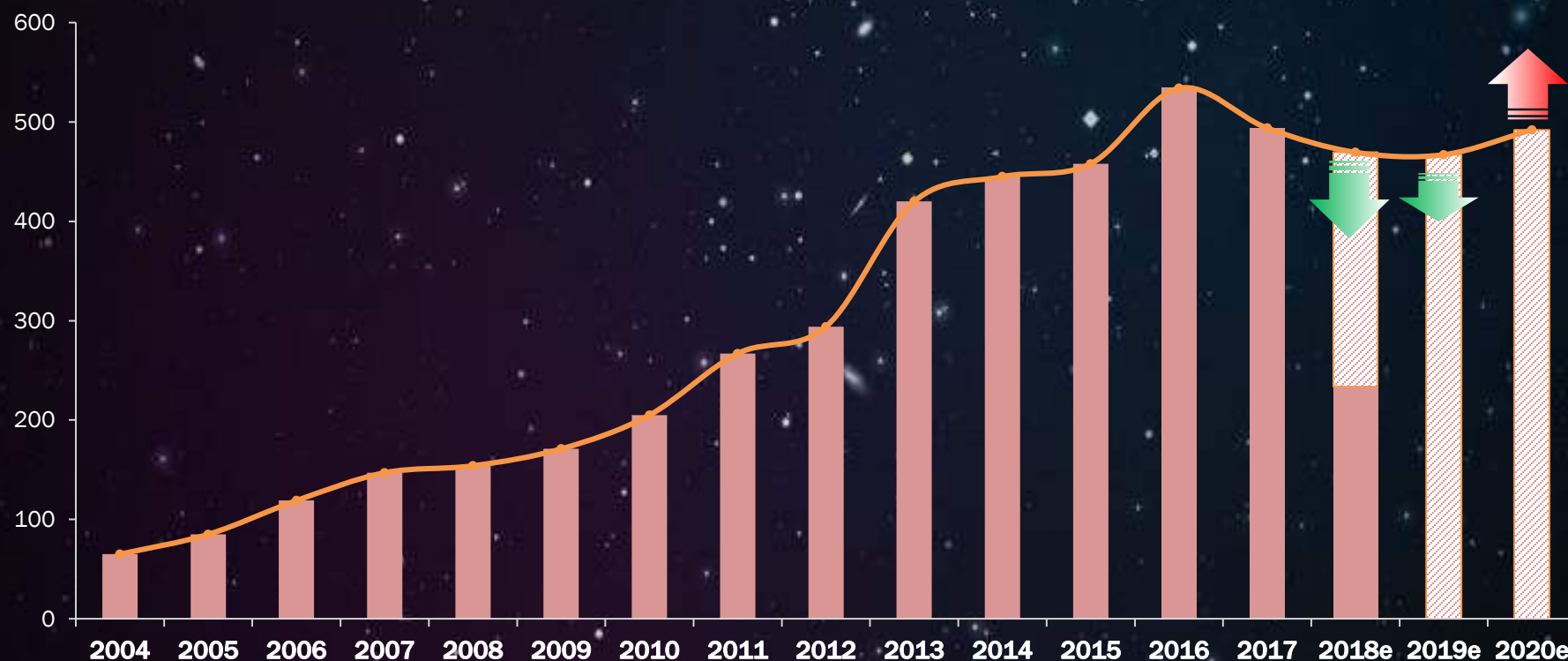


智能终端发展趋势

- ▶ 智能终端发展现状与瓶颈
- ▶ 终端领域技术演进趋势
- ▶ HiAI助力端侧AI发展前行

中国手机市场的发展变迁

中国手机市场已经由高速增长期过渡到存量换机稳定期；
均价持续上升；消费者的选择倾向更好的体验。



手机终端消费品形态的变迁

终端产业发展的两条主线：交互模式+信息服务模式

Feature Phone



Connect people to
people

1997

Smart Phone



Connect people to
mobile Internet

2007

Intelligent Phone



personal Assistant

2017

苹果引导的智能手机第一波创新红利已结束

中美市场的相反选择



大屏触控



计算性能

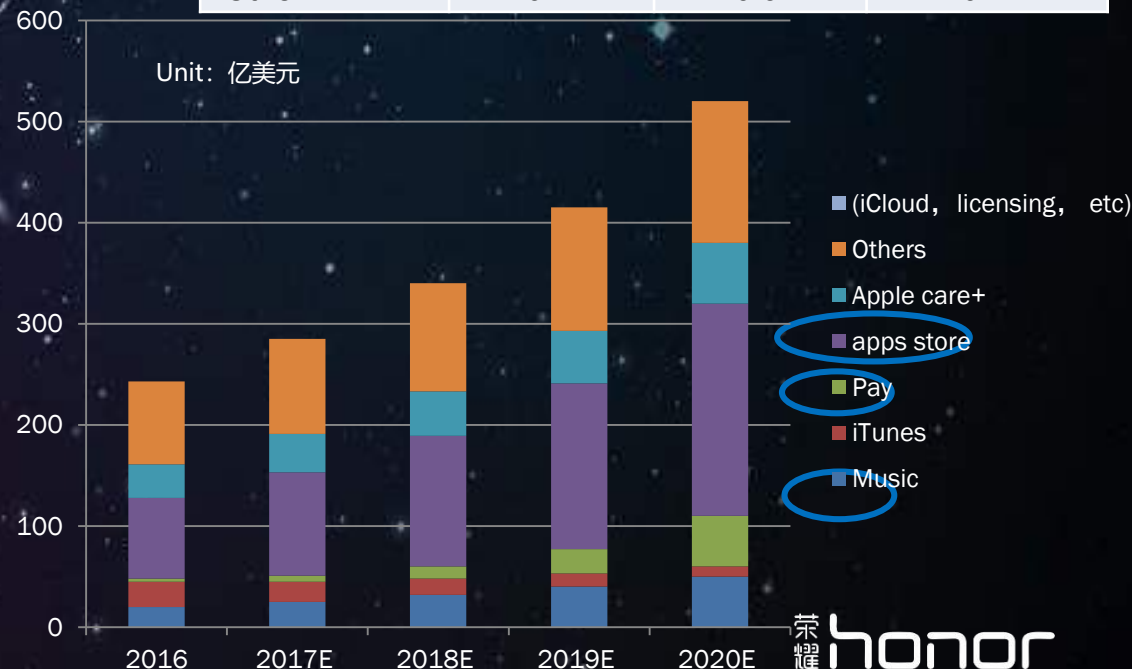


持久顺畅

- 苹果的市场份额有持续下降趋势
- 苹果带来的此波浪潮已近尾声：大屏触控、计算性能、操作流畅三大主要特性已无明显优势
- 苹果逐渐走向以iPhone为平台，主打服务的盈利模式
- 下一个革命性的突破，仍在探索中，呈多样化的趋势

China	Jan'16	Jan'17	% Change
Android	73.9	83.2	9.3
iOS	25.0	16.6	-8.4
Windows	0.9	0.1	-0.8
Other	0.3	0.1	-0.2

USA	Jan'16	Jan'17	% Change
Android	58.2	56.4	-1.8
iOS	39.1	42	2.9
Windows	2.6	1.3	-1.3
Other	0.1	0.3	0.2



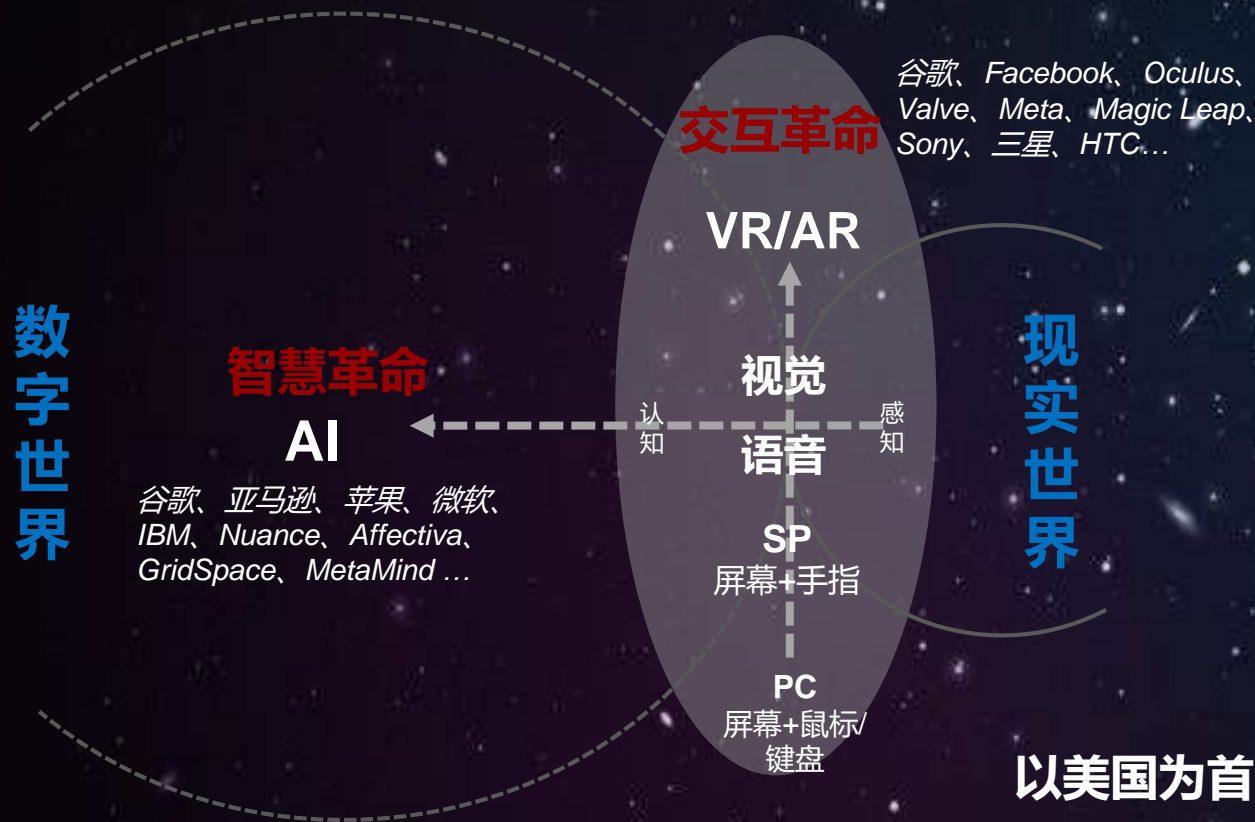
手机终端主要功能配置遭遇发展瓶颈



- ▶ 智能终端发展现状与瓶颈
- ▶ 终端领域技术演进趋势
- ▶ HiAI助力端侧AI发展前行

5-10年内颠覆式创新将带来用户体验的再次跨越式升级

终端用户体验将发生最基础性的改变：(1) 交互的革命；(2) 信息服务的革命——人工智能



新一代终端用户体验

1. 更自然的交互（触屏 + 智能语音 + 动作识别 + 计算视觉）
2. 超越现实的显示（360度、沉浸式、增强现实）
3. 个性化的智能服务（自动场景识别、高智能推荐、无时无刻服务）

以美国为首的技术创新公司正在推动产业的人工智能与交互革命，每一场革命都有可能成为改写产业的“黑天鹅”

未来通信革命



2018 广东互联网大会

2018 GUANGDONG INTERNET CONFERENCE

同期举办：2018 全球未来科技大会（中国·广州站）

全息通信

Holographic

Communication

感知通信

Perception

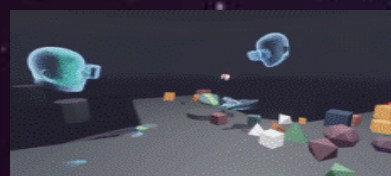
Communication



Facebook next G VR social and communication



Voice Interaction



Perception Interaction



Simple Avatar



Real person Avatar



Realtime holoportation Capture, transport, playback, record, edit..

Record



Digital expression emotion

Playback



Smartphone with depth camera



荣耀 honor

未来通信对ICT管道的影响：将带来100M到G比特的带宽需求（大带宽、低时延）

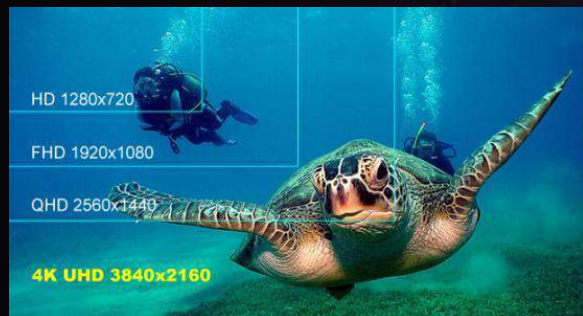


2018广东互联网大会

2018 GUANGDONG INTERNET CONFERENCE

同期举办：2018 全球未来科技大会(中国·广州站)

- 平面数据是现有视频系统的数据基础，是传统H.264/265解决的压缩问题
- 球面数据是现在VR视频应用的数据基础，是工业界的热点
- 体数据是未来AR/VR应用的数据基础，是学术界研究的热点，工业界也开始跟进
- 海量存储，超大带宽（M2M）：AR/VR相比2D视频带来3-4倍存储需求，5D光场技术带来4-6倍数据，HMD头盔清晰度改进为8K带来4-5倍数据，总计40 x – 100x的海量数据存储、传输和处理的需求



平面数据（2D）

- 一帧4K超高清图像24MByte
- 4K@30fps视频 11Gbps
- 采用H.265压缩后20Mbps



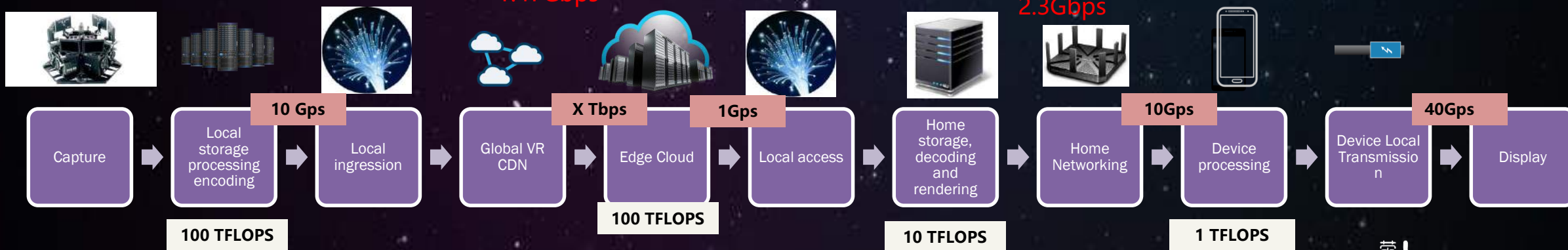
球面数据（2.5D 全景）

- 按照未来16k的极限分辨计算（16kx8k），一帧球面图像将达到384MByte
- 120fps，达到368Gbps
- 参考H265压缩能力在250x，压缩后带宽1.47Gbps



体数据（3D 可全球移动）

- 按照能够在1m的距离内达到视网膜分辨率的要求，在2mx1mx1m空间范围描述一个人的表面积计算，一帧体数据图像将达到1.2GByte
- 120fps，达到1.15Tbps
- 即便按照更高的500x压缩能力，压缩后带宽依然达到了2.3Gbps



端到端的带宽需要100Mbps以上才能达到真实的临场感

荣耀 HONOR

AR/VR对终端行业的影响

AR/VR将是继智能手机后，消费者领域最大的产业变革

智能革命

计算智能 → 感知智能 → 认知智能



输入：触屏 → 语音 + 手势 + 视觉交互
输出：5" 屏 → 100" 屏 → 360° 全浸入

人机交互革命

VR: 感知世界 AR/MR: 认知世界+沟通世界+情感连接



手机/平板/PC/电视等都可能被AR逐步替代，
引起芯片/OS/生态等革命式产业颠覆，引起行业巨头新一轮竞争

 **ARM**
高性能 高功耗 中性能 低功耗

高性能 低功耗
AlwaysOn

芯片架构之争

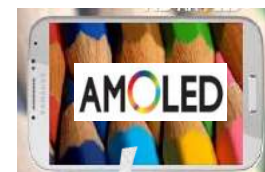
高通/英伟达/Intel

 Microsoft HoloLens
PC/生产力
 Google Tango
Mobile/信息获取

生产 消费 娱乐...

OS&生态之争

谷歌/微软/百度



显示产业颠覆

三星/苹果/MagicLeap



服务入口颠覆

谷歌/微软/FB/百度

中期(~5年)AR/VR将推动手机及PC升级换代，长期(~10年)将可能颠覆一切带屏设备 

- ▶ 智能终端发展现状与瓶颈
- ▶ 终端领域技术演进趋势
- ▶ HiAI助力端侧AI发展前行

端侧AI已成为业界公认的发展趋势

「Apple Jeff Williams」我们放在手机手表里的神经引擎，对未来至关重要，这些将帮助开发者在AI领域创造越来越多的应用，所以我们认为手机将是一个主要的平台

「Qualcomm Jeff Gehlhaar」手机将是未来五年人工智能最为普遍的载体

「Google Jeff Dean推荐」为什么说未来的深度学习是小、轻、快？

19 Mar 2018 | 13:00 GMT

Smartphones Will Get Even Smarter With On-Device Machine Learning


It's time for deep learning algorithms to come down from the cloud and get into your gadgets





View from the Marketplace ?

On-Device Processing and AI Go Hand-in-Hand

As on-device processing becomes more powerful, and AI grows more prevalent, our future will increasingly be defined by the convergence of these two game-changing trends

by MIT Technology Review Insights March 13, 2018

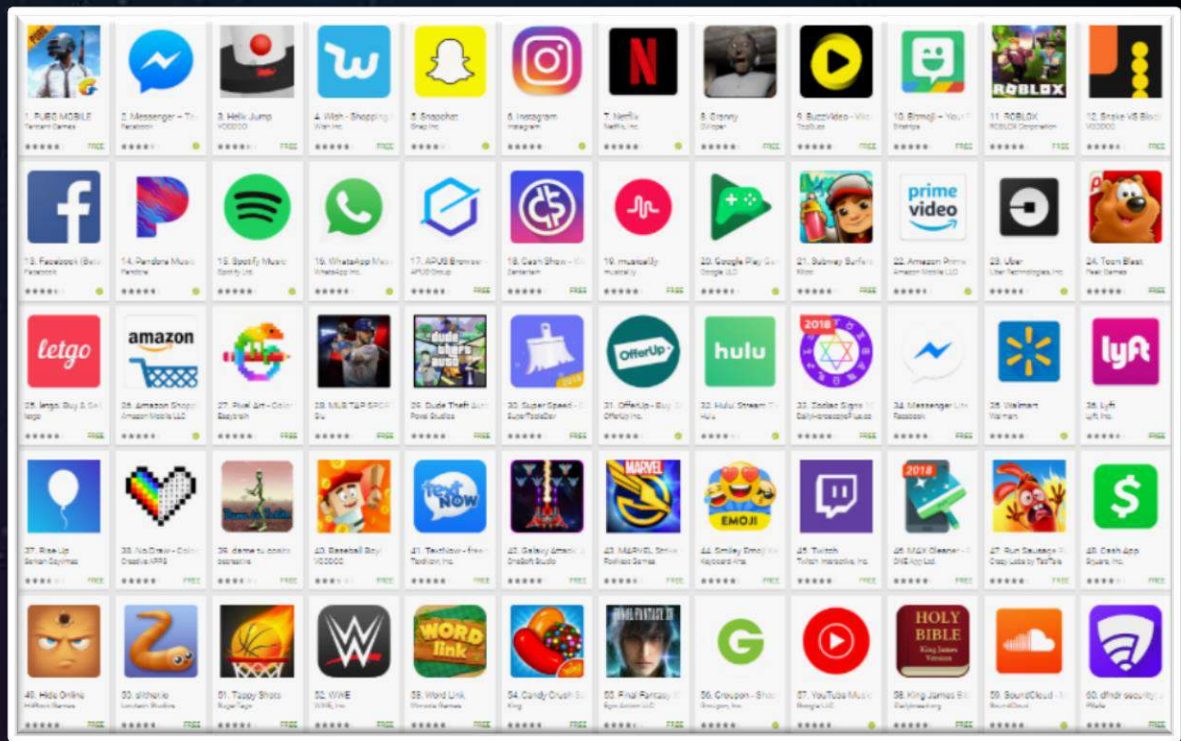
 Help Net Security
January 4, 2018

80% of smartphones will have on-device AI capabilities by 2022

端侧AI应用面临着巨大的挑战

- ❖ 计算密集、复杂， 计算需求巨大， 实时性非常挑战
- ❖ 运行环境受限， 功耗、内存、存储空间非常挑战
- ❖ 越来越多的应用都带AI， 应用场景不确定
- ❖ 模型和算子变化快
- ❖ 前端训练平台五花八门



HiAI Foundation架构



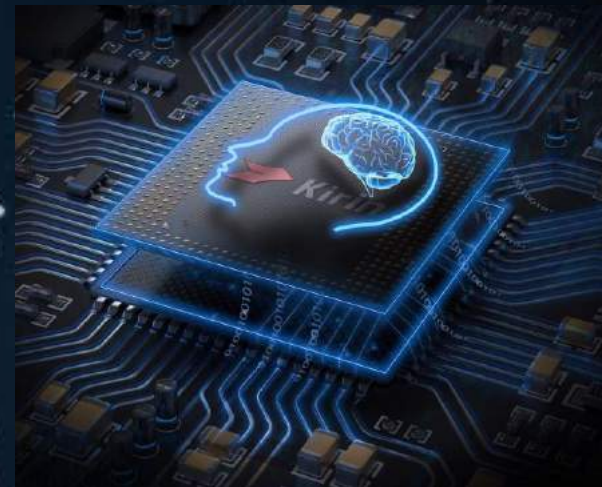
Cloud

HUAWEI HiAI Service



Device

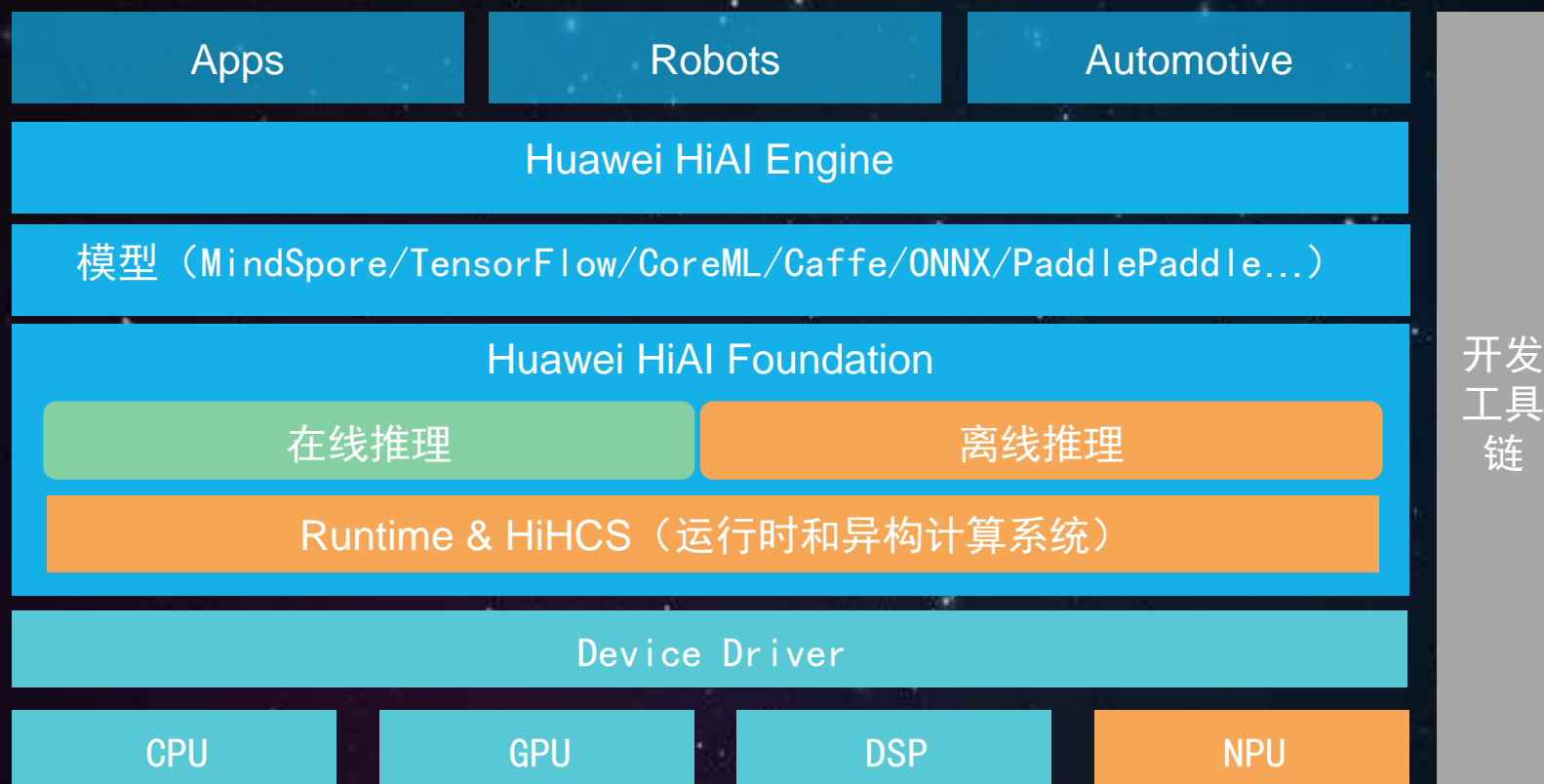
HUAWEI HiAI Engine



Chip

HUAWEI HiAI Foundation

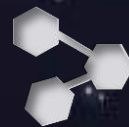
HiAI Foundation架构



工具链



完善的文档



丰富的API



直观的DEMO

HiAI Foundation带给开发者&用户的价值

 2018广东互联网大会
2018 GUANGDONG INTERNET CONFERENCE
同期举办：2018 全球未来科技大会(中国·广州站)



实时



隐私



成本

实时多人姿势识别

 2018 广东互联网大会
2018 GUANGDONG INTERNET CONFERENCE
同期举办：2018 全球未来科技大会(中国·广州站)



荣耀 honor

成功案例-Prisma

 2018 广东互联网大会
2018 GUANGDONG INTERNET CONFERENCE
同期举办：2018 全球未来科技大会(中国·广州站)



荣耀 honor

成功案例-抖音




普通



HiAI使能

HiAI Foundation可以赋能极其丰富的端侧应用场景

	短视频、直播	人脸识别、手势识别、人像分割、人体姿势识别、视频风格化
	社交平台	照片分类、图像识别、图像超分辨
	AR	深度估计、光线估计、环境理解、SLAM
	拍照、修图	美颜、图像增强
	购物	识图购物
	翻译、文字处理	拍照翻译、OCR、分词、命名实体识别、文字情绪识别、文字智能回复

- 关键技术一：专用指令集和计算库，高效执行神经网络算子



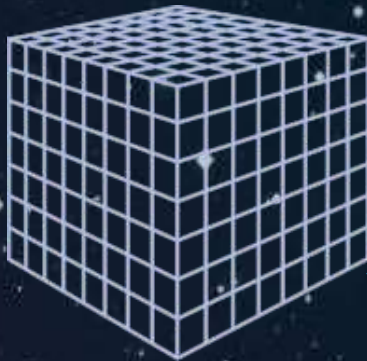
CPU=标量

- 通用计算
- 逻辑控制



GPU=矢量（2D）

- 图形图像处理计算与渲染
- 大规模并行计算



NPU=张量（ $\geq 3D$ ）

- 专用AI指令集
- 更大规模并行计算

Convolution
Deconvolution
Pooling
Relu
Normalize
BatchNorm
FullConnection
Sigmoid

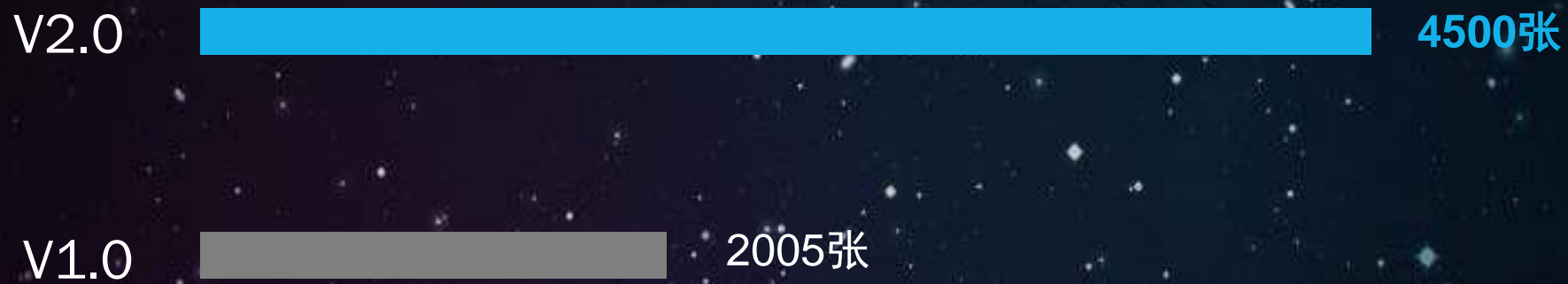
- 关键技术二：离线编译，轻量部署；层间融合，快速推理



- 关键技术三：离线模型运行，local ram，高效节能



HiAI Foundation V2.0性能大幅提升



HiAI Foundation V2.0解锁能力更强



HiAI Foundation 版本迭代快

	V1.0 麒麟970	V1.5 麒麟970	V2.0 麒麟980
框架与API	<ul style="list-style-type: none">○ Caffe/TensorFlow○ Huawei HiAI API	<ul style="list-style-type: none">○ Caffe/TensorFlow○ Android NN	<ul style="list-style-type: none">○ Caffe/TensorFlow、Android NN○ 模型并发运行○ CPU/NPU混合模型调度
算子兼容性	<ul style="list-style-type: none">○ 支持42个算子	<ul style="list-style-type: none">○ 支持90个算子	<ul style="list-style-type: none">○ 支持147个算子
工具链	<ul style="list-style-type: none">○ 命令行工具	<ul style="list-style-type: none">○ 增加图形化工具(IDE)○ 增加轻量模型校验&转换工具	<ul style="list-style-type: none">○ 更多功能的图形化工具(IDE)○ 更加精准的轻量模型校验&转换工具○ INT8量化工具○ 混合模型分割工具
前后向兼容		<ul style="list-style-type: none">○ 增加HiAI Foundation版本与模型兼容性检测○ 增加模型在线编译	<ul style="list-style-type: none">○ 增加INT8模型兼容性检测○ 增加INT8量化模型在线编译

HiAI Foundation V2.0模型分析工具化，告别手工核查

- IDE图形化工具，在IDE开发环境中支持模型自动算子检查和转换
- 提供轻量化模型预分析工具，无需安装IDE也可对模型预分析和转换



模型分析报告

分析时长 00:02 提交时间 2018-06-19 16:48:04 通过算子 785/788 版本 DDK v150

全部 788 通过 785 未通过 3

Reshape 1 Softmax 1 Squeeze 1

序号	算子类型	算子名称	检查结果	建议
785	Squeeze	InceptionV3/Logits/SpatialSqueeze	不支持	不支持该算子。对于 Array 类型算子，请使用：Concat, PadV2,
787	Softmax	InceptionV3/Predictions/Softmax	不支持	不支持非最后一层的 Softmax 层
788	Reshape	InceptionV3/Predictions/Reshape_1	未通过约束检查	1. 不支持 Reshape 作为最后一层;

```
C:\ModelTest>java -jar ./OperatorsCheck.jar -t tensorflow -p ./inceptionv3.pb
Running OS: windows 7
模型类型: Tensorflow
模型路径: ./inceptionv3.pb
输出路径: C:\Users\w00216018\devecoide\hiainmodel\optimizer

开始检查 DDK v100 相关算子
2018-06-13 16:45:06.148872: I tensorflow/core/platform/cpu_feature_guard.cc:140] Your CPU supports instructions t
mpiled to use: AVX
· Squeeze: 不支持该算子。对于 Array 类型算子，请使用：Concat, Const, Identity, Placeholder, ConcatV2 或者 Shape
· Reshape: 不支持该算子。对于 Array 类型算子，请使用：Concat, Const, Identity, Placeholder, ConcatV2 或者 Shape
更多信息请查看: C:\Users\w00216018\devecoide\hiainmodel\optimizer\w100\report.html

该模型未通过DDK v100 算子检测

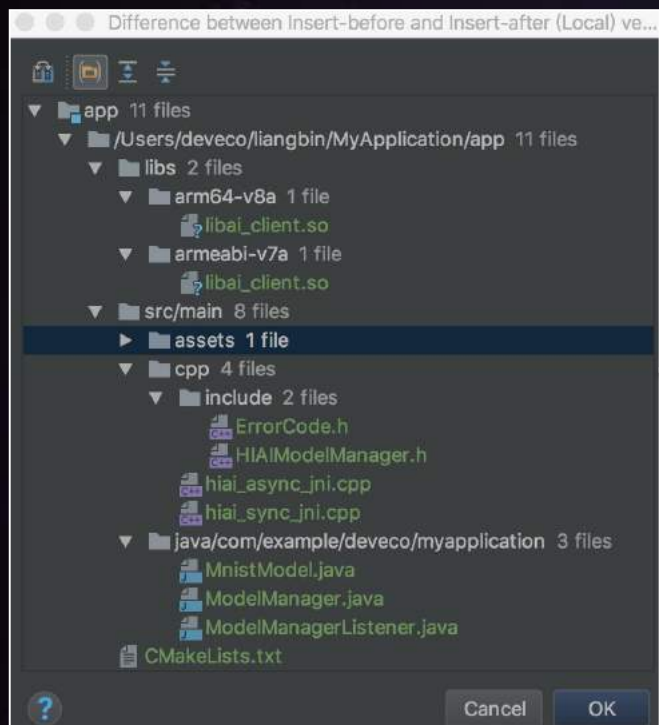
开始检查 DDK v150 相关算子
· Squeeze: 不支持该算子。对于 Array 类型算子，请使用：Concat, PadV2, Const, Slice, Split, Identity, Placeholder,
· Reshape: 1. only support faltten, not reshape;
更多信息请查看: C:\Users\w00216018\devecoide\hiainmodel\optimizer\w150\report.html

该模型未通过DDK v150 算子检测

Complete!
```

IDE离线模型转换成功后：

◆ 自动添加资源 (libai_client.so、模型文件等)



◆ 自动生成代码 (jni、模型封装)

```
package com.example.deveco.myapplication;

import android.content.res.AssetManager;
import android.util.Log;

public class MnistModel {

    /** user load model manager sync interfaces ****/
    public static int load(AssetManager mgr){
        return ModelManager.loadModelSync( modelName: "Mnist", mgr);
    }

    public static String[] predict(float[] buf){
        return ModelManager.runModelSync( modelName: "Mnist", buf);
    }

    public static int unload(){
        return ModelManager.unloadModelSync();
    }

    /** load user model async interfaces ****/
    public static int registerListenerJNI(ModelManagerListener listener){
        return ModelManager.registerListenerJNI(listener);
    }

    public static void loadAsync(AssetManager mgr){
        ModelManager.loadModelAsync( modelName: "Mnist", mgr);
    }

    public static void predictAsync(float[] buf) {
        ModelManager.runModelAsync( modelName: "Mnist", buf);
    }

    public static void unloadAsync(){
        ModelManager.unloadModelAsync();
    }
}
```

• 离线模型shape信息可视化

• 自动统计模型管理接口耗时

The background of the entire image is a deep space scene. In the upper left, a portion of a galaxy with a bright yellow and orange core is visible against a dark blue and black sky filled with distant stars. The central focus is a black hole, depicted as a dark, circular event horizon. Surrounding it is a glowing, turbulent accretion disk with swirling patterns of orange, yellow, and red. A powerful, blue, ethereal beam of light or energy originates from the top left and terminates in a bright white point just above the black hole's horizon. The Chinese text '有朋友 有未来' is superimposed in white, bold characters across the middle of the image, partially overlapping the blue beam and the glowing disk.

有朋友 有未来