嵌入式系统联谊会
www.esbf.org

威视锐® V3 Technology,.Ltd. INNOVATION FOR RESEARCH

利用FPGA加快嵌入式人工智能系统的部署

北京威视锐科技　　姚远

公司简介

威视锐® V3 Technology,.Ltd.

北京威视锐科技有限公司，成立于 2008年的国家级高新技术企业，专注于提供创新性科研和教学解决方案，主要面向无线通信频谱安全、人工智能视觉感知与核生化监测防护等领域。

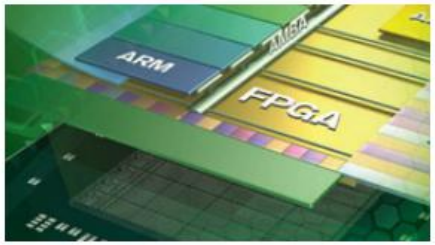
威视锐是全球最大的可编程芯片厂商XILINX的认证设计伙伴和授权培训中心，也是全球领先的模拟芯片厂商ADI的中国区大学计划推广合作伙伴。同时，威视锐也是IBM研究院和微软研究院的长期技术提供商。

IBM　Microsoft®　ANALOG DEVICES 超越一切可能™　XILINX® | ALLIANCE PROGRAM CERTIFIED MEMBER

软件定义无线电SDR


软件定义测试仪SDI


软件定义视觉平台SDVision



软件定义SoC平台SDSoC

典型客户

1000+ 行业客户

200+ 大学和研究机构

20+ 国家和地区

无线通信频谱安全

核心产品　　行业定制

学术合作　　联合实验室

面向不同应用的全系列SDR平台

实时软件无线电平台，

加快下一代无线通信系统的验证

人工智能视觉感知

从终端到云端的机器学习推断

机器学习的应用正迅速地扩展至越来越多的终端市场，在终端、在云端或者在那些基于端处理与基于云的数据分析相结合的混合解决方案中。Xilinx 为部署高级高效率神经网络、算法及应用提供各种开发堆栈及硬件平台。

核生化监测防护

放射性元素测量仪器　　核设施环境检测　　高能物理科研和教学仪器

学院

电子工程师训练营

PART 4

V3学院是威视锐旗下的教育培训品牌，专注于提供系统化工程教育培训，解决高校人才培养和企业用人需求的矛盾。V3学院的专业设置主要有嵌入式系统、人工智能和IC设计等几个方向的国内紧缺专业。

目前，V3学院拥有北京和上海两大培训基地，西安、南昌等多个培训中心。每年参加V3学院培训或者研讨会的在校学生数量超过2000人，在职工程师数量超过1000人。

同时，V3学院通过免费视频、在线直播、微信课堂等多种在线方式，帮助学员完成远程学习和训练。公司2008年成立以来，网络上免费的培训资料影响近万人。

ANALOG DEVICES
超越一切可能™

XILINX Authorized Training Provider

arm

目录

CONTENTS

1

AI和深度学习背景

- 发展简史
- 行业背景
- 平台对比
- FPGA优势

2

云端智能应用

- 业务需求
- 产品形态
- 开发案例
- 机遇挑战

3

终端智能应用

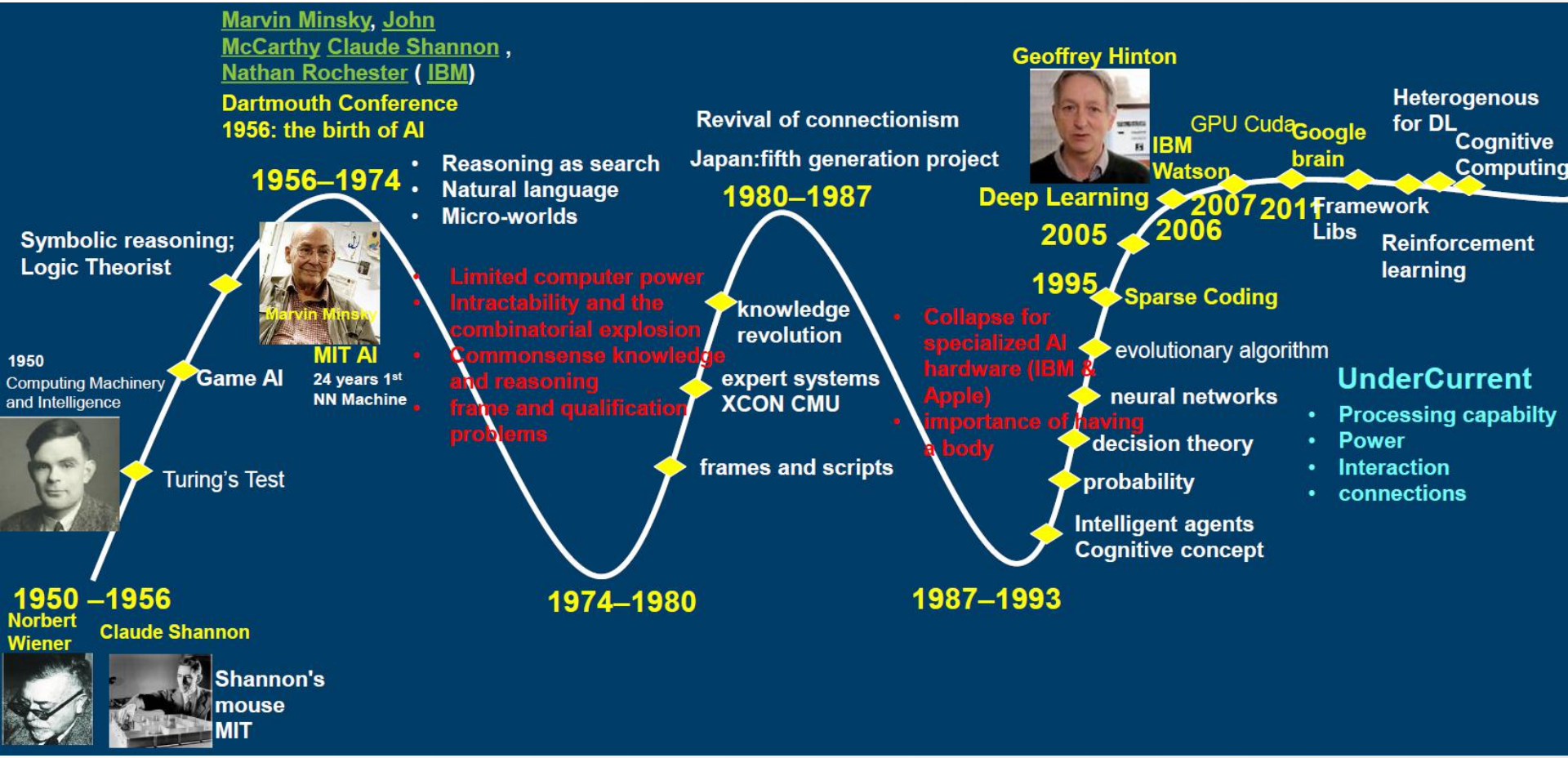
- 面向低延迟
- 面向工业检测

PART 1

AI背景介绍

发展历史　行业应用　平台对比　FPGA优势

AI发展简史


Marvin Minsky, John McCarthy Claude Shannon , Nathan Rochester (IBM)
Dartmouth Conference
1956: the birth of AI
1956–1974
Reasoning as search
Natural language
Micro-worlds
Symbolic reasoning; Logic Theorist
Marvin Minsky
MIT AI
24 years 1st NN Machine
1950
Computing Machinery and Intelligence
Game AI
Turing's Test
1950 –1956
Norbert Wiener
Claude Shannon
Shannon's mouse MIT
Limited computer power
Intractability and the combinatorial explosion
Commonsense knowledge and reasoning
frame and qualification problems
1974–1980
Revival of connectionism
Japan:fifth generation project
1980–1987
knowledge revolution
expert systems XCON CMU
frames and scripts
Collapse for specialized AI hardware (IBM & Apple)
importance of having a body
1987–1993
Geoffrey Hinton
Deep Learning
2005
2006
1995
Sparse Coding
evolutionary algorithm
neural networks
decision theory
probability
Intelligent agents
Cognitive concept
GPU Cuda
IBM Watson
Google brain
2007
2011
Framework Libs
Heterogenous for DL
Cognitive Computing
Reinforcement learning
UnderCurrent
Processing capabilty
Power
Interaction
connections

AI视觉大事件

人工智能 | 机器学习 | 计算机视觉

2012年，ImageNet上图像识别率一下子提高十几个百分点（以往通常是1年1~2个点）；

2016年，谷歌下了一盘棋叫AlphaGo；

2017年， 人工智能在自动驾驶（百度）、医疗诊断（IBM）方面不断有突破

2018年， AI芯片批量上市，行业门槛降低，出现成熟应用场景

应用场景

- **安防监控**

  智能交通、智慧安防、结构化视频、人脸识别

- **汽车/机器人**

  辅助驾驶、自动驾驶、新零售、生产自动化

- **智慧金融**

  智能投资顾问、金融预测与反欺诈融资授信、安全监控预警、智能客服以及安全支付

- **医疗电子**

  社区医院、智能分诊、人工智能参与的智能问诊、基因分析和精准医疗

技术要素

深度学习核心：

学习算法的设计，你设计的大脑到底够不够聪明；
要有高性能的计算能力，训练一个大的网络；
必须要有大数据；

适当的用户场景；
成熟的商业模式。

深度学习计算平台

CPU	面向逻辑运算和事务处理 首要目标是人机界面和多任务调度。
GPU	面向密集的、高并行的计算 首要目标是运算以及数据吞吐量。
CPU+GPU	异构运算结构，协同工作 深度学习的主流架构，大量开源资源
ARM+FPGA	半定制化的可编程硬件电路 可以根据网络结构和运算加速，效率更高
TPU和各种SoC	全定制化的加速芯片 Google为Tensor Flow专门定制，效率最高

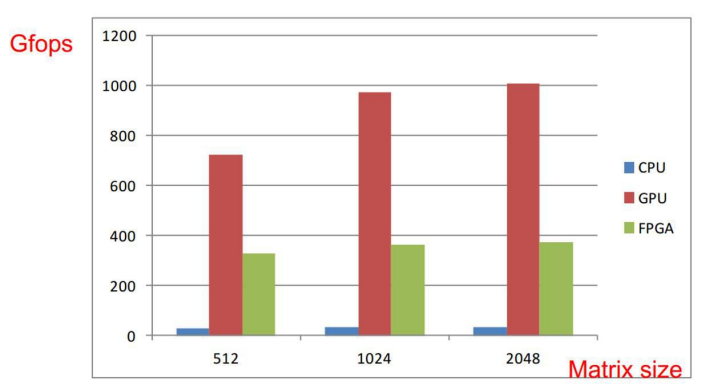
FPGA Vs. GPU

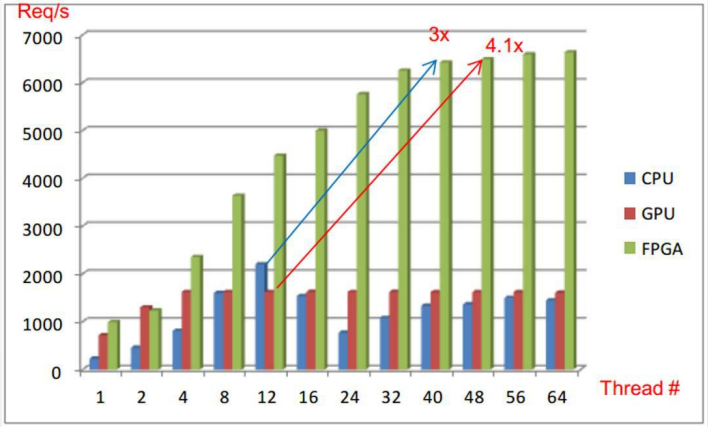
- **Cloud端**

  峰值性能、灵活性、功耗

- **Edge端**

  体积、功耗、实时性





FPGA优势

体积
功耗

动态
配置

结构
定制

可转
ASIC

FPGA芯片结构特别适合模型预测，但是开发难度和门槛限制了其广泛应用。

云端部署

PART 2

业务需求　产品形态　开发案例　机遇挑战

云端的业务需求

全球七大数据中心

阿里巴巴、亚马逊、百度、脸书(Facebook)、谷歌、微软和腾讯

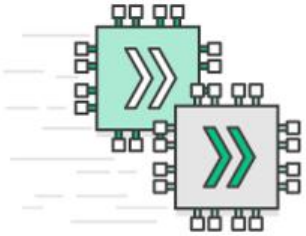
高计算量负载业务

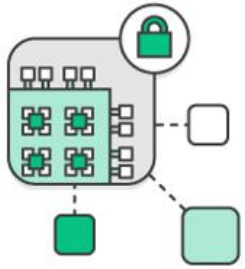
基因测算、金融分析、视频处理、大数据、安全加密、机器学习推断、流媒体、视频直播

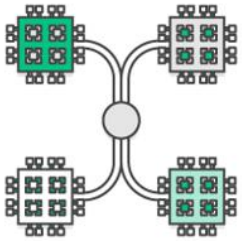
FPGA 的优势

动态配置 | 重复调用 | 可池化或隔离

利用动态可重配置技术，FPGA能在一秒之内快速切换成不同的设计方案，面对下一个工作负载进行硬件优化。

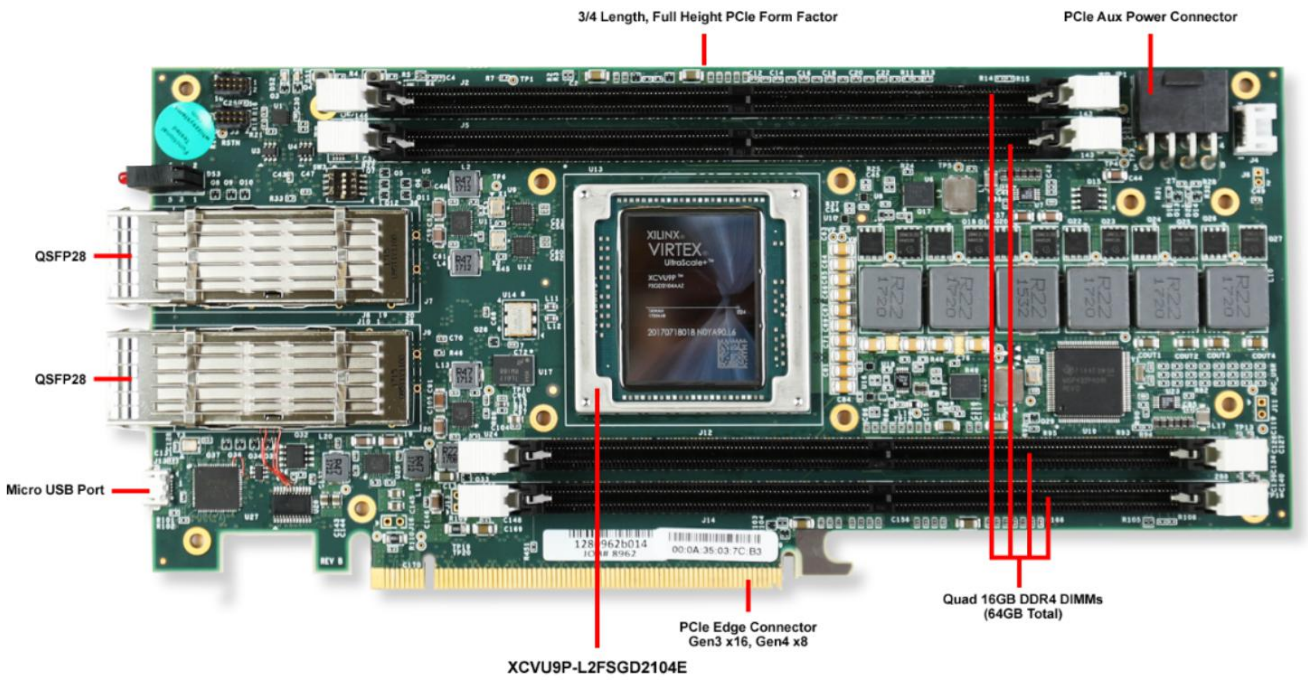






产品形态


XILINX
ALL PROGRAMMABLE.


3/4 Length, Full Height PCIe Form Factor
PCIe Aux Power Connector
QSFP28
QSFP28
Micro USB Port
XILINX
VIRTEX
UltraScale+
Quad 16GB DDR4 DIMMs
(64GB Total)
PCIe Edge Connector
Gen3 x16, Gen4 x8
XCVU9P-L2FSGD2104E

Amazon EC2 F1 实例

在 AWS 云中运行可自定义 FPGA

快速部署自定义硬件加速

借助 F1 实例，您只需在 AWS 管理控制台中单击几下，即可在 AWS 上部署自定义 FPGA。F1 实例可以有一个或多个与其关联的 AFI，让您能够快速灵活地在同一个实例上运行多个加速。此外，F1 实例还为您提供了编程、模拟、调试和编译硬件加速代码所需的易于访问的工具。

可预测的性能

FPGA 通过一种专用的 PCI Express (PCIe) 结构连接到您的 F1 实例，从而使各个 FPGA 能够共用同一内存空间，并能够以高达 12GBps 的速度相互通信。PCI Express 结构与其他网络相隔离，且 FPGA 不会跨实例、用户或账户进行共享。此设计可确保您在使用 FPGA 时只有您的逻辑在其上运行，有助于提供一致的性能。

改变 FPGA 的经济性

借助 F1 实例，您无需自行购买 FPGA 或购买专门的硬件来运行 FPGA，即可为您的工作负载部署硬件加速，从而大幅降低为应用程序部署硬件加速所需的成本。如此一来，您便能够将 FPGA 用于更多工作负载，如基因组学研究和财务风险建模。

可使用您现有的 FPGA 算法

您可以轻松地将您现有的加速算法引入 AWS，并在 F1 实例中使用它们。F1 实例中的 FPGA 和 HDK 中的开发人员工具与硬件加速代码和采用 Verilog 和 VHDL 等常用硬件设计语言或 C 和 Go 等高级语言的设计工具兼容。

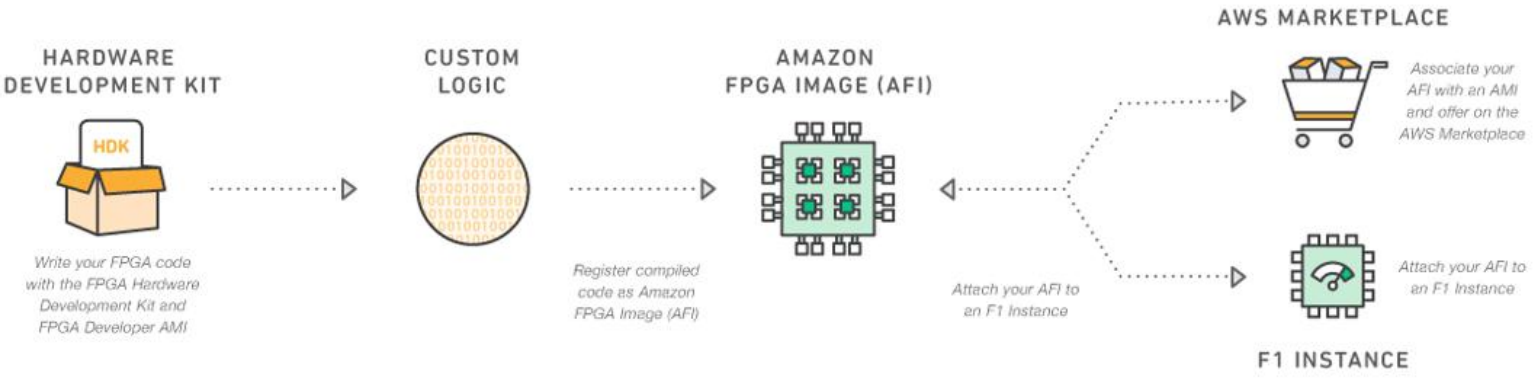
F1实例配置

F1 实例详细信息

实例类型	FPGA 卡	vCPU	实例内存 (GiB)	SSD 存储 (GB)	增强型联网	优化的 EBS
f1.2xlarge	1	8	122	470	是	是
f1.16xlarge	8	64	976	4 x 940	是	是

- Xilinx UltraScale+ VU9P，采用16纳米制程工艺制造。
- 64 GiB ECC保护内存，配合288位总线（四DDR4通道）。
- 专用PCIe x 16 CPU接口。
- 约250万逻辑元素。
- 约6800套数字信号处理（简称DSP）引擎。
- 提供虚拟JTAG接口用于调试。

工作方式



开发　　部署　　提供　　购买

产品应用

财务分析

金融服务行业对多种应用程序的 HPC 功能需求一直在不断增加，包括风险建模和分析、针对安全性的事务分析、高频交易等。金融服务组织可以使用 F1 实例来提高风险建模和分析的准确性，从而显著改进他们的决策制定流程。

大数据搜索和分析

许多大数据应用程序对数据分析和搜索的数量、多样性和速度要求不断提高，导致客户正在寻求硬件加速来满足这些要求。对于这些应用程序，客户可以利用 F1 实例的增强性能来满足其大数据分析和搜索要求。

基因组学研究

必须由基因组学研究人员处理的生物数据的数量和复杂性不断增加，逐步达到了 PB 级范围。研究人员和临床医生必须非常快速地处理这些数据集，以满足医生及其患者的需求。对于此类有时间要求的使用案例，F1 实例是理想的解决方案。

实时视频处理

高性能广播级视频应用程序 (如图片处理、视频分析及视频转码和压缩) 需要使用实时分析功能。F1 实例是满足这些应用程序要求的理想解决方案，且不会影响视频质量。

安全性

F1 实例对于许多安全性应用程序来说非常有用，其中包括防篡改、信息保证和可信关系管理解决方案。

机遇挑战

开启了FPGA as a Service （FaaS）的新时代

国内的BAT和华为也推出类似的服务

为行业加速用户提供新的选项，可能会出现新的商业模式

PART 3

终端部署

CNN的特点　性能优化　SDSoC　PowerAI

终端的场景

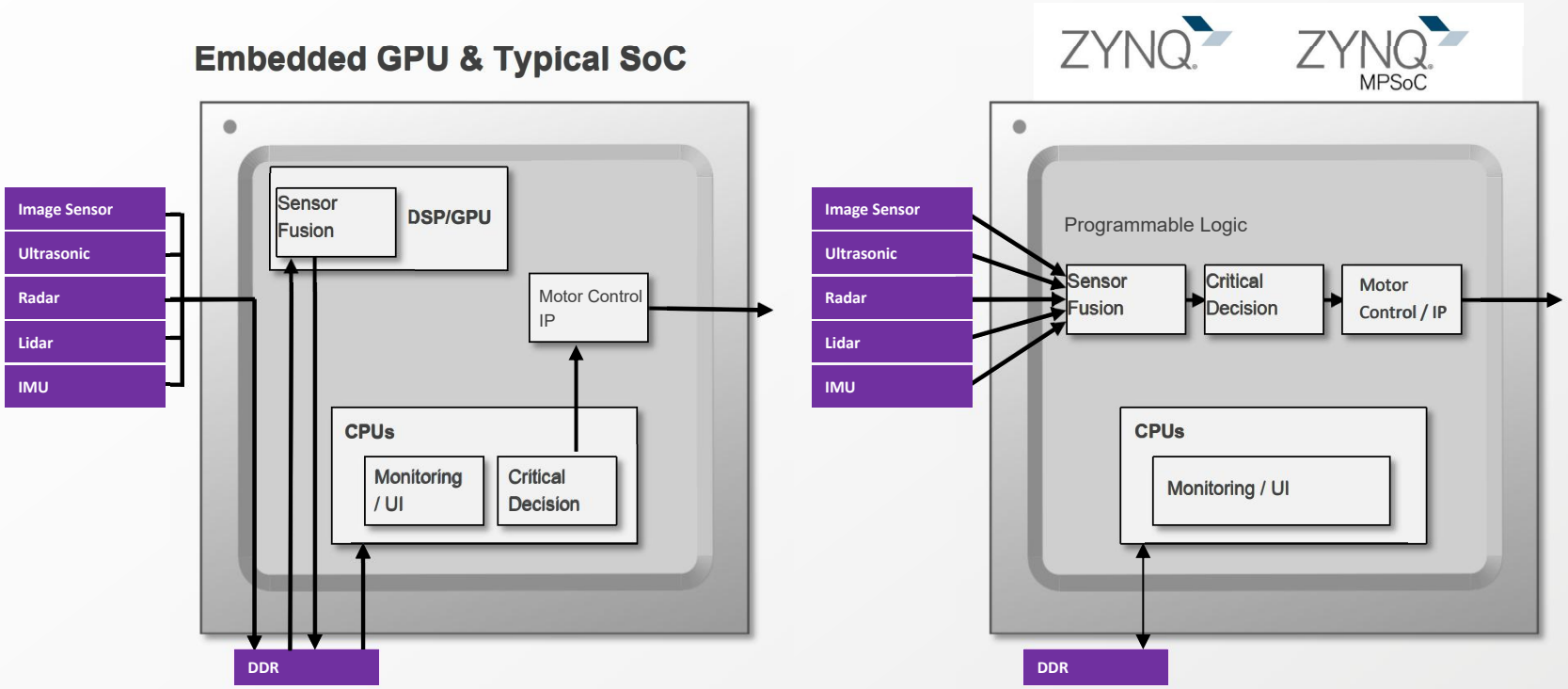
规模化应用： 消费电子、安防、零售 **SoC方案**

差异化应用： 工业自动化、机器人 **FPGA和嵌入式GPU方案**

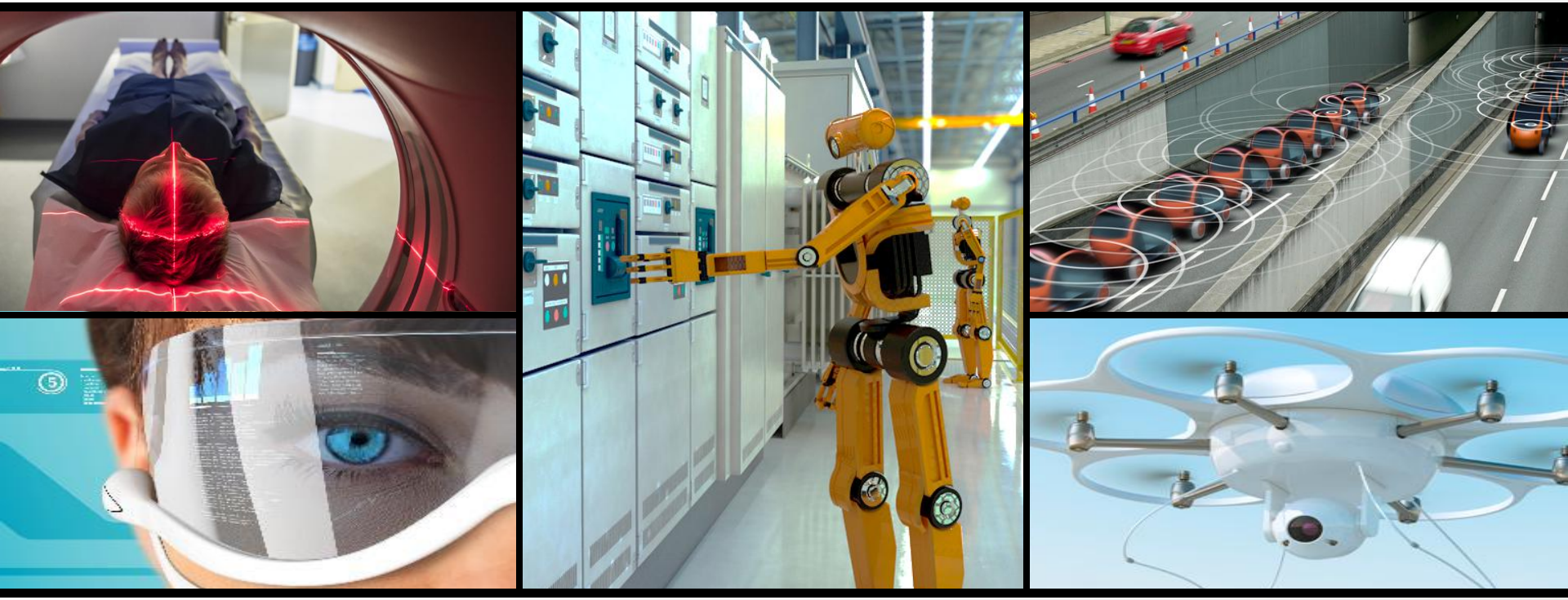
从终端到云端的机器学习推断

机器学习的应用正迅速地扩展至越来越多的终端市场，在终端、在云端或者在那些基于端处理与基于云的数据分析相结合的混合解决方案中。Xilinx 为部署高级高效率神经网络、算法及应用提供各种开发堆栈及硬件平台。

响应时间的优势

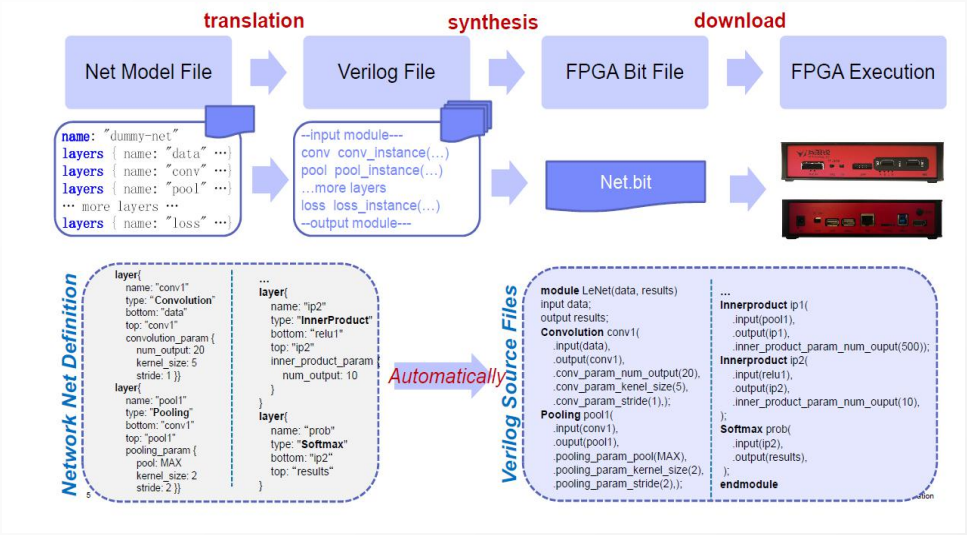

Embedded GPU & Typical SoC
Image Sensor
Ultrasonic
Radar
Lidar
IMU
Sensor Fusion
DSP/GPU
Motor Control IP
CPUs
Monitoring / UI
Critical Decision
DDR
ZYNQ
ZYNQ MPSoC
Image Sensor
Ultrasonic
Radar
Lidar
IMU
Programmable Logic
Sensor Fusion
Critical Decision
Motor Control / IP
CPUs
Monitoring / UI
DDR

基于机器学习的视觉应用



DeepRED：面向低延迟应用解决方案

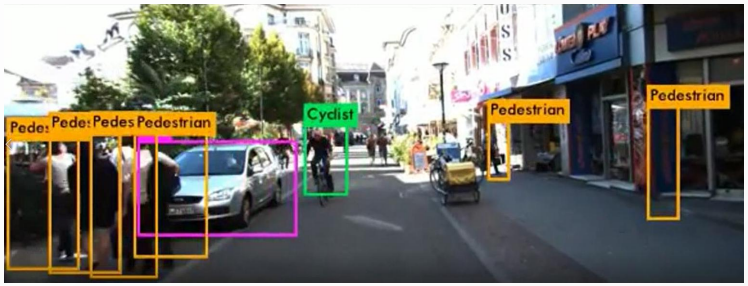
IBM PowerAI Vision

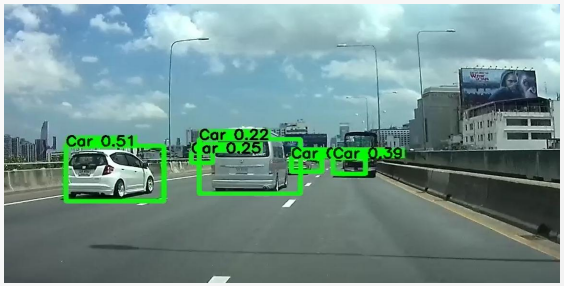


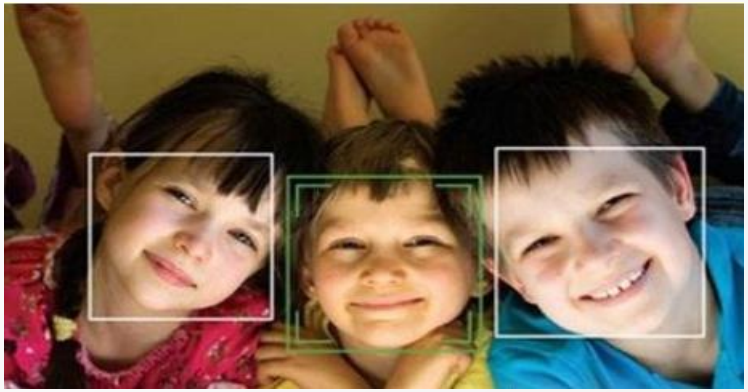
IBM PowerAI工具AccDNN软件（云端）和 Xilinx Vivadao 设计流程

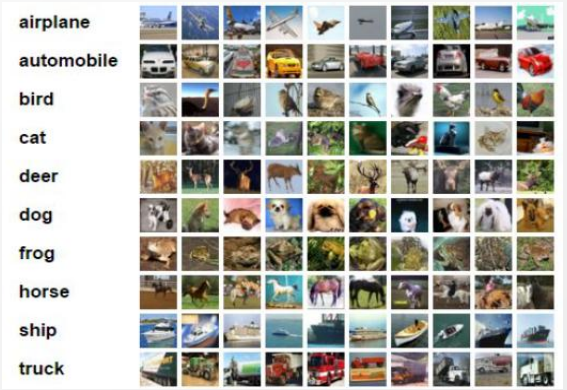
典型应用： 低延迟、快速响应、功耗受限、桌面GPU级算力需求

应用示例


Pedestrian
Cyclist
Pedestrian
Pedestrian


Car 0.51
Car 0.22
Car 0.25
Car 0.39




airplane
automobile
bird
cat
deer
dog
frog
horse
ship
truck

PowerAI Inference Engine service for FPGA embedded system

IBM

Developer

Pre-trained models & weights files by Caffe

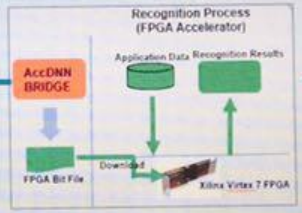
Choose hardware and upload trained DNN result

Estimate FPGA resource and performance by profiling

Change AccDNN parameters and do optimization

SuperVessel Cloud

Generate FPGA IP for DNN



Encrypt the FPGA IP

Download
FPGA SDK from HW provider website

Download
FPGA IP for DNN

Compile FPGA SDK with DNN IP

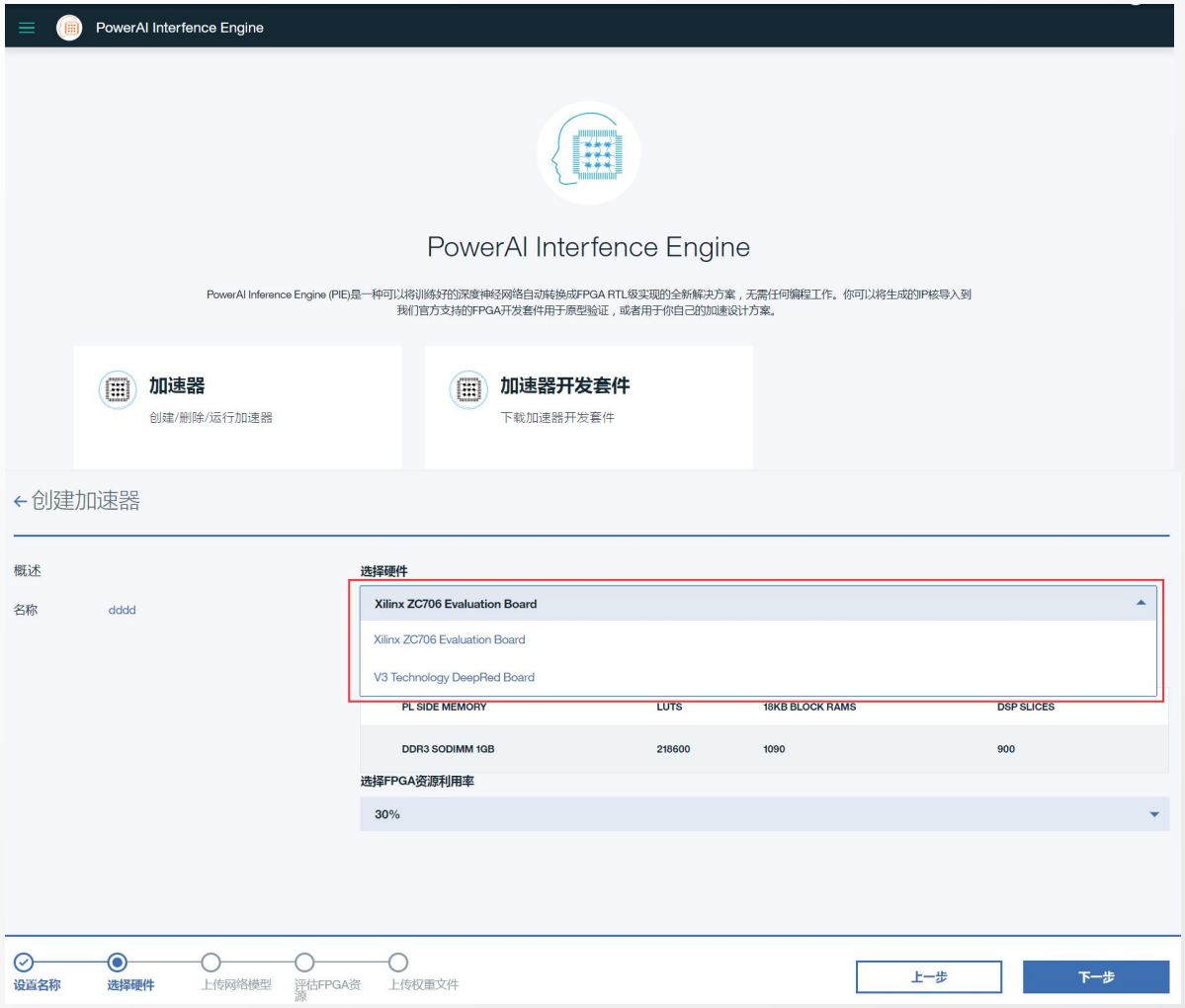
Flash FPGA

Third party embedded system

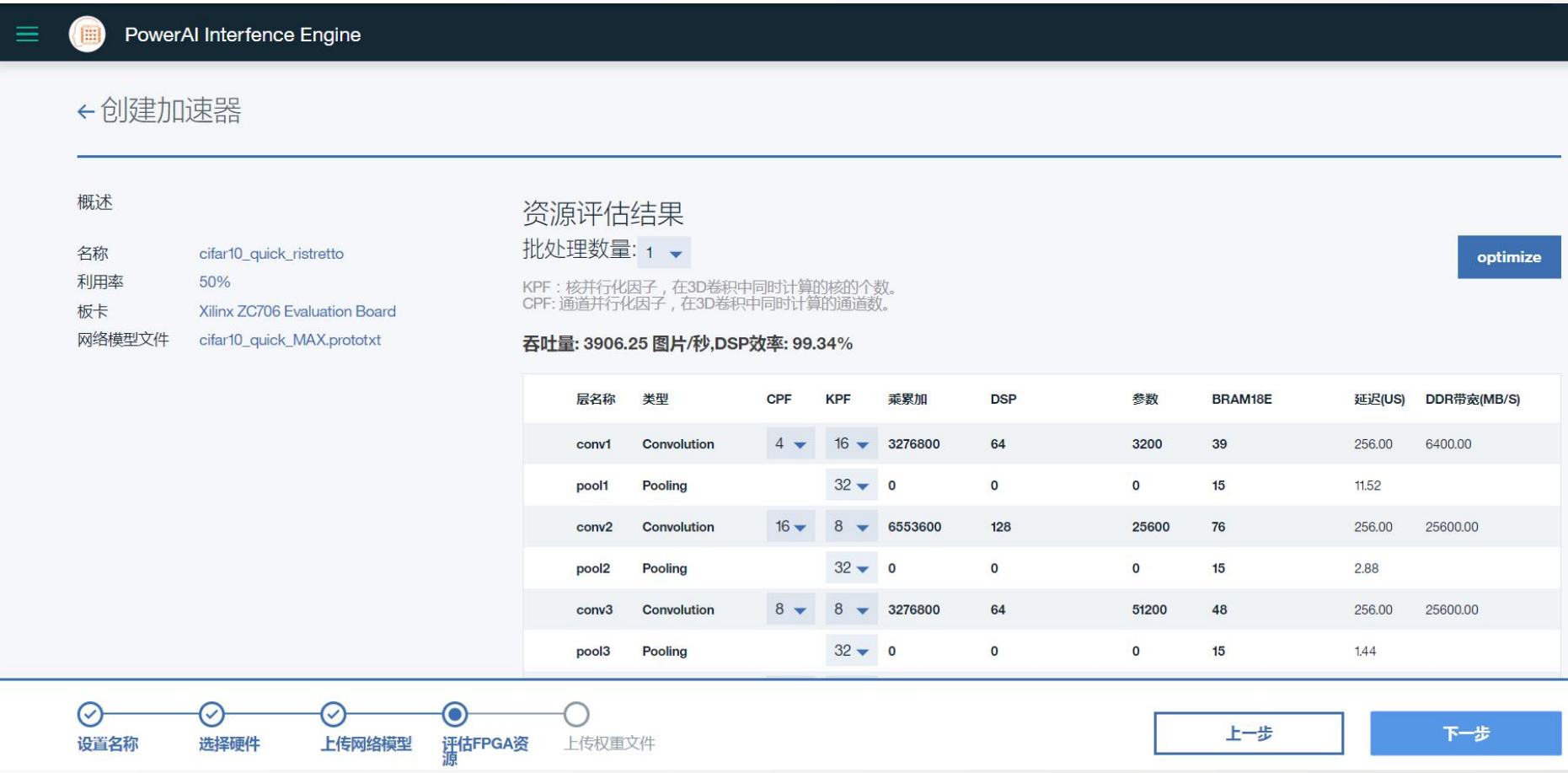
威视锐 V3 Technology. Ltd. INNOVATION FOR RESEARCH

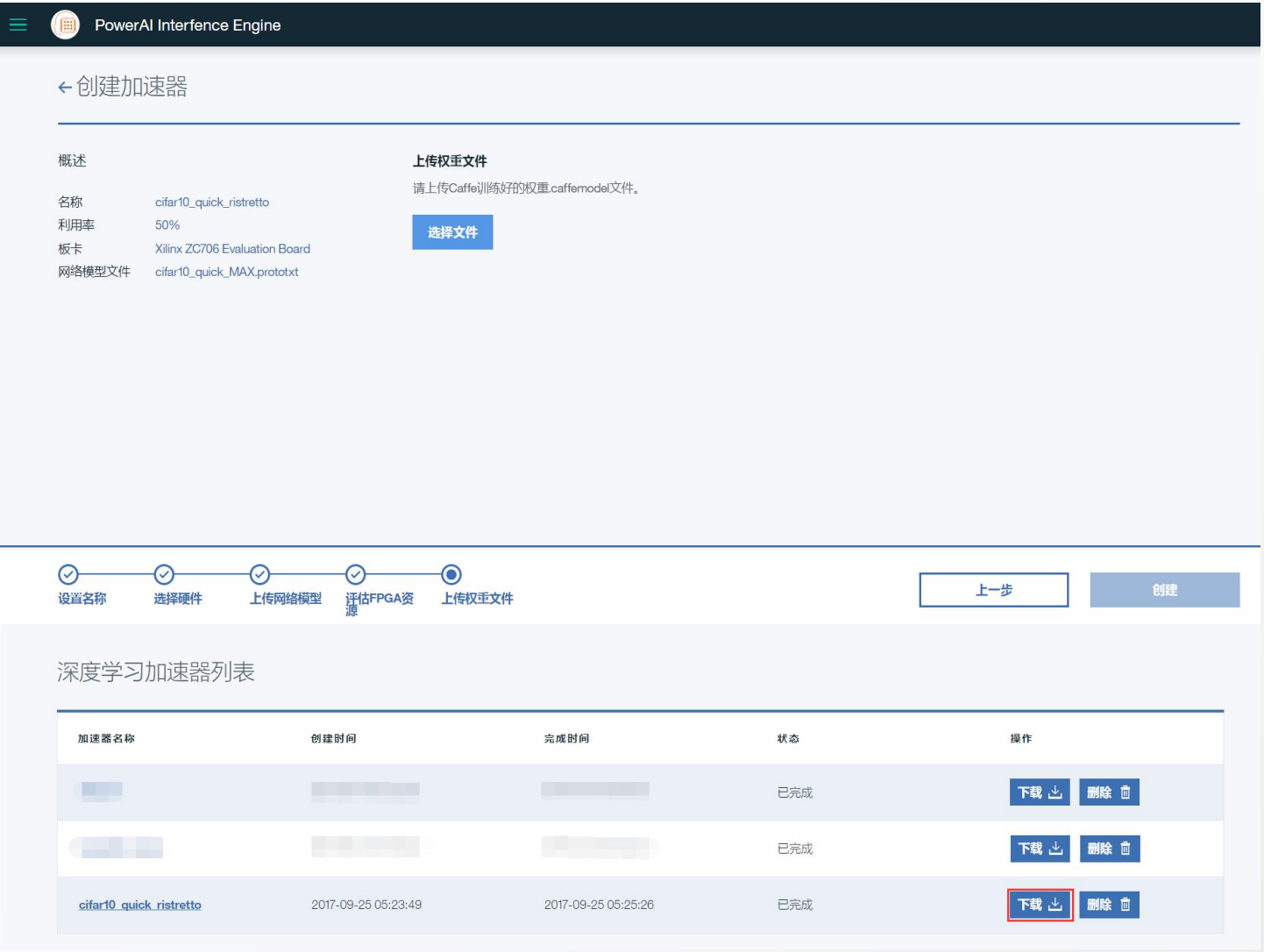
IBM PIE工具开发流程

步骤1： 登录PowerAI 云服务器



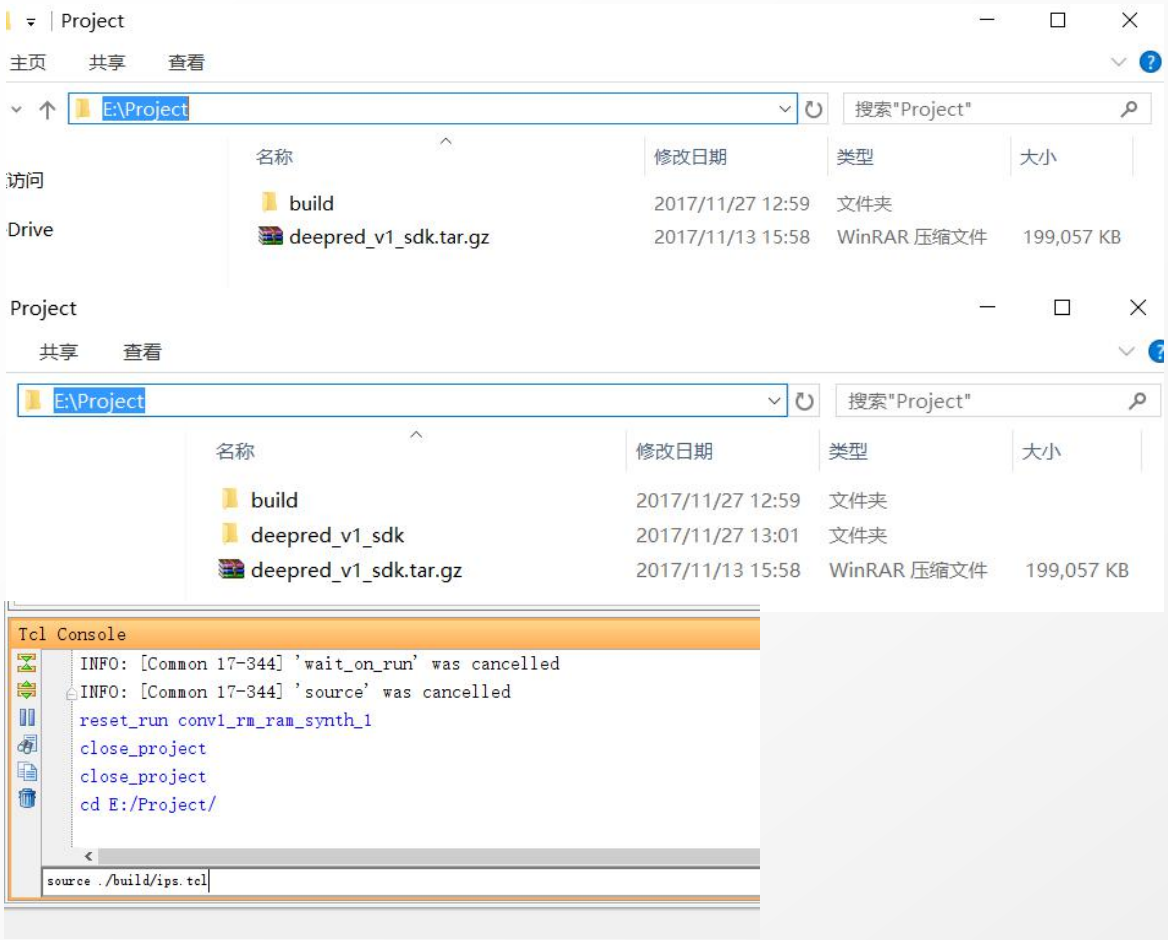
步骤2： 上传网络模型文件，评估资源利用效率

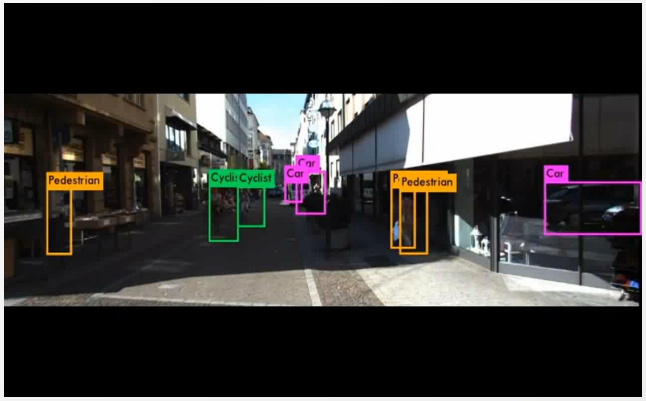




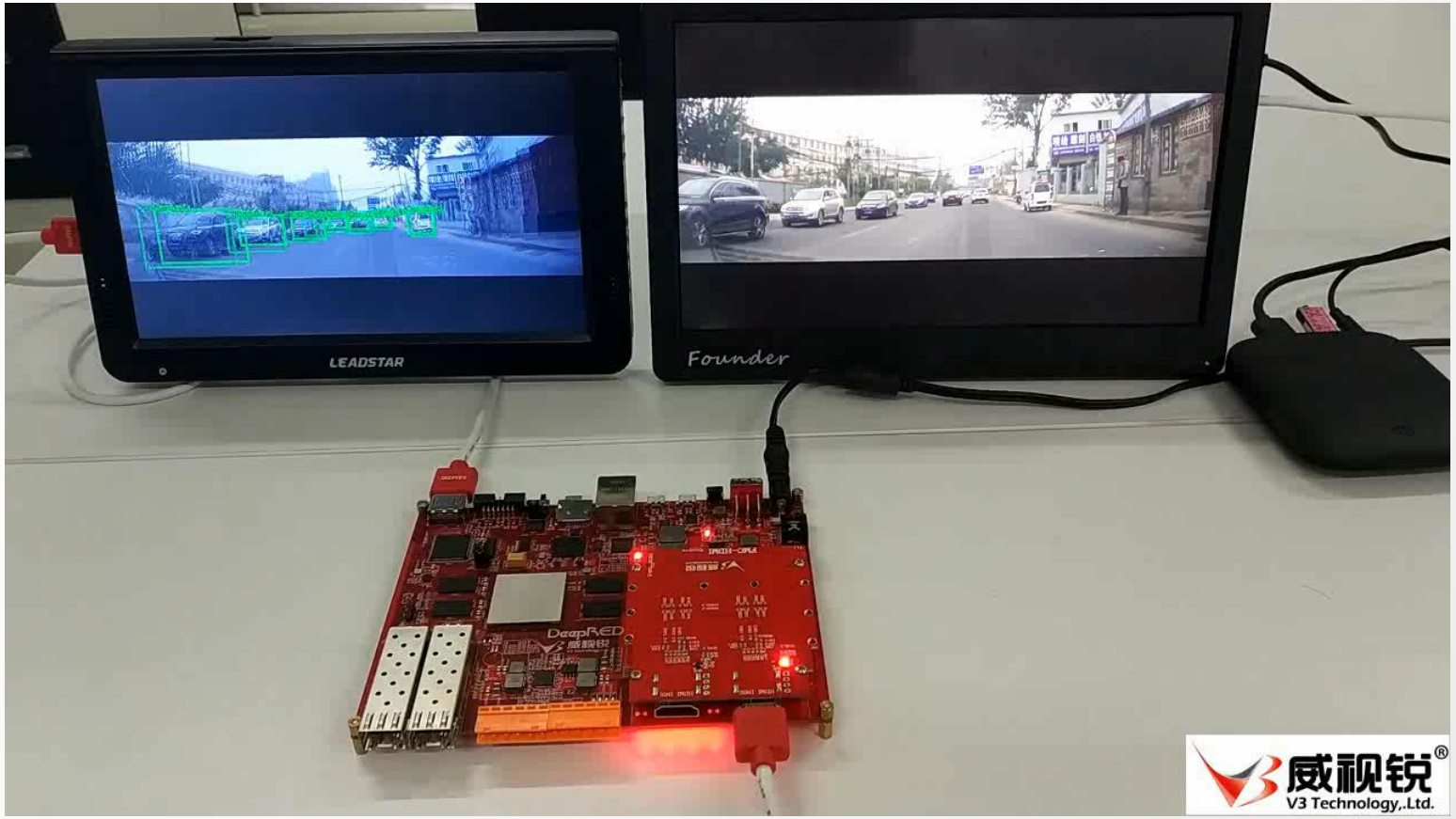
步骤3： 上传权重文件，生产IPcore

利用威视锐SDK，生成FPGA下载文件





实际路测



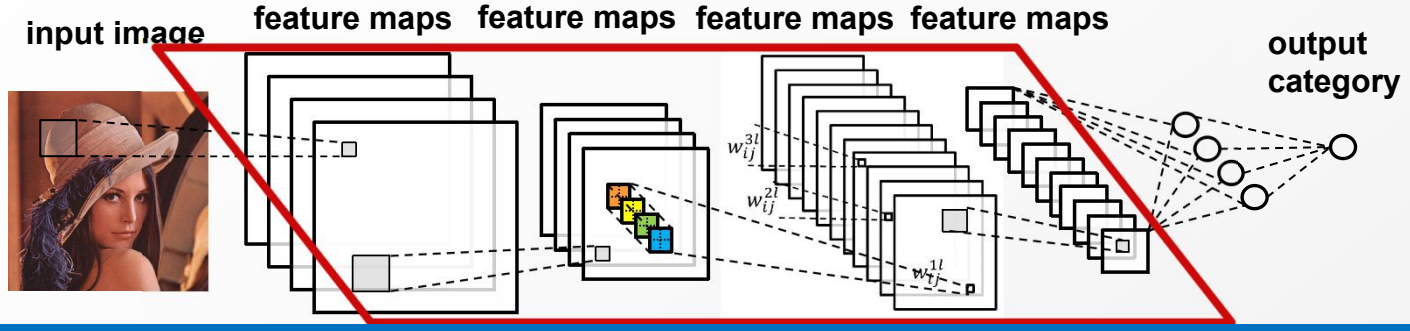
面向工业检测应用解决方案-EagleGo

Xilinx SDSoC软件工具

威视锐 EagleGo: 基于Xilinx Zynq 7020 FPGA，
集成视频输入输出和扩展IO

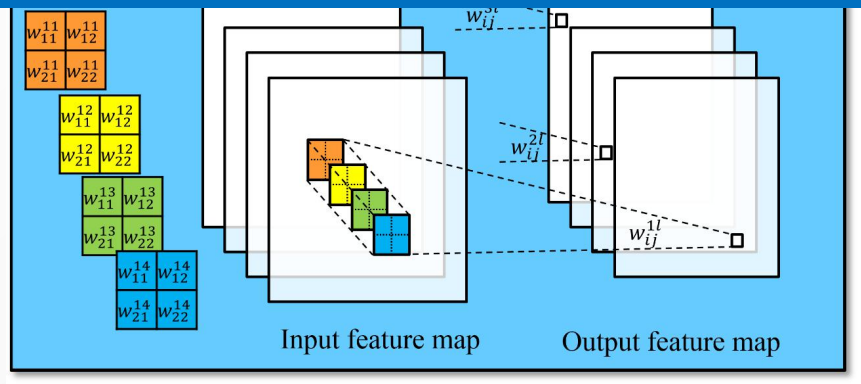
典型场景： 体积小、功耗低、可定制

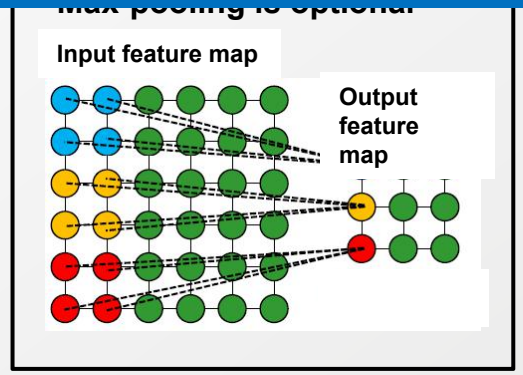
CNNs网络典型结构



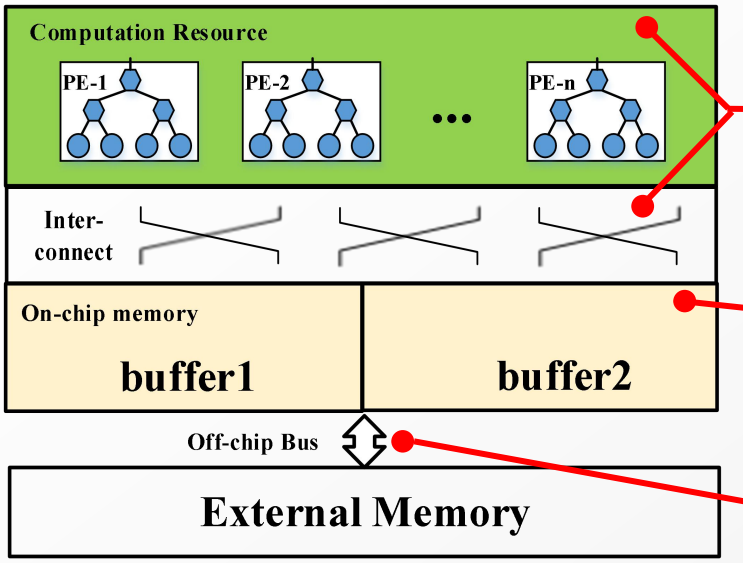
卷积层的计算量累计超过整个算法的90%！

[1] A. Krizhevsky, etc. Imagenet classification with deep convolutional neural networks. NIPS 2012.
[2] J. Cong and B. Xiao. Minimizing computation in convolutional neural networks. ICANN 2014





FPGA上硬件实现的挑战



挑战：资源利用率限制

解决方案：循环展开&流水线

挑战：片上存储器限制

解决方案：循环平铺

挑战：带宽限制

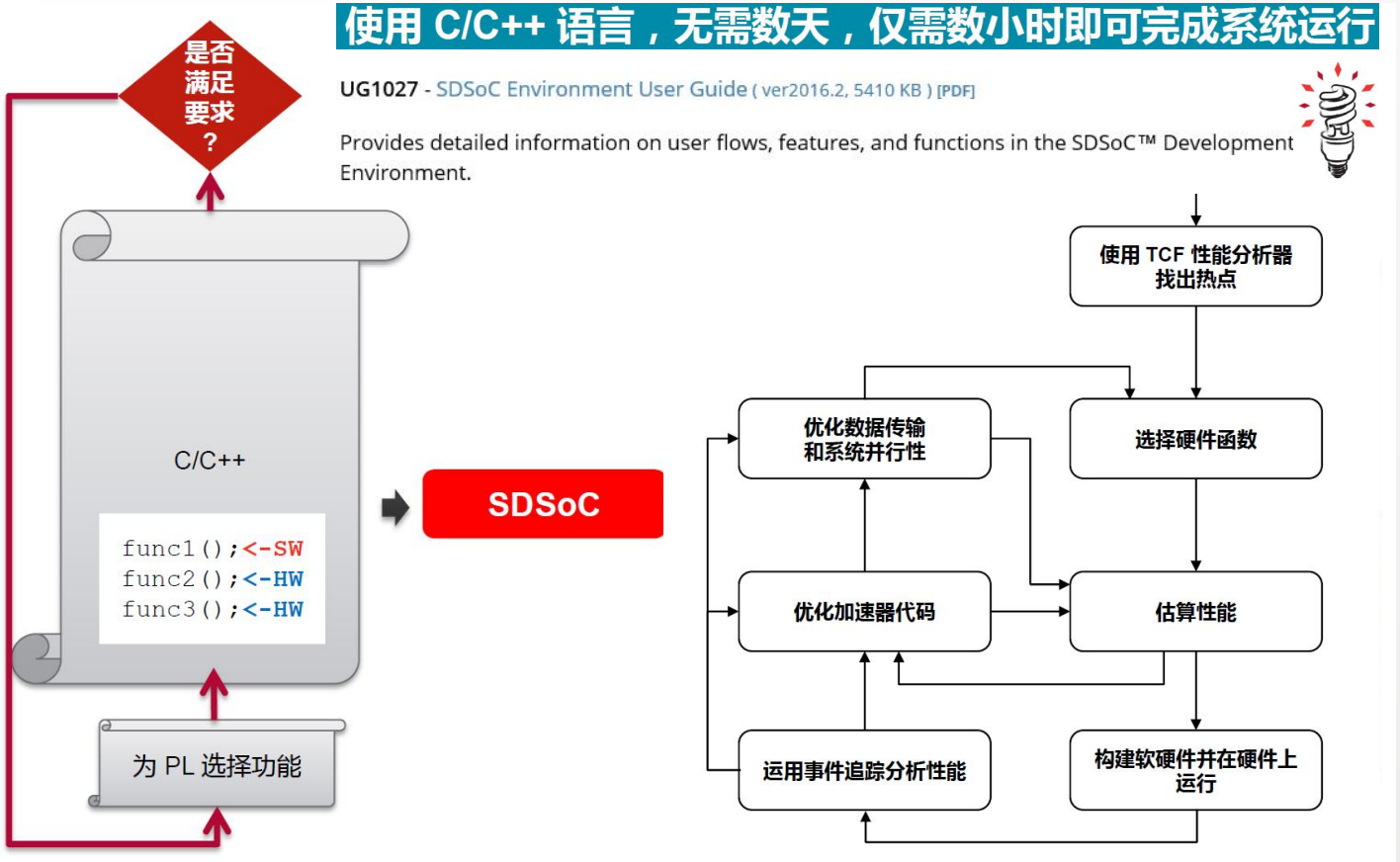
解决方案：数据复用

系统解决方案：联合优化

-- 将计算和通信进行匹配

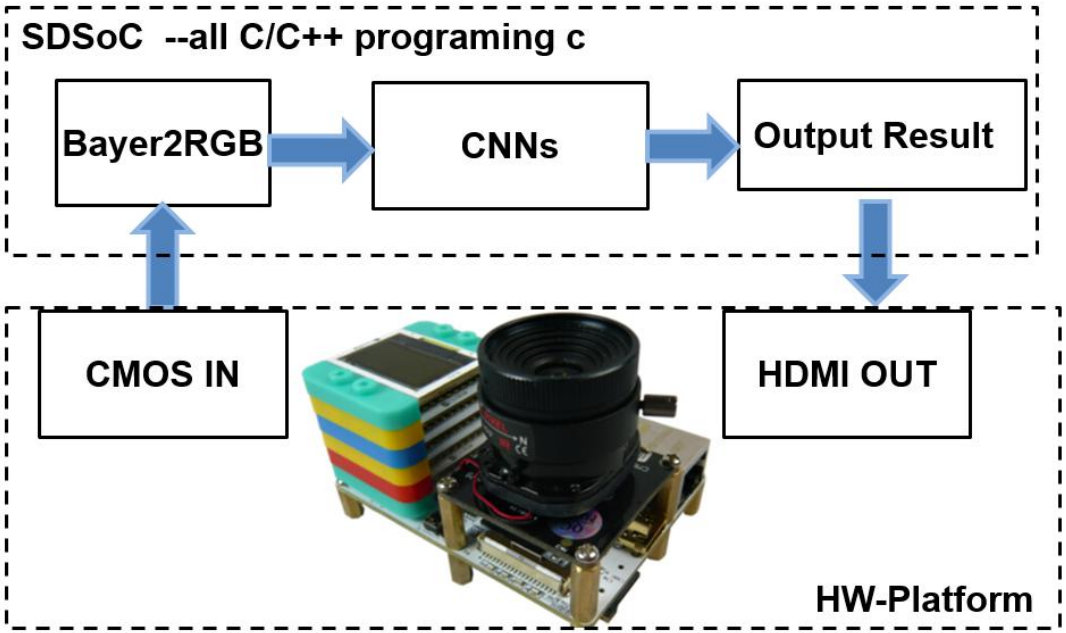
系统优化

SDSoC开发流程


使用 C/C++ 语言，无需数天，仅需数小时即可完成系统运行
UG1027 - SDSoC Environment User Guide (ver2016.2, 5410 KB) [PDF]
Provides detailed information on user flows, features, and functions in the SDSoC™ Development Environment.
是否满足要求？
C/C++
func1();<-SW
func2();<-HW
func3();<-HW
为 PL 选择功能
SDSoC
使用 TCF 性能分析器找出热点
优化数据传输和系统并行性
选择硬件函数
优化加速器代码
估算性能
运用事件追踪分析性能
构建软硬件并在硬件上运行

实现案例：物体分类摄像头

```
float result[10]={0};
// resize 1080P to 32X32  --||
zoom(pVideoBuffer, tgt_img);
conv1(tgt_img,rst1);
conv2(rst1,rst2);
//image process
identi_cal((float*)rst2,result);
//classification
int item_class=0;
int i;
float Pmax = 1.0;
for(i=0;i<10;i++)
    {
       if(result[i]>Pmax)
         {
           Pmax=result[i];
           item_class=i;
         }
     }
```



实现案例：SDSoc技术实现卷积核加速

```
14    int i,j,x,y;
15    for (i = 0; i < 28; i++) {
16      for (j = 0; j < 28; j++) {
17  #pragma HLS PIPELINE
18        float result = 0;
19        for (x = 0; x < A_NCOLS; x++) {
20          // multiply accumulate broken into individual operators
21          // so that AutoESL can infer two FP operators
22  #pragma HLS UNROLL
23        for(y = 0; y < A_NCOLS; y++) {
24           float product_term = in_A[x][y] * in_B[x+i][y+j];
25           result += product_term;
26        }
27        }
28        out_C[i*28+j] = result;
29      }
30    }
31  }
```

功能演示：基于Caffe的物体分类（ZYNQ 7020）

AI领域合作伙伴


Microsoft


IBM


Bai du 百度


megvii


DEEPHI TECH
深 鉴 科 技


DEEPGLINT
格 灵 深 瞳

创新应用方向


更 多 信 息
威视锐：http://www.v3t.com.cn
V3学院：http://www.v3edu.org
V3学院微信课堂（服务号）