# Huawei's mobile processors

Dezső Sima

Vers. 2.1

November 2018

# Huawei's mobile processors

# 1. Overview

# 1. Overview (1)

## Huawei [1]

- Huawei Technologies Co. Ltd. is a Chinese multinational networking and telecommunications equipment and services company headquartered in Shenzhen China.
- It was founded in 1987 and became the world's largest telecom equipment manufacturer.
- The name Huawei (华为) means Chinese achievement.
- Recently it has about 180 000 employees.

Figure: Huawei's logo [1]

Figure: Huawei's headquarter in Shenzen [1]

# 1. Overview (2)

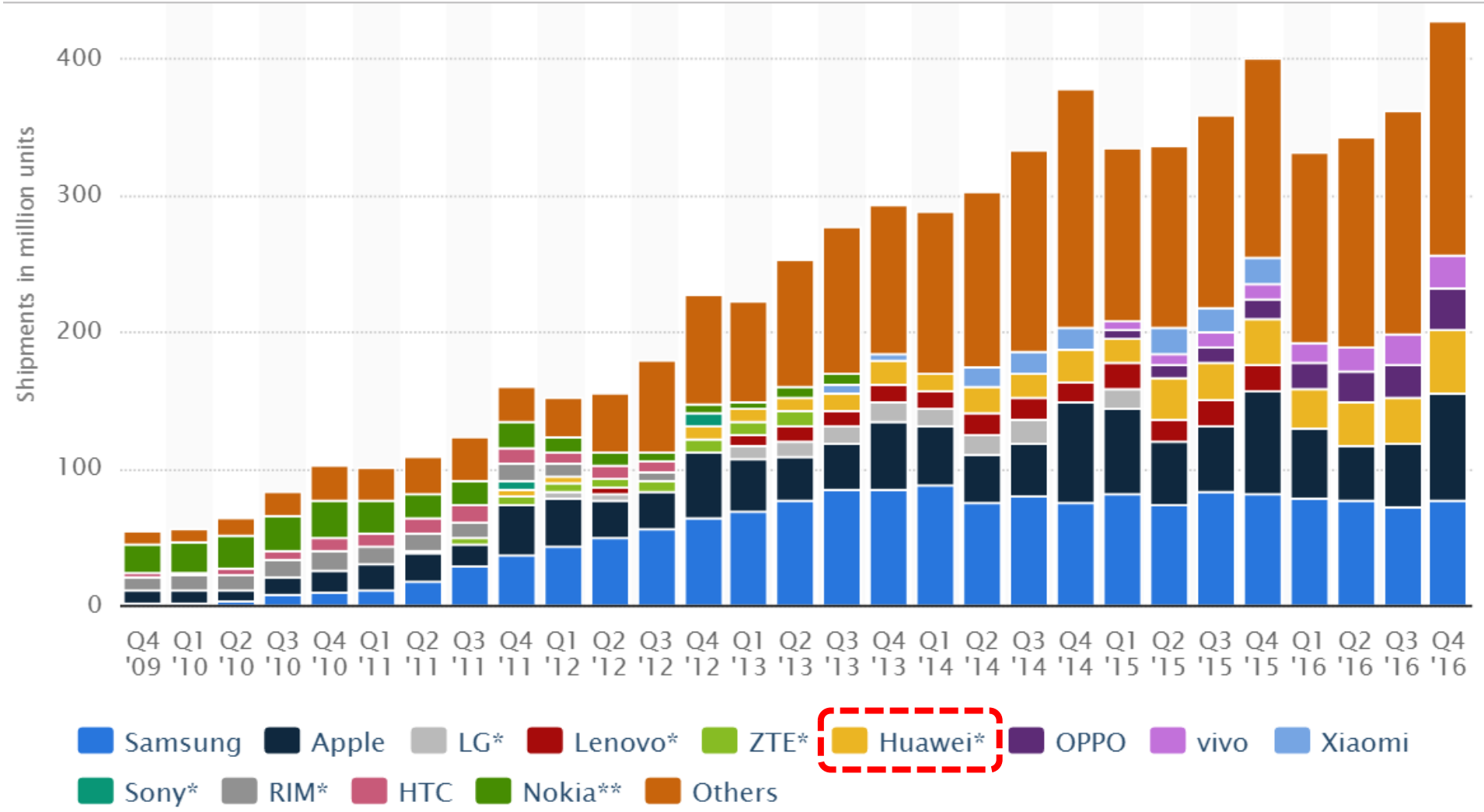
## HiSilicon [2]

- HiSilicon, a global fabless semiconductor and IC design company, headquartered in Shenzhen, China.
- It was founded in 2004 as a subsidiary of Huawei.
- It has offices and research centers worldwide, the number of the employees is about 7000.
- HiSilicon purchases licenses mainly from ARM and designs in the first line application processors for mobiles.

  It filed over 5,000 patents.

# 1. Overview

## Worldwide smartphone shipments by vendors Q4/2009 to Q4/2016 (million units) [3]



# 1. Overview (4)

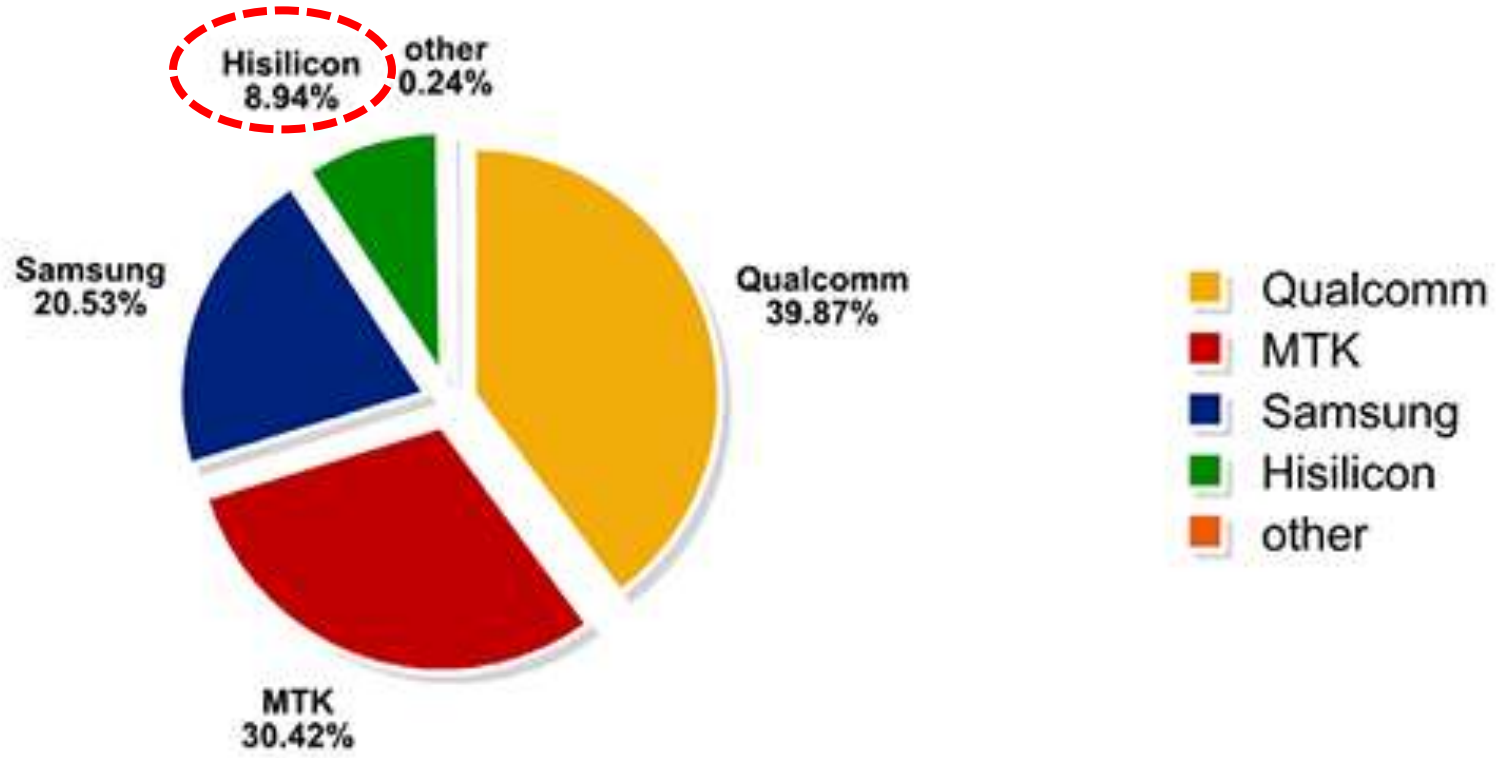
## Worldwide market share of application processors in 2015 used in smartphones (based on revenue) [4]

| Vendor | Market share | Processor lines (examples) | Cores | ISA |
|---|---|---|---|---|
| Qualcomm (USA) | 42 % | Snapdragon 200-800 | Qualcomm designed Krait cores<br>ARM Cortex A line | ARMv7<br>ARMv7/v8 |
| Apple (USA) | 21 % | Apple A7-A9<br>Apple A10 | Apple designed Cyclone core<br>2xbig./4x LITTLE cores | ARMv8 |
| MediaTek (Taiwan) | 19 % | Helio x10<br>Helio X20 | 8xARM Cortex A53 (ARM big.LITTLE)<br>2xARM Cortex A72/8x A53 (ARM big.LITTLE) | ARMv7<br>ARMv8 |
| Samsung (S. Korea) | | Exynos | ARM Cortex A line | ARMv7 |
| Spreadtrum (China) | | SC77xx/88xx | ARM Cortex A5/A7 | ARMv7 |

[Source: Strategy Analytics]

# 1. Overview (5)

## Android smartphone chips worldwide share by vendor in Q3/2016 [5]

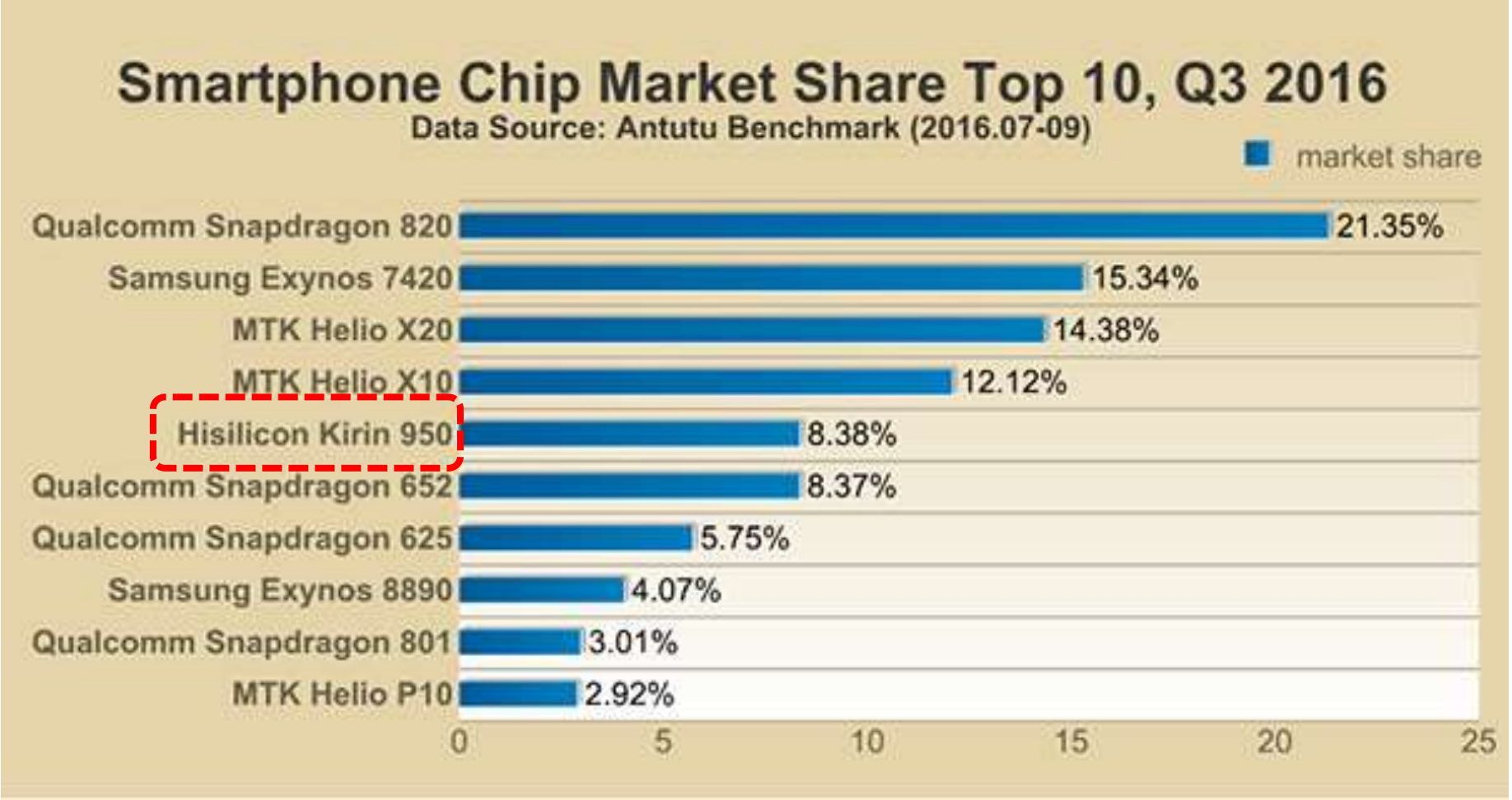



Data source: Antutu Benchmark (2016.07-09)

HiSilicon: Fabless IC design company, subsidiary of Huawei's

Subsidiary: Leányvállalat

Top ten smartphone application processor models sold in Q3/2016 [5]



# 1. Overview (7)

## The meaning of the word "kirin [6]

- Qilin or Kirin, is a mythical creature known in various East Asian cultures.
- It is thought to be a good omen to occasion prosperity or serenity.

Figure: Qing dynasty kirin-shaped incense burner (tömjén füstölő)

# 1. Overview (8)

## HiSilicon's early 32-bit up to quad-core smartphone application processors [7]

| Model Number | Intro. | CPU | Techn | GPU | Memory techn. up to | Conectivity | Utilizing devices |
|---|---|---|---|---|---|---|---|
| K3V1 (Hi3611) | 2009 | 1x ARM9E 0.8 GHz | 130 nm | | | | |
| K3V2 (Hi3620) | 2012 | 4x A9 1.4 GHz | 40 nm | Vivante GC4000 @ 240 MHz | 2x LPDDR3 @500MT/s | GSM, WCDMA | • Huawei Honor 2<br>• Huawei Ascend P2<br>• Lenovo A376 etc. |
| K3V2E | 2013 | 4x A9 1.5 GHz | 40 nm | Vivante GC4000 @ 240 MHz | 2x LPDDR3 @500 MT/s | GSM, WCDMA | • Huawei Honor 3 |

# 1. Overview (9)

## HiSilicon's 32-bit quad- and octa core smartphone application processors [7]

| Model Number | Intro. | CPU | Techn | GPU | Memory techn. up to | Conectivity | Utilizing devices |
|---|---|---|---|---|---|---|---|
| Kirin 910 | 1H/2014 | 4x A9<br>1.6 GHz | 28 nm | Mali-450 MP4<br>@ 533 MHz | LPDDR3 | LTE cat.4 | • Huawei P6 S<br>• Huawei MediaPad X1<br>etc. |
| Kirin 910T | 1H/2014 | 4x A9<br>1.8 GHz | 28 nm | Mali-450 MP4<br>@ 700 MHz | LPDDR3 | LTE cat.4 | • Huawei Ascend P7 |
| Kirin 920 | 2H/2014 | 4x A15 1.7 GHz<br>4x A7 1.3 GHz | 28 nm | Mali-T628 MP4<br>@ 600 MHz | LPDDR3<br>1600 MT/s | LTE cat.6 | • Huawei Honor 6 |
| Kirin 925 | Q3/2014 | 4x A15 1.8 GHz<br>4x A7 1.3 GHz | 28 nm | Mali-T628 MP4<br>@ 600 MHz | LPDDR3<br>1600 MT/s | LTE cat.6 | • Huawei Ascend Mate 7<br>• Huawei Honor 6 Plus |
| Kirin 928 | n.a. | 4x A15 2.0 GHz<br>4x A7 1.3 GHz | 28 nm | Mali-T628 MP4<br>@ n.a MHz | LPDDR3<br>1600 MT/s | LTE cat.6 | • Huawei Honor 6<br>Extreme Edition |

# 1. Overview (10)

## HiSilicon's 64-bit octa core midrange smartphone application processors [7]

| Model Number | Intro. | CPU | Techn | GPU | Memory techn. up to | Conectivity | Utilizing devices |
|---|---|---|---|---|---|---|---|
| Kirin 620 | 2015 | 8x A53<br>1.2 GHz | 28 nm | Mali-450 MP4<br>@ 533 MHz | 1x LPDDR3<br>1600 MT/s | LTE cat.4 | • Huawei P8 Lite<br>• Huqwei Honor 4X<br>• Huawei Honor 4C etc. |
| Kirin 650 | Q2/2016 | 4x A53 2.0 GHz<br>4x A53 1.7 GHz | 16 nm | Mali-T830 MP2<br>@ 900 MHz | 2x LPDDR3<br>933 MT/s | LTE cat.6 | • Huawei P9 Lite<br>• Huawei Honor 5C |
| Kirin 655 | Q4/2016 | 4x A53 2.12 GHz<br>4x A53 1.7 GHz | 16 nm | Mali-T830 MP2<br>@ 900 MHz | 2x LPDDR3<br>933 MT/s | LTE cat.6 | • Huawei P8 Lite<br>• Huawei Mate 9 Lite<br>• Huawei Honor 6X etc. |
| Kirin 658 | Q1/2017 | 1x A53 2.36 GHz<br>3x A53 2.12 GHz<br>4x A53 1.7 GHz | 16 nm | Mali-T830 MP2<br>@ 900 MHz | 2x LPDDR3<br>933 MT/s | LTE cat.6 | • Huawei Ascend Mate 7<br>• Huawei Honor 6 Plus |

## HiSilicon's 64-bit octa core smartphone application processors [7]

| Model Number | Intro. | CPU | Techn | GPU | Memory techn. up to | Conectivity | Utilzsing devices |
|---|---|---|---|---|---|---|---|
| Kirin 930 | Q1/2015 | 4x A53 2.0 GHz<br>4x A53 1.5 GHz | 28 nm | Mali-T628 MP4<br>@ 600 MHz | 2x LPDDR3<br>1600 MT/s | LTE cat.6 | • Huawei P8<br>• Huawei MediaPad X2<br>• Huaweu MediaPad M2 |
| Kirin 935 | Q1/2015 | 4x A53 2.2 GHz<br>4x A53 1.5 GHz | 28 nm | Mali-T628 MP4<br>@ 680 MHz | 2x LPDDR3<br>1600 MT/s | LTE cat.6 | • Huawei P8 Max<br>• Huawei Honor 7<br>• Huaweu Mate S |
| Kirin 940 | Q3/2015 | 4x A72 2.2 GHz<br>4x A53 1.x GHz | 16 nm | Mali-T880 MP4<br>@ MHz | 2x LPDDR4<br>1600 MT/s | LTE cat.7 | • Huawei Honor 7<br>• Huawei Ascend Mate 8 |
| Kirin 950 | Q4/2015 | 4x A72 2.3 GHz<br>4x A53 1.8 GHz | 16 nm | Mali-T880 MP4<br>@ 900 MHz | 2x LPDDR4<br>1600 MT/s | LTE cat.10 | • Huawei Honor 8<br>• Huawei Honor V8<br>• Huawei Mate 8 etc. |
| Kirin 955 | Q2/2016 | 4x A72 2.5 GHz<br>4x A53 1.8 GHz | 16 nm | Mali-T880 MP4<br>@ 900 MHz | 2x LPDDR4<br>1600 MT/s | LTE cat.10 | • Huawei P9<br>• Huawei P9 Plus<br>• Huawei Honor V8 etc. |
| Kirin 960 | Q4/2016 | 4x A73<br>2.362 GHz<br>4x A53 1.844<br>GHz | 16 nm | Mali-G71 MP8<br>@ 1037 MHz | 4x LPDDR4<br>(16-bit)<br>1800 MT/s | LTE cat.12 | • Huawei P 10<br>• Huawei P10 Plus<br>• Huawei Mate 9<br>• Huawei Honor V9 etc. |
| Kirin 970 | 09/2017 | 4x A73 2.36 GHz<br>4x A53 1.84 GHz | 10 nm | Mali-G72 MP12<br>@ 746 MHz | 4x LPDDR4<br>(16-bit)<br>1833 MT/s | LTE cat.18 | • Huawei Mate 10<br>• Huawei Mate 10 Pro<br>• Huawei Honor V10 |
| Kirin 980 | 10/2018 | 2x A76 2.60 GHz<br>2x A76 1.92 GHz<br>4x A55 1.80 GHz | 7 nm | Mali-G76 MP10<br>@ 720 Mhz | 4x LPDDR4<br>(16-bit)<br>2133 MT/s | LTE cat.21 | • Huawei Mate 20<br>• Huawei Mate 20 Pro |

# 2.1 The Kirin 950

## 2.1 The Kirin 950

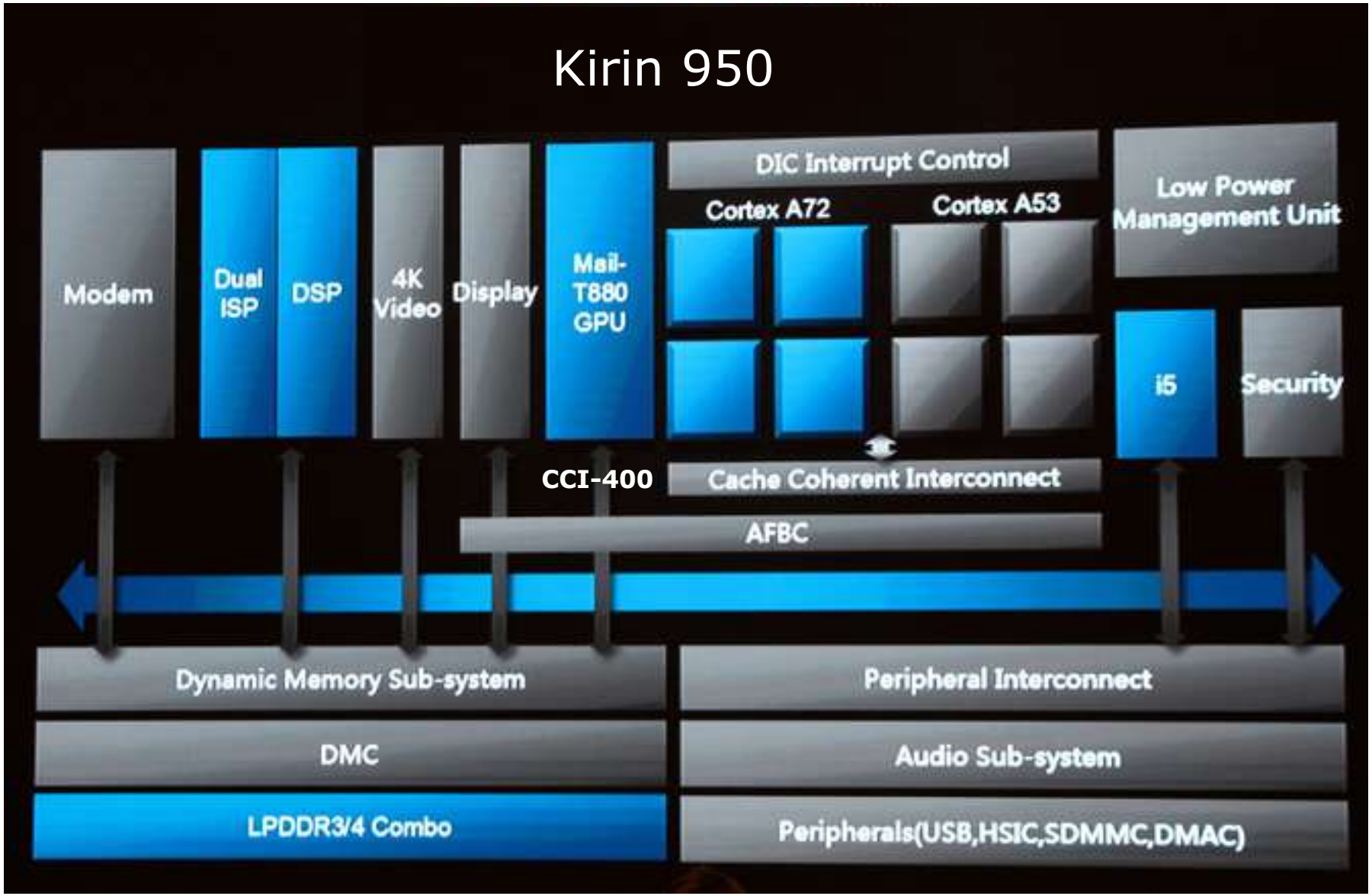
- Introduced in 11/2015.
- It is the application processor of the Huawei P8.
- The Kirin 950 has only a few enhancements over the previous Kirin 940 model, as can be identified in the Table contrasting the Kirin 930/940/950 models, as follows:
  - about 10 % higher A72 speeds (2.4 GHz vs. 2.2 GHz).
  - enhanced GPU (Mali T880 vs. the T860) and
  - enhance connectivity (Cat. 10 LTE vs. Cat 7 LTE).

# 2.1 The Kirin 950 (2)

## Comparing the Kirin 930/940 and 950 models [8]

| | KIRIN 930 | KIRIN 940 | KIRIN 950 |
|---|---|---|---|
| CPU | Quad A53 + Quad A57 (up to 2.0GHz) | Quad A53 + Quad A72 (up to 2.2GHz) | Quad A53 + Quad A72 (up to 2.4GHz) |
| RAM | Dual-channel LPDDR3 | Dual-channel LPDDR4 (25.6GB/s) | Dual-channel LPDDR4 (25.6GB/s) |
| GPU | ARM Mali T628 GPU | ARM Mali T860 GPU | ARM Mali T880 GPU |
| DSP | Tensilica HiFi 3 DSP | Tensilica HiFi 4 DSP | Tensilica HiFi 4 DSP |
| ISP | 32MP ISP | Dual ISP (32MP) | Dual ISP (42MP) |
| Video Encode | 1080p | 4K | 4K |
| Modem | Dual SIM Cat. 6 LTE | Dual SIM Cat. 7 LTE | (Dual-SIM) LTE Cat.10 |
| Sensor Hub | i3 Co-Processor (Sensor Hub) | i7 Co-Processor (Sensor Hub + Connectivity + Security) | i7 Co-Processor (Sensor Hub + Connectivity + Security) |
| External Component Interfaces | eMMC 4.51 / SD 3.0 (UHS-I)<br>BT 4.0 Low Energy<br>Dual-band a/b/g/n Wi-Fi<br>USB 2.0 | UFS 2.0 / eMMC 5.1 / SD 4.1 (UHS-II)<br>MU-MIMO ac Wi-Fi<br>BT 4.2 Smart<br>USB 3.0<br>NFC | UFS 2.0 / eMMC 5.1 / SD 4.1 (UHS-II)<br>MU-MIMO ac Wi-Fi<br>BT 4.2 Smart<br>USB 3.0<br>NFC |
| Release | Q2 2015 | Q3 2015 | Q4 2015 |

## Block diagram of the Kirin 950 [9]



# 2.1 The Kirin 950 (4)

Remark

Blue parts in the block diagram designate upgraded units of the Kirin 950 vs. the Kirin 930 model, whereas gray parts mark units existing before.

# 2.1 The Kirin 950 (5)

## Specific features of the Kirin 950 to be discussed

a) AFBC (ARM Frame Buffer Compression
b) The i5 co-processor
c) c) Volta
d) UFS 2.0 (introduced already in the Kirin 940).

## a) AFBC (ARM Frame Buffer Compression)

- AFBC (ARM Frame Buffer Compression) is a lossless real-time video compression and decompression method that minimizes the amount of data to be transferred between different units implemented on the chip.
- This reduces the effective bandwidth demand and provides a corresponding power saving of up to 50 %.
- ARM announced AFBC in 06/2013 and supports it in their subsequent video processors, GPUs and display processors beginning with the models as seen below.

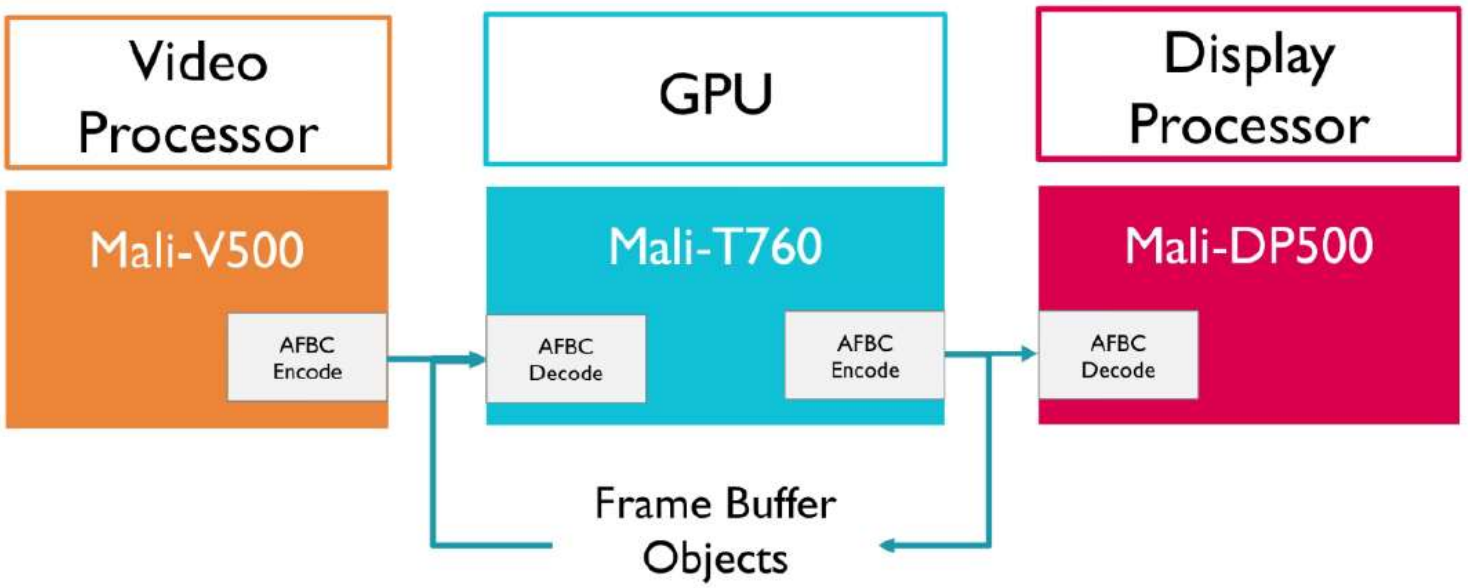
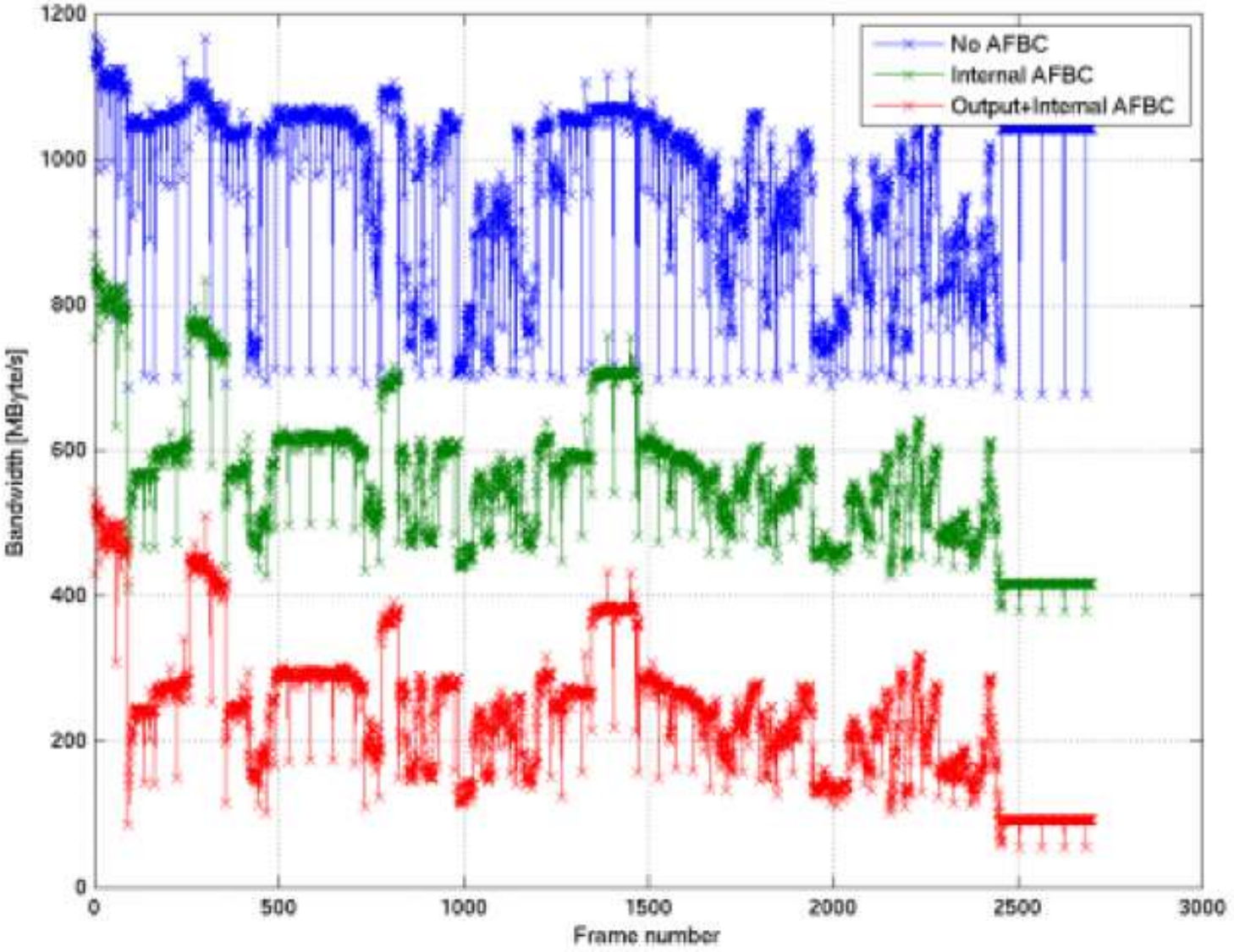



Figure: Use of AFBC compression in processing a video stream [10]

# 2.1 The Kirin 950 (7)

Example: Compression results achieved with AFBC when decoding a 4K H.264 video stream -1 [10]



# 2.1 The Kirin 950 (8)

## Example: Compression results achieved with AFBC when decoding a 4K H.264 video stream -2 [10]

- The blue curve shows the bandwidth without using AFBC.
- The green curve indicates the bandwidth of Mali-V500 when AFBC is used in the internal reference frame compression only.
- Finally, the red curve shows the bandwidth when AFBC is used for the output frame as well with an AFBC capable display processor.
- As seen, the bandwidth reductions are considerable, that is the more than 50 % bandwidth reduction could be achieved for the entire video stream.

# 2.1 The Kirin 950 (9)

## b) The i5 co-processor [9]

- The Kirin 950 integrates a co-processor, designated as the i5 (which has no relation to Intel's i5 processor models).
- It is always on, it controls various sensors of the processor even when the phone is in sleep mode.
- The co-processor carries out efficiently tasks that don't require waking up the other cores, lengthening in this way battery life, like

  navigation
  always-on voice control with speech recognition
  music playback
  etc.

- The i5 is implemented by an ARM M7.
- Previous Kirin models (930/935/940) included also an embedded co-processor, actually the ARM M3, designated as the i3 for supporting the sensor hub.
- The standby power consumption of the i5 is only 6.5 mA vs. 90 mA in case of the previous i3.

## c) VoLTE (Voice over LTE) [11], [12]

- VoLTE is a standard for high-speed wireless communication for mobile phones.
- Traditional telephone calls are limited to a frequency range of 300 Hz to 3.4 KHz.

  VoLTE improves voice quality by widening the frequency range from 50 Hz to 7 KHz.
- This results in better low- and high-frequency limits and a latency reduction from 5 to 6 sec. to 0.5 to 1.5 sec.

  Obviously, a wider frequency range improves also music quality.
- VoLTE was introduced about 2014 and was implemented also in the Kirin 920 to Kirin 940 models.

## UFS2.0

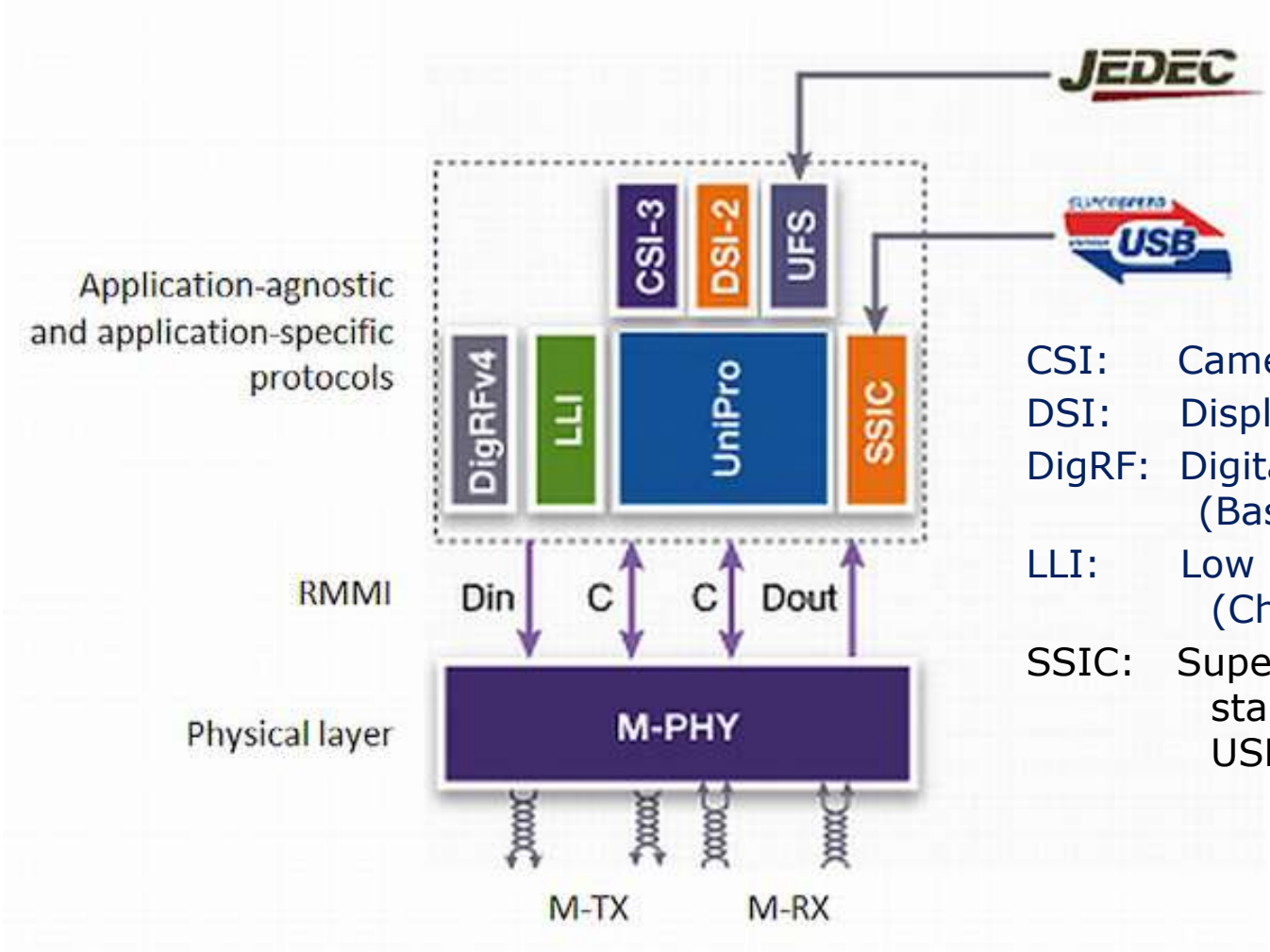
Main enhancements of the UFS interface vs. the eMMC flash interface:

a) the UFS protocol is a layered one and

b) it has an LVDS based serial physical interface whereas eMMC an up to 8-bit parallel interface with single ended (ground based) signaling

as indicated next.

## The layout of the layered UFS protocol [13]




CSI: Camera Serial Interface

DSI: Display Serial Interface

DigRF: Digital RF - BB Interface (Baseband)

LLI: Low Latency Interrface (Chip to Chip)

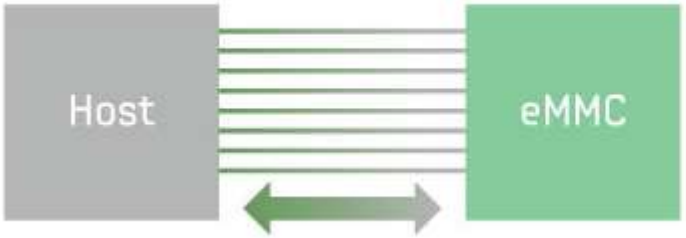
SSIC: SuperSpeed Inter-Chip standard for on-PCB USB communication

## b) Serial, LVDS based physical layer (PHY) between the host and the flash device [14]

- UFS has an LVDS based serial physical interface whereas eMMC an up to 8-bit parallel interface with single ended (ground based) signaling, as indicated below.

eMMC

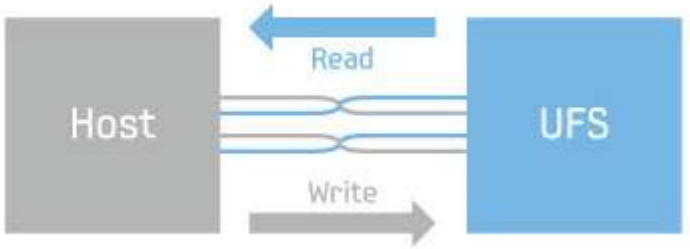
*Parallel x8 interface*




*Half duplex (reads or writes) transactions*

UFS

*Serial LVDS interface*




*Full duplex (reads and writes) transactions*

Figure: Contrasting the eMMC and UFS physical interfaces

- UFS allows full duplex (two-way parallel) reads and writes, whereas eMMC can only perform half-duplex transactions, i.e. either a read or a write operation at a time but no both at the same time.

# 2.2 The Kirin 960

## 2.2 The Kirin 960

- Announced in 10/2016, appearance in devices in Q1/2017.
- It is kernel part of Huawei's P10.
- It is the last 16 nm application processor of Huawei.
- Main features of the Kirin 960 vs. the Kirin 950 are given in the next Table.

# 2.2 The Kirin 960 (2)

## Contrasting the main features of the Kirin 950 and Kirin 960 [15]

| SoC | Kirin 960 | Kirin 955 | Kirin 950 |
|---|---|---|---|
| CPU | 4x Cortex-A73 @ 2.36GHz<br>4x Cortex-A53 @ 1.84GHz | 4x Cortex-A72 @ 2.52GHz<br>4x Cortex-A53 @ 1.81GHz | 4x Cortex-A72 @ 2.30GHz<br>4x Cortex-A53 @ 1.81GHz |
| GPU | ARM Mali-G71MP8 1037MHz | ARM Mali-T880MP4 900MHz | |
| Memory | 2x 32-bit LPDDR4 @ 1866MHz 29.9GB/s | 2x 32-bit LPDDR3 @ 933MHz (14.9GB/s) or 2x 32-bit LPDDR4 @ 1333MHz (21.3GB/s) (hybrid controller) | |
| Interconnect | ARM CCI-550 | ARM CCI-400 | |
| Storage | UFS 2.1 | eMMC 5.0 | |
| ISP/Camera | Dual 14-bit ISP (Improved) | Dual 14-bit ISP 940MP/s | |
| Encode/Decode | 2160p30 HEVC & H.264 Decode & Encode 2160p60 HEVC Decode | 1080p H.264 Decode & Encode 2160p30 HEVC Decode | |
| Integrated Modem | Kirin 960 Integrated LTE (Category 12/13) DL = 600Mbps UL = 150Mbps | Balong Integrated LTE (Category 6) DL = 300Mbps UL = 50Mbps | |
| Sensor Hub | i6 | i5 | |
| Mfc. Process | TSMC 16nm FFC | TSMC 16nm FF+ | |

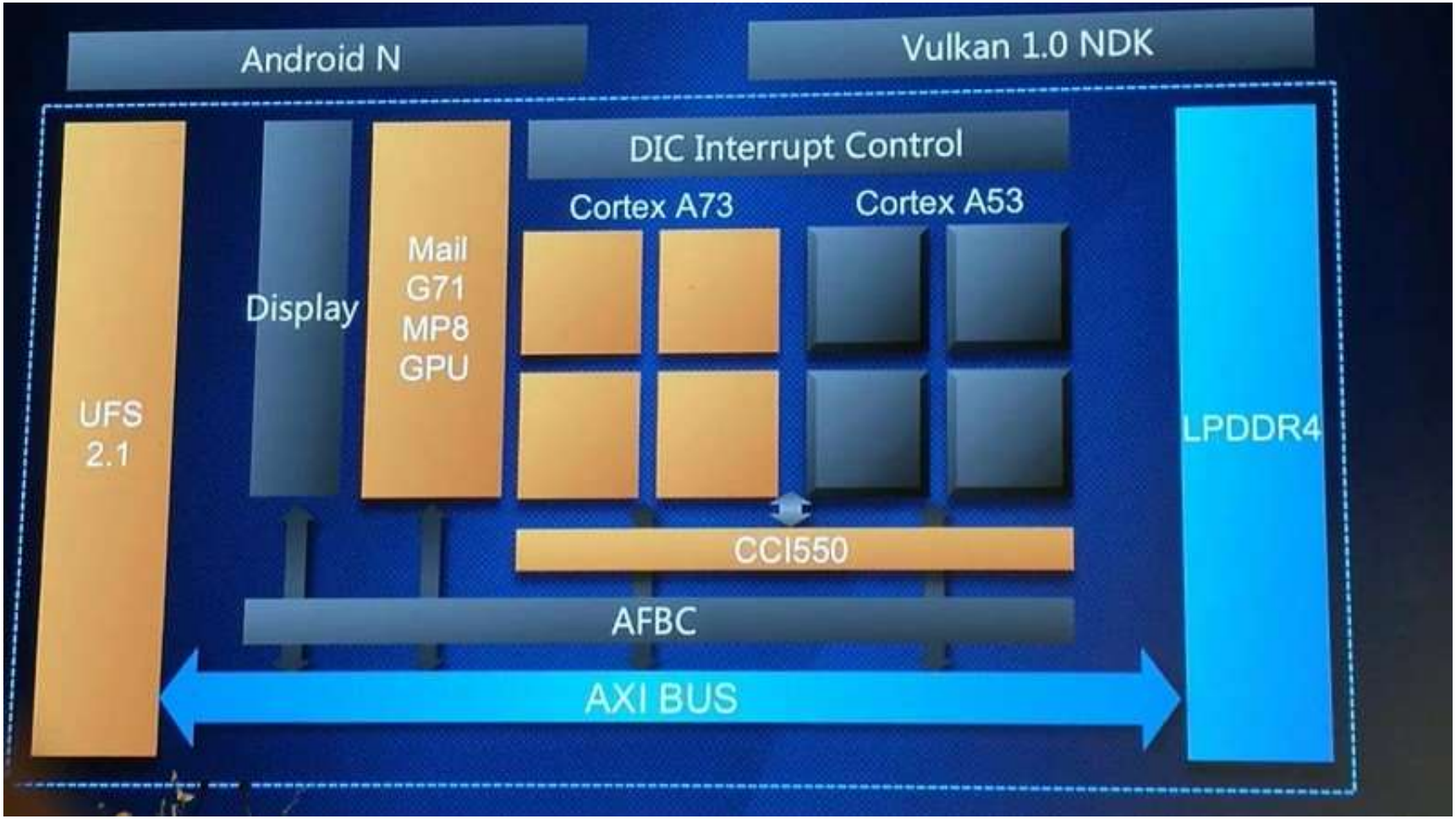
Block diagram of the Kirin 960 [15]

# 2.2 The Kirin 960 (4)

Note

Enhanced units of the Kirin 960 are designated red.

## Simplified block diagram of the Kirin 960 [16]




Enhanced units are designated red.

# 2.2 The Kirin 960 (6)

## Main enhancements of the Kirin 960 vs. the Kirin 950

- Cortex A73 cores instead of A72 cores
- next (3.) generation (Bifrost) Mali G71 GPU (see a)
- CCI-500 cache coherent interconnect instead of CCI-400 (see b)
- a new cat.12/cat.13 modem
- an enhanced co-processor designated as i6
- UFS 2.1 (Universal flash Storage) flash memory instead of UFS 2.0/eMMC 5.1 (embedded MMC) (see c)

Select enhancements

## a) Next (3.) generation (Bifrost) Mali G71 GPU

The Mali G71 GPU belongs to ARM's 3. generation (Bifrost) Mali GPUs, as shown below.

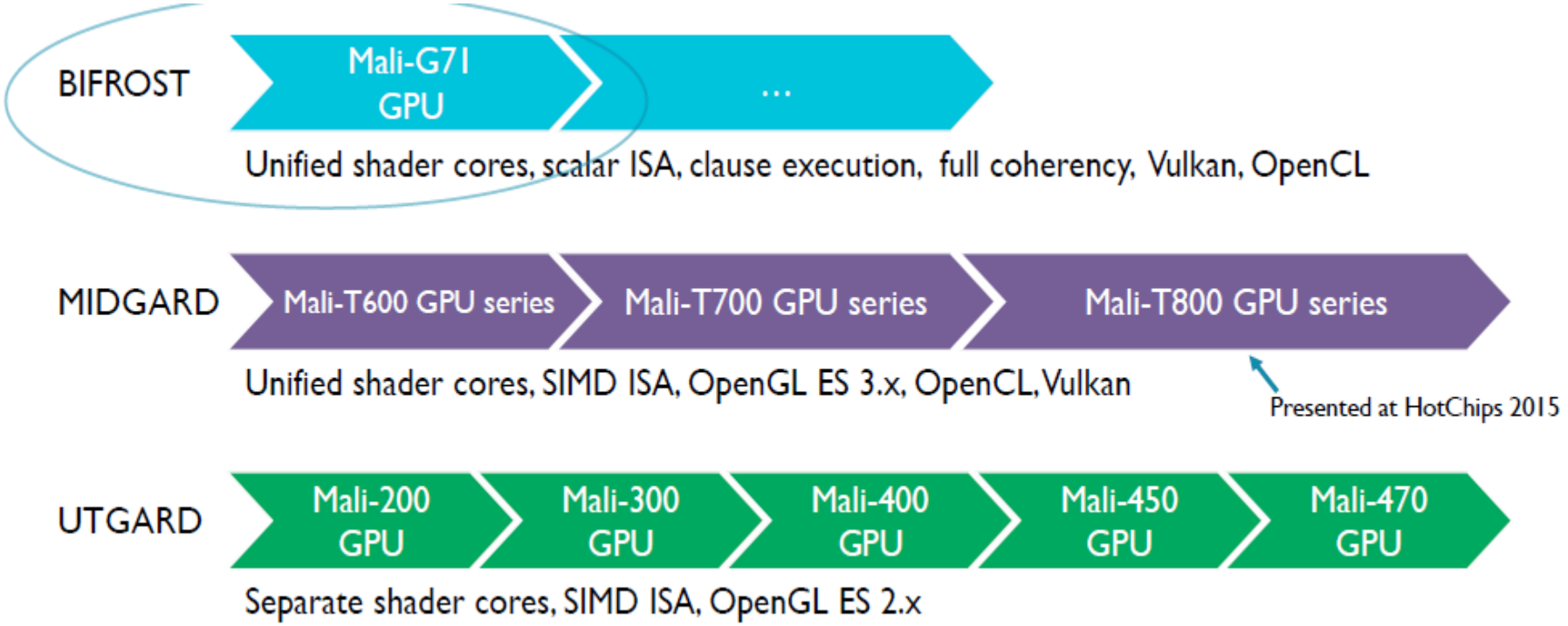



Figure: Three generations of Mali GPUs [17]

These Mali GPU generations are detailed in the Chapter "Samsung's mobile SOCs".

## b) CCI-500 cache coherent interconnect instead of CCI-400

### Use of ARM's CCI-400 interconnect IPs by major SOC providers

**Use of ARM's interconnect IPs targeting mobiles**

**Use of ARM's CCI-400 IP in mobiles of major manufacturers**

- Samsung Exynos 5 Octa 5410 (2013)
- Samsung Exynos 5 Octa 5420 (2013)
- Samsung Exynos 7 Octa 7420 (2015)
- Rockchip RK3288 (2014)
- MediaTek MT6595 (2014)

- Huawei Kirin 950 (2015)
- (Huawei Kirin 960 (2016) uses the CCI-500)

**Use of own proprietary interconnect in the mobiles of major manufacturers**

- SamsungCoherent Interconnect (SCI) in Exynos 8 Octa 8890 (2015)
- MediaTek Coherent System Interconnect (MCSI) in MediaTek MT6797 (2015)

# ARM's cache-coherent interconnects belonging to the CoreLink 500 family

**ARM's cache-coherent interconnects based on the AMBA 4 ACE bus**

**ARM's cache-coherent interconnect belonging to the CoreLink 400 family**

It does not include a snoop filter.

*Typical use in*

*Cortex-A7/A15/A53/A53*
SoCs

*Models*

- *CCI-400 (2010)*

| | |
|---|---|
| Fully coherent CPU clusters up to | 2 |
| No. of LPDDR 4/3 memory channels | 2 |

Suitable already for big.LITTLE configurations

**ARM's cache-coherent interconnects belonging to the CoreLink 500 family**

They include a snoop filter to reduce snoop traffic.

*Cortex-A53/A57/A72*
SoCs

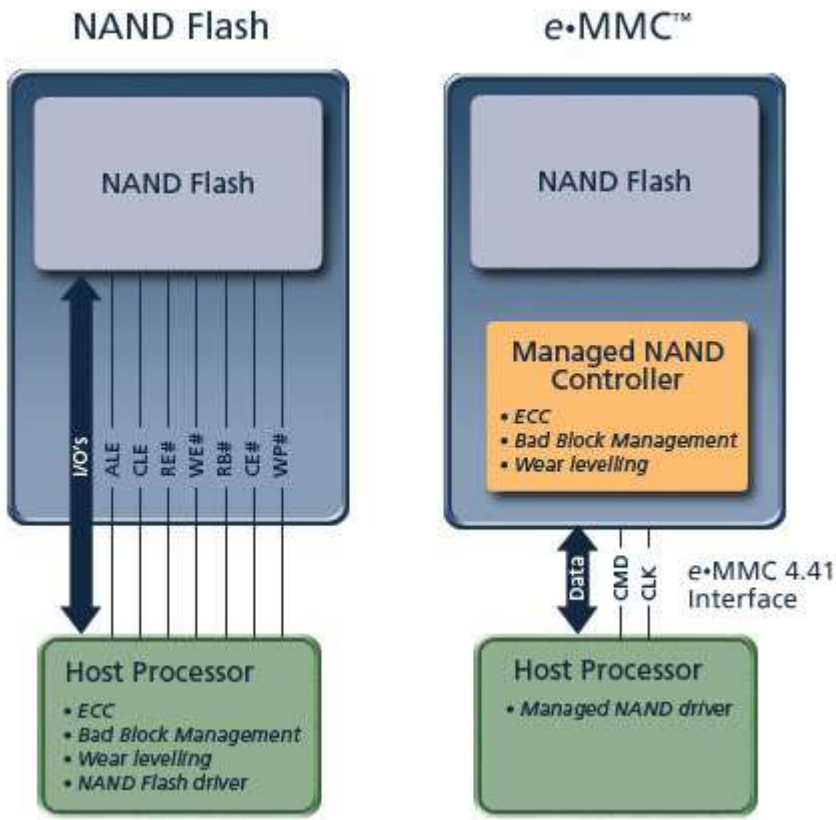
| | CCI-500 (2014) | CCI-550 (2015) |
|---|---|---|
| Fully coherent CPU clusters up to | 4 | 6 |
| No. of LPDDR 4/3 memory channels | 4 | 6 |

## c) UFS 2.1 flash memory instead of UFS 2.0/eMMC 5.1

### The eMMC flash memory

Portable devices, like digital cameras, smart phones or tablets typically store their data on flash memory.

eMMC (embedded Multi-Media Card) refers to a flash memory package that incorporates both the flash memory and the flash memory controller integrated on the same silicon die.



# 2.3 The Kirin 970

## 2.3 The Kirin 970

- Announced in 9/2017, appeared in devices in 10/2017.
- It is the kernel part of Huawei's Mate 10, Mate 10 Pro and Honor V10 (cheaper version).
- It is Huawei's first 10 nm application processor, implemented on a 10x10 $mm^2$ die area and includes about 5.5 billion transistors, about 1.5 billion more than the Kirin 960.

  (By contrast Apple's A10 has 3.1 billion and Qualcomm's Snapdragon 835 3.5 billion transistors).

# 2.3 The Kirin 970 (2)

## Main features of the Kirin 950 – 970 processors [18]

| SoC | Kirin 970 | Kirin 960 | Kirin 950 |
|---|---|---|---|
| CPU | 4x Cortex-A73 @ 2.4GHz<br>4x Cortex A53 @ 1.8GHz | 4x Cortex-A73 @ 2.4GHz<br>4x Cortex A53 @ 1.8GHz | 4x Cortex-A72 @ 2.3GHZ<br>4x Cortex-A53 @ 1.8GHz |
| GPU | Mali-G72 MP12 | Mali-G71 MP8 @900MHz | Mali-T880 MP4 @900MHz |
| Neural Processing Unit (NPU) | Yes | No | No |
| Camera ISP | Dual ISP | ISP | ISP |
| Media processing | 2160p60 HEVC & H.264 Decode<br>2160p30 Encode HDR10 | 2160p30 HEVC & H.264 Decode & Encode<br>2160p60 HEVC Decode | 1080p H.264 Decode & Encode<br>2160p30 HEVC Decode |
| RAM | 4x LPDDR4 | 2x LPDDR4 | 2x LPDDR4 |
| LTE modem | LTE Cat 18 | LTE Cat 12 | LTE Cat 6 |
| Flash interface | UFS 2.1 | UFS 2.1 | eMMC 5.1 |
| Process | TSMC 10nm | 16nm FinFET | 16nm FinFET |

# 2.3 The Kirin 970 (3)

## Main features of the Kirin 970 vs. the previous Kirin 960

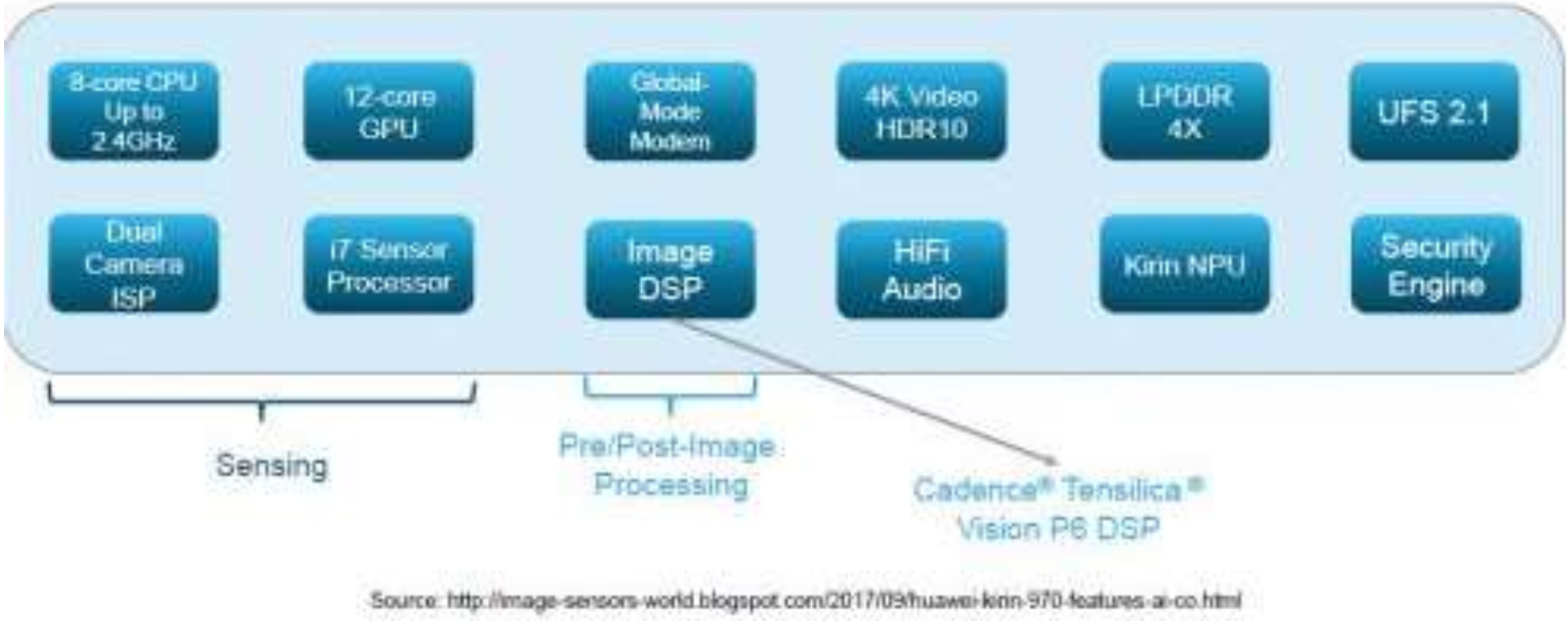
- ARM's Mali-G72 GPU with a 12 core design (MP12)
- Neural Processing Unit (NPU)
- Dual ISP
- 4 channel LPDDR4 memory
- 4.5G modem with Cat. 18 download speed for up to 1.2 Gbps

# 2.3 The Kirin 970 (4)

## HiAI mobile computing architecture

It provides the most efficient use of heterogeneous computing architecture based on the use CPU, GPU, ISP, DSP and NPU.

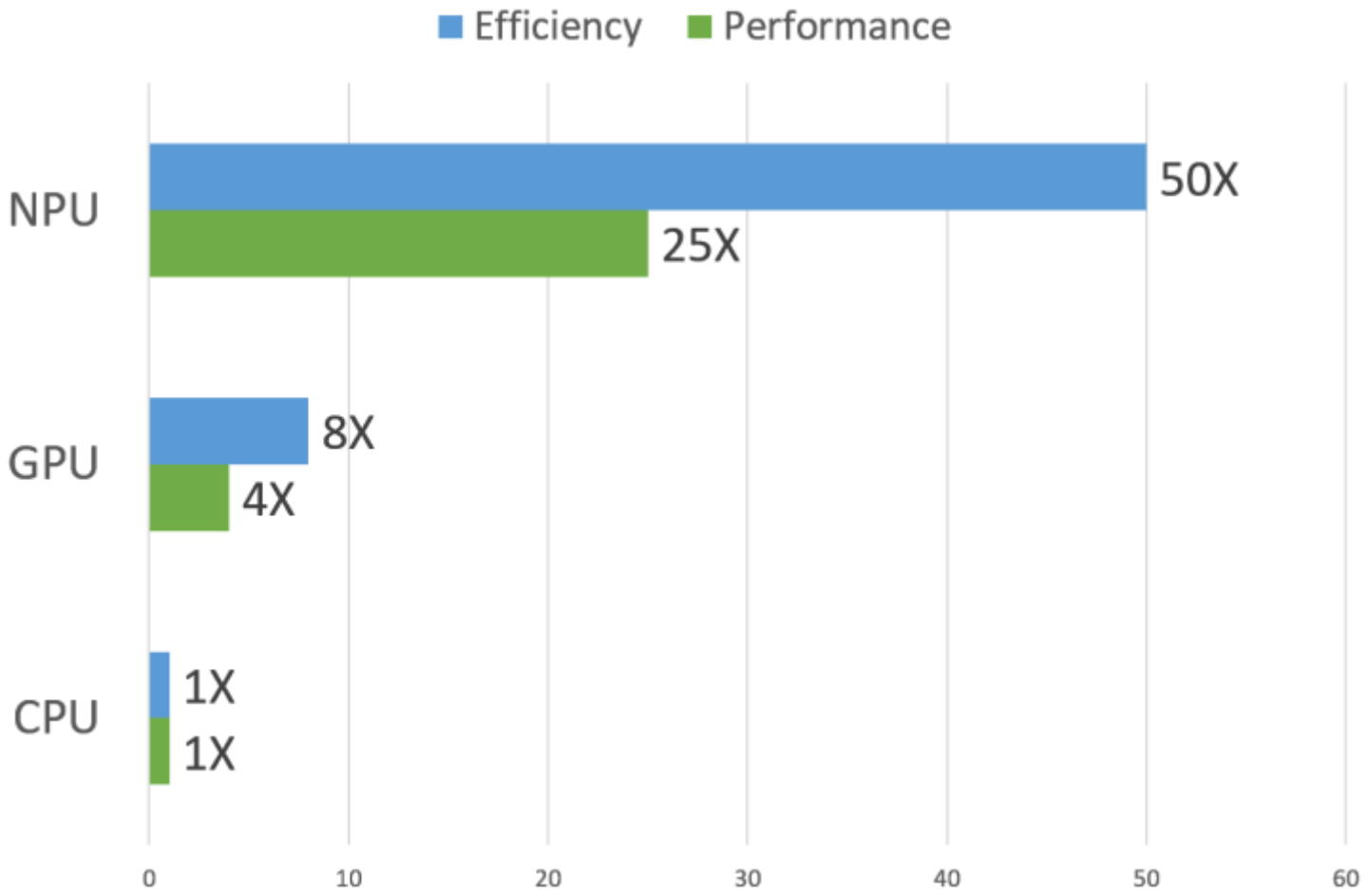
## Main functional blocks of the Kirin 970 [19]




Source: http://image-sensors-world.blogspot.com/2017/09/huawei-kirin-970-features-ai-co.html

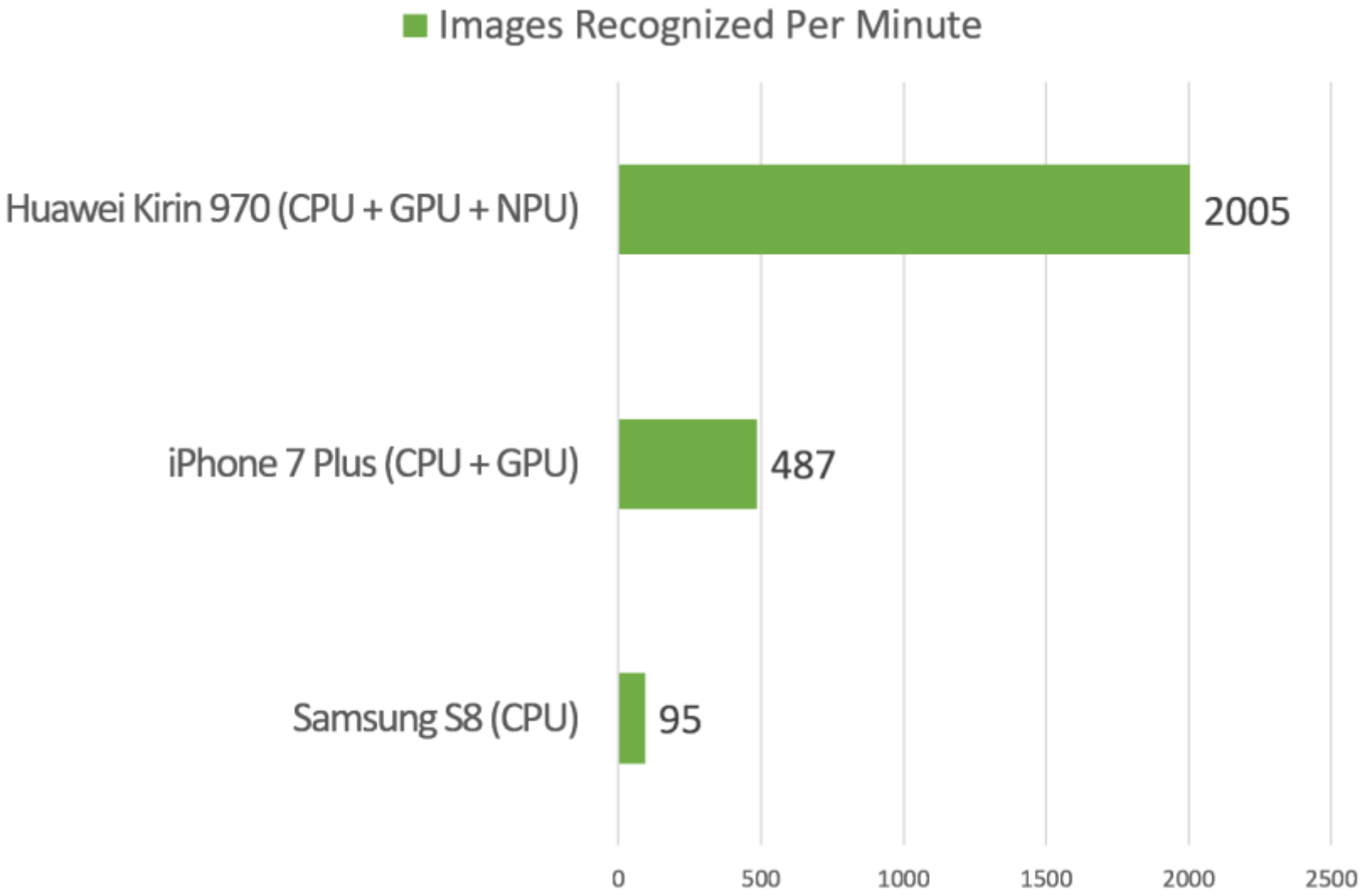
# 2.3 The Kirin 970 (6)

Efficiency and performance of an NPU vs. GPU and CPU for AI computing [20]



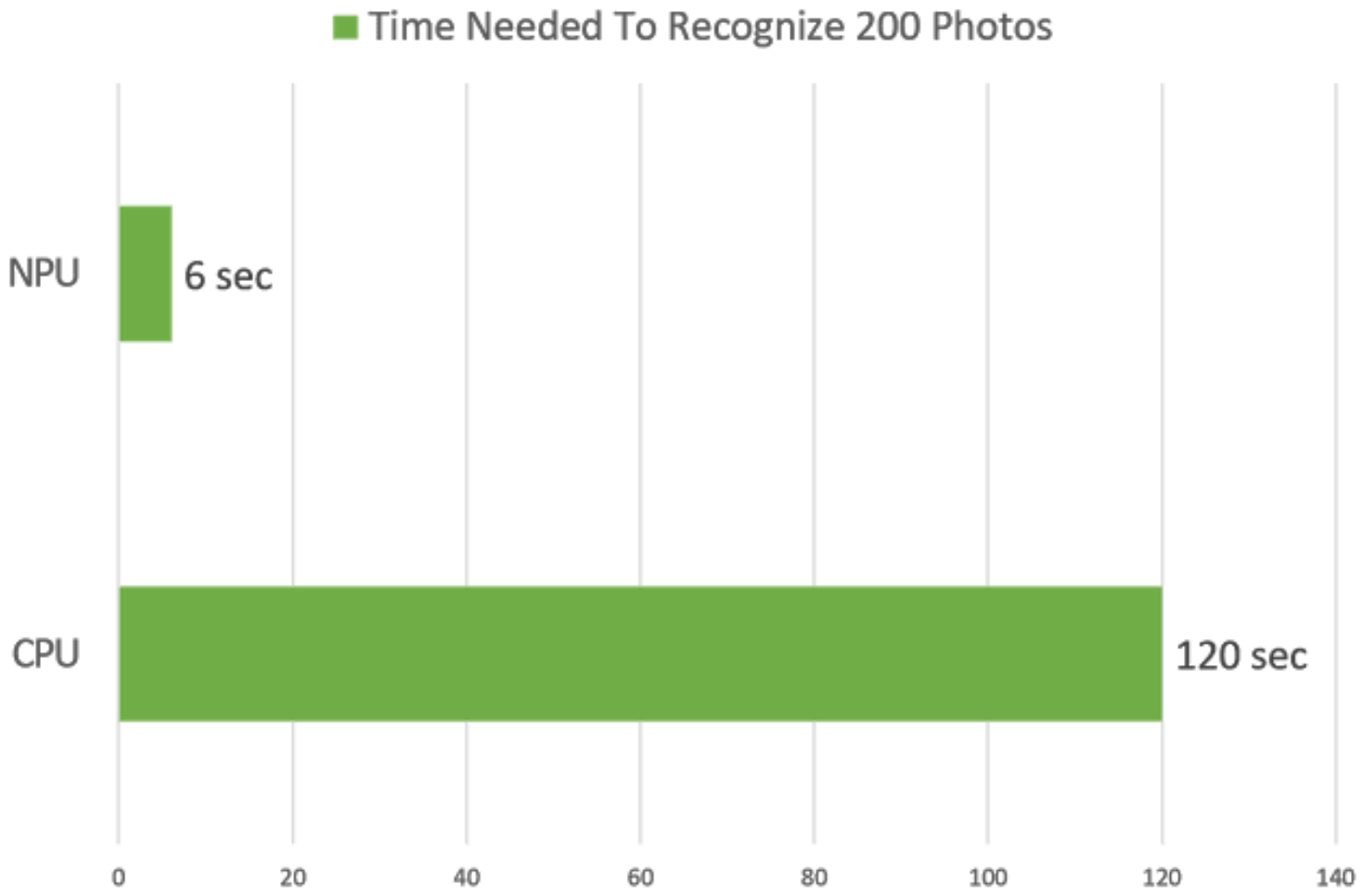
# 2.3 The Kirin 970 (7)

## Performance of image recognition in select smartphones [20]



# 2.3 The Kirin 970 (8)

Time to recognize 200 photos with and without an NPU in the Kirin 970 [20]



# 2.3 The Kirin 970 (9)

## The Neural Processing Unit (NPU)

- Kirin 970's NPU is licensed from Cambricon Technologies Ltd. (a Chinese upstart), actually it is their 1A NPU design.
- It can perform 1.92 TFLOPS FP16 operations.
- At this time there are no technical details available regarding the 1A only relating to its predecessor, a prototype design (on 65 nm technology).
- Nevertheless, to give an insight of Kirin 970's NPU, below we discuss briefly the previous prototype design.

Remark

Cambricon is named after the Cambrian explosion, a relatively short epoch when animal diversity sudden exploded, roughly 540 million years ago, as indicated below.

Phylum:
Group of
related organisms

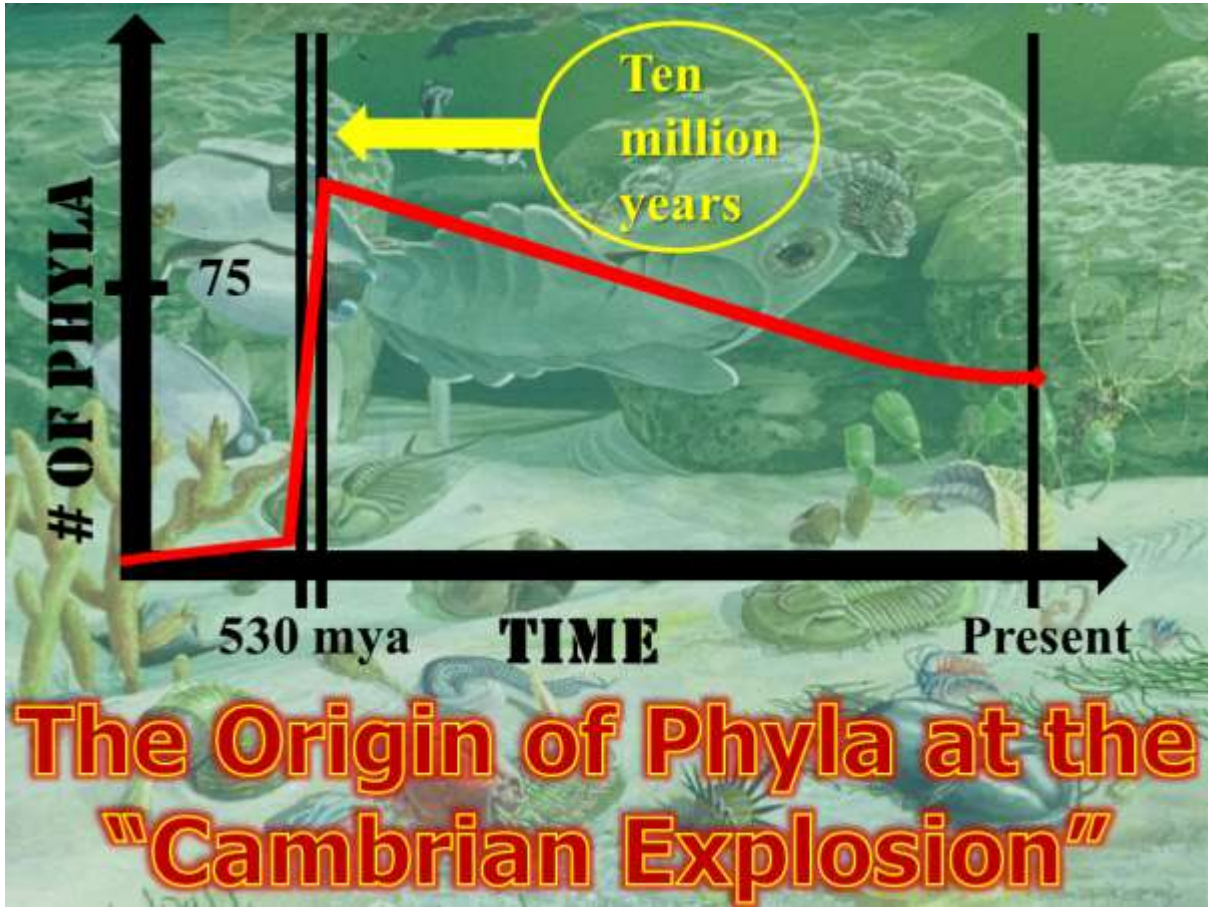



Figure: The Cambrian explosion [21]

## Overview of a previous design to the 1A NPU (a prototype design) [22]

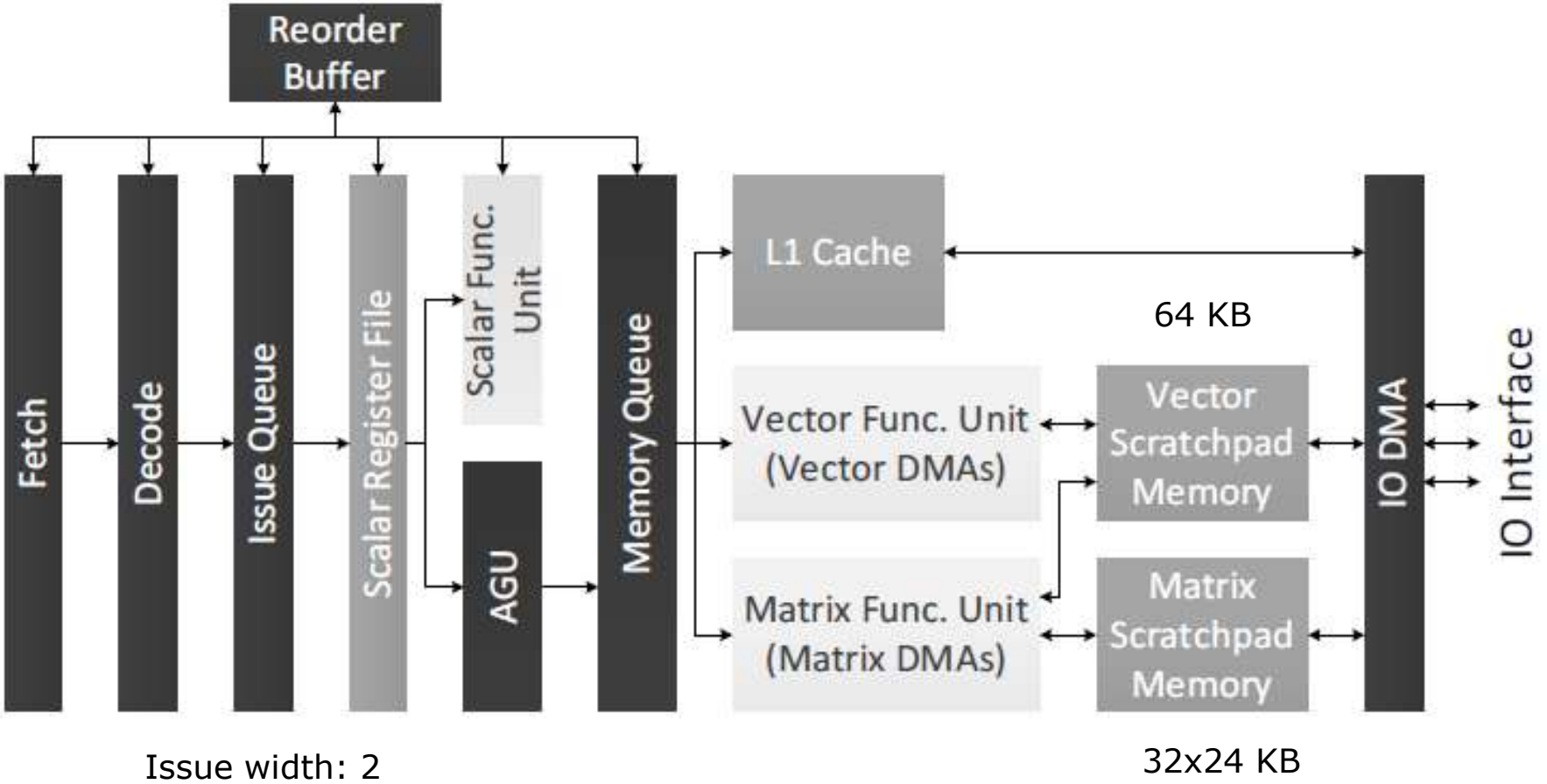
- Key feature of the prototype NPU design is the capability to perform matrix operations by means of 1024 multipliers and 1024 adders, while data are kept an an on-die scratchpad memory of 768 KB.
- The similarity to GPUs is that both NPUs and GPUs are capable to perform operations on matrices, whereas the significant difference between NPUs and GPUs is that GPUs hold the data to be processed in the main memory, whereas the implemented prototype NPU accelerator keeps data in an on-die scratchpad memory.

  This results in notable higher performance and efficiency compared to GPU execution.
- A further difference between NPUs and GPUs is that GPUs have to process basically FP32 data (for 3D representation) or FP64 data if scientific computing by means of OpenCL or CUDA) is supported, whereas NPUs process mainly FP16 or FP8 data considering the point how precisely NN weights need to be represented.

## Overview of the instruction types of Cambricon [22]

| Instruction Type | | Examples | Operands |
|---|---|---|---|
| Control | | jump, conditional branch | register (scalar value), immediate |
| Data Transfer | Matrix | matrix load/store/move | register (matrix address/size, scalar value), immediate |
| | Vector | vector load/store/move | register (vector address/size, scalar value), immediate |
| | Scalar | scalar load/store/move | register (scalar value), immediate |
| Computational | Matrix | matrix multiply vector, vector multiply matrix, matrix multiply scalar, outer product, matrix add matrix, matrix subtract matrix | register (matrix/vector address/size, s-calar value) |
| | Vector | vector elementary arithmetics (add, subtract, multiply, divide), vector transcendental functions (exponential, logarithmic), dot product, random vector generator, maximum/minimum of a vector | register (vector address/size, scalar value) |
| | Scalar | scalar elementary arithmetics, scalar transcendental functions | register (scalar value), immediate |
| Logical | Vector | vector compare (greater than, equal), vector logical operations (and, or, inverter), vector greater than merge | register (vector address/size, scalar) |
| | Scalar | scalar compare, scalar logical operations | register (scalar), immediate |

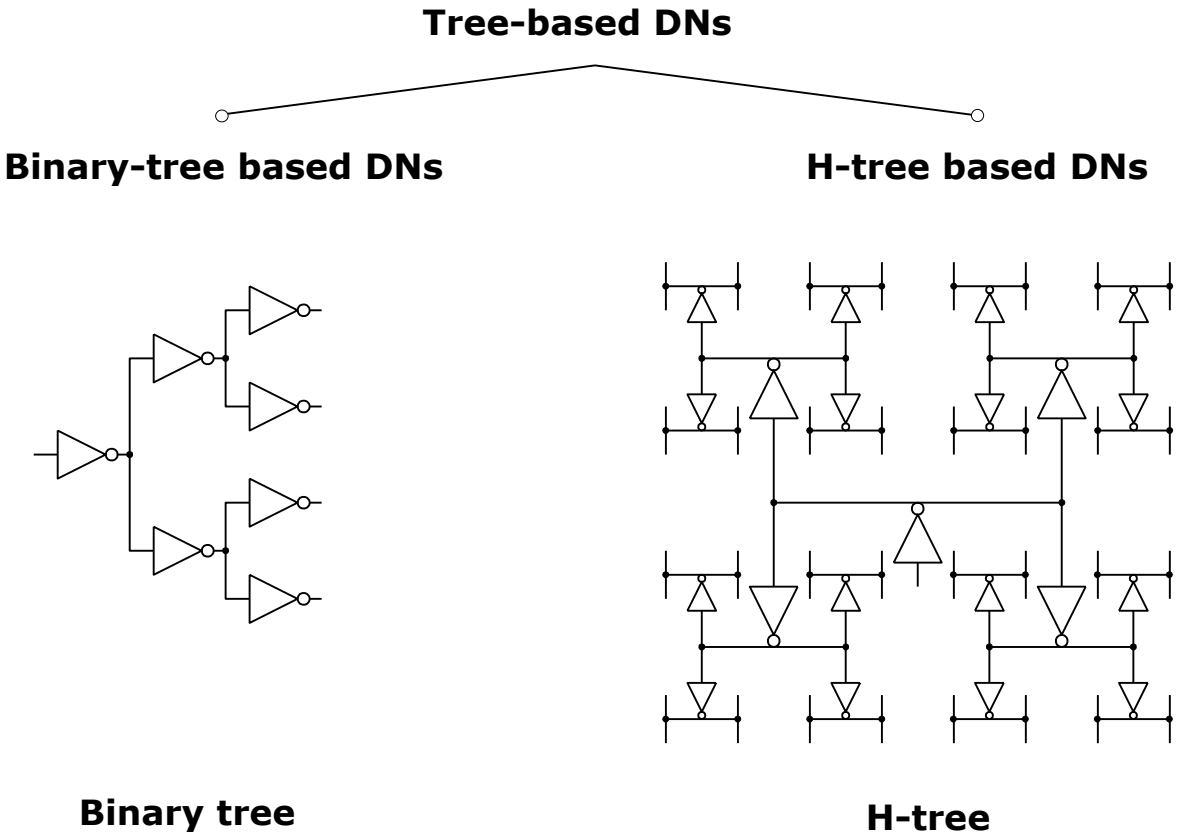
## Block diagram of the Cambrion prototype NPU [22]



## Implementation of the matrix computational unit [22] -1

- The 1 K adders and 1 K multipliers has been divided into 32 separate computational blocks to avoid excessive wire congestion and power consumption on long distance data transfers.
- Instead, the 32 computational blocks are connected through an h-tree bus that broadcasts input values to each computational block and collects output values from each block.

Remark: Tree-based distribution networks (DNs) [23]



## Implementation of the matrix computational unit [22] -2

- Each computational block possesses a 24 KB scratchpad, thus 32 computational blocks have altogether 32 x 24 = 768 KB scratchpad memory.

## Open Mobile AI Platform [24]

- To attract 3rd party interest in the Ai capabilities of the Kirin 970 Huawei announced an "Open Platform for mobile AI" initiatives.
- This allows developers to use either Huawei's own APIs, Google's TensorFlow or Facebook's Caffe2.

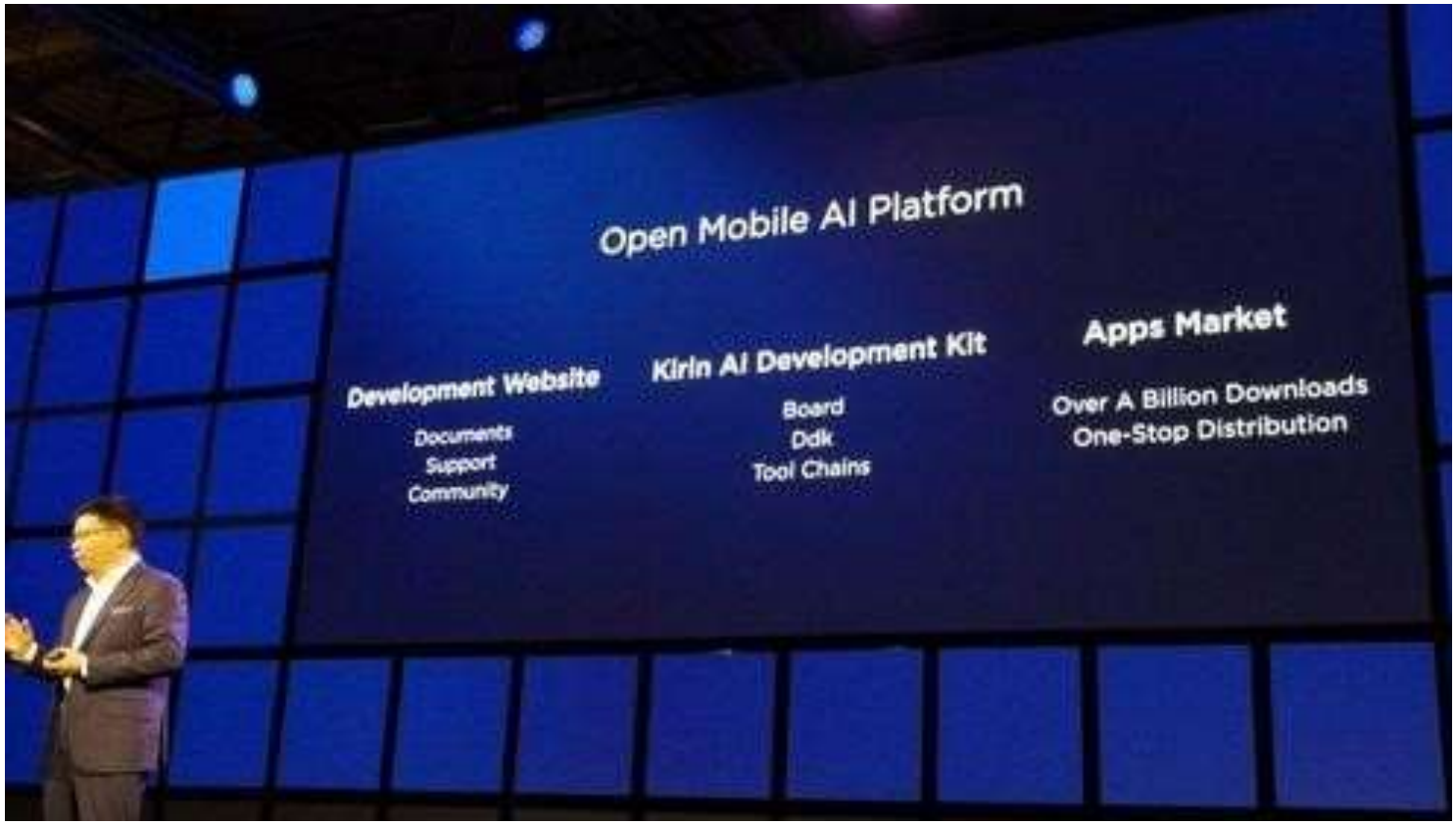
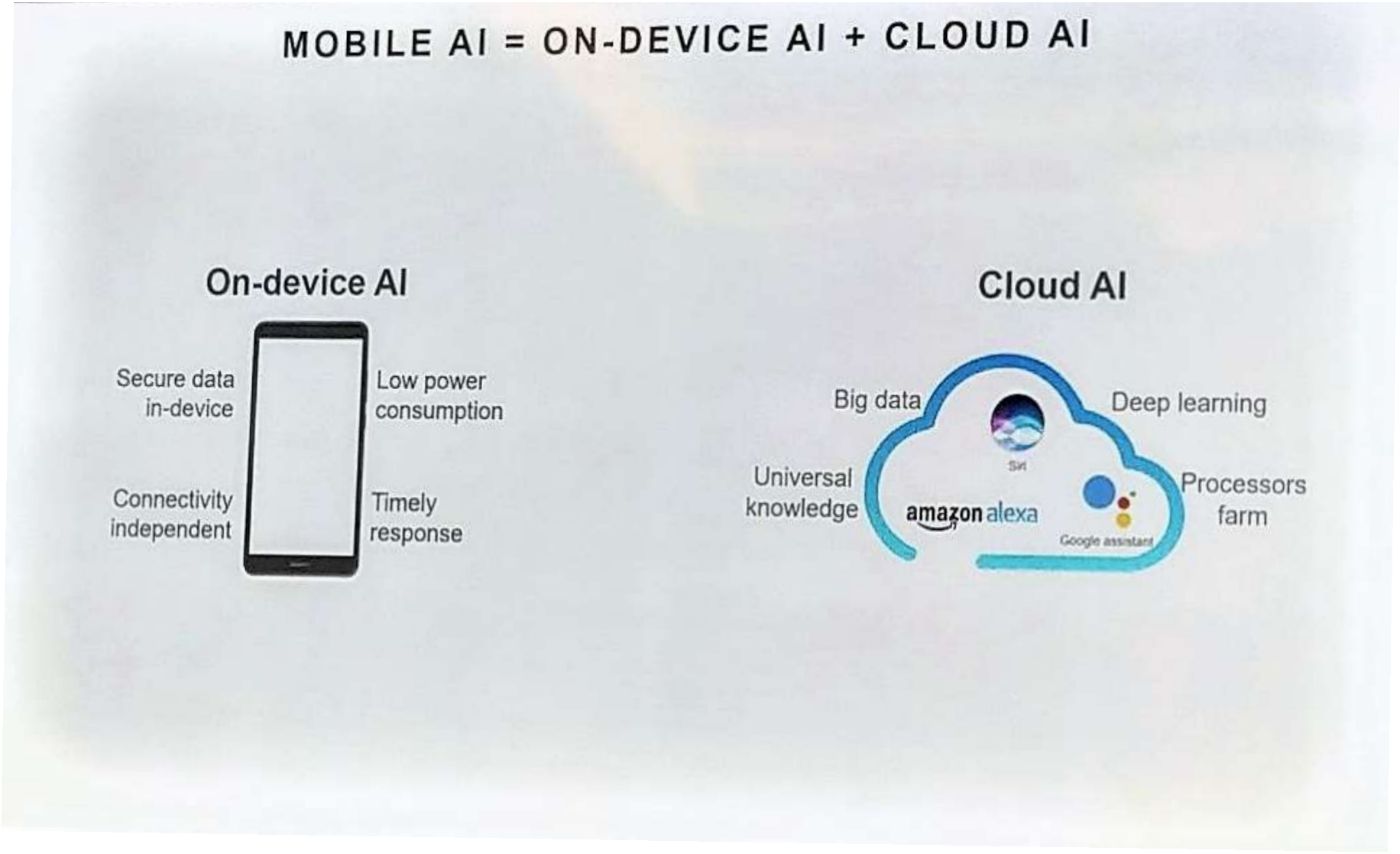



Figure: Huawei's announcement of the "Open Mobile AI platform" in 9/2017 [25]

## Positioning of on-device AI [20]



# 2.4 The Kirin 980

## 2.4 The Kirin 980

- Announced in 08/2018, appeared in devices in 10/2018.
- It is the kernel part of Huawei's Mate 20, Mate 20 Pro and Honor V11??? (cheaper version).
- It is Huawei's first 7 nm application processor, implemented on a die area of 74.1 $mm^2$ and includes about 6.9 billion transistors, about 1.4 billion more than the Kirin 970.

  (By contrast Apple's 7 nm A12 has the same number of transistors (6.9 billion) as the Kirin 980)

# 2.4 The Kirin 980 (2)

Design process of Huawei's Kirin 980 [26]

# 2.4 The Kirin 980 (3)

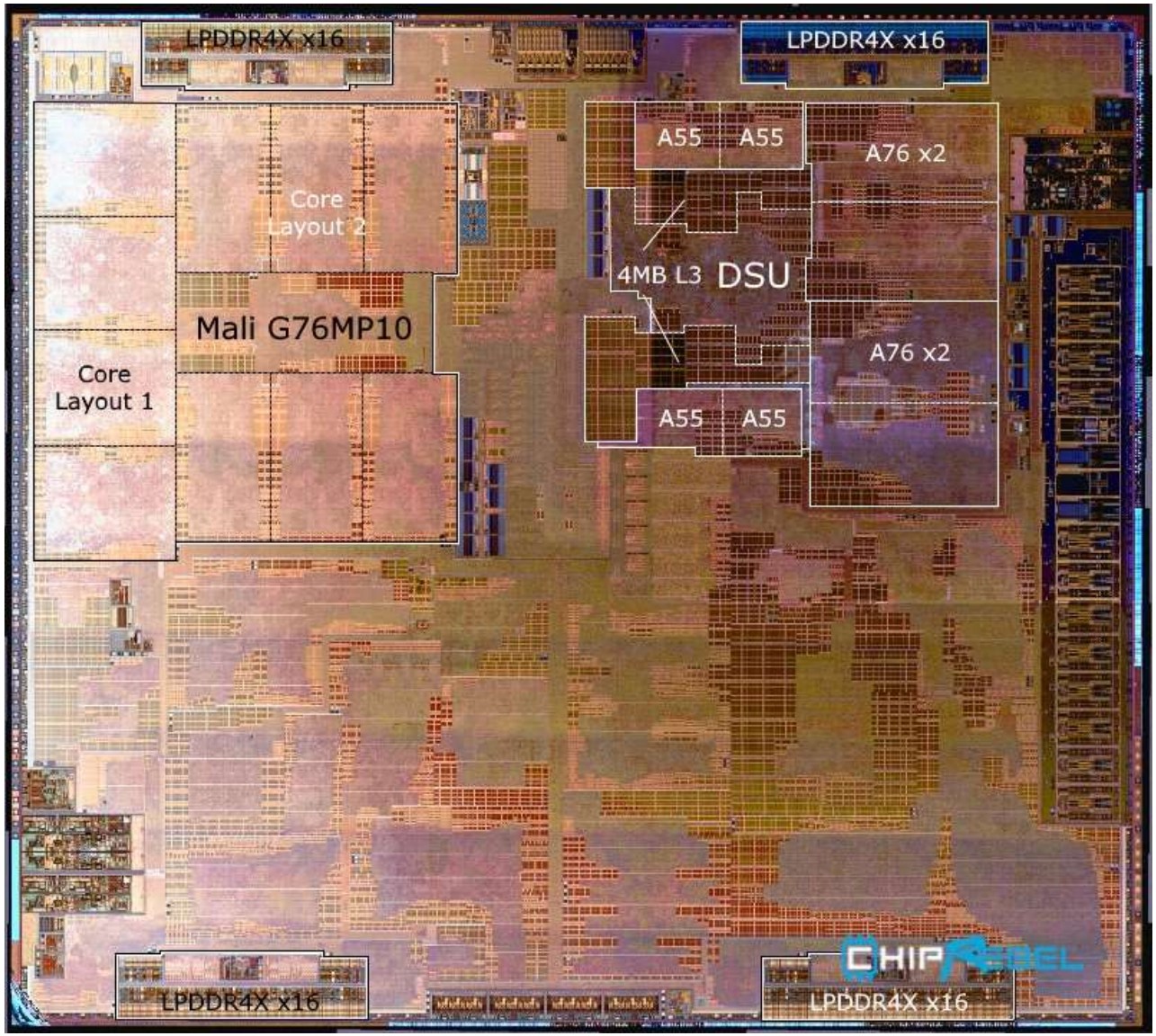
Evolution of transistor counts in Huawei's Kirin line [26]

# 2.4 The Kirin 980 (4)

Cortex A-76-based processor complex of the Kirin 980 processor [26]

# 2.4 The Kirin 980 (5)

## Die micrograph of the Kirin 980 [27]



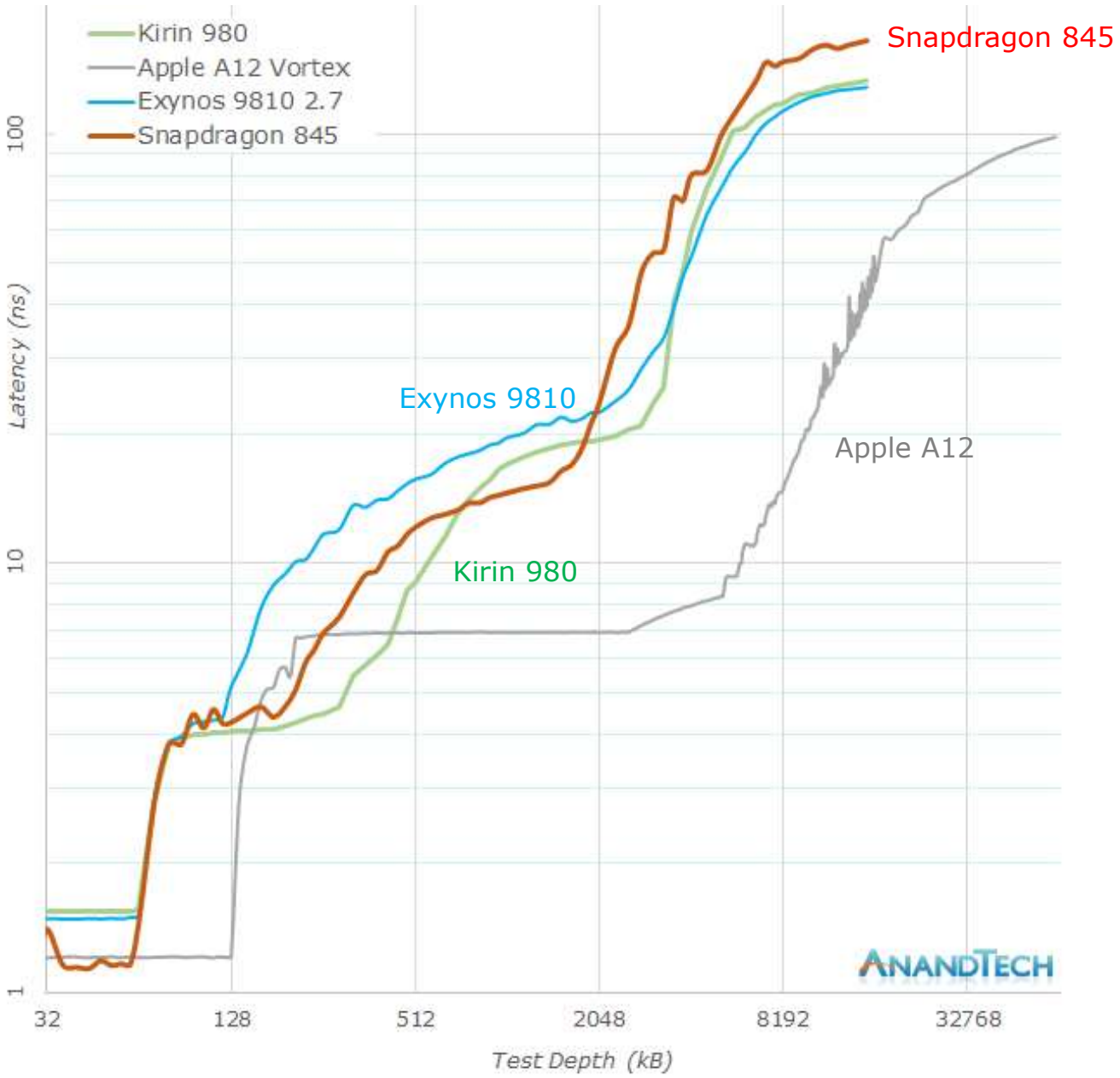
# 2.4 The Kirin 980 (6)

## Main features of the Kirin 960 to 980 processors [26]

| HiSilicon High-End Kirin SoC Lineup | | | |
|---|---|---|---|
| SoC | Kirin 980 | Kirin 970 | Kirin 960 |
| **CPU** | 2x A76 @ 2.60 GHz<br>2x A76 @ 1.92 GHz<br>@ 512KB L2's<br>4x A55 @ 1.80 GHz<br>@ 128KB L2's<br>4MB DSU L3 | 4x A73 @ 2.36 GHz<br>4x A53 @ 1.84 GHz<br>2MB L2 | 4x A73 @ 2.36GHz<br>4x A53 @ 1.84GHz<br>2MB L2 |
| **GPU** | ARM Mali-G76MP10 @ 720 MHz | ARM Mali-G72MP12 @ 746 MHz | ARM Mali-G71MP8 @ 1037MHz |
| **LPDDR4 Memory** | 4x 16-bit Channel<br>LPDDR4X @ 2133MHz<br>34.1GB/s | 4x 16-bit Channel<br>LPDDR4X @ 1833 MHz<br>29.9GB/s | 4x 16-bit Channel<br>LPDDR4 @ 1866MHz 29.9GB/s |
| **Storage I/F** | UFS 2.1 | UFS 2.1 | UFS 2.1 |
| **ISP/Camera** | New Dual ISP +46% speed<br>10-bit pipeline | Dual 14-bit ISP | Dual 14-bit ISP Improved) |
| **Encode/Decode** | 2160p60 Decode<br>2160p30 Encode | 2160p60 Decode<br>2160p30 Encode | 1080p H.264 Decode & Encode<br>2160p30 HEVC Decode |
| **Integrated Modem** | Kirin 980 Integrated LTE<br>(Category 21/18)<br>DL = 1400 Mbps UL = 200 Mbps | Kirin 970 Integrated LTE<br>(Category 18/13)<br>DL = 1200 Mbps UL = 150 Mbps | Kirin 960 Integrated LTE<br>(Category 12/13)<br>DL = 600Mbps UL = 150Mbps |
| **Sensor Hub** | i8 | i7 | i6 |
| **NPU** | Dual @ >2x perf | NPU | No |
| **Mfc. Process** | TSMC 7nm | TSMC 10nm | TSMC 16nm FFC |

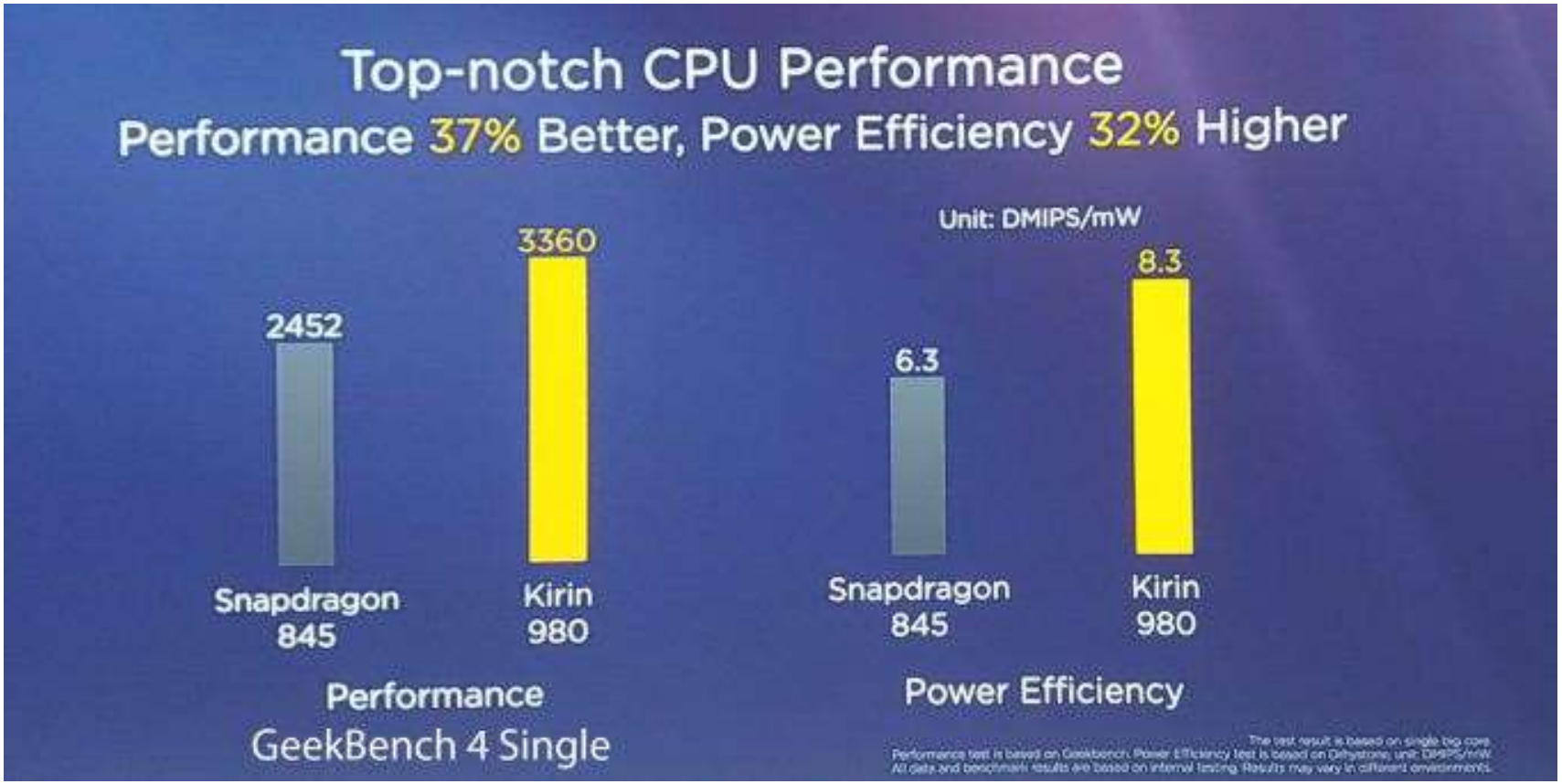
# 2.4 The Kirin 980 (7)

Cache and memory latency of select processors [27]

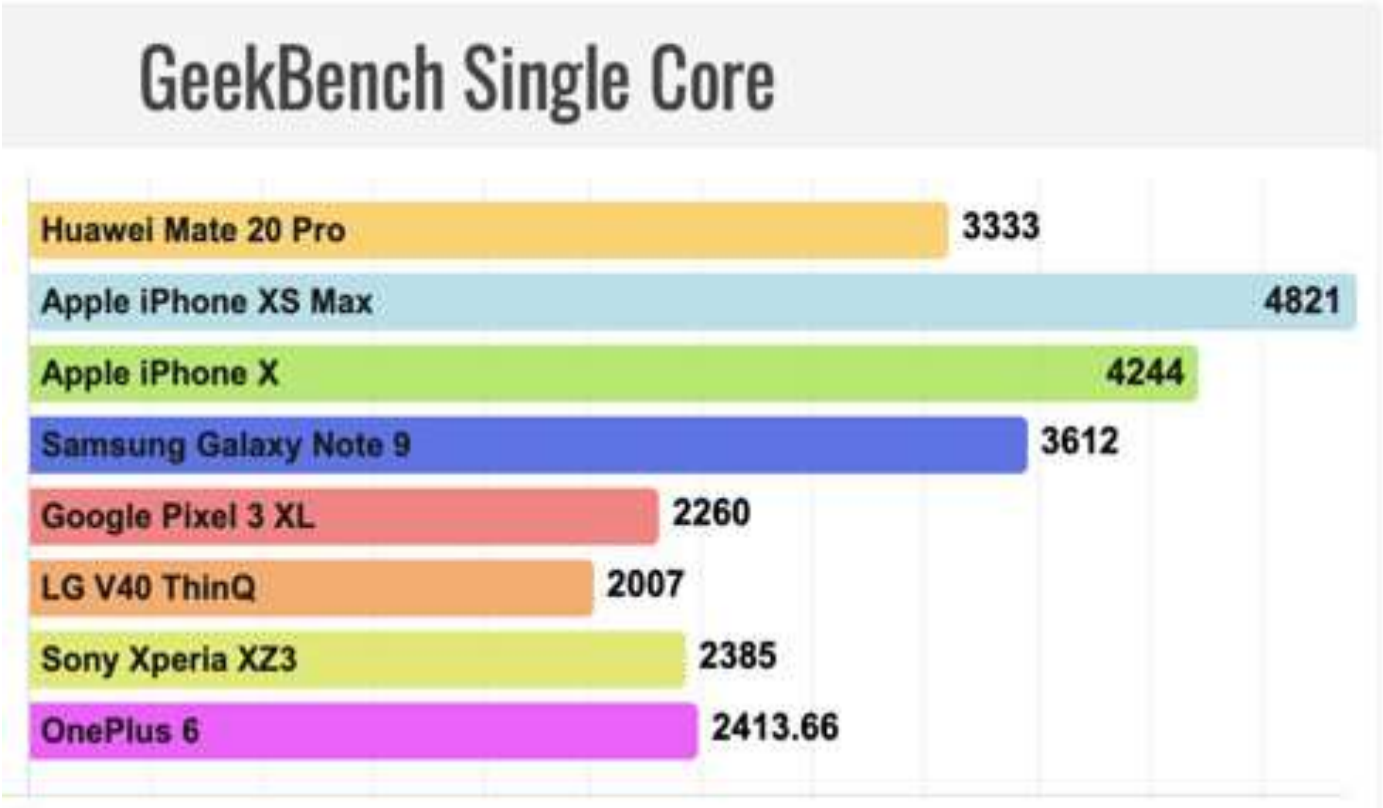


# 2.4 The Kirin 980 (8)

Performance and power efficiency of Kirin 980 vs. Qualcomm's Snapdragon 845 [26]



## Geekbench 4 Single core benchmark scores [28]




Geekbench 4 measures CPU performance vs. a reference platform (Core i7-6600U (Skylake)) with an assigned score of 4000

## Geekbench 4 Multicore benchmark scores [28]

Huawei Kirin 980

Apple A12 Bionic

Apple A11 Bionic

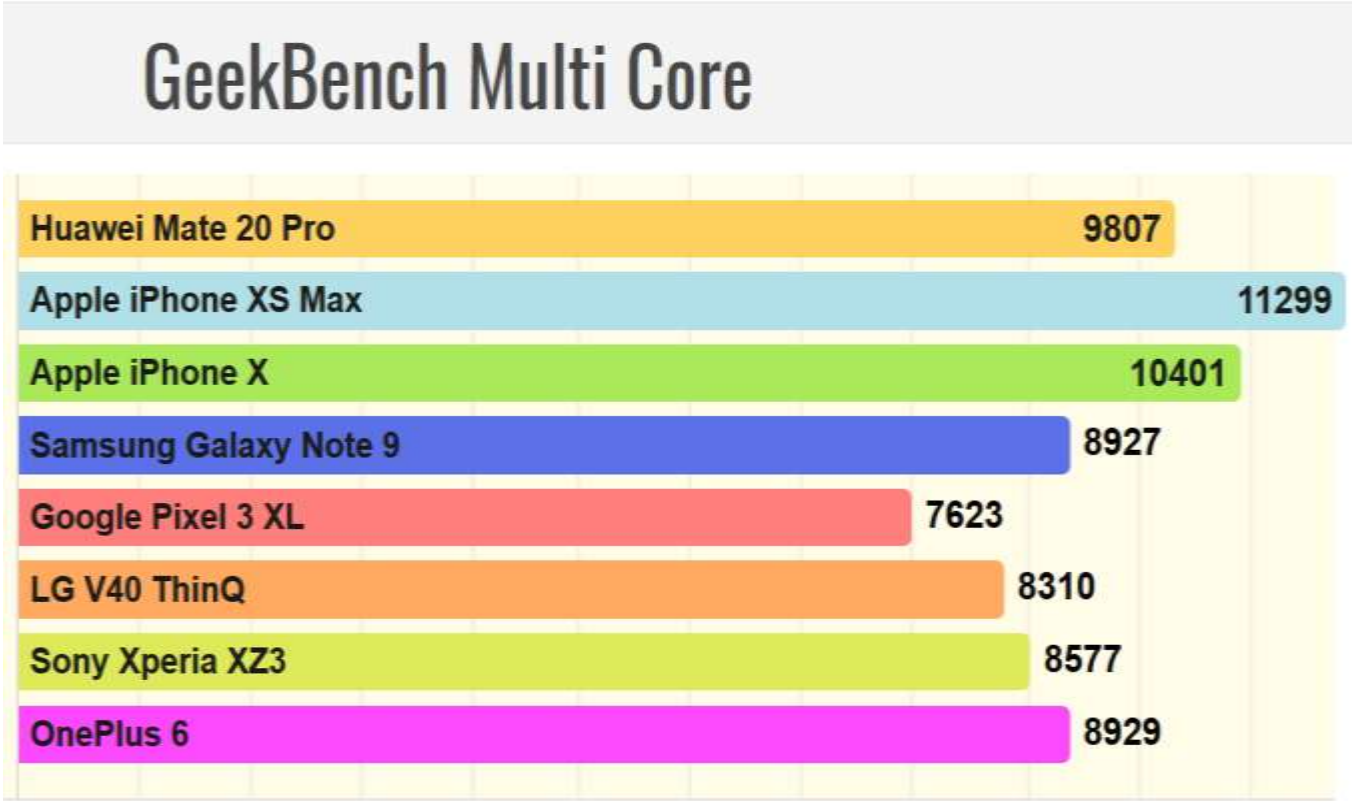
Samsung Exynos 9810

Qualcomm Snapdragon 845

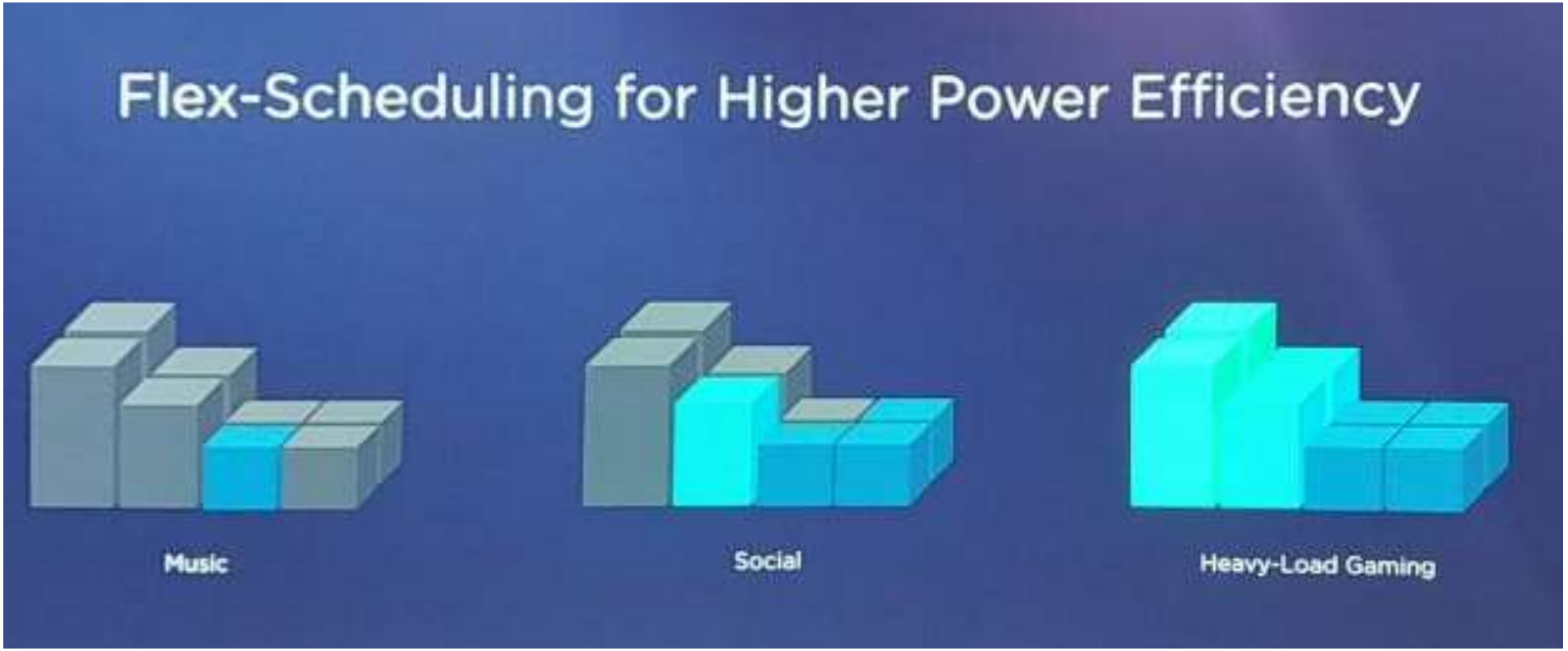
Qualcomm Snapdragon 845

Qualcomm Snapdragon 845

Qualcomm Snapdragon 845



Core scheduling examples for a few use cases [26]

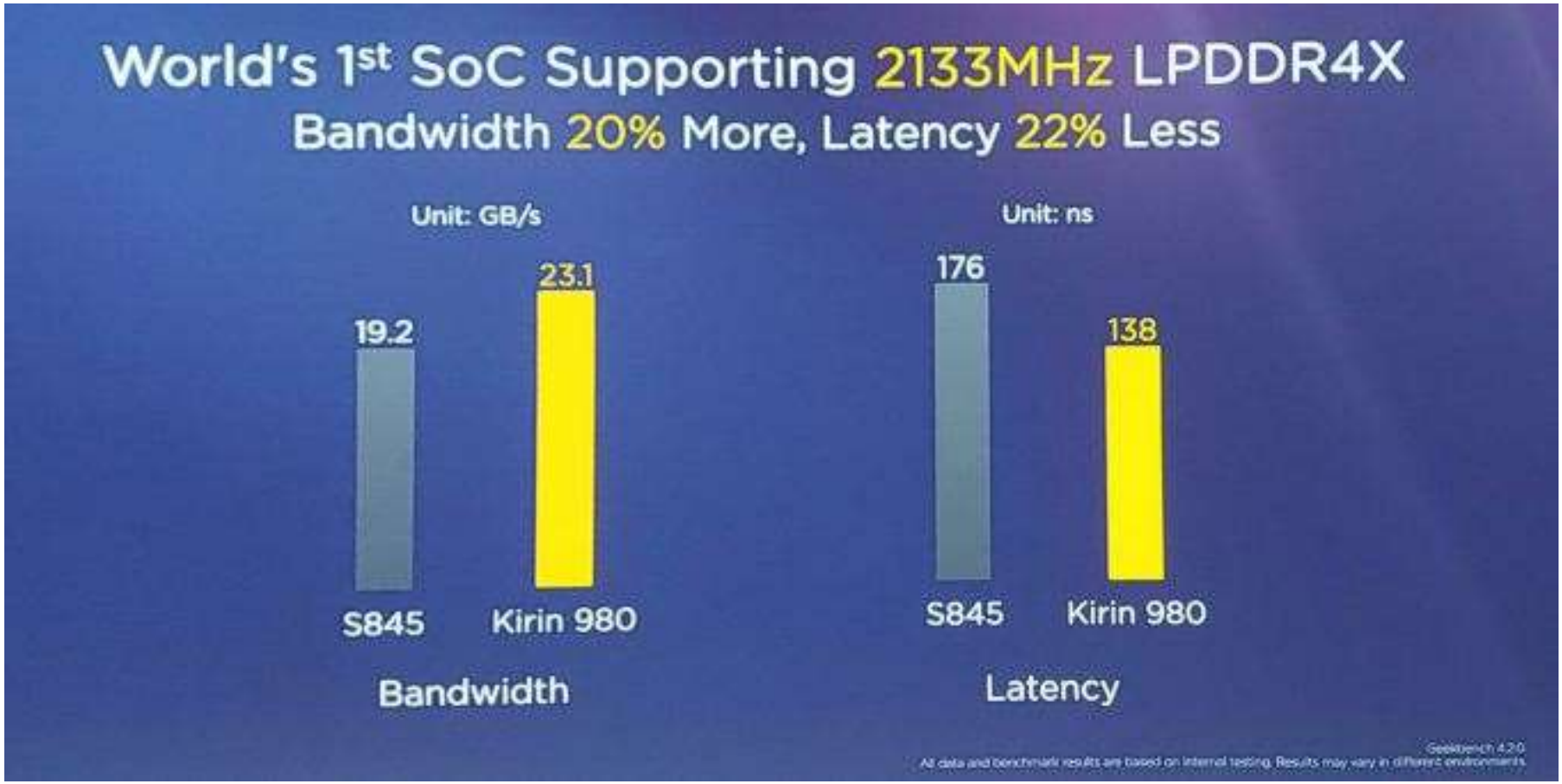



Allocation of the cores: Inclusive core allocation (called also Global Task Scheduling (GTS))

Gray means not allocated.

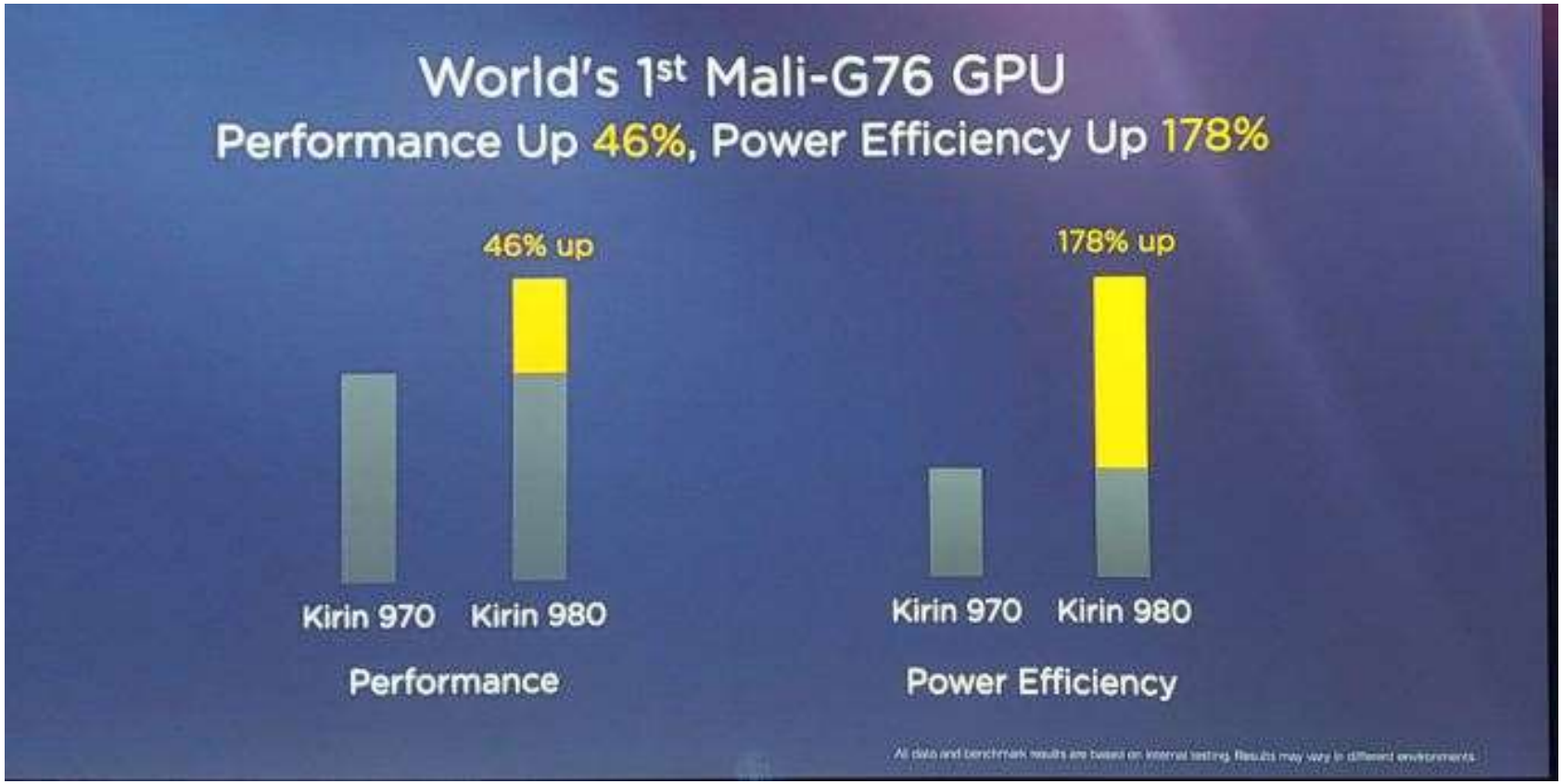
# 2.4 The Kirin 980 (12)

Memory bandwidth and latency of the Kirin 980 vs. Qualcomm's Snapdragon 845 [26 ]

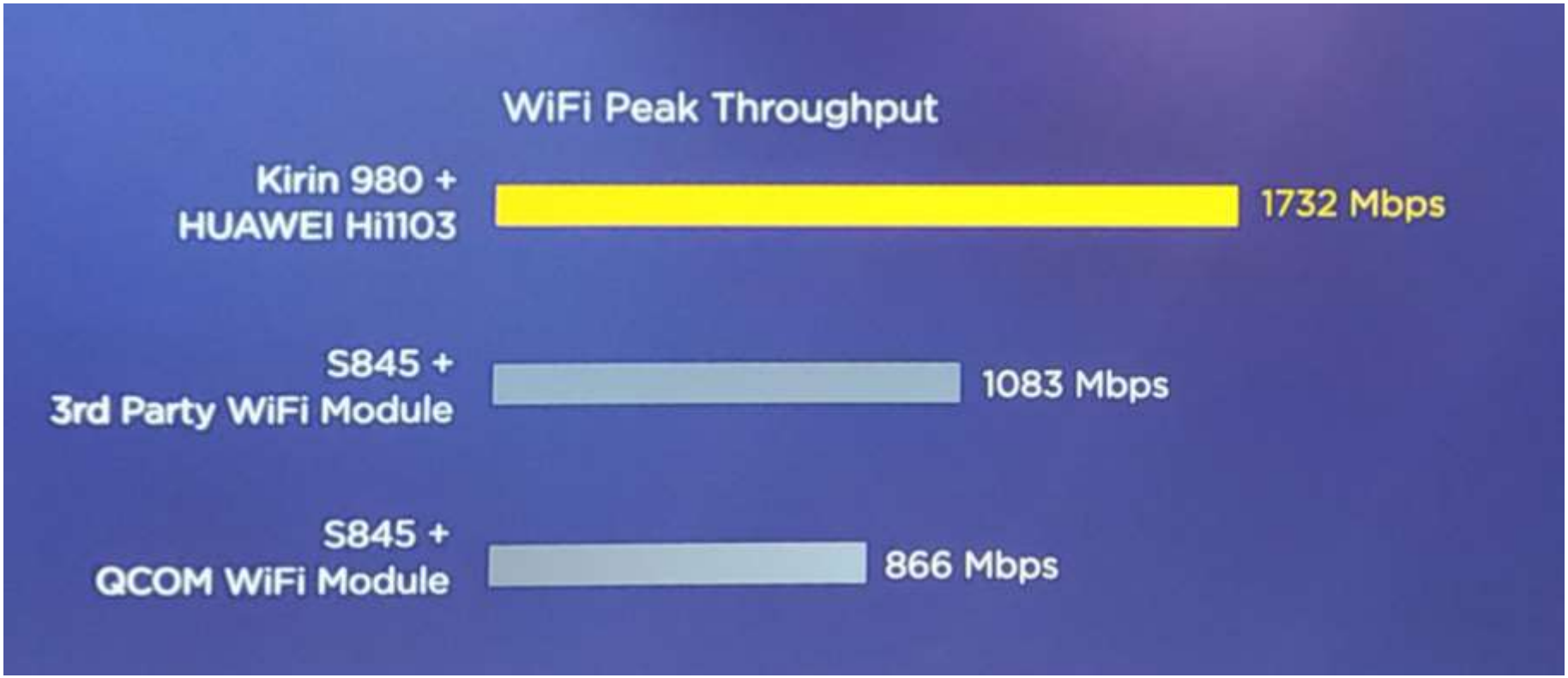



Note: Kirin 980 is world's 1st SoC supporting 2133 MT/s LPDDR4x [26]

Performance and power efficiency of the Mali-G76 GPU vs. the Mali-G72 used in the Kirin 970 [26]



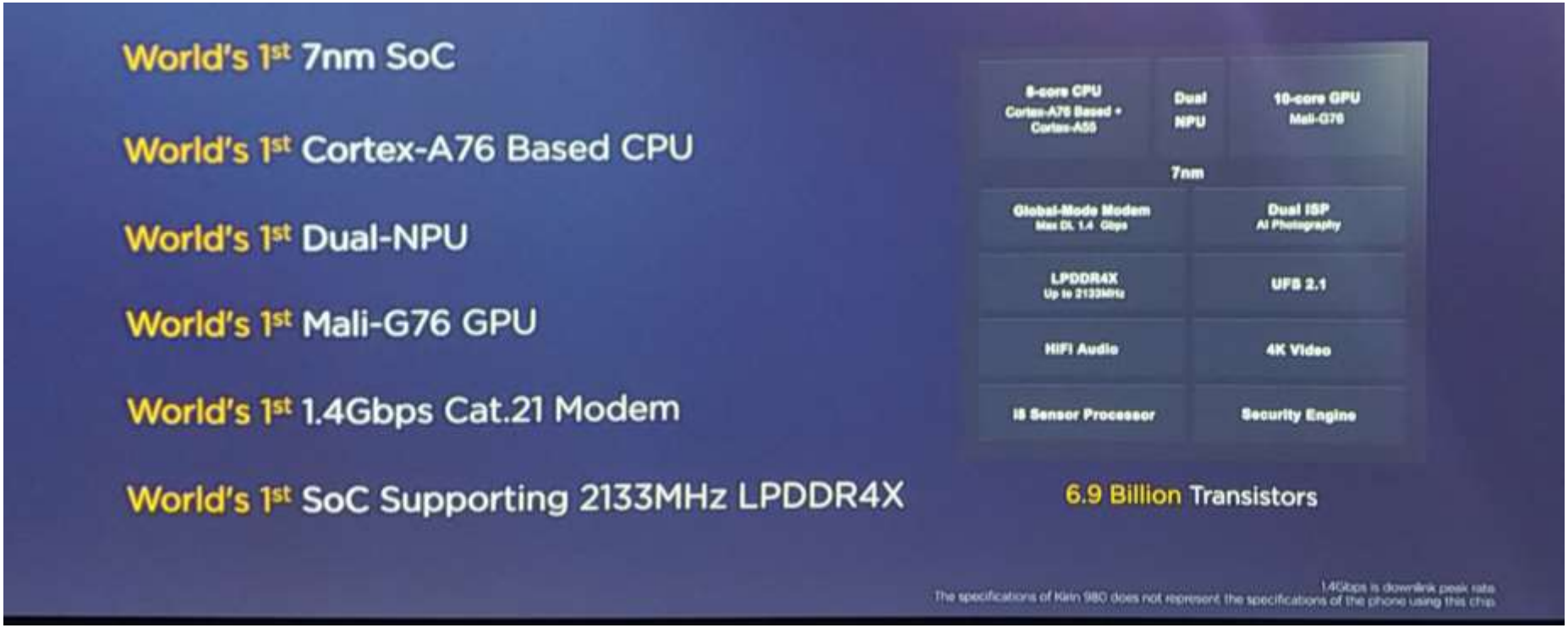
## WiFi speed of Kirin 980 vs. Qualcomm's Snapdragon 845 [26]




According to Huawei it is the world's fastest WiFi at the time being [26]

# 2.4 The Kirin 980 ()

## Outstanding features of the Kirin 980 [5]



# 3. References

# 3. References (1)


[1]: Wikipedia, Huawei, https://en.wikipedia.org/wiki/Huawei

[2]: HiSilicon Company Overview, http://www.hisilicon.com/en/AboutUs/Overview

[3]: Smartphone shipments by vendor worldwide from 4th quarter 2009 to 1st quarter 2017 (in million units), Statista, https://www.statista.com/statistics/271490/quarterly-global-smartphone-shipments-by-vendor/

[4]: Smartphone Apps Processor Revenue Reached $4.7 Billion in Q1 2014 says Strategy Analytics, Strategy Analytics, July 14 2014, http://www.strategyanalytics.com/default.aspx?mod=pressreleaseviewer&a0=5543

[5]: Chip Brand Distribution and Market Shares, Q3 2016, Antutu, Nov. 25 2016, http://www.antutu.com/en/doc/107644.htm

[6]: Wikipedia, Qilin, https://en.wikipedia.org/wiki/Qilin

[7]: Wikipedia, HiSilicon, https://en.wikipedia.org/wiki/HiSilicon

[8]: Lucic K., Huawei HiSilicon Kirin 940 And Kirin 950 SoC Specifications Surface, Android Headlines, March 12 2015, https://www.androidheadlines.com/2015/03/huaweis-hisilicon-kirin-940-kirin-950-soc-specifications-surface.html

[9]: Does not blow black | Interpretation Huawei Kirin 950 strength exactly geometry, GoodChinaBrand, Nov. 18 2015, http://y.tt/51374800070en.html

[10]: Hugosson O., Mali-V500 video processor: reducing memory bandwidth with AFBC, ARM Community, Sept. 11 2013, https://community.arm.com/graphics/b/blog/posts/mali-v500-video-processor-reducing-memory-bandwidth-with-afbc

# 3. References (2)


[11]: Nickinson P., Live from Beijing: Huawei details the Kirin 950 chipset, Android Central, Nov. 4 2015, http://www.androidcentral.com/live-beijing-huawei-details-kirin-950-chipset

[12]: Wikipedia, Wideband audio, https://en.wikipedia.org/wiki/Wideband_audio

[13]: Loberg C., Understand and perform testing for MIPI M-PHY compliance, EDN Network, Febr. 19 2013, https://www.edn.com/design/test-and-measurement/4407197/Understand-and-perform-testing-for-MIPI-M-PHY-compliance

[14]: eMMC to UFS: How NAND Memory for Mobile Products Is Evolving, Samsung Newsroom, Febr. 26 2015, https://news.samsung.com/global/emmc-to-ufs-how-nand-memory-for-mobile-products-is-evolving

[15]: Cutress I., Frumusanu A., Huawei announces the HiSilicon Kirin 960: 4xA73 + 4xA53, G71MP8, CDMA, AnandTech, Oct. 19 2016, http://www.anandtech.com/show/10766/huawei-announces-hisilicon-kirin-960-a73-g71

[16]: Brutális teljesítményt ígér a Kirin 960, iPon, Oct. 19 2016, https://ipon.hu/hir/brutalis_teljesitmenyt_iger_a_kirin_960/38924

[17]: Davies J., The Bifrost GPU architecture and the ARM Mali-G71 GPU, Hot Chips 28, Aug. 2016, https://www.hotchips.org/wp-content/uploads/hc_archives/hc28/HC28.22-Monday-Epub/HC28.22.10-GPU-HPC-Epub/HC28.22.110-Bifrost-JemDavies-ARM-v04-9.pdf

[18]: Sims G., Huawei announces the Kirin 970 – new flagship SoC with AI capabilities, Android Authority, Sept. 2 2017, https://www.androidauthority.com/huawei-announces-kirin-970-797788/

# 3. References (3)


[19]: Esteve E., Tensilica Vision P6 DSP is Powering Huawei Kirin 970 Image, SemiWiki, Nov. 17 2017, https://www.semiwiki.com/forum/content/7140-tensilica-vision-p6-dsp-powering-huawei-kirin-970-image.html

[20]: Han Y.C., Huawei's Neural Processing Unit (NPU) And The Big Picture, Sept. 25 2017, http://www.thetechrevolutionist.com/2017/09/huaweis-neural-processing-unit-npu-and.htm

[21]: The Cambrian Explosion, Genesis Park, http://www.genesispark.com/exhibits/fossils/cambrian/

[22]: Liu S. & al., Cambricon: An Instruction Set Architecture for Neural Networks, Proc. 2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture, pp. 393-405

[23]: Clock Distribution, Acsel-lab.com, July 28 2004, http://www.acsel-lab.com/Projects/clocking/clock_distribution.htm

[24]: Low C., Mobile AI is Huawei's not-so-secret weapon, Engadget, Sept. 6 2017, https://www.engadget.com/2017/09/06/huawei-mobile-ai-ifa-2017/

[25]: Kirin 970 bets on the back of artificial intelligence: in addition to chips, there are data, ecology and clouds, China IT News, Sept. 11 2017, http://www.fonow.com/view/192681.html

[26]: Frumusanu A., HiSilicon Announces The Kirin 980: First A76, G76 on 7nm, AnandTech, Aug. 31, 2018, https://www.anandtech.com/show/13298/hisilicon-announces-the-kirin-980-first-a76-g76-on-7nm

# 3. References (4)

[27]: Frumusanu A., The Mate 20 & Mate 20 Pro Review: Kirin 980 Powering Two Contrasting Devices, AnandTech, Nov. 16, 2018, https://www.anandtech.com/show/13503/the-mate-20-mate-20-pro-review

[28]: Hristov V., Huawei Mate 20 Pro Performance Benchmarks: Kirin 980 inside, Android's first 7nm chip, PhoneArena, 16 Oct., 2018, https://www.phonearena.com/news/Huawei-Mate-20-Pro-Performance-Benchmarks-Kirin-980-inside-Androids-first-7nm-chip_id110051