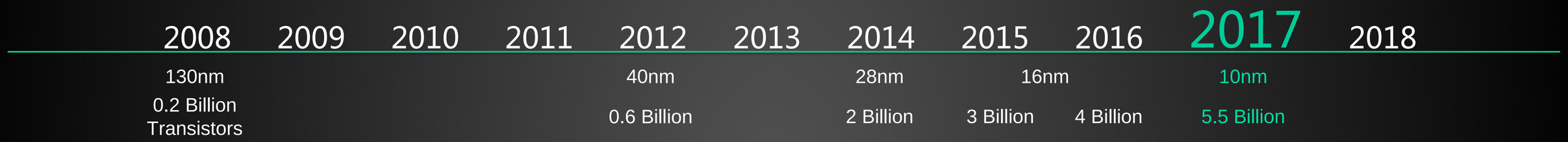
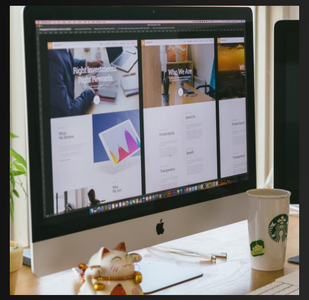
Huawei Smartphone AI

Mikko Terho
VP Technical Planning and Site Manager
Huawei R&D Finland

Oliopäivät 7.12.2017 TUT Tampere

A Decade of Kirin Evolution





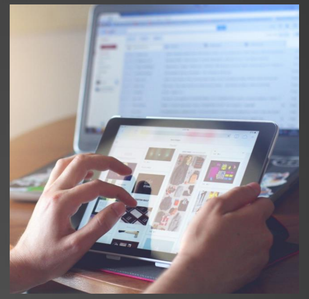
64bit

Fingerprint

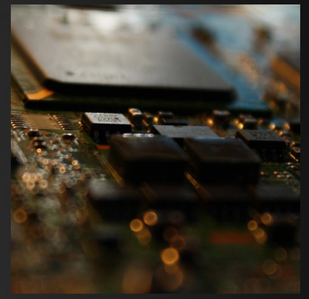
Dual Camera

4G

VoLTE



FinFET



SoC Strategy

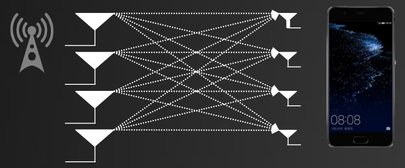
– Continuous Integration, Heterogeneous innovation

User Experience First

– To enhance the user experience as the primary purpose

3 Key Aspects in Mobile SoC - Communications

Technologies to improve the data transmission speed

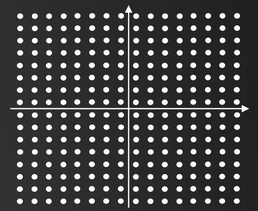


4*4 MIMO

Compared to 2 * 2MIMO, the same bandwidth, Peak rate increased by 1 times.
Improve the user experience in the weak signal scene

5CC CA

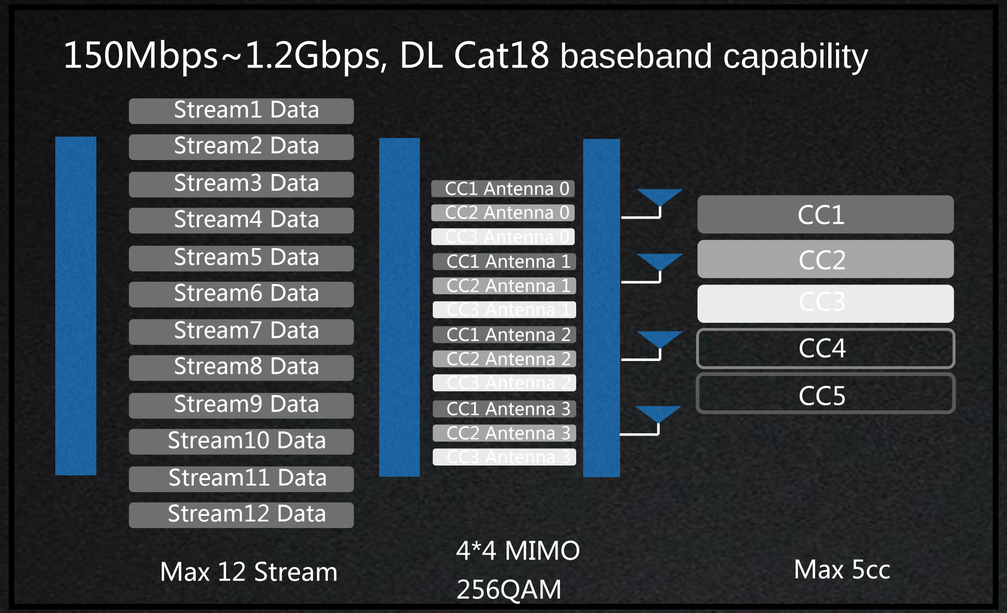
The user peak rate increases linearly with the number of carriers
Aggregate operators' scattered carrier resources



256QAM

Improve the efficiency of spectrum utilization
Compared to 64QAM, 33% increase in peak rate

LTE Category 18



Cat18 1.2Gbps

- 4x4MIMO+256QAM+DL 3CC

Cat16 1Gbps

- DL 5CC + 256QAM

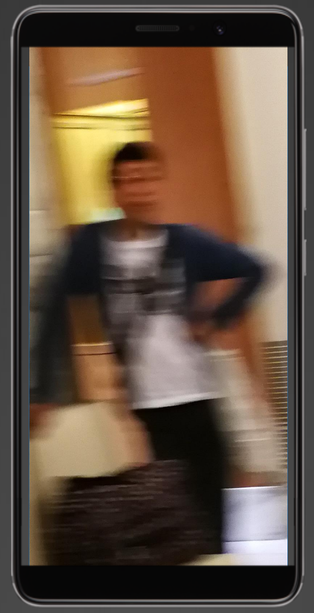
Cat12 600Mbps

- DL 4CC

3 Key Aspects in Mobile SoC - Multimedia

Frustrated with photo taking?

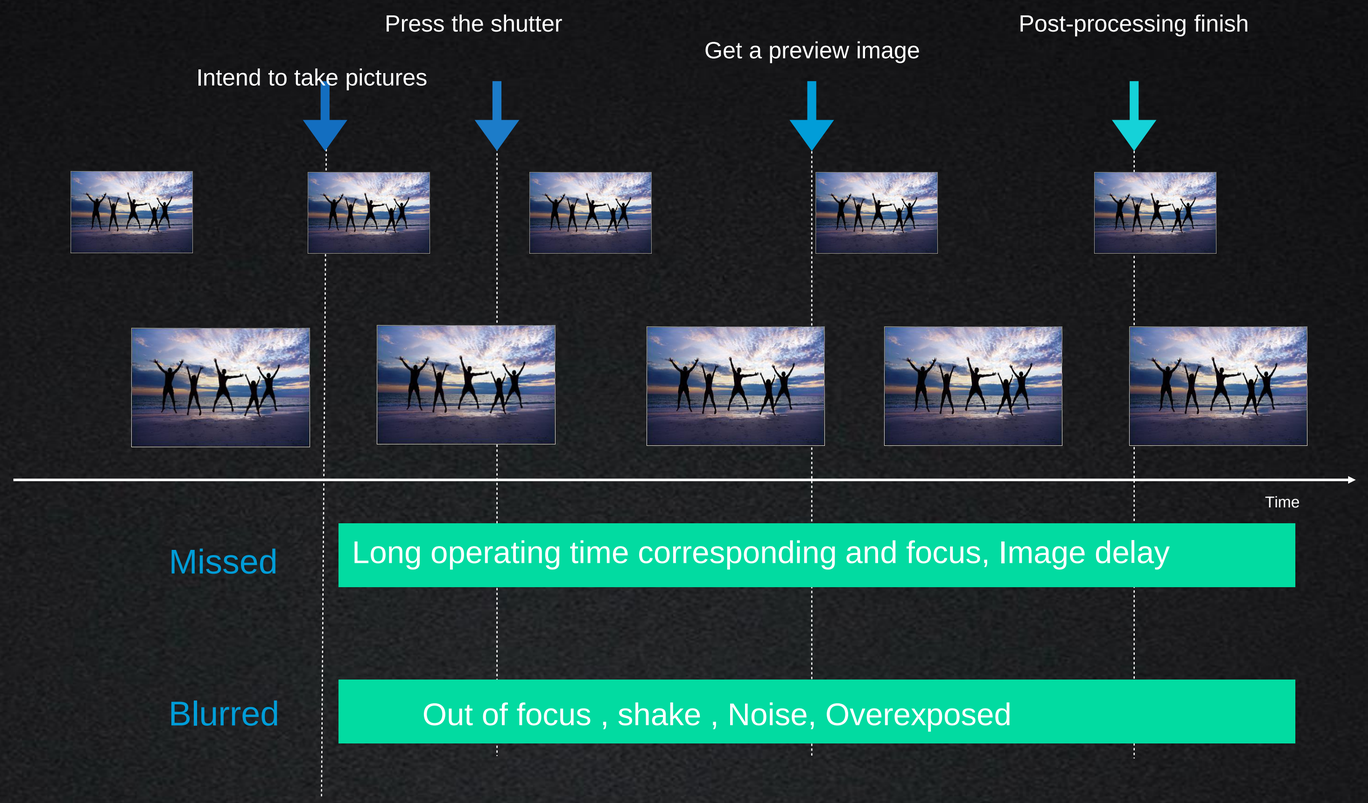
Missed



blurred

Not Clear

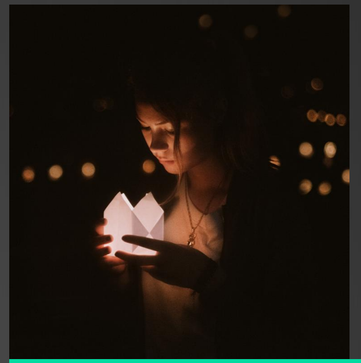
The reasons for these unsatisfied photos ?



The whole process to enhance the camera experience

Fast response

ISP enhance throughput by 25%
Camera Processing response increased by 30%
Camera dual channel parallel processing
ZSL in Color-monochrome dual camera



Fast focus

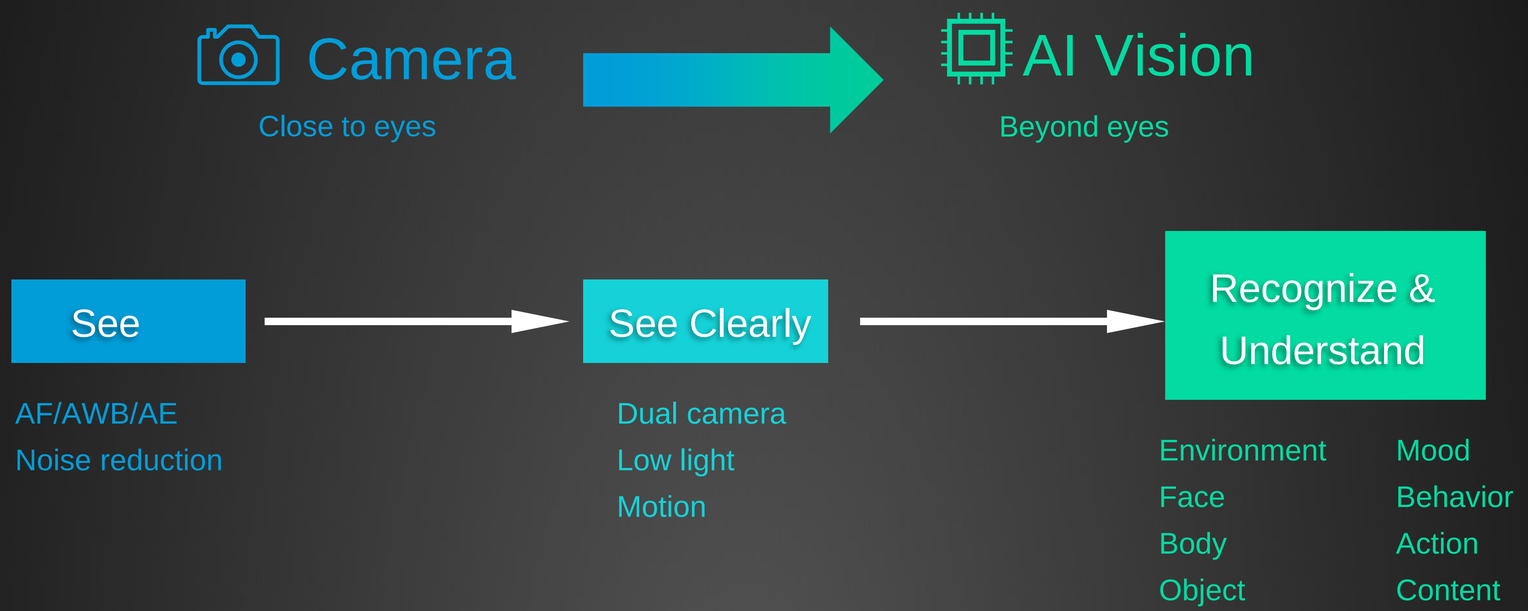
4-Hybrid Auto-focus
Focus self-calibration capability
Face tracking ,Point light , Flat area focusing optimization

Motion detection

Motion detection , include static , slow, medium and fast
Hardware-based face detection
Intelligent camera scene detection

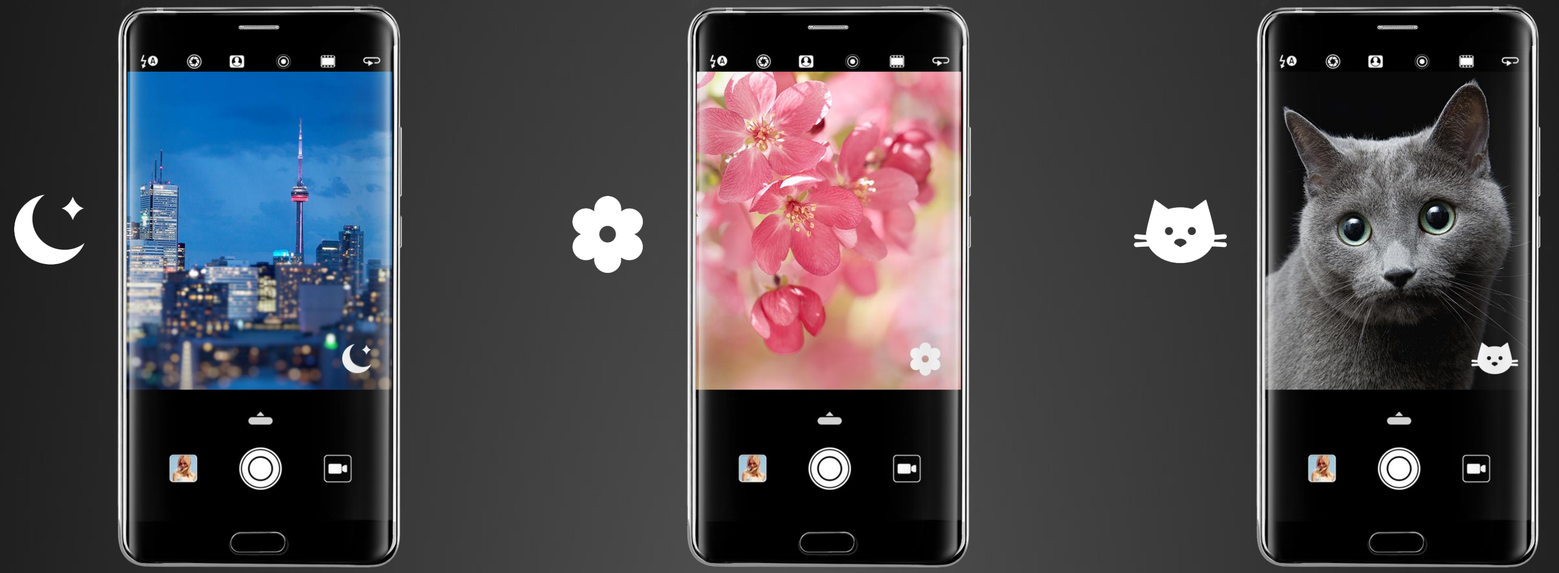
Low light process

Improved noise reduction low light, stage lighting and other scene detection
Improved low light camera strategy


Camera
Close to eyes
AI Vision
Beyond eyes
See
See Clearly
Recognize & Understand
AF/AWB/AE
Noise reduction
Dual camera
Low light
Motion
Environment
Face
Body
Object
Mood
Behavior
Action
Content

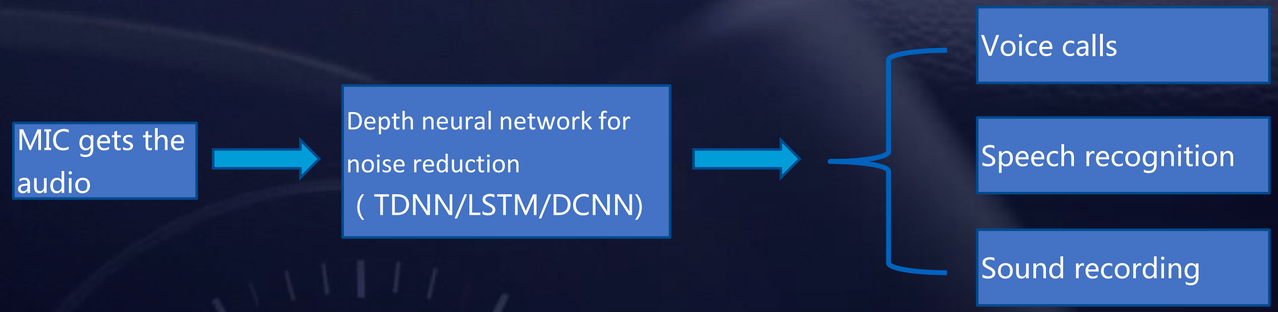
AI Vision engine: AI Camera

Identify top 14 scenarios and automatically adjust camera parameters for better pictures



AI Noise Reduction

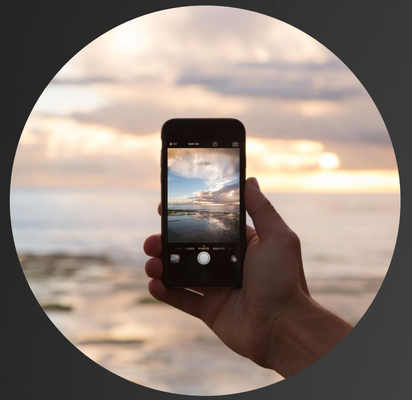
to ensure the quality of audio from the source



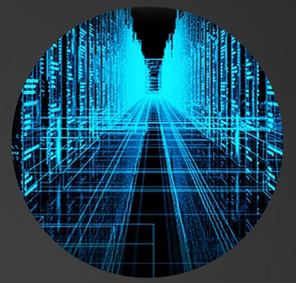
The voice recognition in heavy traffic noise raised from 80% to 92%

All data and benchmark results are based on internal testing. Results may vary in different environments.

3 Key Aspects in Mobile SoC - Intelligence



Mobile AI



Knowledge

Processing

Real-time processing where knowledge is the key

Mobile AI Knowledge Models

Big Data

Training

Update

Common
- Image
- Voice
- text

Personal
- Bioinformatics
- Voice
- Sensor
- Behavior

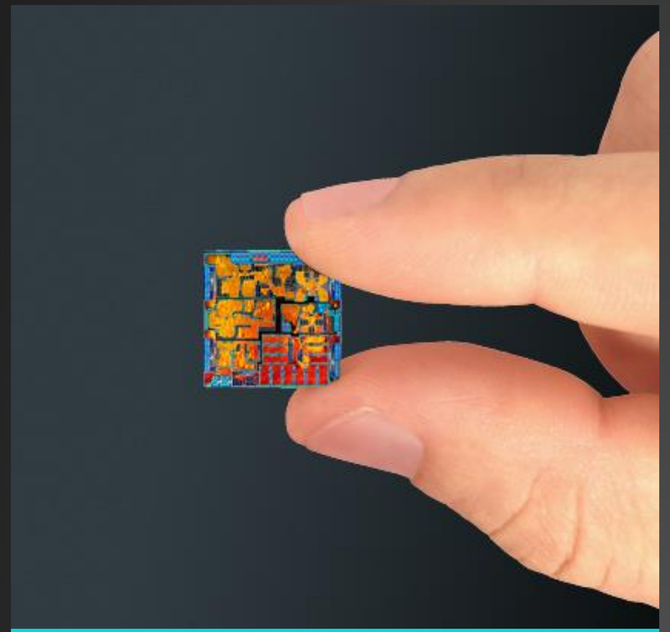
Overcoming the Challenges of Mobile AI

Space

Thermal

Power

Need to meet the physical requirements



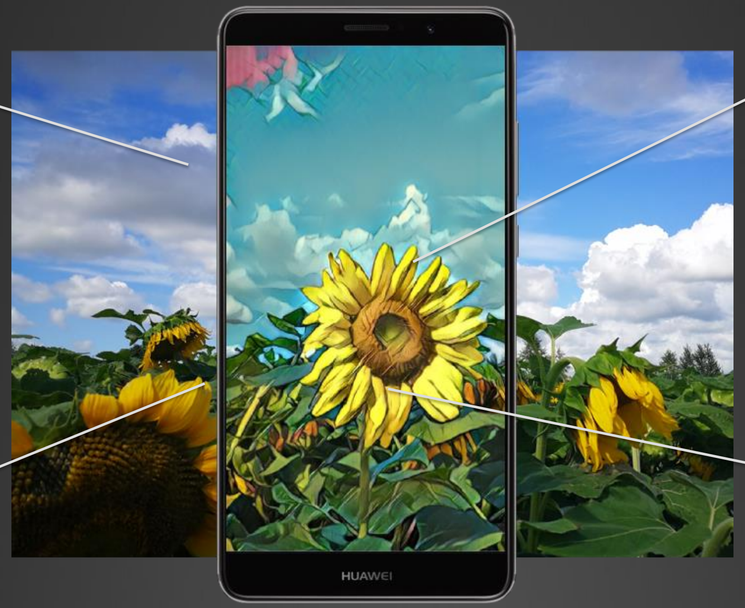
The Size

5.5 Billion Transistors 10x10mm

Energy

2000-4000mAh for the whole device

Accelerating the AI App



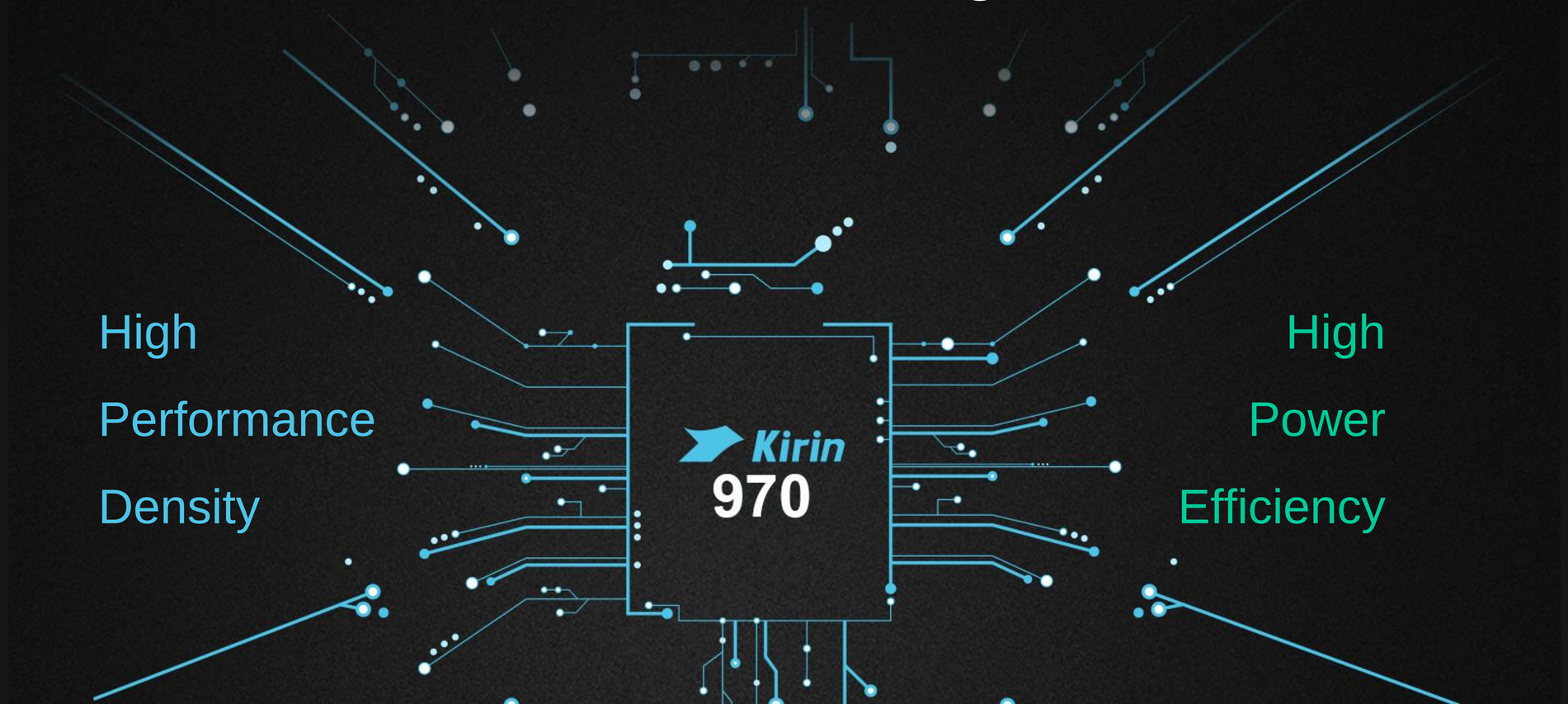
CPU
- Task scheduling
- Load balance
- Memory allocating

NP
- AI Computing

GPU
- UI rendering
- Graphics processing

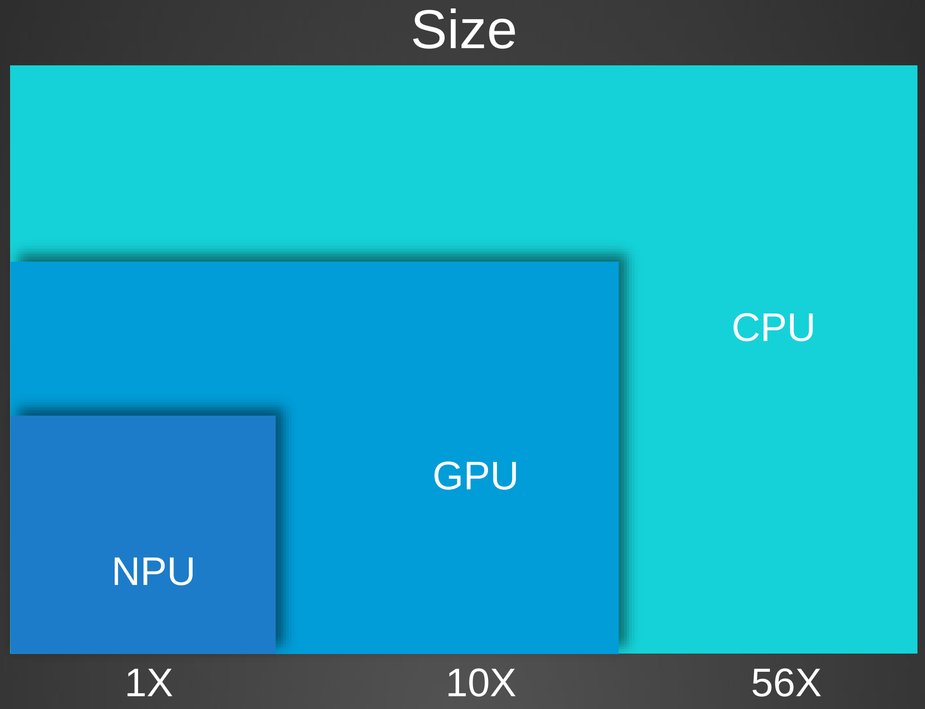
ISP/DSP
- Camera 3A
- Image processing

NPU（Neural Network Processing Unit）



High Performance Density of Kirin 970 NPU

25X Performance = Half of CPU

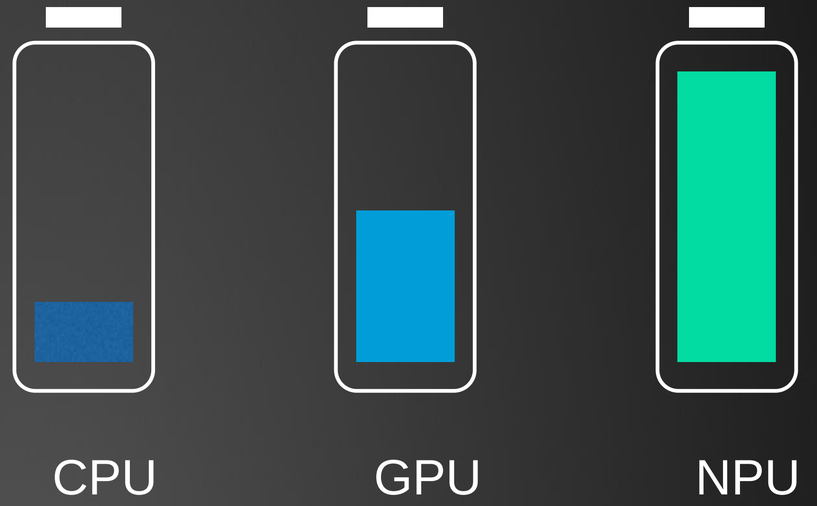


NPU: Neural Network Processing Uni

High Power Efficiency of Kirin 970 NPU

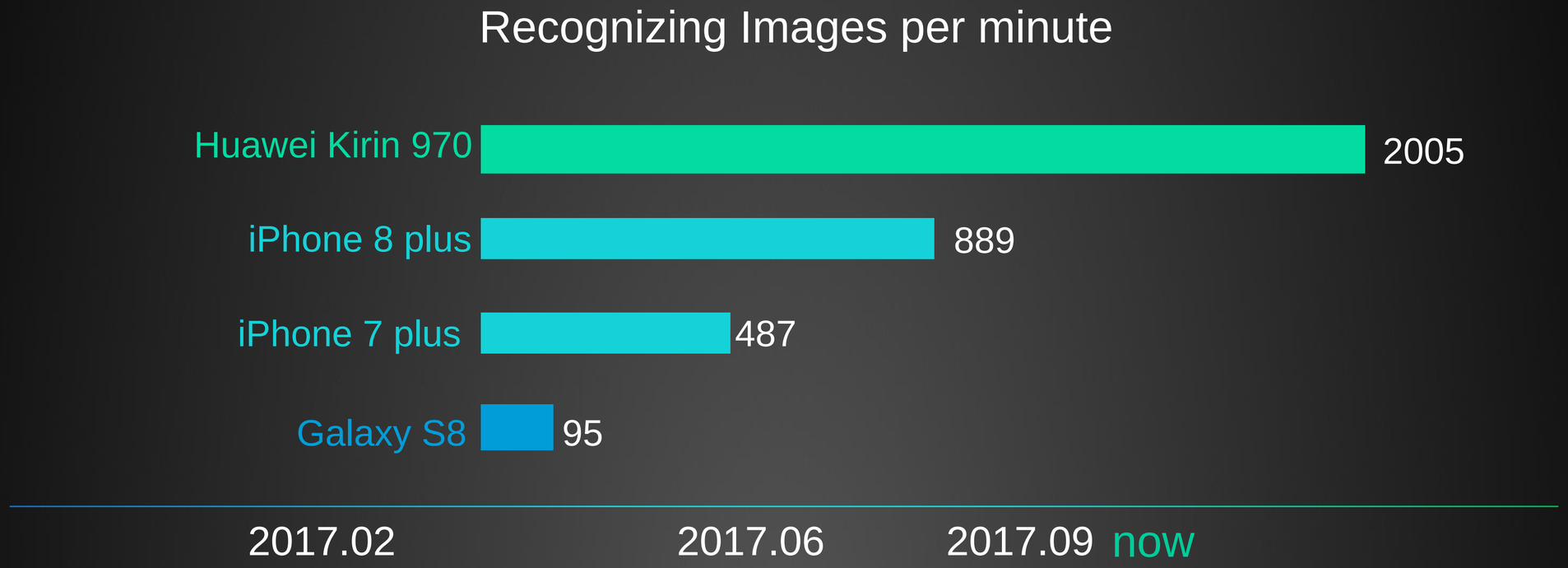
Power consumption is only 1/50 of the CPU

1/50



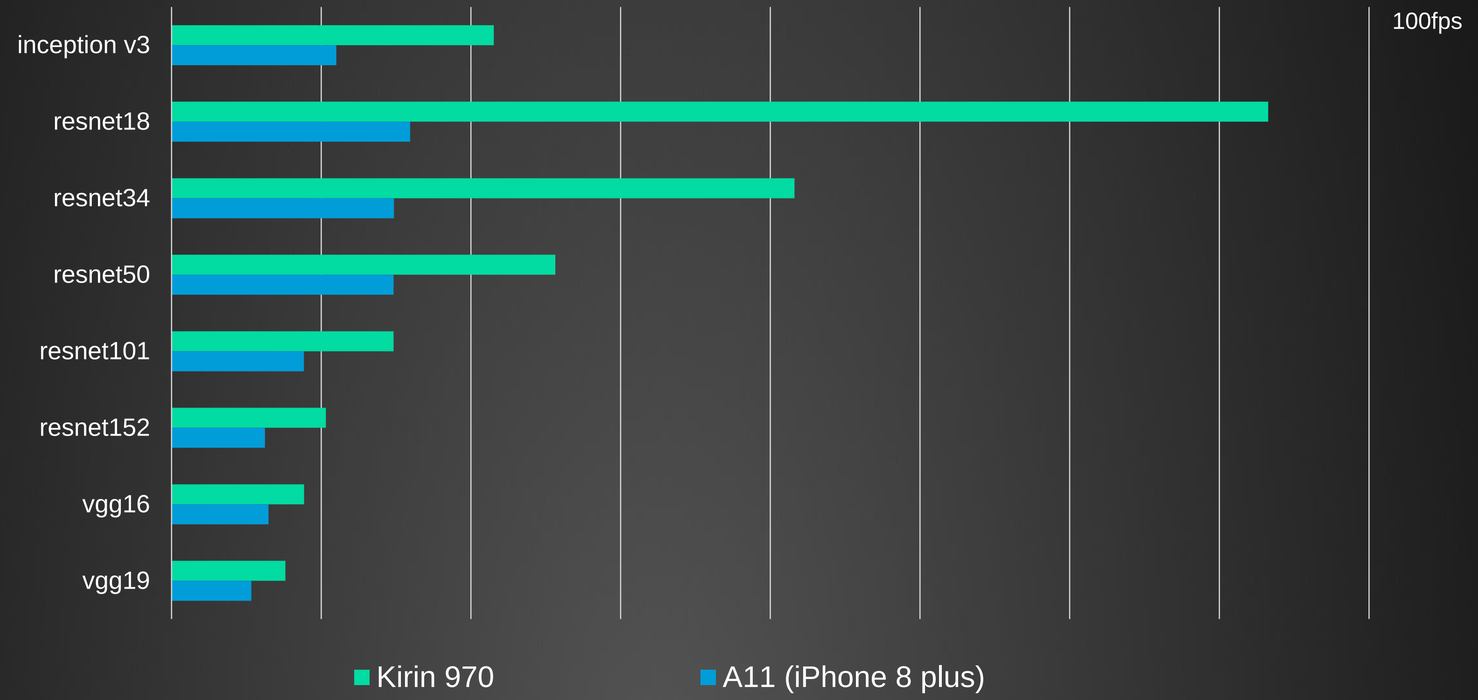
NPU: Neural Network Processing Unit

Rapid increase in AI performance



All data and benchmark results are based on internal testing. Results may vary in different environme

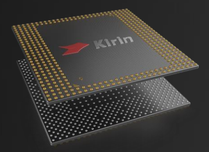
More Performance Comparison



All data and benchmark results are based on internal testing. Results may vary in different environments.

How to Bring NPU benefits to end-users?

Top users requirements



EMUI AI ENGINE

Vision Engine

Performance Engine

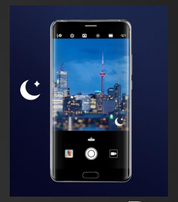
Experience Engine

Energy Engine

3rd parties App Engine

TensorFlow

Caffe

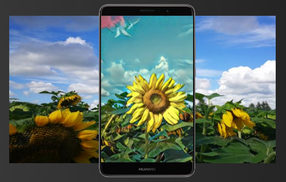


Better Camera

Better Performance

Better Experience

Longer Battery Life



Better 3rd party Apps

Huawei can provide End-to-End AI solution (Chipset, Device, Cloud)

Screen

Camera

Mic.

Headset

Sensors

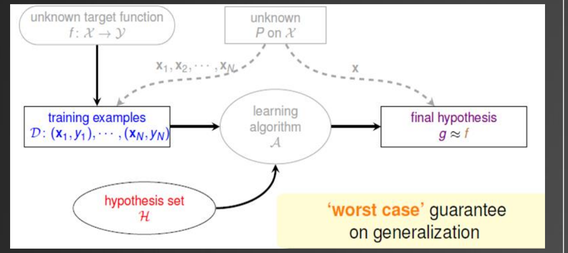
HiBoard

HiTouch

HiVoice

Device

On-device Intelligence



Chipset

CPU

GPU

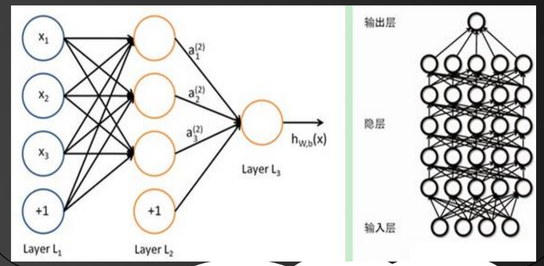
NPU

DSP/ISP

Cloud

Open Platform

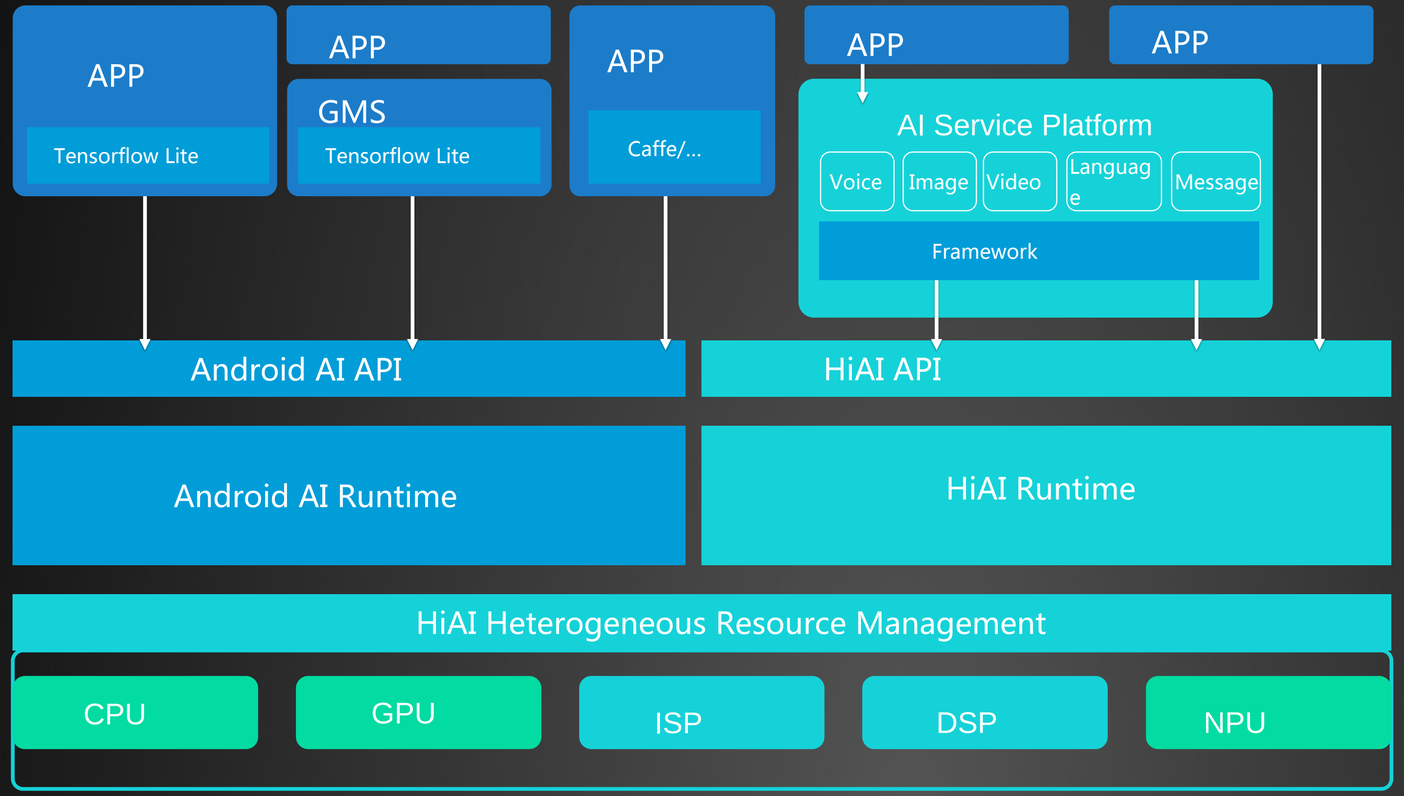
Huawei Cloud AI Capabilities



Big Data

Huawei Mobile AI architecture and applications

- Multi-APP mode
- On-line & Off-line
- Rich API
- Multi-Framework



Tools

AI Development and management

HiAI compilation

Simulator

One More challenge of On-device AI

Training

Models Compression

Inference

How to fit large NN models in a small device?

We offer a free tool of mobile model compression

- Pruning the network
- Quantize the weights
- Huffman code
- ...

Mobile AI Frontier – On-Device Learning

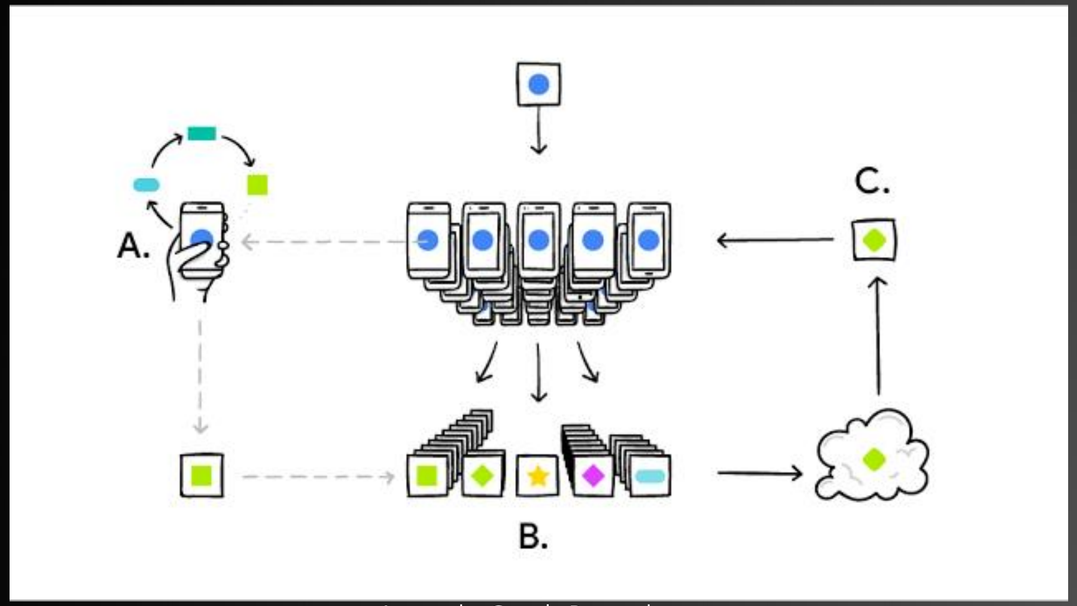


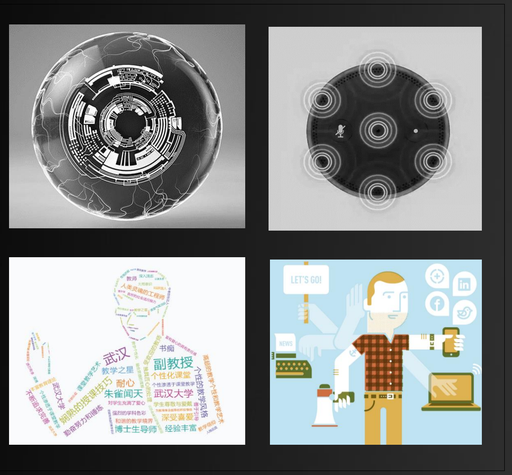
Image by Google Research

Federated Learning

1. Downloads model from cloud
2. Improves it by learning from data on your phone
3. Summarizes the changes as a small focused update
4. Sends the updated model to cloud using encrypted communication
5. The update is immediately averaged with other user updates to improve the shared model

Mobile AI Frontier – Continuous Learning

For on-device continuous learning, devices carry massive information about people, learn personal features and make predictions in real-time: collecting data from pictures, GPS, speech, music, text, apps.



Embedded continuous learning models on device

Data

Inference calculations

Push pre-processed features to the cloud

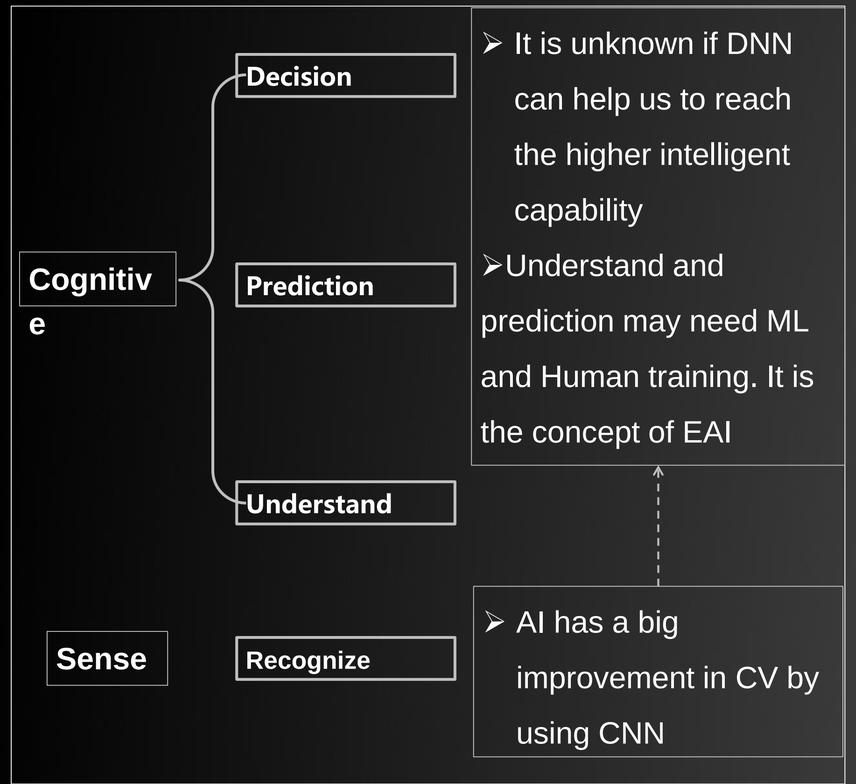
Update weights

Training in the cloud

Three Trends of AI

From Shallow to Deep

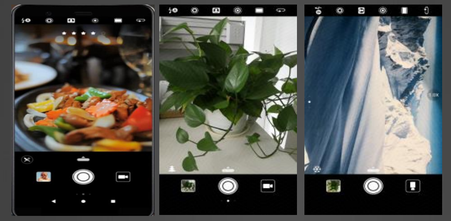
AI from sense to cognitive, understand, to decision



From Specific to General

Current AI models are designed to solve specific problems. DNN makes people believe that general models will be popular

Has to train a OCR for different cases



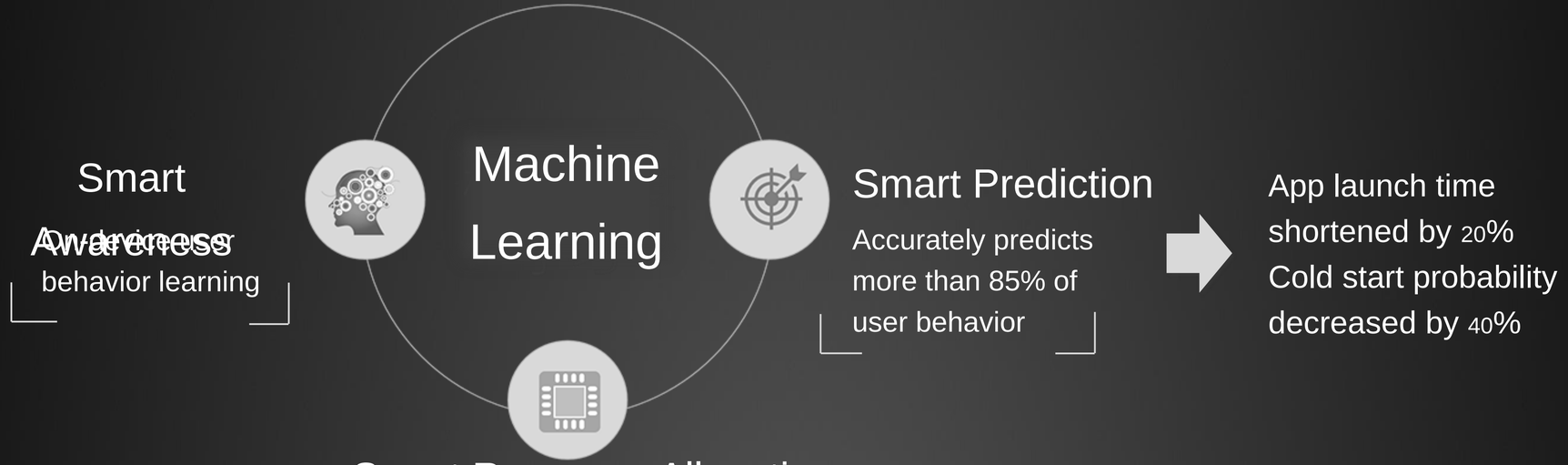
AI aid camera can develop 13 scenarios

Driven by Data to Drive by Knowledge

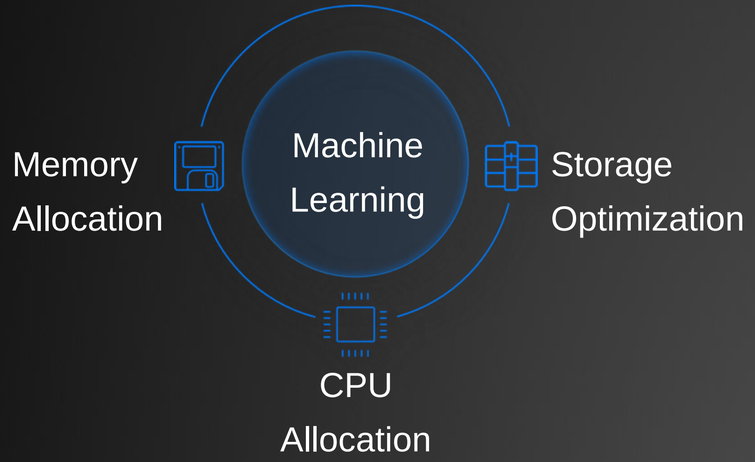
AI development summarized in three stages:

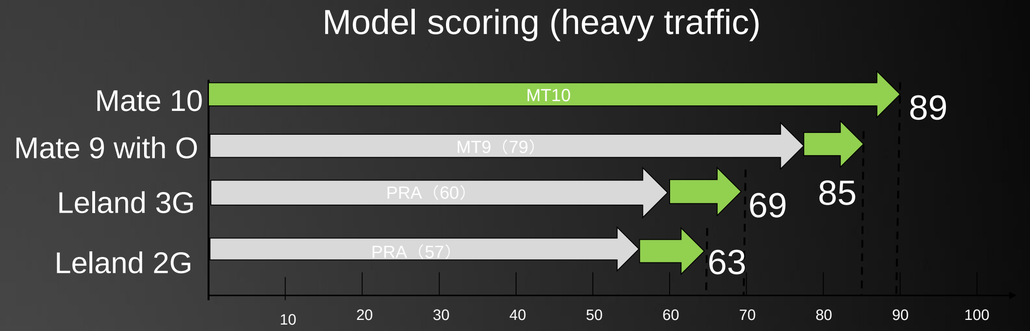
- **Rule Based AI :** Experience, training by people, normally in small, has low accuracy rate.
- **Data Based AI :** Depends on data, self-selection, parameters adjustment by human. Its components are independent , with good accuracy rate, but hard to generalize.
- **Knowledge based AI :** need both data and experience knowledge. Can be self-trained with better accuracy and general availability.

Looking back at EMUI 5.0 & 5.1, already AI-powered Significantly improvement for Android’s lifetime performance



AI-based performance engine
‘Born fast, stays fast’, more than ever





*Based on lab model testing - aging simulation 18 months

- Enhanced F2FS, Background running apps loading improvement, gallery scrolling, contact improvement, low memory management
  - AI engine overall performance improvement +12%
  - Video smoothness + 20%, App response + 15%

AI Experience engine Smart tips *

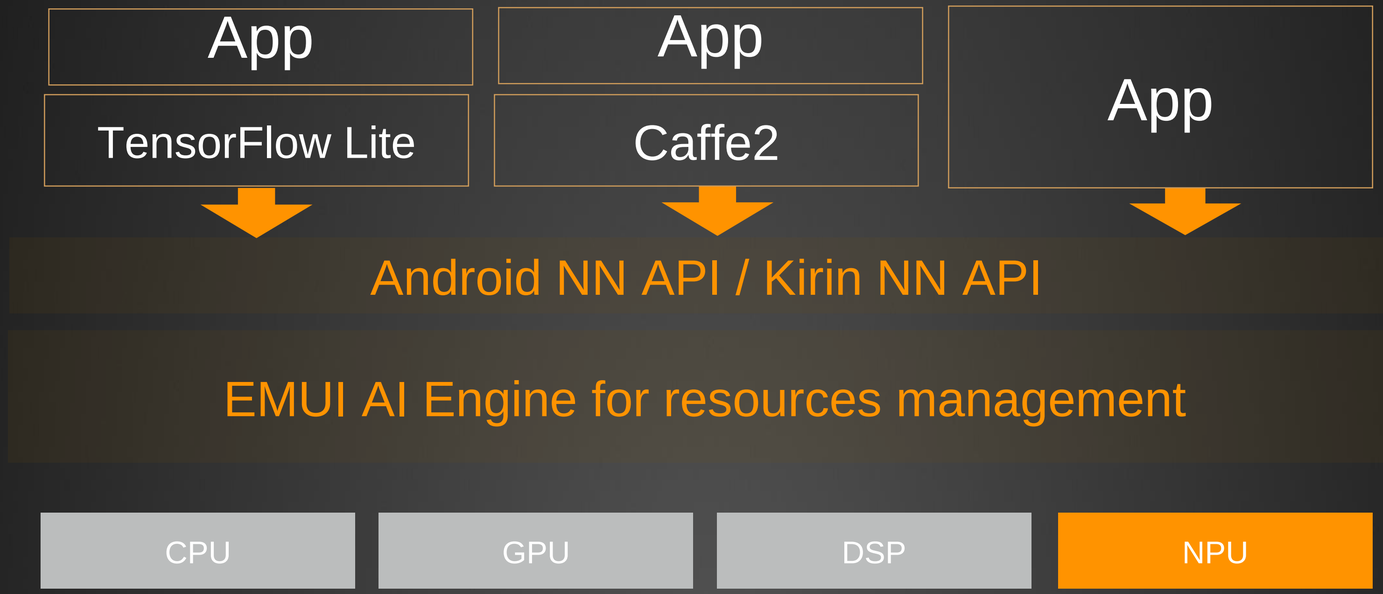
Example:

Intelligently identify user reading environment, gently remind eye-care mode could be enabled

* Eye-Care Mode, knuckles Screenshot, large iris aperture mode, scan business cards and so on has been added to the smart tips

3rd parties AI Apps Engine: Open Ecosystem

3rd parties Apps Empowered by AI Computing Platform



Accelerate 3rd-parties AI frameworks
Ex.: Microsoft translation, perfectly smooth

Real-time translation replacing original text in AR mode while pointing at any text, any format in any language with the camera

Accelerated, no latency thanks to NPU + 3rd party AI framework capabilities

Fast Connectivity

- Global-mode
- LTE Cat18/13 up to 1.2Gbps
- High Speed Railway Optimized

Powerful Computing

- HiAI Architecture
- 4 x A73 + 4 x A53 CPU
- Mali G72MP12 GPU
- Dedicated NPU
- Image DSP

Audiovisual Experience

- Dual ISP
- AI Vision
- 32-bit 384k Audio
- AI Noise Reduction

Long Battery Life

- TSMC 10nm
- i7 Sensor Processor
- LPDDR 4X
- Fine-tuned Power Management