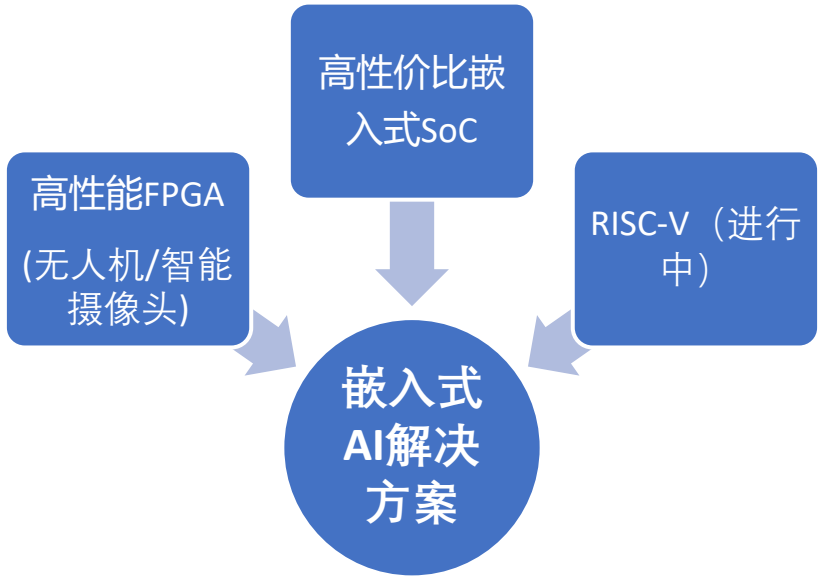
PerfXLab
澎 峰 科 技

针对RISC-V的轻量级深度学习推理框架InferXLite

张先轶，向春阳，张宾，褚双伟
PerfXLab 澎峰科技
xianyi@perfxlab.com

2019中国RISC-V论坛

澎峰科技—嵌入式AI解决方案

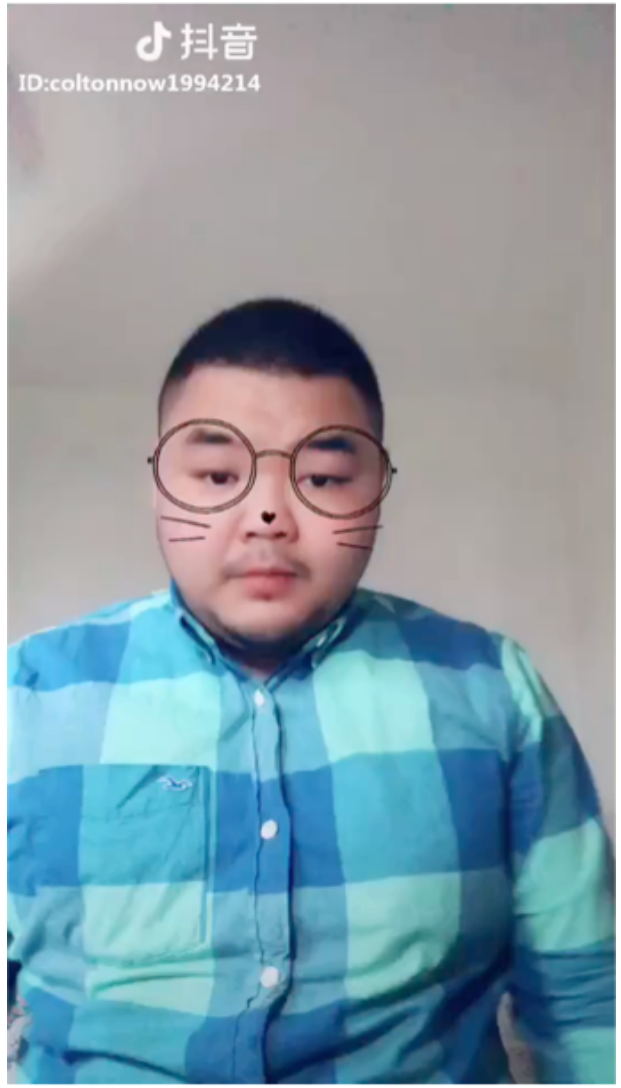


合作伙伴

AVIC
CASC 中国航天 CASC
HRG 哈工大机器人集团 HIT ROBOT GROUP
REG 中国中铁
北京大学
北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY
seeed studio The IoT Hardware Enabler

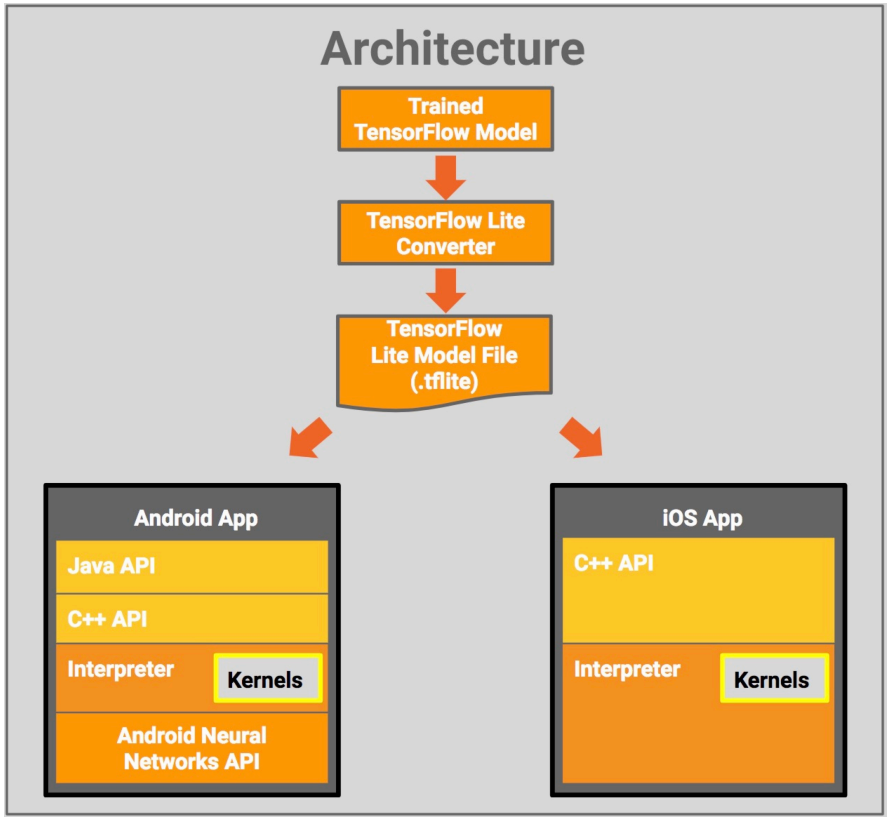
嵌入式深度学习

- 服务器：模型训练
- 嵌入式设备：模型推理
  - 实时性
  - 可靠性
  - 数据隐私
- 以美颜类App为例
  - 人脸检测
  - 人脸关键点



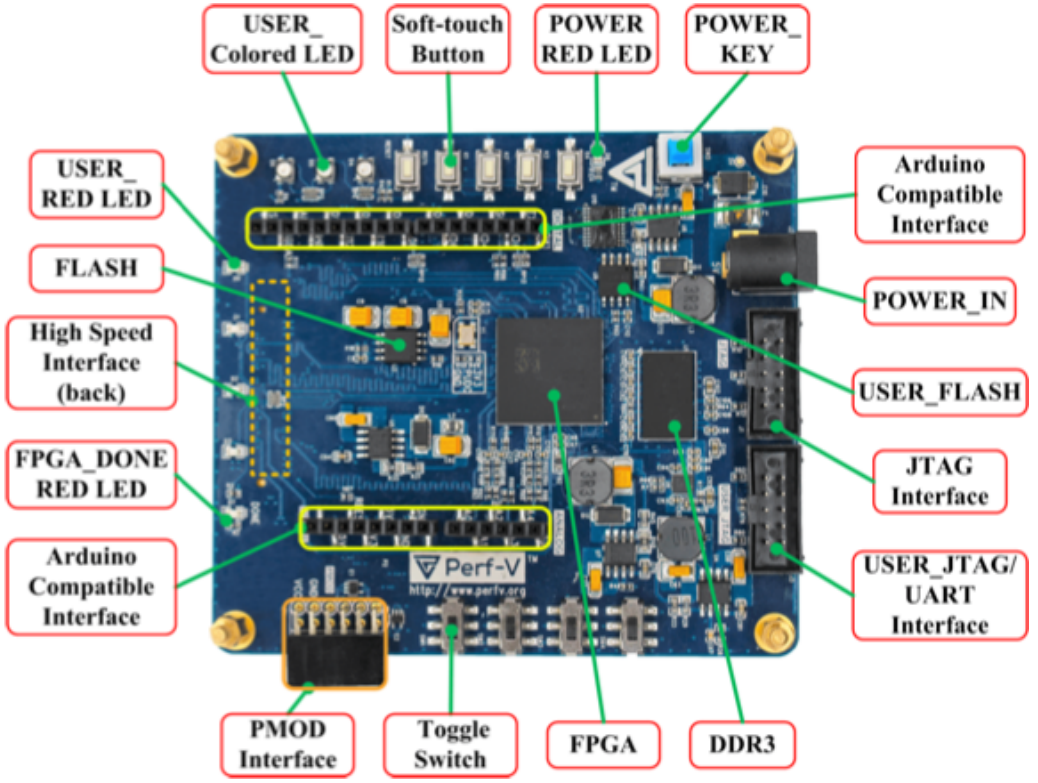
嵌入式深度学习推理框架

- 众多推理框架
  - 腾讯 NCNN
  - ARM OpenAI Tengine
  - TensorFlow Lite
  - 百度 Paddle Lite
  - …
- 特点
  - 开发语言：C++，Tensor类
  - 硬件平台：ARM CPU，嵌入式GPU等
  - OS：Android，Linux



RISC-V移植深度学习推理框架

- RISC-V MCU级别
  - 以Perf-V开发板为目标
- 已有框架移植困难
  - 资源受限
- 尝试 “新轮子”

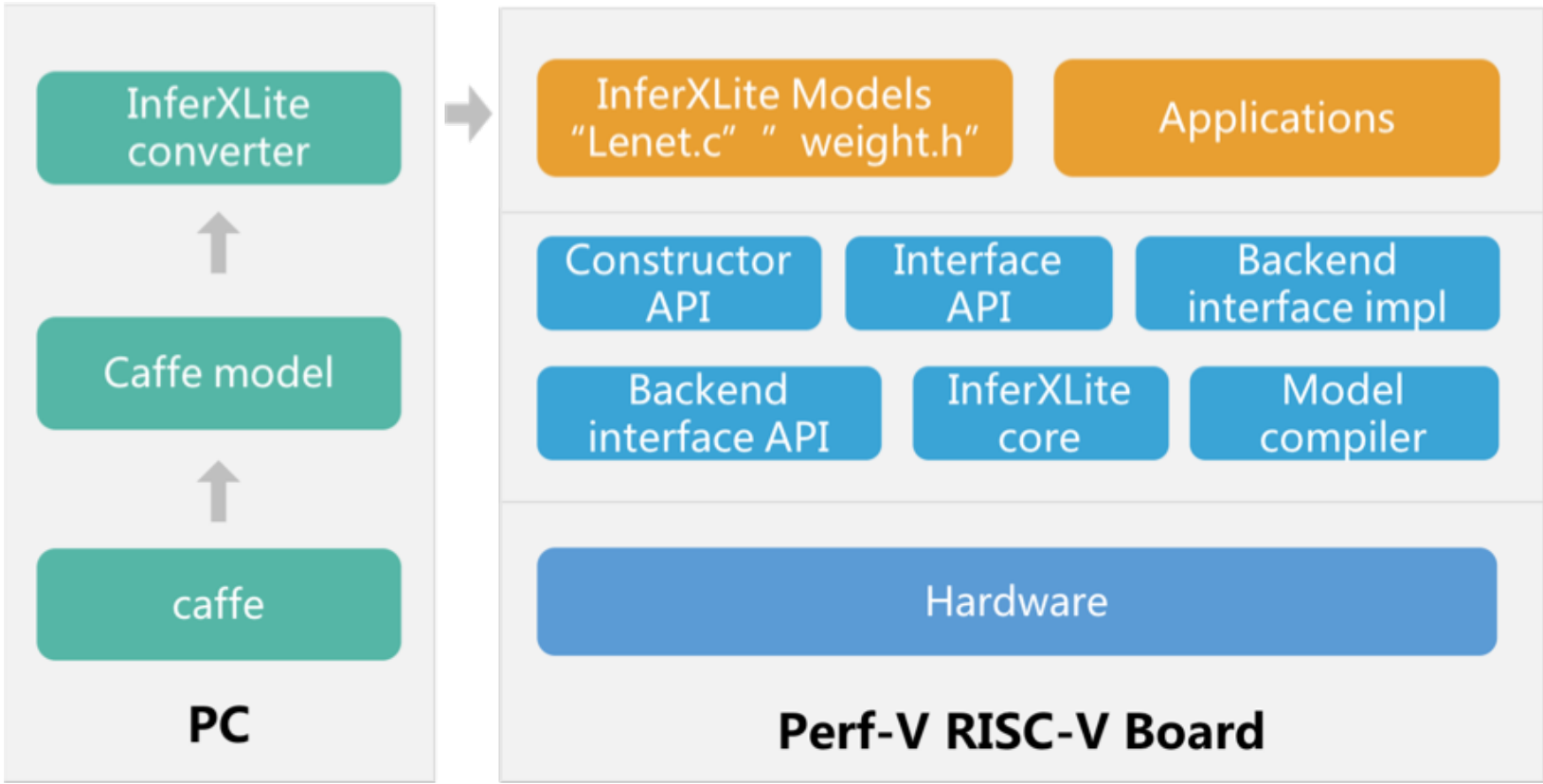


Perf-V RISC-V开发板

http://perfv.org/

RISC-V轻量级深度学习推理框架InferXLite

- C语言实现
  - 降低依赖
- 支持Caffe模型转换
- 模型表示层
  - 如何存储模型
- 接口层
- Backend实现层
  - 卷积实现
  - Pooling实现等

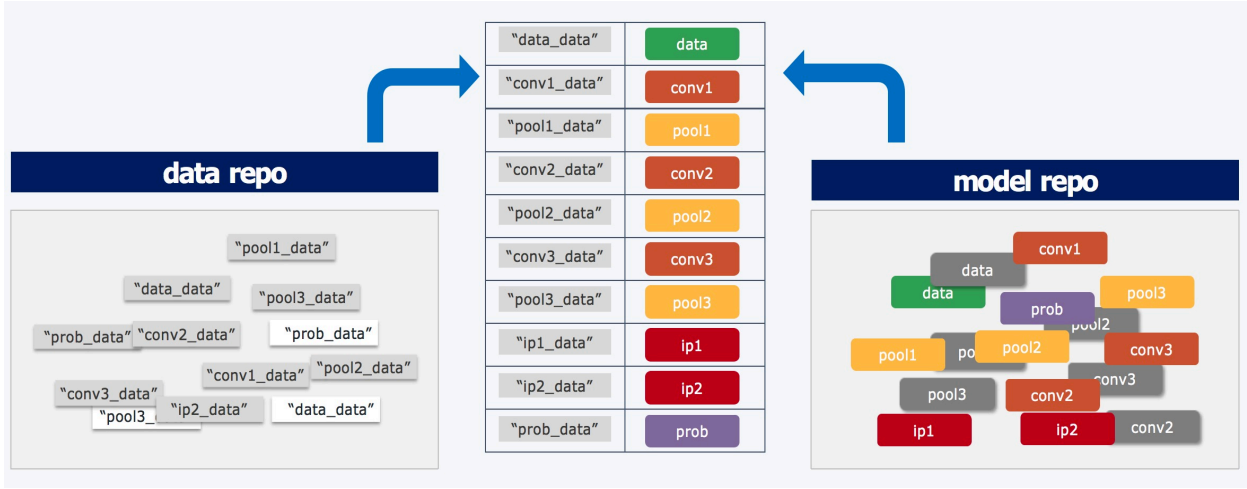


InferXLite RISC-V平台模型表示

- Layer算子
  - C函数调用
  - 函数参数:Kernel size等
- 网络的Layer间依赖
  - 函数调用顺序
  - 通过名称查找依赖
- 权重文件
  - 全局数组
- 模型层间优化
  - Layer合并等，模型转化时完成

```
#include "interface.h"
#include <stdbool.h>

struct inferx_handler;
void LeNet(char* path, char* model, char* data, int *shape, int nshape,
    void* pdata, void** pout,int *len, struct inferx_handler *hd)
{
    inferx_net_preprocess(data,model,hd);
    inferx_input(shape,pdata,"data_data","data",model,data,hd);
    inferx_convolution(1,12,3,3,1,1,0,0,1,1,1,true,false,"data_data","conv1_data","conv1",model,data, 1, 0.0, hd);
    inferx_pooling(2,2,2,2,0,0,"MAX","conv1_data","pool1_data","pool1",model,data,hd);
    inferx_convolution(12,24,3,3,1,1,0,0,1,1,1,true,false,"pool1_data","conv2_data","conv2",model,data, 1, 0.0, hd);
    inferx_pooling(2,2,2,2,0,0,"MAX","conv2_data","pool2_data","pool2",model,data,hd);
    inferx_convolution(24,48,3,3,1,1,0,0,1,1,1,true,false,"pool2_data","conv3_data","conv3",model,data, 1, 0.0, hd);
    inferx_pooling(2,2,2,2,0,0,"MAX","conv3_data","pool3_data","pool3",model,data,hd);
    inferx_inner_product(48,96,true,false,"pool3_data","ip1_data","ip1",model,data,hd);
    inferx_inner_product(96,10,true,false,"ip1_data","ip2_data","ip2",model,data,hd);
    inferx_softmax(1,"ip2_data","prob_data","prob",model,data,hd);
    inferx_finalize("LeNet",hd);
    return;
}
```

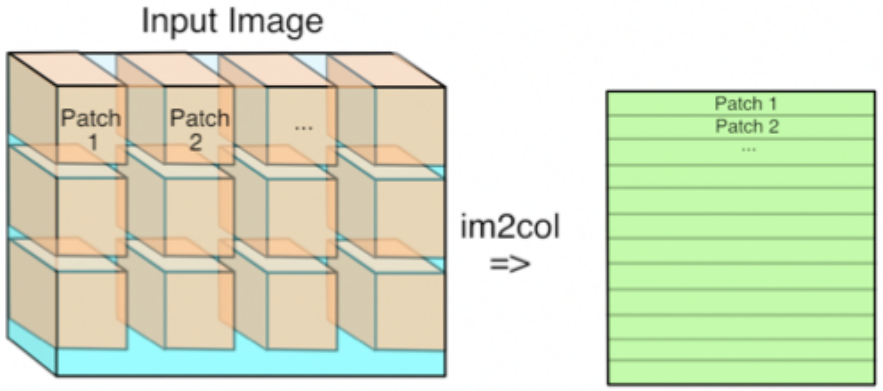


InferXLite RISC-V平台Backend层实现

- 支持算子
  - 卷积
  - 池化
  - RELU激活
  - 全连接
  - Softmax
  - Concat
  - …

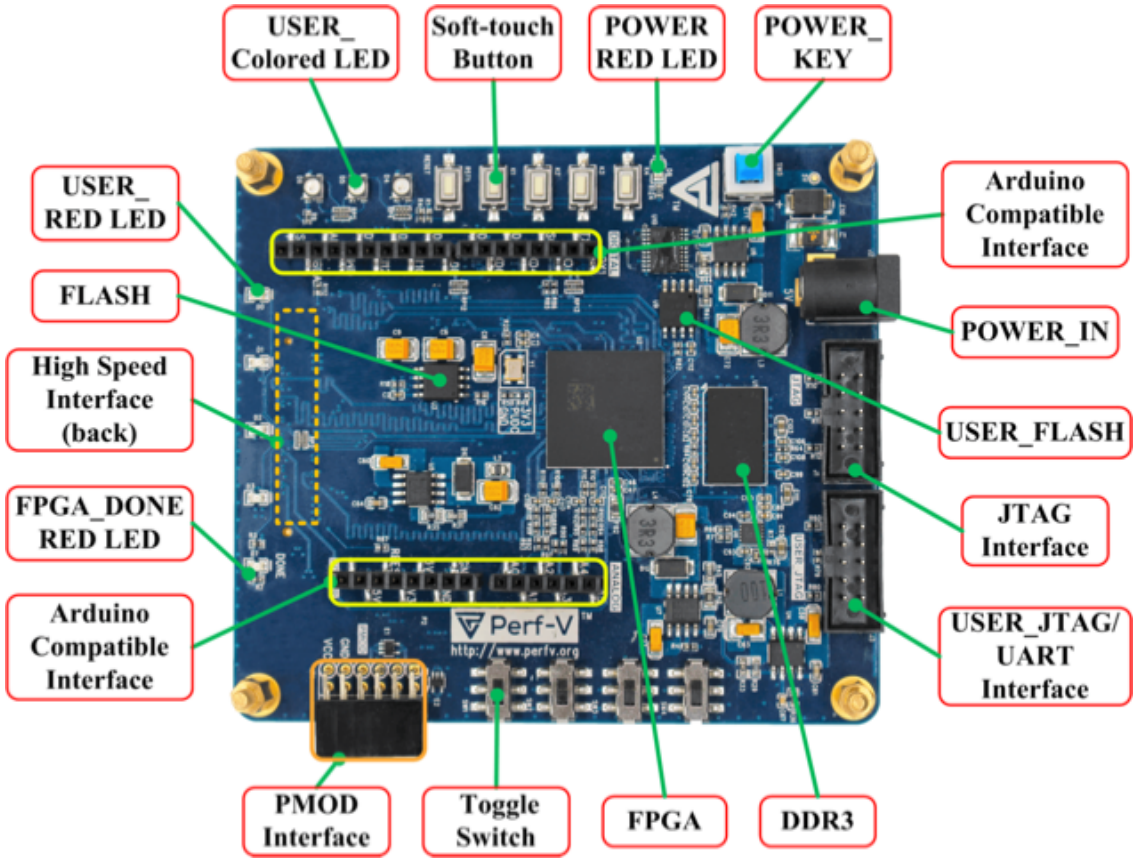
InferXLite RISC-V平台卷积实现

- 基于矩阵乘法
  - 可以利用BLAS实现
  - 利用加速IP
- 直接法
  - 按照卷积定义
  - Memory临时空间需求少
  - InferXLite RISC-V使用
- Winograd法
  - 特殊规格3x3卷积
  - 访存需求更大



实验平台

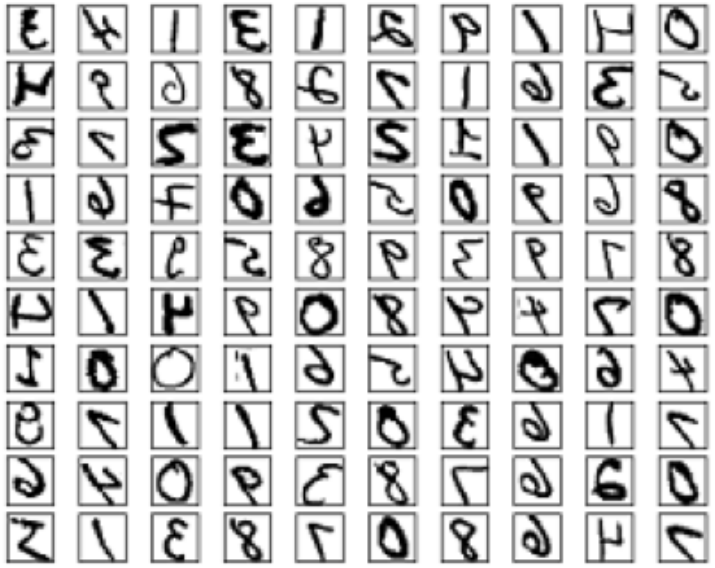
- Perf-V RISC-V开发板
  - Xilinx Artix-7 100T
  - 蜂鸟 E200 软核
    - RV32IMAC
    - 主频约 20MHz
  - DTCM: 512KB
  - ITCM: 32KB



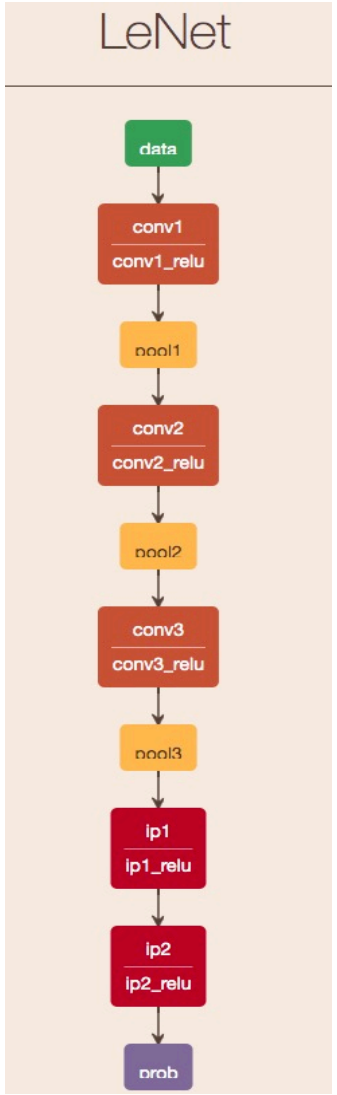
Perf-V RISC-V开发板

http://perfv.org/

实验模型: LeNet

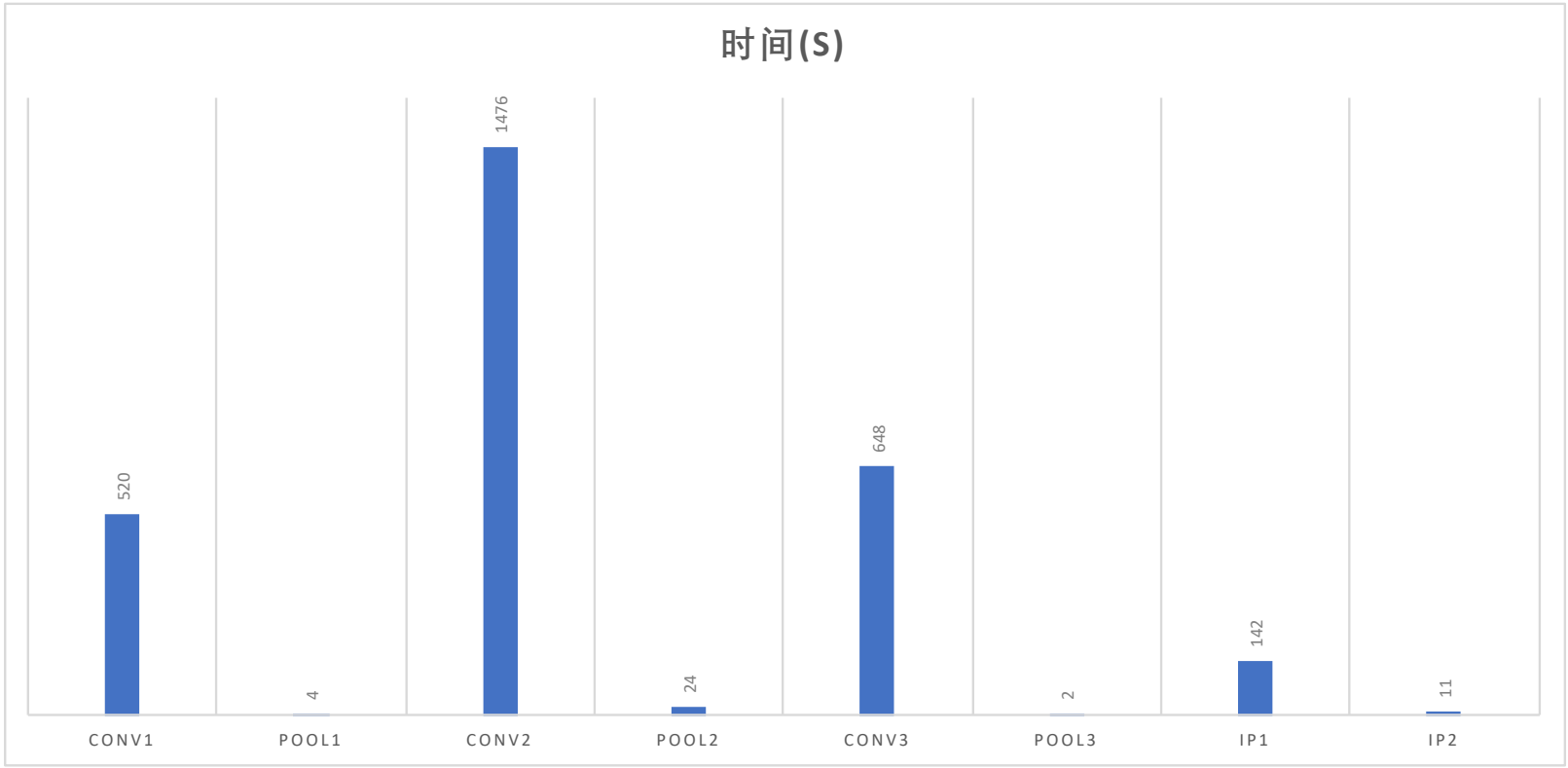


layer	Kernel Size	Stride	Output channel
conv1	3x3	1	12
pool1	2x2	2	N/A
conv2	3x3	1	24
pool2	2x2	2	N/A
conv3	3x3	1	48
pool3	2x2	2	N/A
Ip1	N/A	N/A	96
Ip2	N/A	N/A	10



实验结果

- 推理结果正确
- 运行时间过长
  - 模型使用FP
  - 而硬件只有软浮点
  - Backend实现简单

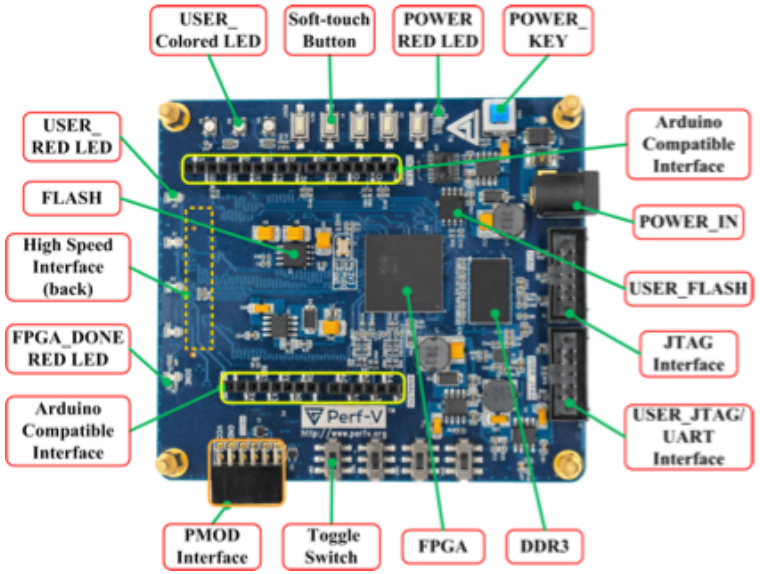


小结

- InferXLite深度学习推理框架移植
  - 模型表示
  - Backend实现
  - 不依赖于额外软件
- RV32IM
  - 能跑，结果正确
  - 速度？
- 下一步工作
  - 量化模型，使用int指令
  - SIMD扩展？
  - 更高主频芯片？

张先轶@PerfXLab

北京 海淀



Perf-V RISC-V开发板

http://perfv.org/

扫一扫上面的二维码图案，加我微信