



# 超高性能图像处理方案 ——基于FPGA加速卡

樊平

深维科技 创始人/CEO

2019/12/13



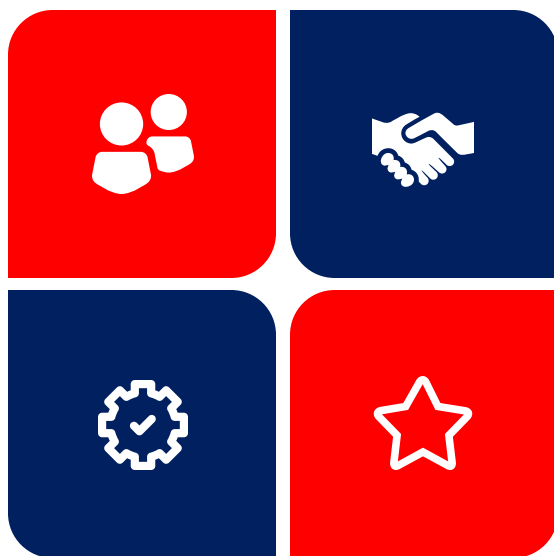
# 目录

- > 深维科技简介
- > 为什么需要图片加速
- > 我们为大家提供了什么产品
- > 部署是个大问题
  - >> 如何线下使用
  - >> 云计算进入FPGA异构加速时代
- > 总结
- > 我们的联系方式

# 深维科技简介

成立于2016年。核心团队对FPGA及图像、视频处理领域有十年以上的丰富经验。

公司专注在数据中心业务，主要集中于图像、视频处理、大数据和HPC等领域



赛灵思联盟成员  
赛灵思数据中心业务ISV  
AWS技术合作伙伴

在FPGA异构计算加速领域，  
拥有快速开发、全栈优化的  
独特能力。

# 图像处理，挑战巨大

移动应用与用户生产内容（UGC）正在驱动数据中心图像处理的业务负载快速增加...



缩略图生成



像素处理(Crop, Sharpen, etc.)

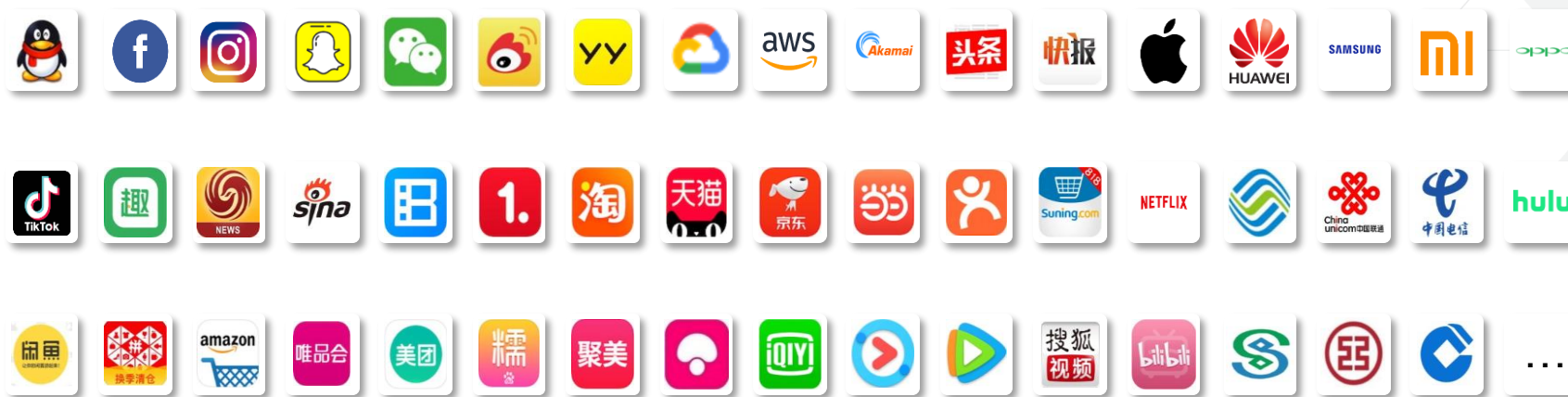


图片转码




智能分析 (Classification, OD, etc.)

众多的应用迫切需要更高性能、更低成本的算力方案



## 数据中心的核⼼问题：⽤户体验与服务成本

 TCO高昂

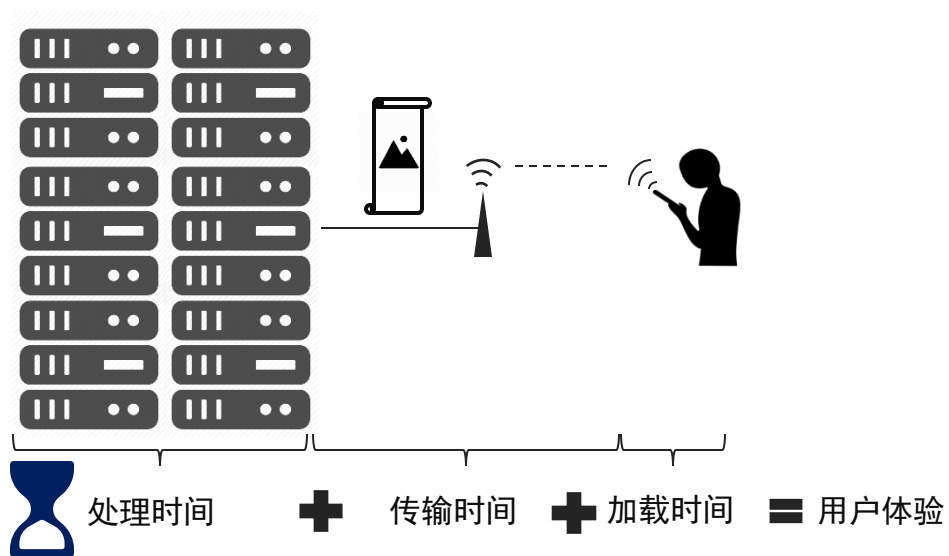
 用户体验差

服务器  
电费  
场地  
人员

\$ +

---

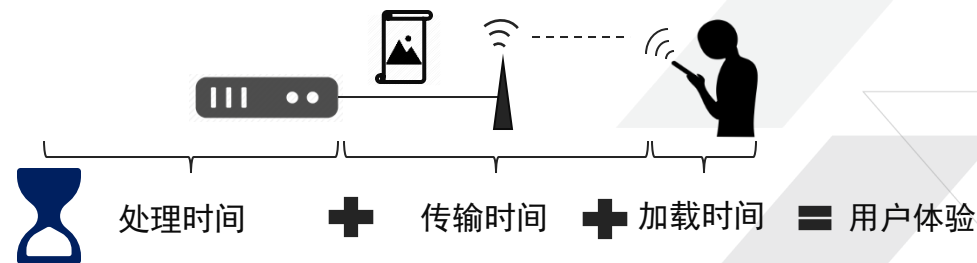
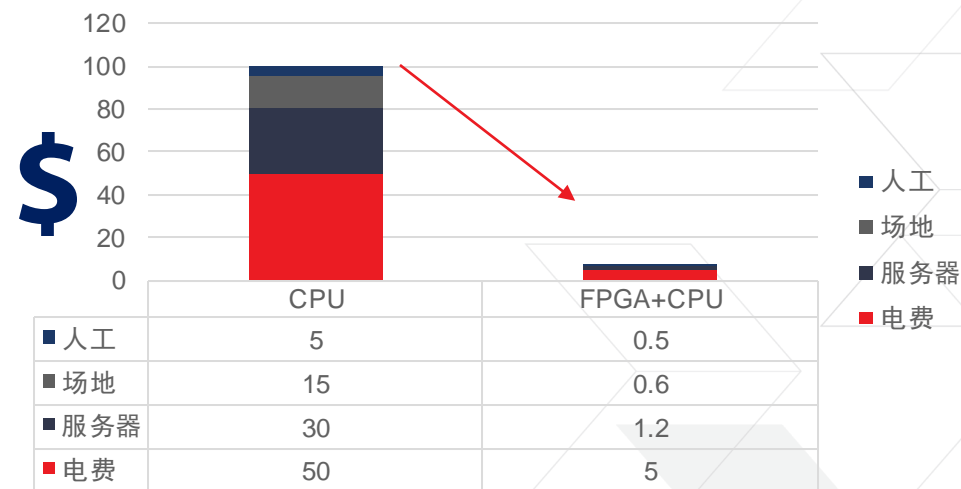
数据中心总成本（TCO）



## CPU处理方案

 TCO降低**5倍**

 处理延迟降低**20倍**



## DeePoly FPGA+CPU处理方案

深维给大家带来什么改变？

# 更快的JPEG2JPEG缩略图生成方案



20倍

并发性能



20倍

处理延迟



5倍

TCO缩减

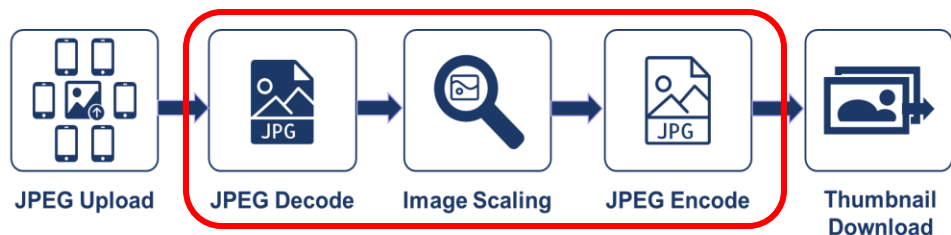


10倍

高效能

- Based on JPEG2JPEG benchmark
- TPS: Transaction per second
- TCO: Total cost of ownership

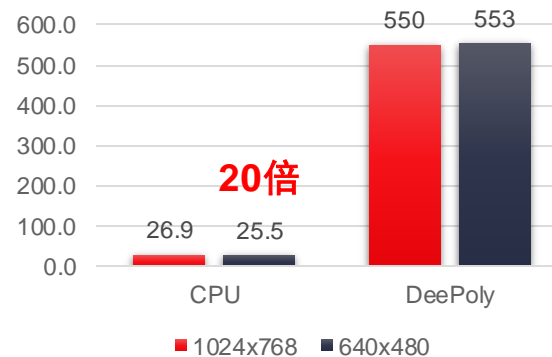
# JPEG2JPEG缩略图加速方案



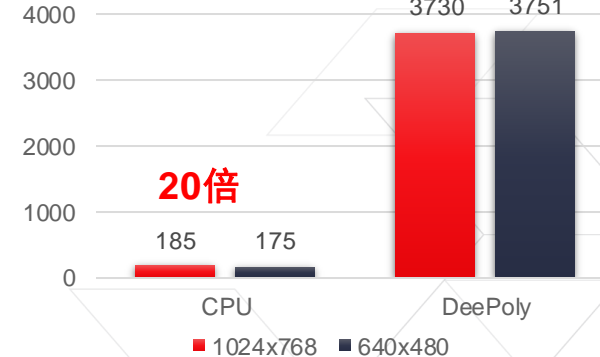
根据不同终端屏幕分辨率进行图像缩放处理



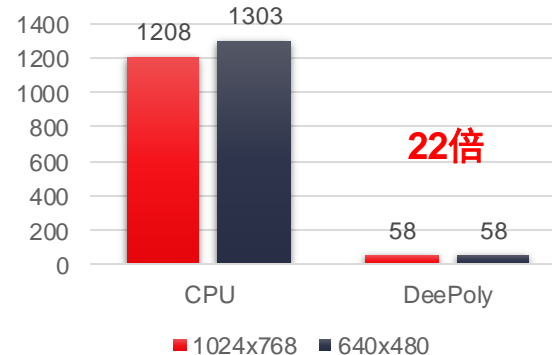
## TPS (Img/sec)



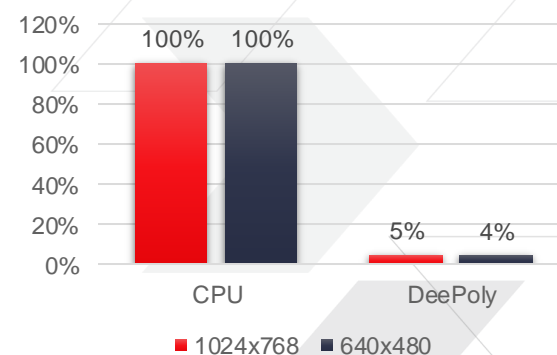
## Throughput (Mbps)



## Latency (ms)



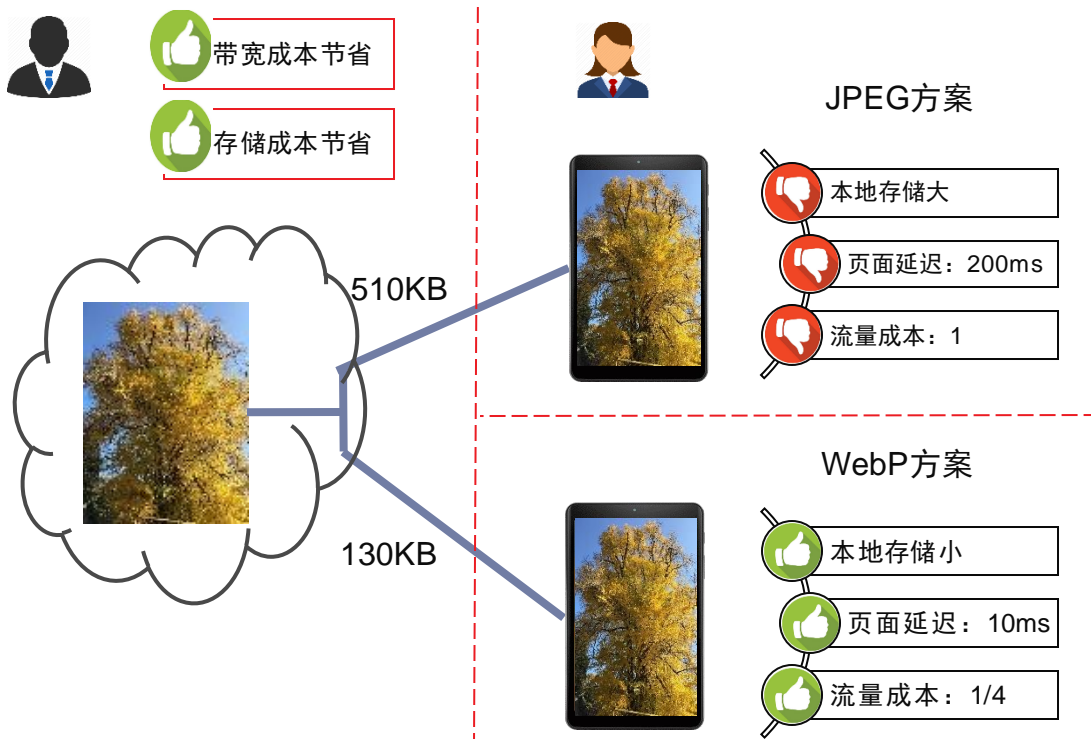
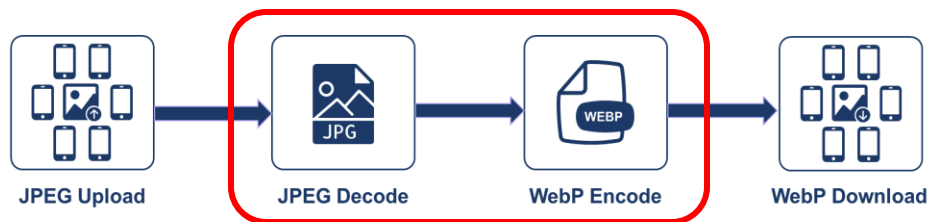
## CPU Utilization (%)



Input: 10000\*4096x2160 (Avg size 868K), Output: 1024x768, 640x480  
FPGA: 1pcs Xilinx Alveo U200, CPU: 2\*Intel(R) Xeon(R) CPU E5-2650v3



# JPEG2WebP转码加速方案



**WebP** 相较于JPEG，在同等图像质量情况下，压缩略可以提升25%-34%



代价是，WebP的计算复杂度也提升了**10倍**，我们该如何应对??

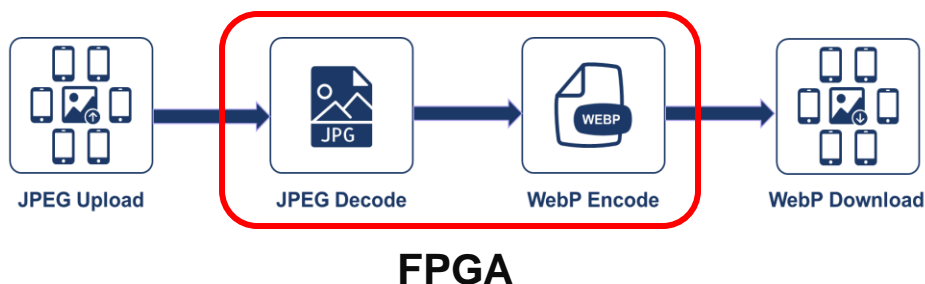
Format	Lenna	Kodak	Tecnik	Image_crawl
WebP: Average File Size (Average SSIM)	26.7 KB (0.864)	46.5 KB (0.932)	139.0 KB (0.939)	9.9 KB (0.930)
JPEG: Average File Size (Average SSIM)	37.0 KB (0.863)	66.0 KB (0.931)	191.0 KB (0.938)	14.4 KB (0.929)
Ratio of WebP to JPEG file size	0.72	0.70	0.73	0.69

WebP和JPEG的平均文件大小（SSIM对应图像质量为JPEG Q=75）

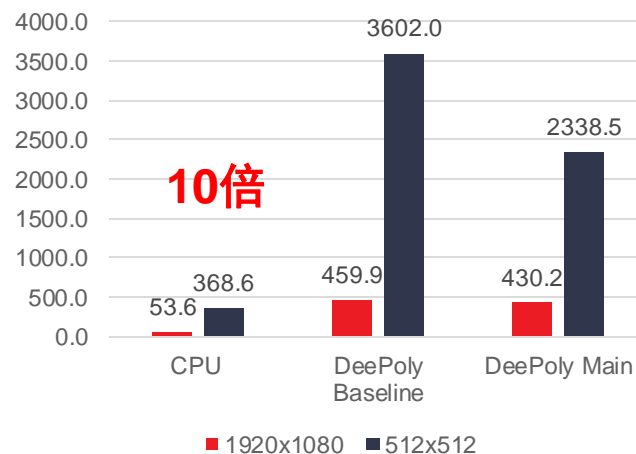
# ThunderImage JPEG2WebP转码方案

## 更快的WebP转码方案

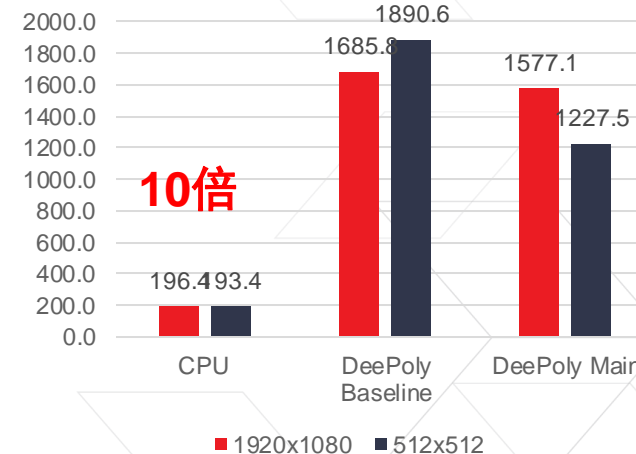
Product Type	WebP Mode	Bit-Accurate
Baseline	M4	No
Main	M4	Yes
Ultra	M6	Yes



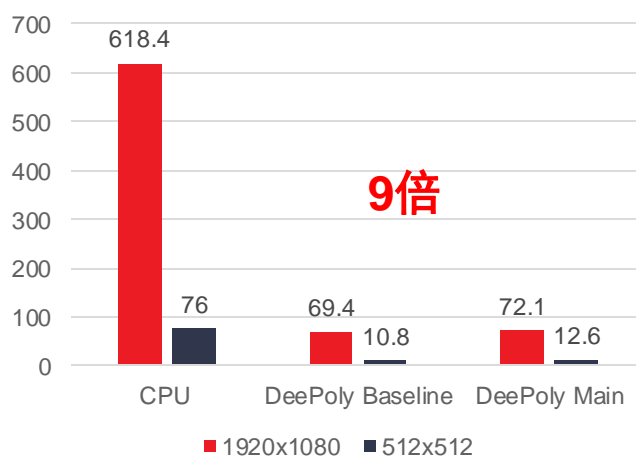
TPS(Img/Sec)



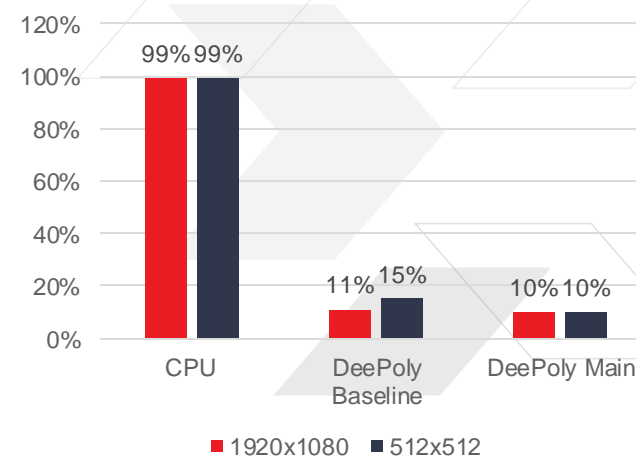
Throughput(Mbps)



Latency(ms)



CPU Utilization(%)



Baseline Mode, FPGA: 1pcs Xilinx Alveo U200 CPU: 2\*Intel(R) Xeon(R) CPU E5-2680v4

# 深维科技的旗舰产品 — ThunderImage



## 性能提升

转码、缩放性能提升 多至**20倍**



## 最佳QoS

处理时延大幅缩减 多至**20倍**



## 算法丰富

支持 WebP(M4/M6), JPEG, J2K and etc.  
ImageMagick Box filter, Lanczos, etc



## 简单易用

与ImageMagick 及 OpenCV无缝兼容  
与cpu算法结果严格比特一致

## 云平台支持



## OS支持



## CPU支持



## 硬件平台支持

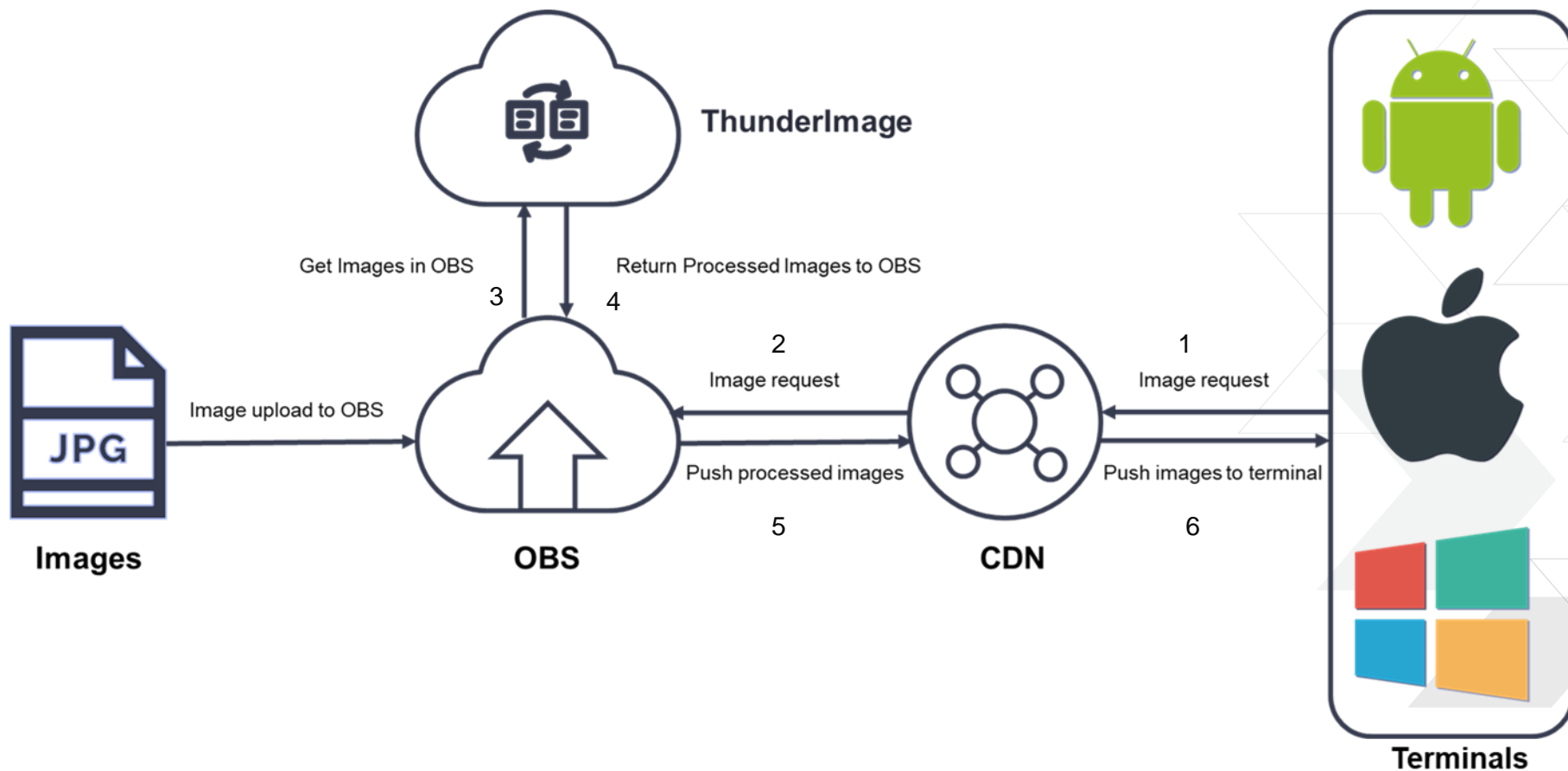


部署是个大问题

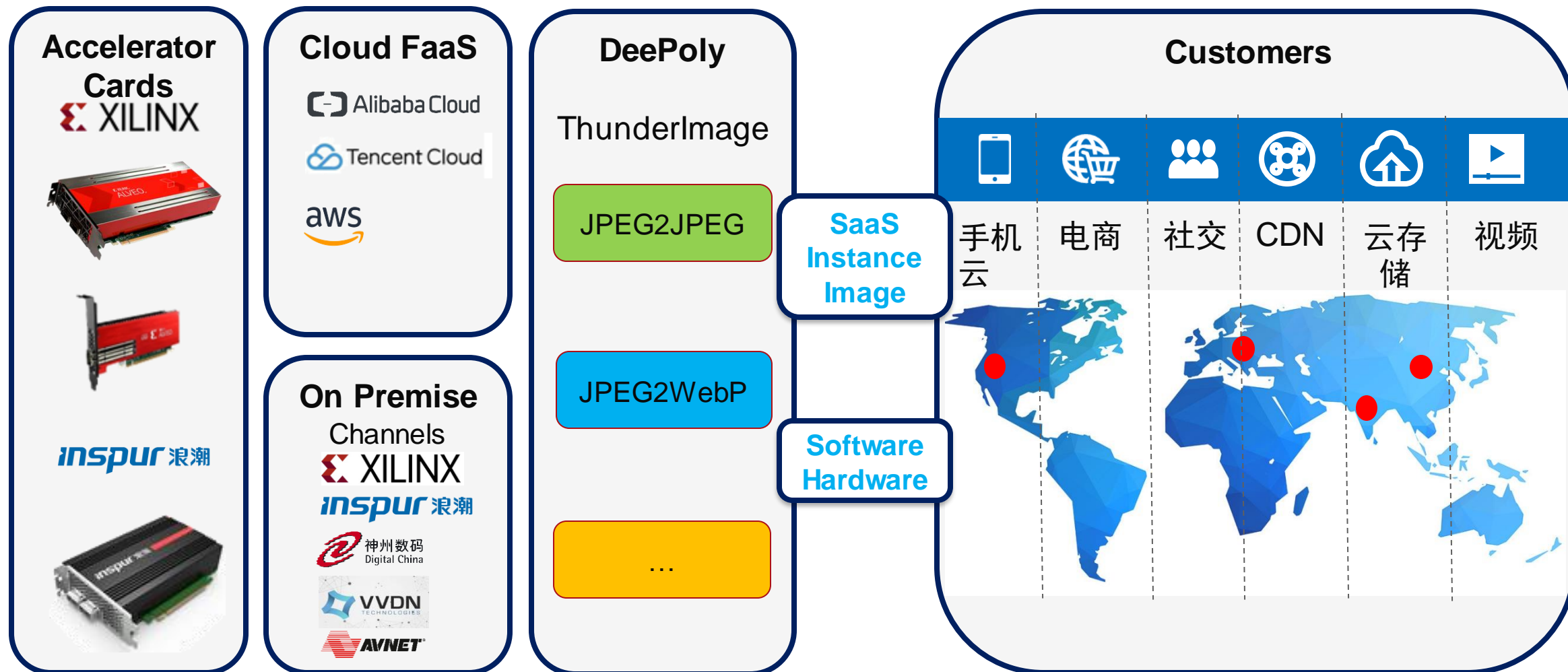
# 如何与生产环境集成

## 与OBS进行集成

- ✓降低带宽开销
- ✓降低存储开销
- ✓提升QoS



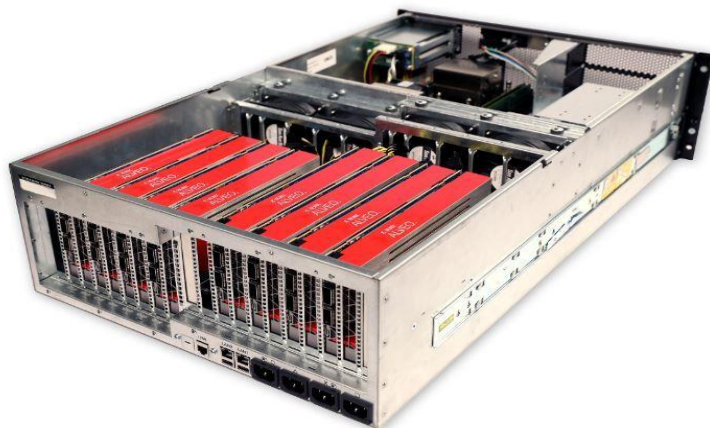
# 部署：公有云方案与私有云方案



# 私有云部署

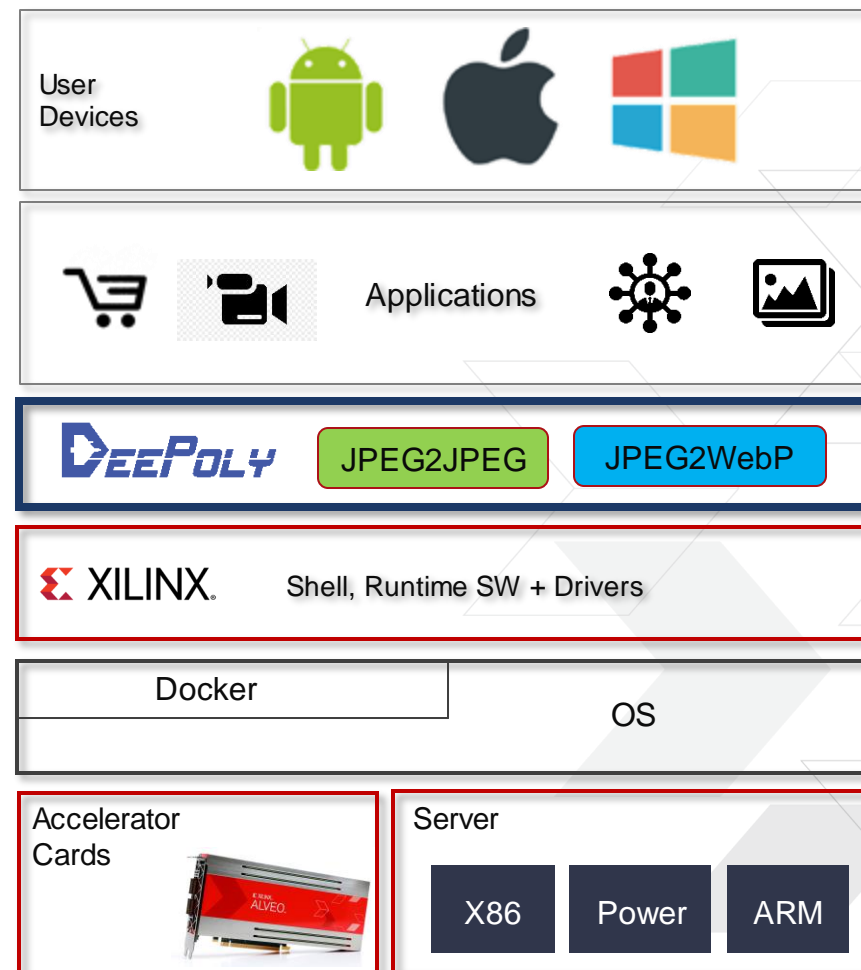


# 线下私有部署



FPGA加速卡	型号
	Xilinx Alveo U50
	Xilinx Alveo U200
	Xilinx Alveo U280

硬件环境



线下私有部署



# Alveo U50 – 最高计算密度，便捷部署

- > 更优的处理性能
- > 部署和维护更容易



**12倍**  
并发度



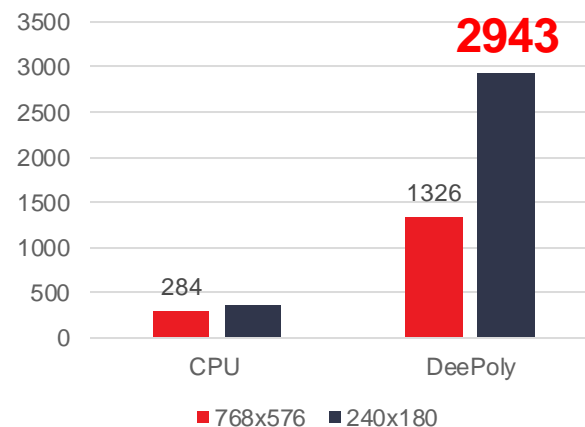
**25倍**  
低延迟

## Xilinx Alveo U50 Key Specifications

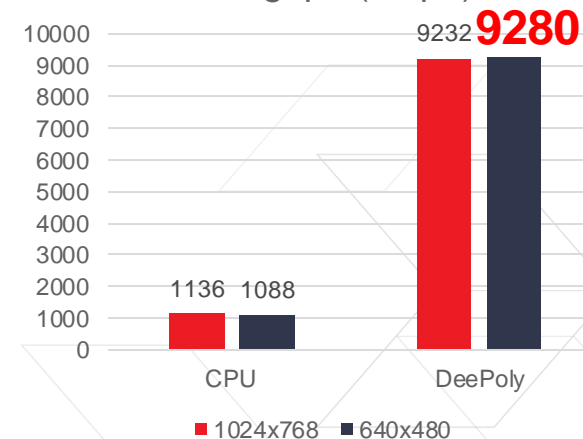


UltraScale+ Architecture  
Low-profile form factor  
8GB HBM2 Memory, 460GB/sec  
PCIe Gen4, CCIX, PCIe Gen3  
QSFP 28 (100GbE)  
< 75W

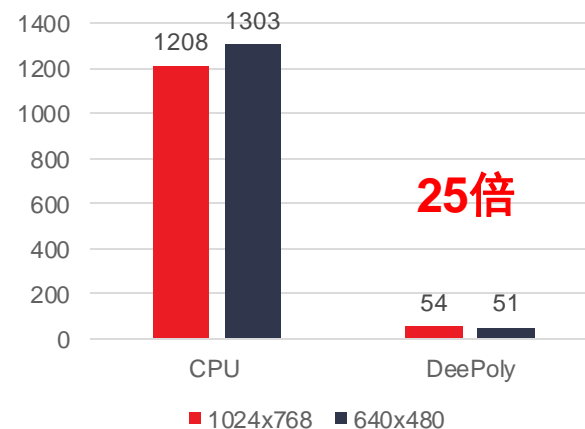
TPS(Img/Sec)



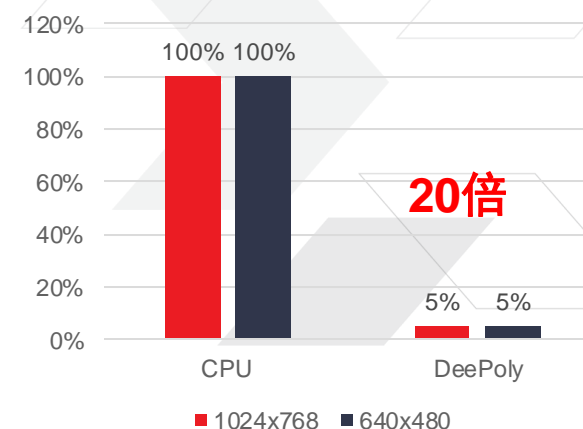
Throughput(Mbps)



Latency(ms)



CPU Utilization(%)

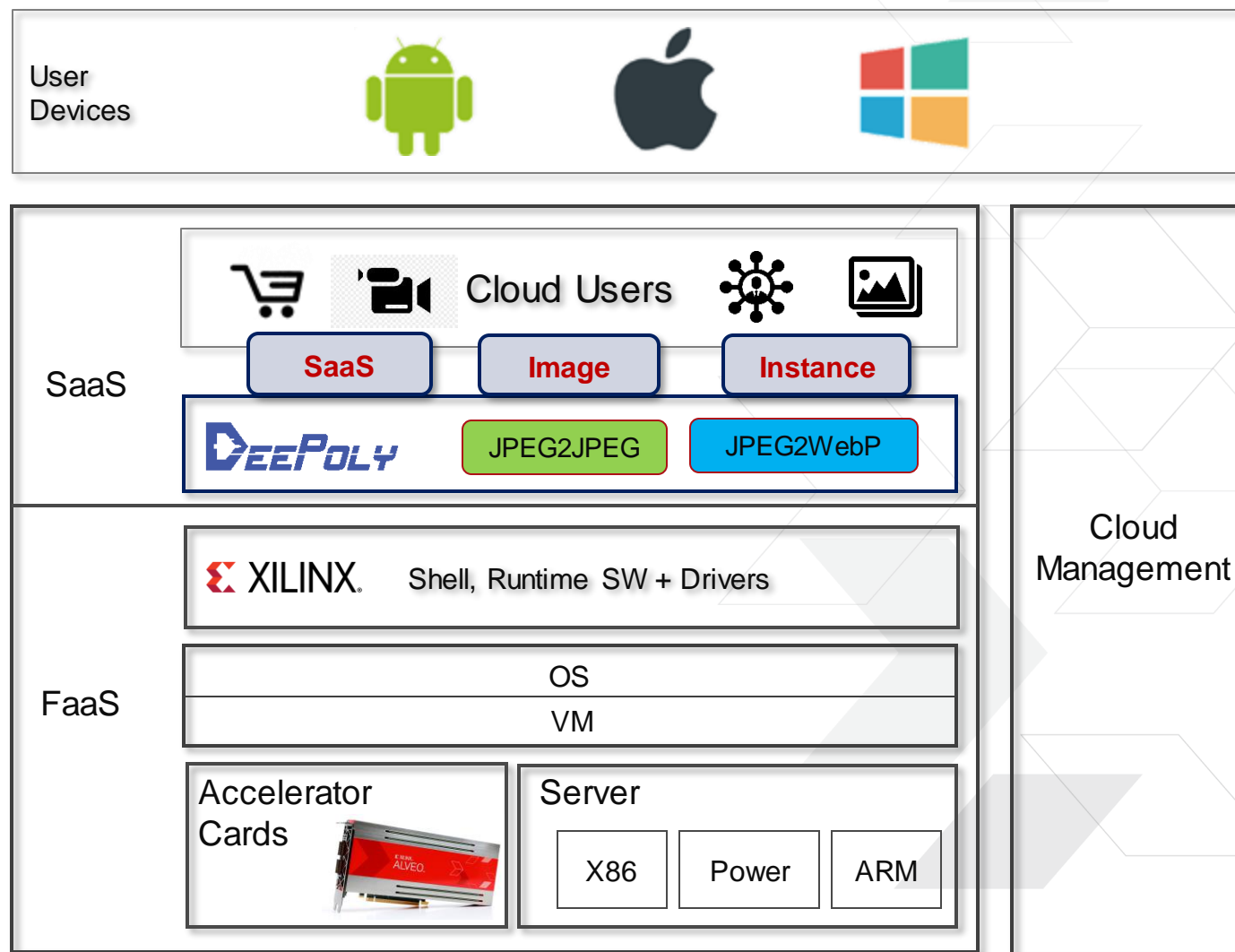


FPGA: 1pcs Xilinx Alveo U50 CPU: 2\*Intel(R) Xeon(R) CPU E5-2650v3

# 公有云部署

# 公有云部署

平台	腾讯云	FX4
	阿里云	F3
	AWS	F1
功能	JPEG2JPEG	缩放
	JPEG2WebP	M4
		M4 b2b
服务	SaaS	易部署
		起步费用低
	Instance	集成灵活
	Image	更优成本
		最强性能
		深度定制

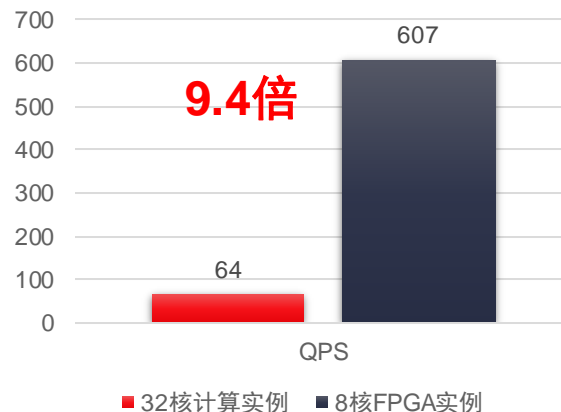


# 腾讯云FaaS首发JPEG缩略图镜像服务

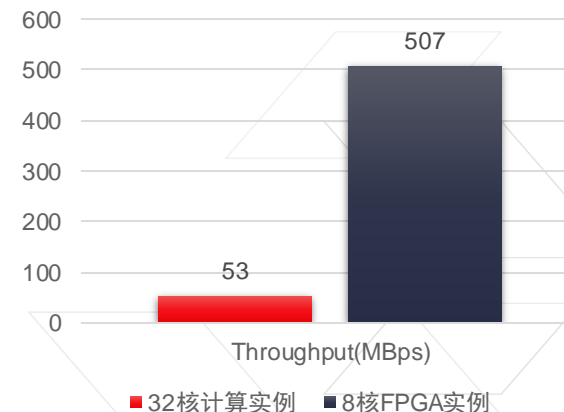
- > 腾讯云FaaS实例FX4
- > ThunderImage JPEG缩略图镜像服务
- > 高并发性能
- > 低处理延迟

腾讯云市场 (Tencent Cloud Marketplace) 页面展示了 ThunderImage JPEG 缩略图处理加速系统。该服务由北京深维科技有限公司提供，具有超高性能、20x 吞吐量提升和 5x 延迟降低的特点。当前价格为 ¥0，并提供免费使用选项。页面还列出了商家的联系方式、服务时间以及相关的服务协议。

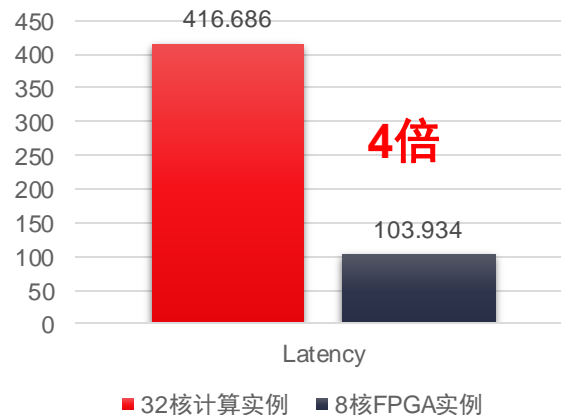
## 单实例并发性能比较



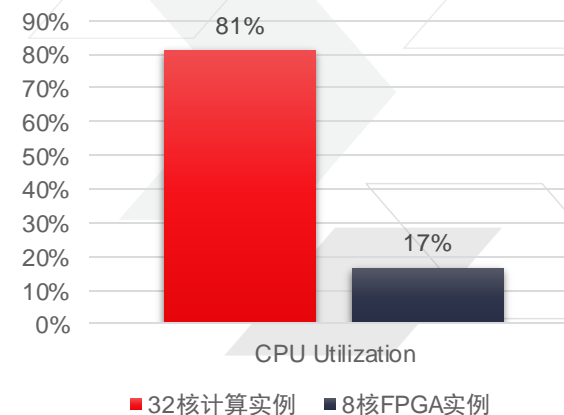
## 单实例吞吐性能比较



## 处理延迟比较



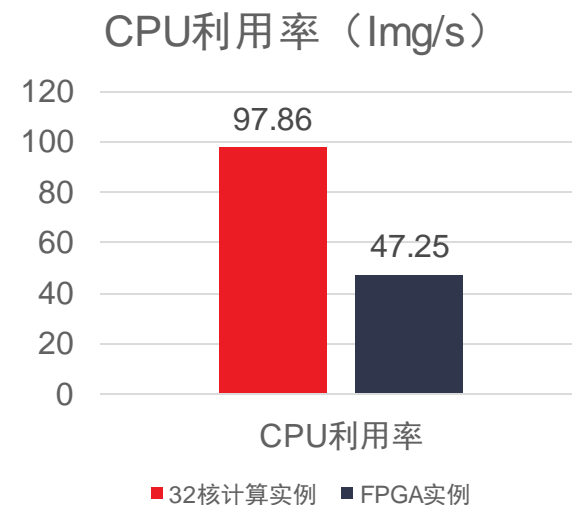
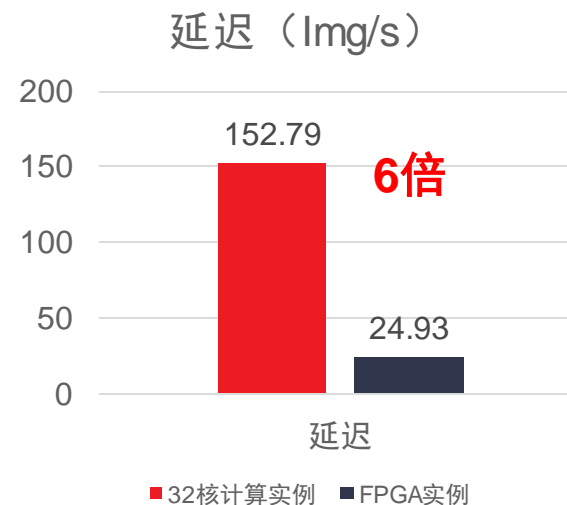
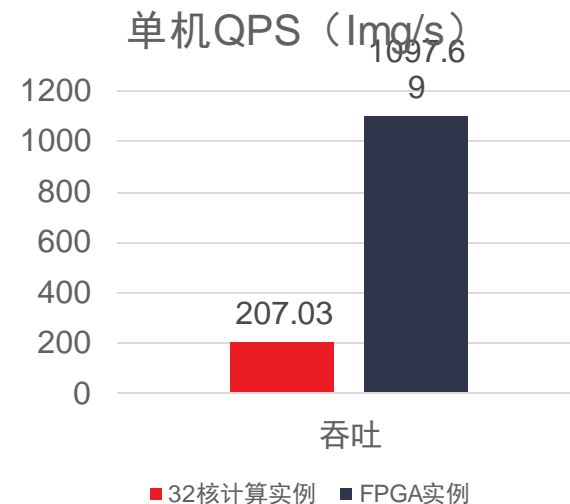
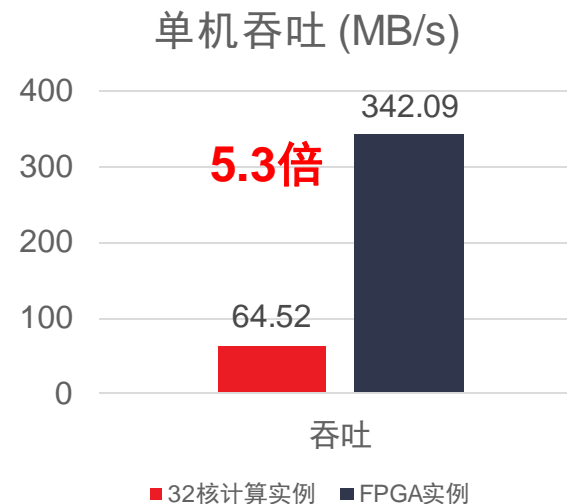
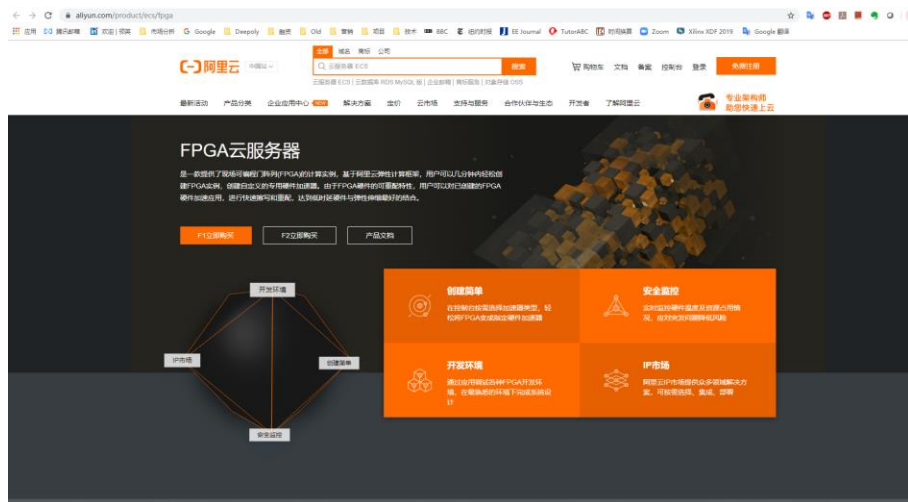
## 单实例CPU利用率比较



腾讯云8核FaaS实例 vs. 腾讯云ECS32核高性能实例

# 阿里云F3首发JPEG2WebP镜像服务

- > 阿里云F3实例
- > ThunderImage JPEG2WebP镜像服务
- > 高并发性能
- > 低处理延迟
- > 结果与CPU结果完全一致



阿里云 f3-c8f1.2xlarge vs. 阿里云 ecs.c5.8xlarge 实例

# 成功案例

# 客户案例1



客户

社交网络



Scenario

云相册的缩略图生产



ThunderImage

JPEG Decoder, Scaler,  
JPEG Encoder



部署方式

线下私有部署



高性能

单节点吞吐性能提升10倍



降低TCO

TCO降低50%



更优的服务质量QoS

延迟降低10倍

高负载下QoS保持稳定表现



更佳的可维护性

小规模集群，维护方便

# 客户案例2



客户

某视频网站



Scenario

WebP转码



ThunderImage

JPEG Decoder, Scaler,  
WebP Encoder



部署方式

公有云部署



高性能

单节点吞吐性能提升6倍



降低TCO

TCO降低40%



更优的服务质量QoS

延迟降低6倍

高负载下QoS保持稳定表现



更佳的可维护性

基于公有云集成，弹性规模，维护方便



# 总结

# ThunderImage重新定义了图片处理性能标准

## 超高性能

单机高达20倍吞吐提升，20倍延迟降低，10倍功耗降低，5倍TCO降低

## 如何进行本地私有部署

深度优化，极致性能，完整的软件栈支持，丰富的硬件平台选择

## 如何在公有云上部署

丰富产品选择，完满匹配大型客户与中小客户业务特点，主流大云均已支持

# 欢迎大家试用PoC

## 客户支持



Product Support Email

support@deepoly.com



Product Website

www.deepoly.com

## 合作伙伴



## 公有云平台



**Adaptable.**  
**Intelligent.**

Redefine the Image Processing

海量图片，轻松搞定

**DEEPLY**



# 联系方式

欢迎一起探讨FPGA异构计算加速，我的微信号：**makefpgaeasy**



樊平-深维科技

北京 海淀



扫一扫上面的二维码图案，加我微信

# 樊平简介

## > 深维科技，创始人/CEO

- >> FPGA异构计算专家，拥有15年FPGA EDA以及芯片架构设计经验。
- >> 于2016年创办深维科技，致力于发展FPGA异构计算加速技术，为视频、图像智能处理领域提供最具性价比的异构计算解决方案。
- >> 在创办深维科技之前，曾在IBM，Cadence和京微雅格担任研发工程师、技术总监等职务，发明数十项国际和中国专利（含在申）。
- >> 毕业于北京航空航天大学，拥有计算机硕士学位。