SCIENCE
AND
TECHNOLOGY

TENGINE
白皮书

TENGINE WHITE PAPER

01 为什么要用Tengine？

1.1 AIoT应用开发需要全新的工具

在过去20年里，互联网技术的高速发展已经深刻地改变了我们的生活。与互联网相比，当前正在进行的物联网（IoT）技术革命则正在以前所未有的方式、更加全面和彻底的重构着我们的生活和工作方式。毫不夸张的说，这是人类发展历史上从未面临过的美好前景。与此同时人工智能（AI）技术与物联网技术的深度结合、以及深度神经网络(DNN)技术在计算机视觉等领域所取得的突破性进展，在全球范围所有行业内掀起波澜壮阔、势不可当的系统性变革。

出于成本，延时，带宽，功耗，可靠性和隐私性问题的考虑，从2018年开始，越来越多的应用场景把AI计算从云端向终端设备迁移，无论在安防、交通、制造、医疗还是自动辅助驾驶、智能家居等领域，数以亿计的设备都将智能化升级。一个新兴的巨大的物联网AI应用市场（AI-IoT）正在形成。然而，巨大市场机遇的背后，我们看到的是一个严峻问题 — AIoT应用的开发缺乏好用的开发工具，这让AI应用的开发效率大打折扣，甚至已经成为了制约行业发展的最大瓶颈。

首先，现有的AI开发工具都是为服务器或者手机平台开发设计的，市面上找不到一款专为嵌入式场景设计的AI开发工具，因此，要在IoT设备上开发一款AI应用可不是一件容易的事。 因为IoT设备无论在芯片规格，软件环境，系统要求上都与服务器设备有很大差异，甚至与手机相比，其硬件资源环境也要比起手机平台苛刻许多。市面上绝大多数IoT设备都是基于微控制器(MCU)开发的，而MCU大量使用Arm Cortex-M处理器，算力只有几百MOPs，内存RAM往往只有几MB，甚至连操作系统都没有，因此，在MCU上开发AI应用，也成了只有极少数技术专家才能完成的事。

即便是在像手机这样的软硬件环境相对宽松的嵌入式设备上（一般基于Arm Cortex-A处理器），当前的AI开发工具仍然很难满足开发者的需求。比如最常见的视觉类AI应用，对算力的要求都很高，通常Arm CPU能提供50GOPs左右的算力，GPU能够提供几百GOPs的算力，勉强能够满足一般应用场景需求，但是当前的AI软件很难针对AI计算模式做优化，不能调用到芯片里的全部计算资源，导致实际硬件利用率只有30%甚至更低。

其次，AI应用的部署也是AI应用落地的一大难点。AIoT行业还处于早起发展阶段，从模型训练工具到芯片部署工具都有非常多的方案可选，这种竞争促进行业健康发展，但也造成了工具与芯片的双重碎片化。当想要把训练好的算法部署到数量众多的物联网设备上时，设备算力、平台间的兼容性问题都层出不穷。如果能有一套好用的工具，能够贯穿AIoT开发部署的全部流程，避免开发者出现兼容障碍，让开发能力与成本聚焦在业务场景上，将会大大加速AI应用的落地。

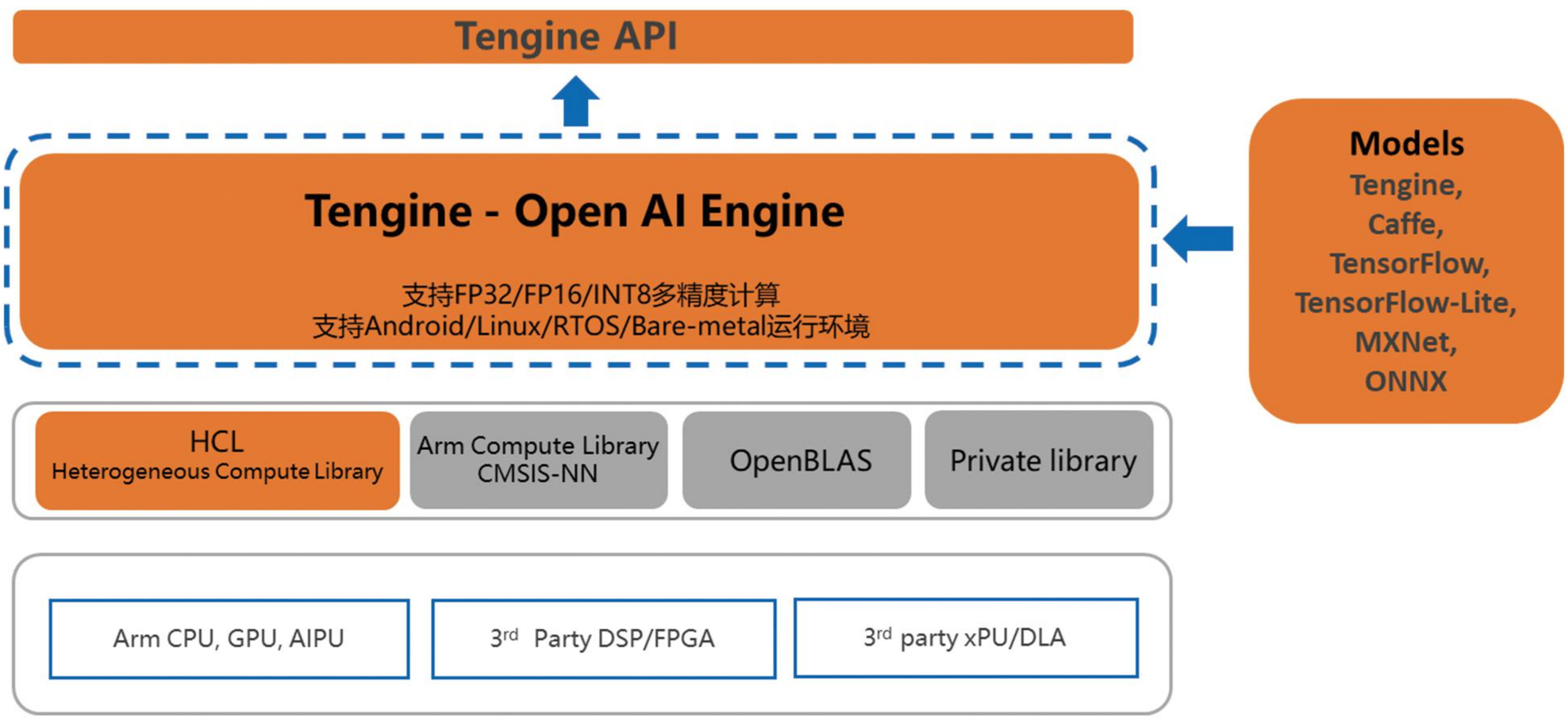
1.2 OPEN AI LAB的愿景

OPEN AI LAB 开放智能机器是由Arm中国和Arm生态伙伴共同发起，致力于提供最好的嵌入式AI应用开发工具与平台，解决嵌入式设备AI计算能力不足，芯片工具碎片化等一系列问题，助力AIoT应用生态蓬勃发展。OPEN AI LAB开发了Tengine，一个兼具性能与兼容性的开放的AIoT应用开发平台，赋能所有AIoT应用场景，提升AIoT开发部署体验，让AI应用开发更简单。

02 Tengine 开放AIoT应用开发平台

2.1 专为IoT场景设计的AI应用部署解决方案

Tengine的设计包含了IoT场景所需的全部特性：



Tengine功能框图

- 支持主流IoT设备

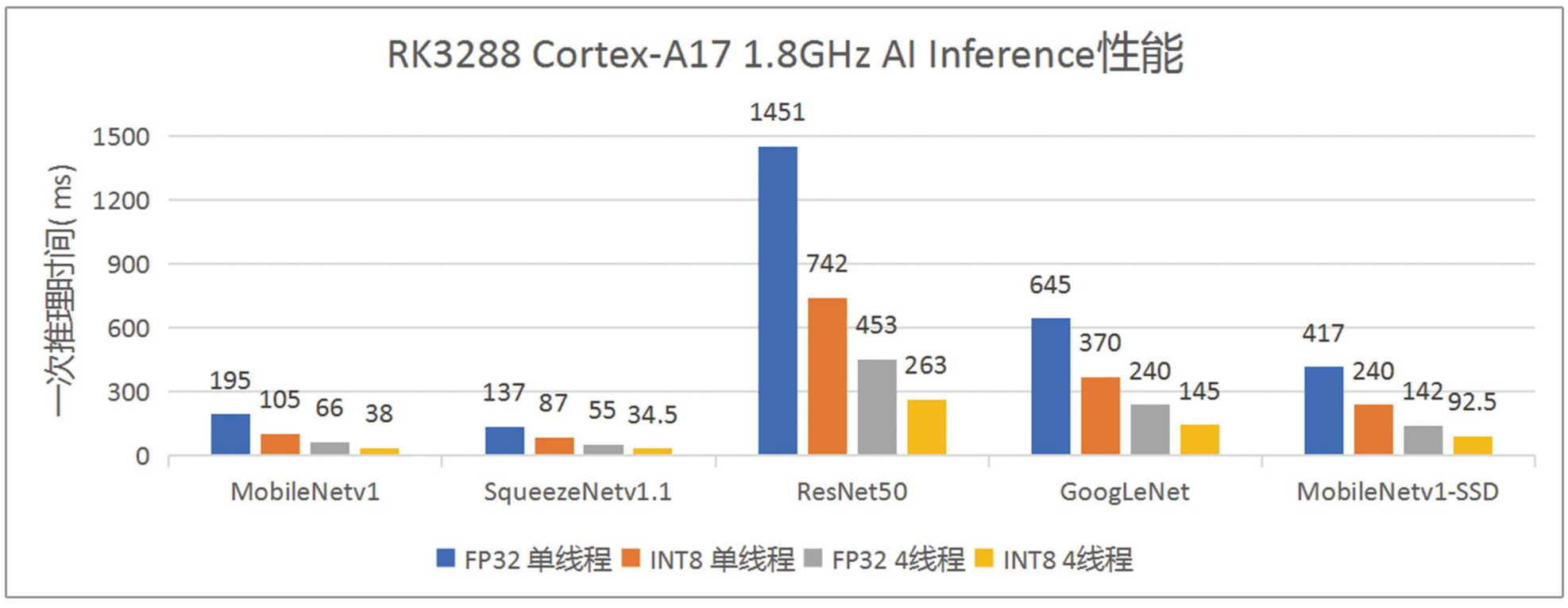
IoT设备大量采用Arm Cortex-A CPU或Arm Cortex-M CPU，因此Tengine设计之初就考虑了对Arm Cortex-A/M CPU的支持(Cortex-M CPU上的Tengine版本称为Tengine-Lite)。为了提高软件通用性，适应复杂多样的IoT设备芯片与软件环境，Tengine采用C++语言编写(Tengine-Lite采用纯C语言编写)，不依赖任何其他第三方库文件，支持主流IoT操作系统和环境，如Android，Linux，RTOS，Baremetal等，方便开发者在不同IoT设备上移植和部署。

- 轻量

当前IoT设备只有AI推理需求，并且要求成极低成本，程序占用ROM/RAM越小越好，因此Tengine只提供AI推理功能从而大幅简化框架设计，并且采用轻量模块化设计方案，便于扩展和裁剪，在AP上Tengine最小程序体积可以做到300KB，在MCU上最小体积可以做到20KB。

- 高效

AIoT设备往往计算资源有限，但应用场景对算力的需求却很高，因此OPEN AI LAB专门开发了针对Arm CPU的高性能计算库HCL（Heterogeneous Compute Library），充分挖掘Arm设备的AI计算能力。HCL库的核心技术是加速卷积等最耗时的算子，支持GEMM/Direct/Winograd等多种卷积计算模式，通过用手工调优汇编，针对Arm CPU微构架做极致优化，支持适配Armv7-A/v8-A/v8.2A全系列CPU。同时，HCL计算库支持FP32/FP16/INT8多种精度，支持纯浮点，端到端量化和混合精度三种计算模式，让开发者对性能与精度有更多的组合选择。



Mobilenetv1 1.0, SqueezeNet, ResNet50, GoogLeNet为标准224x224输入

mobilenetv1-SSD为300x300输入

• 支持异构计算

嵌入式设备的芯片往往是多个计算单元并存，CPU，GPU，DSP，AIPU等等。如何协同调用各种计算设备发挥最佳算力是IoT应用开发的重要需求。Tengine通过异构计算技术能够同时调用CPU，GPU，DSP，AIPU等不同计算单元来完成AI网络计算。Tengine提供两种方式的异构计算：

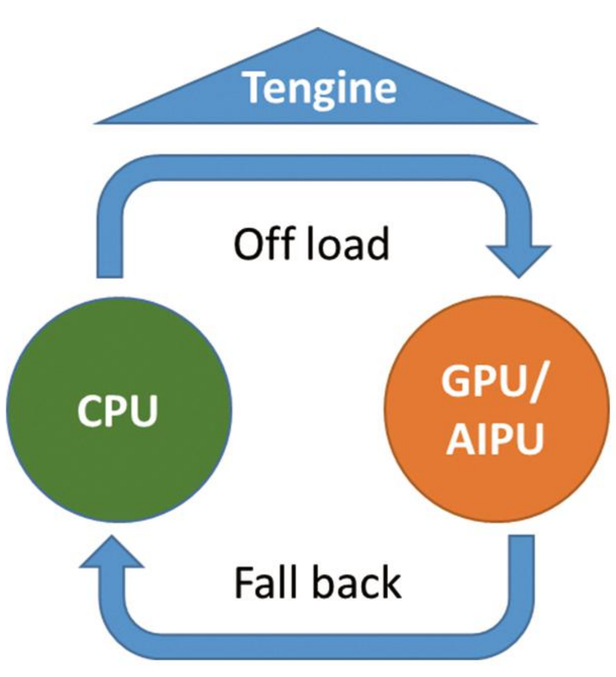
1.在不同的计算单元上运行不同的网络，比如在CPU上运行人脸检测网络，在GPU上运行人脸特征提取网络。

2.把一个网络切分成多个子网络然后通过调度将子网络的计算分配到不同的计算单元上，例如通过把卷积层，全连接层等常规算子放在GPU上计算来降低CPU计算负载，同时可以把Permute，Flatten，Priorbox等GPU不支持的算子放在CPU上执行。

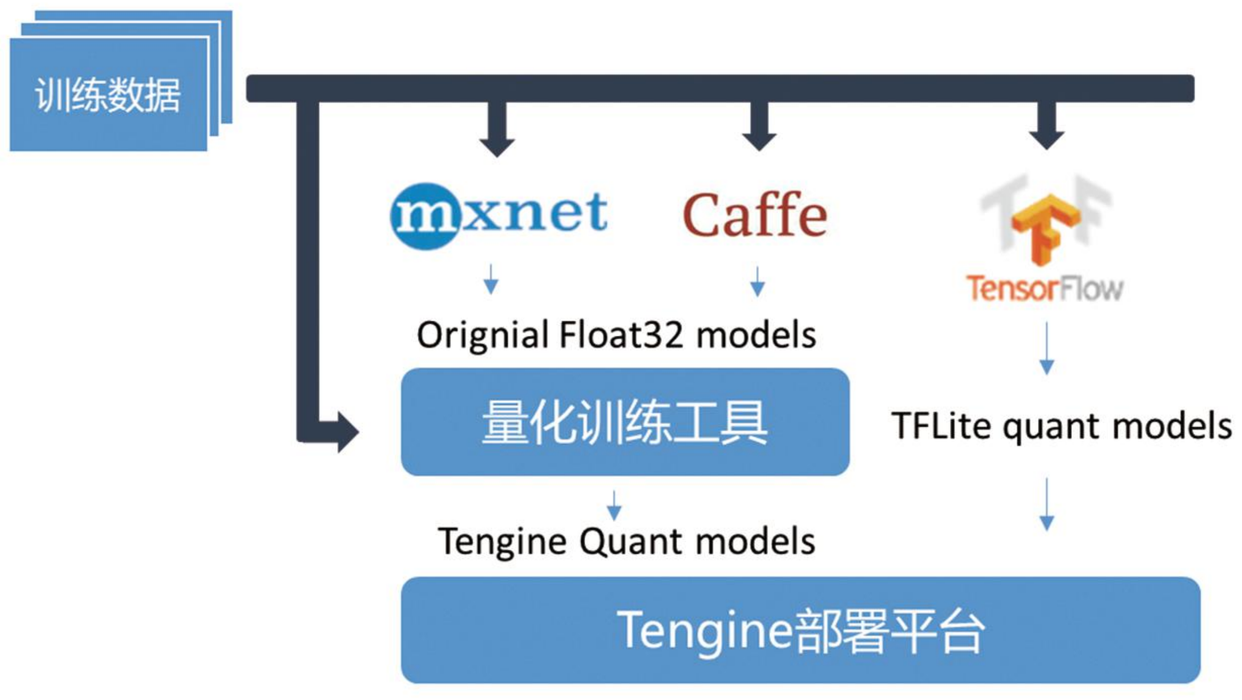
• 支持量化训练

众所周知采用 INT8(8位整型数)精度做AI推理是提升推理性能的一大利器，但这一利器却不是你想用就能用。因为当前主流AI训练框架都采用FP32(32位单精度浮点数)作为输出模型格式。因为没有配套的量化训练工具，AI开发者在部署时通常只能直接对浮点模型进行量化保存后再用于部署（训练后量化），这种浮点数到整型数的转换会带来模型的精度的损失，将直接导致最终的AI推理结果不可靠。尽管深度神经网络算法本身对这种量化误差有一定的容忍度，在某些情况下尤其是大网络的情况下对推理结果的影响偏差能够在接受范围内，但在那些对精度要求严格的场景如金融支付、人脸识别等应用场景下，这样的偏差是不能容忍的。解决量化误差的有效方法就是在训练时对前向计算进行量化建模，保证重训练时的权重更新充分考量到量化误差带来的影响，最终得到权重为整数的模型。

然而，对于量化训练技术，目前几大主流训练框架中只有TensorFlow提供了相关技术方案，且仅提供对TF模型的支持，其他框架都未发布提供量化训练技术方案。而OPEN AI LAB开发的量化训练工具， 可以支持对MXNet和Caffe模型进行重训练，彻底解决了开发者量化部署的难题。



异构计算示意图



Tengine量化训练工具使用流程

2.2 开放的应用开发平台

AIoT产业蓬勃发展离不开众多应用开发者的贡献与创新，为了方便广大开发者进行开发和使用我们最大限度的兼容了各种框架格式及主流的嵌入式芯片，让开发者能够自由选择训练框架与芯片平台。

Tengine应用开发生态

Tengine除了已经支持原生的TensorFlow， MXNet， Caffe的模型格式外，还通过ONNX模型实现了对Pytorch和PaddlePaddle的支持，下一步Tengine还会去支持Darknet与Kaldi模型。

2.3 跨芯片的统一开发平台

Tengine致力于为开发者提供跨硬件设备的统一的开发平台，因此Tengine在不同硬件设备上的API都尽可能保持一致。无论是资源极端受限的MCU还是常规的Cortex-A系列应用处理器，以及Arm中国周易AIPU，海思NNIE，瑞芯微RK3399Pro NPU，开发者都能通过简洁易用的Tengine API把不同的硬件算力调用起来。未来Tengine会持续去支持所有主流的AIoT芯片或加速器。

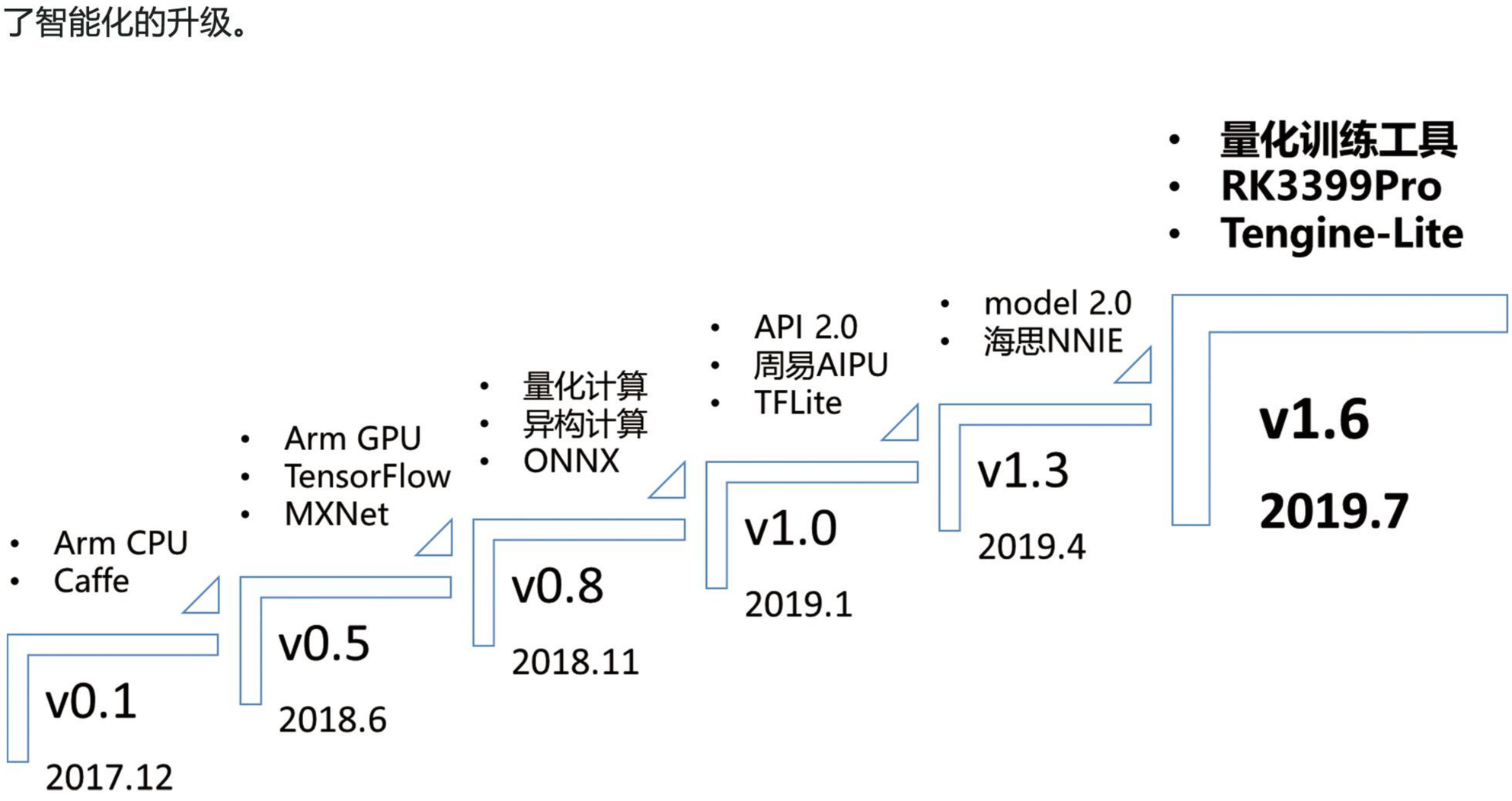
```
init_tengine();

// RK3288, Raspberry Pi
graph_t graph = creat_graph(NULL, “tengine”, model_file);

prerun_graph(graph)

run_graph(graph,1);

release_tengine();
```

```
// MCU  STM32F4 ..
creat_graph(NULL, “tiny”, model_mem);

// 周易AIPU
creat_graph(NULL, “zhouyi”, model_file);

// Hi3519av100, Hi3516cv500
creat_graph(NULL, “nnie”, model_file,model_config);

// RK3399Pro
creat_graph(NULL, “rk3399pro”, model_file);
```

在不同芯片间移植AIoT应用程序，只需要修改一行Tengine代码即可调用芯片算力

2.4 项目开发进度

Tengine项目从2017年年中立项，并于同年12月正式对外发布并开源，至今已经历经4个大版本的升级。最新的1.6版本已在2019年7月正式对外发布。在这近两年的时间里，Tengine的开发团队始终不忘初心、逐梦前行，致力于实现为AI开发者提供最好的开发工具的愿景并且取得了令人瞩目的成绩。目前Tengine已经成为了深受开发者喜爱的可靠的AIoT应用开发平台，数以百万计的设备通过Tengine实现了智能化的升级。



• 历史版本回顾

v0.1 支持Arm Cortex-A CPU推理，支持解析Caffe模型

v0.5 支持Arm GPU推理，支持解析TensorFlow与MXNet模型

v0.8 支持Arm Corte-A CPU混合精度计算与CPU/GPU异构计算，支持解析ONNX模型

v1.0 支持Arm中国周易AIPU，API升级为API 2.0，更好的支持AI加速器特性，支持解析TensorFlow-Lite模型

v1.3 支持海思Hi35系列NNIE加速器，Tengine模型格式升级为2.0，更好的支持量化与混合精度的数据模型

v1.6 支持RK3399Pro NPU加速器，推出Caffe与MXNet模型的量化训练工具，推出可在MUC上运行的Tengine-Lite

• 未来开发计划

未来Tengine将支持包括高通骁龙芯片在内的更多的带AI加速器的芯片平台，支持更多模型如Darknet，提供Pytorch模型转换工具及更好的量化训练工具，进一步提升开发者的体验并和广大开发者一起丰富Tengine的生态。

03 开发者资源

3.1 如何获取Tengine

为满足开发者的不同需要，Tengine提供了多个版本，包括Github开源版本，Tengine Explore开发者版，及EAIDK版本。

- EAIDK版

对于初次接触AIoT应用开发的开发者，我们推荐使用EAIDK开发套件。EAIDK是OPEN AI LAB推出的AI开发板，自带完整功能的Tengine预编译版本，并且提供大量应用案例，如自动辅助驾驶，人脸识别抓拍等。这些应用案例将帮助开发者快速掌握AIoT应用开发技能。访问www.eaidk.com获取EAIDK开发板。

- Tengine Explore开发者版

对于已经有一定应用开发基础，想开发自己AI应用的开发者，我们推荐使用Tengine Explore开发者版，这是专为Tengine高级开发者提供的包含完整功能的预编译版本，提供Tengine全部最新功能，包括所有CPU性能算子，GPU加速库，NPU加速库等。OPEN AI LAB会定期在Tengine开发者网站www.tengine.org.cn上发布包含树莓派等主流开发板的Tengine安装包。

- Github开源版本（https://github.com/OAID/Tengine）

对于已经熟练掌握AIoT应用开发，想进一步研究Tengine框架原理的开发者，我们推荐参考Github开源项目。这是Tengine框架的参考实现，包含了框架的主体代码与Armv8-A构架FP32精度的性能优化算子，支持OpenCL调用嵌入式GPU，同时也支持在x86平台上运行调试。

	开源项目	EAIDK	Tengine Explore开发者版本	
功能	FP32精度CPU算子 针对Armv8构架优化性能	完整版	完整版	
地址	https://github.com/OAID/Tengine	www.eaidk.com	www.tengine.org.cn	
支持硬件	Arm芯片	EAIDK-610 EAIDK-310	树莓派3B+ RK3399 RK3288 RK3399Pro Hi3519av100 Hi3516cv500	更多设备增加中 树莓派4B 全志A40i …

3.2 社区资源

- 获取Tengine最新信息

开发者网站 www.tengine.org.cn

- 案例分享与技术讨论

极术社区　www.aijishu.com

- 源代码学习与参考应用

https://github.com/OAID/Tengine

https://github.com/OAID/Tengine-lite

https://github.com/OAID/Tengine-APP

- 开发者技术交流

QQ群　829565581 (Question:Tengine Answer:openailab)

北京，上海，杭州，深圳开发者社团

OPEN AI LAB

开放智能机器（上海）有限公司，简称OPEN AI LAB，于2016年12月，由Arm中国发起成立 。OPEN AI LAB聚焦端侧人工智能，致力于推动芯片、硬件、算法、软件整个产业链的深度合作，加速人工智能产业化应用部署和应用场景边界拓展，为最终实现万物智能贡献力量。在学校教育领域，OPEN AI LAB依托“Arm 大学教育计划”，已经形成覆盖计算机视觉，语音，AIOT，SLAM等领域的综合实验室建设方案，并提供校外实习实践，专业共建等服务。

电话：021-80181176

邮箱：developer@openailab.com

网址：www.openailab.com

地址：上海市徐汇区宜州路188号B8栋3层

深圳市南山区科技园北区清华紫光信息港B栋801

北京市海淀区北四环西路58号理想国际大厦