

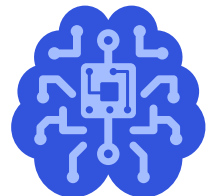
March 2018

@qualcomm\_tech

Qualcomm

# Achieving AI @Scale on Mobile Devices

Qualcomm Technologies, Inc.



Mobile is the largest computing platform in the world

~7.8 Billion

Cumulative smartphone unit shipments forecast between 2018-2022

Source: IDC Aug. '18





Qualcomm

**30** Years of driving the evolution of wireless

---

**#1** Fabless semiconductor company

---

**#1** in 3G/4G LTE modem

---

**842M**

MSM™ chipsets shipped FY'16

Source: Qualcomm Incorporated data. Currently, Qualcomm semiconductors are products of Qualcomm Technologies, Inc. or its subsidiaries. MSM is a product of Qualcomm Technologies, Inc. IHS, Mar. '17; Strategy Analytics, Mar. '17

Qualcomm Technologies' success is based on

# Technology leadership

## Modem

3G (CDMA, GSM)

4G/LTE

5G

## Connectivity

Bluetooth Smart

Bluetooth Mesh

802.11ac/ax

802.11ad

802.11n

802.15.4

DSRC

Powerline

GNSS/Location

Peripherals (PCIe, USB.)

## Connectivity



## Computing



## Processors

CPU

GPU

DSP

## Multimedia

Video

Speech

Display processing

Media processing

Camera processing

Audio processing

Computer Vision

AR / VR

## RF

Power amps

Acoustic filters

RF switches

## Other

Machine learning

NoC / MemCtrl

Power management

Security

Touch

Fingerprint

# Chipset complexity is growing dramatically

More frequent and more complex design cycles



Comparative scale

Circa 2000



**<10 Million**  
Transistors on a chip

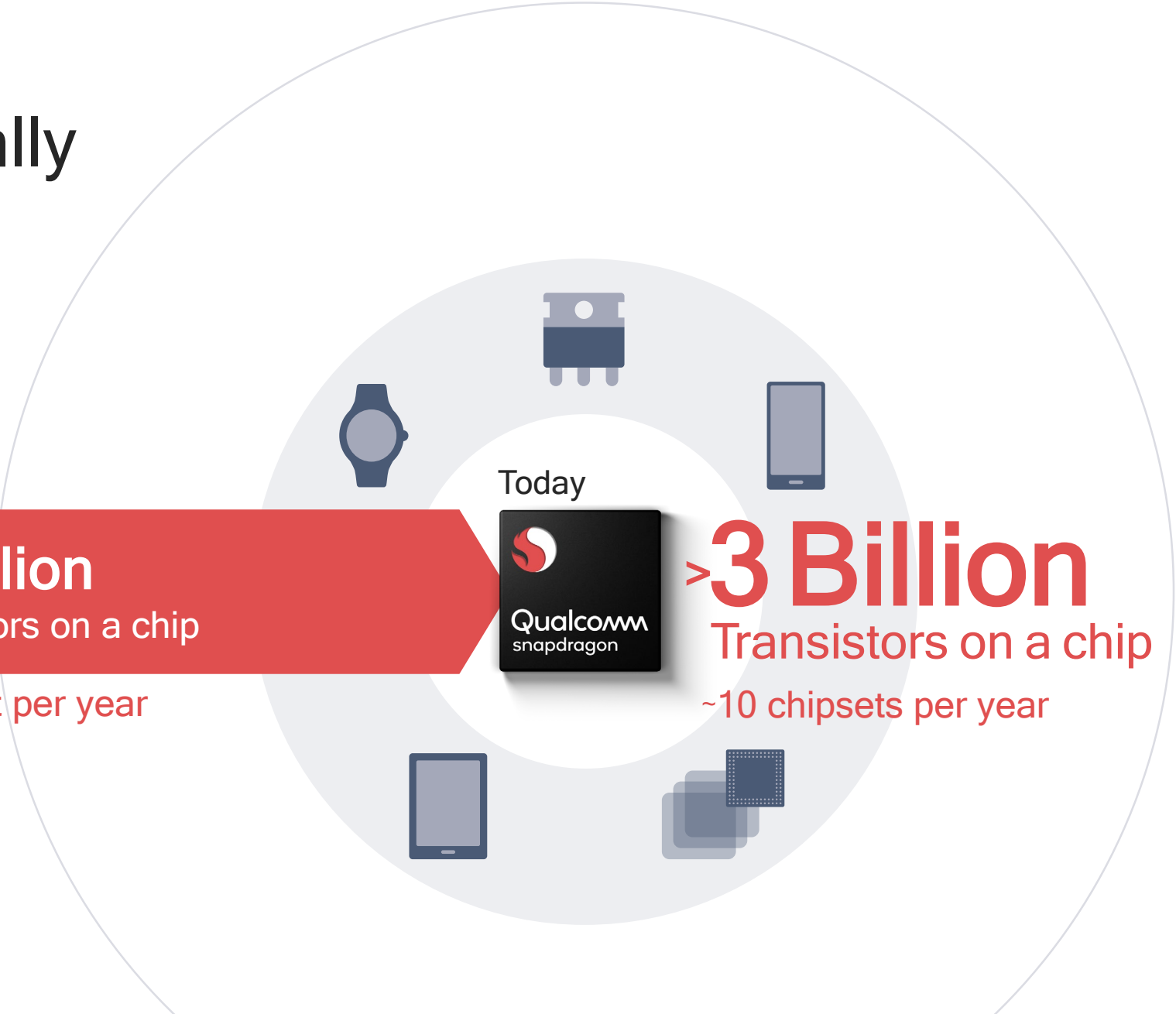
1 chipset per year

Today



**>3 Billion**  
Transistors on a chip

~10 chipsets per year



# A mobile processor today—Snapdragon 835

Highly integrated and complex SoC using 10nm process technology

## Snapdragon X16 LTE

World's first announced gigabit-class LTE modem

## Qualcomm® Hexagon™ DSP

Snapdragon Neural Processing Engine support

## Qualcomm® Kryo™ 280 CPU

Our most power efficient architecture to date



## Qualcomm® Adreno™ Visual Processing

25% faster graphics rendering  
60x more display colors\*

## Qualcomm Spectra™ Camera ISP

Smooth zoom | Fast-autofocus  
True to life colors

## Qualcomm® Mobile Security

First to support full biometric suite

\* Compared to Snapdragon 820

# Mobile scale changes everything

Mobile computing



Smart homes



Industrial IoT



Wearables



Smart cities



Automotive



Healthcare



Networking



Extended reality

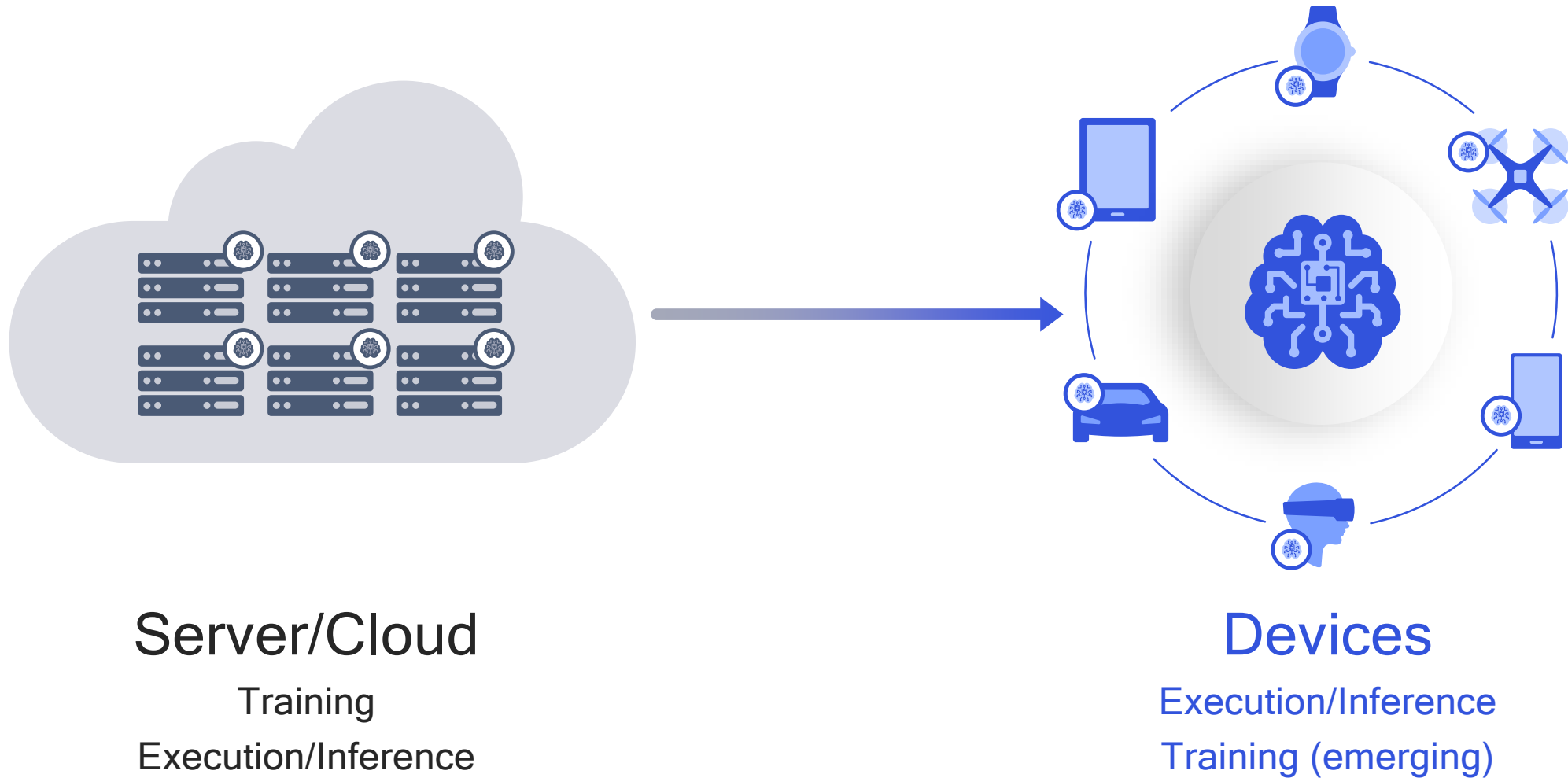


Rapid replacement cycles

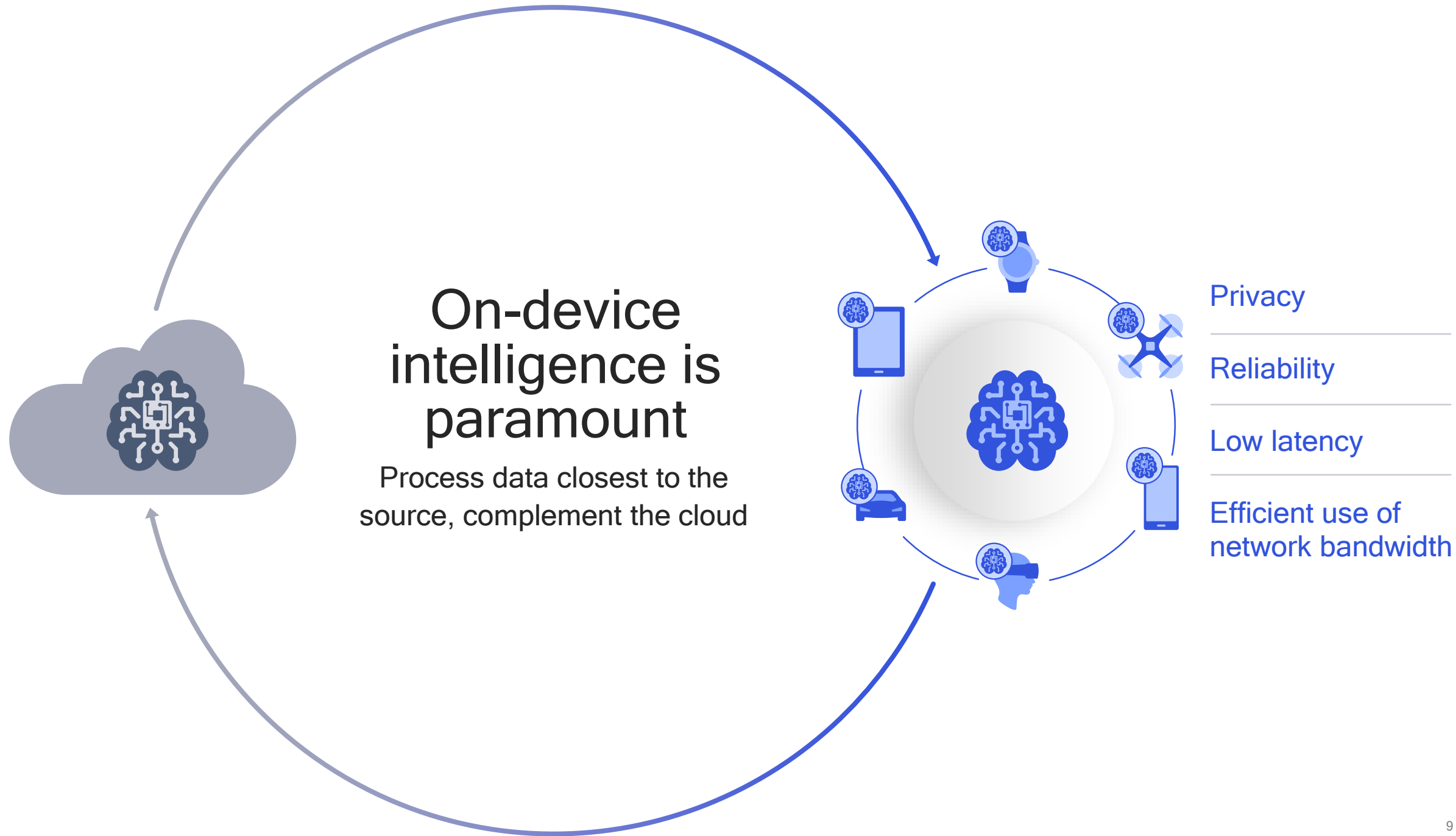
Superior scale

Integrated and optimized technologies

# Intelligence is moving to the device

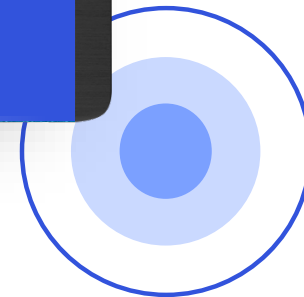
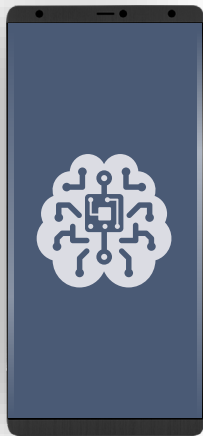
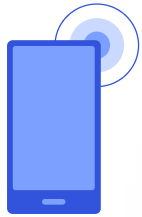
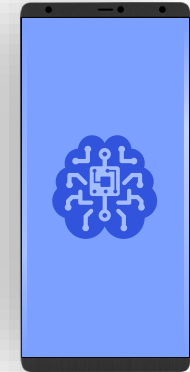
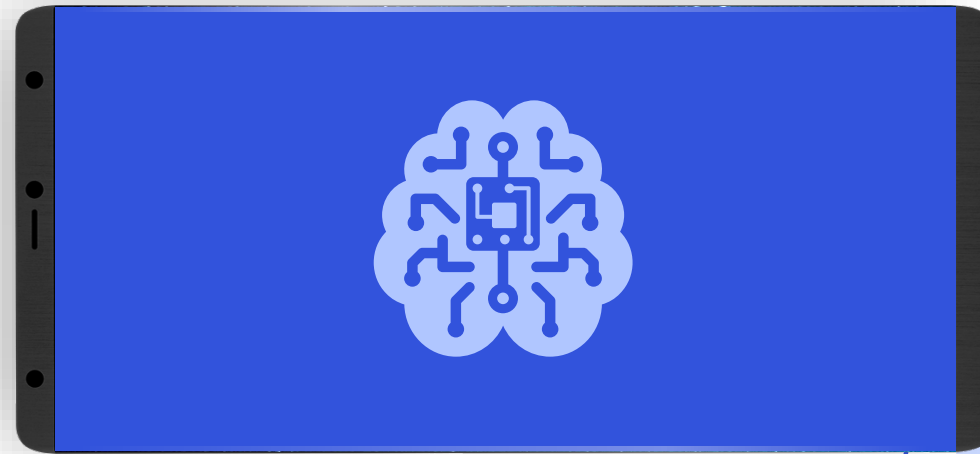








# Mobile is becoming the pervasive AI platform

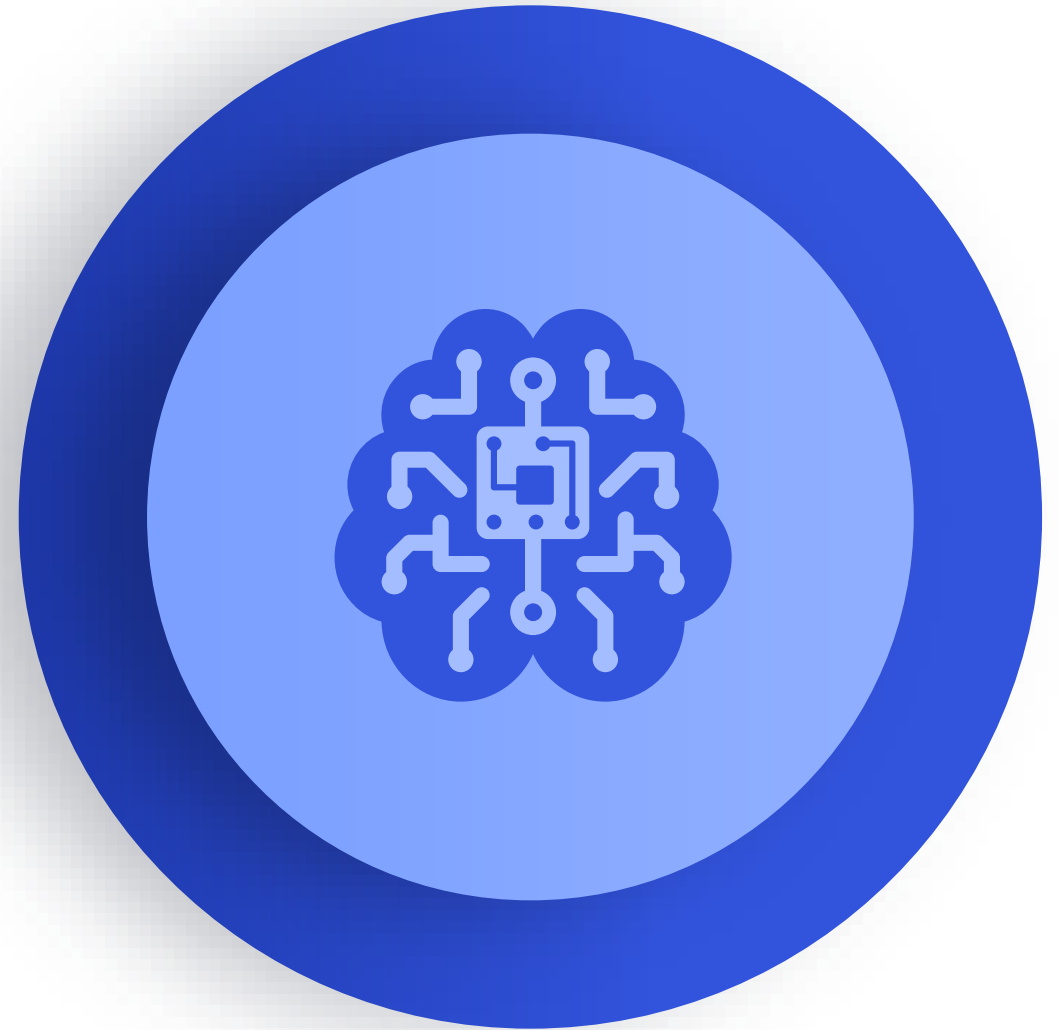


Qualcomm



# Qualcomm Technologies is accelerating on-device AI

Making efficient on-device machine learning possible for highly responsive, private, and intuitive user experiences



# Qualcomm® Artificial Intelligence Platform

The platform for efficient on-device machine learning

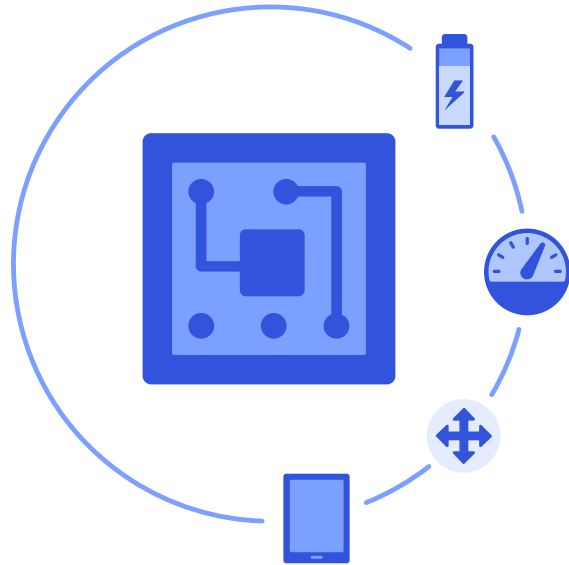
A high-performance platform designed to support myriad intelligent-on-device-capabilities that utilize:

- Qualcomm® Snapdragon™ mobile platform's heterogeneous compute capabilities within a highly integrated SoC
- Innovations in machine learning algorithms and enabling software
- Development frameworks to minimize the time and effort for integrating customer networks with our platform



# Making on-device intelligence pervasive

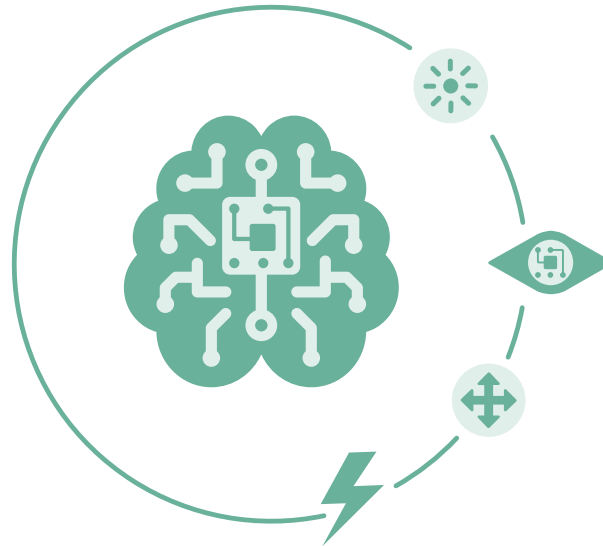
Focusing on high performance HW/SW and optimized network design



## Efficient hardware

Developing heterogeneous compute to run demanding neural networks at low power and within thermal limits

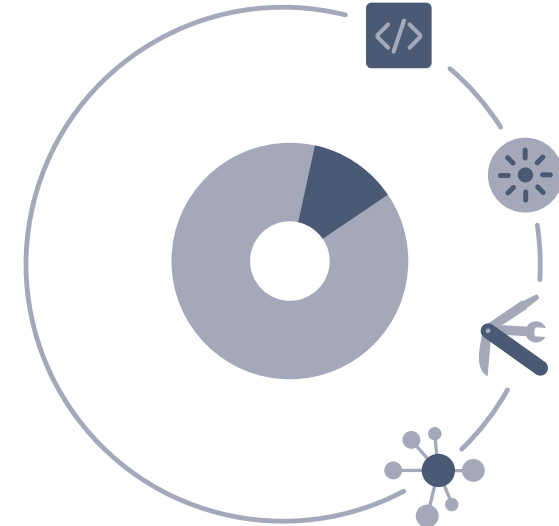
Selecting the right compute block for the right task



## Algorithmic advancements

Algorithmic research that benefits from state-of-the-art deep neural networks

Optimization for space and runtime efficiency



## Software tools

Software accelerated run-time for deep learning

SDK/development frameworks

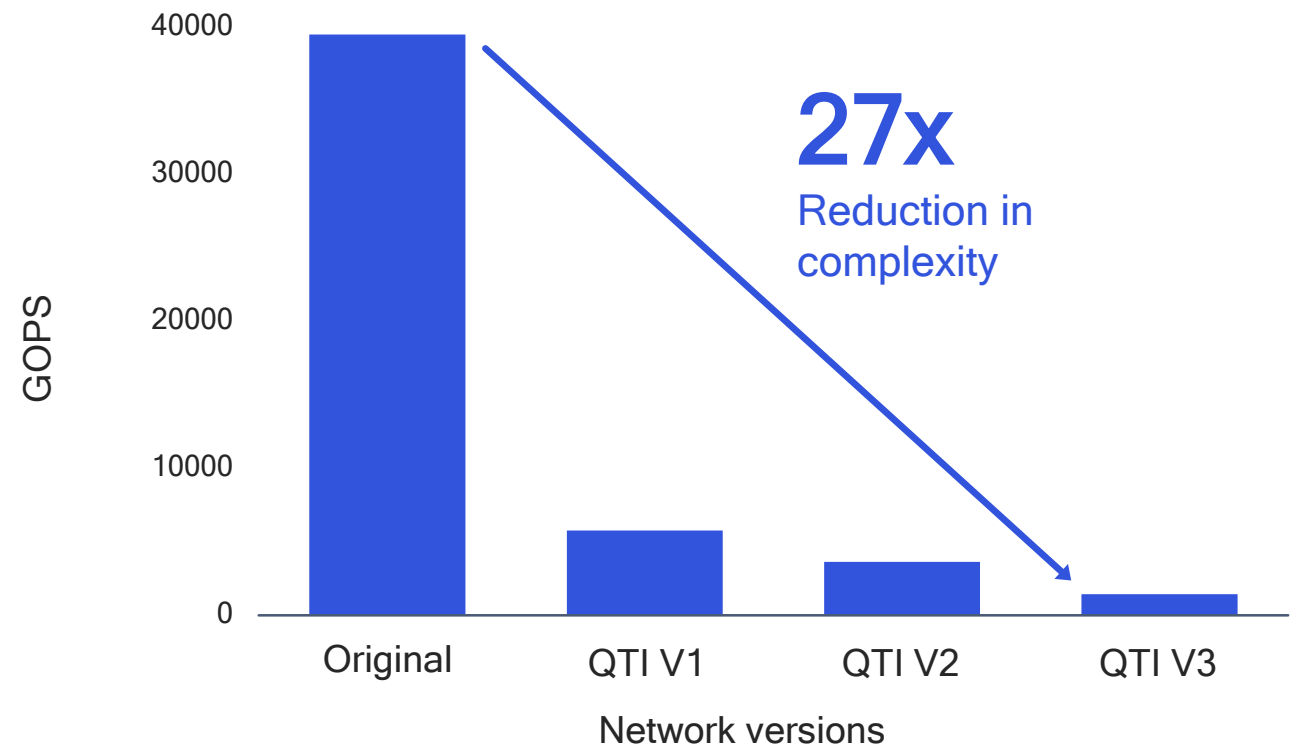
# Algorithmic enhancements for space and runtime efficiency

Improve performance by addressing model complexity

## Neural network optimizations for embedded

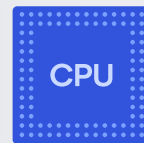
- Improved network architecture
- Focus on memory and storage
  - Reduce bit widths
  - Model compression
  - Leverage sparsity
- Architecture learning

## Required operations per image

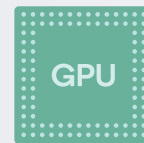


# Snapdragon Neural Processing SDK

Software accelerated runtime for the execution of deep neural networks on device



Qualcomm®  
Kryo™ CPU



Qualcomm®  
Adreno™ GPU



Qualcomm®  
Hexagon™ DSP

## Efficient execution on Snapdragon

- Takes advantage of Snapdragon heterogeneous computing capabilities
- Runtime and libraries accelerate deep neural net processing on all engines: CPU, GPU, and DSP with vector extensions



## Model framework/Network support

- Convolutional neural networks and LSTMs
- Support for Caffe/Caffe2, TensorFlow, and user/developer defined layers



Offline  
conversion tools



Analyze  
performance



Sample  
code

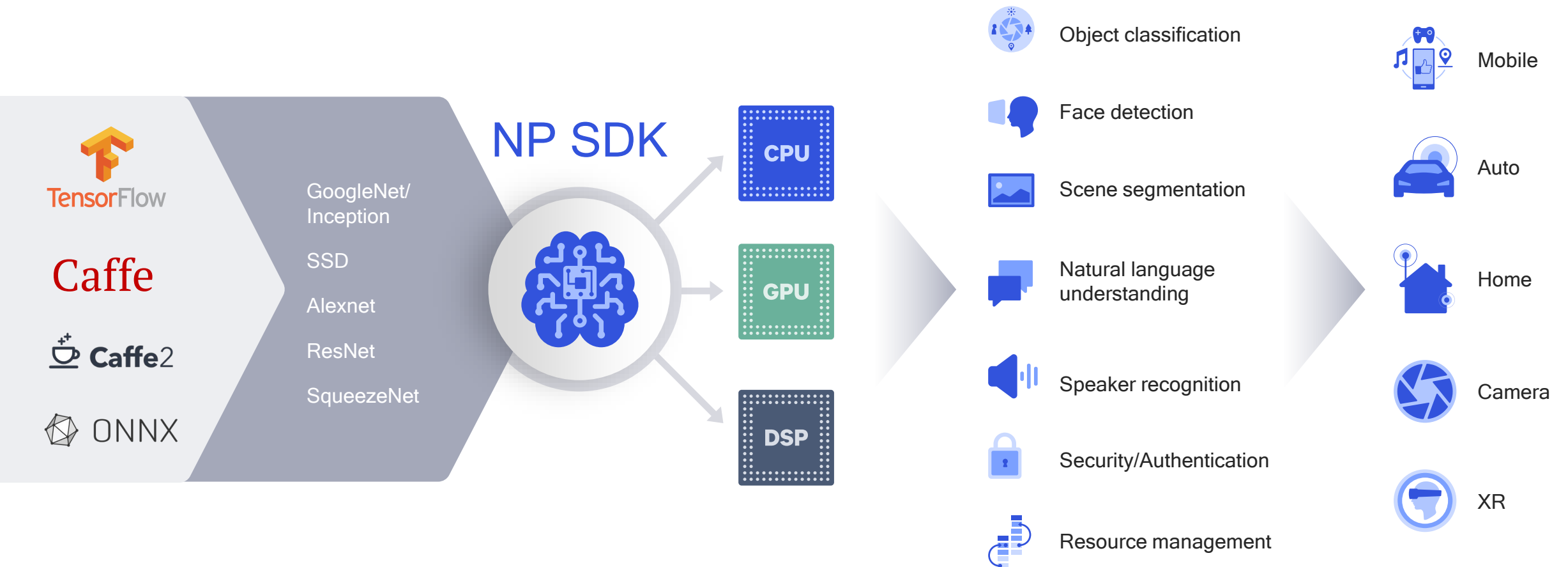


Ease of  
integration

## Optimization/Debugging tools

- Offline network conversion tools
- Debug and analyze network performance
- API and SDK documentation with sample code
- Ease of integration into customer applications

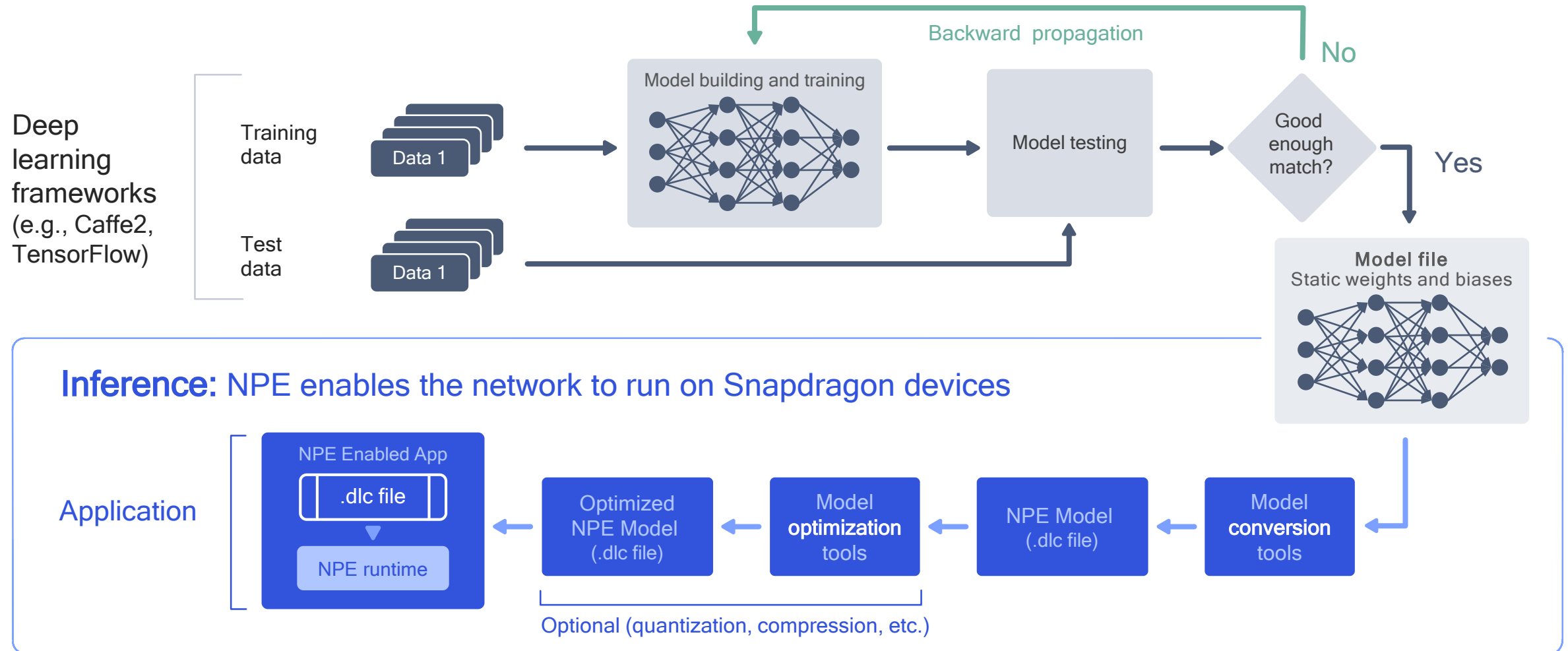
# Robust software and tools simplify development



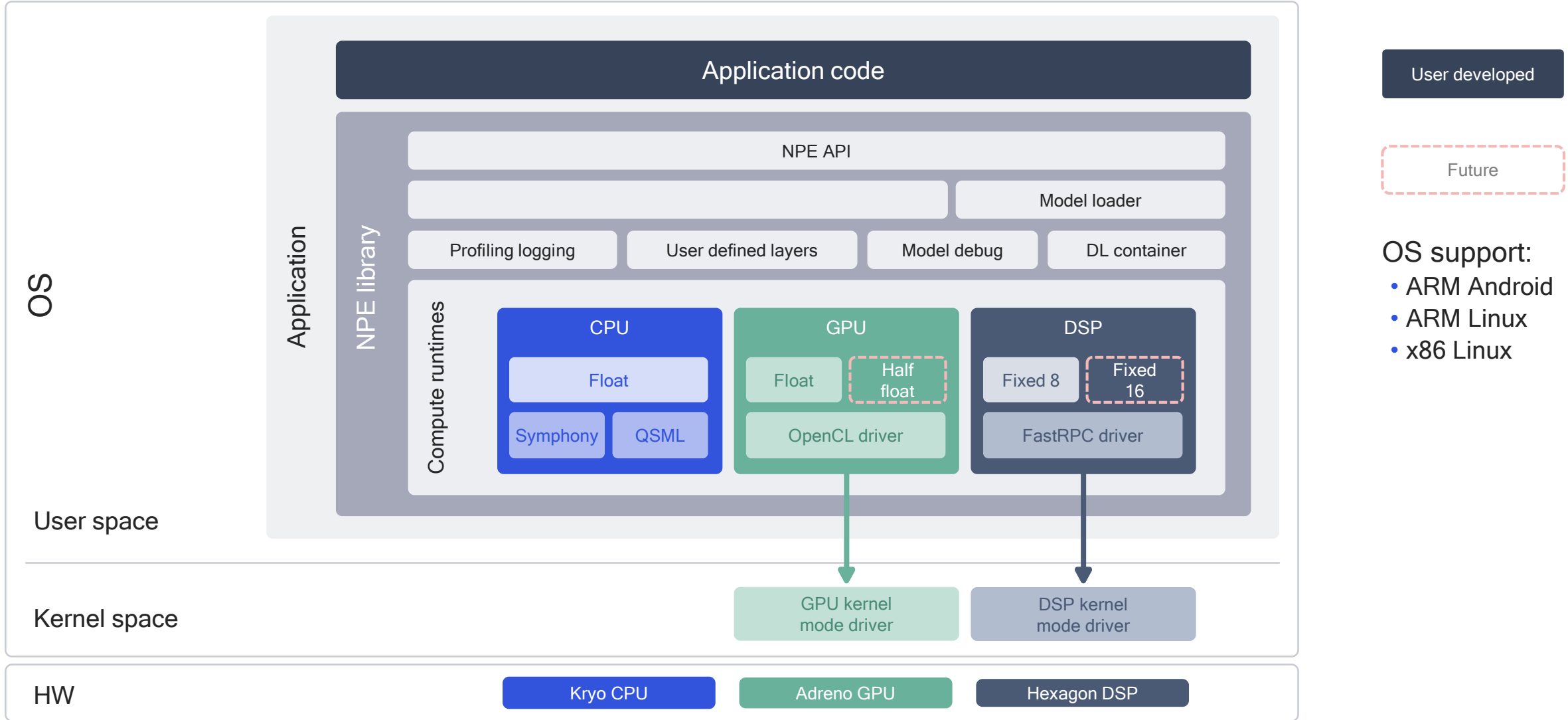


# Model to runtime workflow: Training and inference

Training: Machine learning experts build and train their network to solve their particular problem



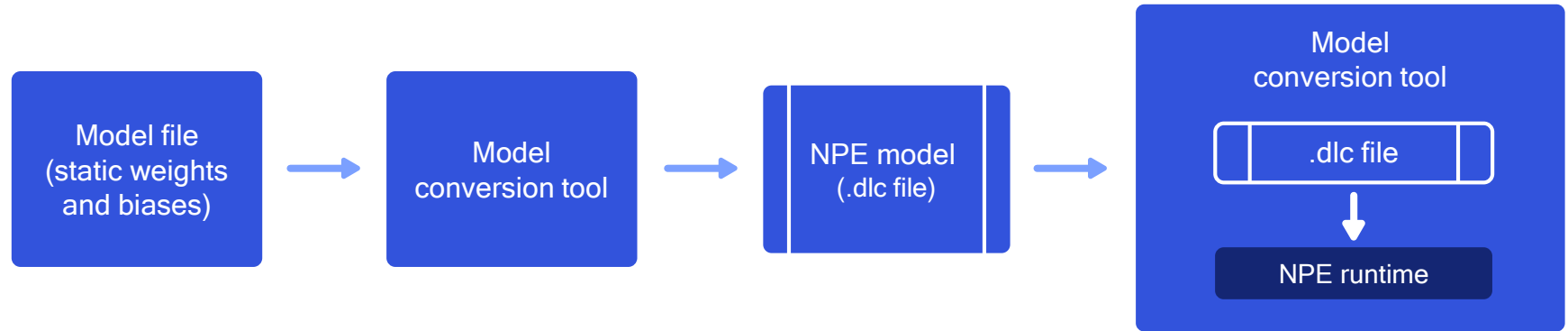
# High-level software architecture for the Snapdragon NPE



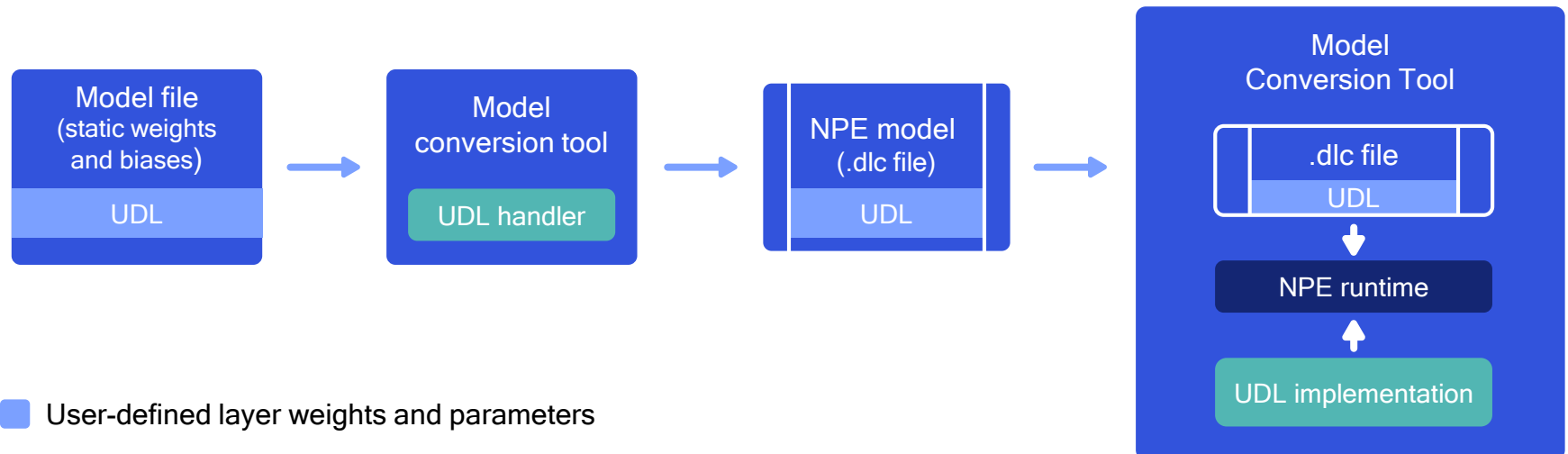
# User Defined Layer (UDL) workflow

Supports prototyping of layers not yet supported by the Snapdragon NPE

## NPE workflow

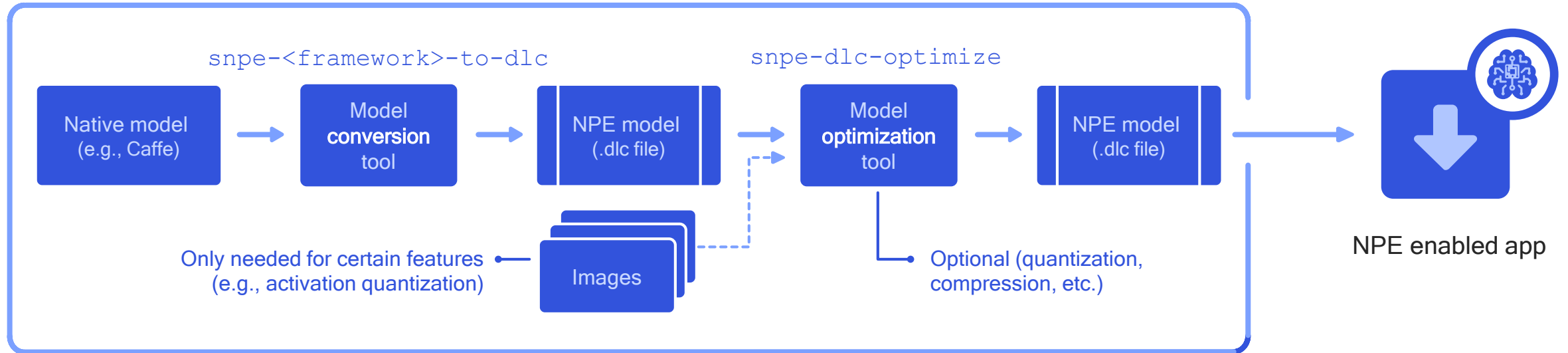


## NPE workflow with UDL



■ User-provided implementation    ■ User-defined layer weights and parameters

# Converting and quantizing a model



`snpe-<framework>-to-dlc`  
(`<framework>` = Caffe | Caffe2 | TensorFlow)

- Input is the model in native framework format
- Output is a converted but not optimized NPE DLC file

`snpe-dlc-optimize`

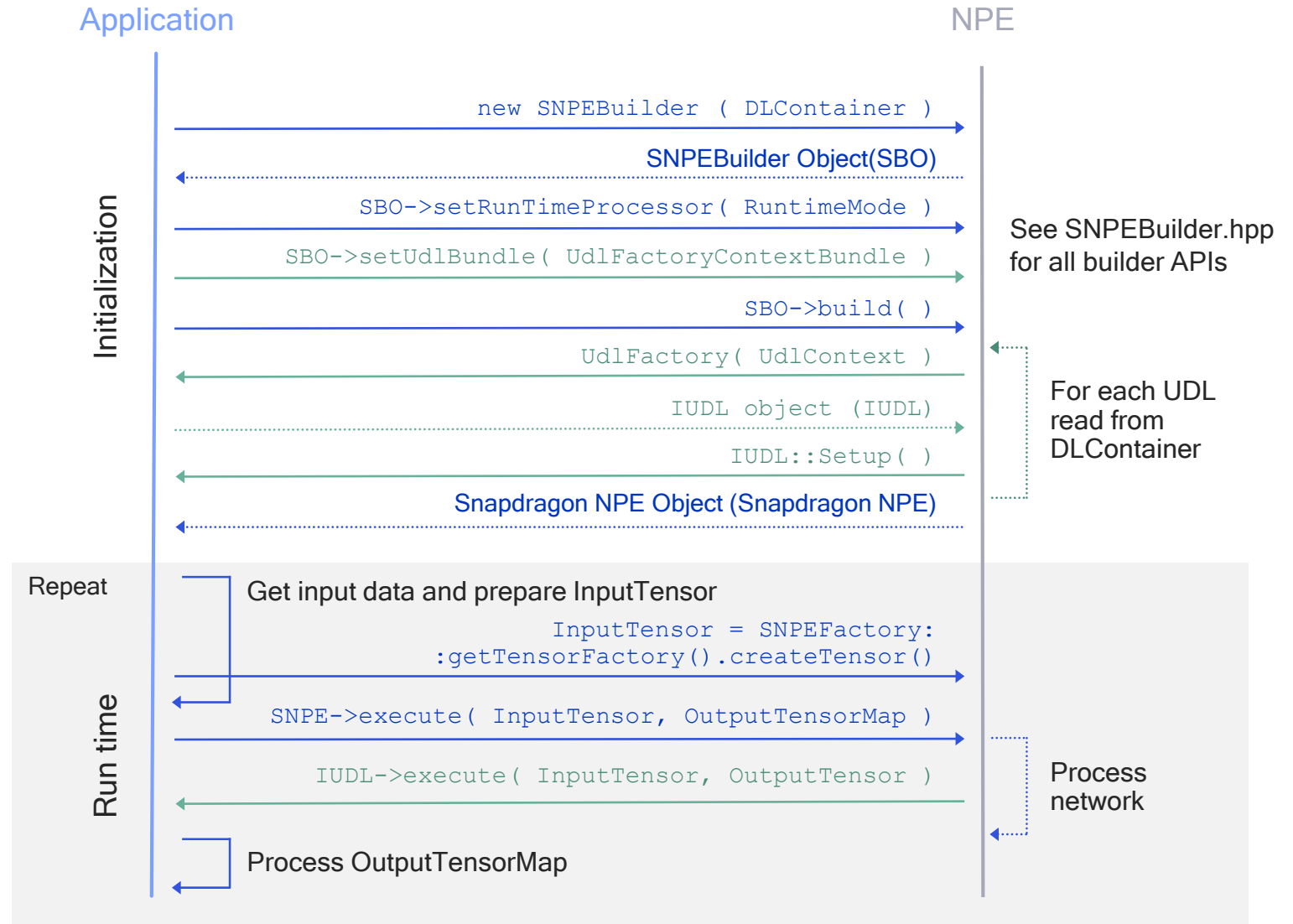
- Converts non-quantized DLC models into 8-bit quantized DLC models
  - Additionally implements further optimizations such as SVD compression
- Quantized model is necessary for fixed-point NPE runtimes (e.g. DSP)

# API example usage

NPE provides a simple C++ API with the following functionality

- Load a DLC model and select the runtime
- Execute the model
- Debug support
  - Dump the output of all layers in a model
- Collect performance metrics
  - Per-layer timing

Green API calls are only required for UDLs





100s

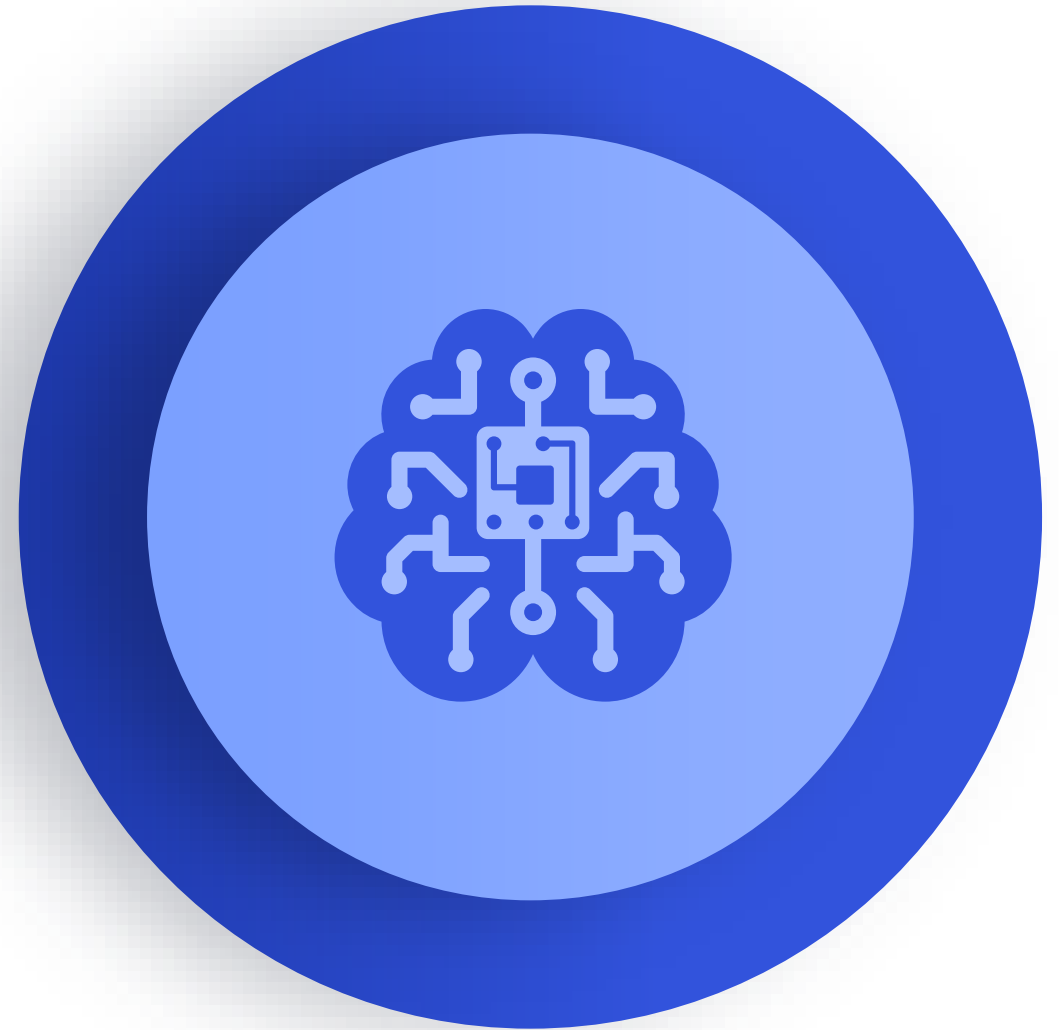
of Millions  
of units

AI + Qualcomm Technologies

**Massive scale**

Learn more at: [developer.qualcomm.com](https://developer.qualcomm.com)

# Qualcomm Technologies and Facebook collaboration for massive AI scale



# Facebook + Qualcomm Technologies

## On-device AI with Snapdragon

### Facebook and Qualcomm's Caffe2 collaboration

- Demonstrated Caffe2 acceleration with NPE at F8 2017
- 5x performance upside on GPU (compared to CPU)
- Announced commercial support of Caffe2 in July through Qualcomm Developer Network
- Facebook AML has integrated the NPE with Caffe2

### Future Caffe2/NPE research and development

- Continue to work closely with Facebook to optimize key networks for maximum on-device performance
- Enhancements to Caffe2 allowing Snapdragon specific SoC optimizations
- More advanced AI-powered XR applications



“On-device machine learning is made possible by the Qualcomm Snapdragon NPE which does the heavy lifting needed to run neural networks more efficiently on Snapdragon devices.”

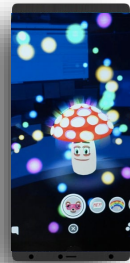
Source: XDA





# Enhancing the Facebook experience through on-device AI

More engaging social media with AI and AR



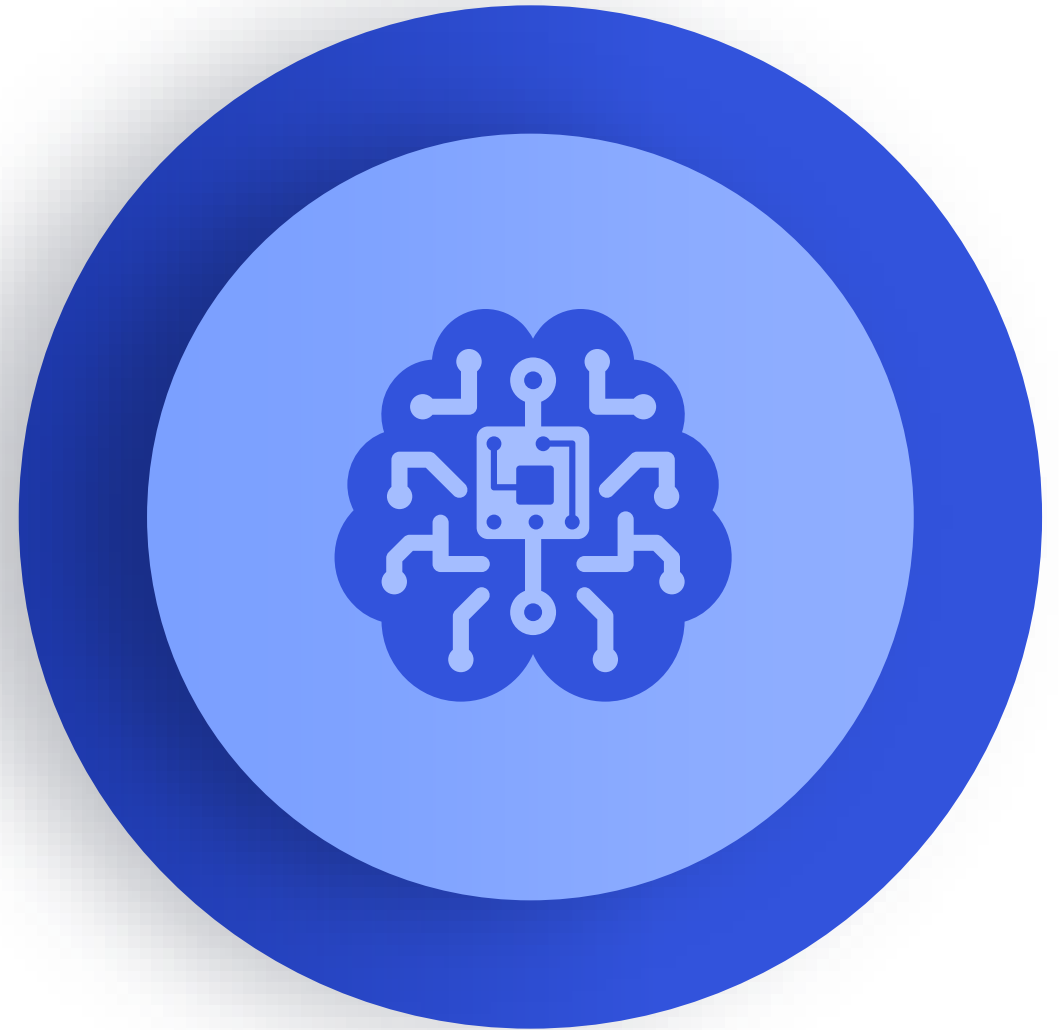
## Augmented reality features potentially powered by AI

- Style transfer and filters
- Frames and masks
- Photo and live videos, including 360°
- Contextual awareness (e.g. location/sensor metadata)

## On-device acceleration benefits

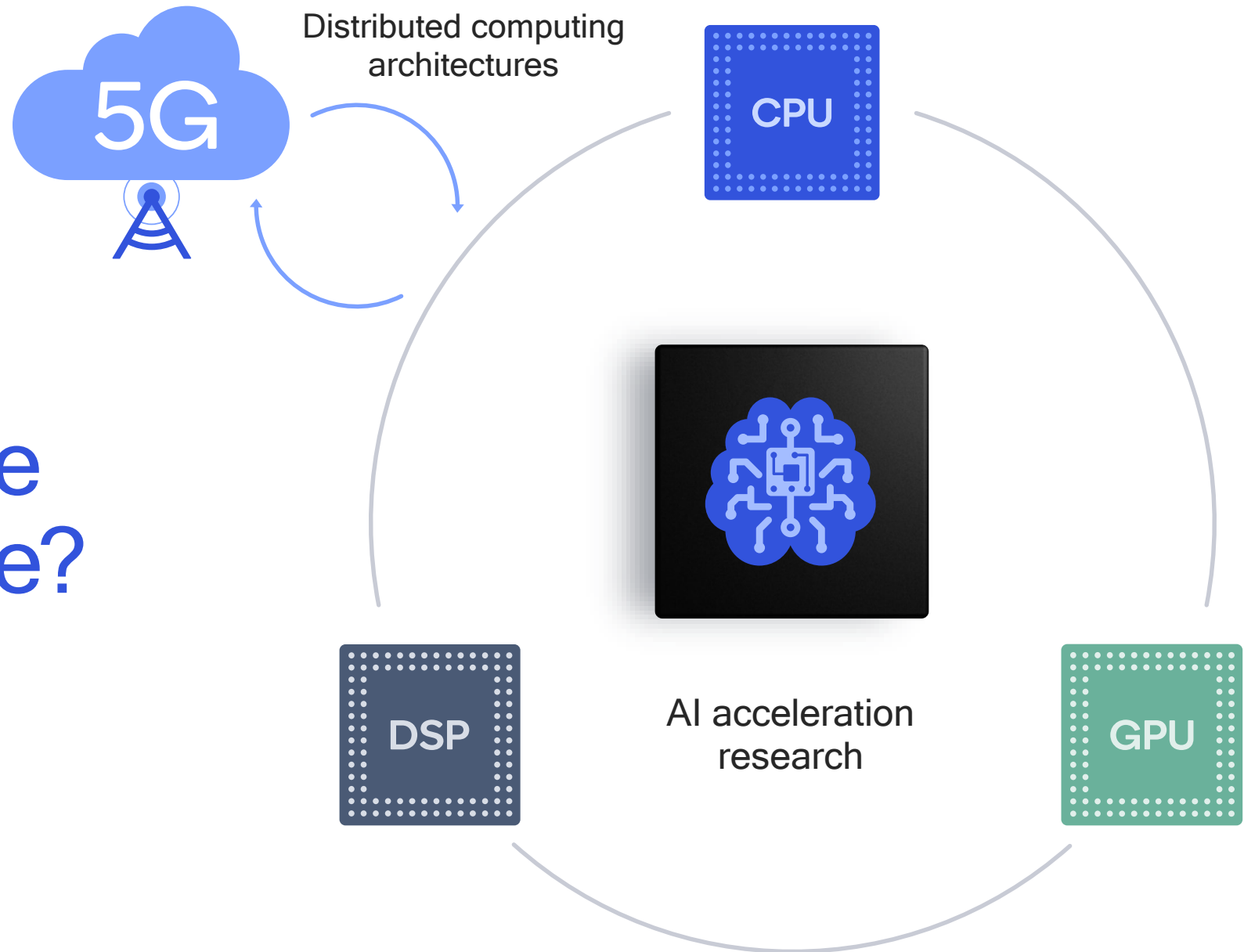
- Smooth UI with increased frame rate
- Increased battery life

What's next



AI hardware

# What does the future look like?



# AI offers enhanced experiences and new capabilities for smartphones

True personal assistance



Extended battery life



Enhanced connectivity



Qualcomm



Superior photography



Natural user interfaces



Enhanced security

A new development paradigm  
where things repeatedly improve

# AI will bring XR closer to the ultimate level of immersion

Creating physical presence in real or imagined worlds





# AI is revolutionizing the car of the future

## Redefining the in-car experience

- Natural user interfaces
- Personalization
- Driver awareness monitoring

## Paving the road to autonomy

- Surround view perception
- Sensor fusion
- Path planning
- Decision making

# What's next?

Specialized  
hardware



Algorithmic  
advancements






Improved  
optimization strategies





# Thank you!

Follow us on:   

For more information, visit us at:

[www.qualcomm.com](http://www.qualcomm.com) & [www.qualcomm.com/blog](http://www.qualcomm.com/blog)

Nothing in these materials is an offer to sell any of the components or devices referenced herein.

©2018 Qualcomm Technologies, Inc. and/or its affiliated companies. All Rights Reserved.

Qualcomm is a trademark of Qualcomm Incorporated, registered in the United States and other countries. Other products and brand names may be trademarks or registered trademarks of their respective owners.

References in this presentation to “Qualcomm” may mean Qualcomm Incorporated, Qualcomm Technologies, Inc., and/or other subsidiaries or business units within the Qualcomm corporate structure, as applicable. Qualcomm Incorporated includes Qualcomm’s licensing business, QTL, and the vast majority of its patent portfolio. Qualcomm Technologies, Inc., a wholly-owned subsidiary of Qualcomm Incorporated, operates, along with its subsidiaries, substantially all of Qualcomm’s engineering, research and development functions, and substantially all of its product and services businesses, including its semiconductor business, QCT.