XDF XILINX DEVELOPER FORUM

API for Real time/Dynamic Image Processing

Presented By

Ivan Wong
VP and Co-founder
Oct 16th 2018

XILINX

CTAccel & Xilinx partnership

CTAccel Limited

- Founded in Mar, 2016
- Based in Hong Kong & Shenzhen
- Focus on FPGA-based accelerated computing for Data Center
- US Patent
- Core value of our company: Quality & Innovation

- Partner with Xilinx Inc since 2016
- Delivered first Xilinx FPGA based Off-Premise solution in 2016
- Delivered first Xilinx FPGA based Cloud solution in 2017

Market demand and challenges

More Data

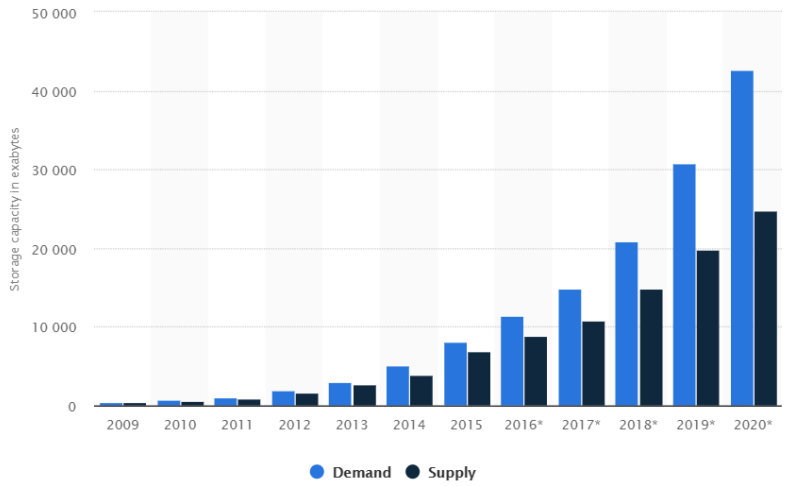
- Users generate more image and video data everyday
- Higher resolution in capture and display

Better Quality

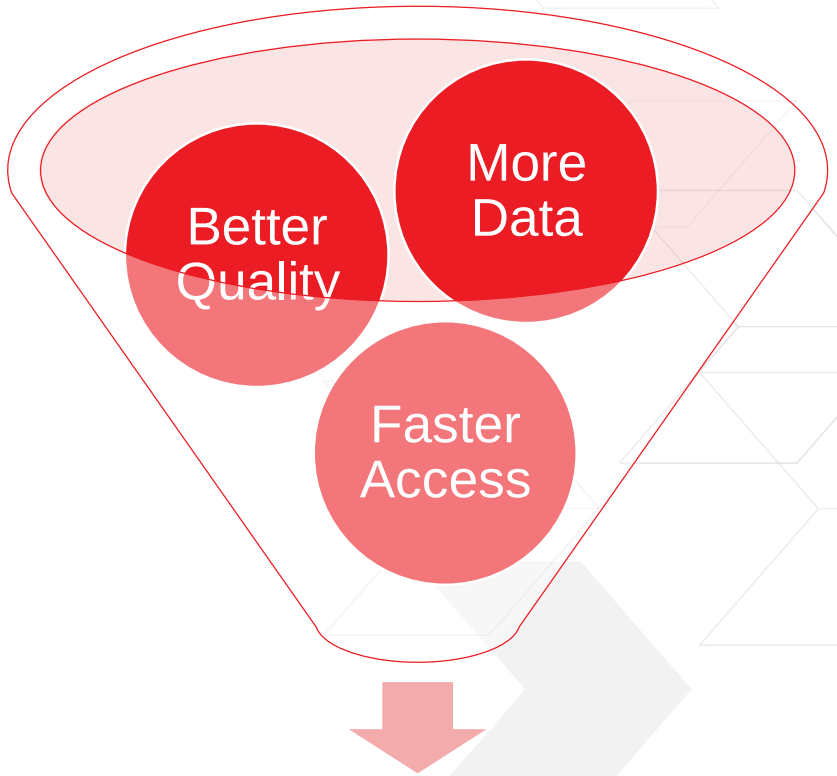
- Users crave better viewing experience

Faster Access

- Customers demand instant access to the resource



Data storage supply and demand worldwide, from 2009 to 2020 (in exabytes)*



Challenges:

- Huge consumption of computational and storage resource
- Server and storage performance IS NOT KEEPING PACE

*Source: https://www.statista.com/statistics/751749/worldwide-data-storage-capacity-and-demand/

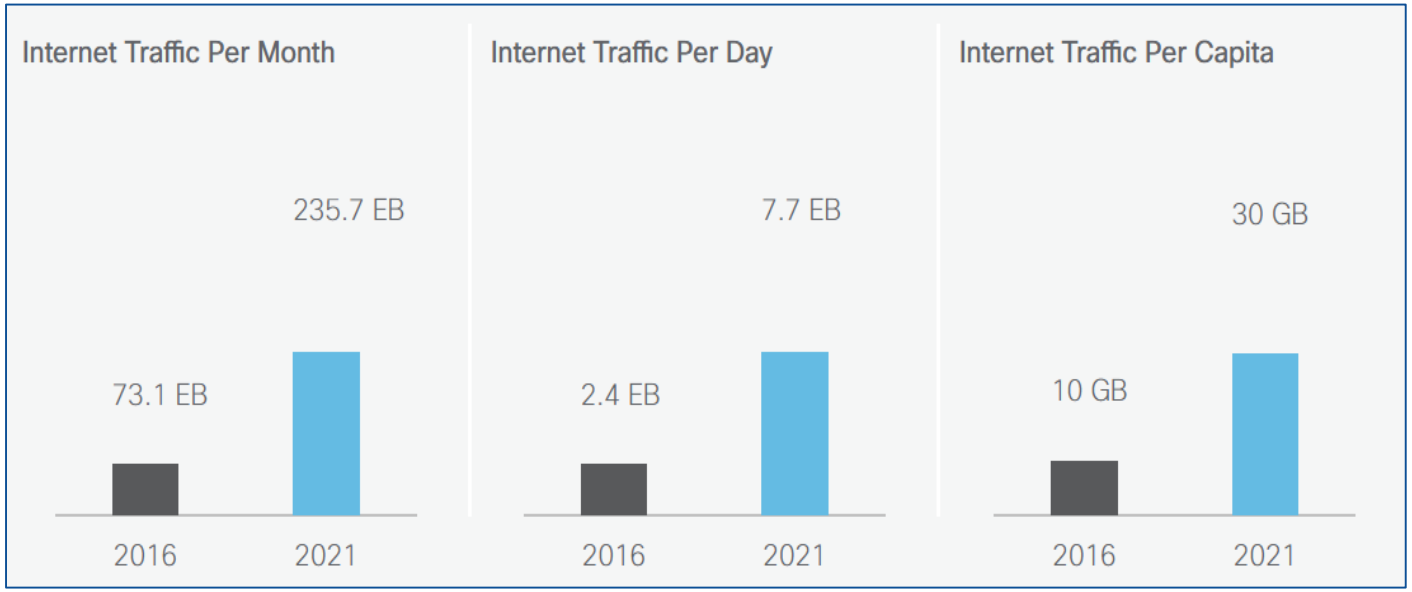
Constraints in existing solutions

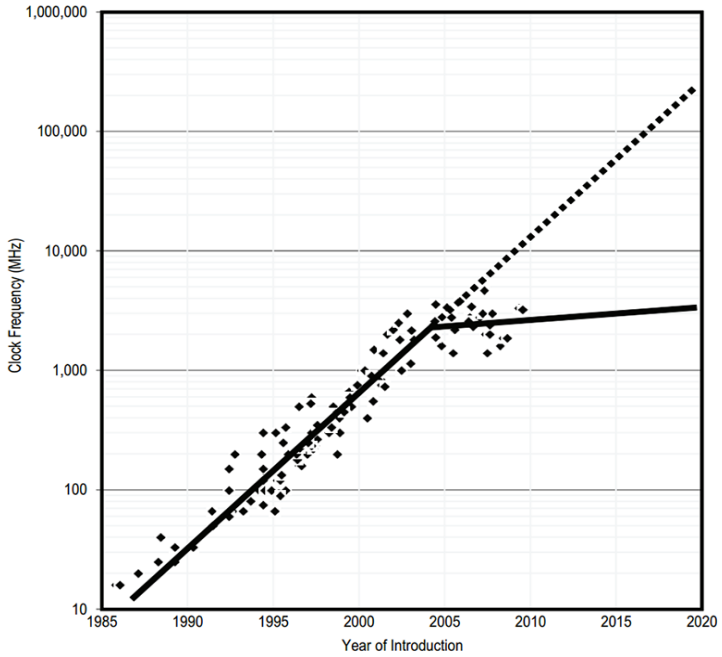
Internet traffic increases by 26% annually, image is a large portion of internet data.

Source: Cisco--VNI Forecast Highlights Tool

CPU performance per core has not increased since 2005

Source: The Future of Computing Performance (2011)

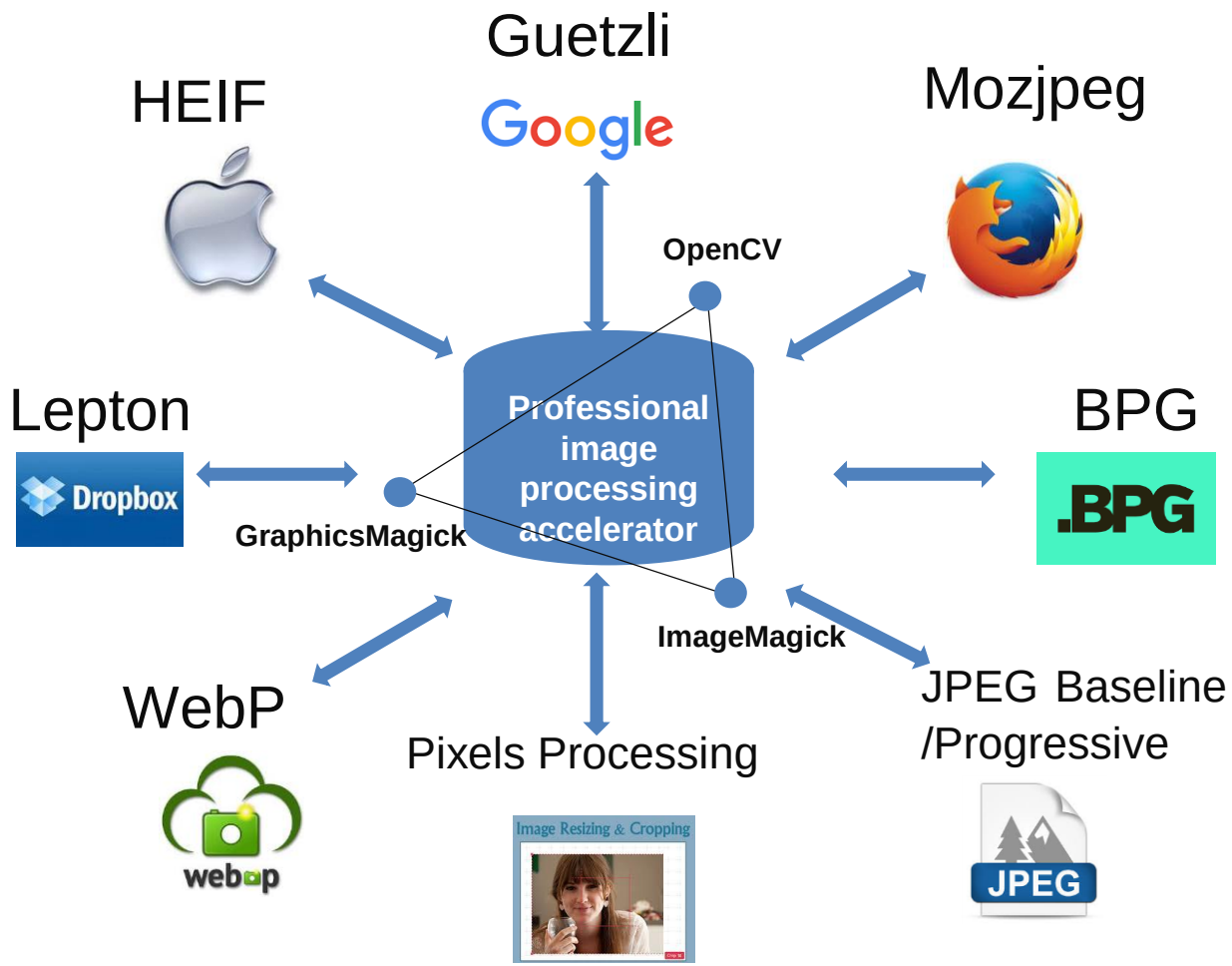




Chip Frequency graphed against year of introduction

How to achieve the fastest image processing with the least amount of computational resources?

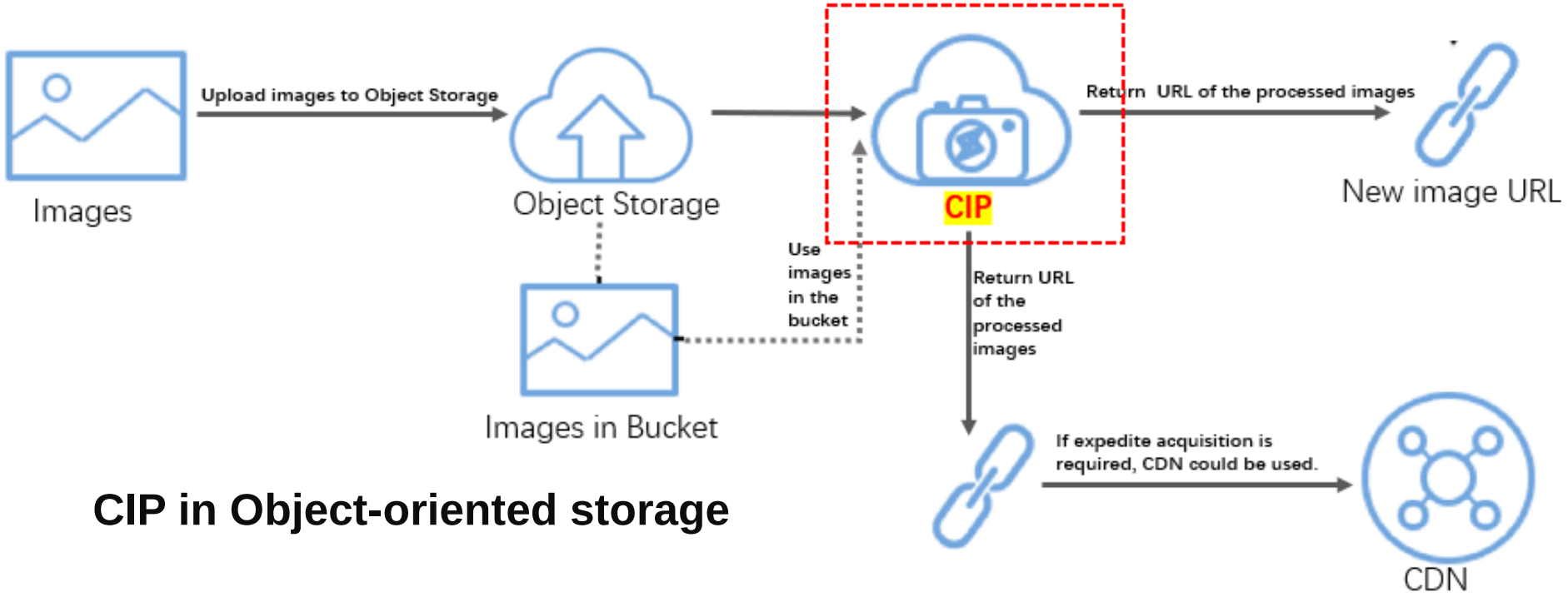
Modern image formats continue to emerge



Where is CTAccel image processor located in customers' production environment?

Scenarios

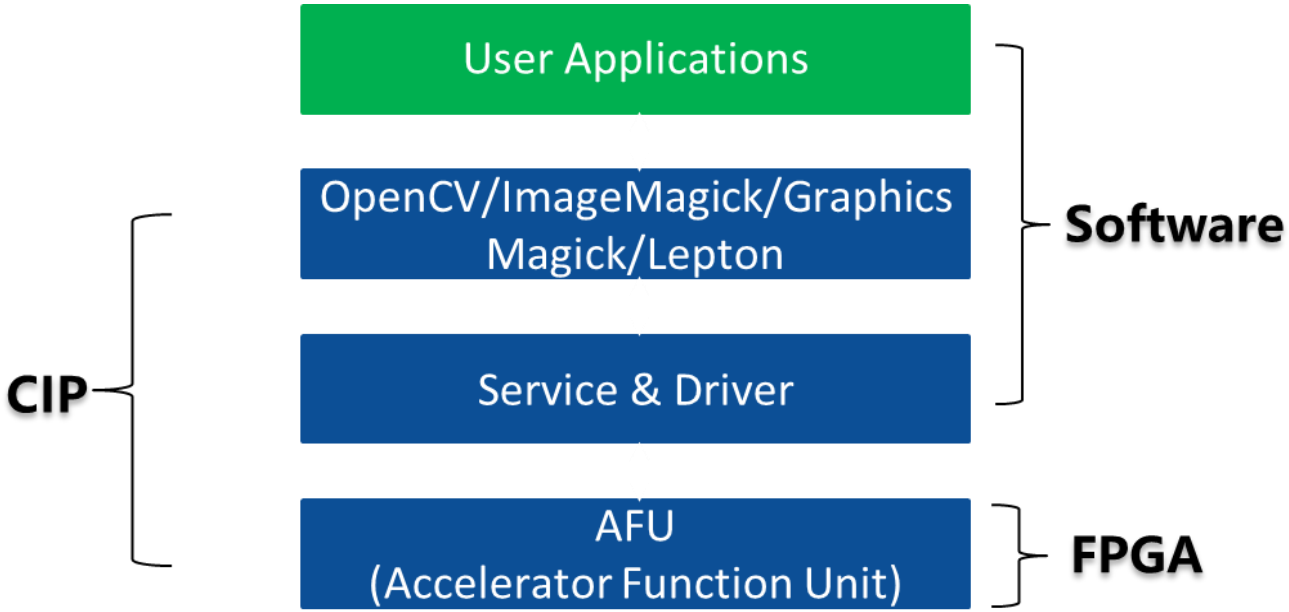
Cloud album	Social network	E-commerce	News app
UGC	Web portals	Cloud storage	……



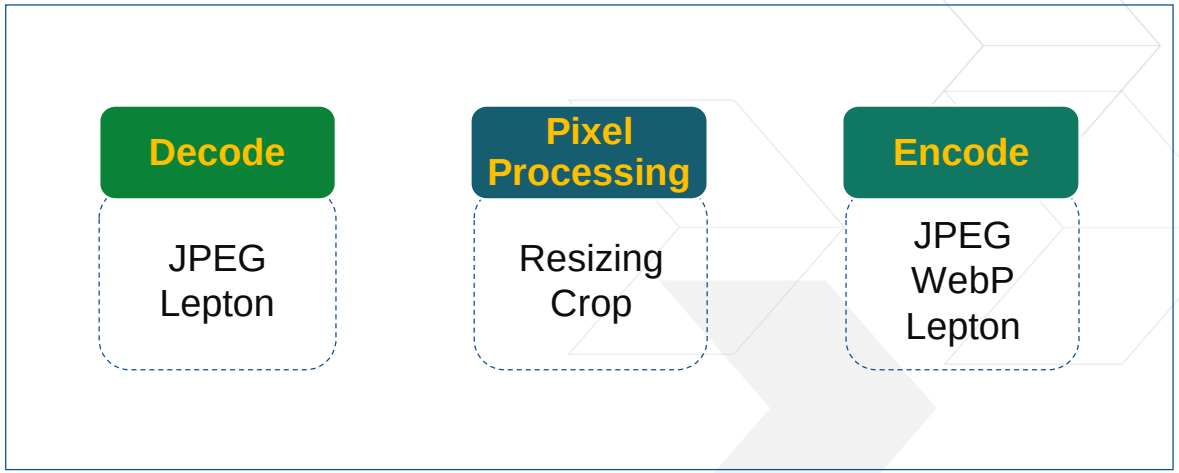
CIP in Object-oriented storage

CIP software stack and accelerated functions

System Stack



Accelerated Function Units



Why Xilinx?

Applications

Genomics

Security

Machine Learning

Video/Image Processing

Big Data Analytics

Financial Analytics

Compute Acceleration

Machine Learning

Video & Image Processing

Data Analytics

Genomics

Increase Service Level

Software-Defined Development Environments

Solution Development Partners & Customers

Cloud, Enterprise to Edge-Computing

Product 1：JPEG 2 JPEG

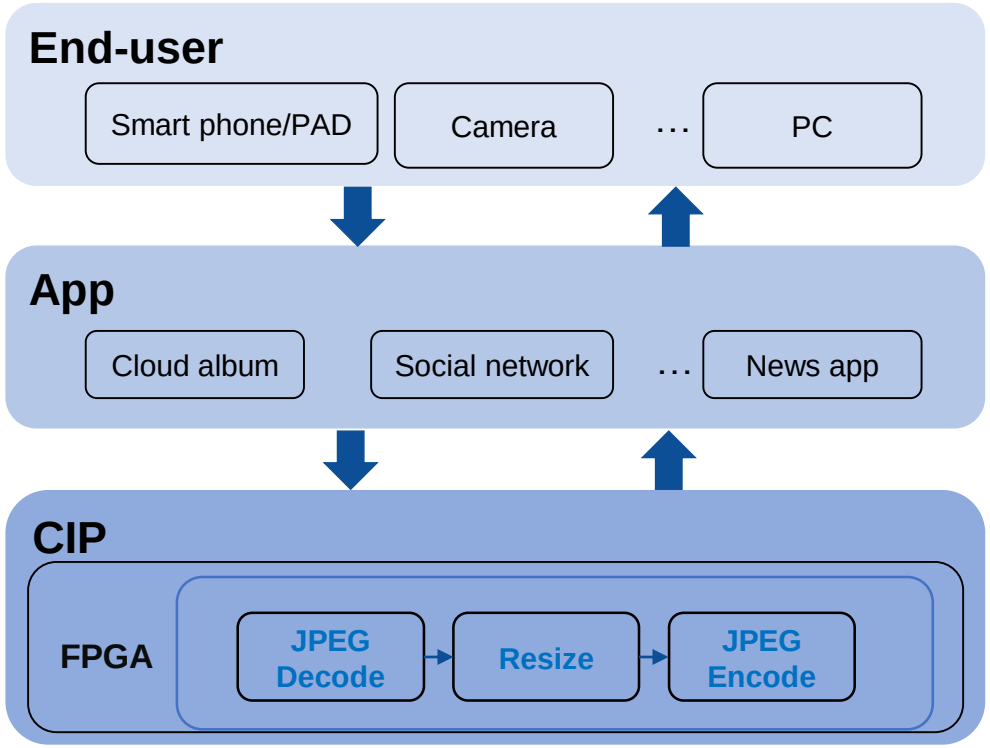
Problems solved

- Accelerate the speed of JPEG image thumbnail generation, reduce image-processing servers.

Key features

- **High Throughput**
- **Low Latency**
- **Low CPU Utilization**

Solution



Scenarios

Internet apps with JPEG

- E-commerce platform
- Cloud album
- Social network with pictures
- Web portals
- News application

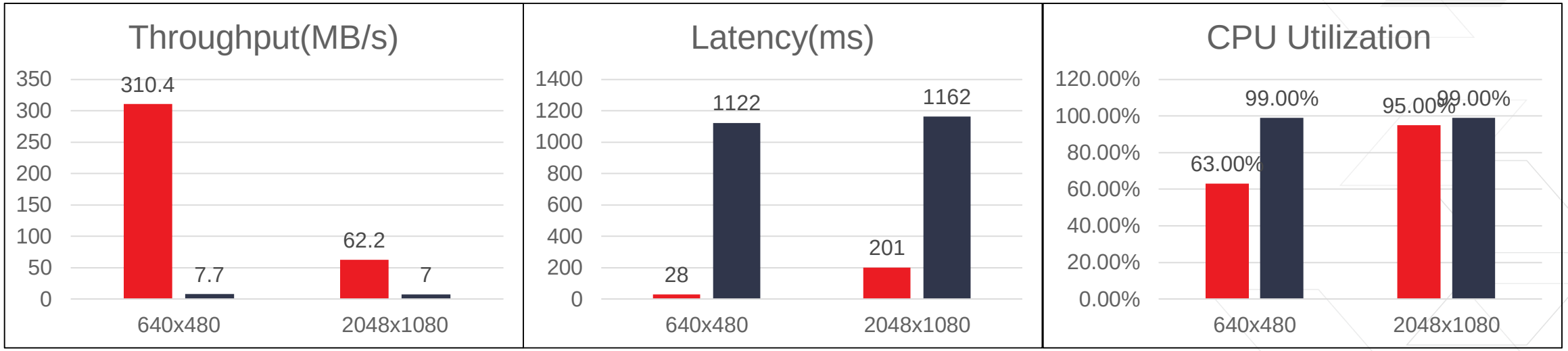
Values

- **TCO reduction：**
  - Reduce image-processing servers
  - OPEX reduction
- **Improve customer experience：**
  - Accelerate the image rendering speed

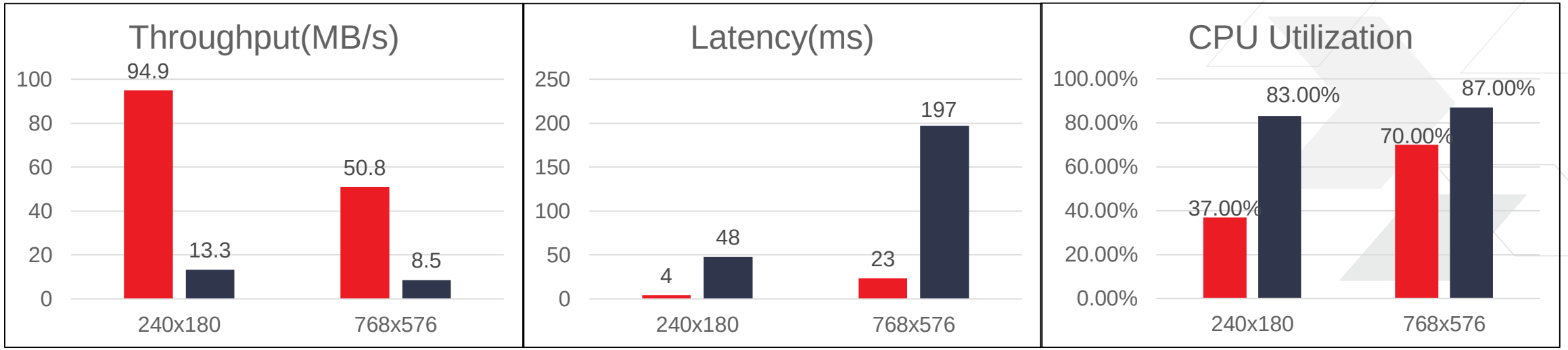
JPEG 2 JPEG performance on cloud

VU9P (HUAWEI fp1c.2xlarge)

8vCPU (HUAWEI fp1c.2xlarge)



Input : 10000*4096x2160 (Average Size 803k)



Input : 10000*1024x768 (Average Size 130k)

Product 2: JPEG 2 WebP

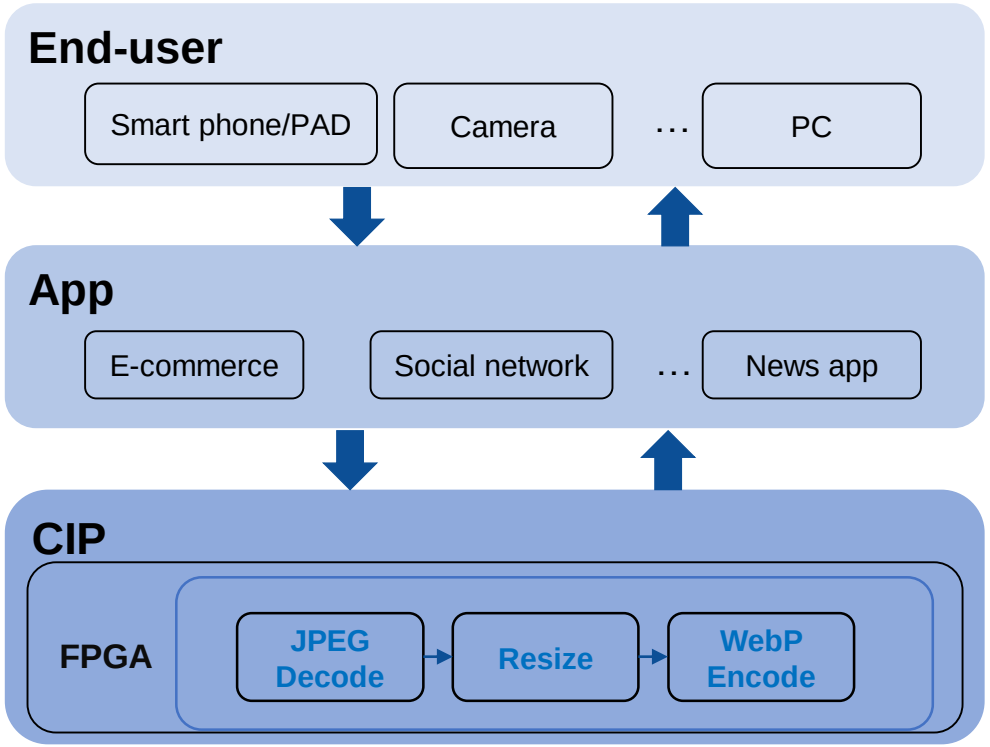
Problems solved

- **Vast amount of computational resources and long time** when processing WebP encoding by CPU;
- Accelerate the transcoding from JPEG to WebP ;

Key features

- **High Throughput**
- **Low Latency**
- **Low CPU Utilization**

Solution



Scenarios

Internet apps with WebP

- E-commerce platform
- Social network with pictures
- Web portals
- News application

Values

- **TCO reduction :**
  - Save CDN traffic
  - Reduce image-processing servers
  - OPEX reduction
- **Improve customer experience :**
  - Accelerate the image rendering speed

JPEG to WebP performance

FPGA Used:

- Single Xilinx Kintex UltraScale+ KU115
- Dual Xilinx Kintex UltraScale+ KU115

➢ QPS

Single FPGA is **5.11** times of CPU
Dual FPGA is **8.39** times of CPU

➢ Latency

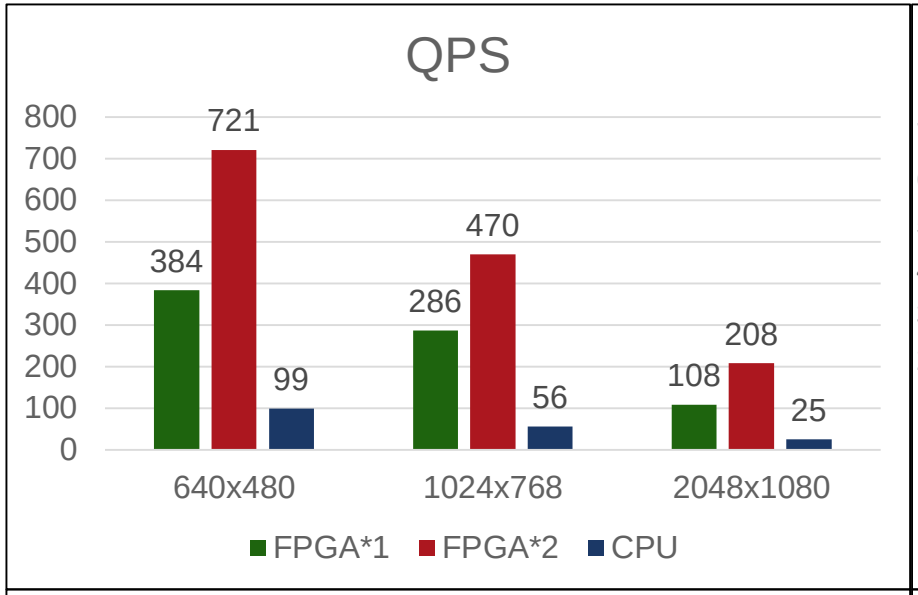
Single FPGA is **0.20** times of CPU
Dual FPGA is **0.12** times of CPU

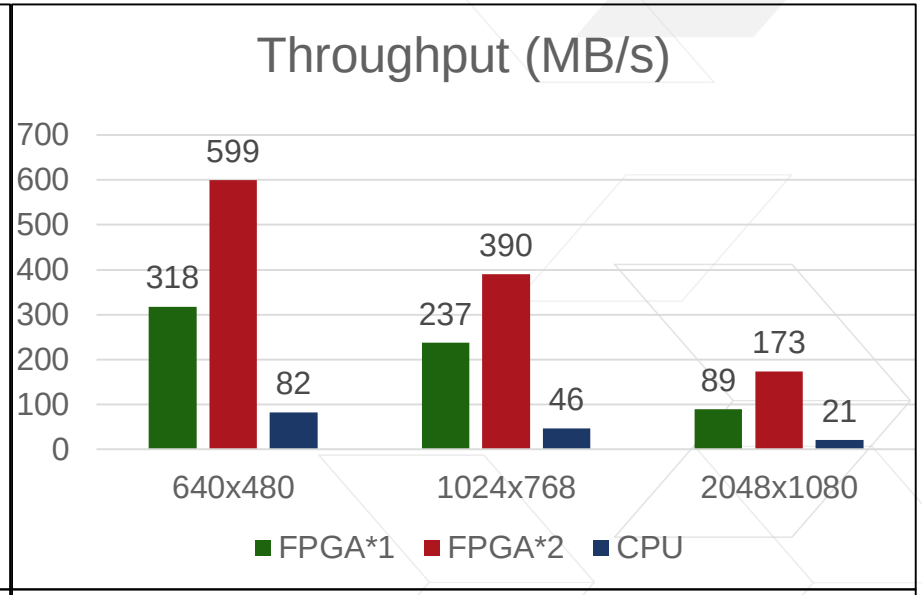
➢ Input:

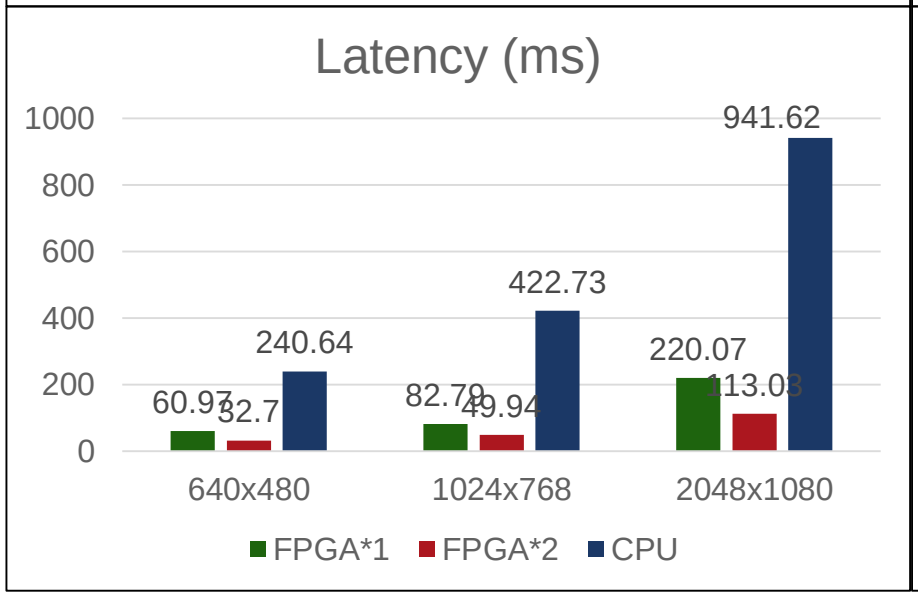
10000*4096 x2160 images

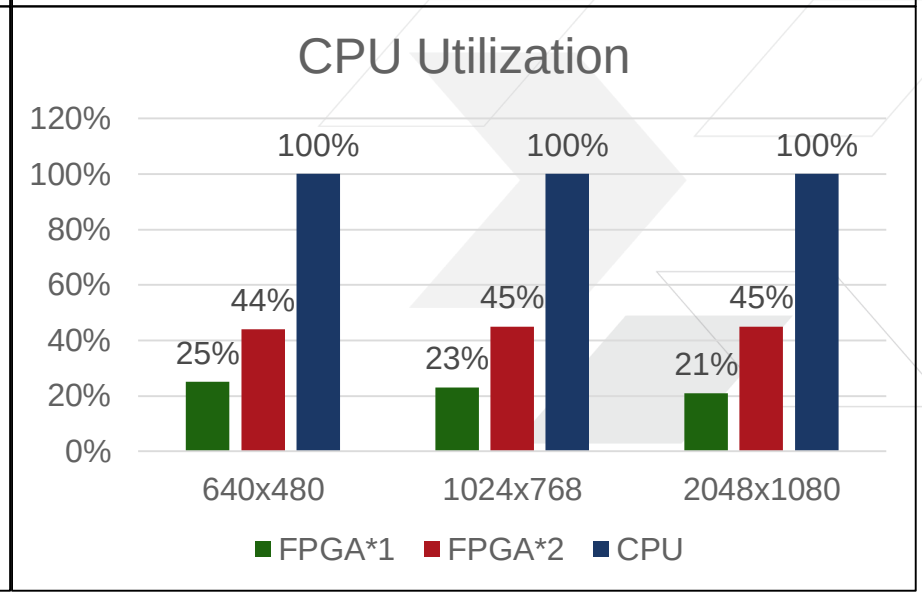
➢ Test environment:

CPU: 2*Intel(R) Xeon(R) CPU E5-2630v2
RAM: 128GB
OS: CentOS Linux release 7.2.1511



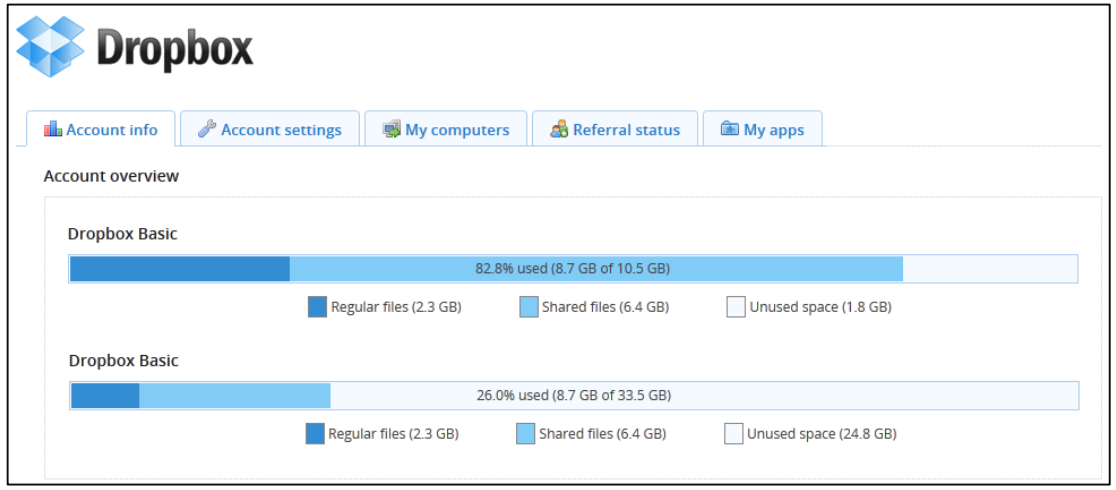






Modern image format-Lepton and its problem

- Lepton is a new image compression format developed by Dropbox and made open source in 2016.
- Lepton compresses JPEG images losslessly. It reduces file size by an average of 22%.
- It preserves the original JPEG file bit-by-bite perfectly, including all metadata.
- Lepton can be applied into the **scenarios with massive images storage**. It can **effectively reduce the storage cost**.



	JPEG	Lepton	Compression Ratio
1024x768 420	1.3G	1.1G	15%
1024x768 422	1.4G	1.1G	21%
1024x768 444	1.6G	1.3G	18%
4096x2160 420	8.3G	6.3G	24%
4096x2160 422	8.7G	6.6G	24%
4096x2160 444	9.5G	7.1G	25%
	Average		**21%**

Problem The downside of Lepton is that it requires heavy computation power for both compression and decompression. A conventional X86 server with dual E5-2630 CPU can only compress JPEG files into Lepton format at a rate of 20 megabytes per second.

Product 3: JPEG to Lepton

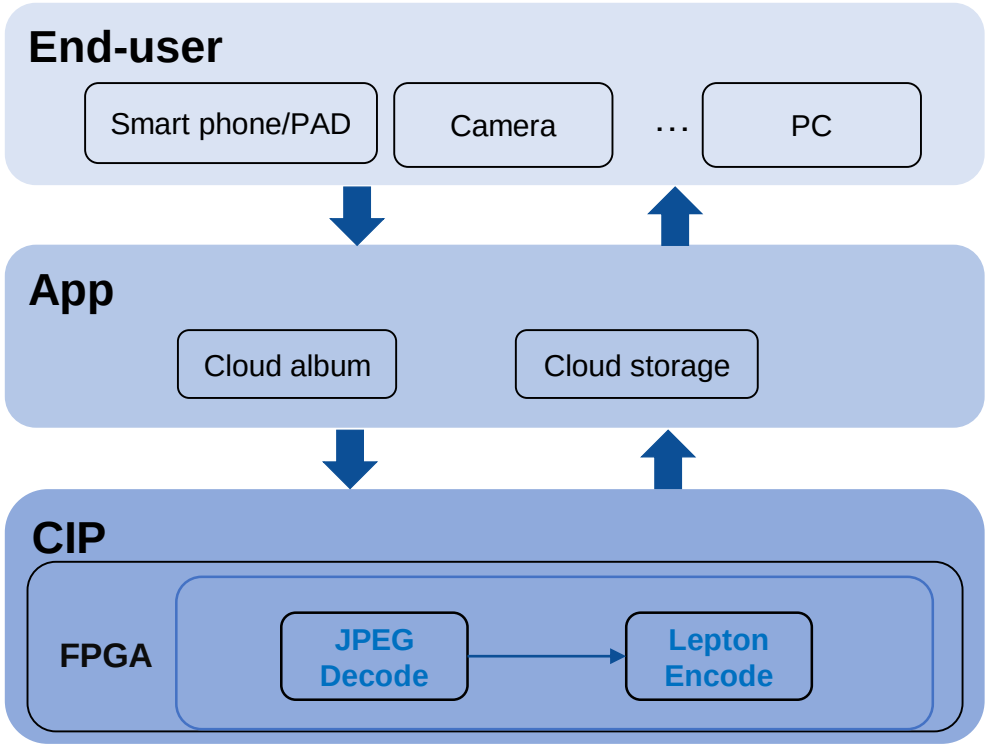
Problems solved

- **Vast amount of computational resources and long time** when processing Lepton encoding by CPU;
- Accelerate the transcoding from JPEG to Lepton ;

Key features

- **High Throughput**
- **Low Latency**
- **Low CPU Utilization**

Solution



Scenarios

Scenarios with massive images storage ：

- Cloud album ;
- Cloud storage ;

Values

- **TCO Reduction ：**
  - Save CDN traffic
  - Reduce image-processing servers
  - Reduce image-storage servers
  - OPEX reduction

JPEG to Lepton performance

FPGA Used:

- Virtex UltraScale+ FPGA VCU1525

- **QPS**

FPGA is **4.2** times of CPU

- **Latency**

FPGA is **0.24** times of CPU

- **Input images:**

Total number: 999 images

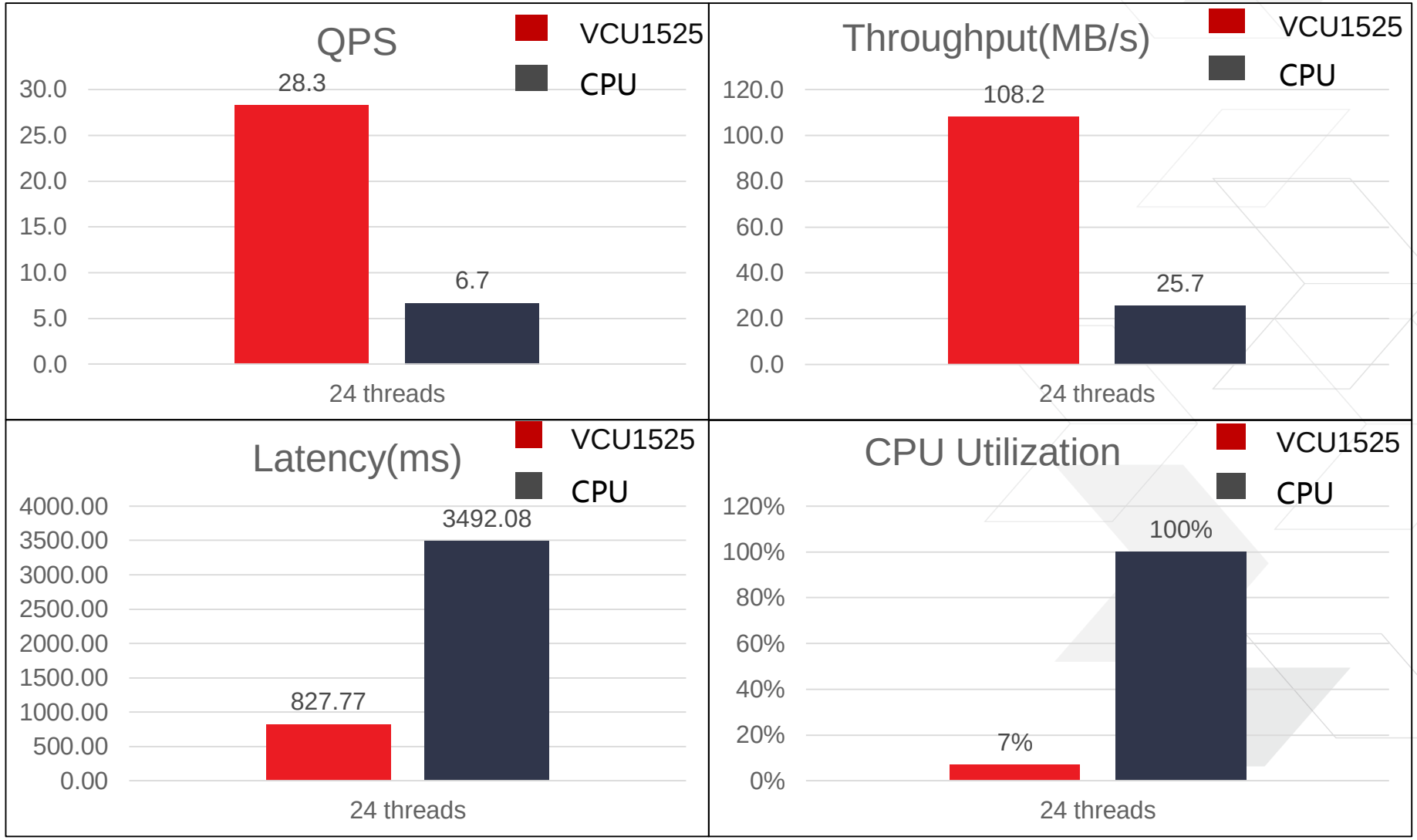
Total file size: 3.8GB

- **Test environment ：**

CPU: Intel(R) Xeon(R) E5-2630v2 x2

RAM: 128GB

OS: CentOS Linux release 7.3.1611



Product 4: HEIC to JPEG

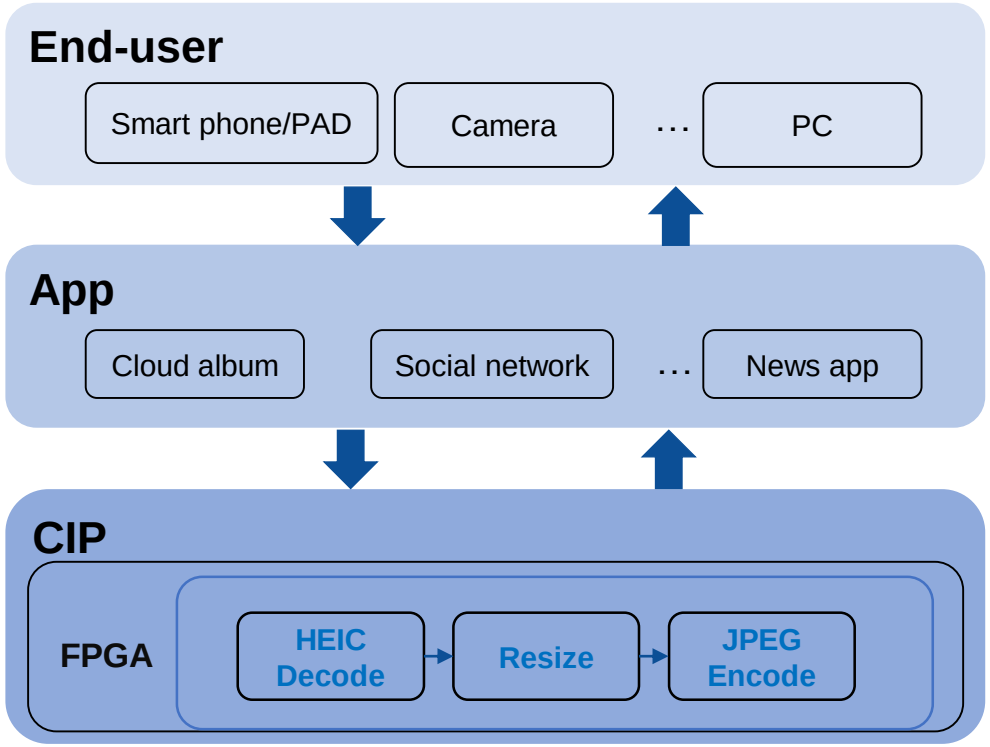
Problems solved

- **Vast amount of computational resources and long time** when processing HEIC decoding by CPU;
- Poor customer experience;

Key features

- **High Throughput**
- **Low Latency**
- **Low CPU Utilization**

Solution



Scenarios

Internet apps with HEIC

- E-commerce platform
- Social network with pictures
- Web portals
- News application

Values

- **TCO reduction :**
  - Reduce image-processing servers
  - OPEX reduction
- **Improve customer experience :**
  - Accelerate the image rendering speed

Summary of CIP's values

High performance

5-10 times throughput promotion compare to CPU

3 times latency reduction compare to CPU

Low power

20W-40W per FPGA card

TCO reduction: improve computing density of DC, reduce racks

Software compatible

Support: ImageMagick, OpenCV, GraphicsMagick, Lepton

Allow seamless migration from software-based implementation to CIP

Ease of maintenance

Remote upgrading

PR(Partial Reconfiguration) technology allows fast and easy context switch in accelerator functionality without rebooting the server

Case study 1

Customer

Leading mobile phone manufacturer in China

Scenario

Image thumbnail generation for cloud album

CIP

JPEG decode

Resize

Xilinx FPGA

Advantages

High performance

- Reduce the total number of server cluster by 50%
- Improve the throughput by 50%

TCO reduction

- Reduce TCO by 25%

Better QoS

- 3x latency reduction
- Deterministic timing

Ease of maintenance

- Smaller cluster of servers

Feedback from customer

- CIP has been deployed in the production environment for more than 16 months
- Maintaining stable operating status

Case study 2

Customer

Famous video portal website in China

Scenario

Transcoding from JPEG to WebP

CIP

JPEG decode

Resize

WebP encode

Advantages

High performance

- 1 server with one CIP accelerator has the equivalent computing capability with 3 servers without CIP

TCO reduction

- Reduce TCO by 50% at least

Improve customer experience

- Reduce latency by xx% (not announced the specific values by our customer)
- Deterministic timing: the same low latency at full-load and no-load

Feedback from customer

- Test period: 2 months
- CIP has passed the test of customer
- CIP will be deployed in the production environment in 2018Q4

Evolution planning of CIP

Ready	Ready	2018Q4	2019
JPEG->Lepton	Lepton->JPEG	HEIC	Guetzli
WebP, M2	WebP, M6		Mozjpeg
Pixels			BPG
JPEG			JPEG-Progressive

Cloud platform support

amazon web services

Alibaba Cloud

HUAWEI

Tencent Cloud

百度云 cloud.baidu.com

Hardware platform support

			
Virtex UltraScale+ VCU1525	Kintex® UltraScale™115	Kintex® UltraScale™ 060	Zynq UltraScale+

New product line of CTAccel

Image

Video

AI

Higher performance density

Improve quality of service

TCO reduction

High-quality FPGA-based accelerated computing solution provider

Xilinx MPSoC with VCU - H264/H265

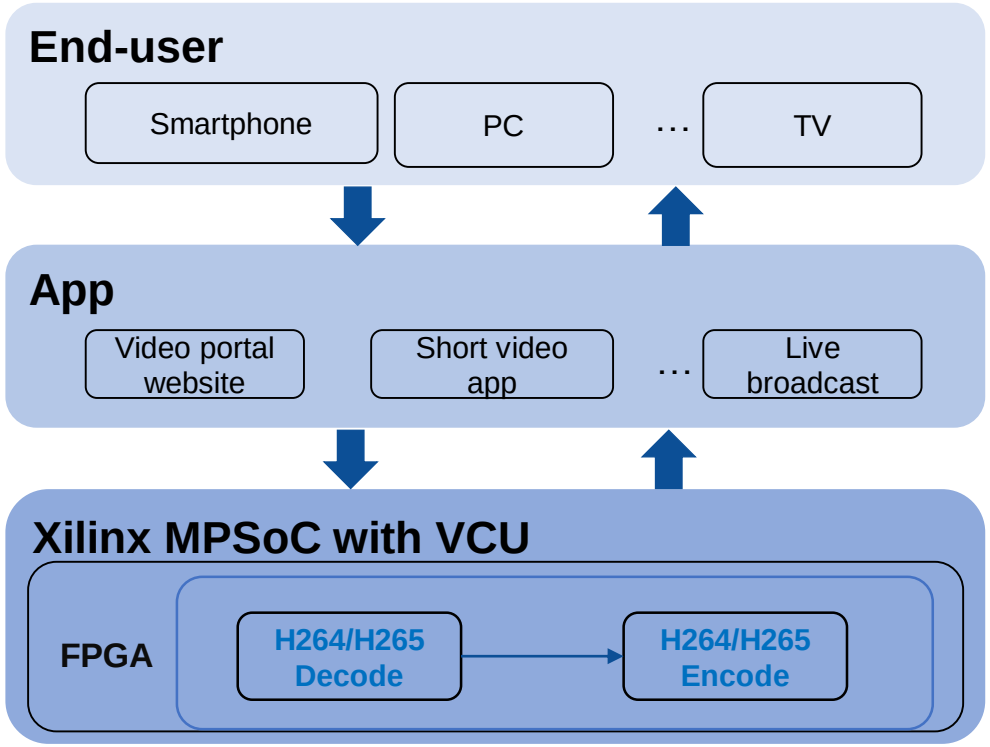
Problems solved

- Accelerate the speed of video processing, improve customer experience

Key features

- **FPS**
- **PSNR**

Solution



Scenarios

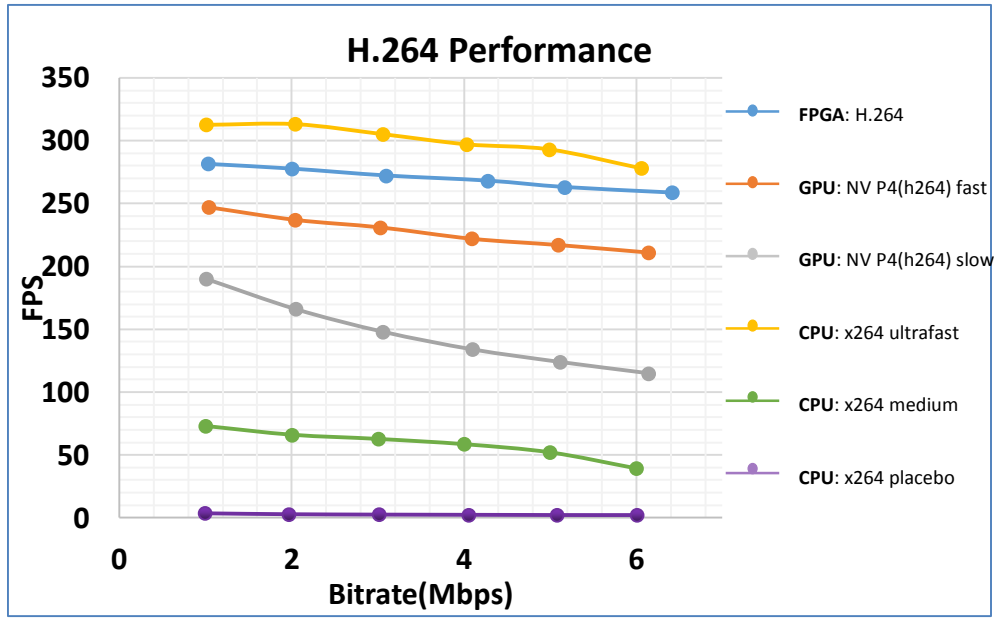
Applications which require video transcoding:

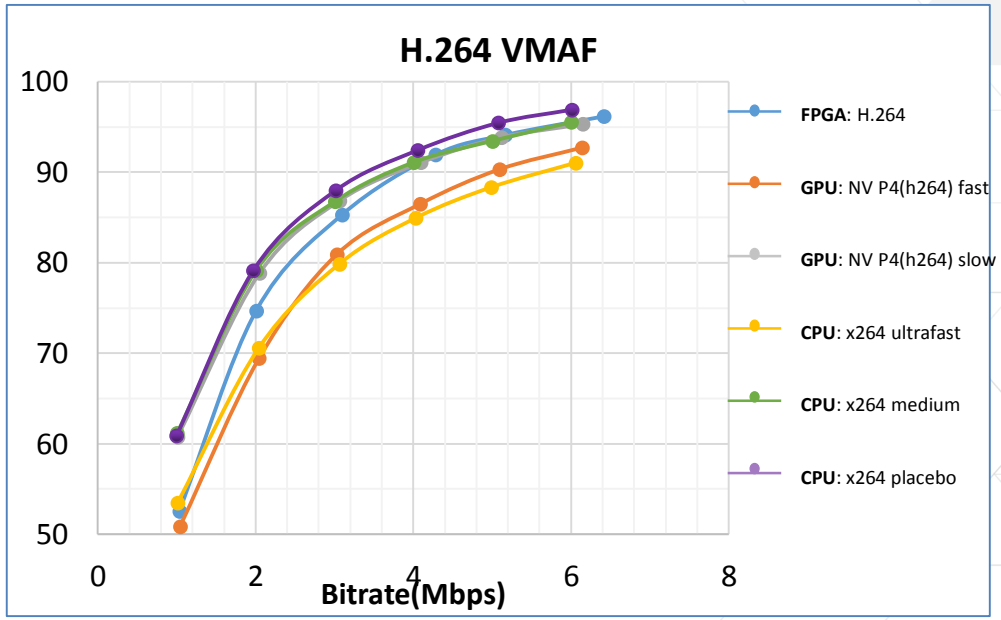
- Video portal websites ;
- Short video apps ;

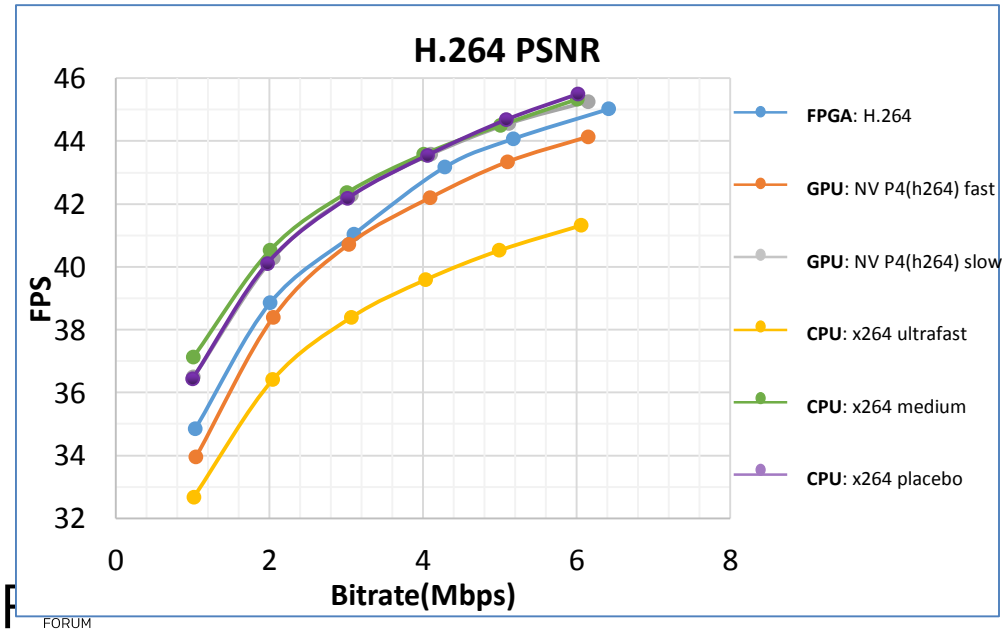
Values

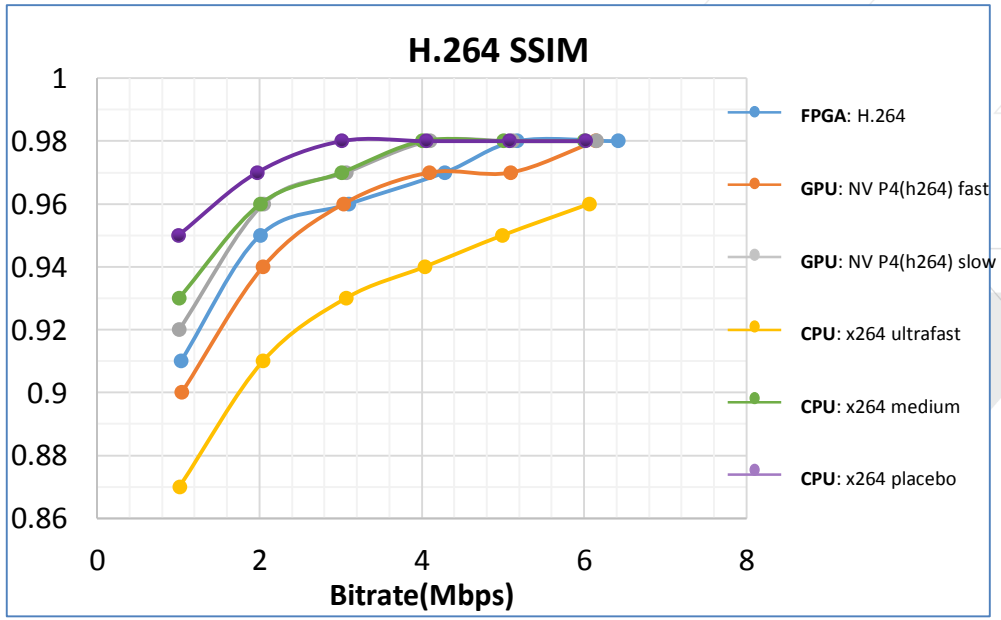
- **TCO Reduction :**
  - Reduce processing servers
  - OPEX reduction
- **Improve customer experience**

Performance & quality of H.264 (MPSoC)

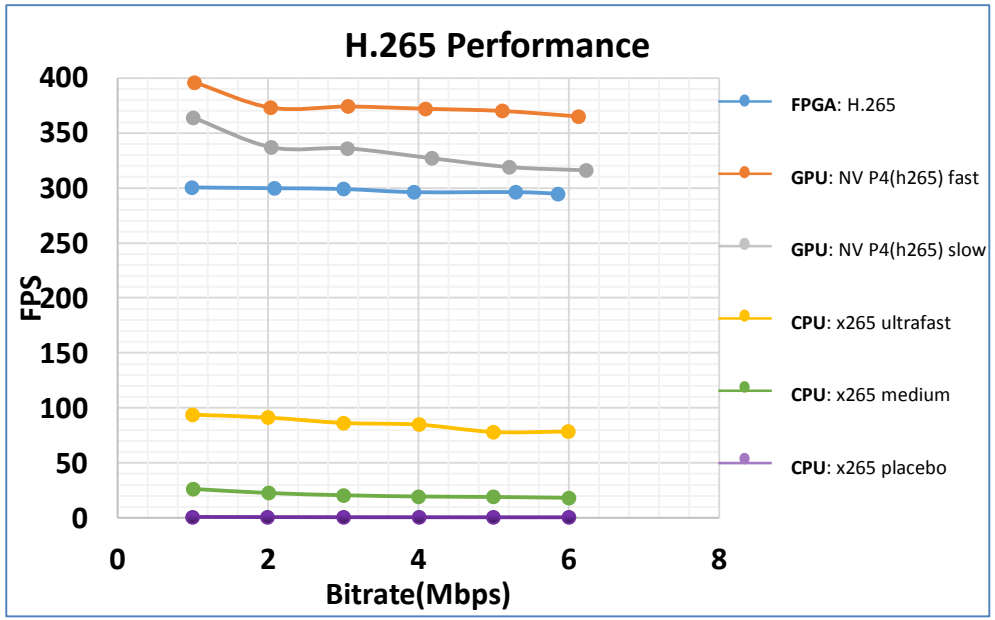


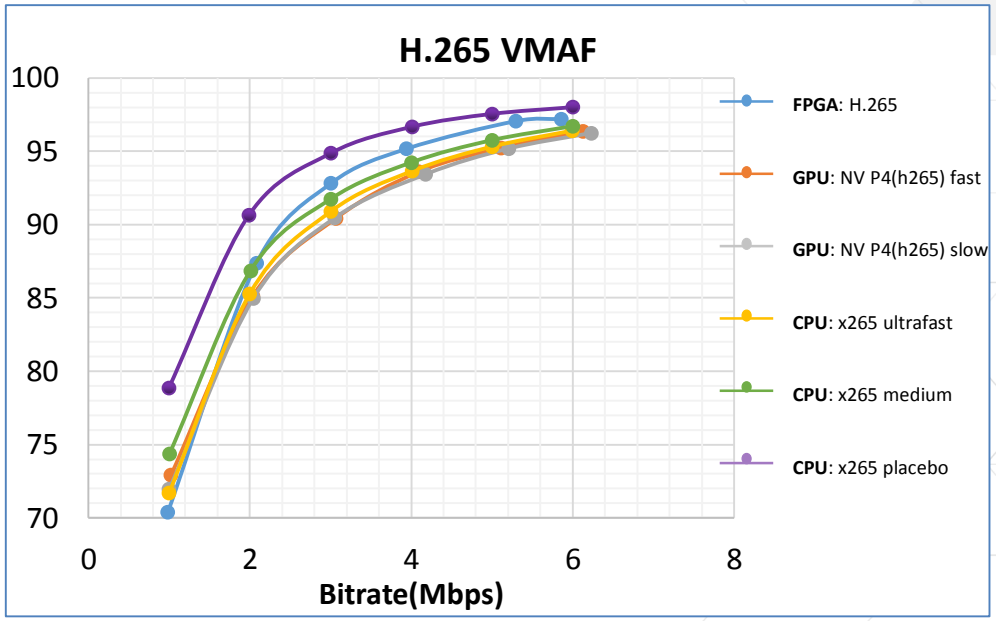


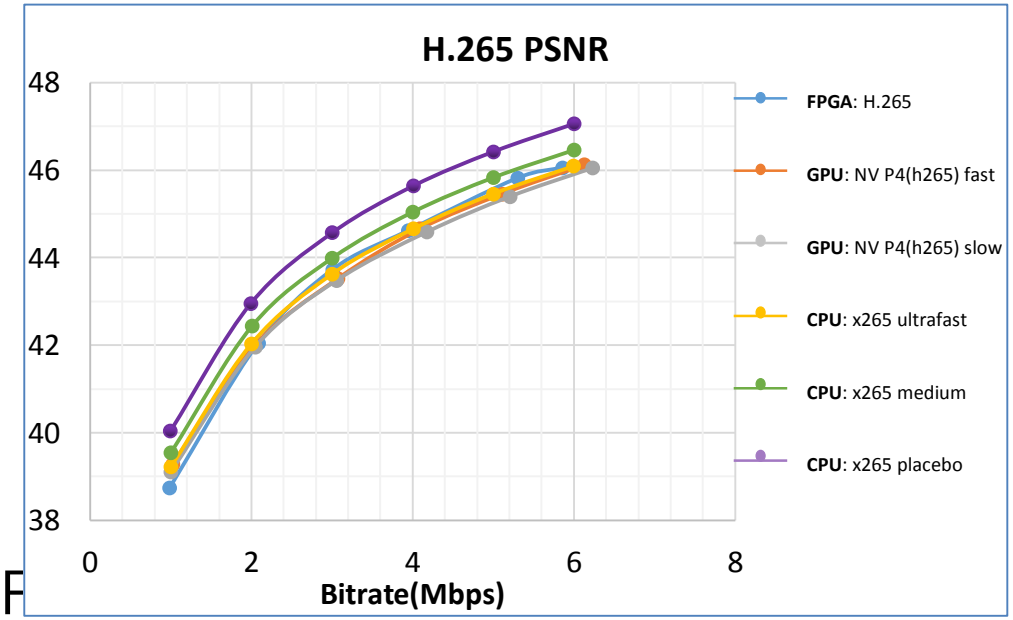


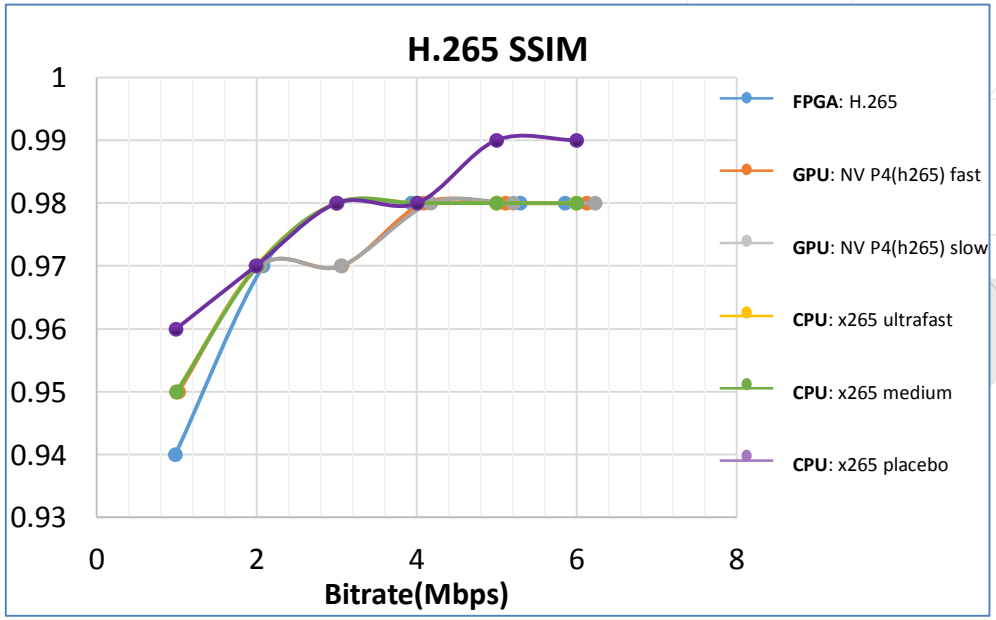


Performance & quality of H.265 (MPSoC)









Summary of H264/H265 encoding (MPSoC)

Codec	Item		Bitrate(Mbps)		
			Low (<2)	Medium(2-5)	High(>5)
H.264	FPGA vs GPU	Performance	Better than GPU		
		Quality	Better than NV P4 fast, worse than NV P4 slow		Better than NV P4 fast , similar to NV P4 slow
	FPGA vs CPU	Performance	Worse thanx264 ultrafast , better than x264 medium and x264 placebo		
		Quality	Better than x264 ultrafast Worse than x264 placebo Worse than x264 medium		Better than x264 ultrafast Worse than x264 placebo Similar to x264 medium
H.265	FPGA vs GPU	Performance	Worse than GPU		
		Quality	<1.6M : worse than GPU 1.6M~2M : better than GPU	Better than GPU	
	FPGA vs CPU	Performance	Better than CPU		
		Quality	<1.6M : worse than CPU 1.6M~2M : better than x265 ultrafast , worse than x265 placebo , worse than x265 medium	Worse than x265 placebo , better than x265 ultrafast and x265 medium	

I-Frame 2 JPEG

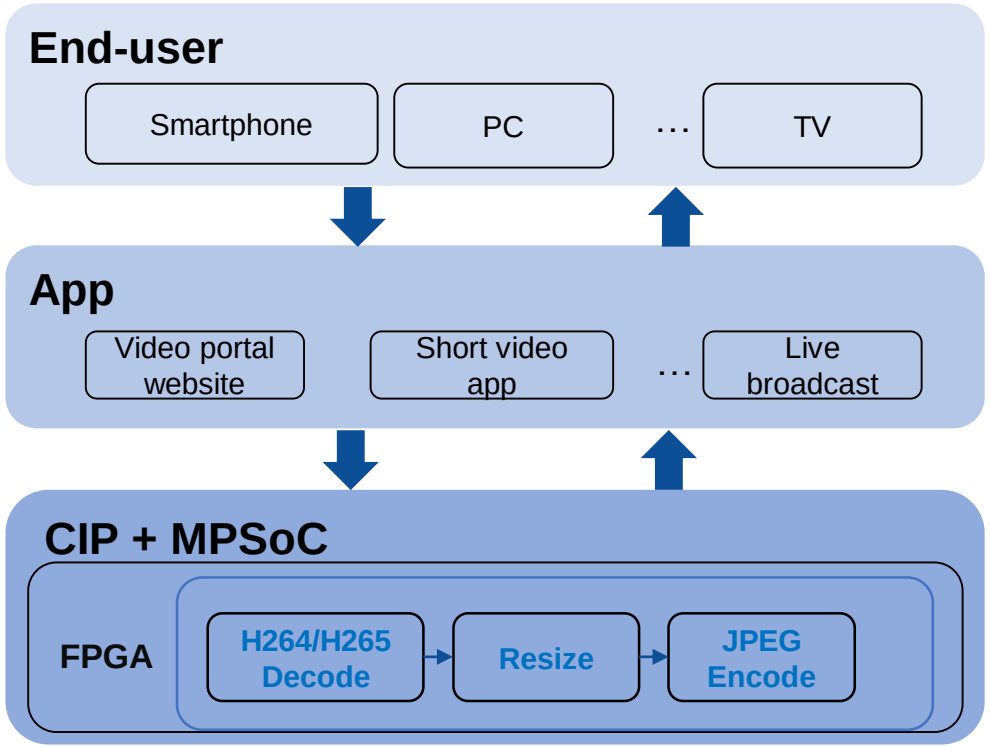
Problems solved

- Accelerate the speed of ”I-Frame 2 JPEG”, improve customer experience

Key features

- **High Throughput**
- **Low Latency**
- **Low CPU Utilization**

Solution



Scenarios

Applications which require “I-Frame 2 JPEG”:

- Video portal websites ;
- Short video apps ;

Values

- **TCO Reduction :**
  - Reduce processing servers
  - OPEX reduction
- **Improve customer experience**

AI accelerator: face detection/recognition

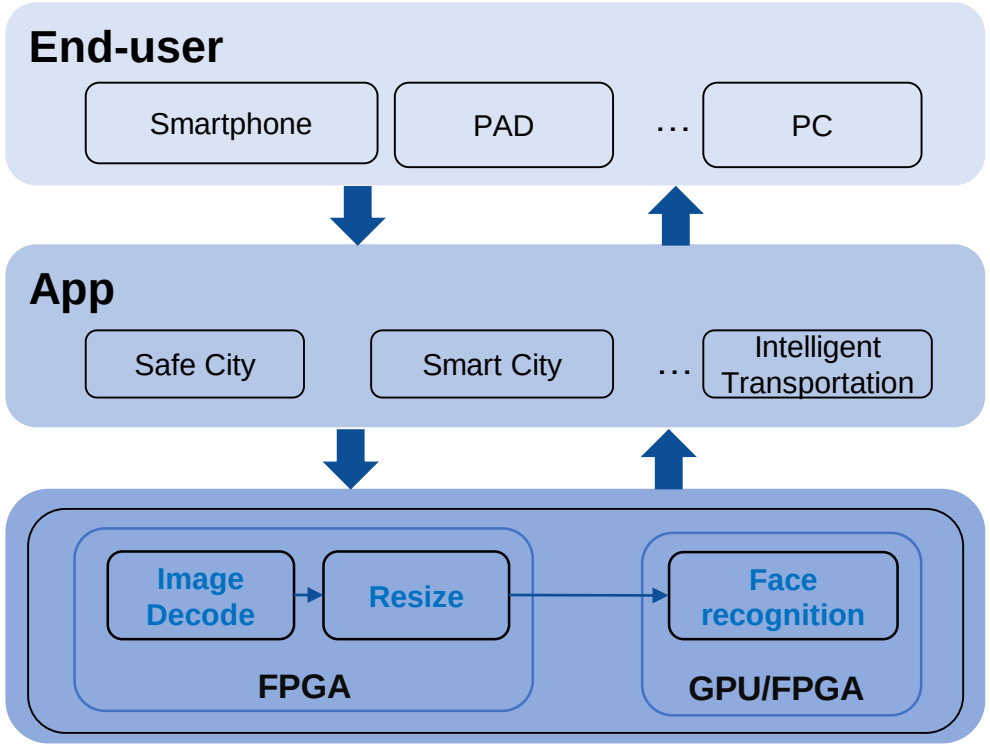
Problems solved

- Accelerate the speed of ”face detection/recognition”, improve customer experience, reduce TCO

Key features

- **High Throughput**
- **Low Latency**
- **Low CPU Utilization**

Solution



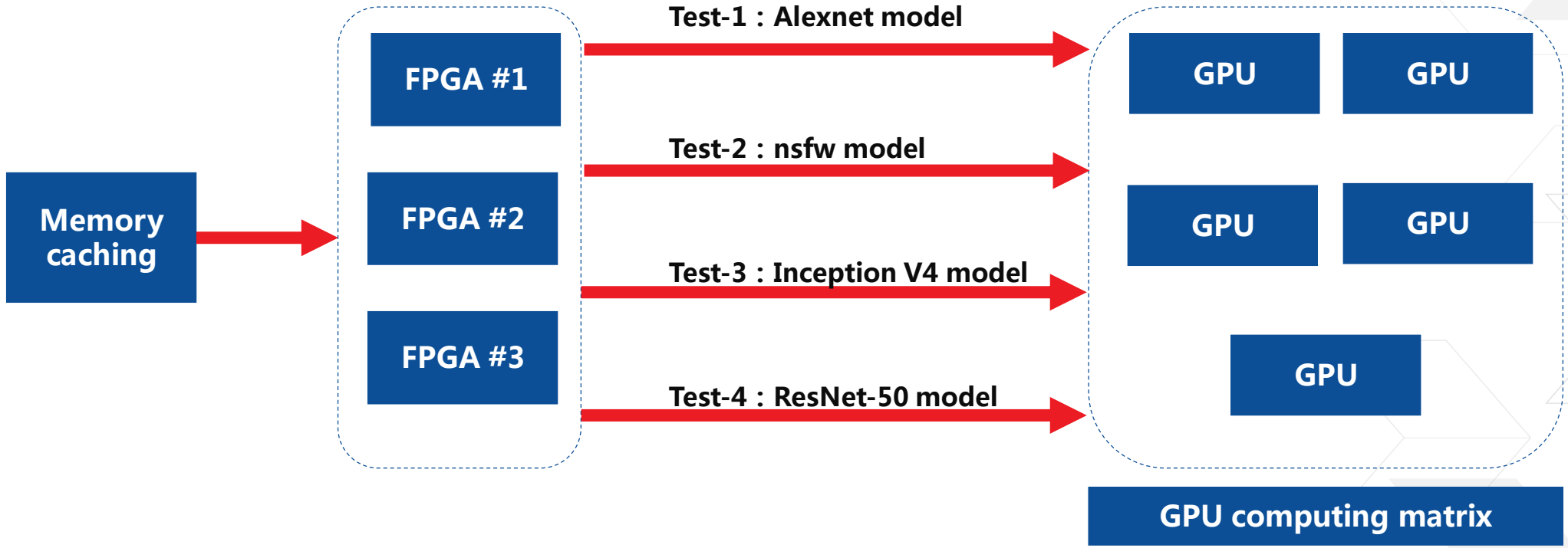
Scenarios

- Smart City ;
- Intelligent Transportation ;
- Safe City ;
- Security system ;

Values

- **TCO Reduction :**
  - Reduce image-processing servers
  - OPEX reduction
- **Improve customer experience**
  - Improve the speed of face detection/recognition

AI accelerator: test scenarios



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

Test data

Image type	Resolution	Quantity
8 Mega Pixel	2647X3278	14708
16 Mega Pixel	5312x2988	9280
1080p	1920x1080	21400

Test result 1-Alexnet

Alexnet model

QPS:

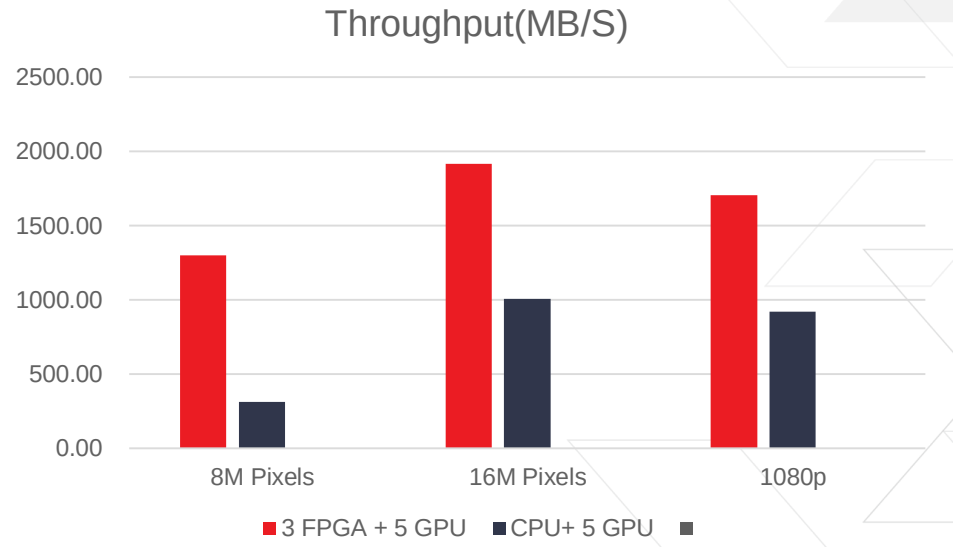
FPGA+GPU is **2.5** faster than CPU+GPU

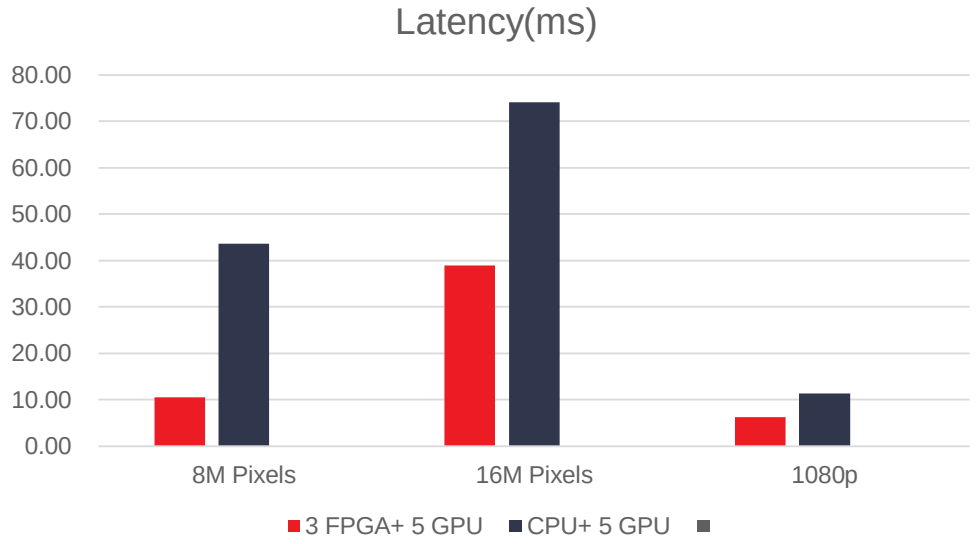
Latency :

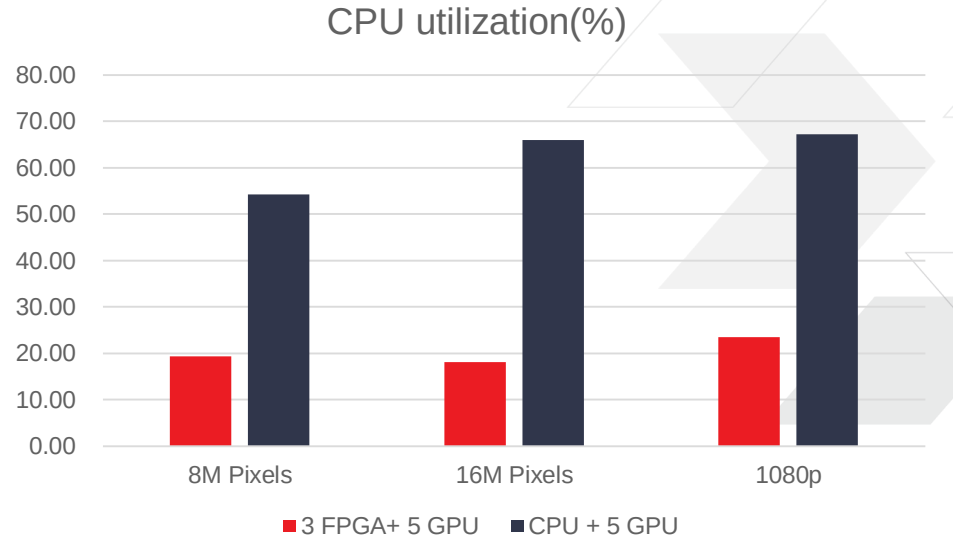
FPGA+GPU is **30%** of CPU+GPU

CPU usage :

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







Test result 2-nsfw

nsfw model

QPS:

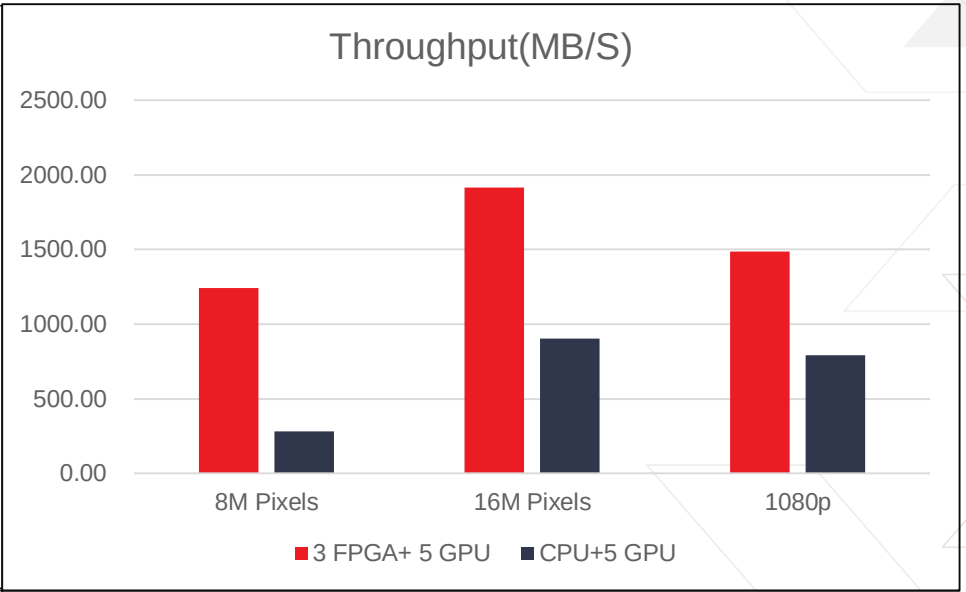
FPGA+GPU is **3** faster than CPU+GPU

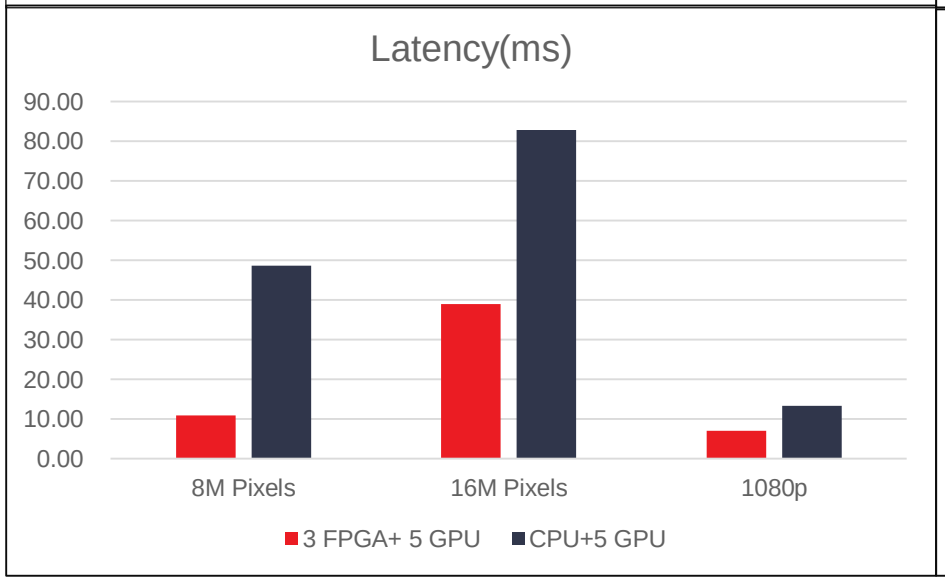
Latency :

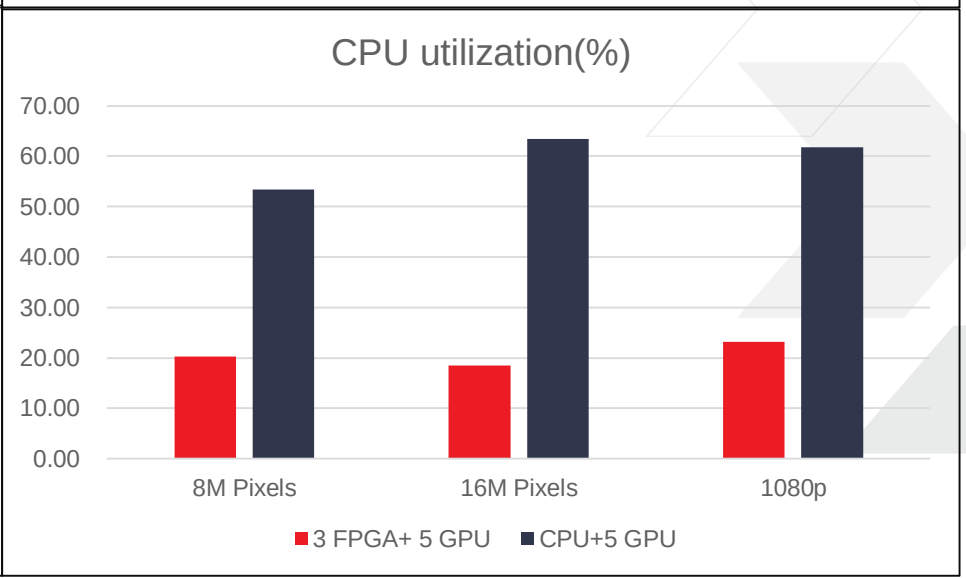
FPGA+GPU is **30%** of CPU+GPU

CPU usage :

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







Test result 3-Inception V4

InceptionV4 model

QPS:

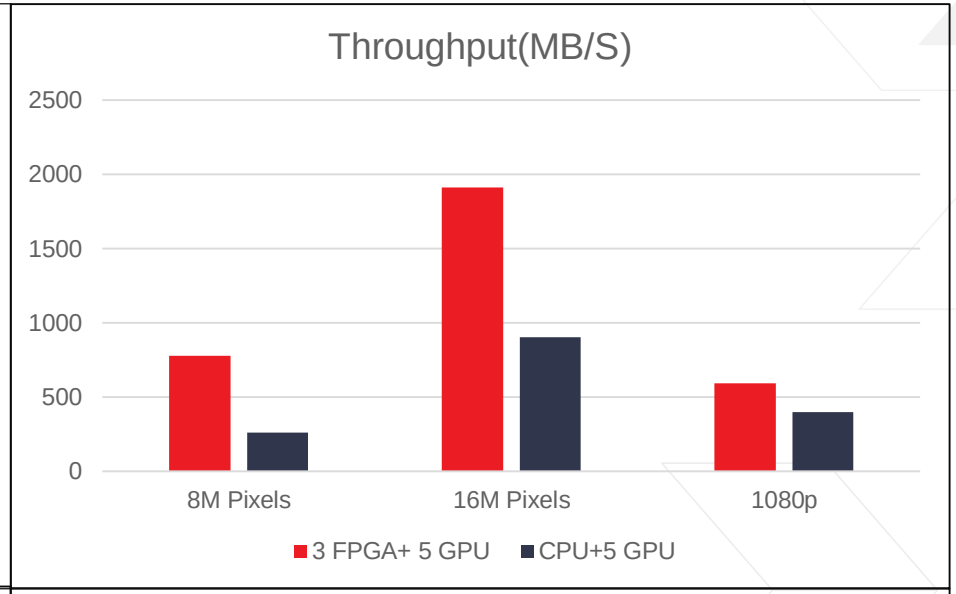
FPGA+GPU is **2** faster than CPU+GPU

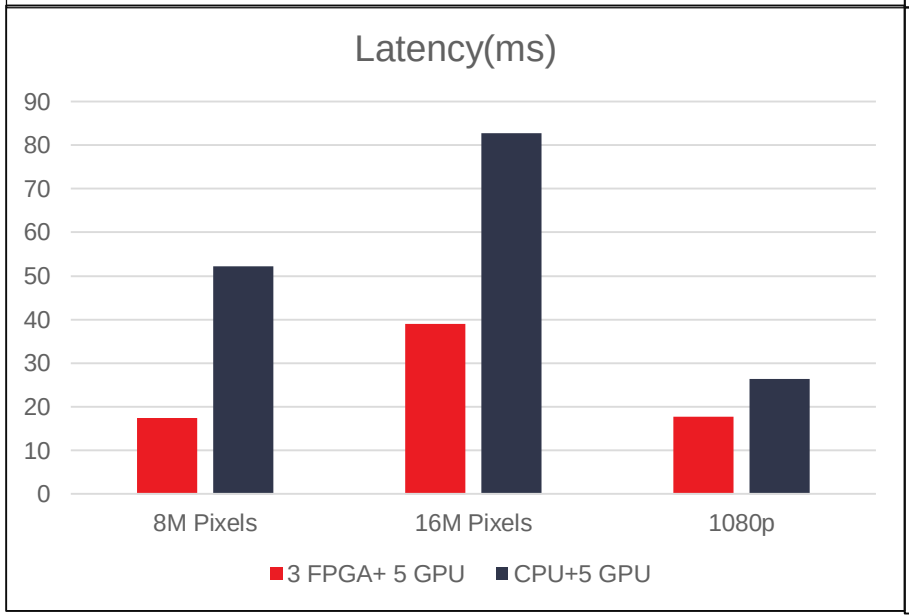
Latency :

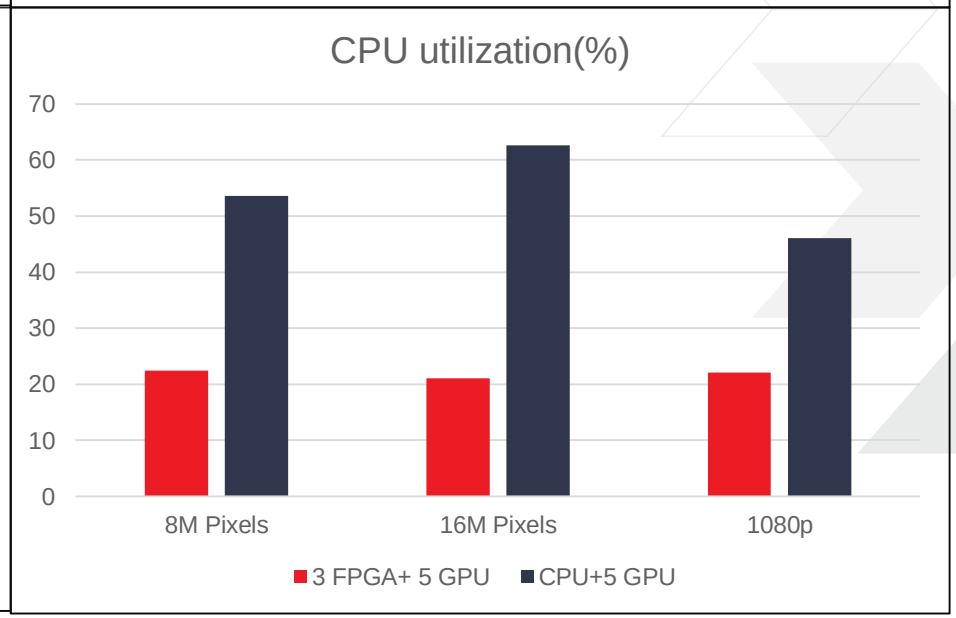
FPGA+GPU is **40%** of CPU+GPU

CPU usage :

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







Test result 4-ResNet-50

ResNet-50 model

QPS:

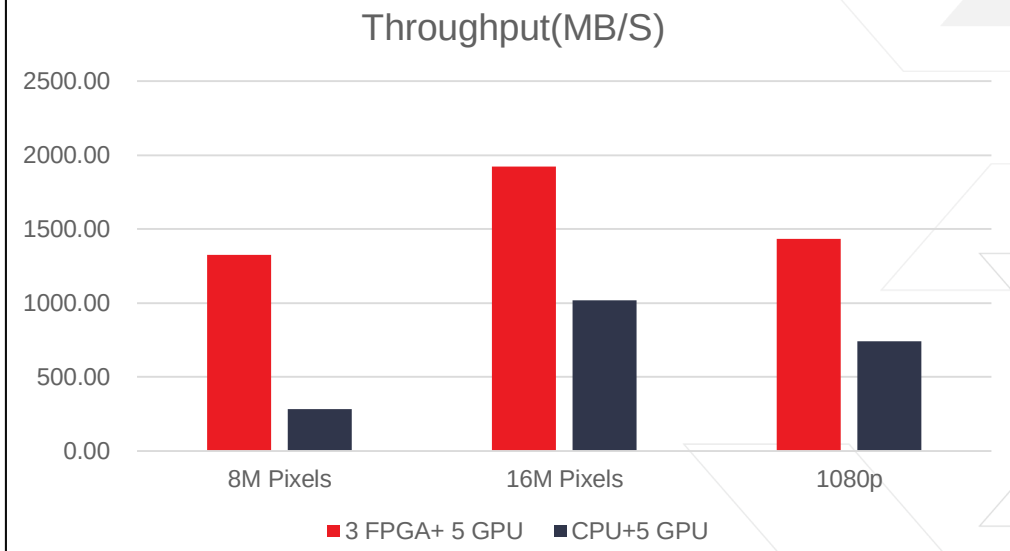
FPGA+GPU is **3** faster than CPU+GPU

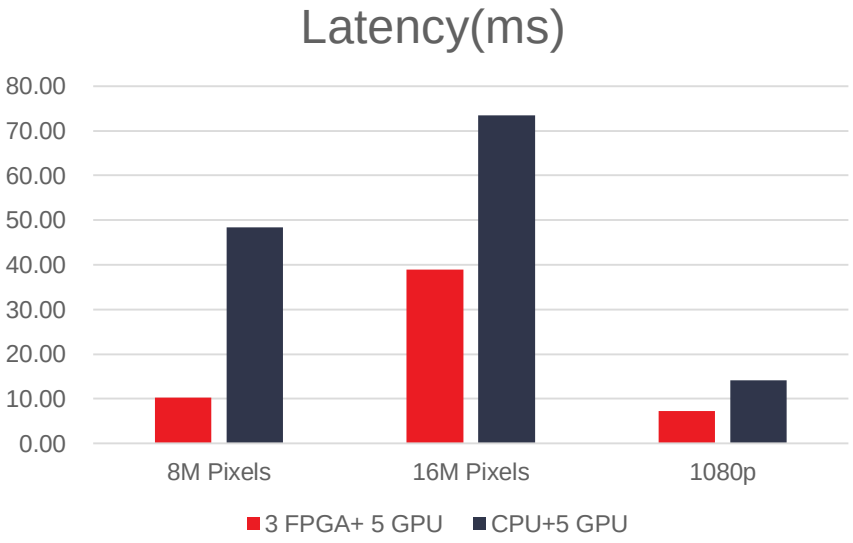
Latency :

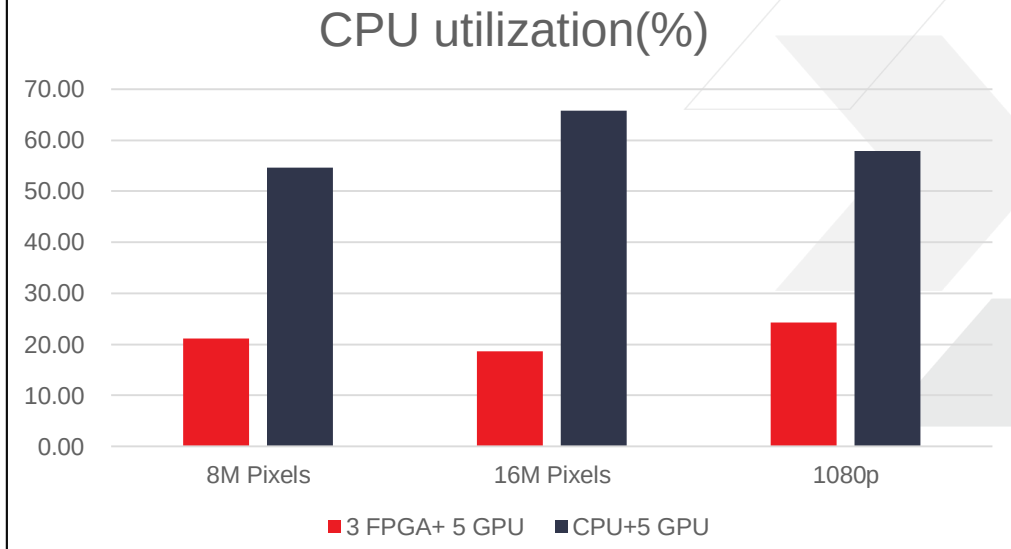
FPGA+GPU is **30%** of CPU+GPU

CPU usage :

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







Test result 5-Accuracy

Model	Category	Accuracy(top1)	Accuracy(top5)
Alexnet	TensorFlow	0.49	0.74
	accel	0.49	0.73
Inceptionv4	TensorFlow	0.80	0.95
	accel	0.79	0.95
ResNet50	TensorFlow	0.73	0.91
	accel	0.72	0.90
nsfw	TensorFlow	0.75	N/A
	accel	0.75	N/A

For any product related enquiries:

Product enquiry email:
info@ct-accel.com

Product website:
www.ct-accel.com

Cloud partners:

AWS

HUAWEI Cloud

Alibaba Cloud

Baidu Cloud

Tencent Cloud

Product hotline:
+86-0755-88914045

FPGA Partner: XILINX®

Adaptable.
Intelligent.