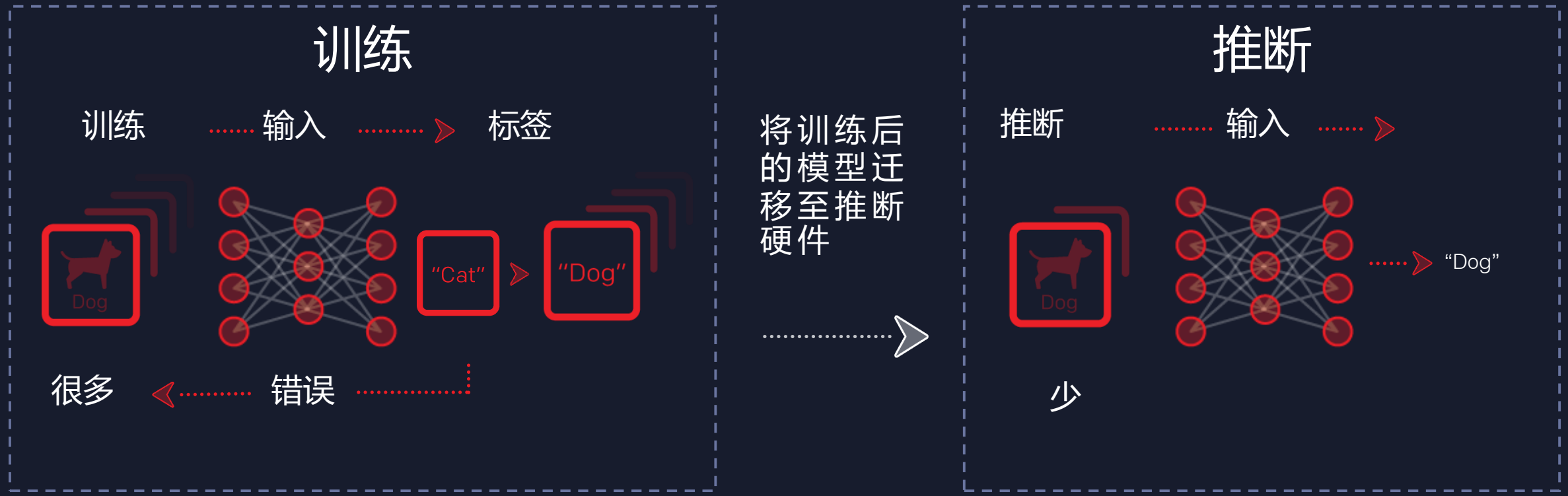
AI 加速

Salil Raje
执行副总裁 - 软件及 IP 产品

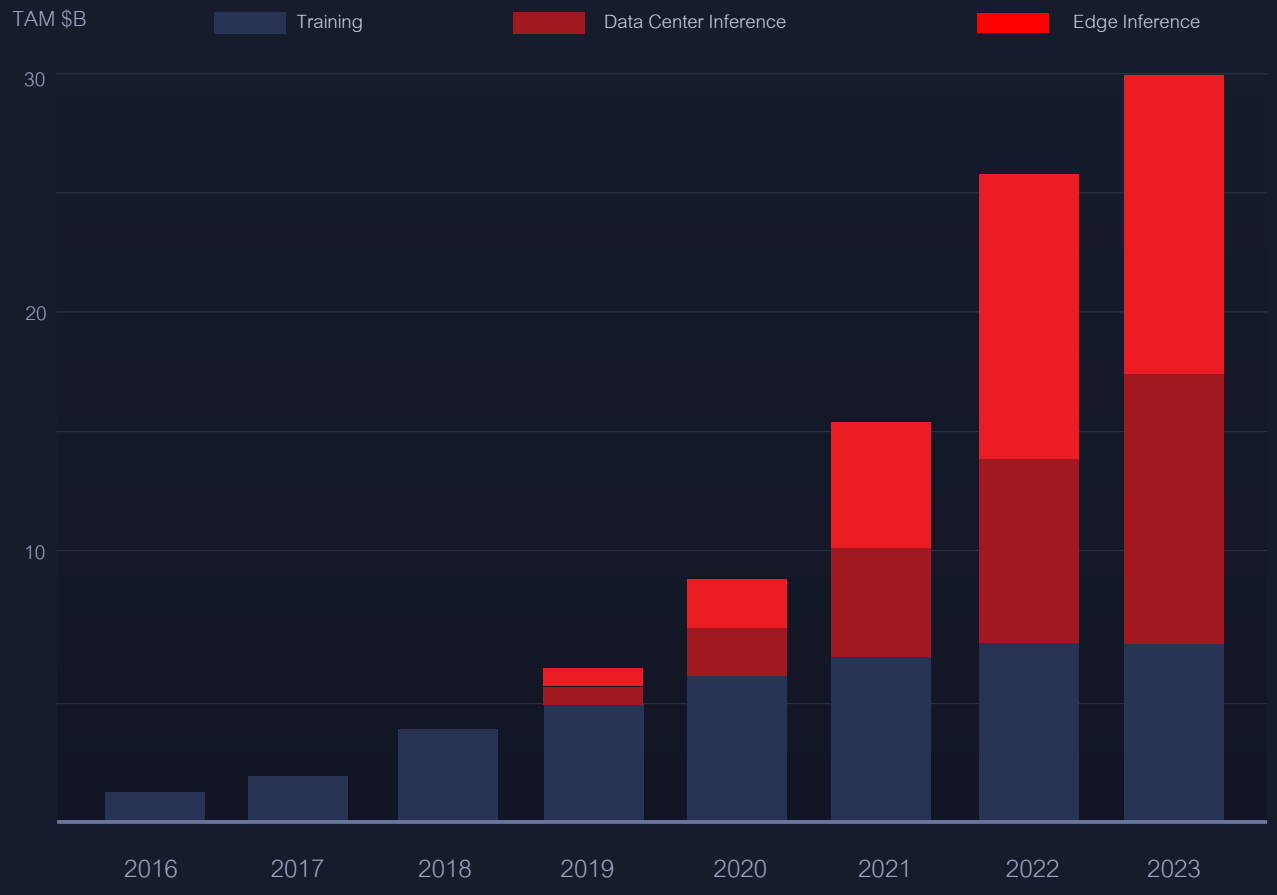
拥抱所有的开发者！

数据科学家	Frameworks: Python, APIs	DEEPHI 深鉴科技 Caffe mxnet FFMPEG TensorFlow
SaaS 开发者	FaaS Platform	aws HUAWEI Aliyun Alibaba Cloud Computing NIMBIX
应用开发者	SDX: C++, OpenCL, Libraries	Linux freeRTOS Xen
嵌入式开发者	Embedded Software: MPSoC	
有一定硬件基础的软件开发者	HLS: C++ IP Functions	
系统集成商	IP Integrator: System Integration	
硬件开发者	Vivado Design Suite: RTL Full Design	

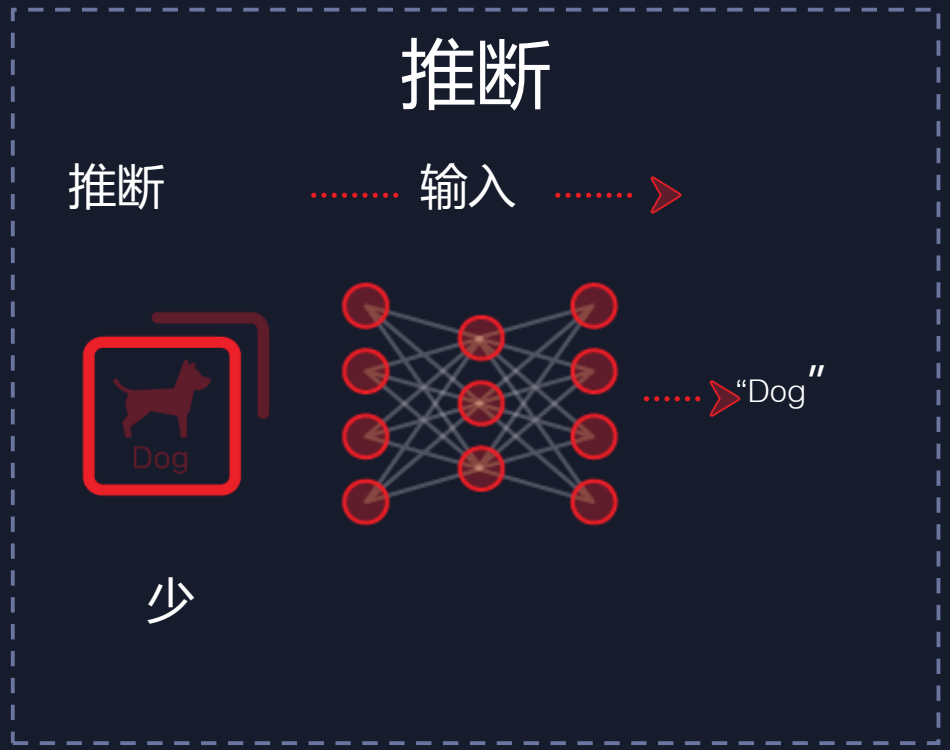
训练 vs. 推断



推断需求预计将持续攀升

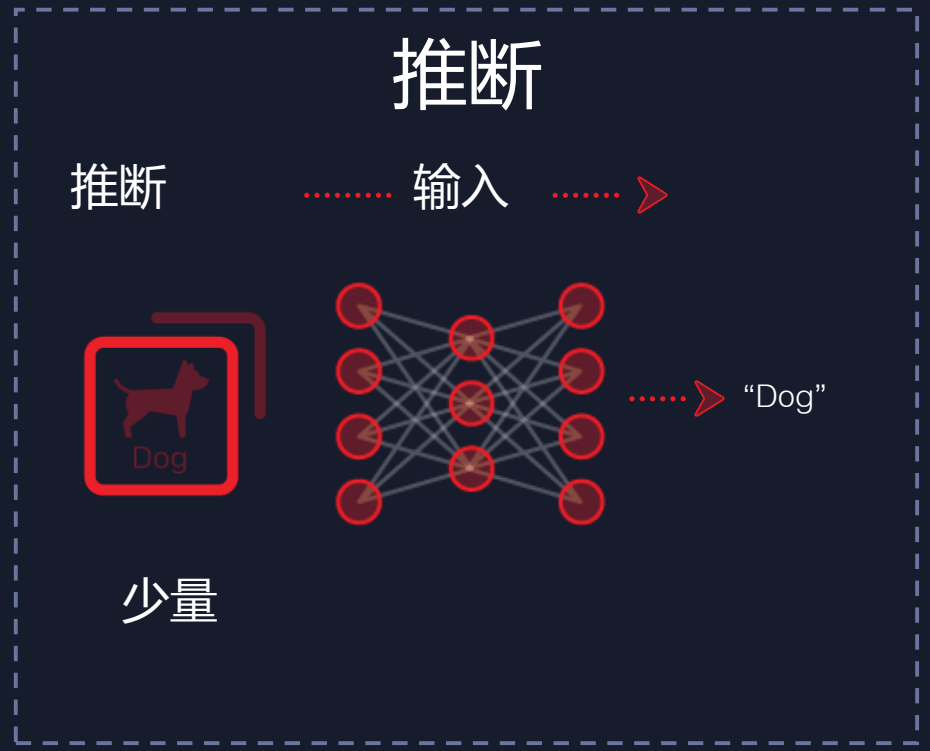


Barclays Research, 2018年5月公司报告



推断的挑战

- 创新的速度
- 低时延时的性能
- 更低的功耗
- 整体应用加速

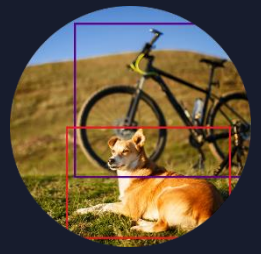


AI 模型创新的速度

各种应用

分类

目标识别

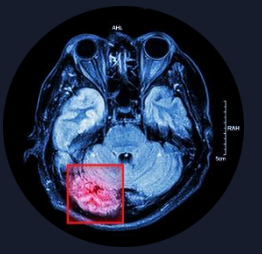


分割

语言识别

推荐引擎

异常检测



CNN

RNN, LSTM

MLP

适用范围广泛的各种模型

AI模型创新的速度：分类

Source:

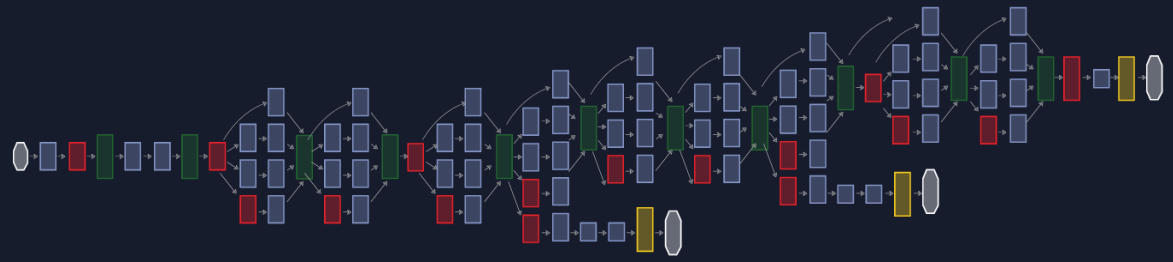
https://arxiv.org/pdf/1605.07678.pdf https://arxiv.org/pdf/1608.06993.pdf

https://arxiv.org/pdf/1709.01507.pdf https://arxiv.org/pdf/1611.05431.pdf

网络的复杂性与日俱增

AlexNet

GoogLeNet



DenseNet

推断趋向低精度

RELATIVE ENERGY COST

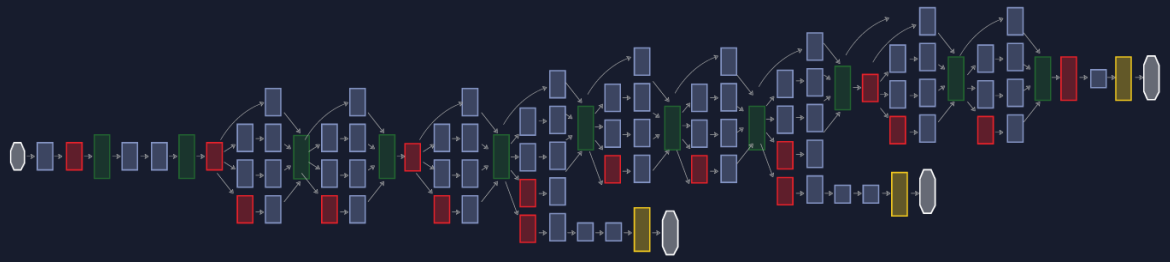
Operation:	Energy (pJ)
8b Add	0.03
16b Add	0.05
32b Add	0.1
16b FP Add	0.4
32b FP Add	0.9

Source: Bill Dally (Stanford), Cadence Embedded Neural Network Summit, February 1, 2017

硅片更新的周期远远跟不上创新的速度

AlexNet

GoogLeNet



DenseNet

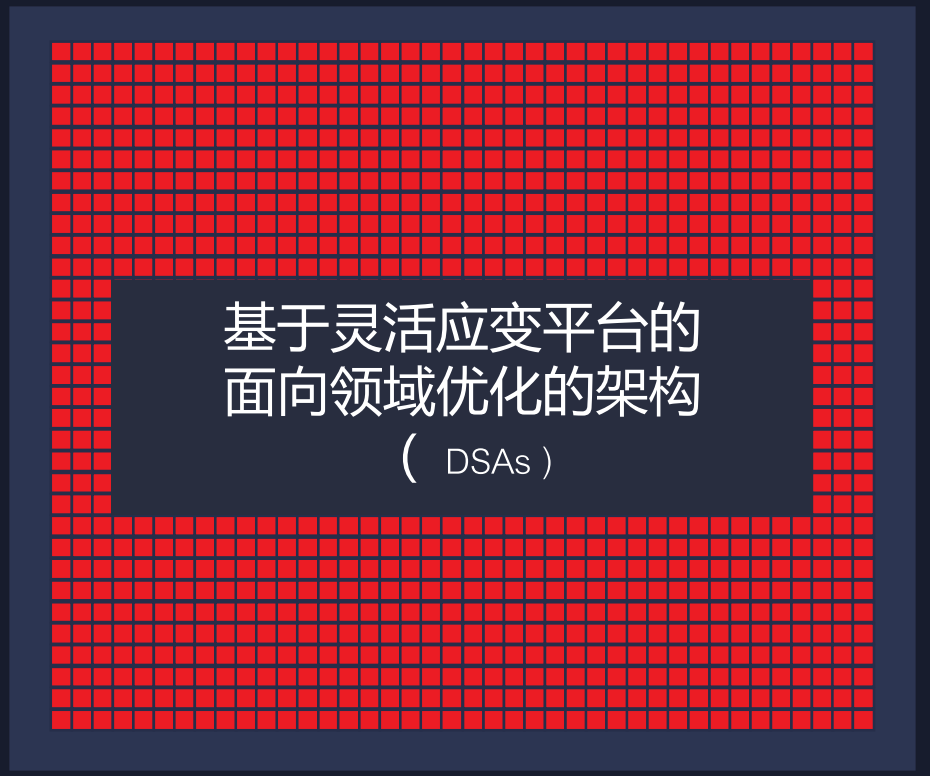
硅片的生命周期

只有灵活应变的硬件才能应对推断所面临的挑战

自定义数据流

自定义存储器层次结构

自定义精度



赛灵思收购深鉴科技

自定义数据流

自定义内存层次结构

自定义精度

DEEPHi 深鉴科技 Now Part of XILINX

剪枝　　量化

- 专利压缩技术
- 减少 DL 加速器占用空间
- 提高每瓦性能

示例: DeePhi LSTM

自定义数据流

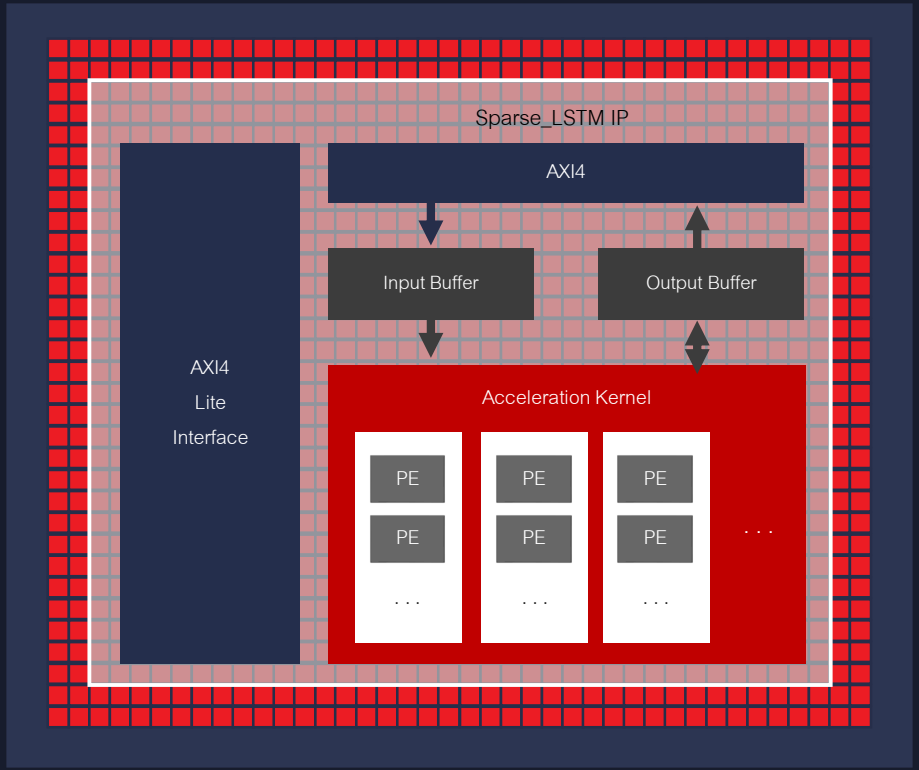
面向语音识别的 LSTM

自定义存储器层次结构

内存中稀疏矩阵的实现

自定义精度

12 bit weights, 16 bit activations



示例： xDNN

自定义数据流

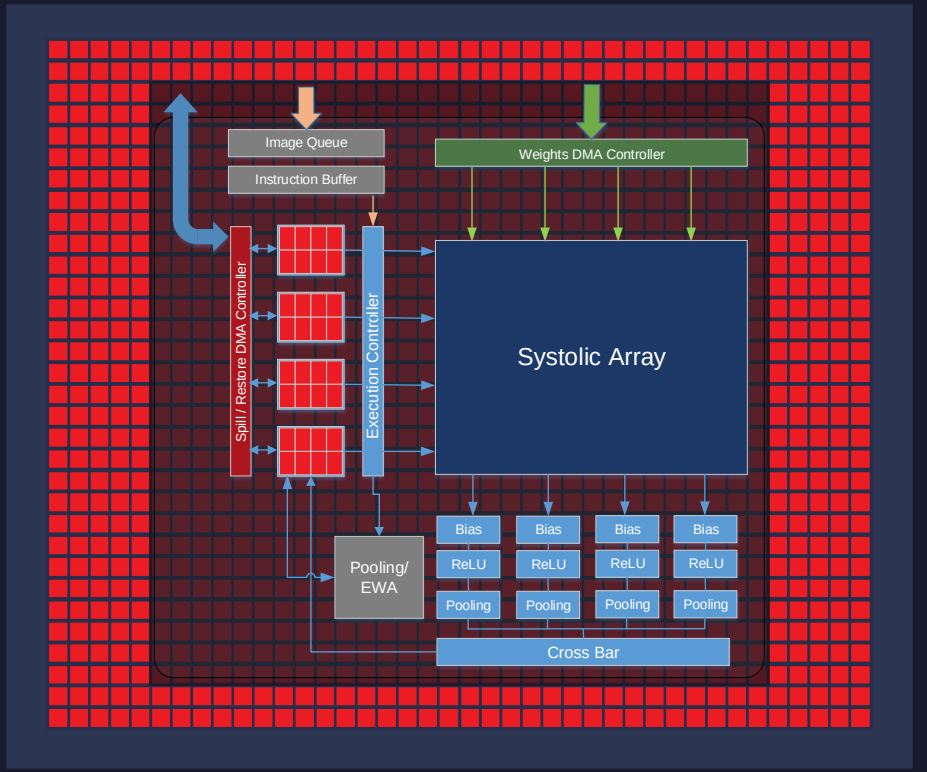
面向最新 CNN 而优化

自定义内存层次结构

优化的片上存储器

自定义精度

Int8



低时延对推断至关重要

时延响应 50 毫秒

CPU/GPU

Input 1	Result 1
Input 2	Result 2
Input 3	Result 3
Input 4	Result 4

高吞吐量 或 低延迟

时延响应仅为 3 毫秒

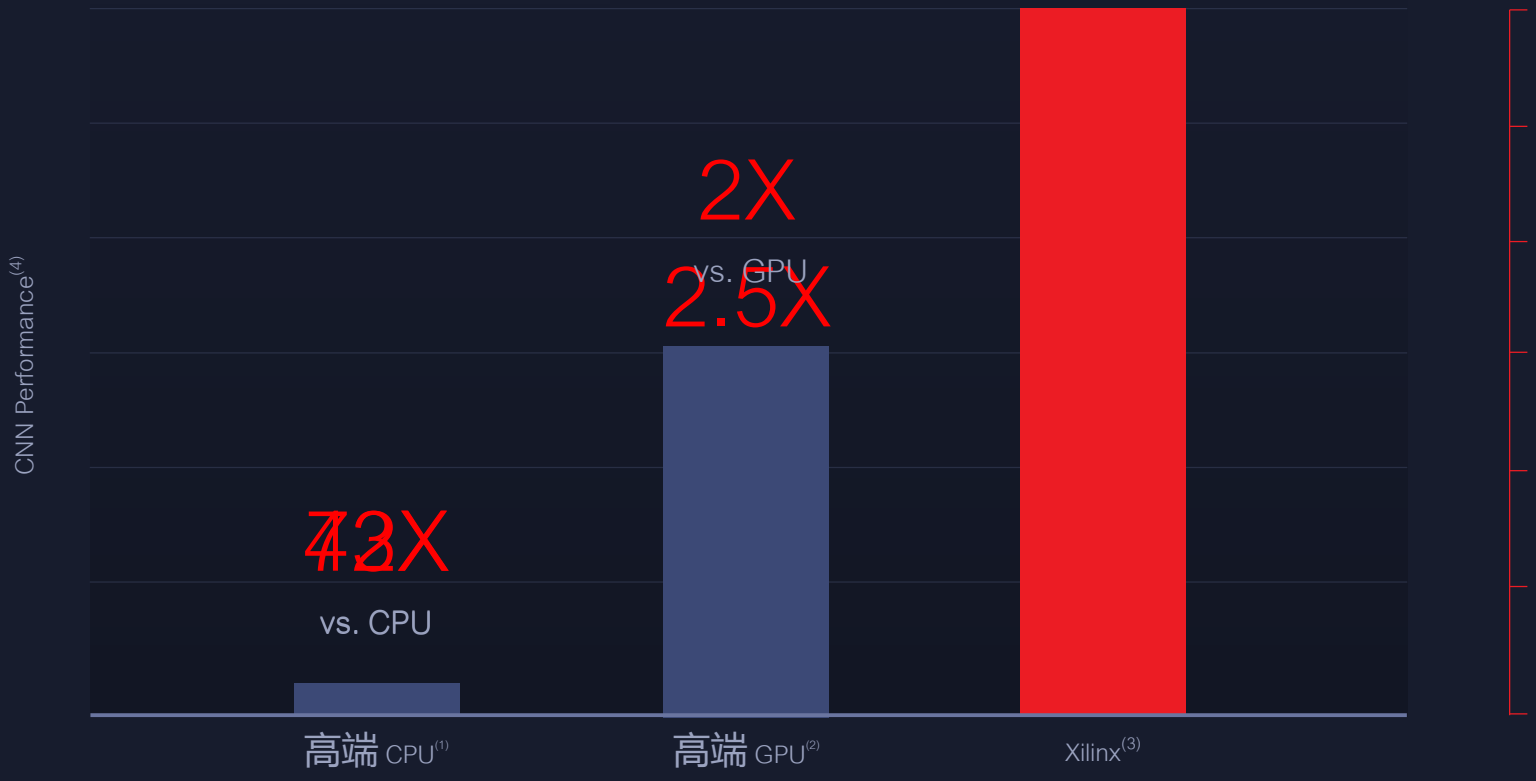
FPGA/ACAP

Input 1	Result 1
Input 2	Result 2
Input 3	Result 3
Input 4	Result 4

高吞吐量 和 低延迟

低时延：赛灵思的独特优势

时延不敏感的推断



AI 推断加速

利用 AI 引擎

大多数的自适应引擎及标量引擎可以支持整体应用的硬件加速

(1) Measured on EC2 Xeon Platinum 8124 Skylake, c5.18xlarge AWS instance, Intel Caffe: https://github.com/intel/caffe

(2) V100 numbers taken from Nvidia Technical Overview, "Deep Learning Platform, Giant Leaps in Performance and Efficiency for AI Services"

(3) Versal Core Series

(4) GoogLeNet V1 throughput (Img/sec)

低时延 CNN推断的性能优势

深鉴科技的剪枝技术

1.3倍-8倍

基于该网络的
性能提升

Sources: Alveo - Published (INT8); Versal - Projected (INT8), 65% PL reserved for whole application; GPU 1 - P4 Published (INT8); GPU 2 - V100 Published (FP16/FP32); GPU 3 - T4 Projected

功耗对于推断应用至关重要

云推断

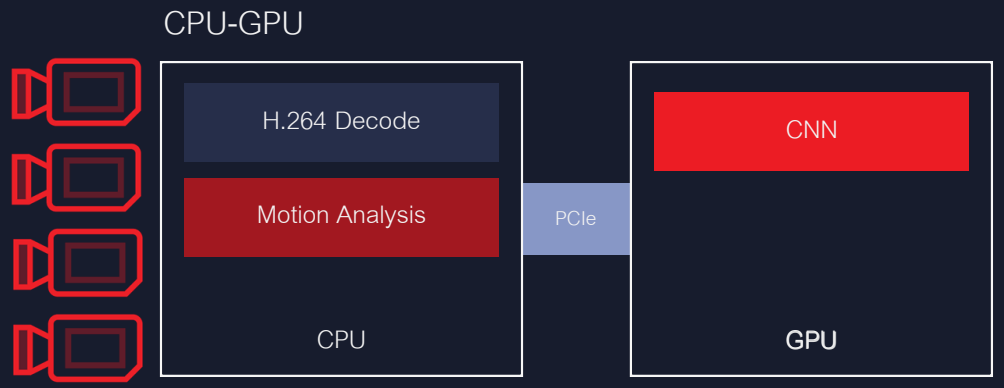
SK Telecom NUGU 个人助理应用

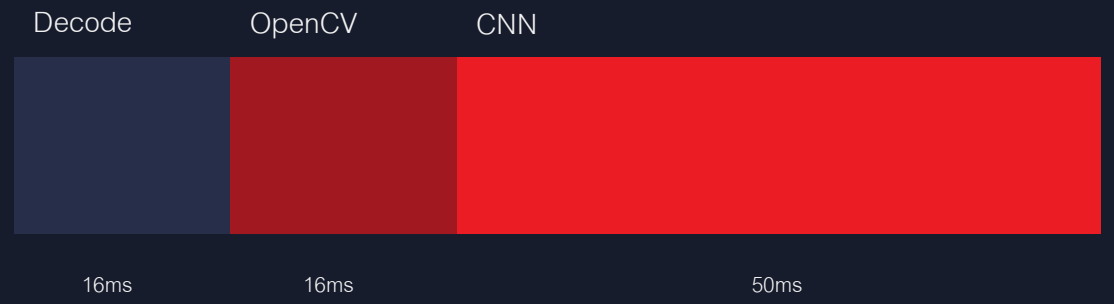
16倍

每瓦性能比GPU

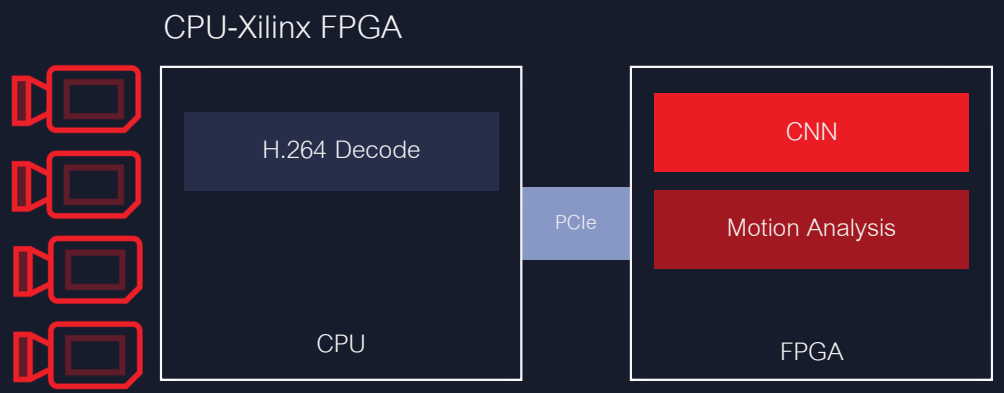
韩国 SK 电讯的 NUGU 个人助理

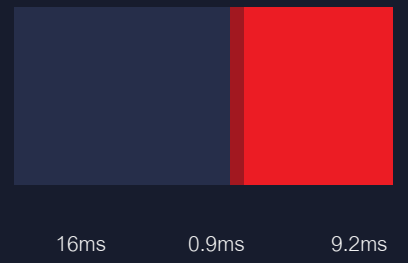
加速整体应用:智慧城市/ 安全





> 功耗: 75瓦　> 时延: 82毫秒　> 吞吐量: 4x12 fps





> 功耗: 50W　> 时延: 26.1毫秒　> 吞吐量: 4x38 fps

加速整体应用:在线视频处理

欢聚时代 YY Inc.

Aupera Making Video Alive

1
Aup2603

Decode H.264	ML / Analytics	Process	Encode HEVC

视频转码 + AI 分析

30
E5 Servers

48 ZU7EV

打造开发者的社区

云

{RESTful API}

python™

端

Caffe

TensorFlow™

mxnet

客户模型

开源模型

加速库

剪枝 / 压缩

编译器和量化工具

Runtime

xDNN

Descartes (LSTM)

Aristotle (DNN)

FPGA-as-a-Service

Alveo

定制开发板

FPGA 及 ACAP

IN SUMMARY

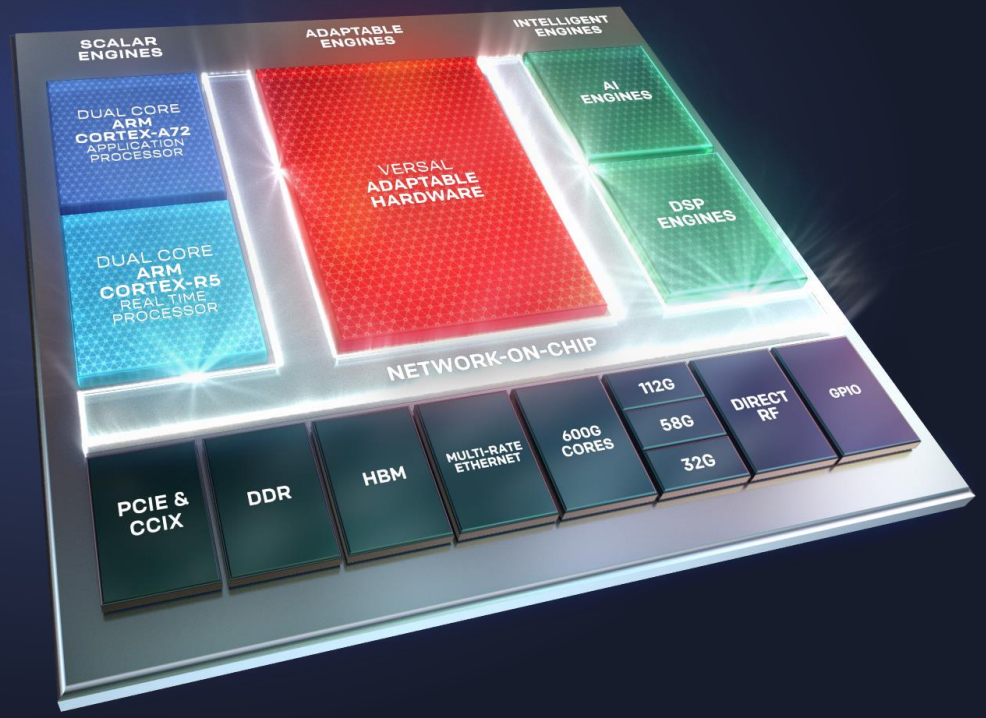
只有赛灵思灵活应变的器件才能提供:

同步 AI 创新的速度

以低时延提供最佳性能

提供最佳的功耗效果

加速整体应用




赛灵思
打造灵活应变、
万物智能的世界