3584	Number of Cores	1
Clock frequency (MHz)	1582	Clock frequency (MHz)	190
Memory	11GB	Memory	8 GB
Memory bandwidth (GB/s)	484	Memory bandwidth (GB/s)	34

Table 1. Specifications of used hardware

The SW environment was made up of the following main components:

- Ubuntu 16.04.5 LTS
- NVIDIA Driver Version: 384.130
- aiWare Evaluation Kit; Almotive’s publicly available NN evaluation SW for aiWare
- cuDNN v7.1.4
- Caffe 1.0.0
- CUDA 9.0.176
- TensorFlow r1.9

Created on	22/09/2017	Author	Almotive
Last modified	24/01/2019	Version	1.4
		Document type	Measurement Report

2.2 USED NEURAL NETWORKS

The following public neural networks were used in the measurements:

- GoogLeNet
- GoogLeNet_seg
- VGG16
- VGG16_seg
- Inception_v4
- Inception_v4_seg
- ResNet_50
- ResNet_50_seg
- Inception_ResNet_v2
- Inception_ResNet_v2_seg
- MobileNet
- MobileNet_seg

The base networks (e.g. GoogLeNet) are the implementations of the classic convolutional networks but without the classification head, so that any kind of image processing network can be built upon them.

The segmentation networks (e.g. GoogLeNet_seg) are examples of the usage of these networks on the task of pixelwise semantic segmentation. In these examples, the base networks are used with a simple deconvolutional segmentation head.

We used both base network and segmentation network. We included segmentation network, as this can provide typical visual output of NN-based object detection. On the other hand, we included base network as that is flexible and allow attaching several kinds of extra layers for different kinds of output (classification, segmentation, bounding box, etc.). Their operations take up about 65-95% of NN operations, depending the architecture of the final network.

The descriptor files of actual neural networks used in this benchmark can be downloaded, see Appendix A – References.

2.3 METHODOLOGY

Measurements were taken with the following limits:

- Input was available in accelerator's local memory.
- Runtime was measured until the result arrived in the accelerator's local memory.
- Only runtime was measured, no power efficiency was considered.

The following operation requirements were applied:

- FPGA GOP requirement = ((int8) multiplications + (int8) additions + (int8) comparisons)/1000000000. GMAC requirement = GOPS/2
- GPU GFLOP requirement = ((float32) multiplications + (float32) additions + (float32) comparisons)/1000000000. GMAC requirement = GFLOPS/2

Int8 and float32 operations are compared due to the limitations of GPUs; the performance of the NNs are substantially the same on both precisions. Int8 operations result only in a minimal precision loss compared to float32.

- MAC utilization = $\text{MAC_busy_time} / \text{total_inference_time} = \text{NN_MAC_requirement} / (\text{total_inference_time} * \text{MAC_capacity})$

The runtime measurements were done on the above NNs with random weights and input data. The input data and weights were prepared in PC memory before runtime measurement; the output data of the network was not read back to keep the measure clean from acceleration-independent data movement times. The runtime measurements were done with PC timer on high number of network inferences (50x at GPU with Caffe and 50x at aiWare). The variance of the runtime values was less than 0.3%.

Created on	22/09/2017	Author	Almotive
Last modified	24/01/2019	Version	1.4
		Document type	Measurement Report

The aim of this benchmark is to compare aiWare and GPU hardware architectures for rendering NN inferences. aiWare is designed for latency constrained, real-time, embedded AI for computer-vision applications, while GPUs were originally designed for real-time 3D graphics rendering.

This benchmark does not include the original resolution of NNs. NNs were run on different input sizes ranging from 640x480 to 4096x2176 pixels (all using 3 input channels). MAC utilization is shown in the results table as it has key importance in embedded applications. MAC utilization is directly proportional with performance and affects power efficiency independently from target technology. MAC utilization is also a good measure on how optimal a hardware architecture is for a given task; which is rendering NN inferences in this benchmark.

2.3.1 GPU

The runtime measurements were done with Caffe as it has provided the highest performance amongst software used in the wider community.

The performance related settings of the GPU were set for maximum performance (i.e. Prefer Maximum Performance, Fan speed fixed to 100%).

2.3.2 AIWARE

The original Caffe models were converted automatically to an aiWare-proprietary format. The conversion and the runtime measurement were done with aiWare SDK. The aiWare hardware IP was deployed on an off-the-shelf Intel FPGA-based PCIe development board.

3 BENCHMARK

3.1 GENERAL CONSIDERATIONS

aiWare2 accelerates NN layer types more efficiently than its predecessor, aiWare IP 1.0. Furthermore – due to architectural optimizations – it can reach higher operational frequencies both in FPGAs and ASICs. Table below shows the MAC Utilization ratios on 2 Megapixel input size with different NN architectures.

NN architecture	aiWare 1.0 @	aiWare2 @	GeForce GTX 1080
	160MHz 3x1920x1056	190MHz 3x1920x1056	Ti@1582MHz 3x1920x1056
GoogLeNet	91.45%	90.02%	40.35%
GoogLeNet_seg	84.77%	84.07%	29.34%
Inception_ResNet_V2	82.66%	90.72%	- *
Inception_ResNet_V2_seg	81.64%	89.62%	- *
Inception_v4	85.60%	92.77%	- *
Inception_v4_seg	84.40%	91.54%	- *
MobileNet	- **	89.88%	45.42%
MobileNet_seg	- **	87.91%	22.68%
ResNet50	80.68%	87.70%	38.24%
ResNet50_seg	78.50%	85.25%	33.25%
VGG16	97.22%	96.79%	88.65%
VGG16_seg	96.36%	96,02%	79.57%
Average value:	86.33%	89.66%	47.19%

Table 2. Comparison of MAC Utilization with different nets and processors

* Insufficient memory on GPU

** MobileNet was not measured with earlier processor version

MobileNet was also added to the CNN architecture list and measured with aiWare2.

Created on	22/09/2017	Author	Almotive
Last modified	24/01/2019	Version	1.4
		Document type	Measurement Report

For “aiWare2 @5,6TMAC”, we have selected the aiWare clock frequency and core number to match the GPU’s GMAC capacity (5,6 TMAC in this benchmark). These attributes of the aiWare hardware IP can be linearly scaled to the values used in this benchmark. A high-performance ASIC implementation of the aiWare hardware IP will be available in 2018 for internal testing purposes.

Created on
Last modified

22/09/2017
24/01/2019

Author
Version
Document type

Almotive
1.4
Measurement Report

3.2 RESULTS

The following table summarizes the main results of our benchmarking:

			GeForce GTX 1080 Ti (5.6TMAC)		aiWare IP			aiWare/GPU
			Caffe		aiWare2 in FPGA @0.19 TMAC		aiWare2 @5.6 TMAC	
Network name	Size	GMAC requirement	FPS	MAC utilization	FPS	MAC utilization	FPS	MAC utilization ratio
								GeForce GTX 1080 Ti (5.6TMAC)
3.2.1 GOOGLNET_BASE								
googlenet base	3x4096x2176	282,09	9,18	45,65%	0,65	94,75%	19,06	2,08x
googlenet base	3x2560x1920	155,56	16,51	45,29%	1,20	95,73%	34,87	2,11x
googlenet base	3x2048x1536	99,56	25,29	44,40%	1,88	96,36%	55,04	2,17x
googlenet base	3x2048x1024	66,37	35,37	41,41%	2,83	96,44%	82,71	2,33x
googlenet base	3x1920x1056	64,17	35,65	40,35%	2,73	90,02%	79,81	2,23x
googlenet base	3x1280x704	28,52	65,53	32,96%	6,36	93,20%	185,53	2,83x
googlenet base	3x800x576	14,58	88,73	22,82%	11,06	82,92%	323,62	3,63x
googlenet base	3x704x480	10,69	102,88	19,40%	15,87	87,23%	462,96	4,50x
googlenet base	3x640x480	9,72	105,60	18,10%	17,59	87,89%	515,46	4,86x
3.2.2 GOOGLNET_SEG								
googlenet seg	3x4096x2176	286,26	-*	-*	0,60	88,16%	17,51	-*
googlenet seg	3x2560x1920	157,86	11,53	32,10%	1,10	88,88%	31,96	2,77x
googlenet seg	3x2048x1536	101,03	17,67	31,49%	1,73	89,61%	50,38	2,85x
googlenet seg	3x2048x1024	67,36	25,08	29,79%	2,59	89,77%	75,82	3,01x
googlenet seg	3x1920x1056	65,12	25,54	29,33%	2,51	84,07%	73,37	2,87x
googlenet seg	3x1280x704	28,94	48,78	24,89%	5,83	86,78%	170,36	3,49x
googlenet seg	3x800x576	14,8	70,52	18,41%	10,21	77,67%	297,62	4,22x
googlenet seg	3x704x480	10,85	83,06	15,89%	14,61	81,51%	429,18	5,13x
googlenet seg	3x640x480	9,87	85,54	14,88%	16,21	82,19%	476,19	5,52x
3.2.3 VGG16_BASE								
vgg16 base	3x4096x2176	2726,61	-*	-*	0,07	98,12%	2,04	-*
vgg16 base	3x2560x1920	1503,64	-*	-*	0,13	98,53%	3,72	-*
vgg16 base	3x2048x1536	962,33	5,57	94,61%	0,20	98,30%	5,81	1,04x

Created on
Last modified

22/09/2017
24/01/2019

Author
Version
Document type

Almotive
1.4
Measurement Report

			GeForce GTX 1080 Ti (5.6TMAC)		aiWare IP			aiWare/GPU
			Caffe		aiWare2 in FPGA @0.19 TMAC	aiWare2 @5.6 TMAC		
Network name	Size	GMAC requirement	FPS	MAC utilization	FPS	MAC utilization	FPS	MAC utilization ratio
								GeForce GTX 1080 Ti (5.6TMAC)
vgg16 base	3x2048x1024	641,55	8,16	92,34%	0,30	98,38%	8,71	1,07x
vgg16 base	3x1920x1056	620,25	8,10	88,65%	0,30	96,79%	8,87	1,09x
vgg16 base	3x1280x704	275,67	18,83	91,52%	0,69	98,08%	20,21	1,07x
vgg16 base	3x800x576	140,97	33,52	83,35%	1,32	95,88%	38,64	1,15x
vgg16 base	3x704x480	103,38	47,73	87,00%	1,81	96,08%	52,80	1,10x
vgg16 base	3x640x480	93,98	50,48	83,65%	2,00	96,46%	58,38	1,15x
3.2.4 VGG16_SEG								
vgg16 seg	3x4096x2176	2730,69	-*	-*	0,07	97,33%	2,03	-*
vgg16 seg	3x2560x1920	1505,89	-*	-*	0,13	97,75%	3,69	-*
vgg16 seg	3x2048x1536	963,77	5,04	85,63%	0,20	97,51%	5,75	1,14x
vgg16 seg	3x2048x1024	642,52	7,41	83,94%	0,30	97,58%	8,63	1,16x
vgg16 seg	3x1920x1056	621,18	7,26	79,57%	0,30	96,02%	8,78	1,21x
vgg16 seg	3x1280x704	276,08	16,97	82,64%	0,69	97,28%	20,03	1,18x
vgg16 seg	3x800x576	141,18	30,23	75,27%	1,31	95,08%	38,28	1,26x
vgg16 seg	3x704x480	103,53	42,61	77,79%	1,79	95,20%	52,30	1,22x
vgg16 seg	3x640x480	94,12	45,39	75,34%	1,98	95,81%	57,87	1,27x
3.2.5 INCEPTION_V4_BASE								
inception_v4_base	3x4096x2176	1451,05	-*	-*	0,13	97,69%	3,83	-*
inception_v4_base	3x2560x1920	800,21	-*	-*	0,24	97,79%	6,94	-*
inception_v4_base	3x2048x1536	512,13	-*	-*	0,37	98,68%	10,95	-*
inception_v4_base	3x2048x1024	341,42	-*	-*	0,56	98,81%	16,45	-*
inception_v4_base	3x1920x1056	330,09	-*	-*	0,55	92,77%	15,97	-*
inception_v4_base	3x1280x704	146,71	11,69	30,25%	1,23	92,86%	35,97	3,07x
inception_v4_base	3x800x576	75,02	17,44	23,07%	2,17	83,64%	63,37	3,63x
inception_v4_base	3x704x480	55,01	18,83	18,27%	3,08	87,02%	89,93	4,76x
inception_v4_base	3x640x480	50,01	19,64	17,32%	3,52	90,51%	102,88	5,23x

Created on
Last modified

22/09/2017
24/01/2019

Author
Version
Document type

Almotive
1.4
Measurement Report

			GeForce GTX 1080 Ti (5.6TMAC)		aiWare IP			aiWare/GPU
			Caffe		aiWare2 in FPGA @0.19 TMAC		aiWare2 @5.6 TMAC	
Network name	Size	GMAC requirement	FPS	MAC utilization	FPS	MAC utilization	FPS	MAC utilization ratio
								GeForce GTX 1080 Ti (5.6TMAC)

3.2.6 INCEPTION_V4_SEG

inception_v4_seg	3x4096x2176	1455,32	-*	-*	0,13	96,28%	3,76	-*
inception_v4_seg	3x2560x1920	802,57	-*	-*	0,23	96,38%	6,82	-*
inception_v4_seg	3x2048x1536	513,64	-*	-*	0,37	97,25%	10,76	-*
inception_v4_seg	3x2048x1024	342,43	-*	-*	0,55	97,35%	16,16	-*
inception_v4_seg	3x1920x1056	331,06	-*	-*	0,54	91,54%	15,71	-*
inception_v4_seg	3x1280x704	147,14	11,00	28,54%	1,21	91,60%	35,39	3,21x
inception_v4_seg	3x800x576	75,24	16,51	21,90%	2,14	82,63%	62,42	3,77x
inception_v4_seg	3x704x480	55,18	18,12	17,63%	3,03	85,95%	88,57	4,88x
inception_v4_seg	3x640x480	50,16	18,67	16,51%	3,46	89,22%	101,21	5,40x

3.2.7 RESNET_50_BASE

resnet_50_base	3x4096x2176	687,06	-*	-*	0,26	93,14%	7,69	-*
resnet_50_base	3x2560x1920	378,89	-*	-*	0,48	93,16%	13,99	-*
resnet_50_base	3x2048x1536	242,49	9,46	40,45%	0,77	96,20%	22,53	2,38x
resnet_50_base	3x2048x1024	161,66	13,61	38,82%	1,16	96,51%	33,91	2,49x
resnet_50_base	3x1920x1056	156,29	13,87	38,24%	1,09	87,70%	31,89	2,29x
resnet_50_base	3x1280x704	69,46	30,33	37,16%	2,54	90,84%	74,29	2,44x
resnet_50_base	3x800x576	35,52	47,82	29,95%	4,13	75,47%	120,77	2,52x
resnet_50_base	3x704x480	26,05	61,31	28,16%	6,18	82,70%	180,83	2,94x
resnet_50_base	3x640x480	23,68	64,72	27,02%	6,82	82,97%	199,60	3,07x

3.2.8 RESNET_50_SEG

resnet_50_seg	3x4096x2176	691,42	-*	-*	0,25	90,42%	7,43	-*
resnet_50_seg	3x2560x1920	381,3	-*	-*	0,46	90,47%	13,50	-*
resnet_50_seg	3x2048x1536	244,03	8,13	34,98%	0,74	93,22%	21,72	2,66x
resnet_50_seg	3x2048x1024	162,69	11,76	33,74%	1,12	93,57%	32,69	2,77x
resnet_50_seg	3x1920x1056	157,29	11,99	33,25%	1,05	85,25%	30,80	2,56x
resnet_50_seg	3x1280x704	69,9	25,93	31,97%	2,46	88,28%	71,74	2,76x

Created on
Last modified

22/09/2017
24/01/2019

Author
Version
Document type

Almotive
1.4
Measurement Report

			GeForce GTX 1080 Ti (5.6TMAC)		aiWare IP			aiWare/GPU
			Caffe		aiWare2 in FPGA @0.19 TMAC		aiWare2 @5.6 TMAC	
Network name	Size	GMAC requirement	FPS	MAC utilization	FPS	MAC utilization	FPS	MAC utilization ratio
								GeForce GTX 1080 Ti (5.6TMAC)
resnet_50_seg	3x800x576	35,75	40,75	25,69%	4,01	73,66%	117,10	2,87x
resnet_50_seg	3x704x480	26,21	53,56	24,76%	5,99	80,68%	174,52	3,26x
resnet_50_seg	3x640x480	23,83	56,53	23,75%	6,59	80,75%	193,05	3,40x

3.2.9 INCEPTION_RESNET_V2_BASE

inception_resnet_v2_base	3x4096x2176	1577,7	-*	-*	0,12	96,12%	3,46	-*
inception_resnet_v2_base	3x2560x1920	870,06	-*	-*	0,22	96,52%	6,30	-*
inception_resnet_v2_base	3x2048x1536	556,84	-*	-*	0,34	97,13%	9,92	-*
inception_resnet_v2_base	3x2048x1024	371,22	-*	-*	0,51	96,53%	14,78	-*
inception_resnet_v2_base	3x1920x1056	358,9	-*	-*	0,49	90,72%	14,36	-*
inception_resnet_v2_base	3x1280x704	159,51	6,47	18,20%	1,07	88,10%	31,42	4,84x
inception_resnet_v2_base	3x800x576	81,57	9,77	14,06%	1,83	76,56%	53,36	5,45x
inception_resnet_v2_base	3x704x480	59,82	11,43	12,06%	2,57	78,98%	75,02	6,55x
inception_resnet_v2_base	3x640x480	54,38	11,86	11,37%	2,89	80,78%	84,46	7,10x

3.2.10 INCEPTION_RESNET_V2_SEG

inception_resnet_v2_seg	3x4096x2176	1581,97	-*	-*	0,12	94,87%	3,41	-*
inception_resnet_v2_seg	3x2560x1920	872,41	-*	-*	0,21	95,25%	6,21	-*
inception_resnet_v2_seg	3x2048x1536	558,34	-*	-*	0,33	95,88%	9,76	-*
inception_resnet_v2_seg	3x2048x1024	372,23	-*	-*	0,50	95,31%	14,55	-*
inception_resnet_v2_seg	3x1920x1056	359,87	-*	-*	0,48	89,62%	14,15	-*
inception_resnet_v2_seg	3x1280x704	159,94	6,23	17,58%	1,06	87,09%	30,95	4,95x
inception_resnet_v2_seg	3x800x576	81,79	9,61	13,86%	1,80	75,83%	54,67	5,47x
inception_resnet_v2_seg	3x704x480	59,98	11,16	11,80%	2,54	78,23%	74,13	6,63x
inception_resnet_v2_seg	3x640x480	54,53	11,67	11,22%	2,85	80,00%	83,33	7,13x

3.2.11 MOBILENET_BASE

mobilenet_base	3x4096x2176	876,09	-*	-*	0,21	93,93%	6,09	-*
mobilenet_base	3x2560x1920	483,14	-*	-*	0,38	94,07%	11,07	-*
mobilenet_base	3x2048x1536	309,21	-*	-*	0,60	94,70%	17,40	-*

Created on
Last modified

22/09/2017
24/01/2019

Author
Version
Document type

Almotive
1.4
Measurement Report

			GeForce GTX 1080 Ti (5.6TMAC)		aiWare IP			aiWare/GPU
			Caffe		aiWare2 in FPGA @0.19 TMAC		aiWare2 @5.6 TMAC	
Network name	Size	GMAC requirement	FPS	MAC utilization	FPS	MAC utilization	FPS	MAC utilization ratio
								GeForce GTX 1080 Ti (5.6TMAC)
mobilenet_base	3x2048x1024	206,14	12,25	44,55%	0,89	94,69%	26,11	2,13x
mobilenet_base	3x1920x1056	199,29	12,92	45,42%	0,85	86,88%	24,77	1,91x
mobilenet_base	3x1280x704	88,58	23,55	36,79%	1,96	89,27%	57,24	2,43x
mobilenet_base	3x800x576	45,29	34,23	27,35%	3,30	76,83%	96,53	2,81x
mobilenet_base	3x704x480	33,22	40,14	23,52%	4,93	84,22%	144,30	3,58x
mobilenet_base	3x640x480	30,2	41,58	22,14%	5,62	87,19%	163,67	3,94x
3.2.12 MOBILENET_SEG								
mobilenet_seg	3x4096x2176	871,45	-*	-*	0,21	95,15%	6,20	-*
mobilenet_seg	3x2560x1920	480,58	-*	-*	0,39	95,32%	11,27	-*
mobilenet_seg	3x2048x1536	307,57	-*	-*	0,61	95,99%	17,74	-*
mobilenet_seg	3x2048x1024	205,05	6,20	22,41%	0,91	95,98%	26,62	4,28x
mobilenet_seg	3x1920x1056	198,24	6,49	22,68%	0,86	87,91%	25,21	3,88x
mobilenet_seg	3x1280x704	88,11	13,03	20,24%	2,00	90,39%	58,28	4,47x
mobilenet_seg	3x800x576	45,05	20,76	16,49%	3,36	77,71%	98,14	4,71x
mobilenet_seg	3x704x480	33,04	25,51	14,87%	5,03	85,34%	146,41	5,74x
mobilenet_seg	3x640x480	30,04	27,40	14,52%	5,71	88,22%	166,39	6,08x

Table 2. Benchmarking results

* Insufficient memory on GPU

Created on	22/09/2017	Author	Almotive
Last modified	24/01/2019	Version	1.4
		Document type	Measurement Report

4 CONCLUSION

Measurements have shown that Almotive's aiWare solution beats even the most powerful commercially available GPUs in terms of performance and efficiency, with ratios ranging from about 104% to even 714%.

Our dedicated NN accelerator HW provides exceptional results both in terms of performance and MAC utilization, due its dedicated architecture providing support for NN layers that comprises 80-99% of processing power requirements of a typical NN.

This benchmark shows the viability of Almotive's aiWare solution for NN acceleration.

Created on	22/09/2017	Author	Almotive
Last modified	24/01/2019	Version	1.4
		Document type	Measurement Report

5 APPENDIX A – REFERENCES

- The actual NN descriptor files used in this benchmark in Caffe `.prototxt` format can be downloaded from [here](#).

Created on
Last modified

22/09/2017
24/01/2019

Author
Version
Document type

Almotive
1.4
Measurement Report

6 APPENDIX B – TERMS AND ABBREVIATIONS

The following terms and abbreviations are used in this document:

Terms	Description
ASIC	Application-Specific Integrated Circuit
NN	Neural Network
FPGA	Field-Programmable Gate Array
GFLOP	Giga Floating Point Operations
GMAC	Giga Multiply-Accumulate Operations
GOP	Giga Operations
GPU	Graphics Processing Unit

Created on
Last modified

22/09/2017
24/01/2019

Author
Version
Document type

Almotive
1.4
Measurement Report

7 APPENDIX C – DOCUMENT HISTORY

Version	Date	Modified by	Description of the modification
1.0	04/10/2017	Tamas Forgacs Andras Juhasz	Initial public version
1.1	04/12/2017	Tamas Forgacs Andras Juhasz	Minor corrections in Table 2. Benchmarking results
1.2	12/12/2017	Tamas Forgacs Andras Juhasz	Measurements in Table 2. Benchmarking results extended with new NNS
1.3	01/02/2018	Tamas Forgacs Andras Juhasz	Measurements in Table 2. Benchmarking results extended with new NNS
1.4	24/01/2019	Peter Frank	Measurements in Table 2. Benchmarking results extended with new NNS

Created on	22/09/2017	Author	Almotive
Last modified	24/01/2019	Version	1.4
		Document type	Measurement Report

Disclaimer

The information in this document is subject to change without notice and describes only the product defined in the introduction of this documentation.

This documentation is intended for the use of Almotive customers only for the purposes of the agreement under which the document is submitted, and no part of it may be used, reproduced, modified or transmitted in any form or means without the prior written permission of Almotive.

The documentation has been prepared to be used by professional and properly trained personnel, and the customer assumes full responsibility when using it.

Almotive welcomes customer comments as part of the process of continuous development and improvement of the documentation.

The information or statements given in this documentation concerning the suitability, capacity, or performance of the mentioned software products are given "as is" and all liability arising in connection with such software products shall be defined conclusively and finally in a separate agreement between Almotive and the customer.

The information disclosed shall not constitute any representation, warranty, assurance, or guarantee by Almotive to the customer of any kind, in particular, with respect to the non-infringement of trademarks, patents, copyrights or any other intellectual property rights, or other rights of third parties.

However, Almotive has made all reasonable efforts to ensure that the instructions contained in the document are adequate and free of material errors and omissions.

Almotive will, if deemed necessary by Almotive, explain issues which may not be covered by the document.

Almotive will correct errors in this documentation as soon as possible.

IN NO EVENT WILL Almotive BE LIABLE FOR ERRORS IN THIS DOCUMENTATION OR FOR ANY DAMAGES, INCLUDING BUT NOT LIMITED TO SPECIAL, DIRECT, INDIRECT, INCIDENTAL OR CONSEQUENTIAL OR ANY LOSSES, SUCH AS BUT NOT LIMITED TO LOSS OF PROFIT, REVENUE, BUSINESS INTERRUPTION, BUSINESS OPPORTUNITY OR DATA, THAT MAY ARISE FROM THE USE OF THIS DOCUMENT OR THE INFORMATION IN IT.

This documentation and the product it describes are considered protected by copyrights and other intellectual property rights according to the applicable laws.

The logo is a trademark of Almotive Kft. Other product names mentioned in this document may be trademarks of their respective owners, and they are mentioned for identification purposes only.

If any provision or part of a provision of this disclaimer shall be, or is found by any court of competent jurisdiction or public authority to be invalid or unenforceable, such invalidity or unenforceability shall not affect the other provisions or parts of such provisions of this disclaimer; all of which shall remain in full force and effect.

Copyright © 2019 Almotive All rights reserved