

# 展望FaaS云服务

## ——FPGA云服务的技术架构和行业应用



高剑林

腾讯公司 FPGA技术负责人  
FPGA云服务技术负责人

# 展望FaaS云服务

- 云服务不是水和电

# 云服务的本质

- IT向全社会的扩散

- IT只在少数行业、少数公司获得应用，大多数传统行业中小型公司没有能力应用IT(缺少开发、运营、迭代的能力，而非购买设备)
- 无开发能力行业使用IT的关键是云

- 使用云服务的方式

- 如果云服务只是提供硬件，将避免不了和自购设备比较，以价格决定是否使用
- 云服务应以服务交付，立即可用，尤其重要的是：随时可以集成到并可拼接入业务

# 云服务的本质

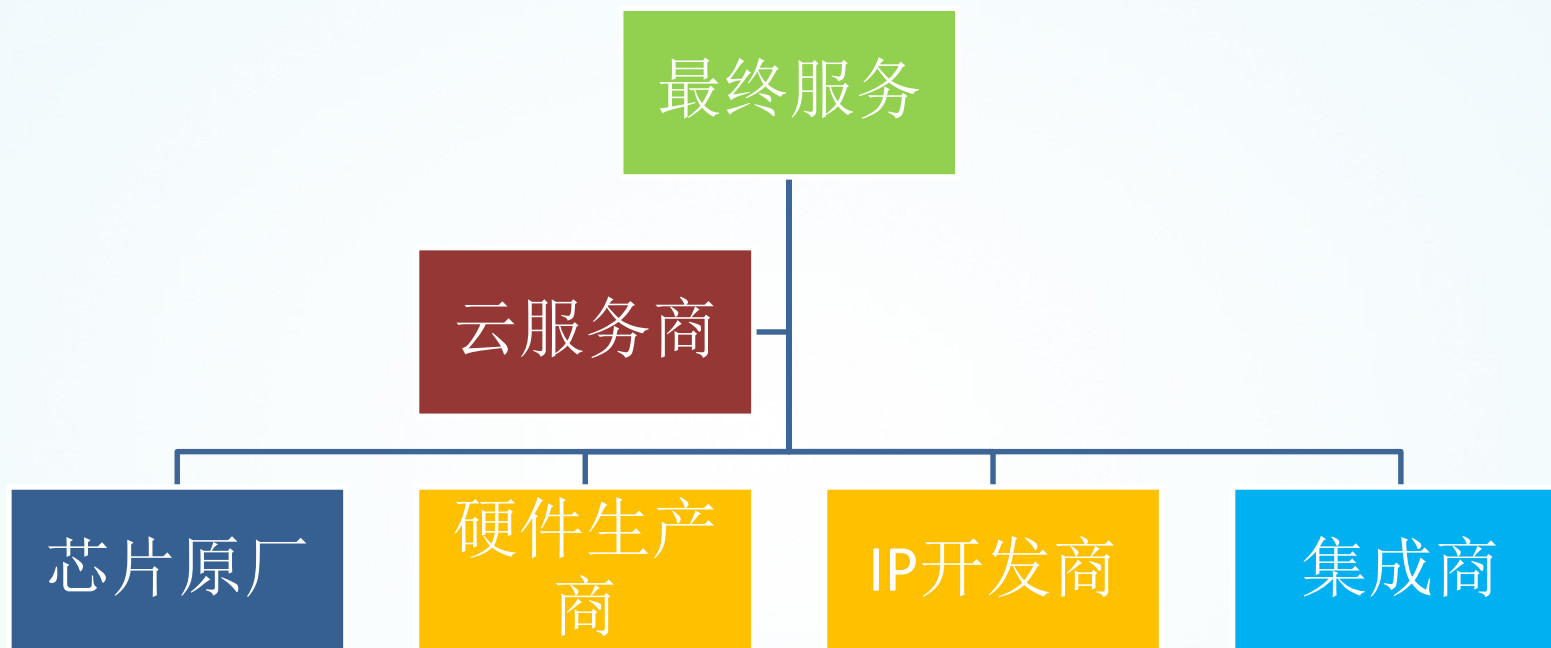
- 人类社会两次协作革命

- 第一次：商品出现和商品经济：大规模的分工和协作成为可能，任何个体和公司都购买商品/零部件、集成到自己的产品、对外售卖商品
- 第二次：云服务出现

- 云服务的本质

- 云是一种协作模式:更快的应用售卖的服务（商品化的服务？），更快的集成到自己的服务，更快的对外提供服务

# FPGA如何提供云服务



- 当前FPGA产业链存在多个环节，需要整合多方的技术和资源，才能形成最终服务化产品



# FPGA如何提供云服务

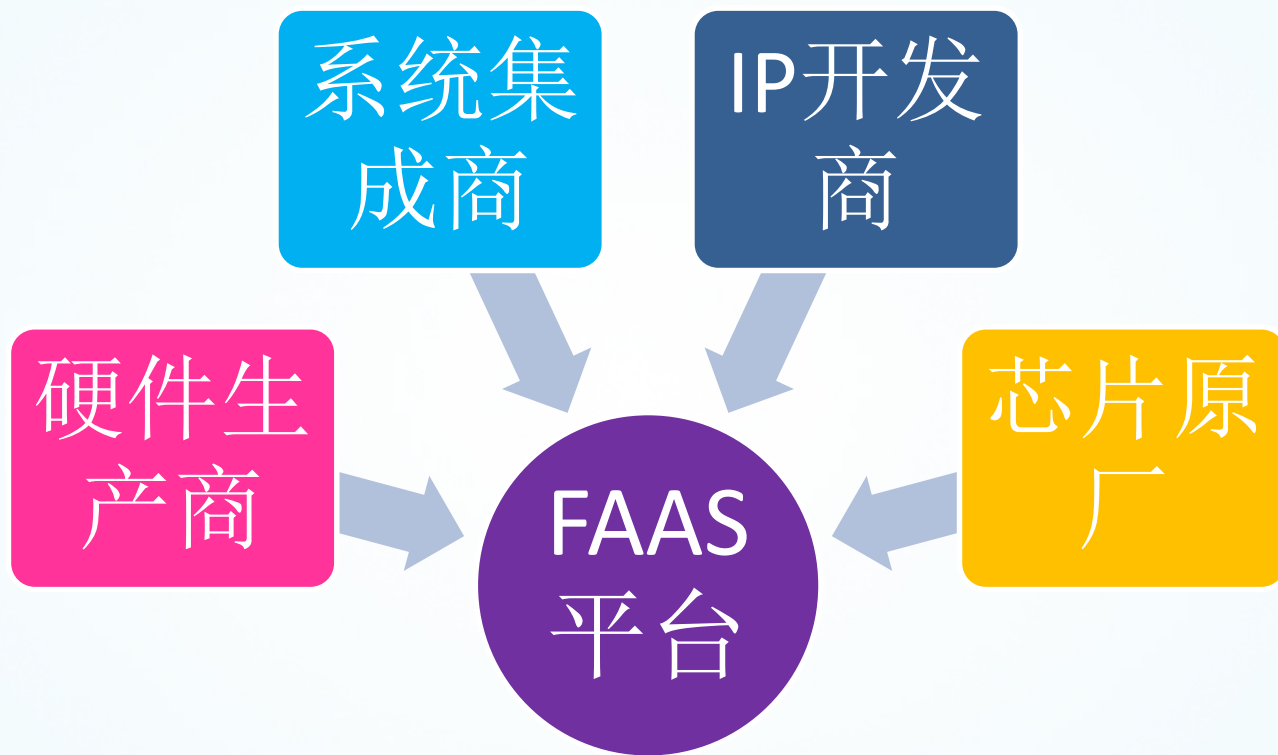
## • 传统供应商的痛点

- 芯片原厂：赛灵思和altera两大厂商只提供芯片，不提供整体方案，增长缓慢，不直接接触用户，方案都是半成品
- IP开发商：以IP出售作为商业模式，不提供硬件和其它服务，IP只是零部件，不能直接使用，IP商担忧产权泄露，不提供验证和测试，需要签NDA并付费，交付周期极长、过程极繁琐
- 集成供应商：技术人员带硬件拜访客户，现场讲解，留下硬件供验证和测试，过程繁琐，成为IDC供应商需要认证和商务谈判，硬件维护麻烦

## • 用户的痛点

- 高端用户（华为、中兴）具有硬件、软件开发能力，自行开发所有组件
- 低端用户完全依赖集成商，价格贵升级慢
- 测试验证周期长，硬件购买需排期

# FAAS的优势





# FAAS的优势

- 对传统供应商的优势

- IP开发商：提供在线验证功能，IP自动部署到生产环境，借助网络提供用户数据上传和计算，结果自动返回
- 集成供应商：不再提供硬件售卖方式，而是通过云提供设备，按需购买，解决硬件的备料问题，解决人员距离导致的响应问题，维护由云平台负责
- 芯片原厂：解决原厂服务小用户的难点，新的模式带来新的用户，恢复增长

- 对用户的优势

- 硬件成本低，采用最新的工艺和技术
- 随时弹性扩容和裁减
- 线上选择IP，测试验证快，和用户生产环境吻合

# FAAS以服务交付的优势

- 易集成

- fpga厂商只做硬件和IP，守着自己的领域，不能扩展
- 使用云服务，可以自己的知识产权和算法为核心，使用他人的IP，使用腾讯云提供的存储、计算、中间件、数据库和网络服务，搭建自己的扩展服务
- 比如：图片编码、解码厂商使用图片识别服务，迅速实现图片安全监测、OCR、人类识别等业务，图片可存入腾讯云的图片云

- 按需使用

- 可以使用次数计费、包月计费、机器数量计费
- 弹性扩容和收缩

- 快速进化

- 开放的开发者社区、使用者社区，行业体系+评价体系
- 打破信息孤岛，联结一切，产生新的生态

- AI的应用
  - ocr的识别
  - 语音计算
- 医疗和基因计算
  - 基因测序的加速服务
  - 病理图片的鉴别
- 金融
  - 量化交易
  - 证券行情分发

# 医疗和基因计算

- 测序成本不断降低并逼近数据分析成本，同时随着近年基因数据量爆发式增长，数据分析成为瓶颈，而传统方法计算速度慢，效率低。样本采集、基因测序、数据分析和鉴别、数据保存各个链条脱节
- 解决方案：基于FPGA异构加速技术，搭建测序工厂、IDC机房、基因数据存储、分析和鉴别一体化的云服务

| BWA+GATK    | 纯CPU方案              | FPGA  | Speedup(x) |
|-------------|---------------------|-------|------------|
| 机器配置        | E5-2620v3,<br>64G内存 | 腾讯FX2 |            |
| BWA_MEM /小时 | 16                  | 7     | 2~3        |
| GATK /小时    | 28                  | 14    | 2          |

# 医疗和基因计算

- 与其他类GATK解决方案不同，支持完整的BWA+GATK flow。
- 全基因计算的CPU+FPGA异构加速方案，搭载业界高性能FPGA，包含 130 万个逻辑单元，2 条16 GB 高速DDR4 内存，支持高速 PCIe 3.0高速接口。
- 对其中关键算法实现进行重写，并行，流水和优化
- 支持腾讯云端快速部署





# ML FPGA加速

## Flexibility : 可编程性适配快速演进的ML算法

- DNN、CNN、LSTM、MLP、reinforcement learning以及决策树等等
- 任意精度支持
- 模型压缩、稀疏网络、更快更好的网络

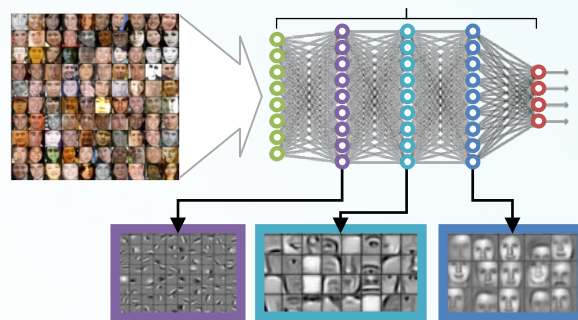
## Performance : 构建实时性AI服务能力

- 与GPU 媲美的加速性能
- 较于GPU数量级的推断延时优势

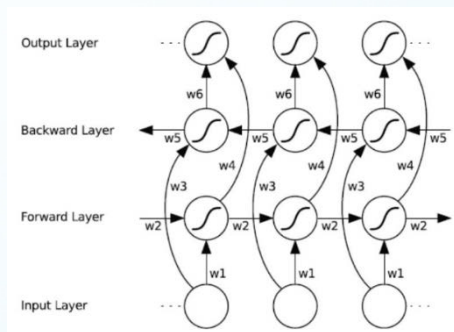
## Development : 开发周期持续优化

- 日益完善的基本算子库
- 通用性：控制计算抽离架构+编译器

## Convolutional Neural Networks



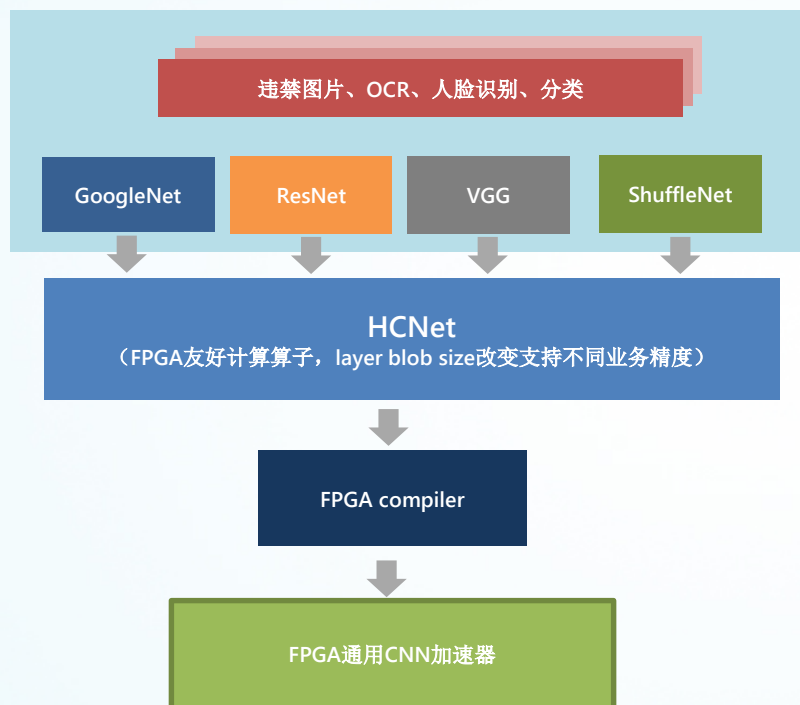
## Recurrent Neural Networks





# ML FPGA加速

## FPGA CNN解决方案



**Flexibility** : 支持业务秒级切换

**Performance** : TCO成本降低

**Development** : 算法硬件协同设计, 新模型支持周期缩短到2周

# 谢谢！