Open Source AceCon 2021 智能云边开源峰会
AI x Cloud Native x Edge Computing
人工智能×云原生×边缘计算

赋能全场景，构建“芯”生态

“周易”AIPU的创新与演进

安谋科技

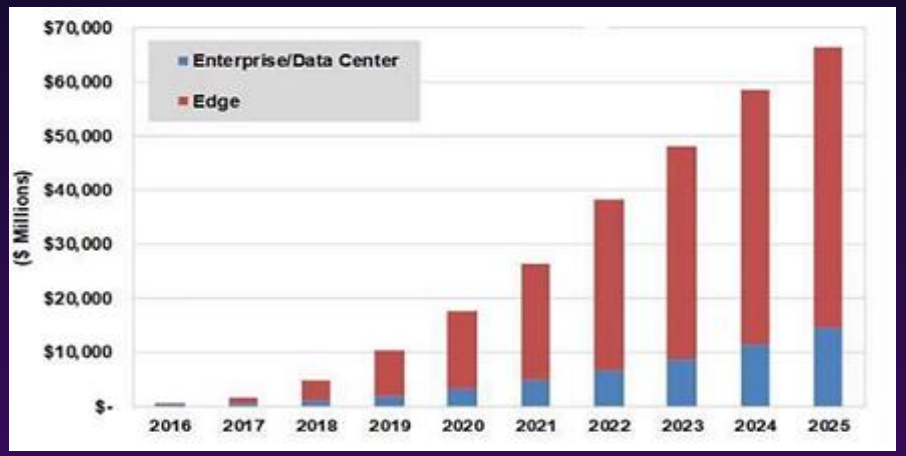
吴彤
Senior AI Technical Marketing Manager

1 智能计算进入新时代

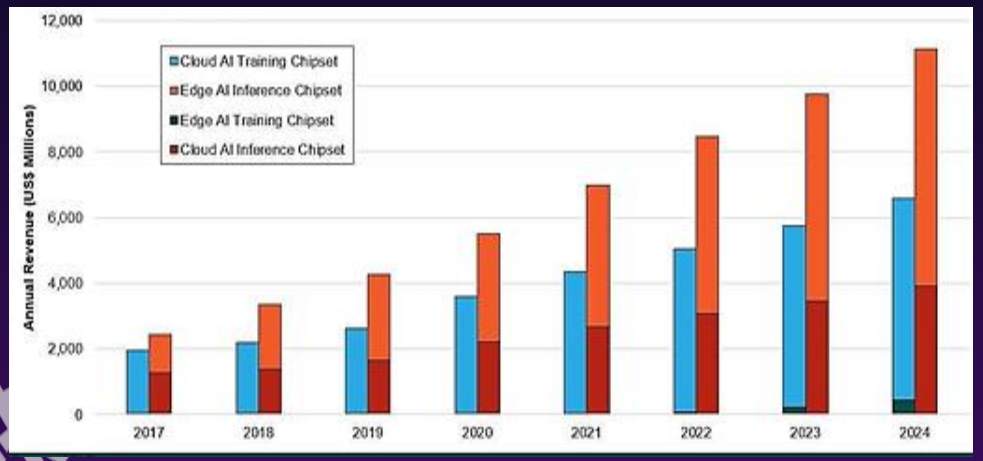
第五波计算浪潮

2 AI芯片市场概览

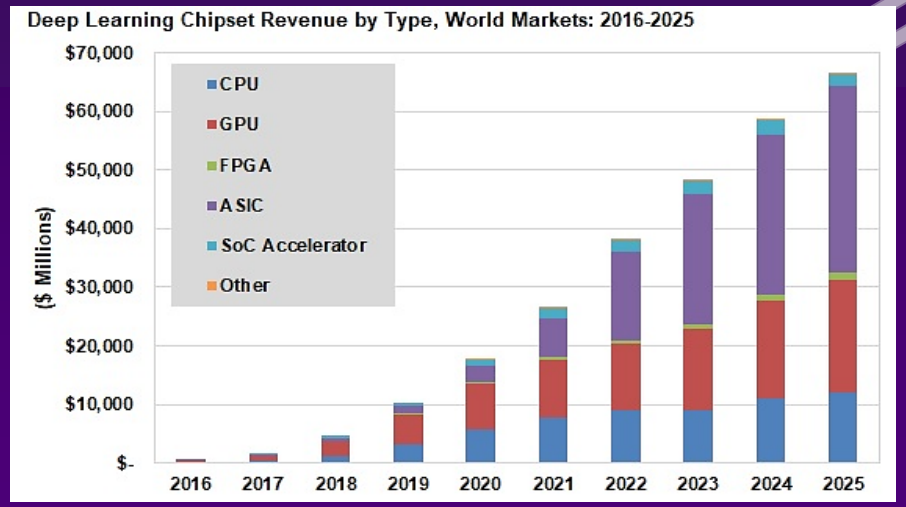
AI芯片市场规模及结构分析



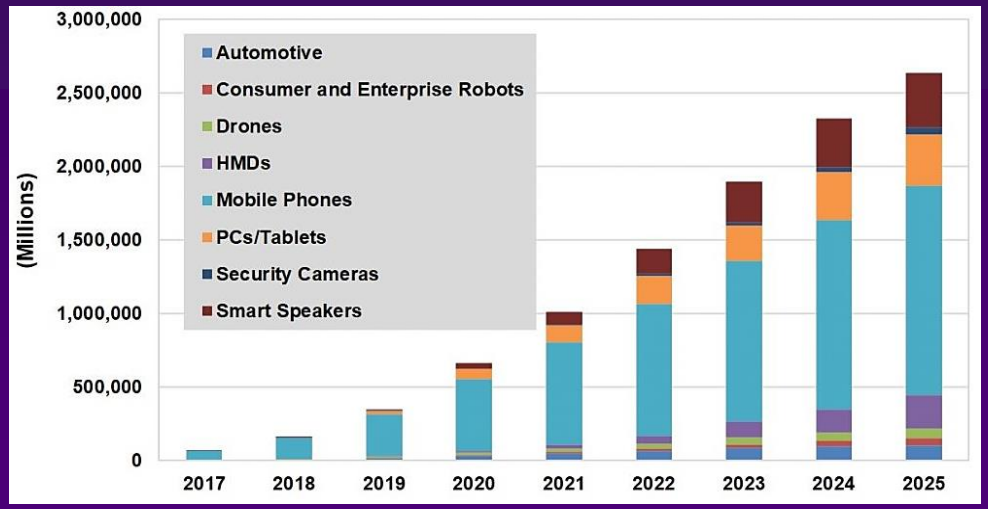
AI芯片市场规模



AI芯片按场景



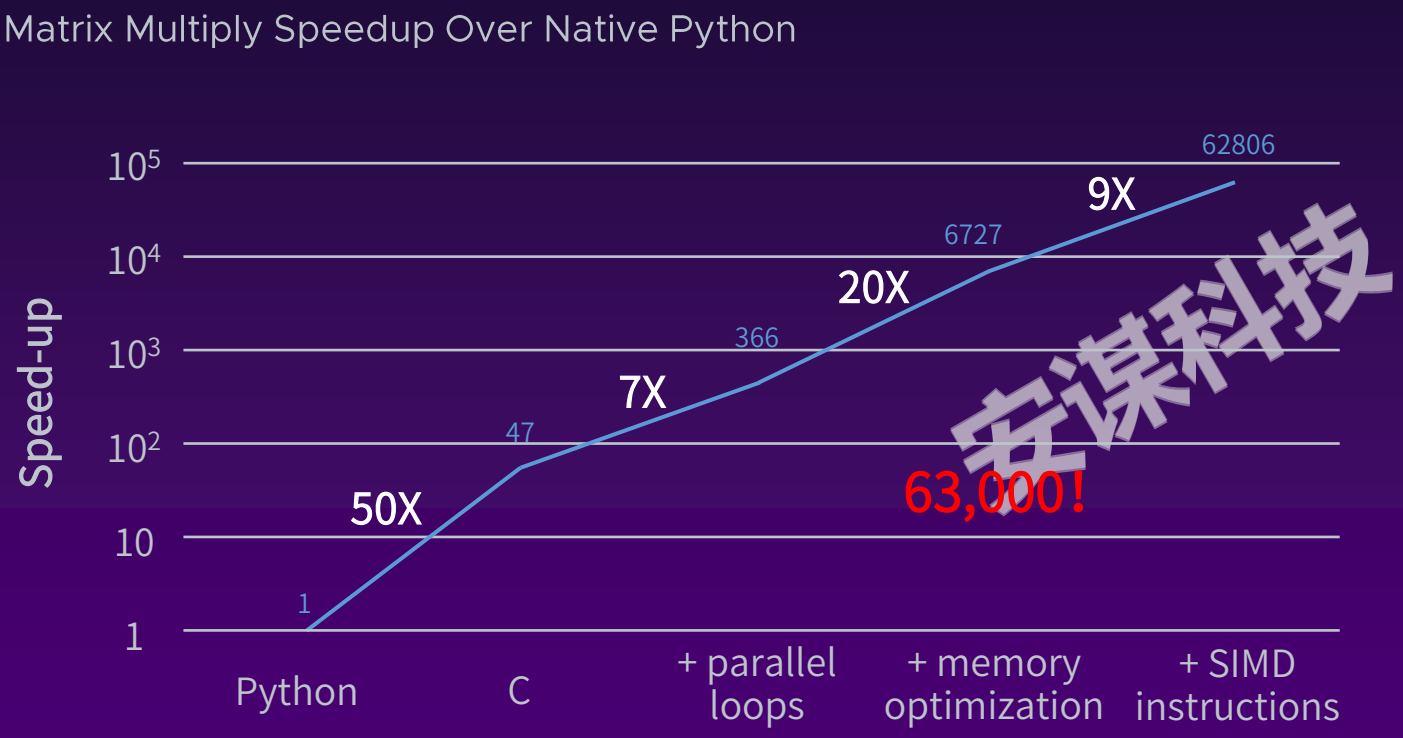
AI芯片按架构



AI芯片按功能

数据来源：Tractica & ABI Research

AI芯片领域专用架构(DSA)的兴起



- DSA将弥补软硬件的性能鸿沟
- DSA将硬件架构进行定制并使其具备特定领域应用特征，使该领域的一系列应用任务都能高效执行
- 典型的DSA架构：
  - 机器学习领域的神经网络处理器
  - 图形图像/虚拟现实领域的图像处理器（GPUs）
  - 可编程网络交换机及接口

数据来源：David Patterson, A New Golden Age for Computer Architecture

AI芯片架构演进趋势(通用 + 专用)

GPU
- Many cores
- Strong scalar/Weak vector
- High freq/power hungry

DSA
(Domain Specific Architecture)
- Many cores
- Ad-hoc instructions

Domain Specific DSP
- Strong vector
- Selective fix function
- Low freq/power hungry

HWA
(Hard-Wired Accelerator)
- Systolic array like
- Standard fix function

GPU+HWA

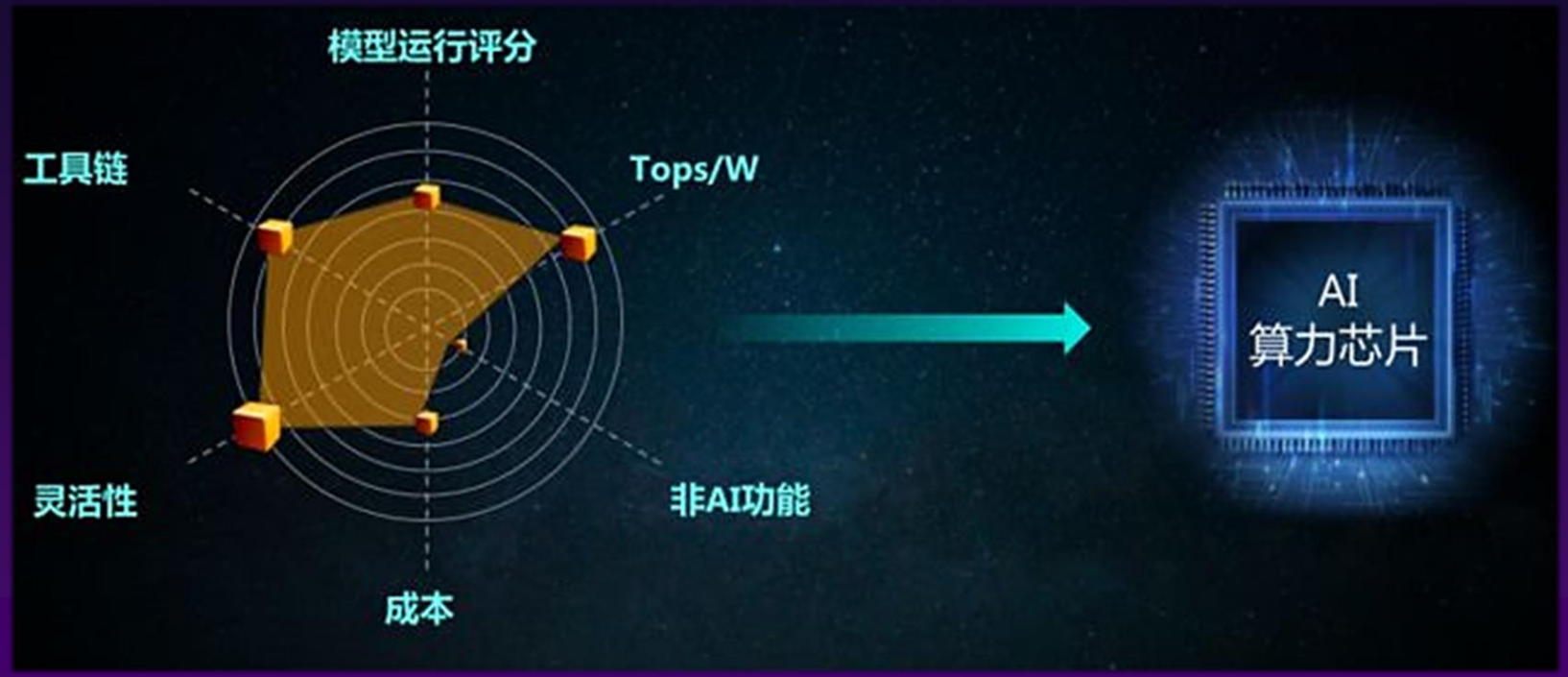
GPU Derivative
- Remove graphic pipeline

DSA+HWA

DSP Derivative
- Add huge MAC array

Host+DSA+HWA
- High perf/watt
- Enough flexibility

AI芯片选型评估标准



模型运行评分：
- MLPerf
- AI-Benchmark
- DAWNBench
- Benchmarking(AIIA)

工具链：
- 支持不同的算法框架
- 性能仿真器
- 准确的调试反馈
- 神经网络量化优化

Tops/W：
- TOPS
- TOPS/W(位宽，功耗，OPS利用率)
- 存储带宽

灵活性：
- 算力能否应对非常规的网络
- 算力能否扩展支持的层种类
- 算力能否兼容将来可能出现的新网络

成本：
- OPS利用率/能效比高，成本低

非AI功能：
- 具有应用处理器核(AP) - SoC芯片
- 包含其它非深度学习的图像处理模块
- 包含视频处理与编解码模块
- 有丰富的嵌入式接口用于数据传输和控制

3 安谋科技“周易”AIPU的创新与演进

赋能每台智能设备

安防

手机

自动驾驶/智能座舱

物联网

家居

新零售

VR/AR

机器人

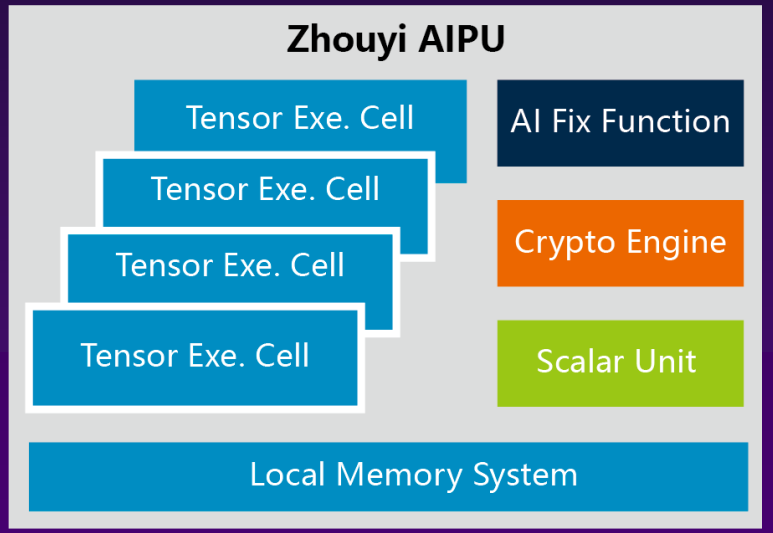
多样性场景 + 通用 /专用结合 - “周易”AIPU

本土研发

安谋科技自研AI IP，自主可控，多个领域大规模商用

创新架构

专门用于深度学习的AI IP, 统一指令集实现各类AI计算功能



完整生态

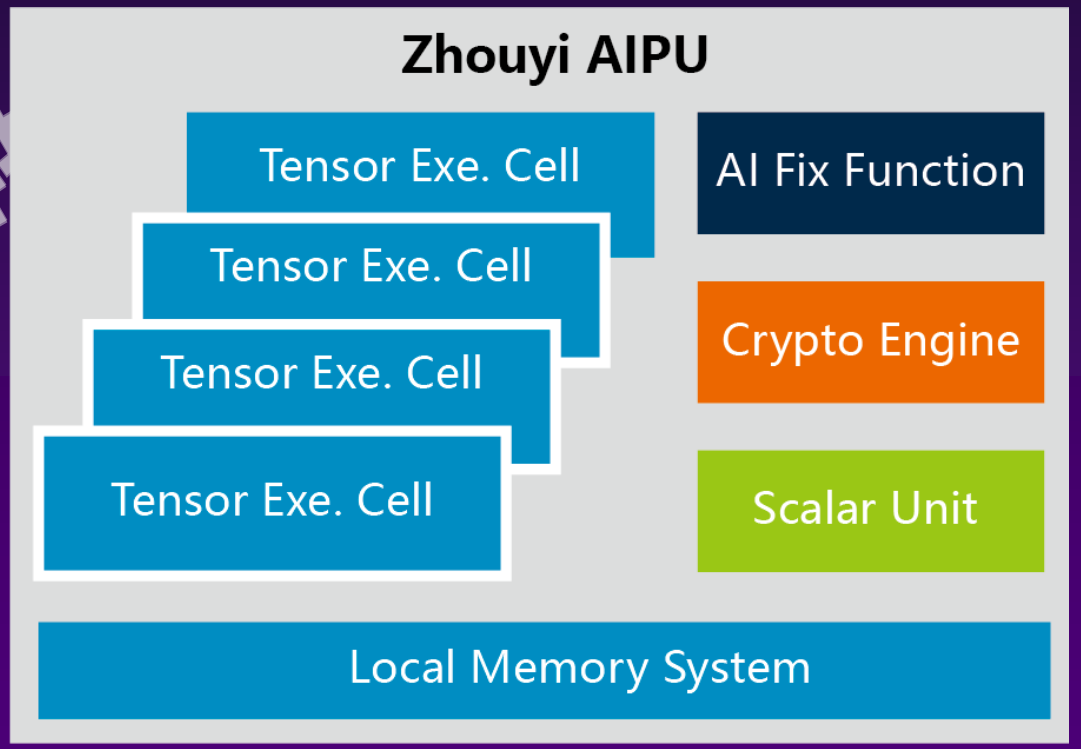
AIPU + 完整工具链 + 算法

PPA

PPA更佳，兼具灵活性及安全性

新设计 – 创新架构，更佳PPA，更安全的AIPU

- ✓ **创新架构**
  - 张量
    - 新的架构设计了AI特定域张量指令集(Tensor Instructions)
  - 高能效比的专用硬件加速
    - 特定AI操作的指令集以实现定制的硬件加速单元标量
  - 标量
    - 用于NN计算所必需的通用标量指令集
- ✓ **PPA最佳平衡**
  - 可扩展性配置
  - 高密度性能
- ✓ **可选的安全扩展**
  - 有效保护用户信息，AI算法
- ✓ **可选的客户扩展**
  - 客户自定义算子(50%客户选择自定义算子)

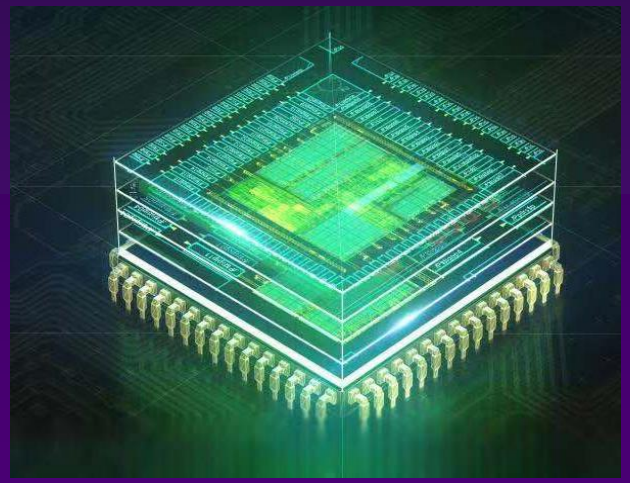


更灵活 – 可编程的AIPU

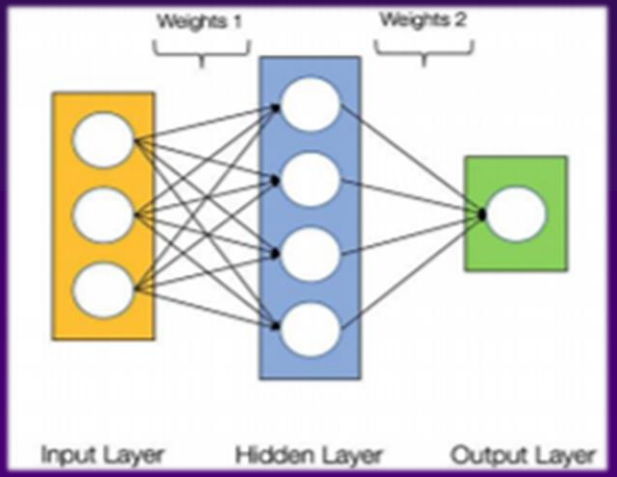
芯片合作伙伴
一颗芯片卖给更多的OEMs

OEM产品
可运行各类的主流算法

灵活的AI解决方案

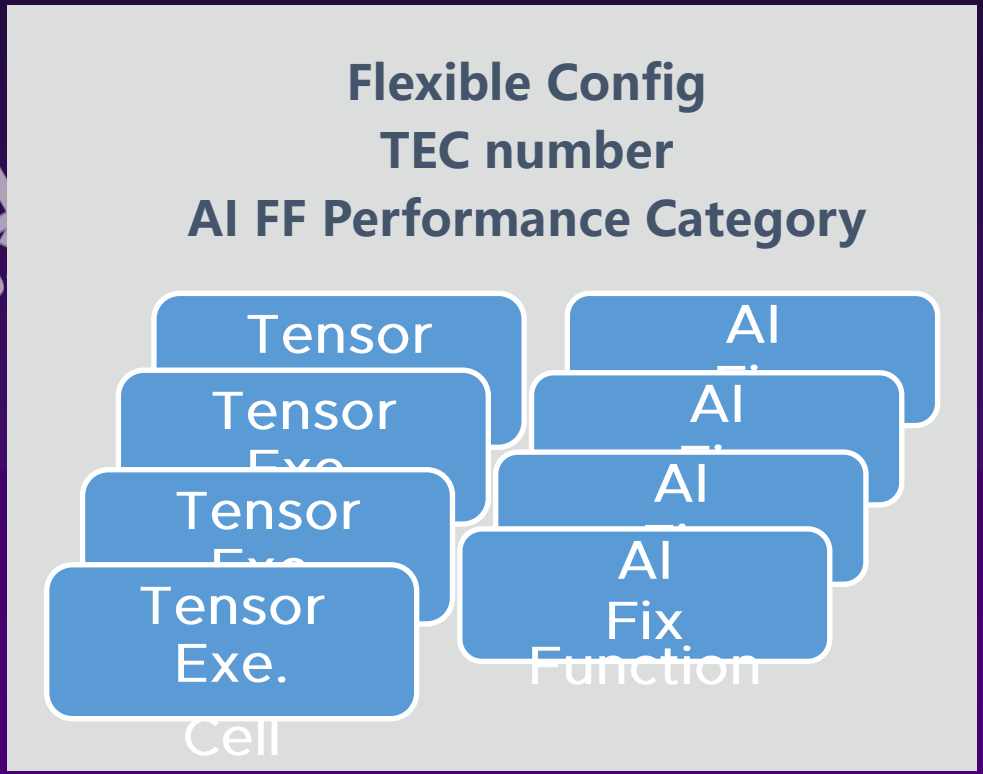


- **主流的算法支持**
  - SSD/Resnet/Segnet…
  - CNN/RNN/LSTM…
- **OEM内部算法或第三方开发**
  - 客户自定义层
  - 高效支持支持所有预定义层及算子(OP)



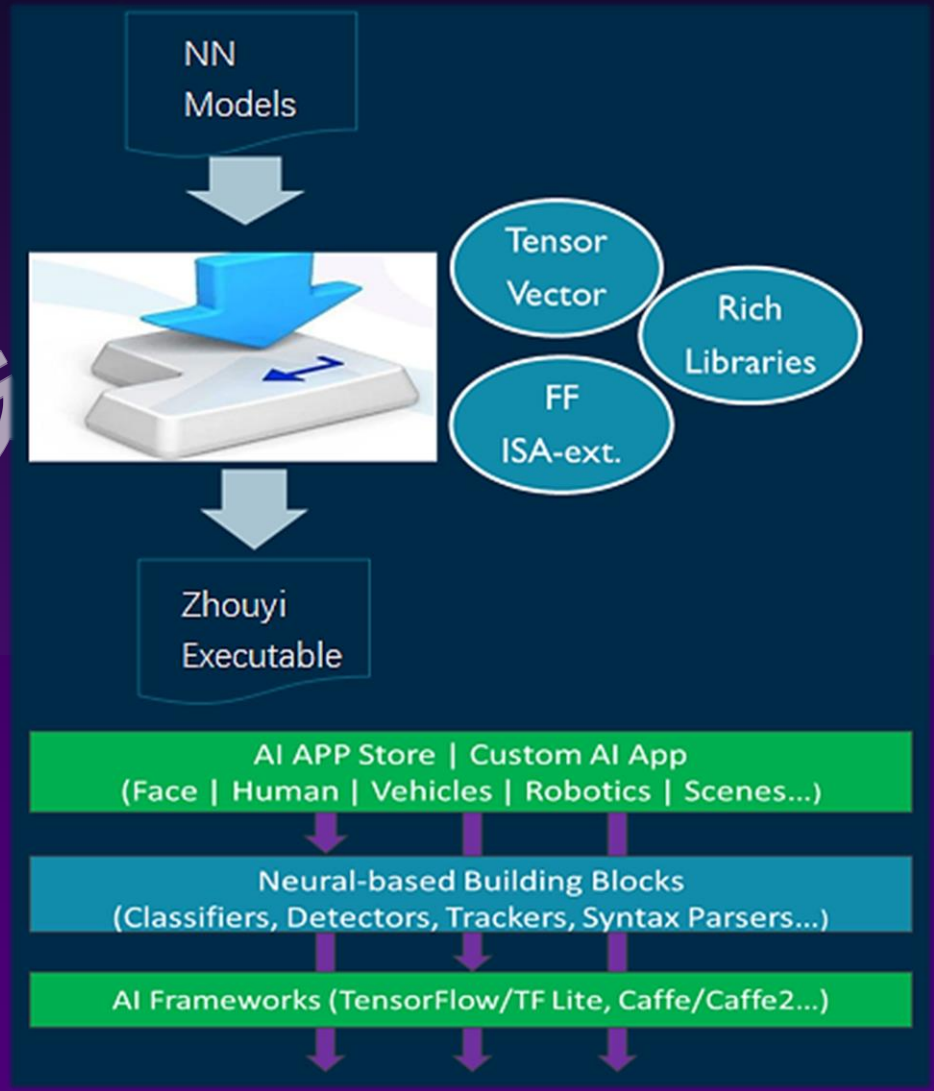
更灵活 – 可扩展的AIPU (0.2TOPS – 上百TOPS)

- **多个TEC(Tensor Execution Cell)**
  - 对任何神经网络层/算子(OP)完全可编程
- **特定 AI 操作指令集(AI Fix Function)**
  - 卷积和内积(Convolution and Inner Product)
  - 池化(Max and Average Pooling)
  - 激活(Activation)
  - BN, Bias, Eltwise
  - 权重/特征图压缩
- **用户硬件自定义扩展(Customize FF)**
  - 客户自定义AI Fix Function
  - 特定操作的差异化
  - 复用”周易”成熟的工具链



更易用 – 完整易用的AIPU SDK

- **完整工具链**
  - 链接库(离线)
    - 模型转换(一键生成)/性能库
  - 运行库
    - Scheduler/RT资源最优利用率
    - Driver and Firmware可配置
  - 仿真器/调试器
- **NBB - Neural-based Building Blocks**
  - 统一的面向AI应用的API
  - 易于生成AI应用
- **支持所有主流AI框架**
- **丰富的AI应用**
  - 基于CPU/GPU/AIPU的应用优化
  - 覆盖10W+人群的开发者社区(极术社区)

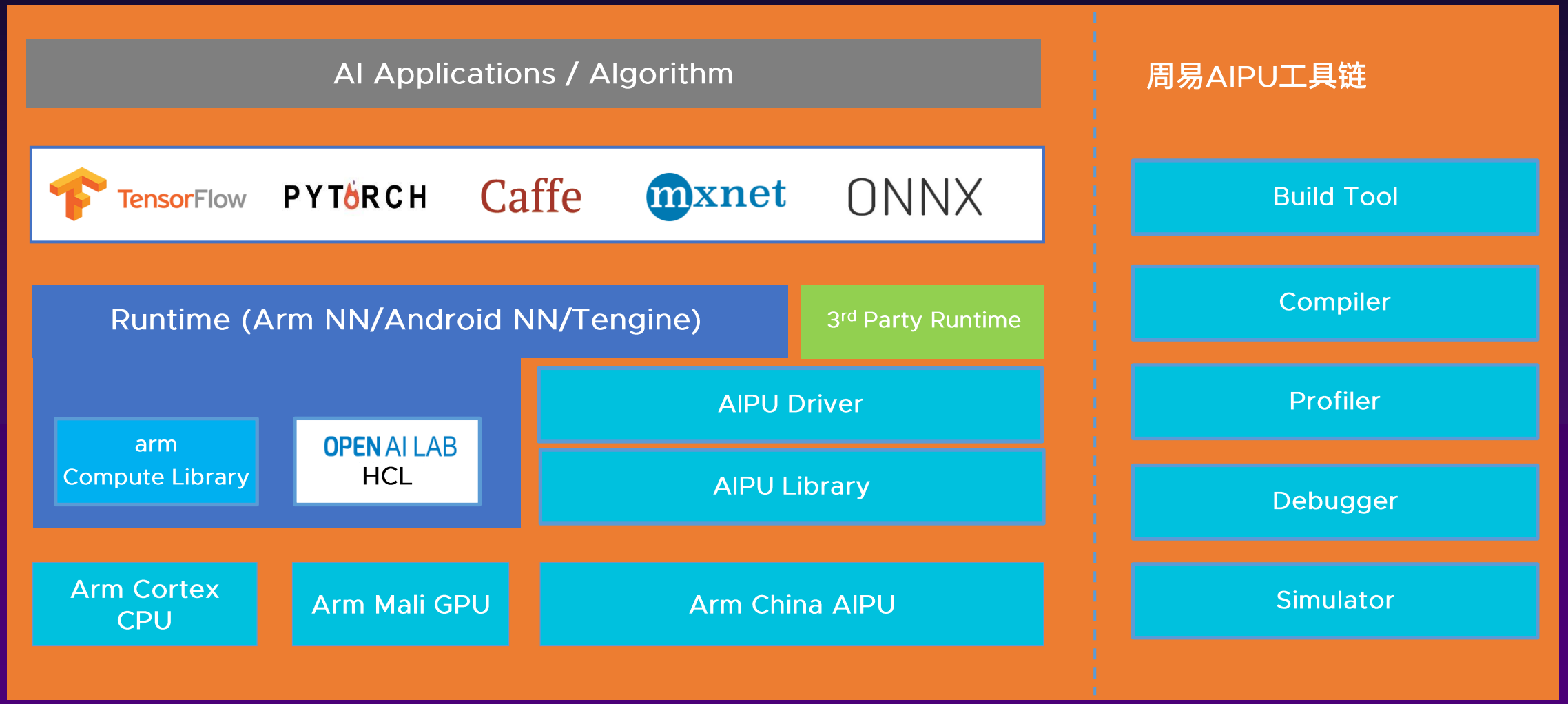


全场景 – 端/边缘/云全覆盖的AI IP

	Device(端)					Edge(边缘)						Cloud(云)
Scene场景	智能音箱	智能门锁	智能TWS	智能手机	刷脸支付终端	AR/VR	智能机器人	智慧商显	智能安防(IPC/NVR)	ADAS & IVI	边缘服务器	数据中心
TAM(2019) 市场规模	¥69.1亿元	¥15亿元 (人脸)	¥68亿元 (安卓)	¥7326亿元	¥50亿元	¥73.3亿元	¥152 亿元	¥789 亿元	¥350亿元	¥725亿元	N/A	N/A
Shipment(2019) 出货量	3682万	110万	792万	3.69亿	1000万	200万	1000万(扫地)	307万	7000万	220万(L2) <40万(IVI)	150万	62万(GPU)
Performance 算力	<1TOPS	<1TOPS	<1TOPS	1–10 TOPS	1–10 TOPS	1–10 TOPS	4–20TOPS	1–10 TOPS	1–20 TOPS	>20TOPS	10- 100 TOPS	>200TOPS
Power 功耗	<10mW	<10mW	<10mW	1- 2W	3 – 10W	3 – 10W	10 - 30W	1- 10W	3 –10W	10 – 100W	10 – 100W	200W
Latency 延迟	<10ms	<10ms	<10ms	10 – 100ms	10 – 100ms	<20ms	<10ms	10 – 100ms	10 – 500ms	<50ms	ms ~ s	ms ~ s
Cost 成本控制	高	高	高	极高	高	高	中	中	高 ~ 低	中	中	低
Main Solution 主流方案	GPU	DSP	DSP	GPU/ ASIC	GPU	GPU/ ASIC	GPU	GPU/ ASIC	GPU/ASIC/FPGA	GPU/FPGA	GPU	GPU/FPGA
Best Solution 最佳方案	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC	ASIC	GPU/FPGA/ ASIC

- “周易”Z1：边缘计算通用， 面向IoT & Edge
- “周易”Z2：边缘计算中高性能场景，算力/性能2X”周易”Z1, 架构/内存优化

全平台 – 全栈式AI异构平台



全支持 – 丰富的模型库及主流算子支持

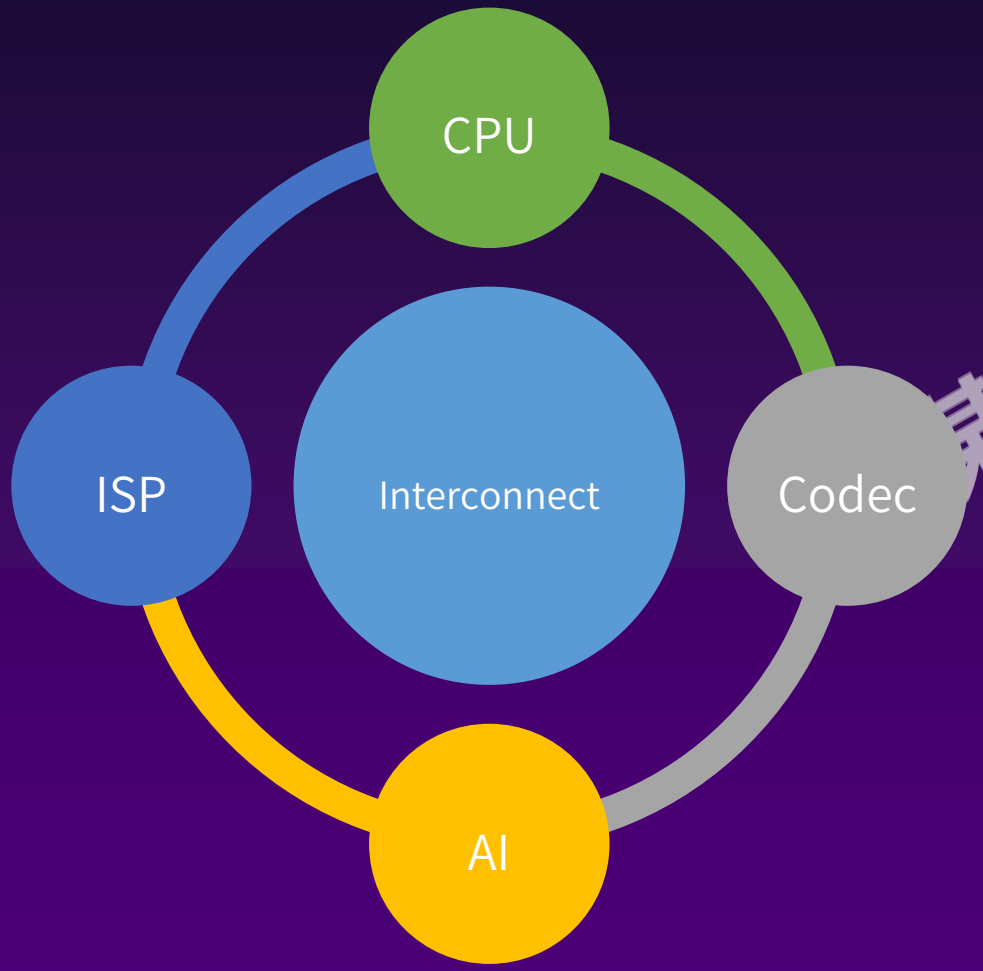
预训练模型库(Model Zoo)：

- Detection
- Classification
- Segmentation
- Super resolution
- Driver monitor
- Speech Recognition
- Text to Speech
- Transformer

算子(OP)：已支持120+主流算子,更多算子支持实现中

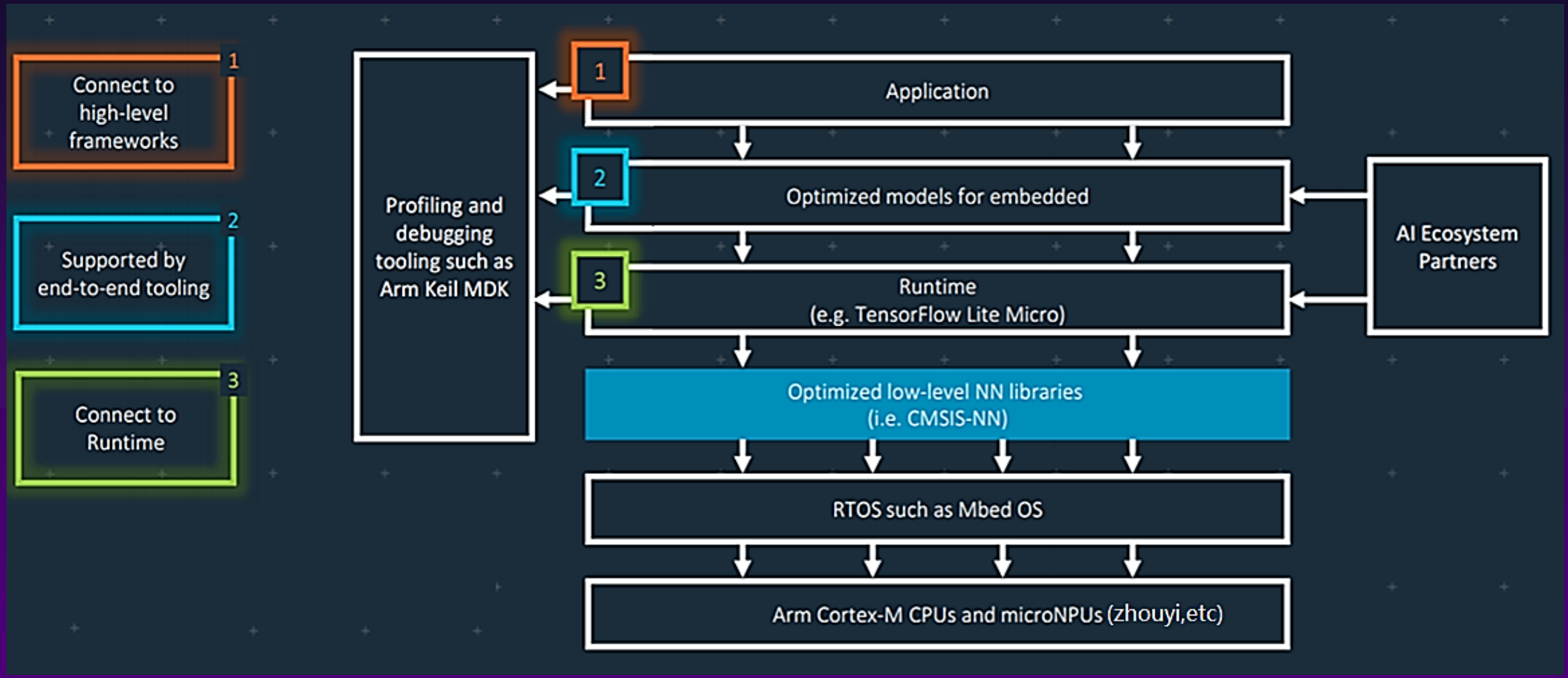
- Conv
- Deconv
- Depthwise conv
- Dilated conv
- FC
- BN
- ReLU
- LeakyRelu
- sigmoid
- Tanh
- Top-k
- Max pooling
- Avg. pooling
- LSTM
- GRU
- Prior box
- Eltwise
- Softmax
- Reorg
- Detection output
- Resize
- Slice
- Region
- Concat
- permute
- Flatten
- Reshape
- Depth to space
- Space to depth
- Transpose
- Moments
- Gather
- math

全方案 – 面向视觉的完整IP解决方案

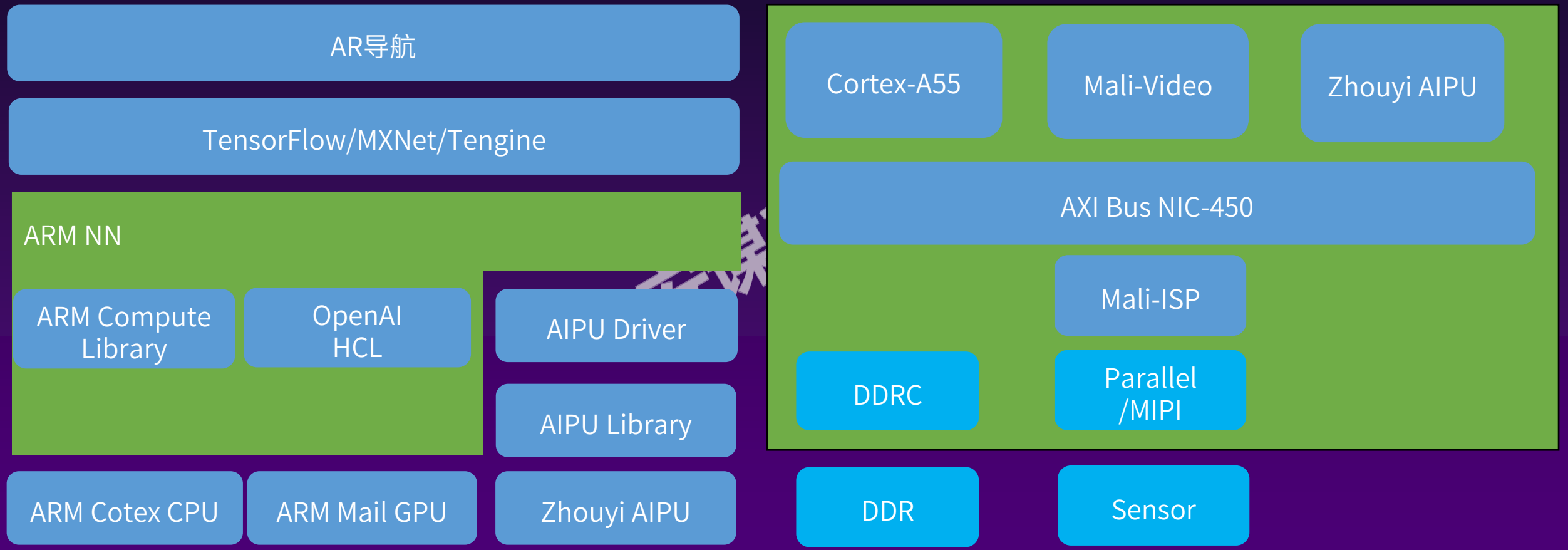


- CPU – Arm® Cortex ® -A系列
- ISP – “玲珑” i3/i5
  - 可配置, 支持HDR, 3NR
- Video Encoder/decoder - v5/v7
  - 可配置, 支持全标准
- AIPU – “周易” Z1/Z2
- SOC总线系列
  - 缓存一致性CoreLink CCI总线
  - NOC互联总线
  - 低延迟CoreLink NIC总线

“周易”Z1 – 面向TinyML场景的更佳解决方案



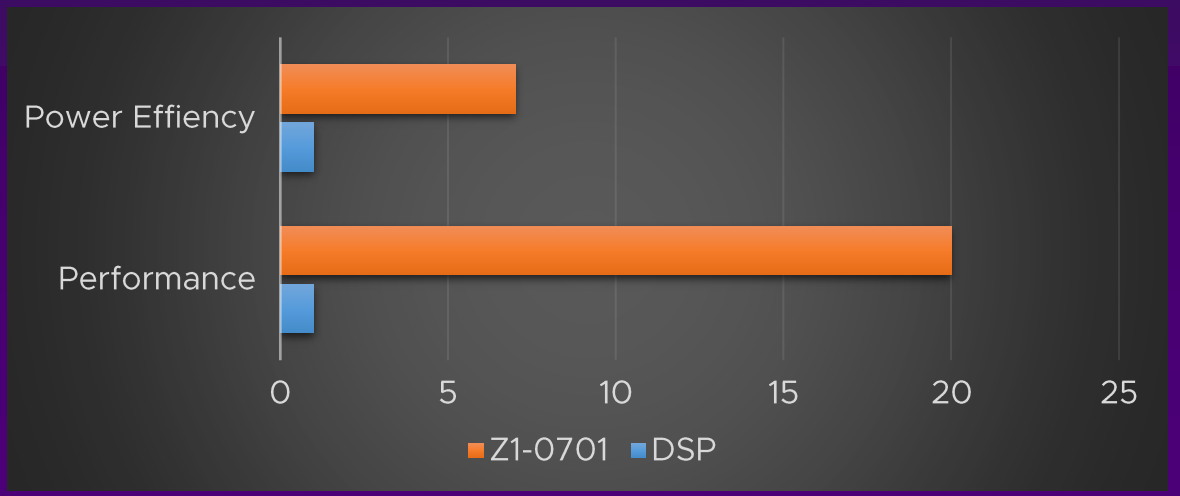
“周易”Z2 – 面向智能座舱及中高端安防的更佳解决方案



“周易”Z1 – 智能家居端到端语音识别首选

- **语音模型的高性能/视觉分析**
  - 本地语音唤醒(KWS)/语音识别(ASR)/自然语言理解(NLP)/翻译(Translation)/检测与识别
- **商用落地能力**
  - 全志R329智能语音芯片(“周易”Z1)即将大规模商用

- **0.1~ 0.2 TOPS @ DVFS**
  - 高能效比的专用硬件加速(“星辰”/A53 /A55 + “周易”)
  - 对任何神经网络层/算子(OP)完全可编程
- **支持各类主流的AI框架**
- **低功耗及更小的面积**



“周易”Z1 – 人体关键点检测应用案例

- 支持主流算法(Open pose, Deep pose, CPM)
- 82 FPS，“周易”Z1-0904(1T)
- 低带宽
- 低成本/高能效比

“周易” Z2 – 边缘计算中高端场景首选

➢ 性能提升

- 单核4TOPS算力，2X于“周易” Z1，多核算力可达128TOPS
- 部分神经网络模型，相同算力配置下性能提升可达100%以上
- 混合精度支持Int8/Int16

➢ 微架构优化

- 芯片面积减少30%

➢ 内存子系统提升

- 128/256bit AXI interface
- CACHE/Local SRAM/Global SRAM
- 可选的SOC系统SRAM

➢ 高级带宽节省技术(ABST)

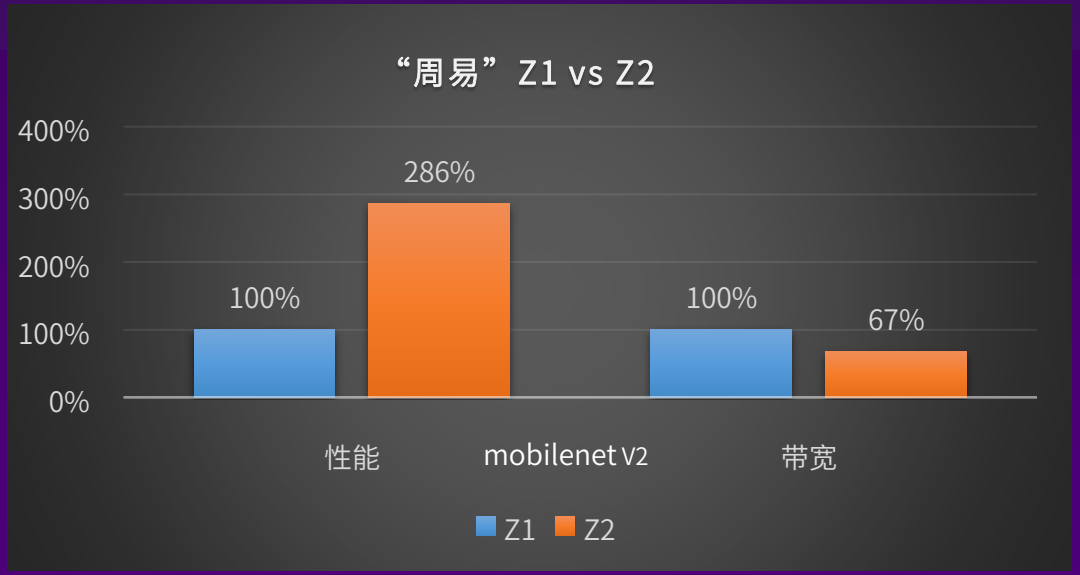
- 权重(weight)和特征图(feature map)压缩

中高端安防

- 8 – 16路高清视频分析
- 高级带宽节省技术(ABST)
- Tier 1安防客户芯片研发中

自动驾驶/智能座舱

- 多核可配置扩展, 最高算力可达128T
- 芯擎(吉利)7nm智能座舱芯片年内量产



“周易”Z2 – 驾驶员疲劳监测(DMS)应用案例

1. 商业化主流驾驶员疲劳监测(DMS)算法
2. Face Detection, Face landmark, Head pose, Gaze等算法融合
3. 70 fps， “周易”Z2-0901(1T)
4. 低带宽
5. 低成本/高能效比

“周易”Z2 – DTV超级分辨率应用案例

- 商业化主流超级分辨率算法
- 1080P->4K@60 fps, “周易”Z2-0901(1T)
- 视频质量评价工具 - VMAF 93分
- 低带宽
- 低成本/高能效比

安谋科技：全球标准，本土创新

周易

面向安防、车载、移动和
AIoT 场景的AI处理器

山海

面向物联网设备的全栈式
信息安全解决方案

安谋科技

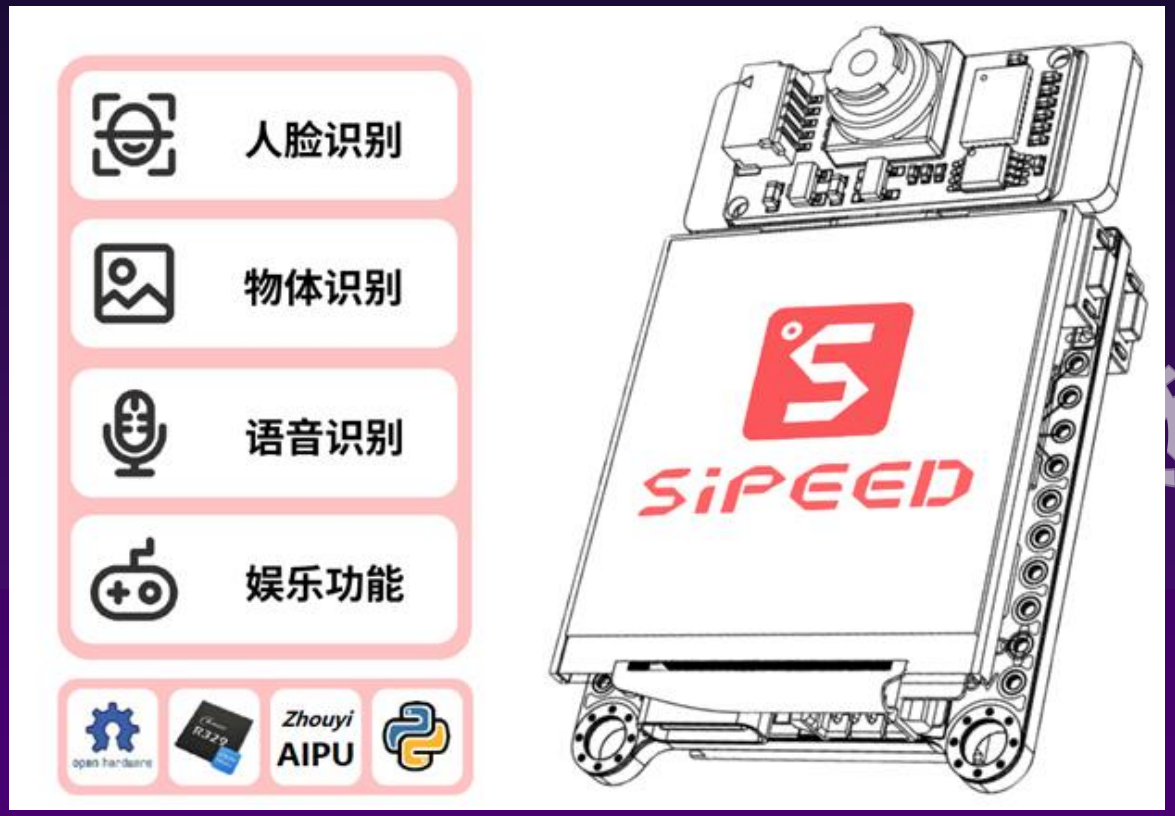
星辰

面向智能、互联、安全IoT应用
需求的处理器

玲珑

包括ISP、VPU等在内的
多媒体处理器

R329芯片AI开发板首发



极术社区： https://aijishu.com/a/1060000000221780

EdgeX & Arm

LF Edge Member Company - Premier (24)

Altran	Arm Holdings	AT&T	Baidu	Charter Communications
Dell Technologies	Dianomic	Equinix	Ericsson	Fujitsu

智能计算产业创新联合体 Open NPU Innovation Alliance

架构（ISA）自由

- 为不同应用场景提供可选和可定制化智能计算方案
- NPU处理器指令集的独立自主可控

统一的标准

- 开源ISA，合力于标准的演进与软硬件生态开发
- 统一开放（开源）的工具链，SDK、软件库，降低总体智能计算应用成本

技术演进

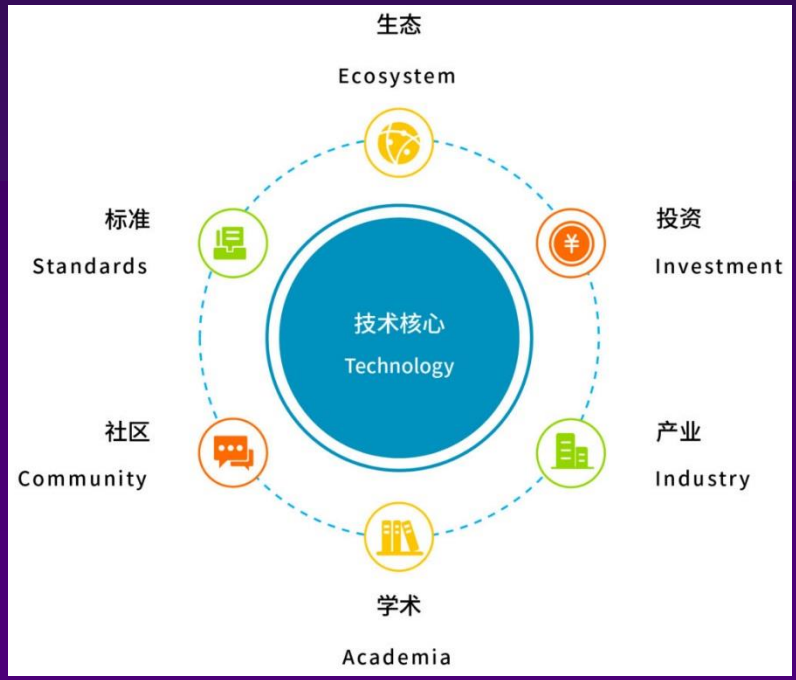
- 推动异构计算中的NPU技术演进，保持前沿创新
- 拓展NPU在智能计算应用领域的适用场景
- 有效满足新算力堆迭、分布计算、超域(xDSA)计算的需求

更多信息：https://www.open-npu.org

会员范围

*暂不接受个人会员申请

- IP公司 / IC设计公司
- 系统厂商
- 学术机构（AI及微电子）
- AI生态系统平台及企业（SW/HW/算法）



边缘计算分论坛

Open Source AceCon

2021 智能云边开源峰会

AI x Cloud Native x Edge Computing

人 工 智 能 × 云 原 生 × 边 缘 计 算

Thank You

安谋科技