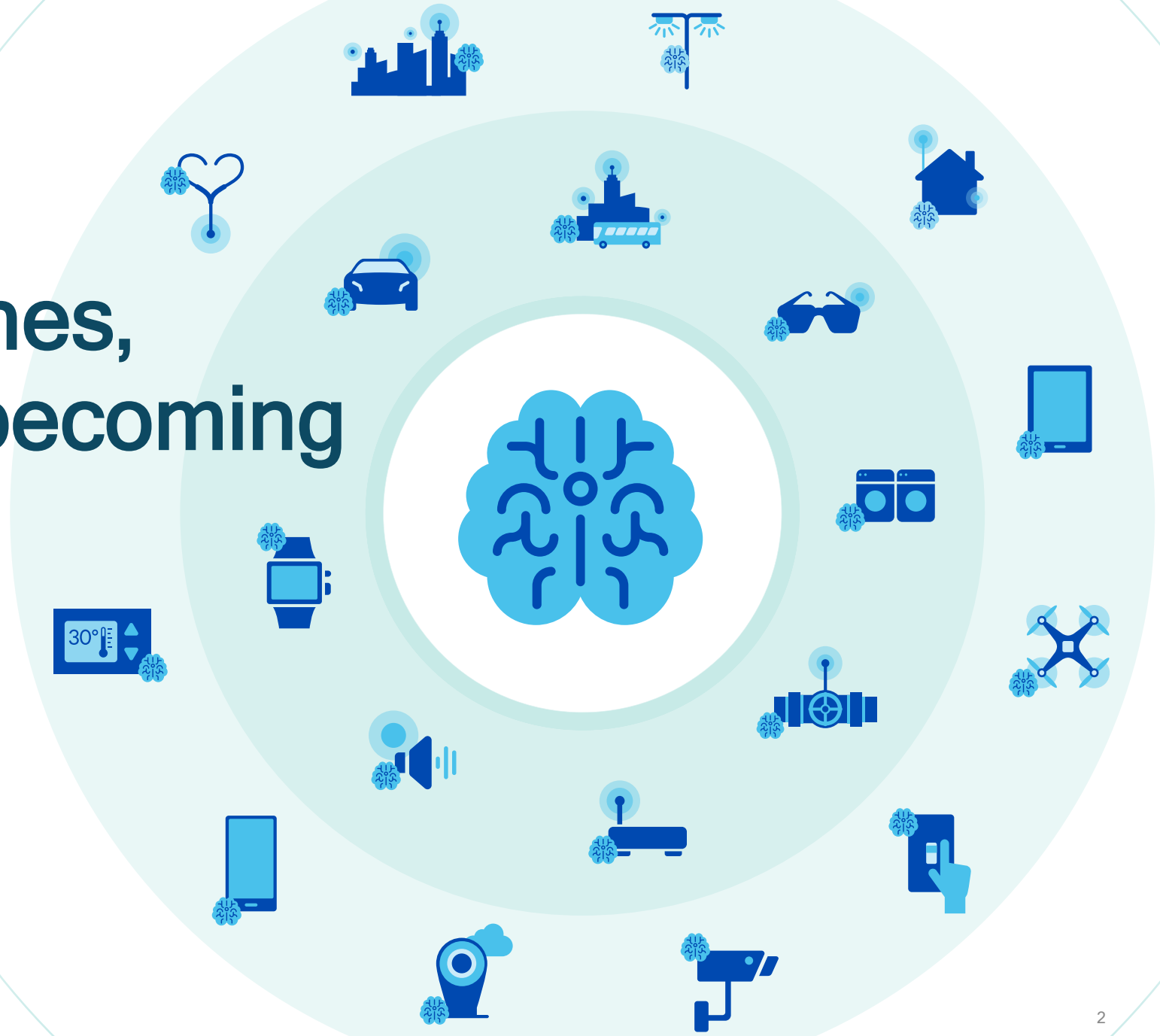
Mobile Innovation – The Key to AI @ The Edge

Gary Brotman
Director, AI Product Management
Qualcomm Technologies, Inc.
October 11, 2017

@GaryZero

Devices, machines, and things are becoming more intelligent



“Embedded and connected systems will reach annual volumes of 24 billion by 2021, with half of all the compute architectures shipping in 2021 supporting and processing AI.”

Source: IDC - Mapping the Future of Silicon for AI – Sept. 2017 (doc #US43117317)

"Heterogeneous architectures will be instrumental in enabling developers to program once and leverage the most efficient solution for AI workloads."

Source IDC – Mapping the Future of Silicon for AI – Sept. 2017 (doc #US43117317)

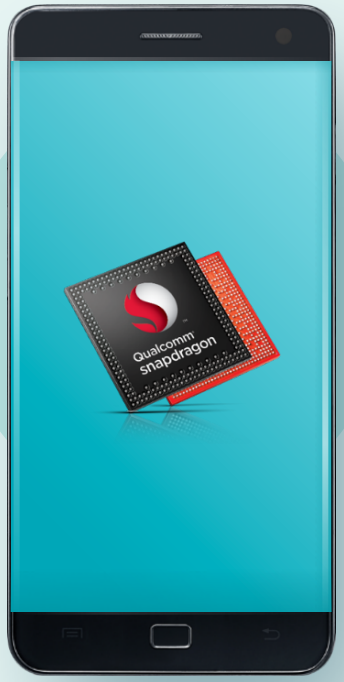
Mobile Drives Innovation at the Edge

Mobile computing

Smart homes

Industrial IoT

Wearables



Smart cities

Automotive

Healthcare

Networking

Extended reality

Rapid replacement cycles

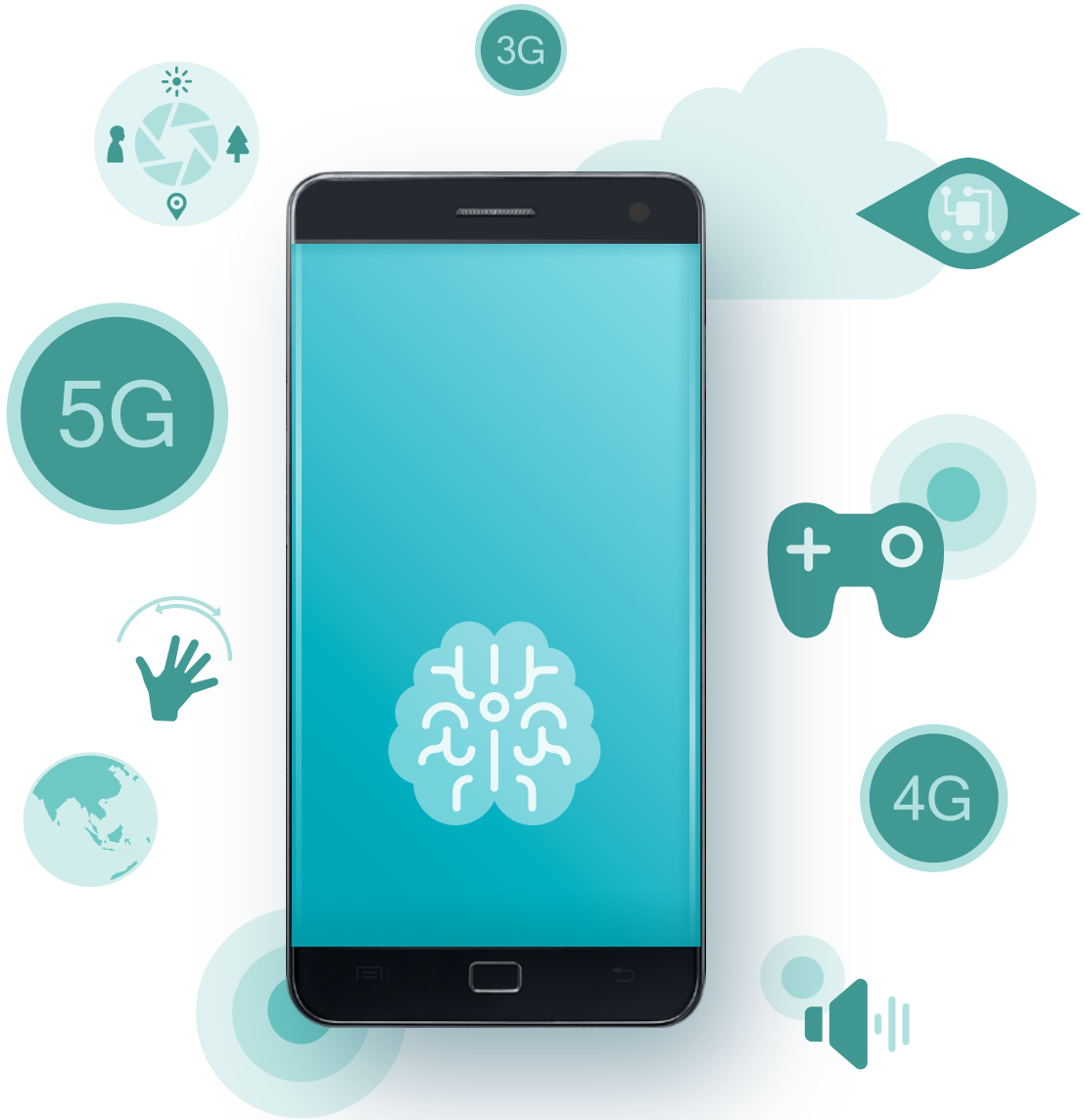
Superior scale

Integrated and optimized technologies

Mobile - THE most pervasive AI platform

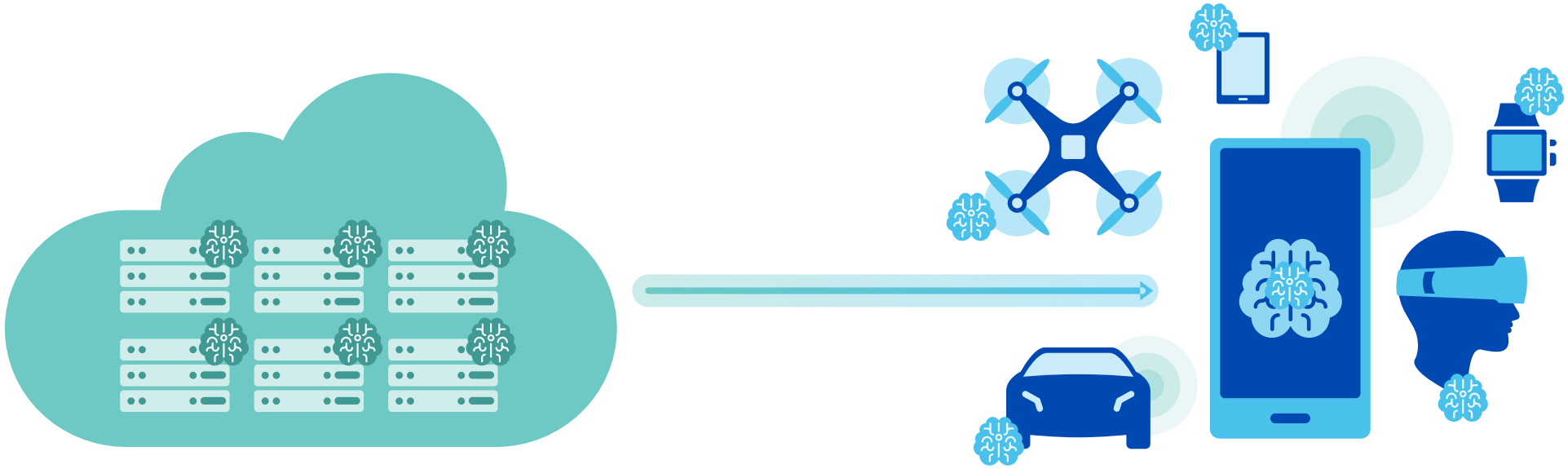
>8.5 Billion

Cumulative smartphone unit shipments forecast between 2017-2021



Source: Gartner Mar. ‘17

Intelligence is moving to the device

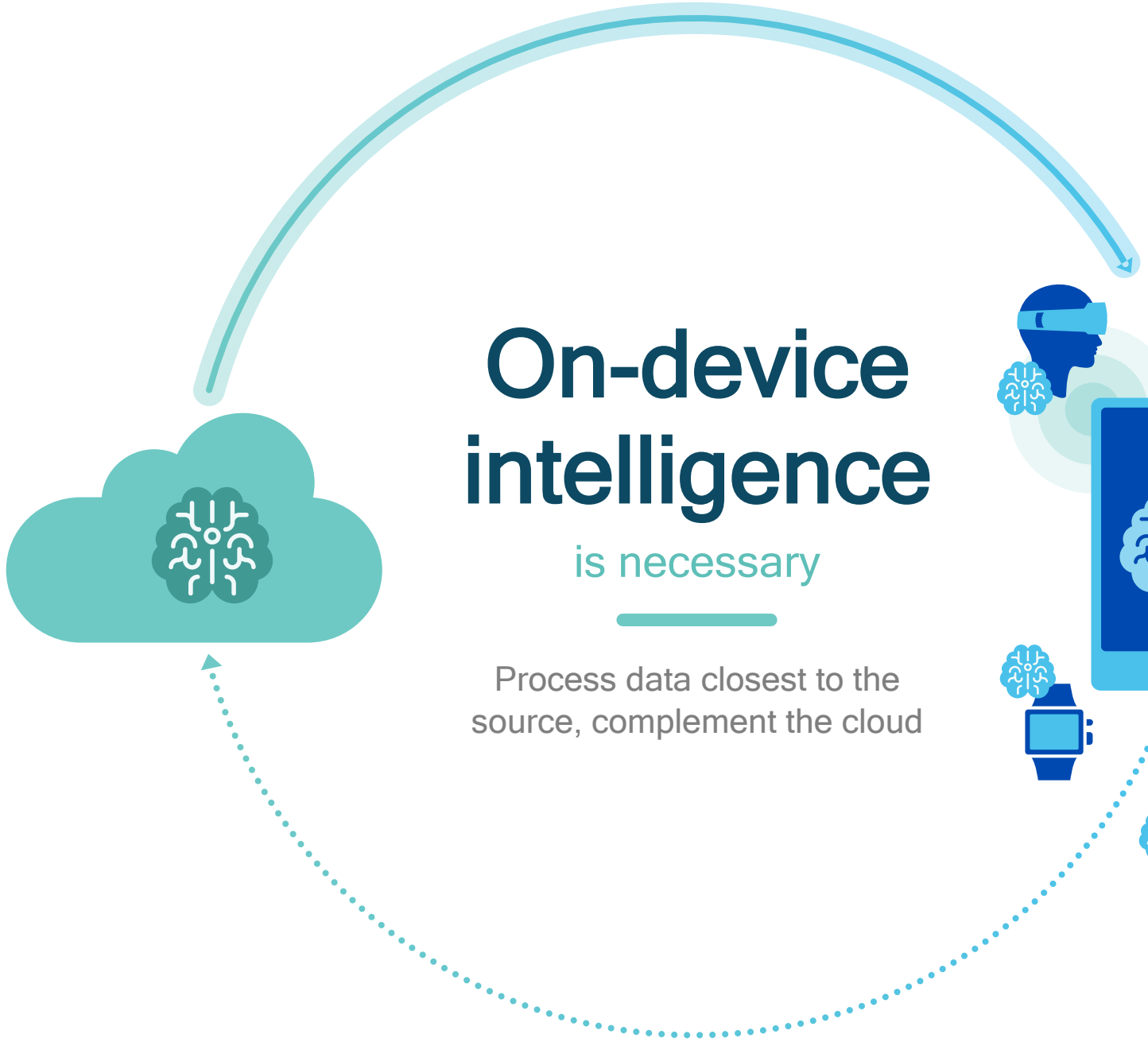


Server/Cloud

Training
Execution/Inference

Devices

Execution/Inference
Training (emerging)



On-device intelligence

is necessary

Process data closest to the source, complement the cloud

Low latency

Privacy

Reliability

Power and thermal efficiency are essential for on-device AI

The challenge of AI workloads

- Very compute intensive
- Large, complicated neural network models
- Complex concurrencies
- Always-on
- Real-time

Constrained environments

- Must be thermally efficient for sleek, ultra-light designs
- Requires long battery life for all-day use
- Storage / Memory bandwidth limitations

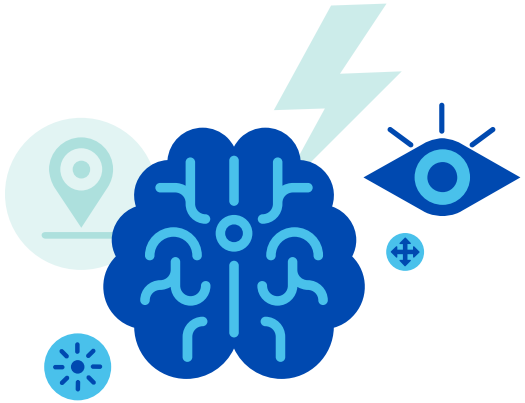
Making on-device intelligence a reality

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

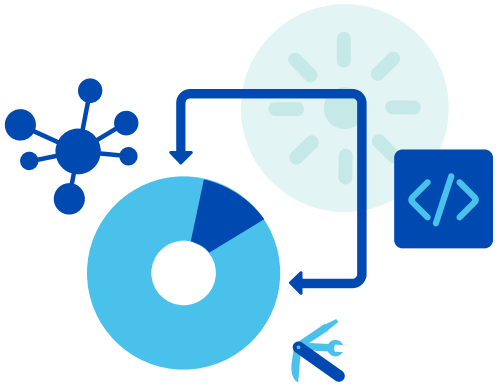
Selecting the right compute block for the right task



Algorithmic advancements

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

Optimization for space and runtime efficiency



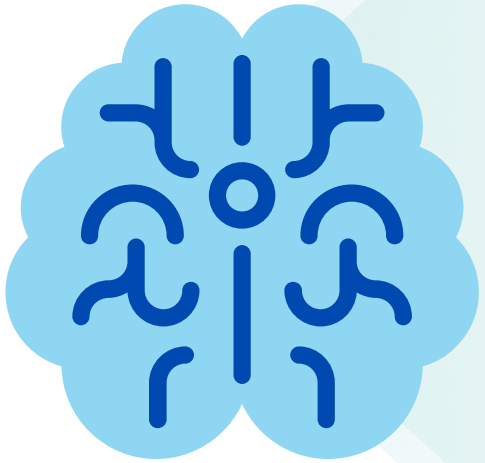
Software tools

Software accelerated run-time for deep learning

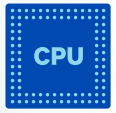
SDK/development frameworks

Snapdragon

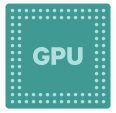
Neural Processing Engine



Available at: developer.qualcomm.com



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 HVX

Model framework/network support

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

Offline conversion tools

Analyze performance

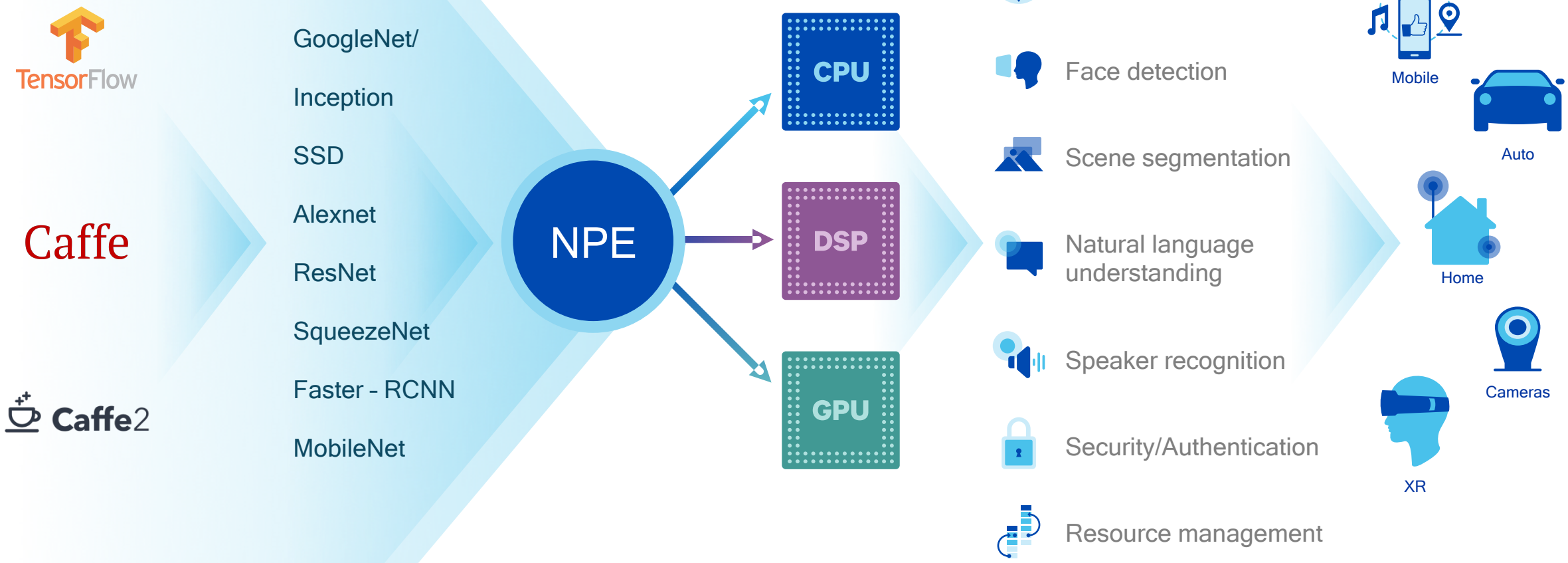
Sample Code

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

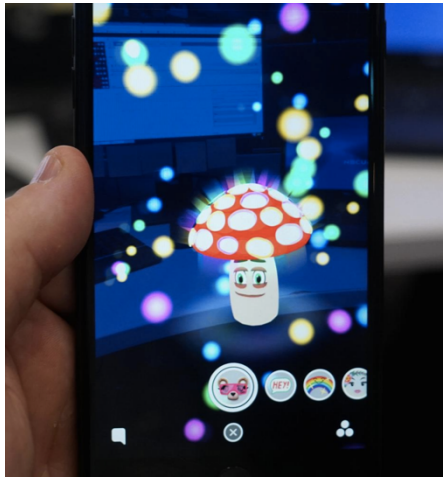
Software accelerated runtime for the execution of deep neural networks on device

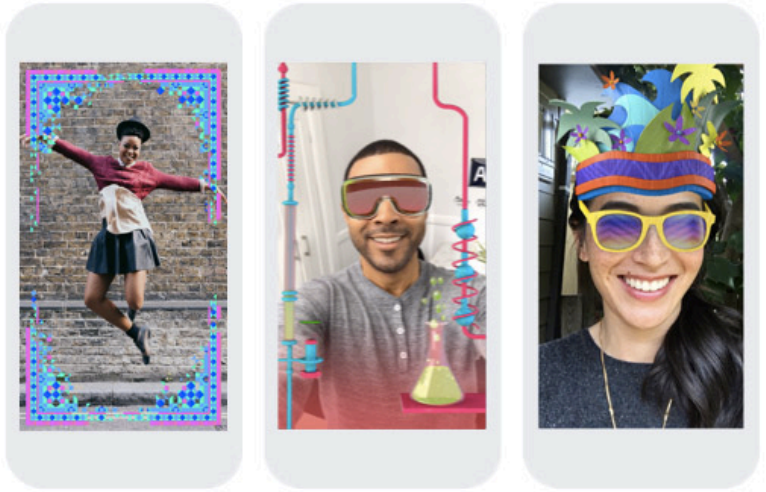

TensorFlow
Caffe
Caffe2
GoogleNet/
Inception
SSD
Alexnet
ResNet
SqueezeNet
Faster - RCNN
MobileNet
NPE
CPU
DSP
GPU
Object classification
Face detection
Scene segmentation
Natural language understanding
Speaker recognition
Security/Authentication
Resource management
Mobile
Auto
Home
Cameras
XR

Social Media Apps: Camera + AR + AI = Consumer Delight

AI workloads moving from cloud to the edge

- Leading social media apps using AR & AI tech in abundance
  - Camera at the center – key to consumer delight and communication
  - Augmented reality features powered by AI
    - Style Transfer & Filters
    - Frames & Masks
    - Photo and live videos, including 360
    - Contextual awareness (e.g. location/sensor metadata)
- Progression of high level tasks
  - Detection, Perception, Understanding, Prediction
- Why on device?
  - Lower latency, privacy, connection independent





AI will bring XR closer to total immersion

Creating physical presence in real or imagined worlds

Visuals
Rendering techniques

Battery efficiency
Managing workload concurrency

Interactions
Natural UI
Depth estimation

Sounds
Audio Enhancement and ID

AI will make the car of the future possible

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

On-device AI Use Cases

Style Transfer

Speaker Recognition

AR/VR

Landmark Detection

Text Recognition

Noise Suppression

Object Location

Object Tracking & Avoidance

Video Summarization

Security & Liveness

ADAS/ Autonomous

Spatial Audio Detection

Computational Photography

Face Detection & Recognition

Natural Language Understanding

Diagnostics & Power

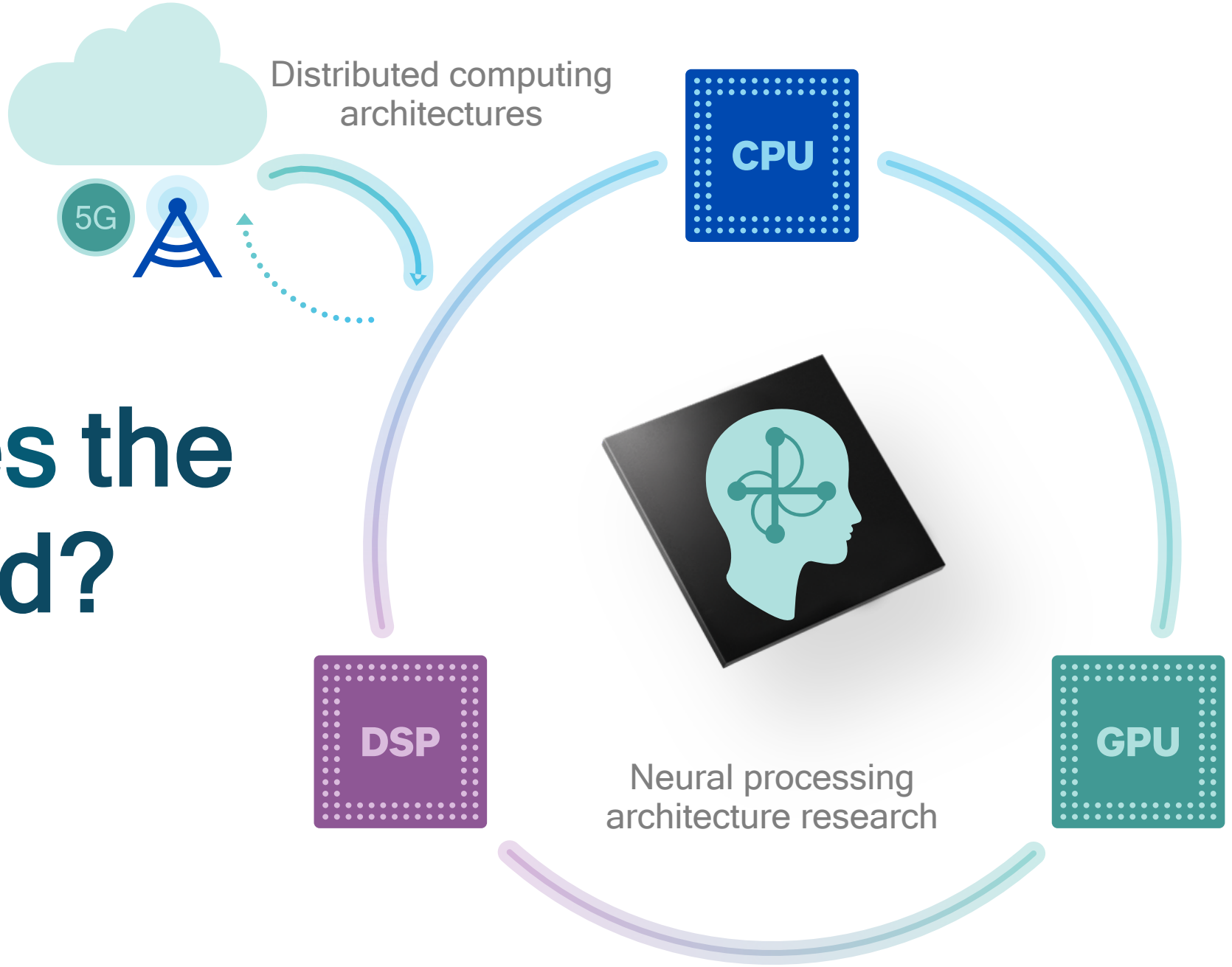
Gestures/Hand Tracking

Language Translation

Voice/Audio Keyword

Scene Classification

What does the future hold?



Thank you

Follow us on:

For more information, visit us at:
www.qualcomm.com & www.qualcomm.com/blog