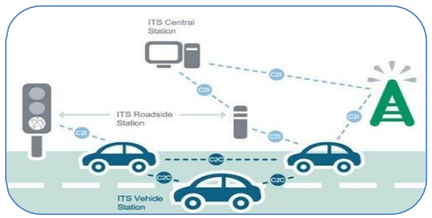
High Quality Accelerate Solution Provider

Demand from Development

Mobile Internet

Smart City

Internet of Vehicle



Smart Traffic

.....

Enormous numbers of pictures and videos are generated by devices

Media file with larger size and higher resolution

More and more real-time image video Analytics demand

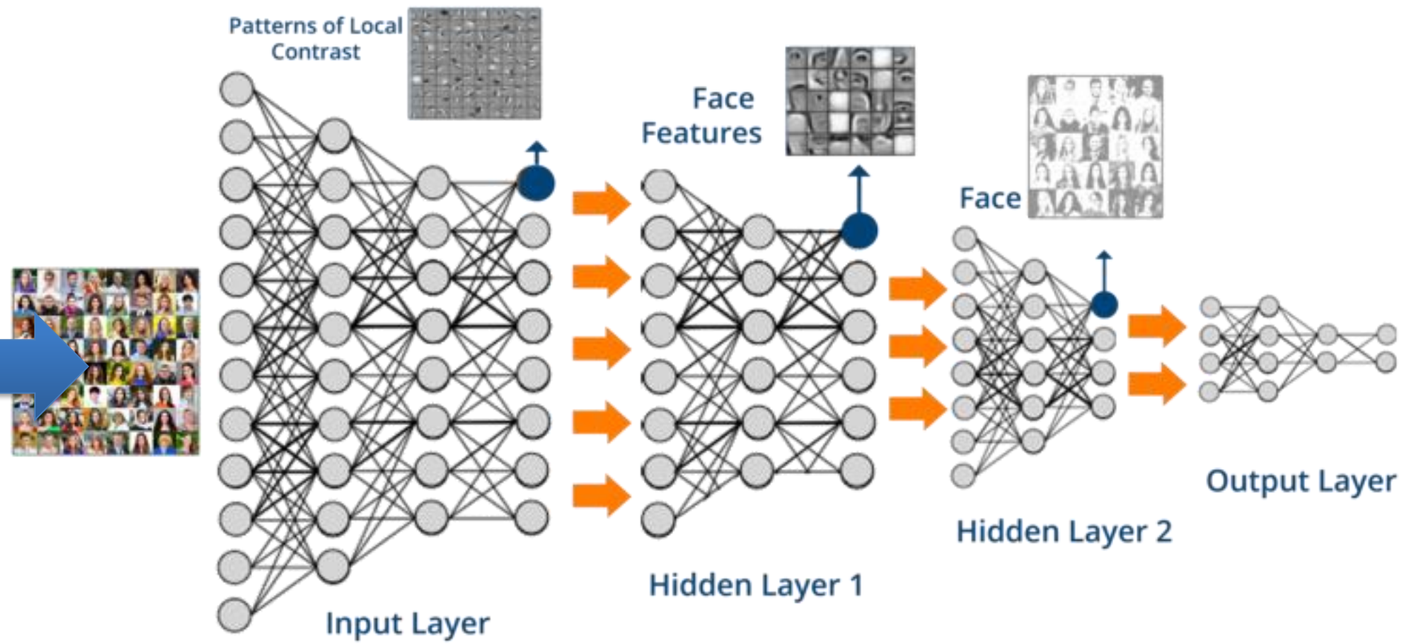
Consume large amount of computing resources

Huge expenditure for bandwidth

Long delay leads to poor customer experience

AI Process Break-down

Preprocessing



Step 1

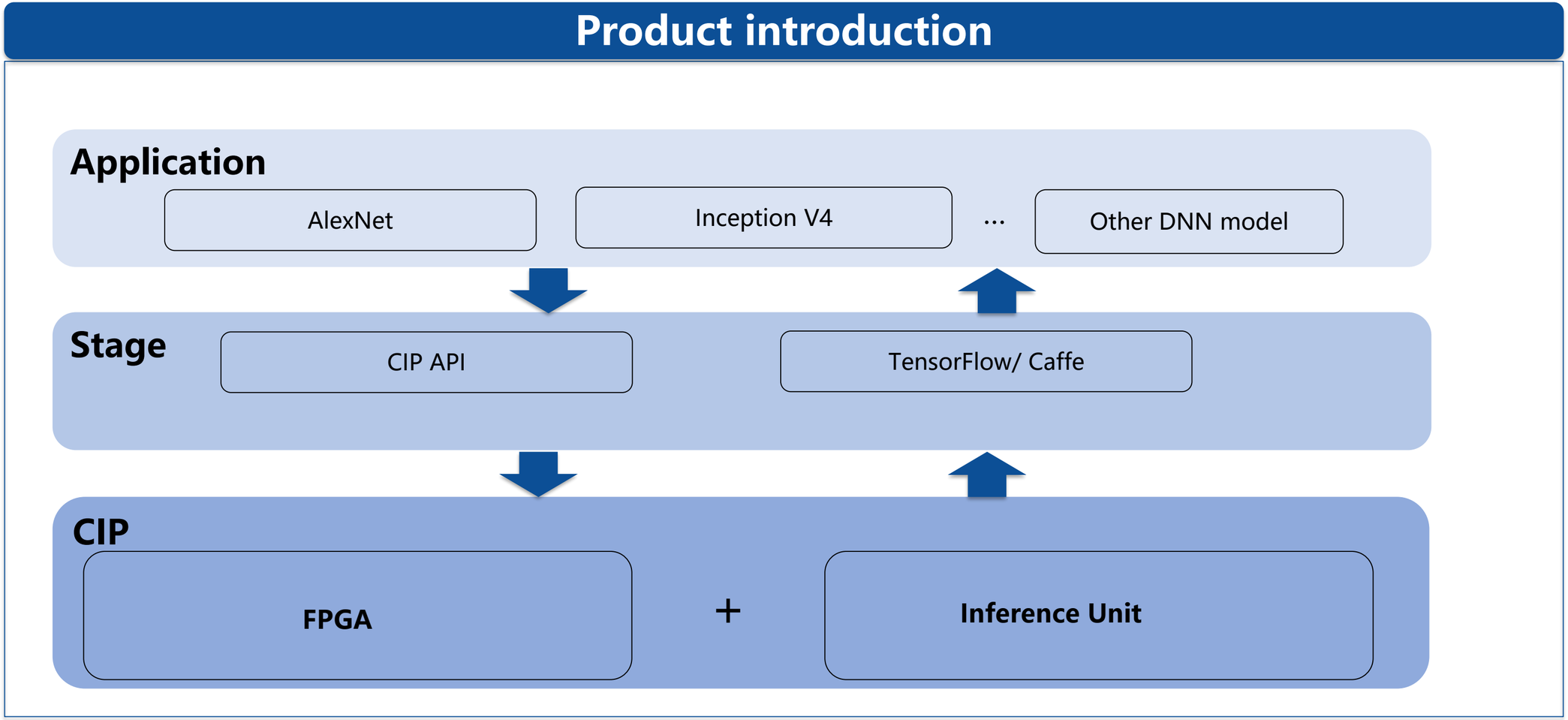
Data reformatting

Step 2

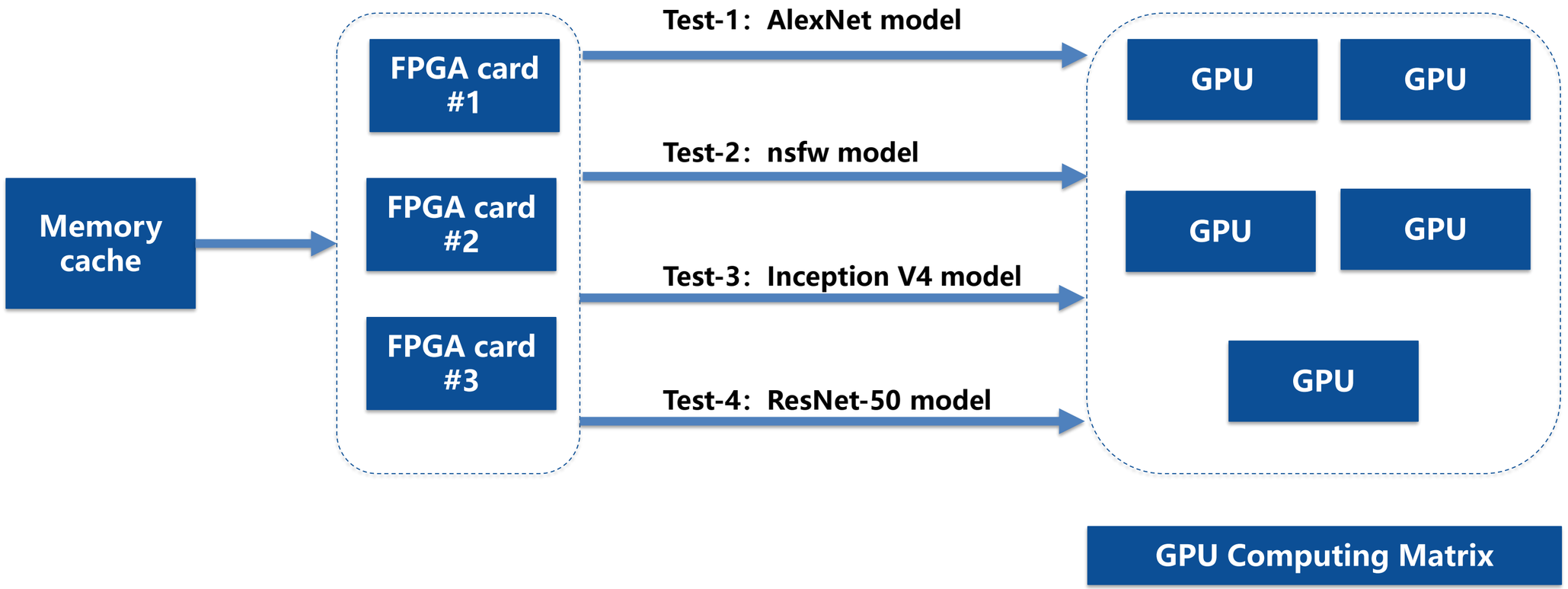
Supply into model

Solution Structure：FPGA+GPU Strong Alliance

Product introduction



Solution Model



Test Content

The test is divided into 2 parts：

- The first part is a performance test, mainly testing throughput, CPU usage and latency.

- The second part is the accuracy test. It mainly tests whether the FPGA replacing the CPU for preprocessing will affect the accuracy of the neural network.

Indicators

Index	Unit	Meaning	Test method
Throughput	MB/s	Amount of data processed per second	Step1: Total processing time for reading all images through the test program: Step2: Throughput = total library size / total processing time of the image
Latency	ms	Processing time of a single picture	Step1: Read the processing time of each image through the test program Step2: Delay = average of single image processing time
CPU utilization	%	Server CPU usage	Test program read from system directly

Test Environment

Test environment:

- CPU: 2*Intel(R) Xeon(R) CPU E5-2690v4 x 2
- RAM: 128GB
- OS: CentOS Linux release 7.4
- Kernel version: 3.10.0-514.2.2.el7.x86_64
- Python version: 2.7

Input:

Pic categories	Resolution	Amount
8 Mega Pixel	2647X3278	14708
16 Mega Pixel	5312x2988	9280
1080p	1920x1080	21400

TensorFlow Application resolution:

Model	Resolution
AlexNet	227*227
nsfw	224*224
InceptionV4	299*299
ResNet_50	224*224

Comparative Analysis of Test Data -Alexnet

Alexnet Model

QPS:

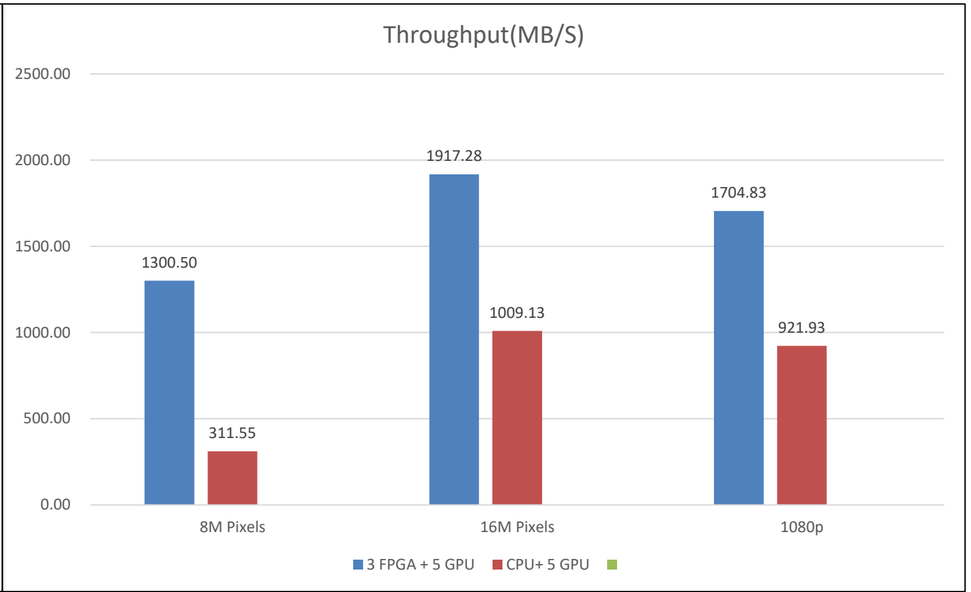
FPGA+GPU solution is **2.5**times of CPU+ GPU solution

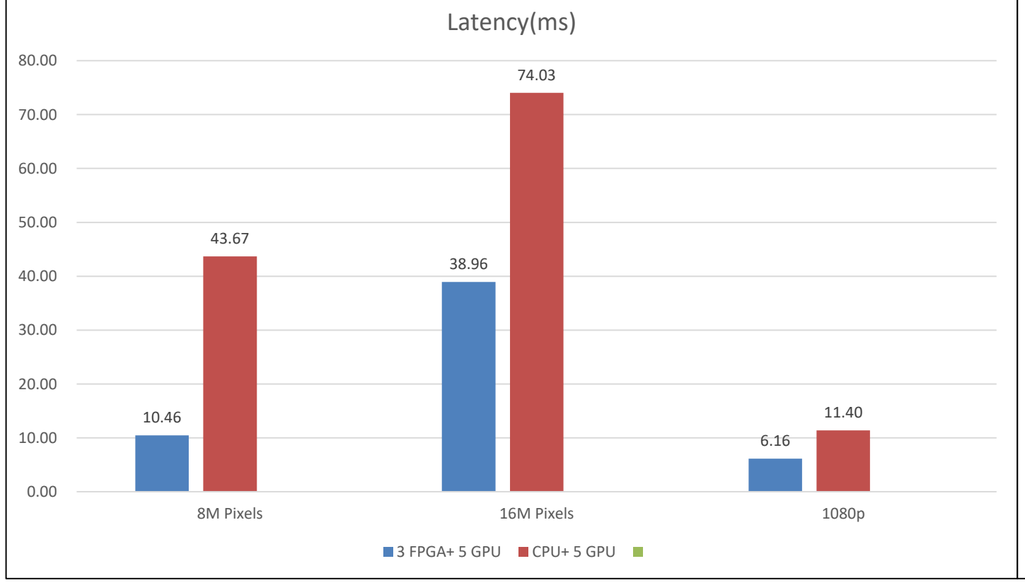
latency:

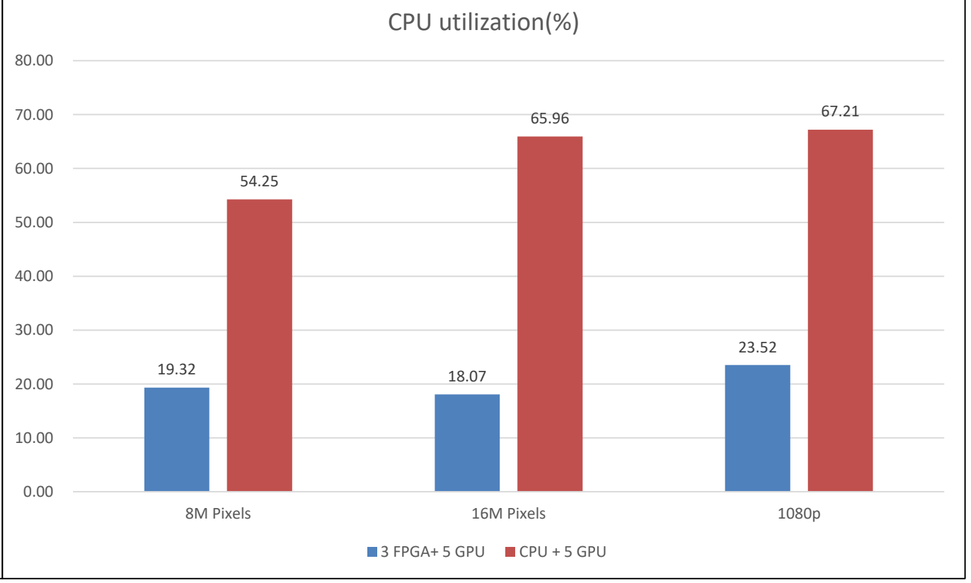
FPGA+GPU solution is **30%** of CPU+GPU solution

CPU usage rate:

FPGA+GPU solution is **20%** of CPU+GPU solution







Comparative Analysis of Test Data-nsfw

Nsfw Model

QPS:

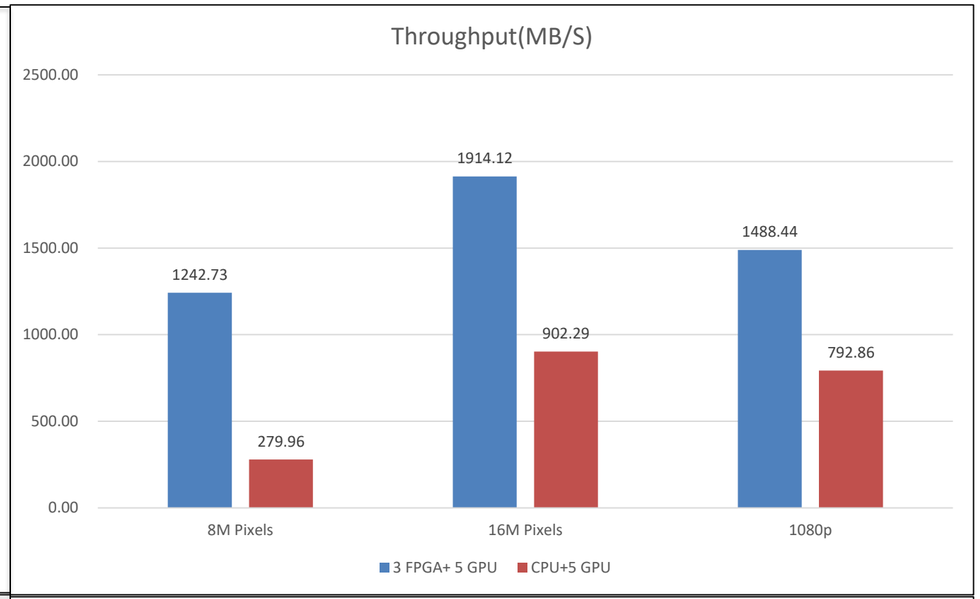
FPGA +GPU solution is **3** times of CPU+GPU solution

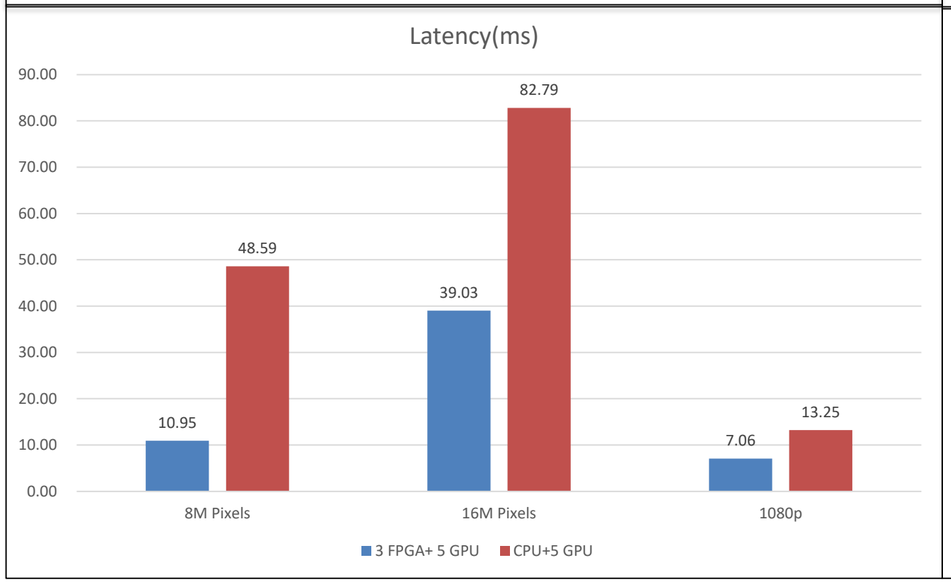
Latecncy:

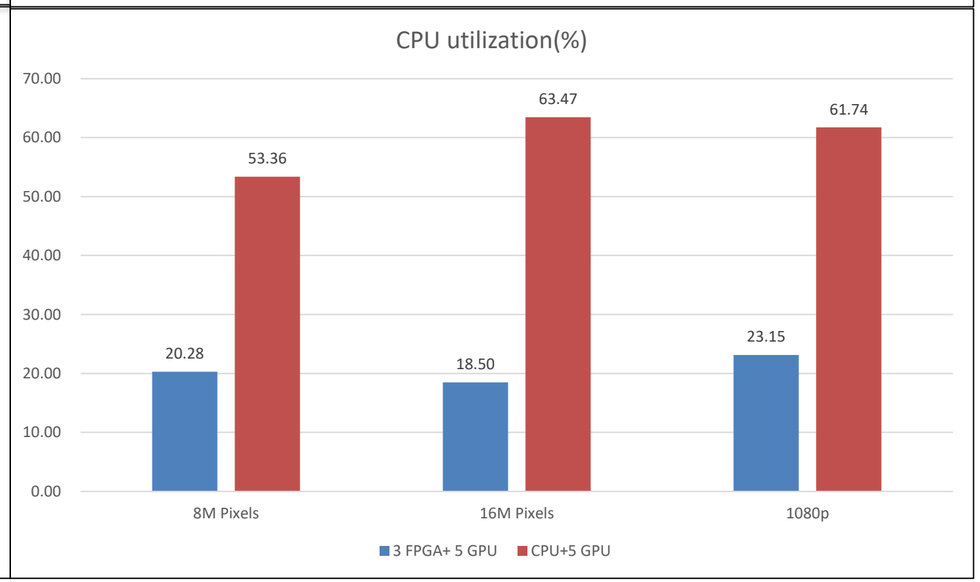
FPGA +GPU solution is **30%** of CPU+GPU solution

CPU usage rate:

FPGA +GPU solution is **20%** of CPU+GPU solution







Comparative Analysis of Test Data-Inception V4

InceptionV4 Model

QPS:

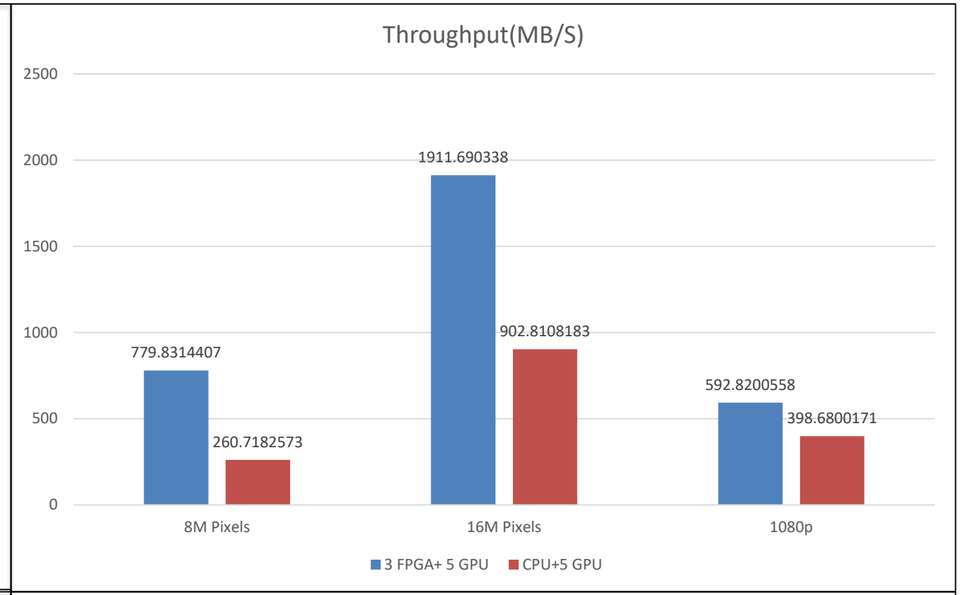
FPGA +GPU solution is **2** times of CPU+GPU solution

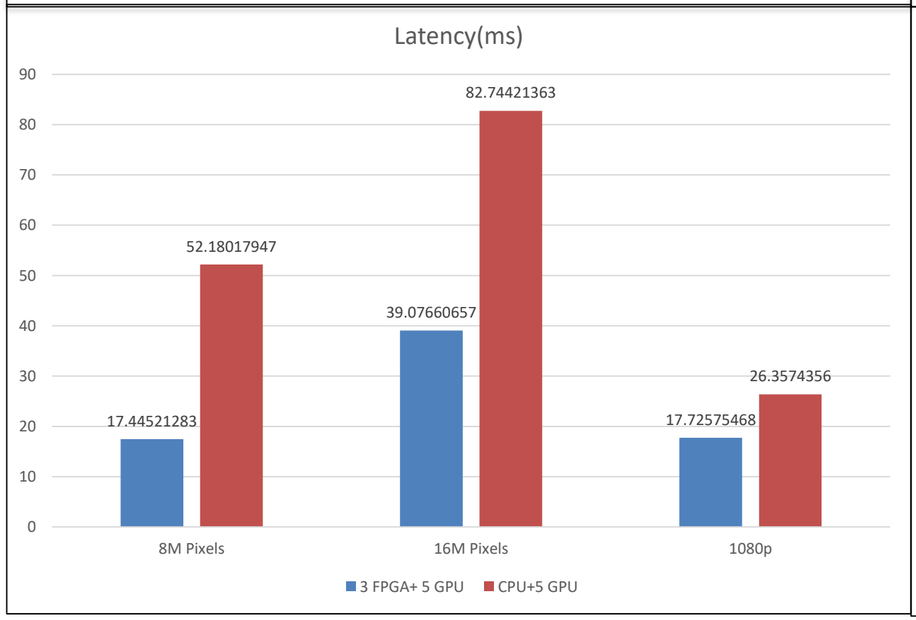
Latency:

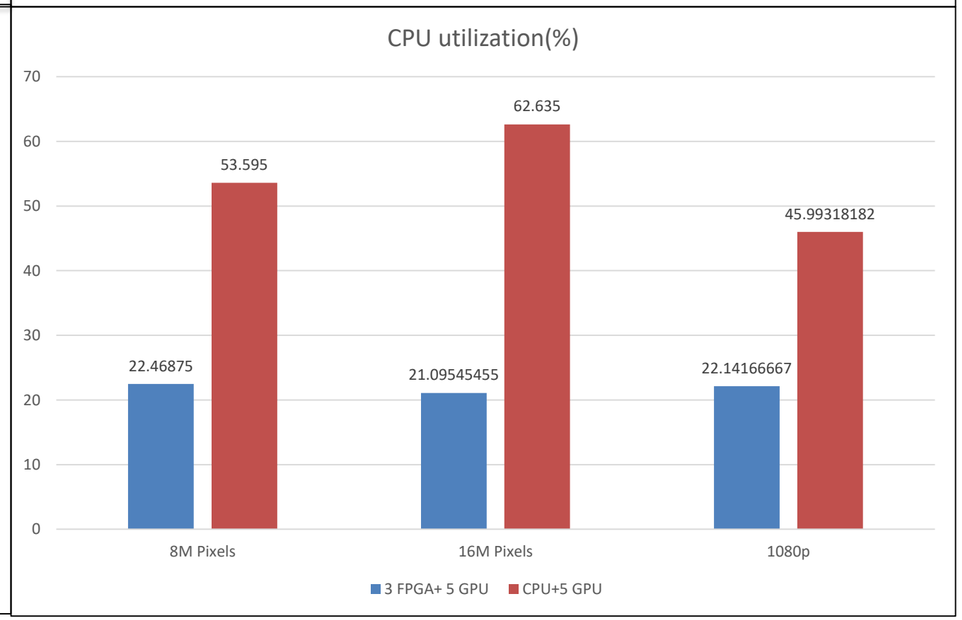
FPGA +GPU solution is **40%** of CPU+GPU solution

CPU usage rate:

FPGA +GPU solution is **30%** of CPU+GPU solution







Comparative Analysis of Test Data-ResNet-50

ResNet-50 Model

QPS:

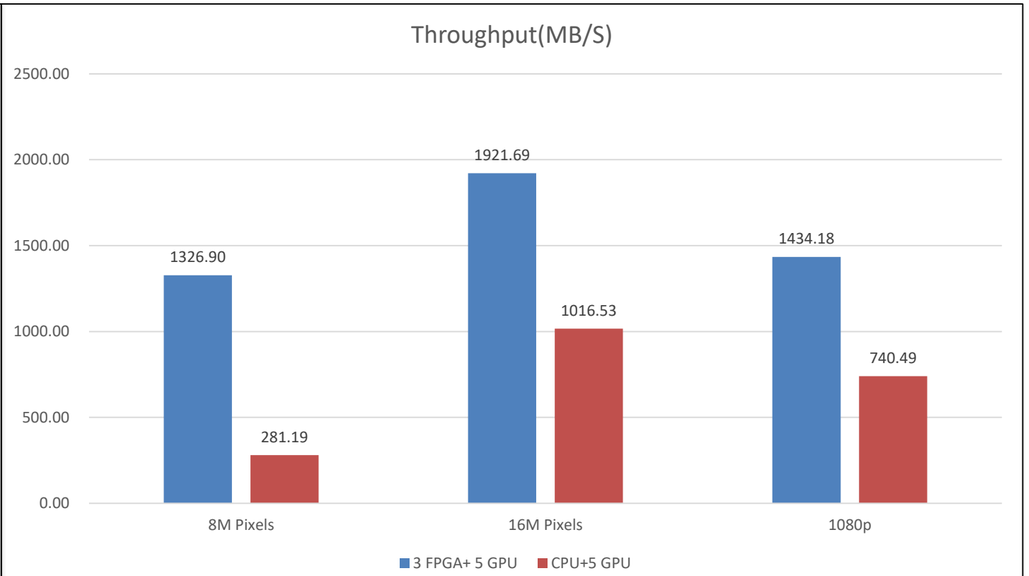
FPGA+GPU solution is **3** times of CPU+GPU solution

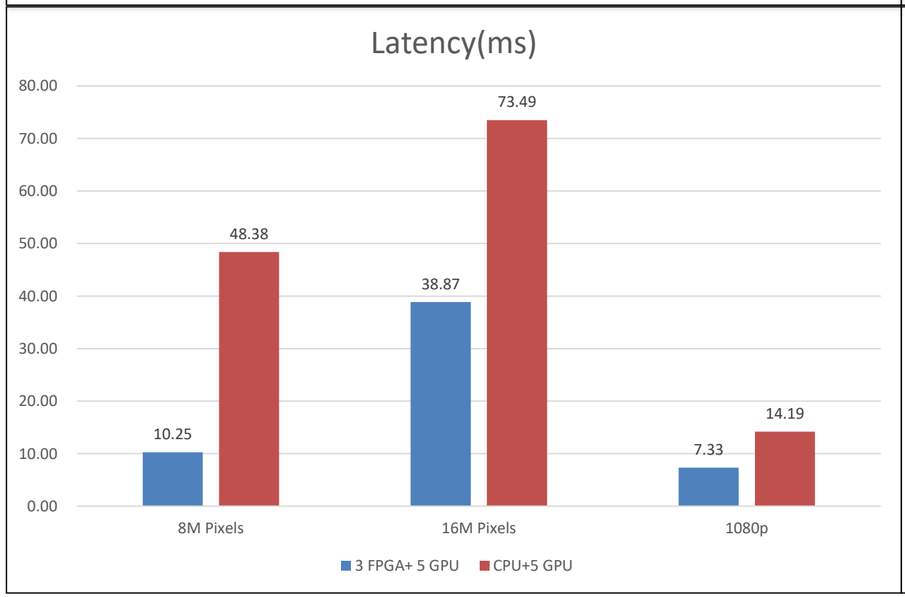
latency:

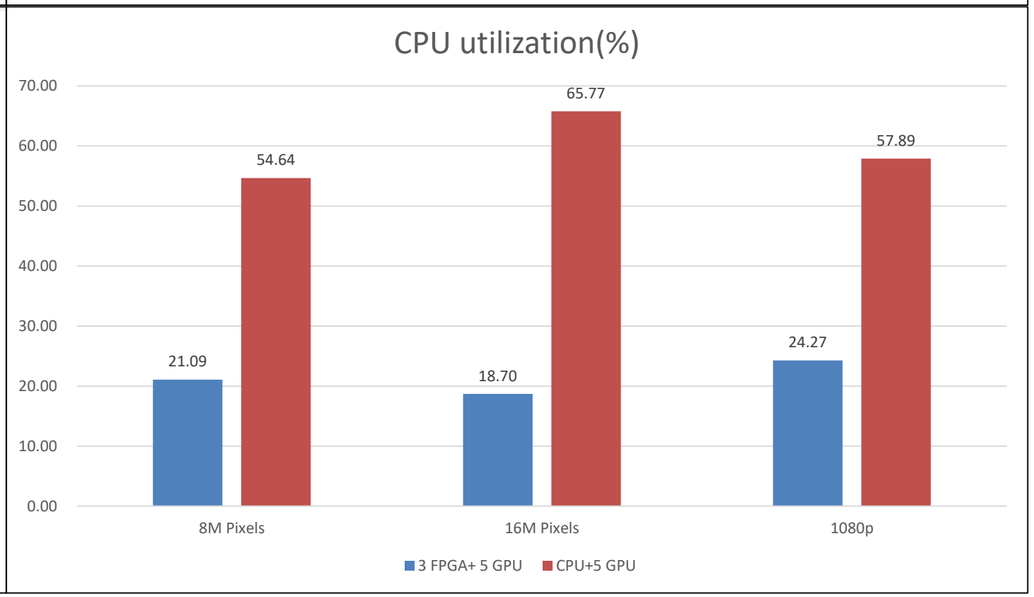
FPGA+GPU solution is **30%** of CPU+GPU solution

CPU Usage rate:

FPGA+GPU solution is **20%** of CPU+GPU solution







Accuracy Test Result

Model	Category	Accuracy(top1)	Accuracy(top5)
Alexnet	tensorflow	0.49	0.74
	CTAccel	0.49	0.73
inceptionv4	Tensorflow	0.80	0.95
	CTAccel	0.79	0.95
ResNet50	Tensorflow	0.73	0.91
	CTAccel	0.72	0.90
nsfw	Tensorflow	0.75	
	CTAccel	0.75	

Thanks!

CTAccel Limited

www.ct-accel.com

Telephone：+86-0755-88914045

E-mail:info@ct-accel.com

Zip code：518000

Address：Rm 2706,Golden Central Tower, No 3037 Jintian Rd, Futian Dist,Shenzhen ,China