
Qualcomm
April 9, 2019
San Francisco, CA
@cristianoamon
Advancing the AI future
Cristiano Amon
President
Qualcomm Incorporated


A world where virtually
everyone and everything
is intelligently connected

Connectivity

2G

Carrier Aggregation

Location

Multimode, multiband transceivers

Envelope tracking

3G

Antenna tuning

5G

LAA

4x4 MIMO

4G

RF Front End

Processing

Sensor processing

Video processing

Machine learning

Computer vision

CPU

GPU

DSP

Audio processing

Display processing

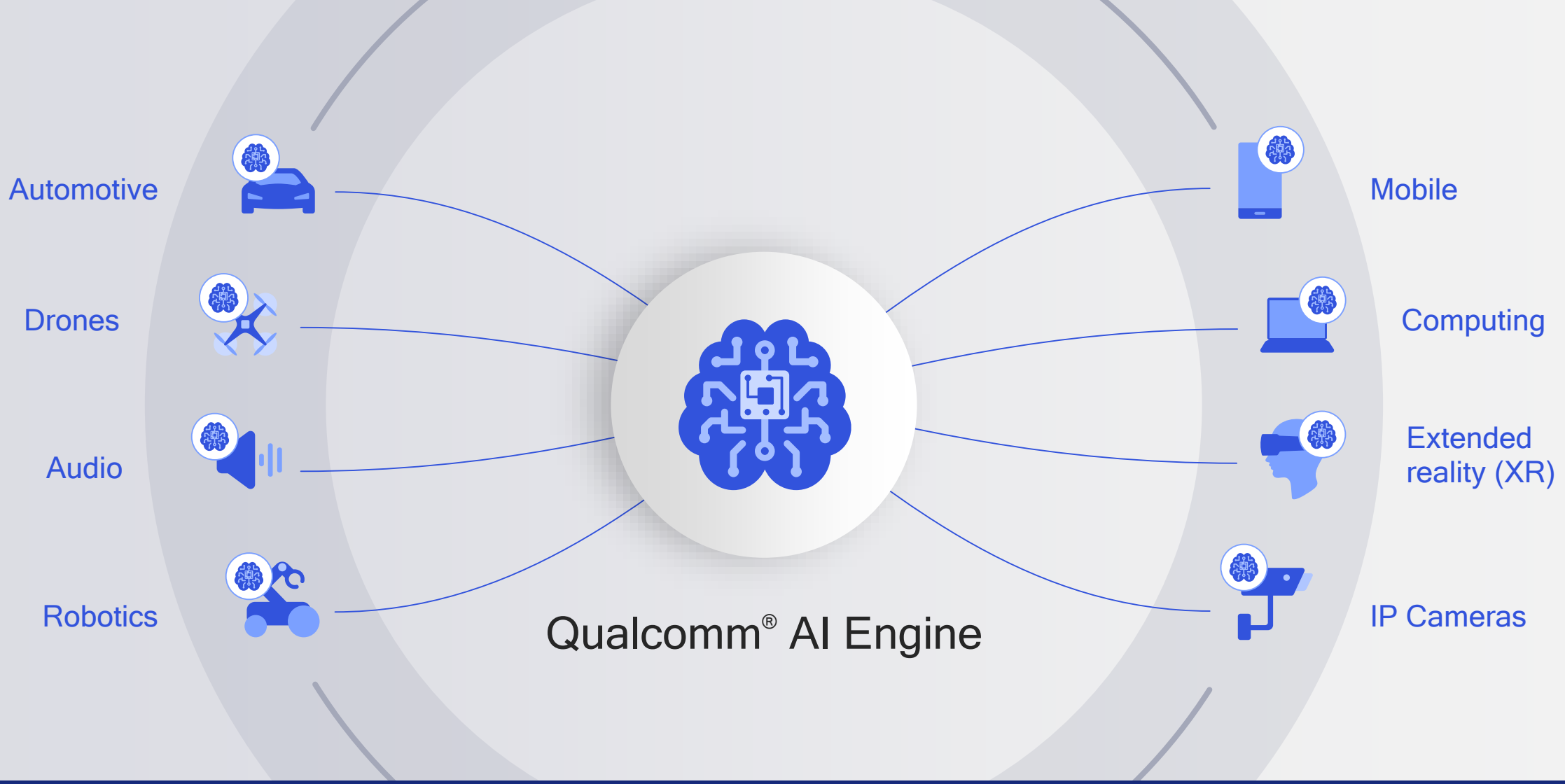
Camera processing

Power management

Language processing

We pioneered the technologies powering this future

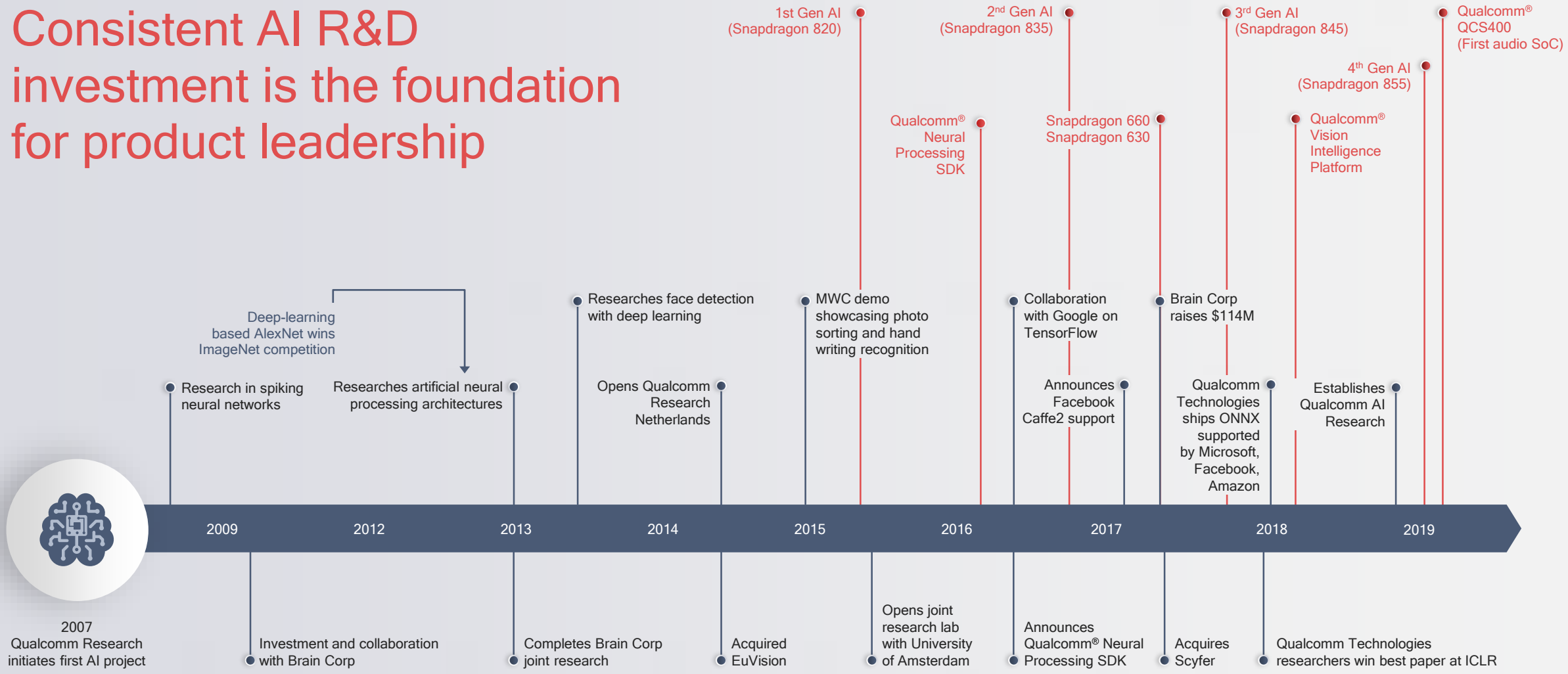
Leadership across advanced connectivity, processing, and systems design



Pioneering on-device intelligence

Qualcomm Artificial Intelligence Research

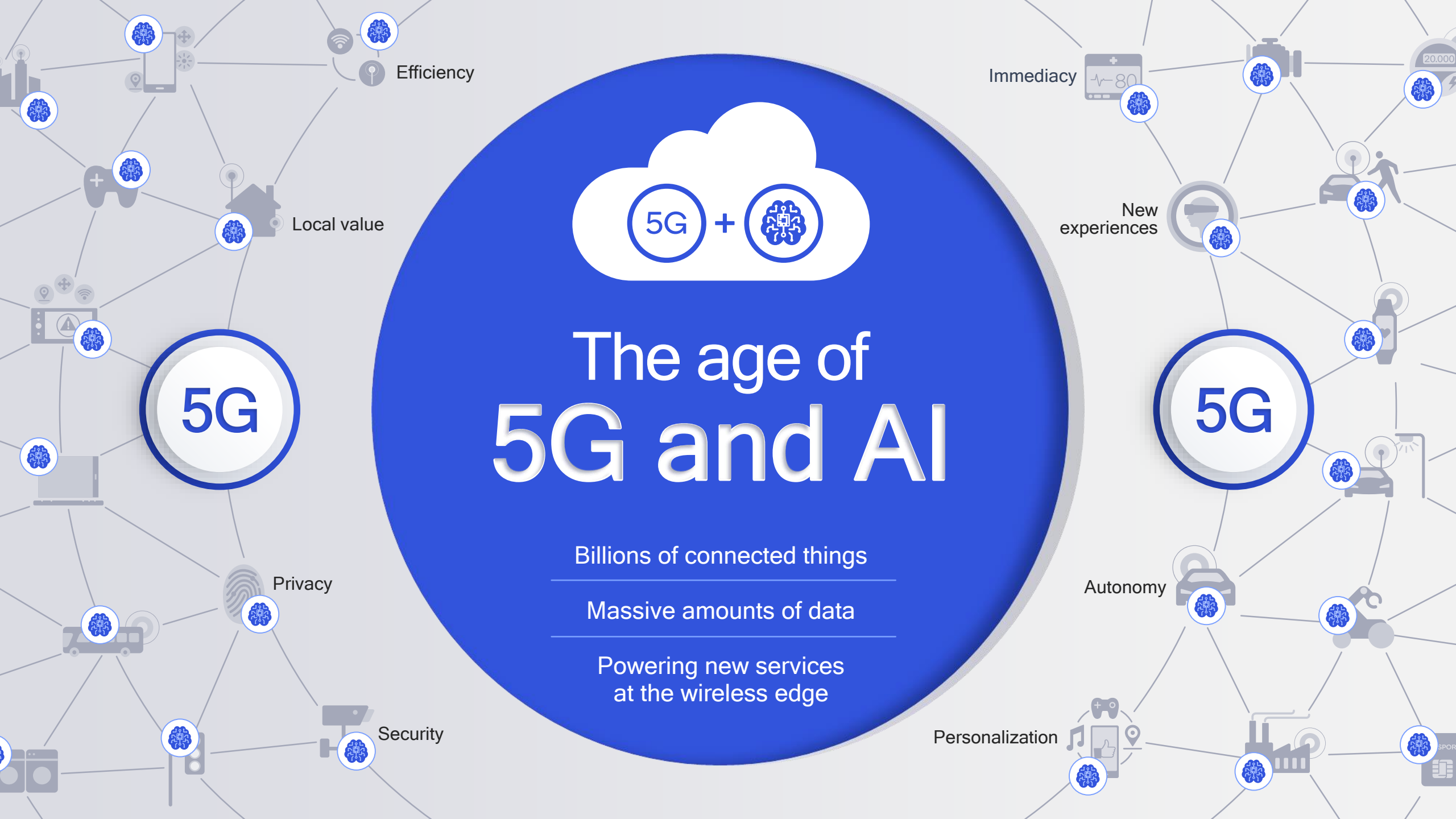
Consistent AI R&D investment is the foundation for product leadership



Collaborating with leading cloud players and ISVs

Cloud

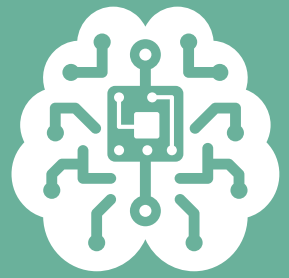
Software


Efficiency
Local value
5G
Privacy
Security
Immediacy
New experiences
5G
Autonomy
Personalization
5G + AI
The age of
5G and AI
Billions of connected things
Massive amounts of data
Powering new services at the wireless edge

$12.3 Trillion

In goods and services enabled by 5G in 2035

Source: The 5G Economy, an independent study from IHS Markit, Penn Schoen Berland and Berkeley Research Group, commissioned by Qualcomm.



$3.9 Trillion

In business value by 2022

Source: Gartner, Inc., *Gartner Says Global Artificial Intelligence Business Value to Reach $1.2 Trillion in 2018*, April 25, 2018.

Driving significant economic impact

5G rollout happening faster than 4G

Source: IHS Report Jan '19, Qualcomm Technologies data

Year 1 announcements underscore tremendous momentum with 5G

On-device intelligence is quickly gaining momentum

Key segments are expected to see full AI attach rates by 2025

10%

AI attach rate

2018

100%

AI attach rate

2025

Mobile

Automotive

XR

PCs / Tablets

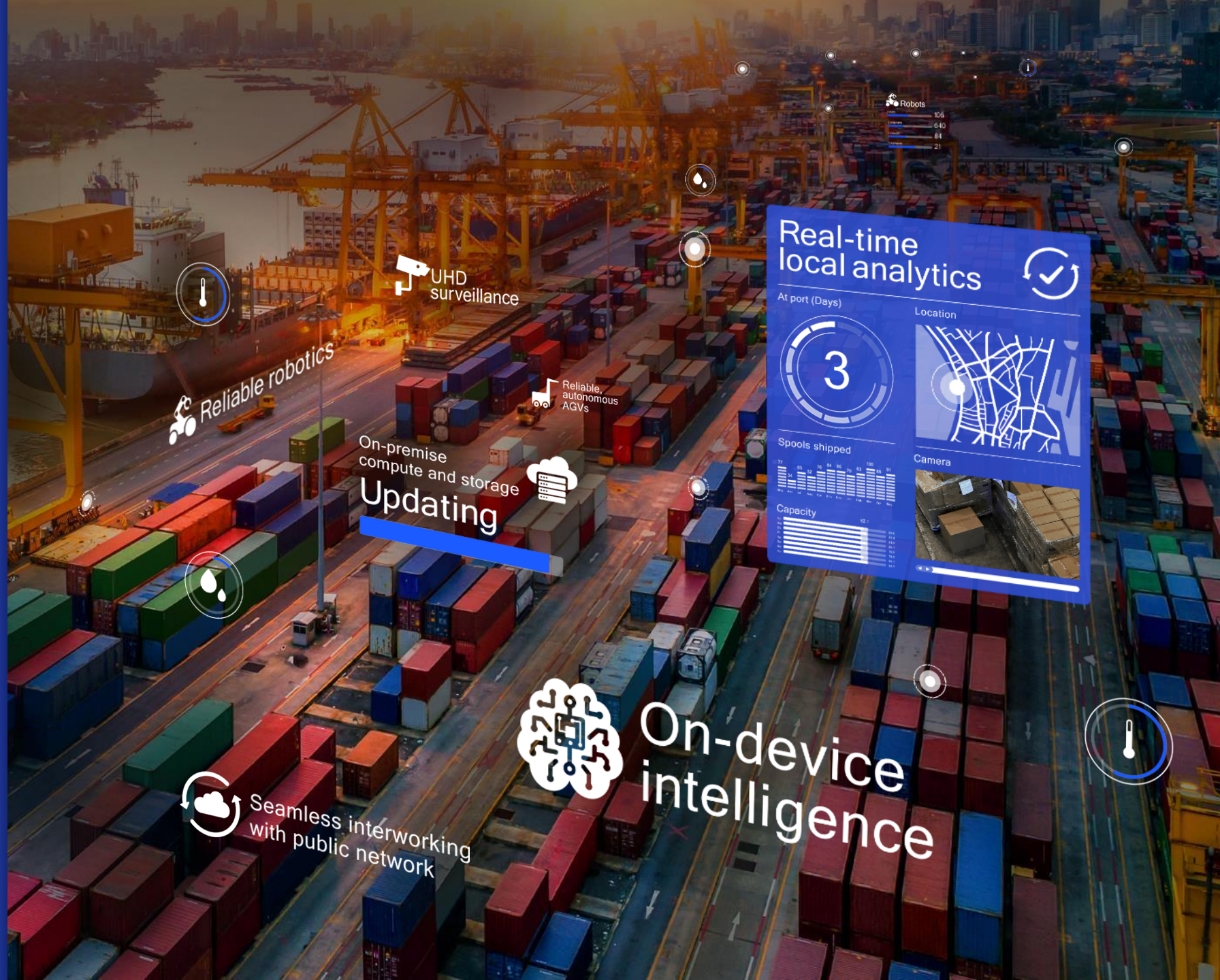
Smart speakers

Source: Tractica, 2019

Driving distributed intelligence

Bringing the cloud closer to devices at the edge

AI will drive transformation across industries



Shaping the future of transportation

Personalized driver settings

Driver awareness monitoring

Greater autonomous capabilities

Fueling a new era of cloud gaming

The cloud is becoming the new console

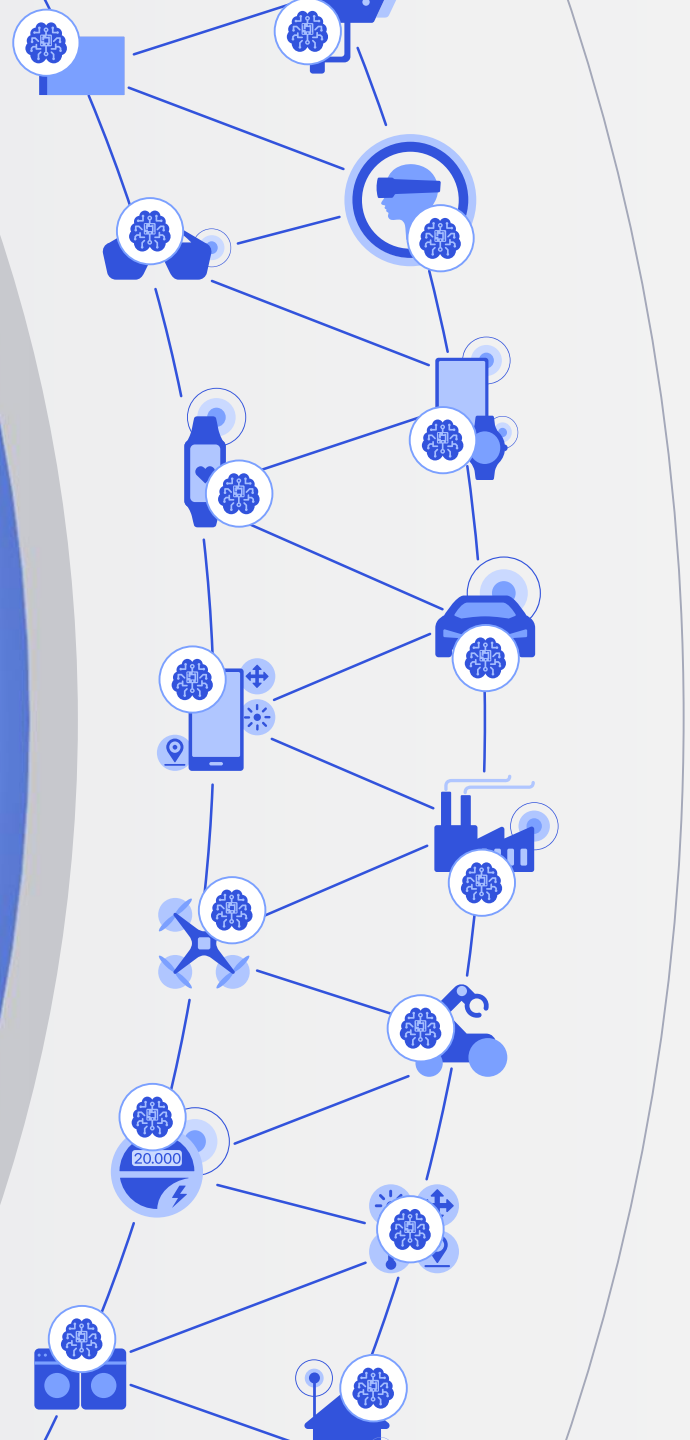
Boundless mobile XR experiences

Powering the factory of the future

The wireless edge realizes the full potential of 5G and AI

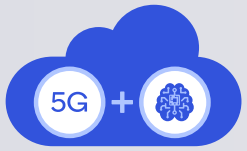
Inventing technology at scale to realize the promise of AI on trillions of connected devices

5G



Qualcomm

Foundational R&D



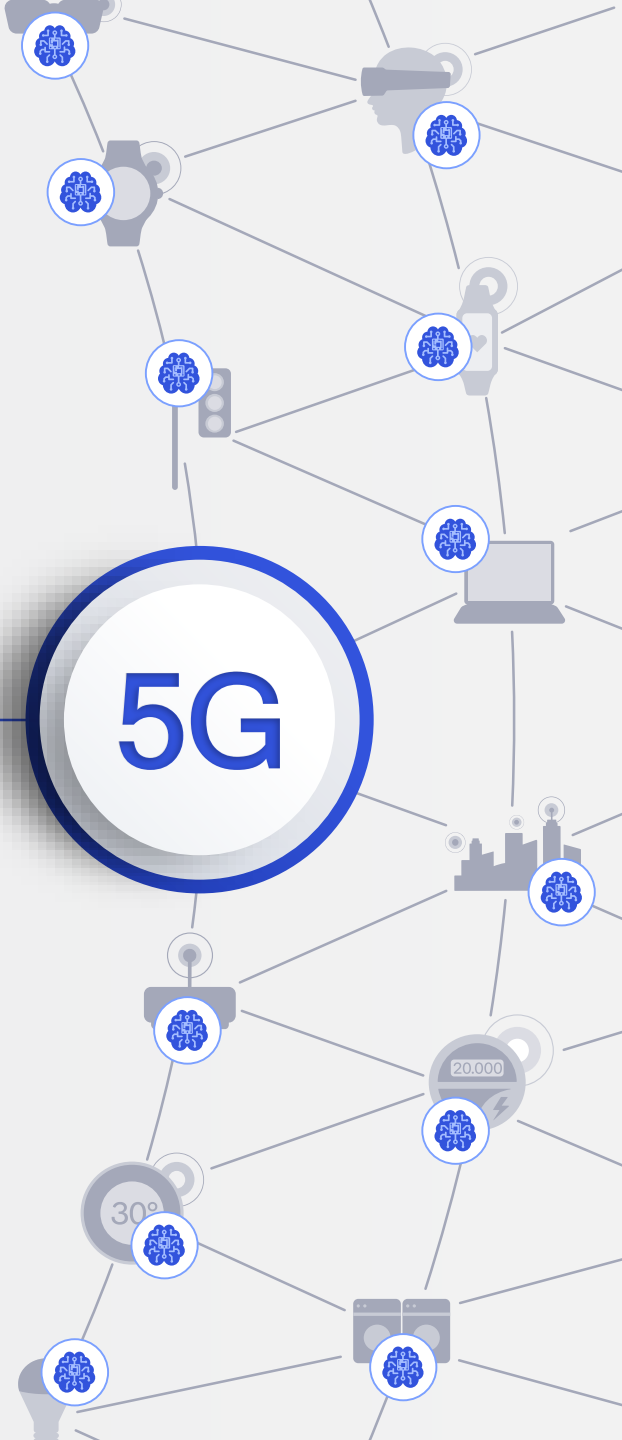
5G + AI technology leadership

Systems design expertise

with AI Engine

Advanced silicon

Ecosystem investment



Uniquely positioned to power the intelligently connected future

April 9, 2019 @qualcomm San Francisco, CA

Qualcomm

Bringing leading performance per watt to the cloud

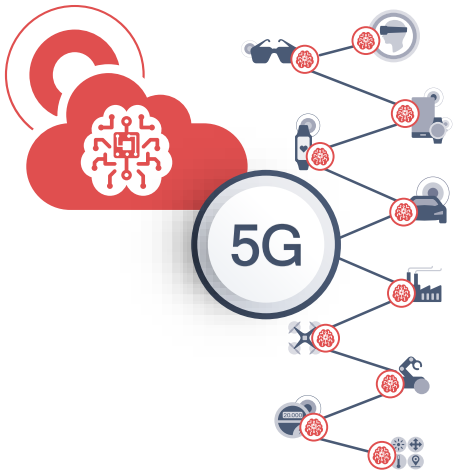
Keith Kressin

SVP, Product Management
Qualcomm Technologies, Inc.

Client AI

Bringing Artificial Intelligence to the Client Edge

5G + AI leadership



Leader in wireless edge development

Qualcomm snapdragon

with AI Engine

10+ years of AI research

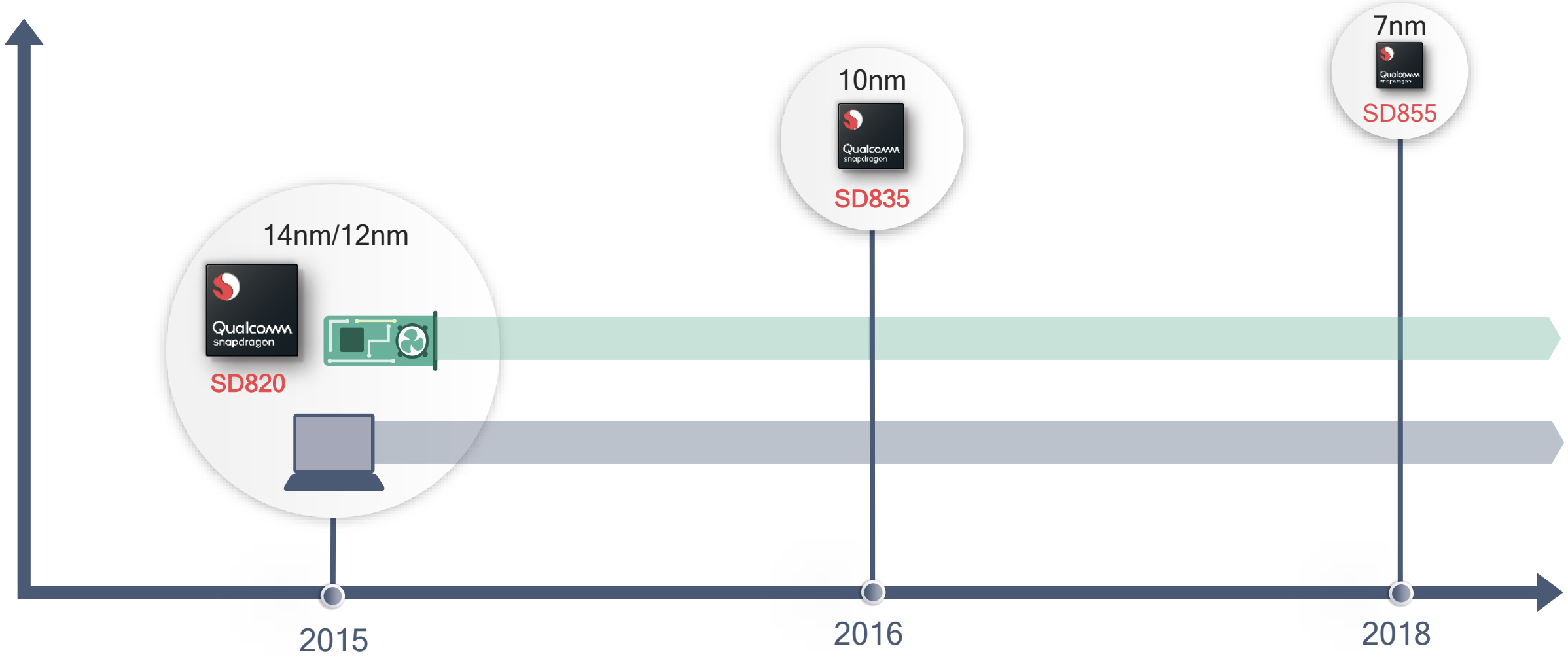
amazon alexa · Baidu 百度 · facebook
Google · Microsoft · Tencent 腾讯
Face++ 旷视 · ellipticlabs · 商汤 sensetime
Thundercomm · Uncanny Vision
mi · vivo · smartisan · oppo
ONEPLUS · MOTOROLA · LG · BLACK SHARK
ASUS · THETA · KEDACOM

Broad industry collaboration

Uniquely positioned to make the intelligent wireless edge a reality

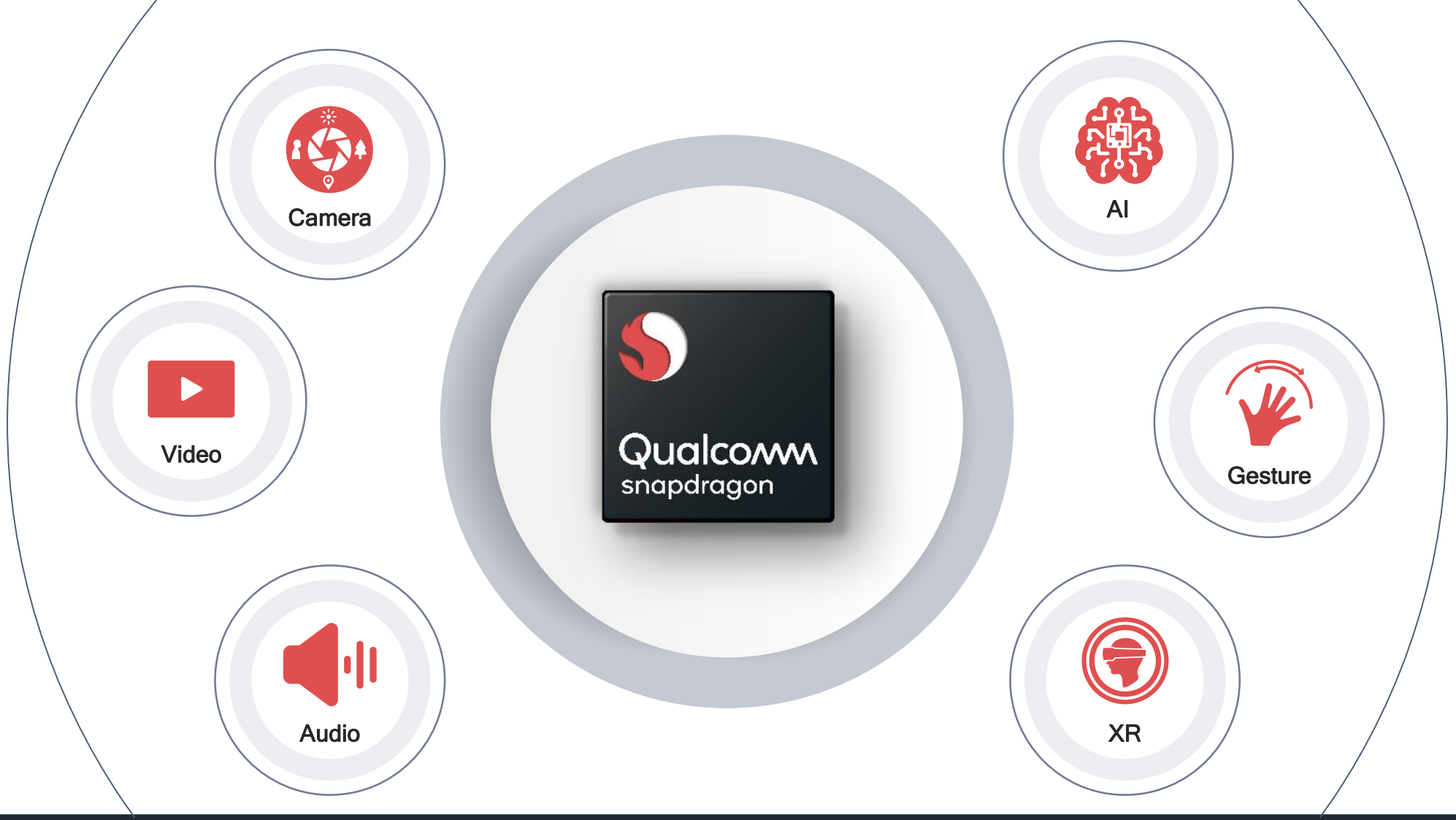
Design expertise

Bringing power efficiency to all mobile clients



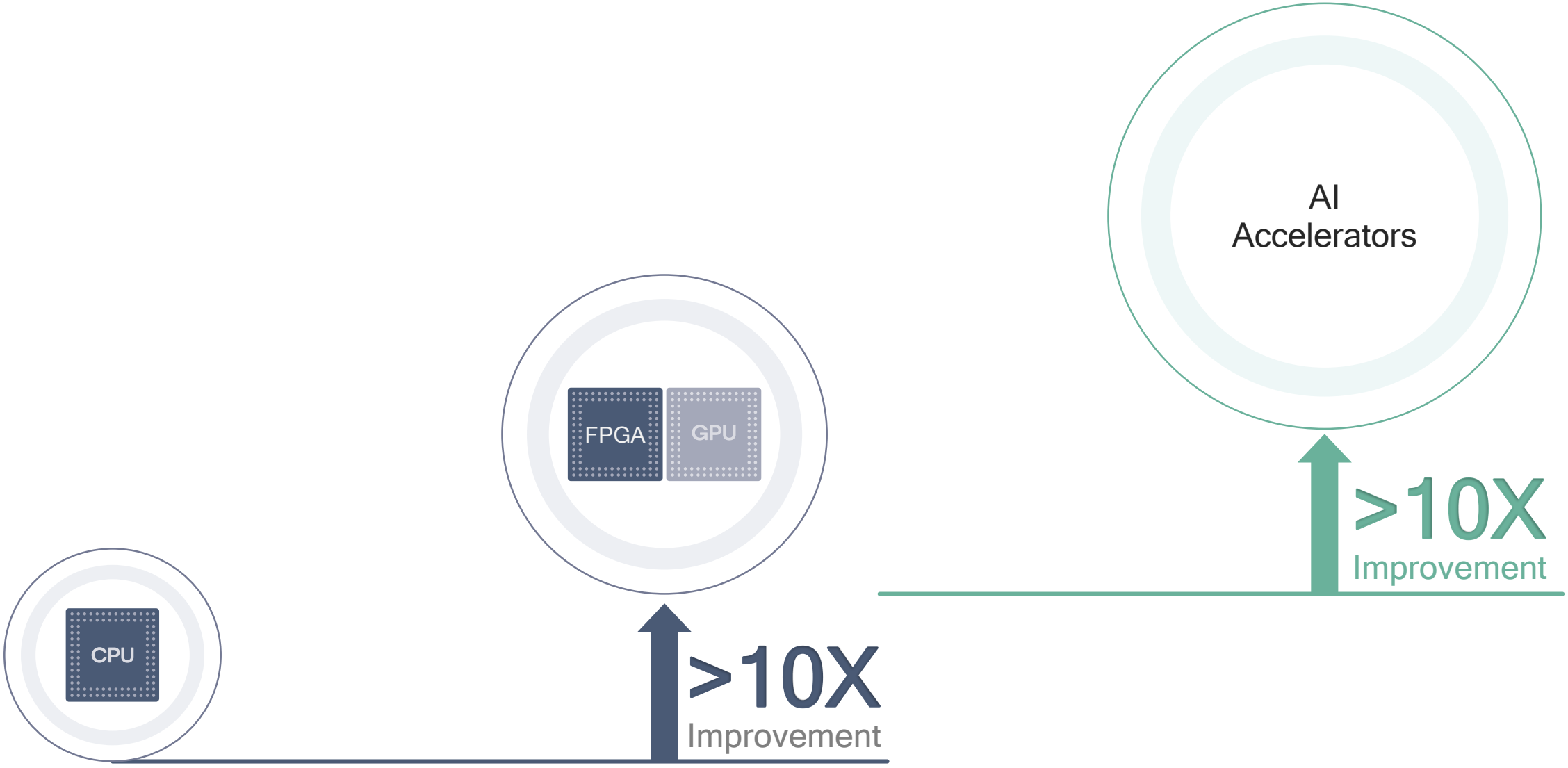
Process node leadership

Scale



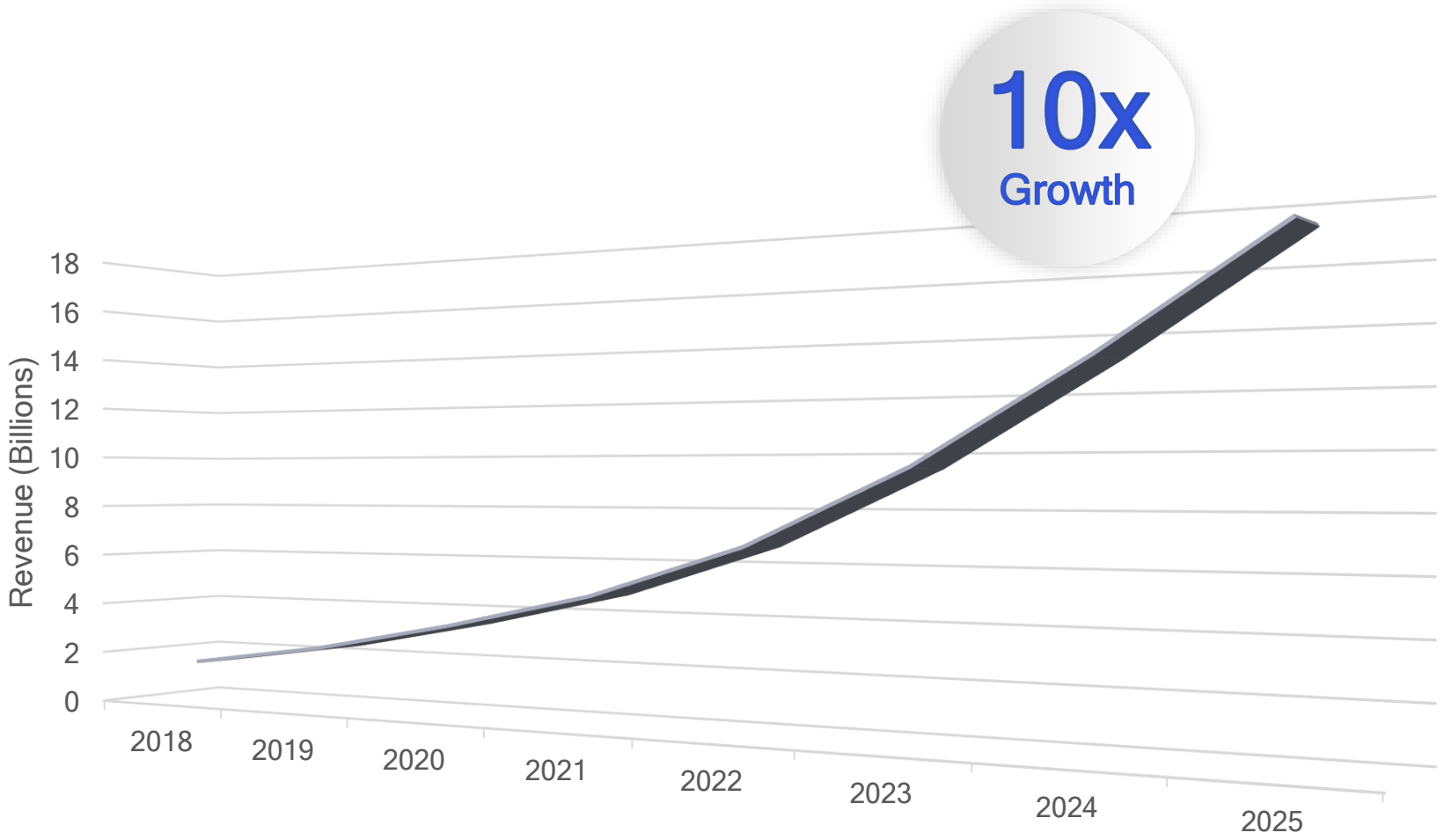
Low power signal processing across all key user experiences

An architecture shift in AI cloud inferencing

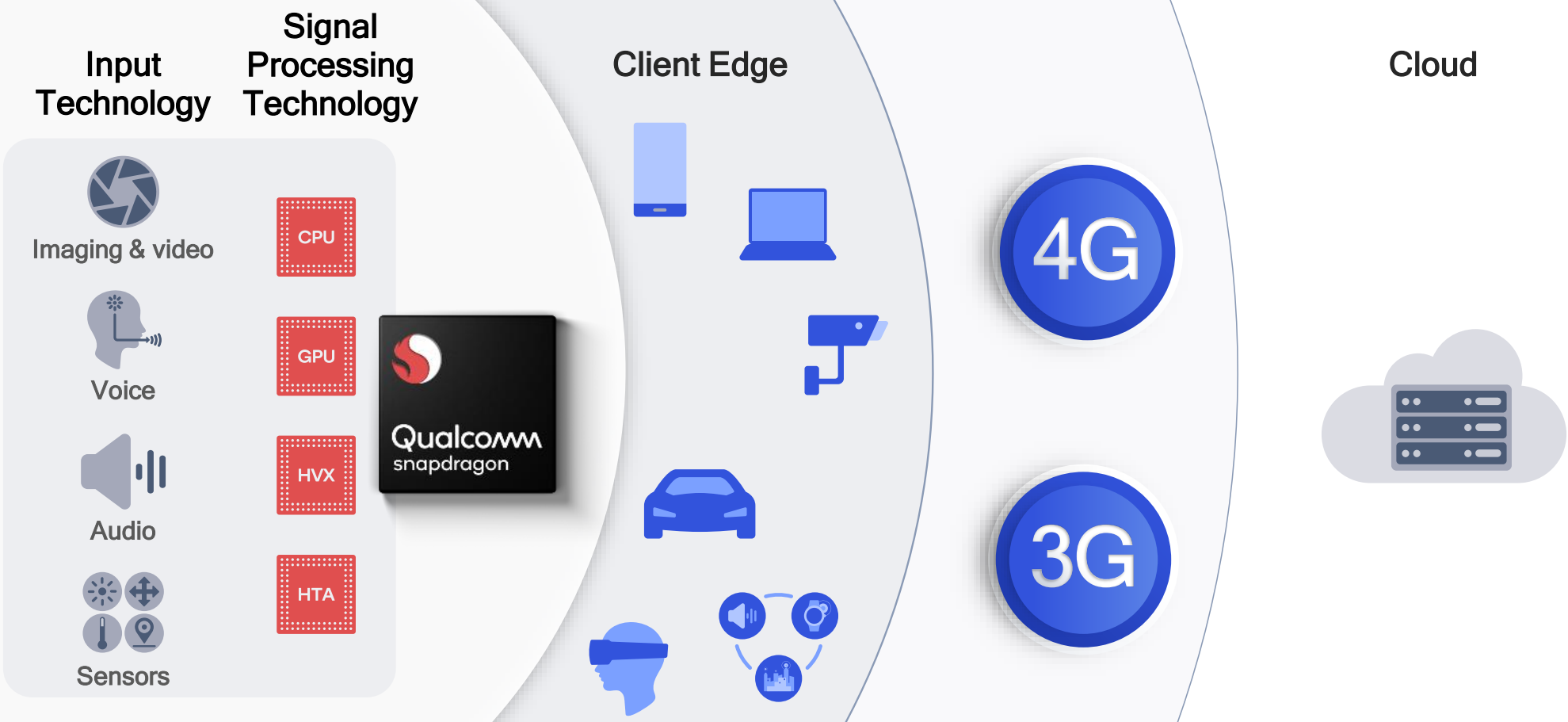


By 2025, market ramp of AI in datacenters

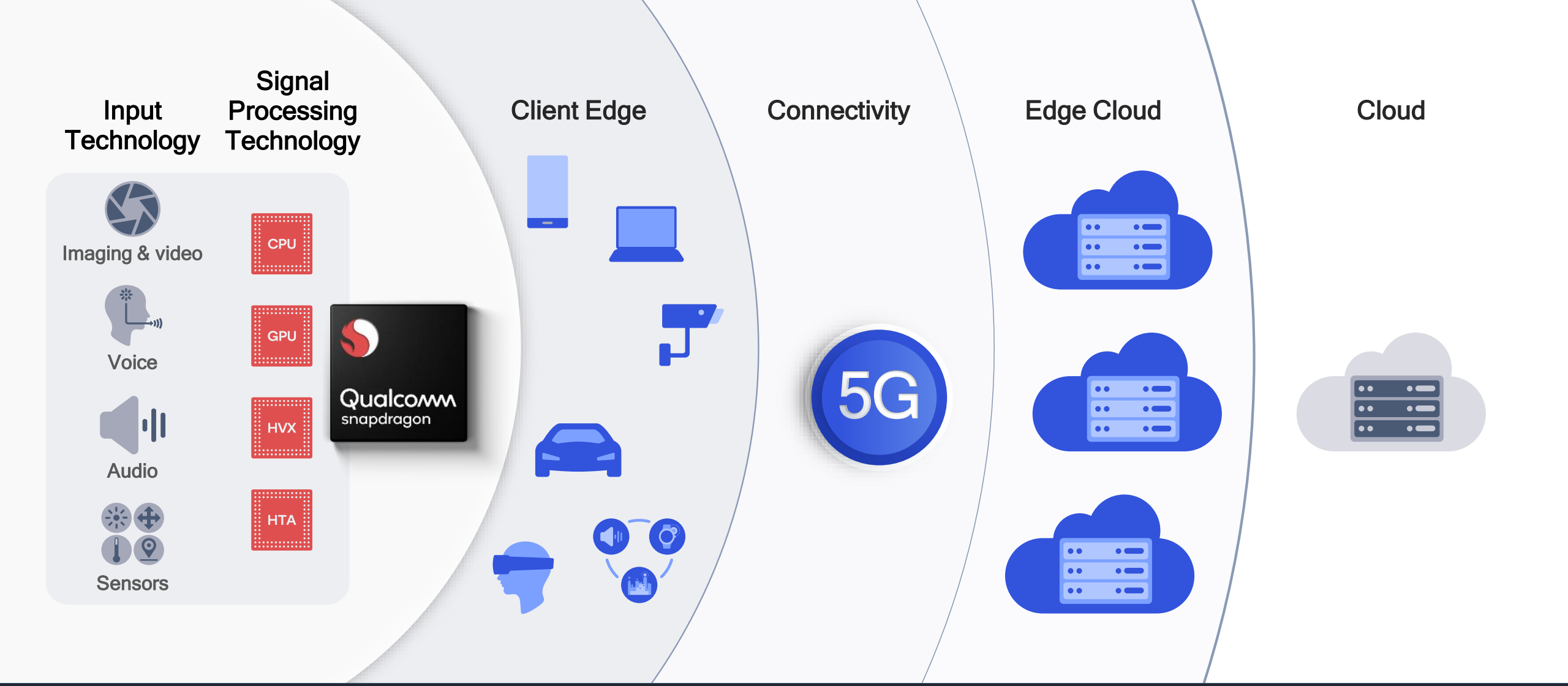
$17 Billion



Source: Tractica



Need to lower latency and increase cloud AI processing performance



5G and more powerful Edge Cloud processing will transform user experiences

April 9, 2019 @qualcomm San Francisco, CA

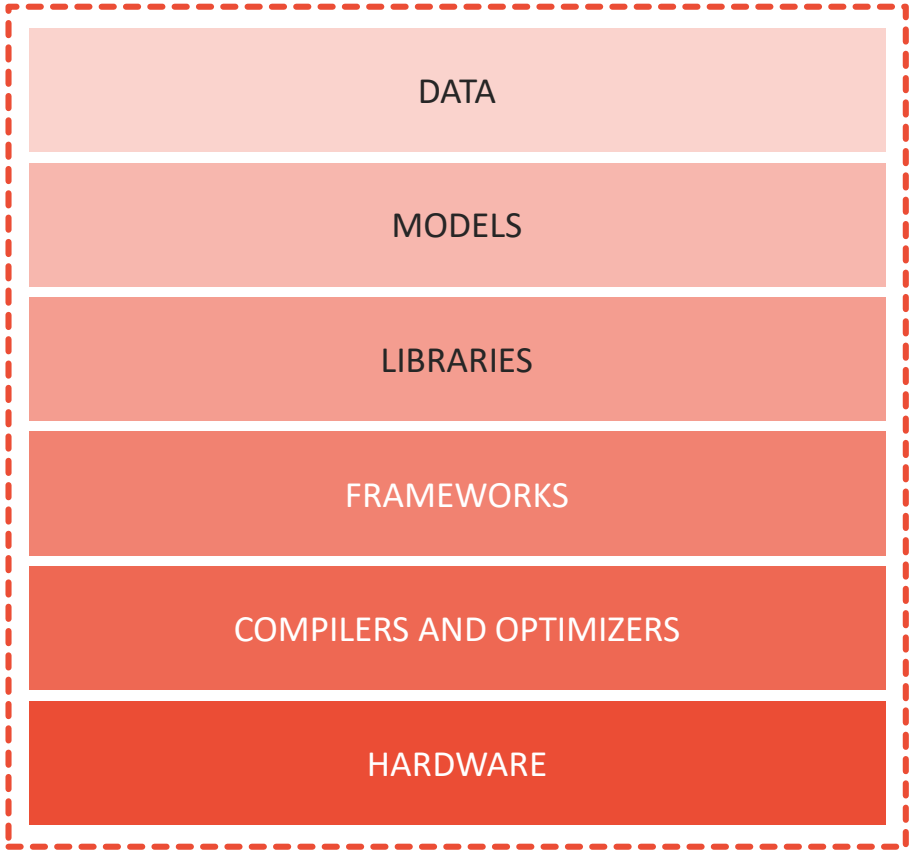
Qualcomm

AI at Facebook

Joe Spisak

Product Manager
Facebook AI

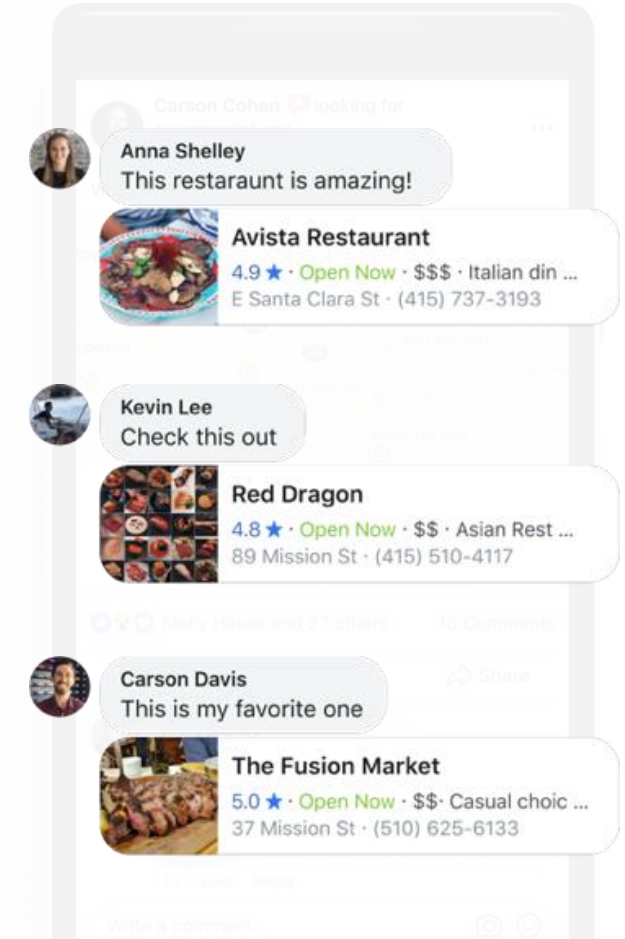
Full Stack Approach



AI Powered Use Cases

Enhancing Existing Products

SOCIAL
RECOMMENDATIONS



Enhancing Existing Products

SOCIAL
RECOMMENDATIONS

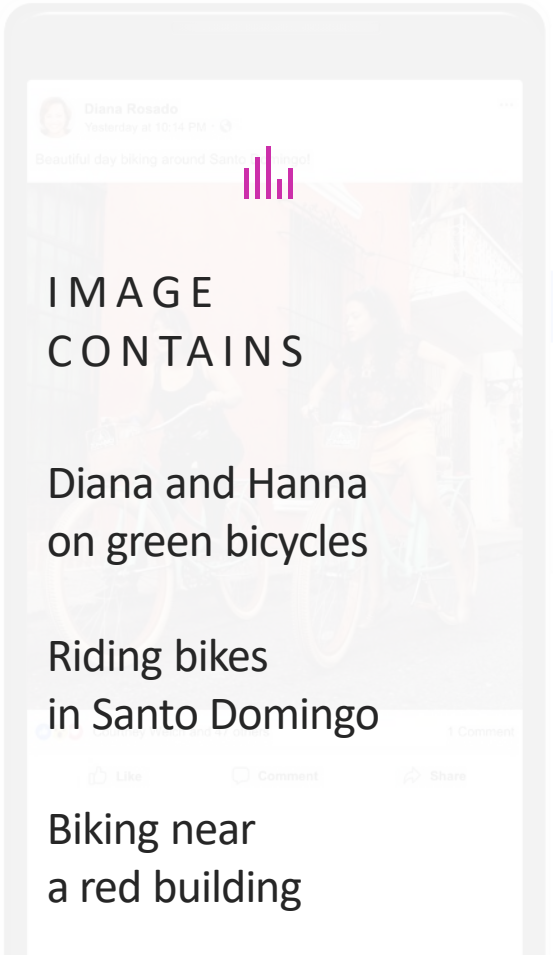
MACHINE
TRANSLATIONS

Enhancing Existing Products

SOCIAL RECOMMENDATIONS

MACHINE TRANSLATIONS

ACCESSIBILITY



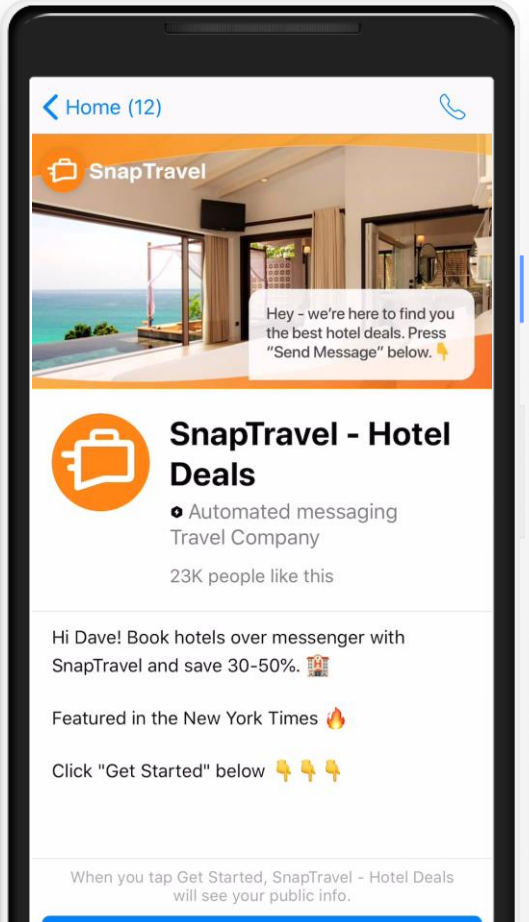
Powering New Experiences

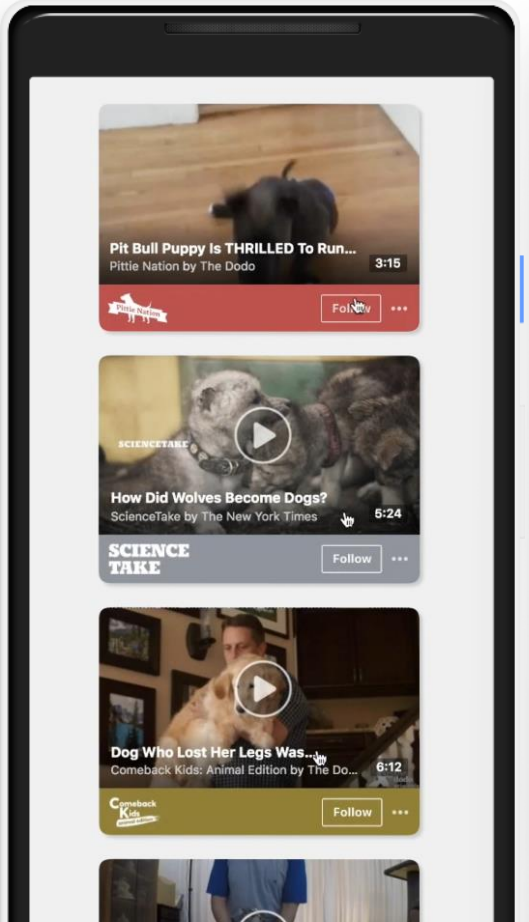
BOTS & ASSISTANTS

GENERATED CONTENT

AR EFFECTS

VR HARDWARE

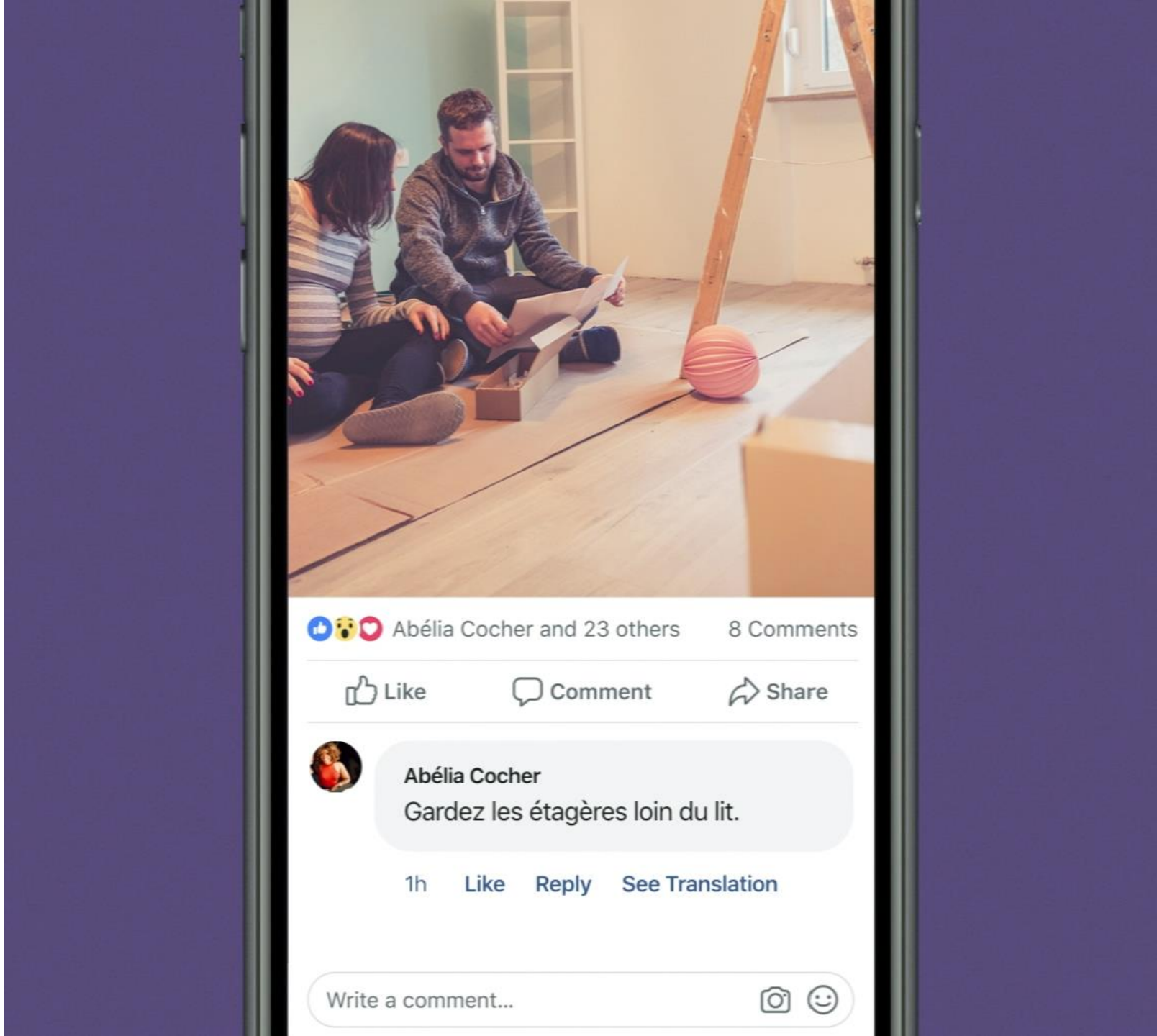




Datacenter AI Trends

5.95B+

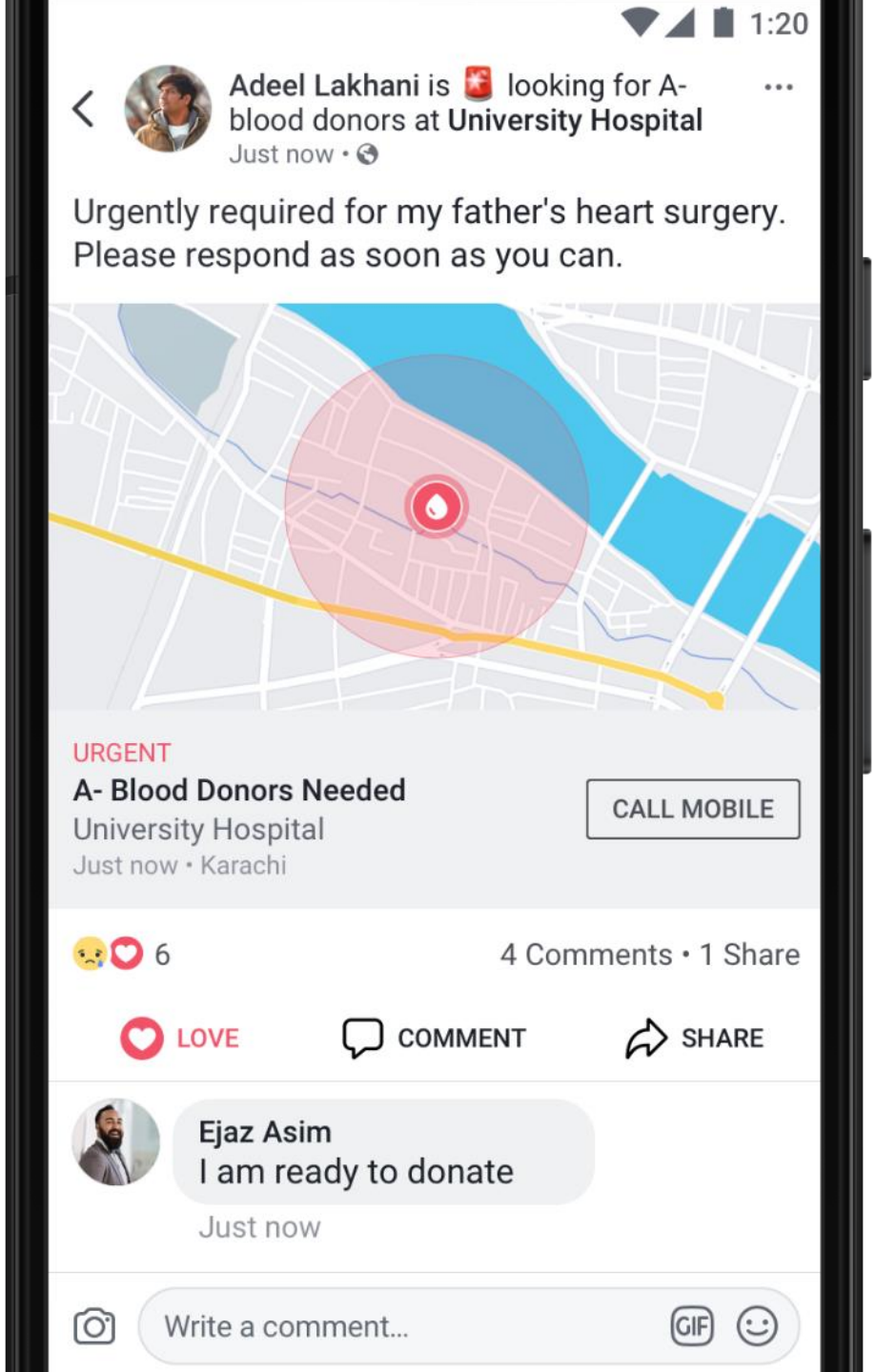
Translations Per Day



NLP In Action

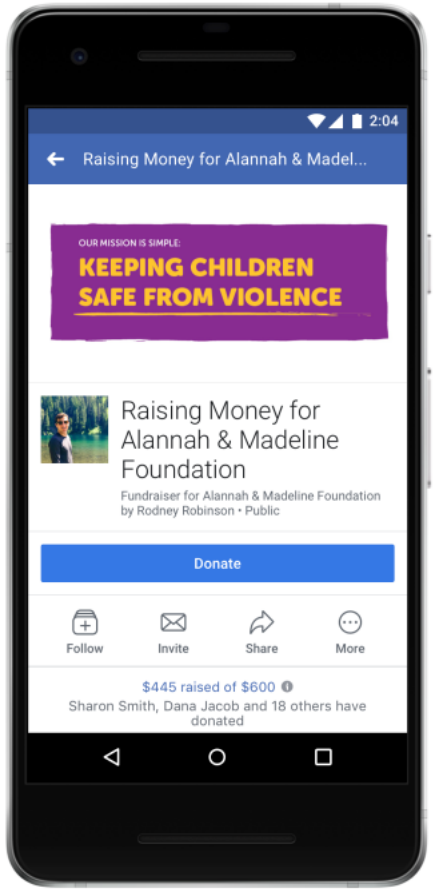
35M+

Blood Donors



$1B+

Raised For Charitable Giving
Powered By AI

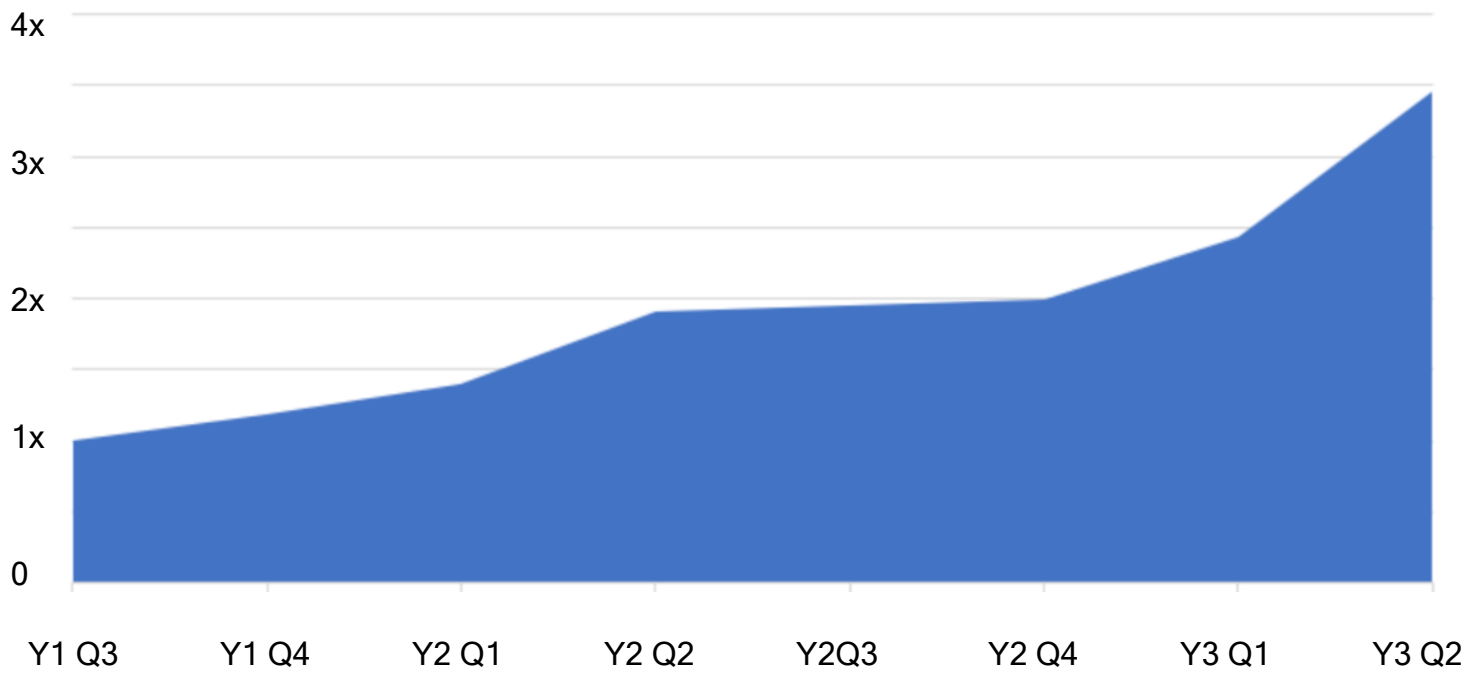


200+ Trillion

Inferences Per Day

Data-Center Power Consumption

Data-center power consumption is doubling every year



Deep Learning Inference in Facebook Data Centers: Characterization, Performance Optimizations and Hardware Implications - Nov 2016, https://www.researchgate.net/publication/329207337_Deep_Learning_Inference_in_Facebook_Data_Centers_Characterization_Performance_Optimizations_and_Hardware_Implications

Key Attributes For Next Generation Inference

Reliability

Latency

Power Efficiency

Modular server

8-socket server

OCP accelerator module

8-accelerator baseboard

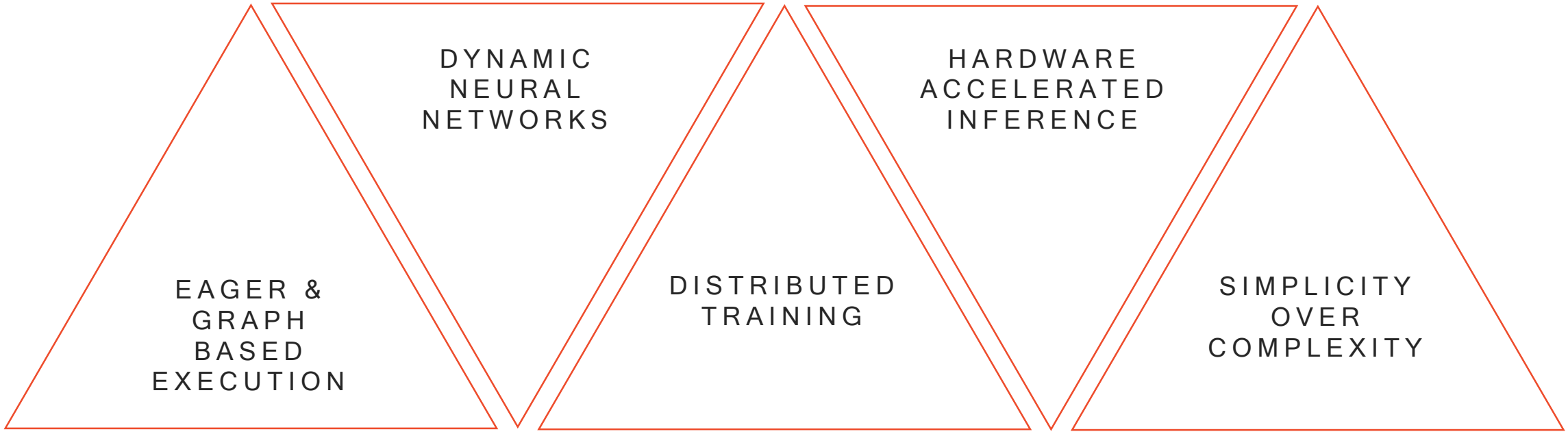
8-accelerator platform

Research to Production

Research to Production

PyTorch

A machine learning framework with an emphasis on:



PyTorch

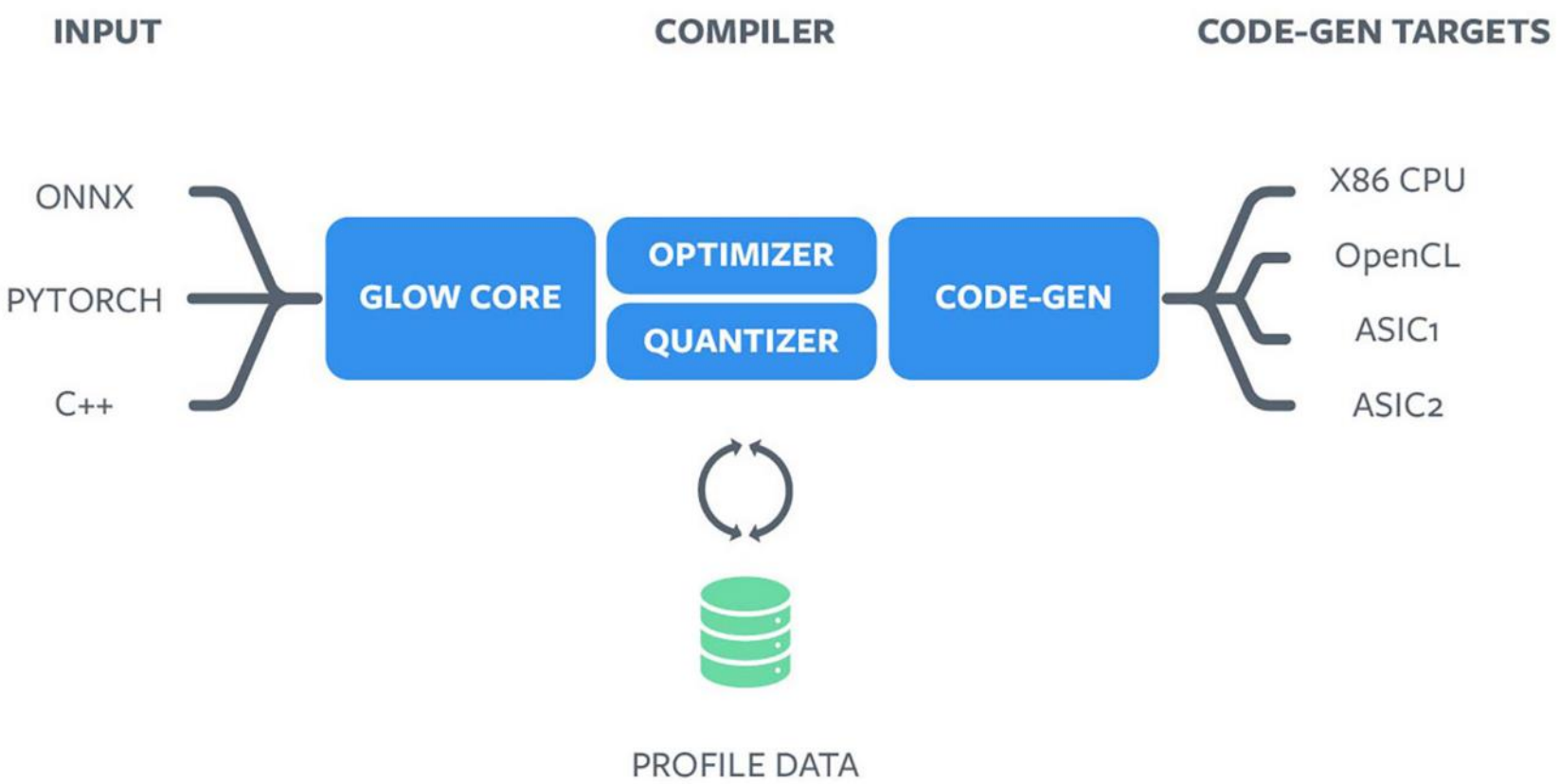
2.8X

INCREASE IN GITHUB CONTRIBUTORS

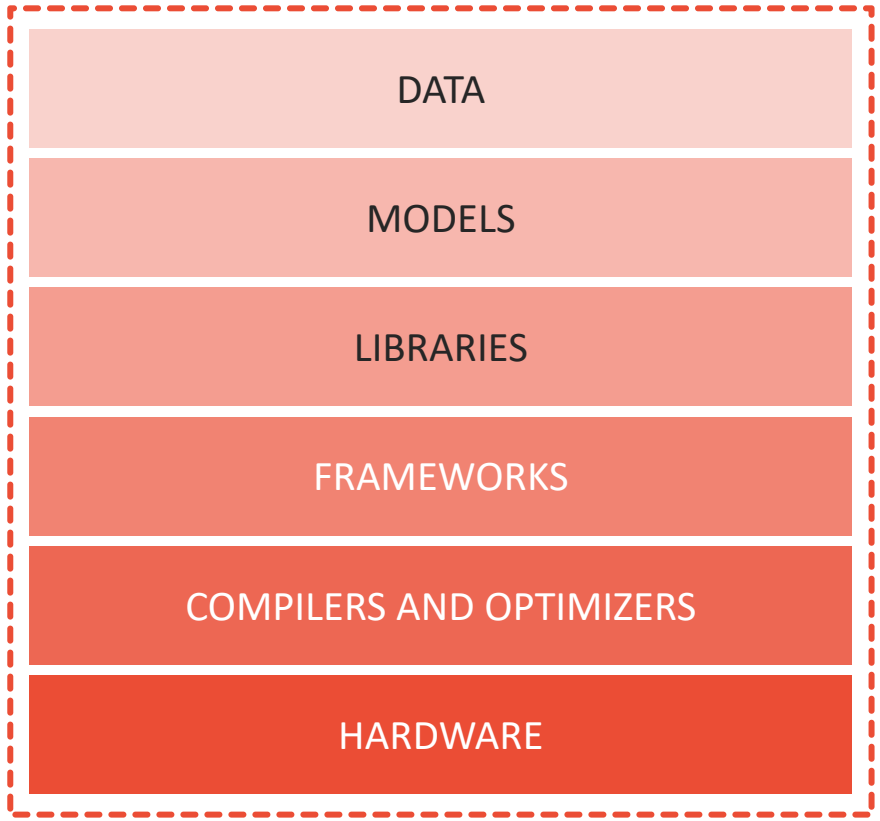
#2

FASTEST GROWING OPEN SOURCE PROJECT

Glow


INPUT
COMPILER
CODE-GEN TARGETS
ONNX
PYTORCH
C++
GLOW CORE
OPTIMIZER
QUANTIZER
CODE-GEN
X86 CPU
OpenCL
ASIC1
ASIC2
PROFILE DATA

Full Stack Approach



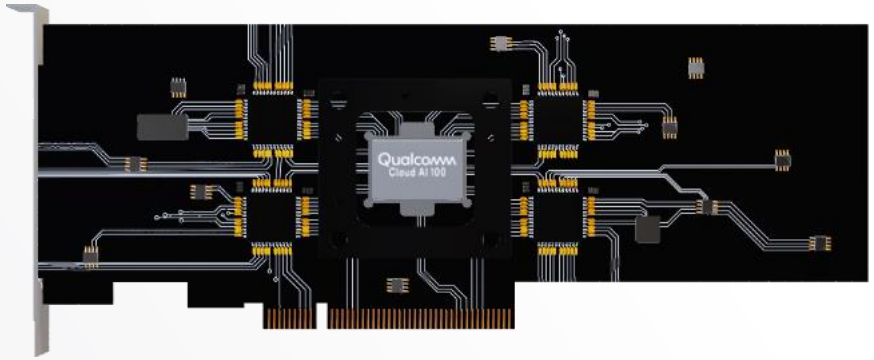
PYTORCH 1.0

GLOW, …

BIG BASIN, TIOGA PASS, …

Software

Full software stack



Features

Newsfeeds · Advertising · Personalized videos · Search · XR · Gaming

Tools

GLOW Compiler · TensorFlow Compiler · Performance Monitoring · Profilers · Debuggers · Card Tuning · Quantizers

Runtimes

PYTORCH · ONNX RUNTIME · TensorFlow

Frameworks

PYTORCH · Keras · mxnet · Cognitive Toolkit · PaddlePaddle · TensorFlow

April 9, 2019 @qualcomm San Francisco, CA

Qualcomm

AI at Microsoft

Venky Veeraraghavan (@venkyv)

Partner Group Program Manager
Microsoft Corp

Azure

AI on a massive global network

54

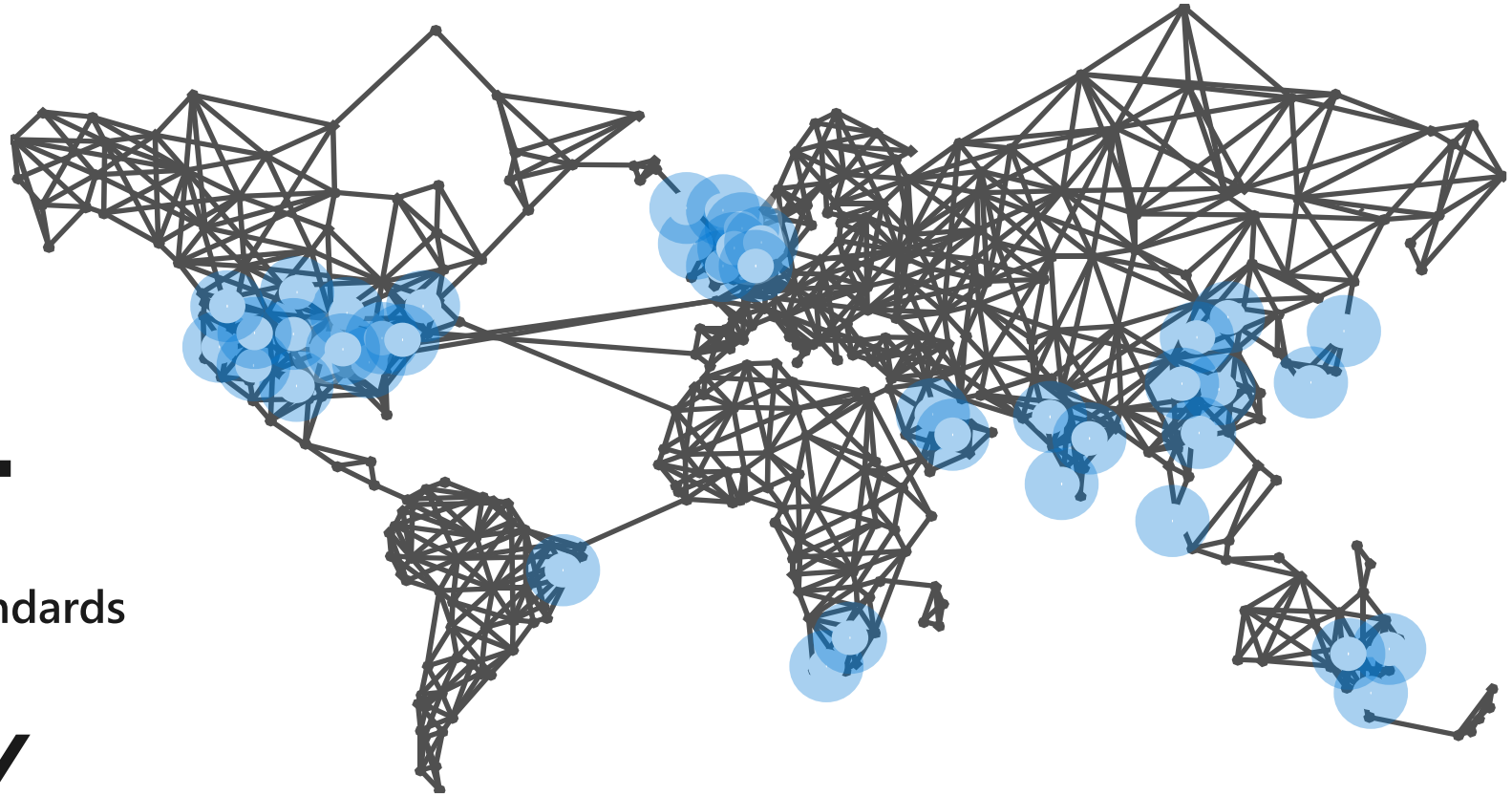
Azure regions

90+

Compliance standards

95%

Fortune 500 use Microsoft Azure



Data

Your data + Microsoft data

Cloud

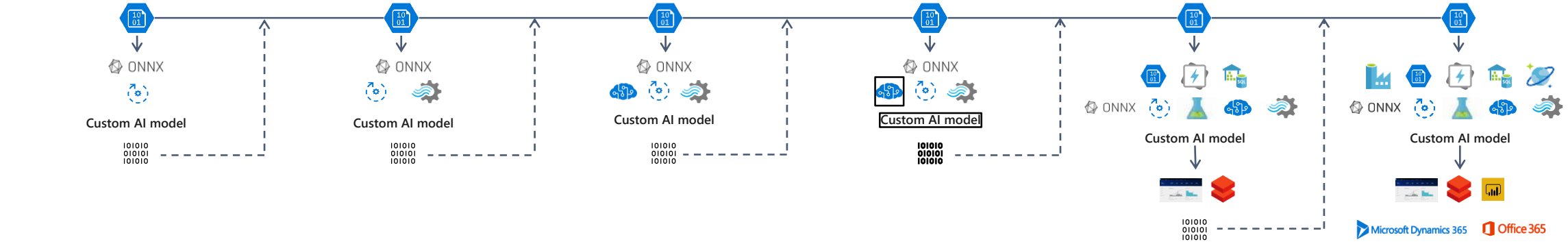
Power of Azure

Models

Breakthrough advancements

Deploying AI to Edge to Cloud

Sensors + control		Sensors to interactive	On-premises infrastructure		Global scale processing
MCU based	IoT devices	Edge devices	Edge appliances	Edge cloud	Hyperscale cloud
Inferencing	Inferencing	Inferencing	Inferencing	Inferencing	Inferencing
Training	Training	Training	Training	Training	Training
Data	Data	Data	Data	Data	Data



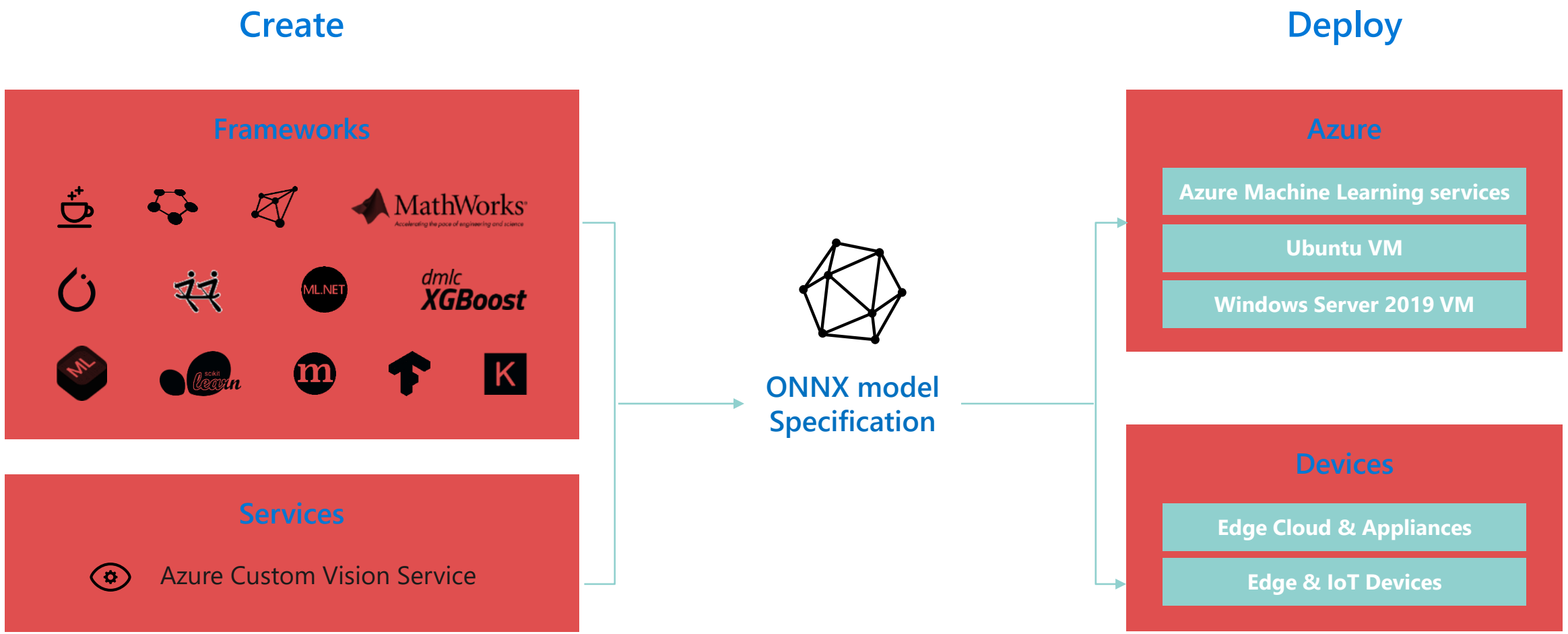
Less aggregated data and insights — **Full spectrum of data collection and aggregation using Azure Storage** — More aggregated data and insights

Less AI functionality — **Azure AI functionality availability based on computational resources** — Most AI functionality


Missing equipment
Flammable equipment near heat source
No hat


No hat

ONNX - open specification for ML models



ONNX Runtime is open source

Windows ML

- Runs on **Qualcomm Snapdragon SDM850** and the new **Snapdragon 8CX**
- Windows ML API allows developers to easily integrate pre-trained ML models into their applications
- Built on top of DirectML, a low-level API in the DirectX family

Always Connected PCs

Adobe

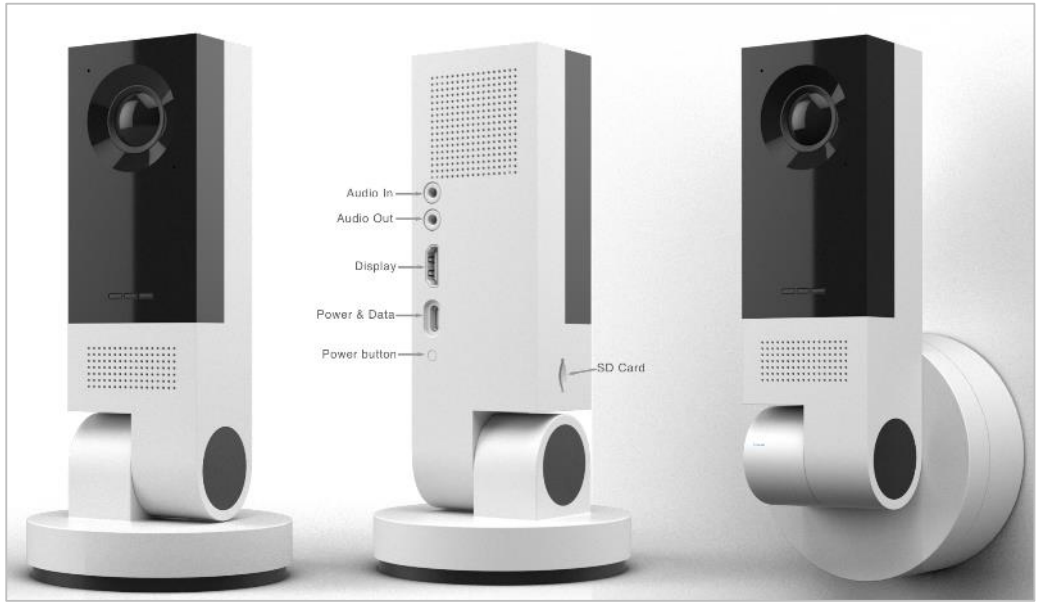
Office

Thundersoft AI Developer Platform

Intelligent Edge AI Devices

- Run AI models on the edge without additional computers or web connection or leverage the cloud
- Create, deploy and manage all your models in the cloud and the edge with Azure ML and Azure IoT Edge

Vision AI Developer Kit

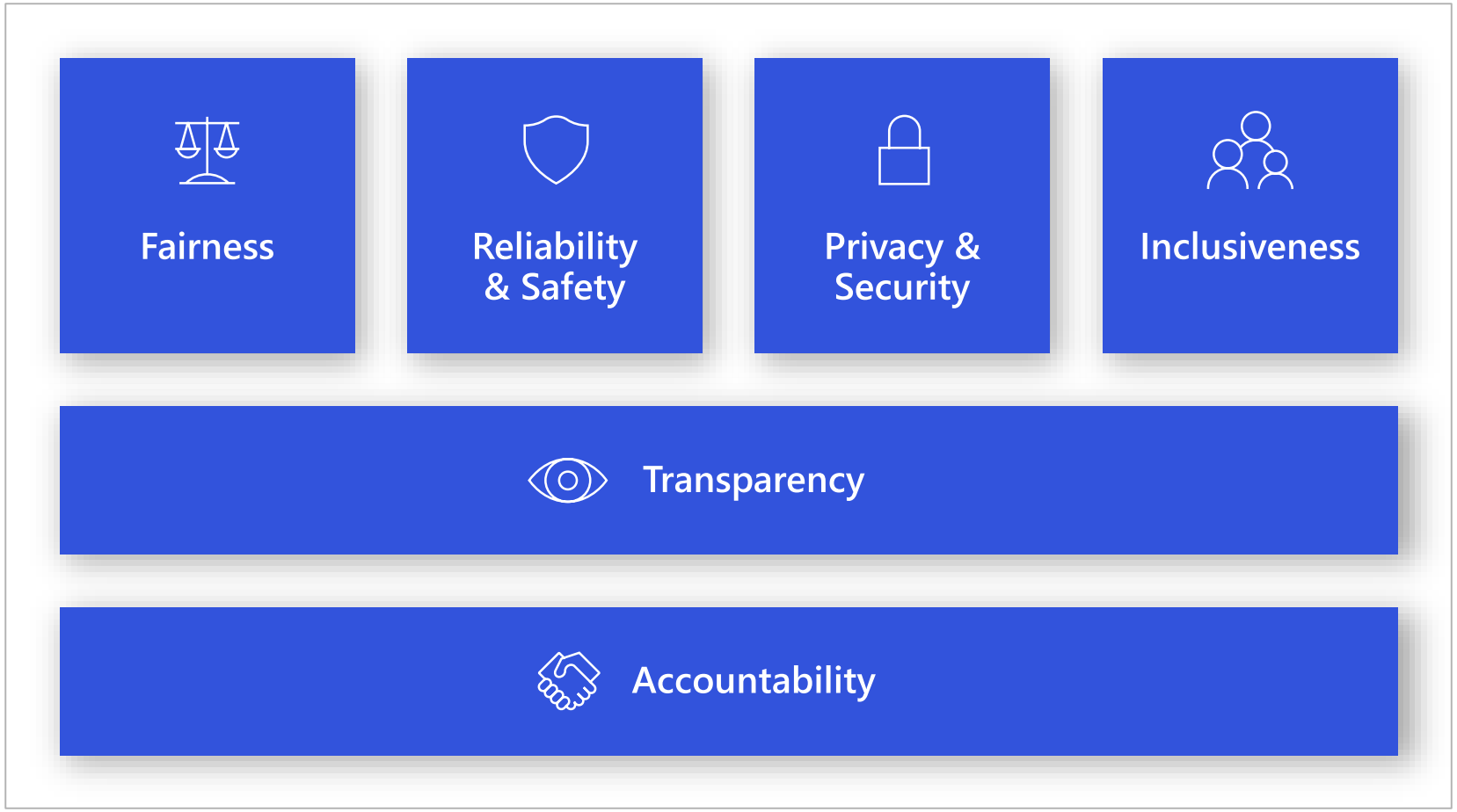


Based on the Qualcomm 603

Hololens 2

Based on the Qualcomm 845

Responsible AI



Better product serving the Broader Population

Responsibility and Social Impact

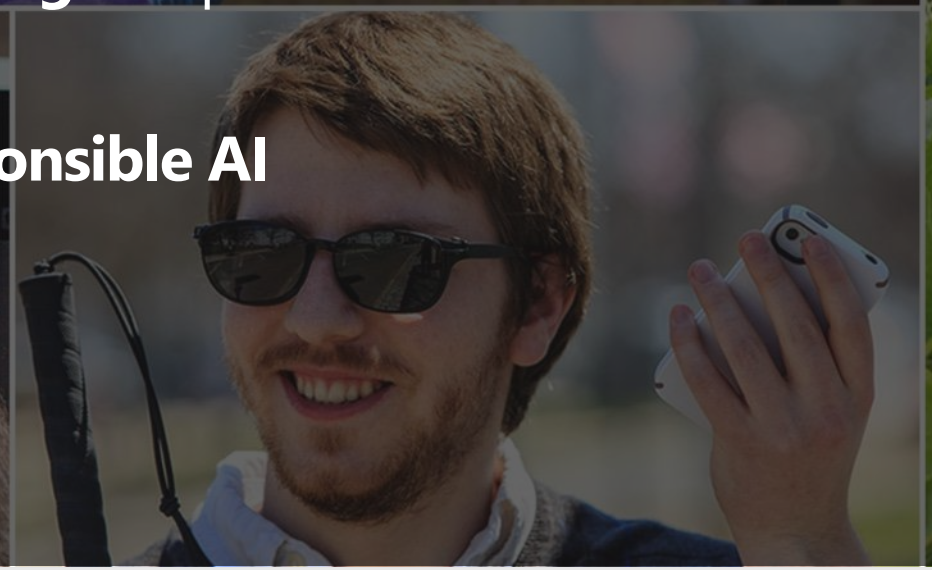
Legal and Policy

Competitive Advantage and Brand

In closing...

Microsoft provides a comprehensive platform for Machine Learning that leverage Qualcomm silicon

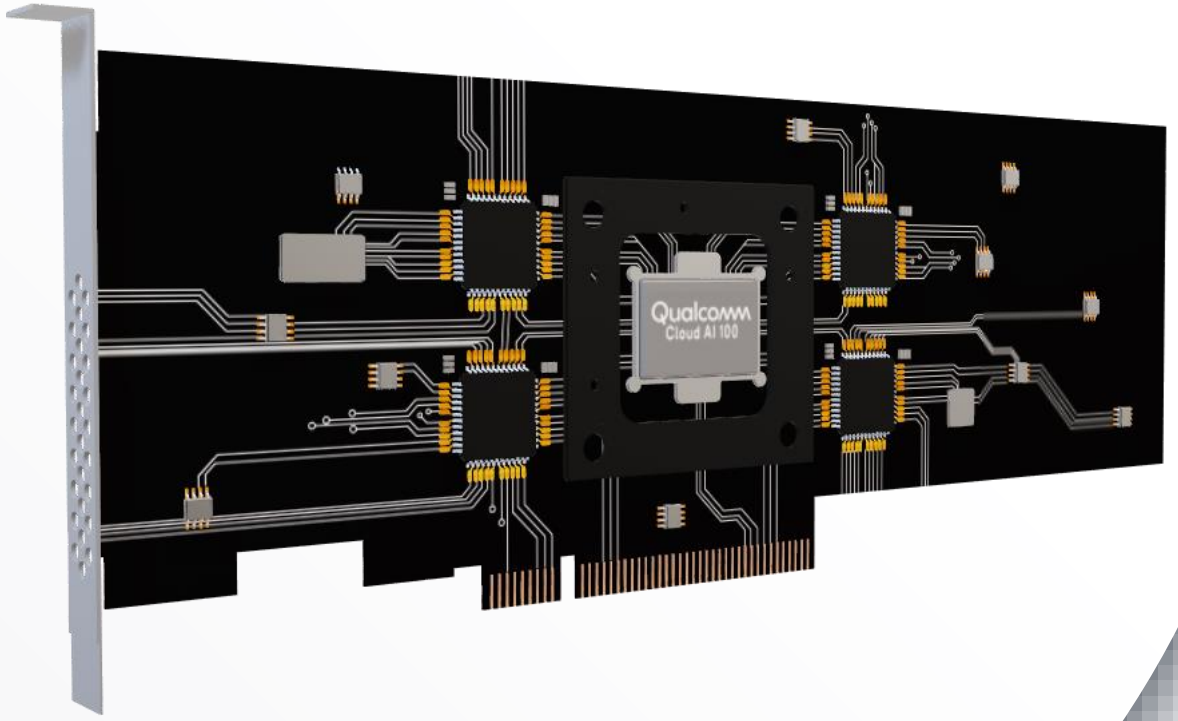
- Hyperscale end-to-end **Cloud AI** platform
- Consistent & Open **Edge AI** platform for IOT and Windows devices
- Foundation for **Responsible AI**



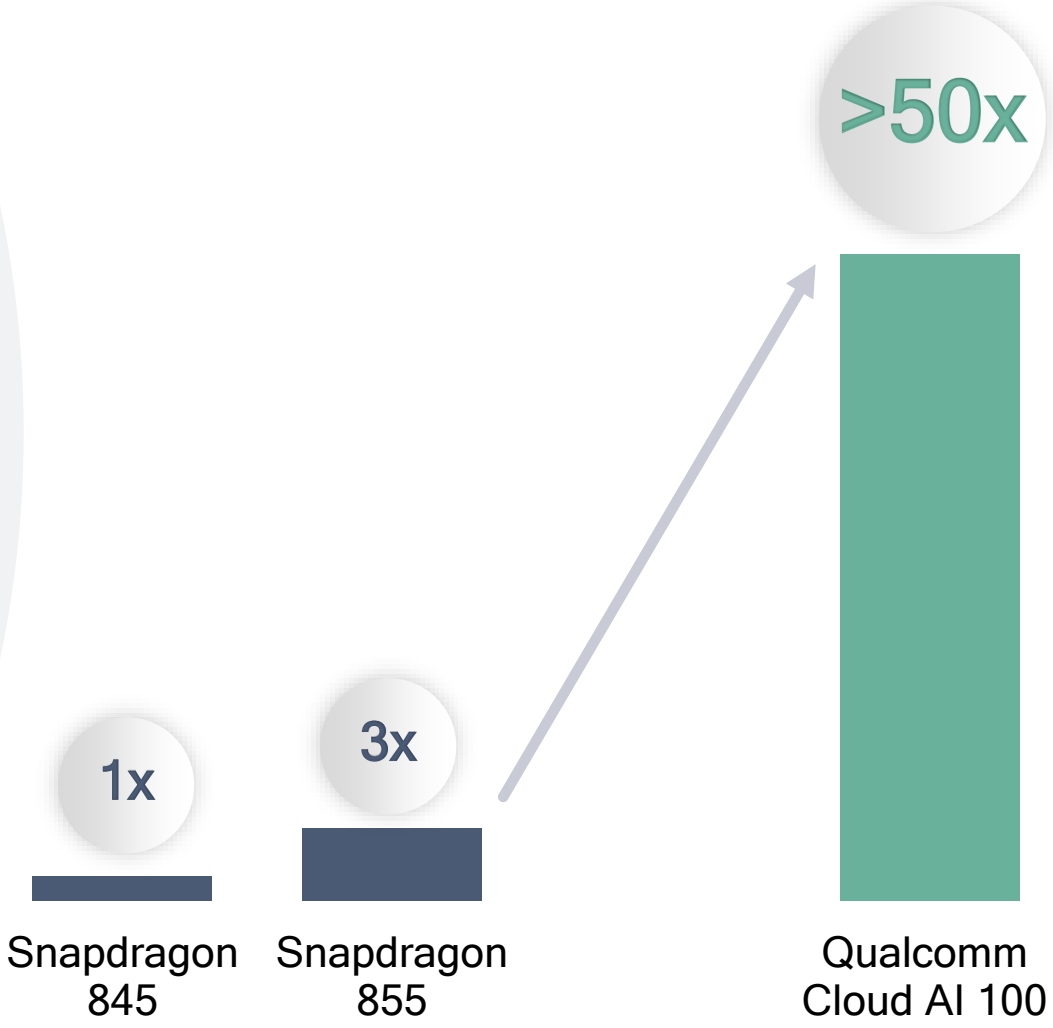
Microsoft

Performance

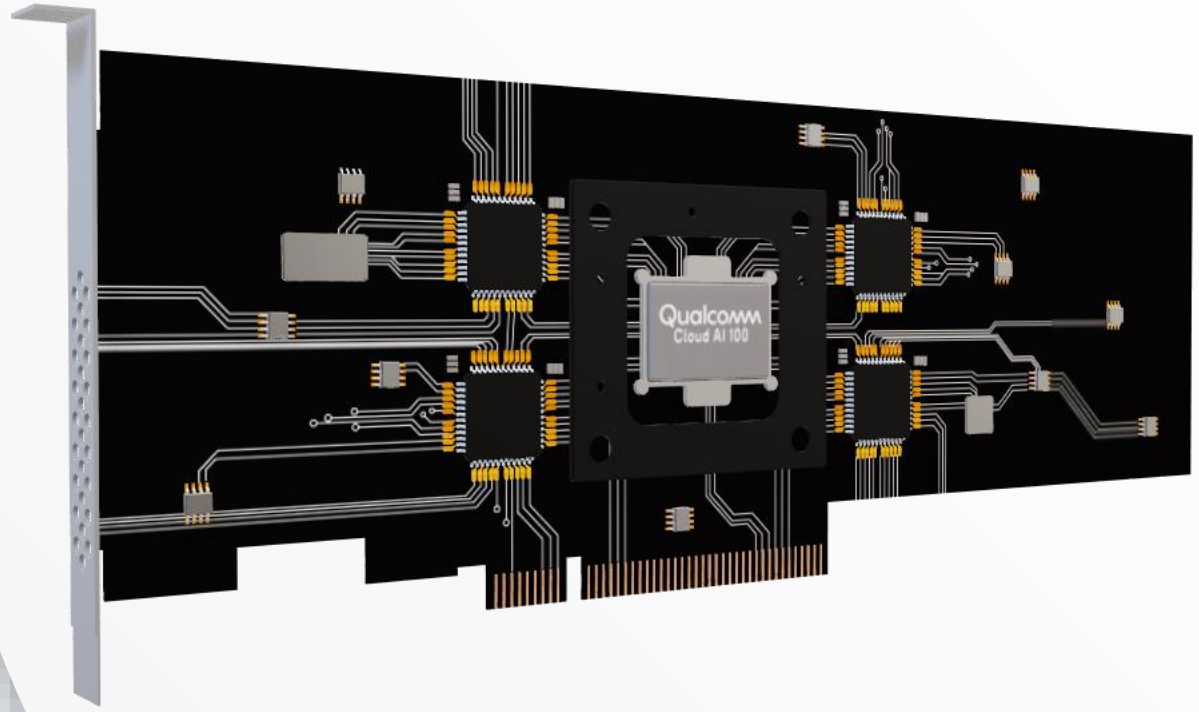
Built from the ground up



Peak AI performance



Qualcomm® Cloud AI 100



- Built on 7nm
- >350 TOPS Peak AI Performance
- Sampling 2nd half of 2019

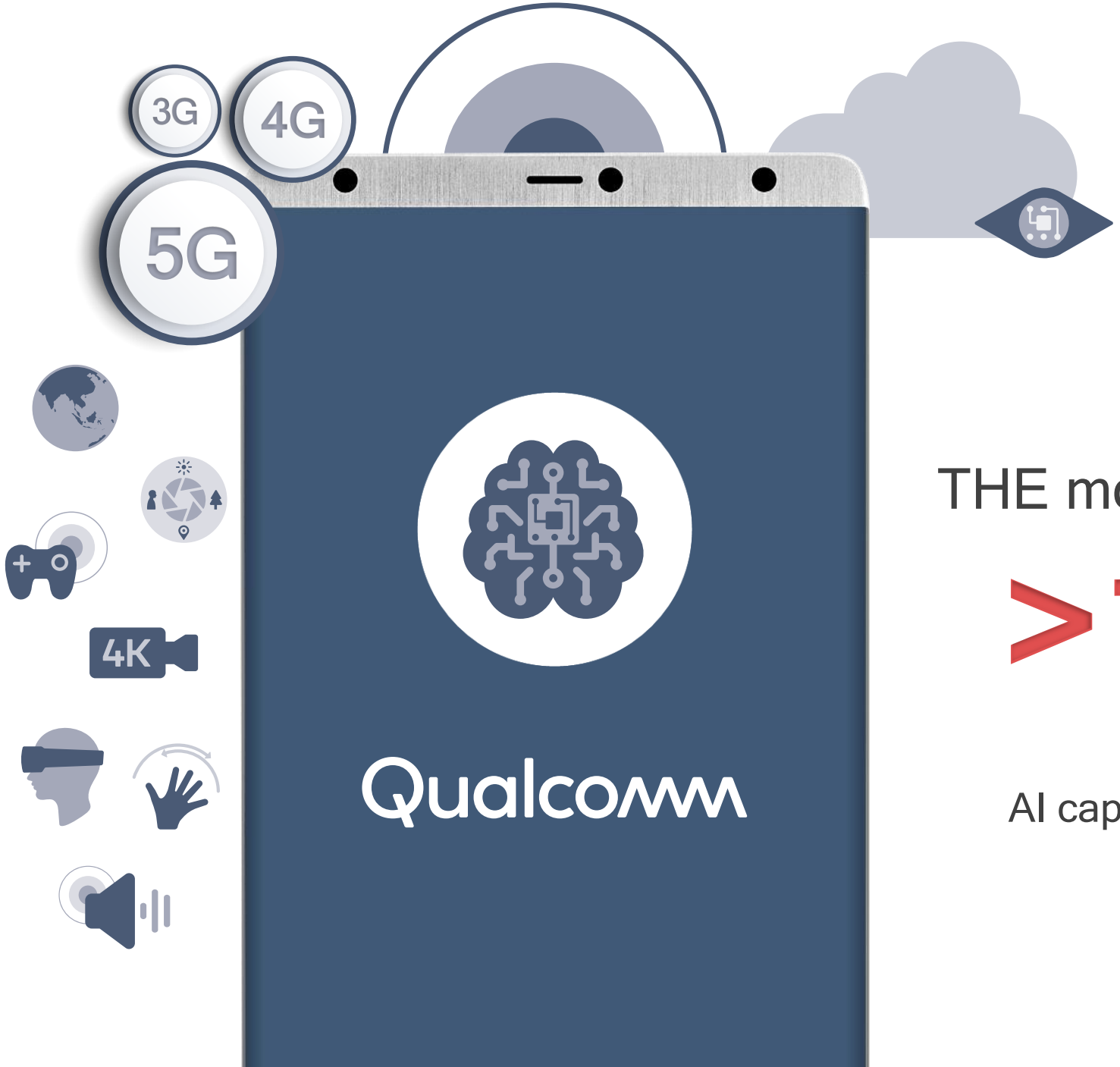
April 9, 2019 @qualcomm San Francisco, CA

Qualcomm

Mobile AI

Ziad Asghar

VP, Product Management
Qualcomm Technologies, Inc. (QTI)



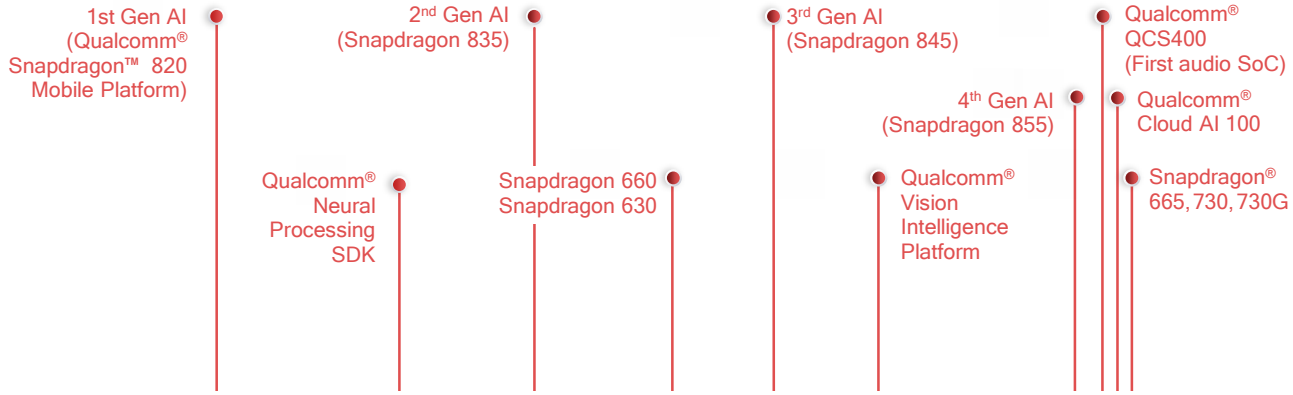
Mobile -
THE most pervasive AI platform

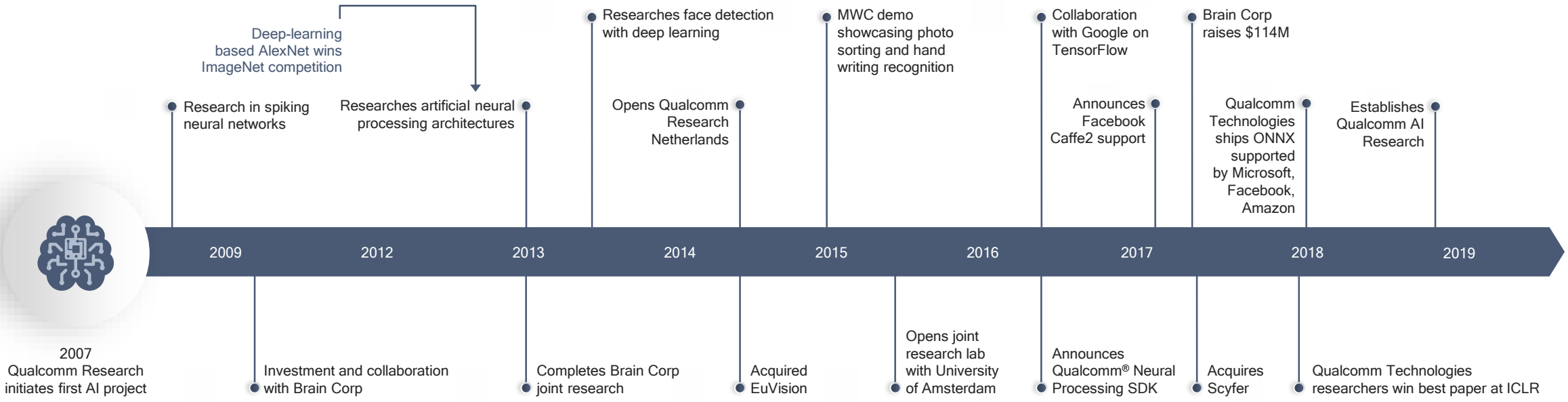
>1 Billion

AI capable devices enabled with QTI technology

Qualcomm Artificial Intelligence Research

Consistent AI R&D investment is the foundation for product leadership





The challenge of AI workloads

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

Power and thermal efficiency are essential for on-device AI

Constrained mobile environment

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

Qualcomm snapdragon 855 mobile platform

Adreno 640

50% More ALUs*
FP32 & FP16

Hexagon 690

New Tensor Accelerator

- QTI designed
- Dedicated to AI
- Multidimensional math and integrated nonlinear functions

4x Vector eXtensions*

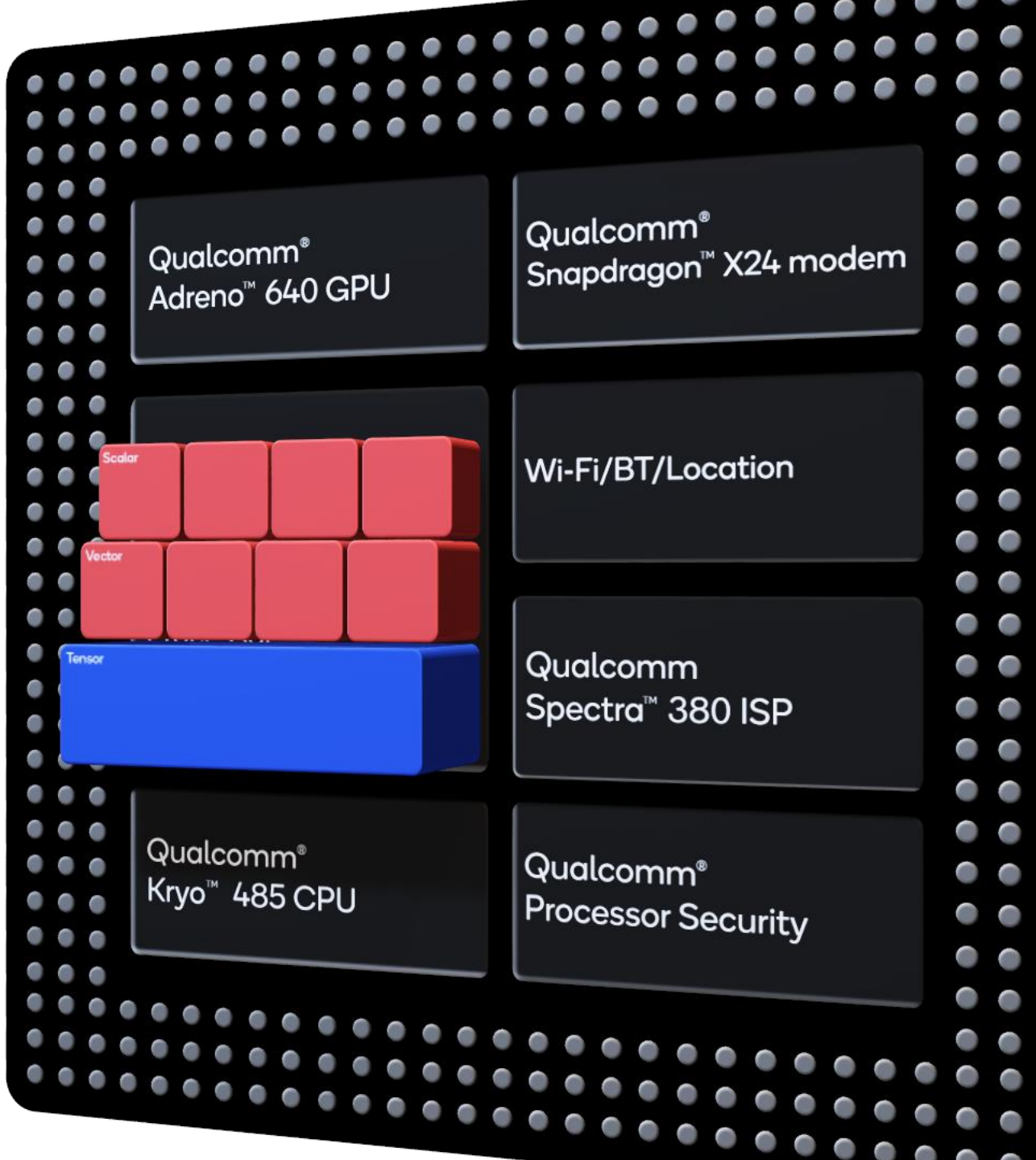
Optimized scalar

Voice Assistant

INT16, INT8 & Mixed

Kryo 485

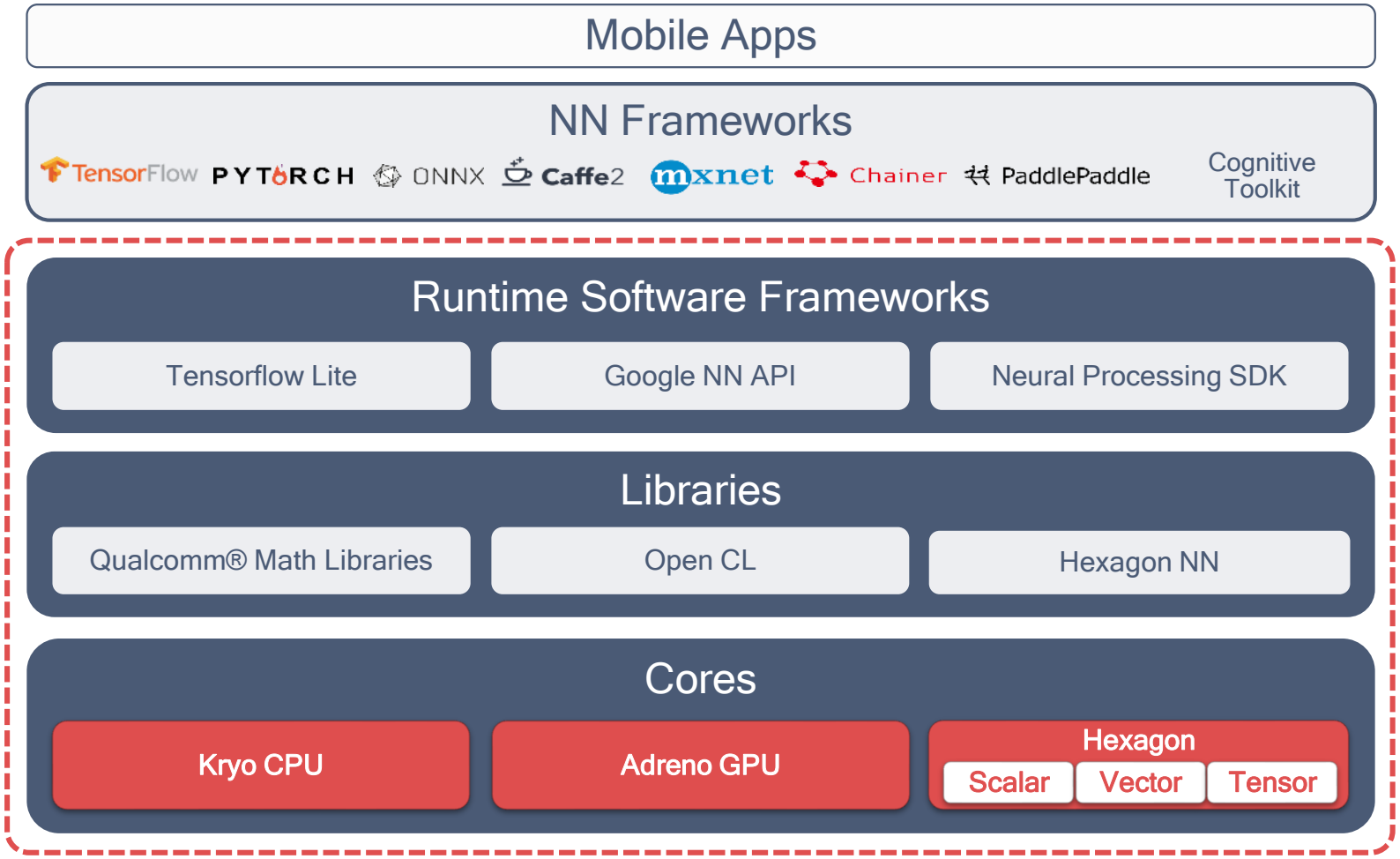
New dot product instructions
FP32 & INT8



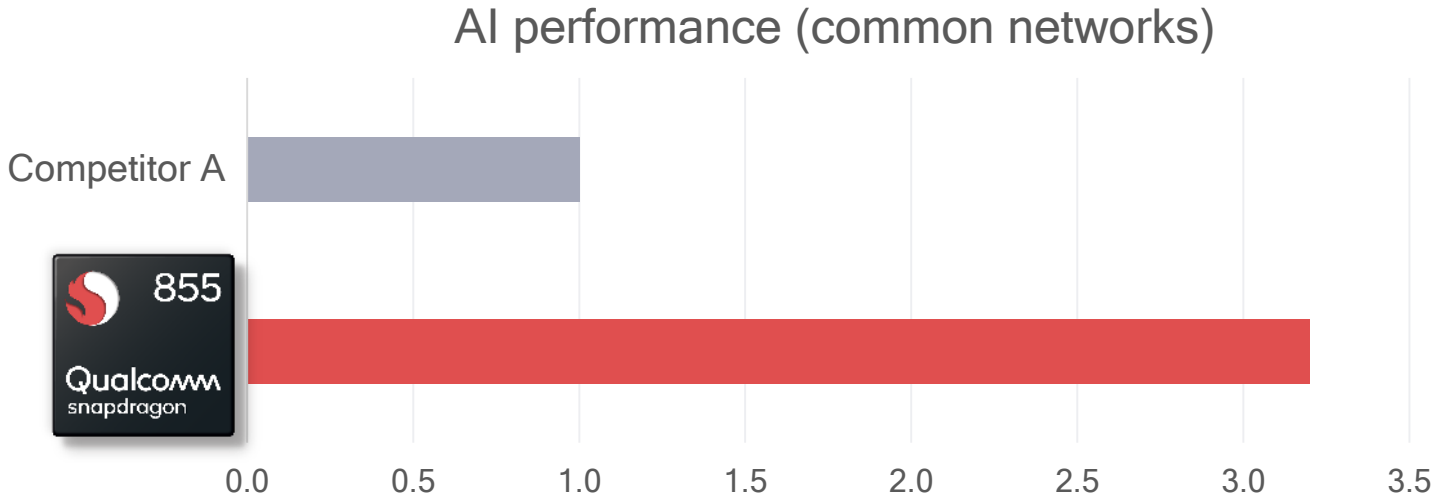
*Compared to Snapdragon 845

Qualcomm® AI Engine

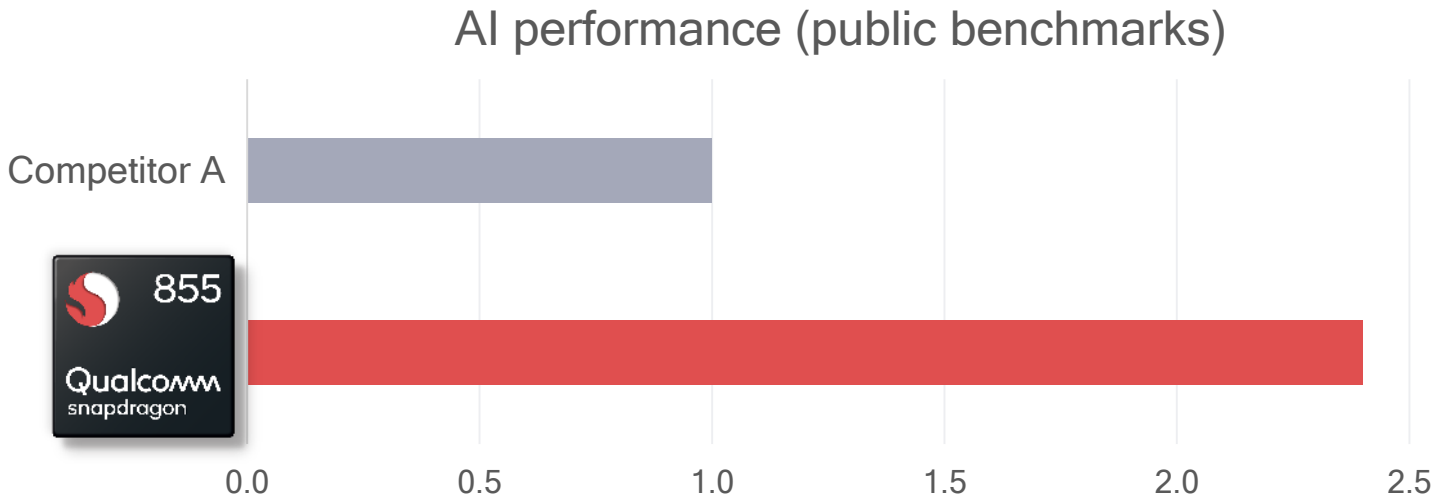
4th Gen
AI Engine



Snapdragon 855 AI performance

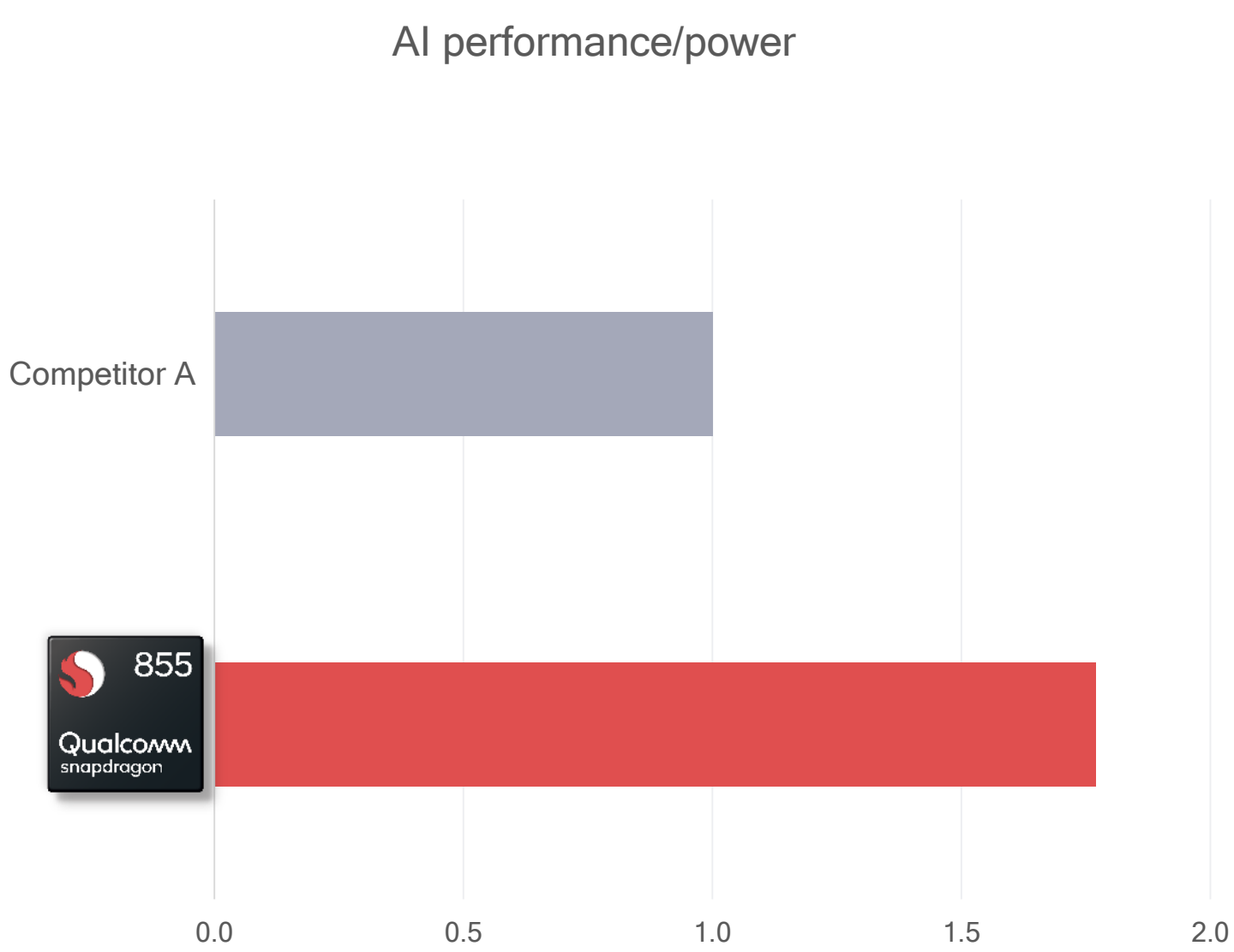


Note: Average over MobileNet, Inceptionv3, Resnet34 on commercial devices.



Note: Average over ETH, NeuroScope, Antutu AI, AIMark on commercial devices.

Snapdragon 855 AI power advantage



Note: Power measured running AIMark.

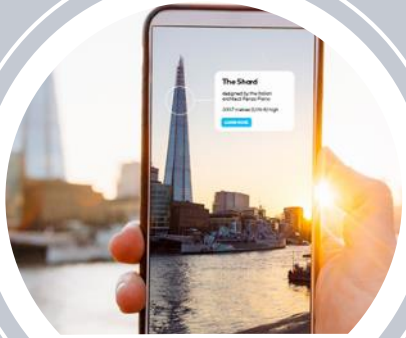
Announcement

Qualcomm snapdragon 665 mobile platform

Qualcomm snapdragon 730 mobile platform

Qualcomm snapdragon 730G mobile platform

Camera



AI

Gaming

Innovating in new tiers

Snapdragon 665 Mobile Platform

Snapdragon 665 Mobile Platform

Stunning pictures at every angle

- Up to 2X faster AI*
- Triple camera
- 48MP

* Compared to Snapdragon 660

Qualcomm
AI Engine
+ Camera

Snapdragon 730 Mobile Platform

Qualcomm snapdragon

Snapdragon 730 Mobile Platform

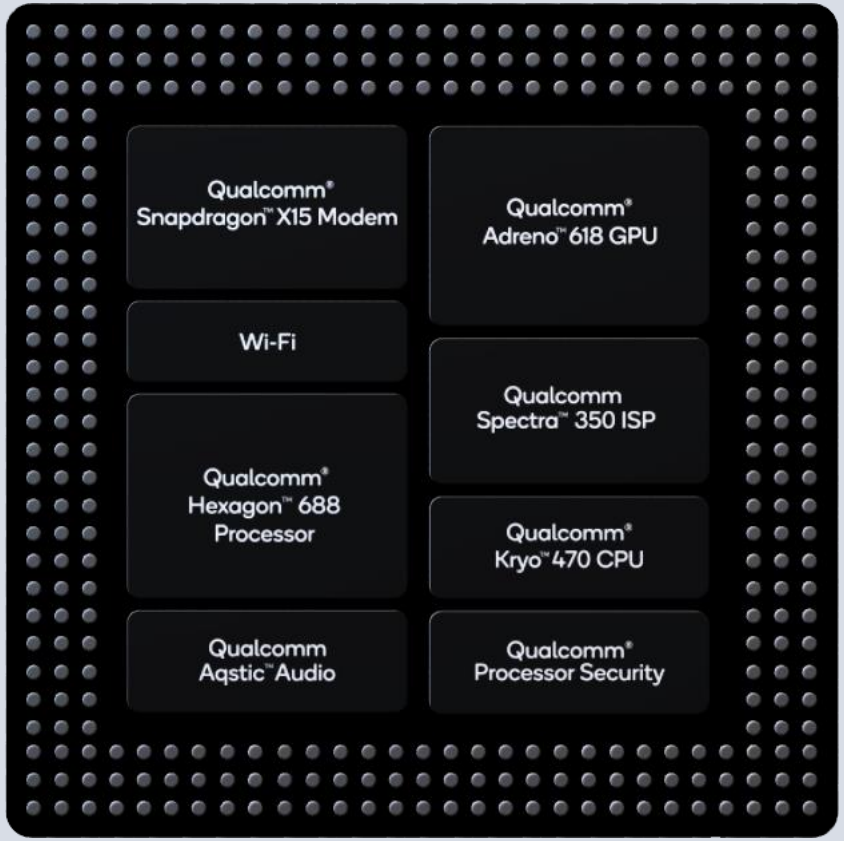
Awe-inspiring video capabilities

- 1st CV-ISP in 7 Series
- 1st Tensor Accelerator in 7 Series
- 1st True HDR in 7 Series

Qualcomm snapdragon 730 mobile platform

Firsts in the 7 series

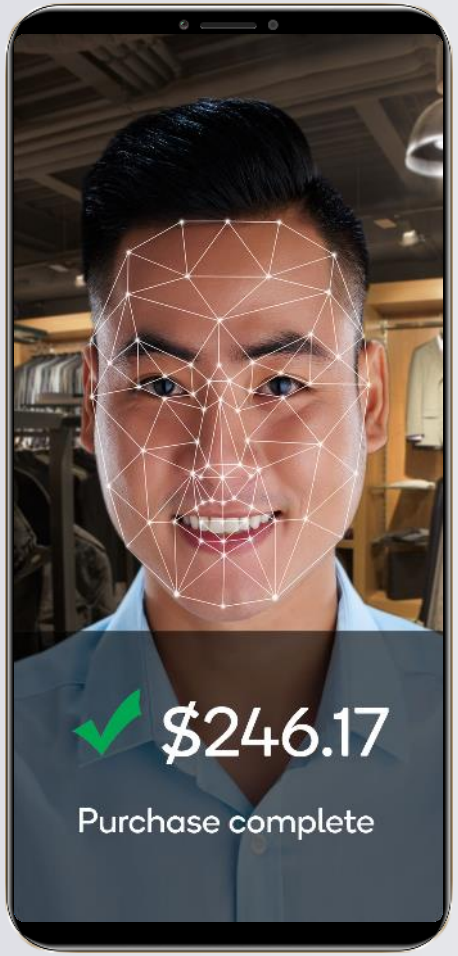
- 1ST 4th Gen Qualcomm AI Engine
- 1ST Tensor Accelerator
- 1ST Qualcomm® apt-X™ Adaptive Audio
- 1ST Wi-Fi 6 ready



- 1ST CV-ISP
- 1ST Qualcomm Spectra 3-series
- 1ST Kryo 4-series
- 1ST 8nm process

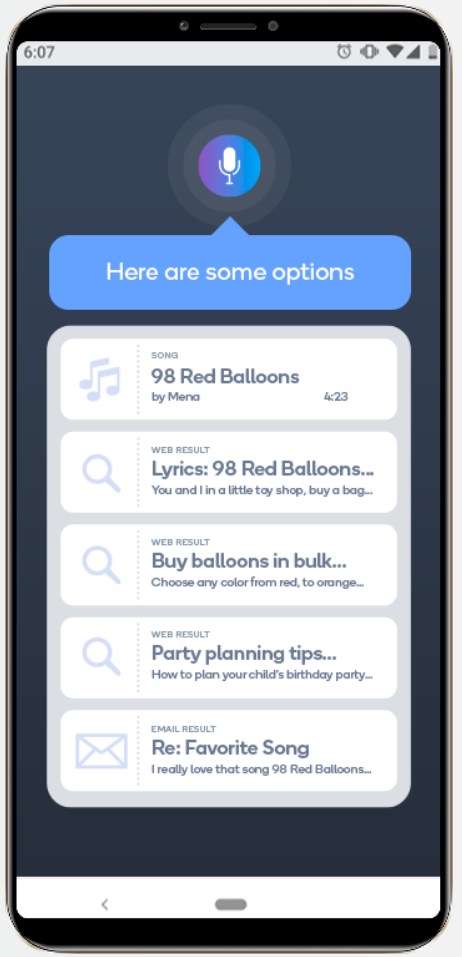
* Compared to Snapdragon 710

Smart cropping



Payments

Augmented Reality



Voice Assistant

On-device AI

Always-on, fast, more intuitive and secure

Snapdragon 730G Mobile Platform

Qualcomm
snapdragon

Snapdragon 730G Mobile Platform

Exceptional gaming performance and more

- Select Snapdragon elite gaming features
- 15% faster graphics rendering than Snapdragon 730

AI support on QTI SoCs

= AI Engine
= AI Capable

Snapdragon 8 Series	SD820, SD835, SD845, SD855
Snapdragon 7 Series	SD710, SD730, SD730G
Snapdragon 6 Series	SD660, SD665, SD630/36, SD670
Snapdragon 4 Series	SD427, SD429, SD439, SD 450
Compute	835, 850, 8cx
IOT	QCS8053, QCS8009, QCS603, QCS605
XR	XR1
Auto	820A, 3rd Gen Qualcomm® Snapdragon Automotive Cockpit Platforms
Voice and Music	QCS400

Style Transfer

Gesture Recognition

Audio Context Detection

Text Recognition

Personalized Avatars

Video Summarization

Computational Photography

Speech Synthesis

AR/VR

Gestures/Hand Tracking

Speaker Recognition

Ultrasound based virtual sensors

Super Nightshot

Text to Speech

Security & Liveness

Face Detection & Recognition

3D Face Payment

ASR

Language Translation

Single & Dual Camera Bokeh

Artistic Style Transfer

Voice Biometrics

Video Style Transfer

Noise Suppression

Object Location

Super Resolution

Natural Language Understanding

Super IQ

Low Light Bokeh

Multi-keyword Detection

Landmark Detection

OIS

Super HDR

Spatial Audio Detection

RAW HDR/HDR +

Diagnostics & Power

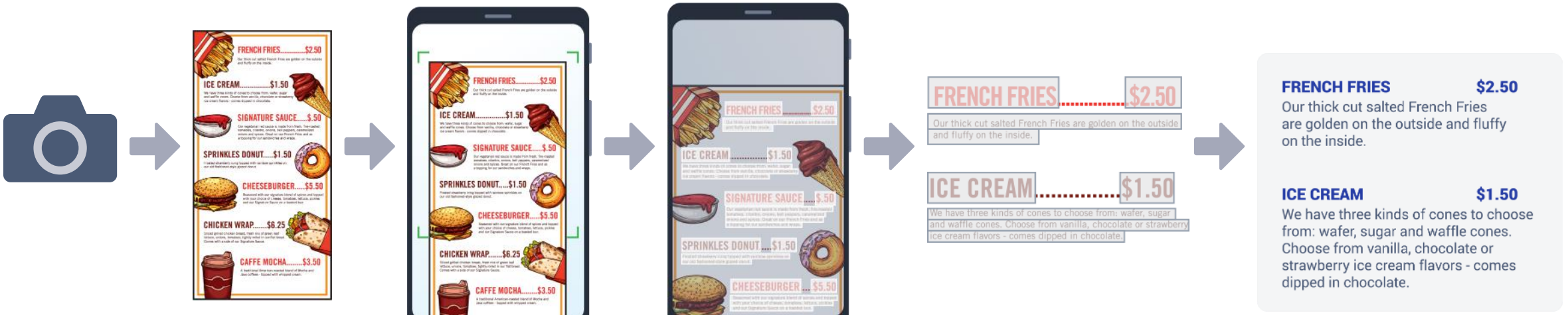
Echo Cancellation

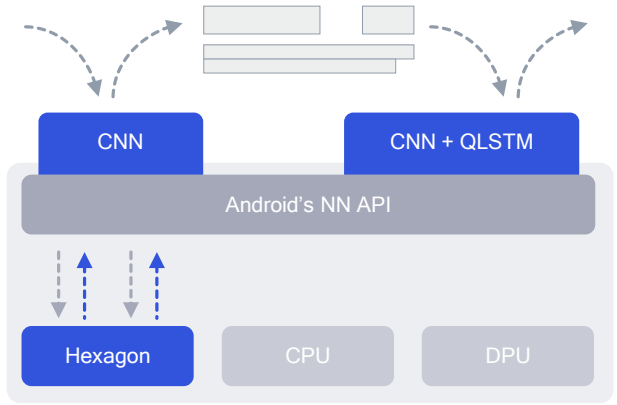
AI-based UI

Scene Classification

Key AI use cases

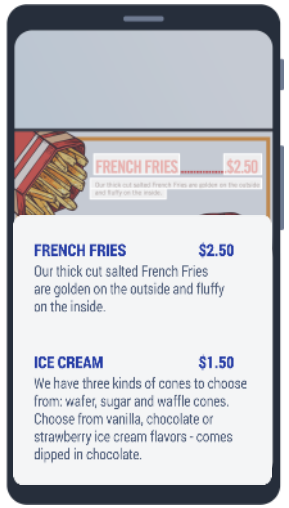
Google Lens Dense Text Copy





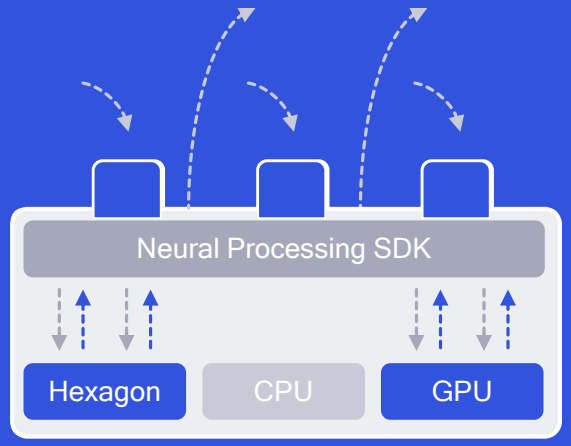
Qualcomm AI Engine

FRENCH FRIES **$2.50**
Our thick cut salted French Fries are golden on the outside and fluffy on the inside.



XR gaming

AI improves gaming experience on the device



Qualcomm AI Engine

Control
- Hands tracking
- Eye tracking
- Body motion
- Object tracking

Display
- Visual style transfer

Audio
- AI audio effects
- AI placed audio

Mechanics
- AI opponent
- AI game assistant

Hearing
- No howling
- Speech recognition & NLP
- Your voice selection

XR World
- People/object detection
- Scene segmentation
- Pose estimation
- Intent detection

AI “On” and “In” the device

A new development paradigm
where things repeatedly improve

Qualcomm snapdragon

with AI Engine

AI software ecosystem

Camera

- 商汤 sensetime
- Face++ 旷视
- ArcSoft
- Thundercomm
- Morpho
- anyvision. A BETTER TOMORROW
- nalbi
- polarr
- LOOM.AI
- Mobius Labs

Audio / Translation

- 科大讯飞 iFLYTEK
- AISPEECH
- 有道 youdao
- ELEVOC 大象声科 | 听得见的智能
- trio.ai 三角兽

Gesture

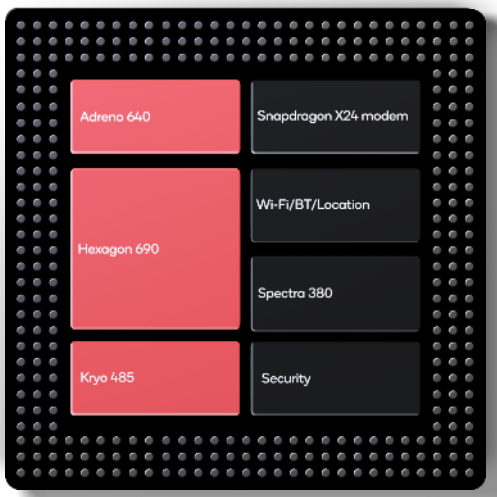
- ellipticlabs
- Qeexo

Augmented reality

- 網易 NetEase www·163·com
- 商汤 sensetime

Automotive

- COGENTEMBEDDED
- PATHPARTNER
- JUNGO CONNECTIVITY
- Zong Mu



Qualcomm
AI Engine

Frameworks

TensorFlow

PYTORCH

ONNX

Caffe2

mxnet

PaddlePaddle

Chainer

Cognitive
Toolkit

OS

Windows

Ecosystem

商汤 sensetime

Face++ 旷视

有道 youdao

ArcSoft

Thundercomm

Uncanny Vision

ellipticlabs

ANYVISION
A BETTER TOMORROW

網易 NetEase
www.163.com

有态度的门户

Nalbi

AISPEECH

ELEVOC
大象声科

科大讯飞 iFLYTEK

Tencent 腾讯

Baidu 百度

Google

amazon

Alibaba Group
阿里巴巴集团

Microsoft

Features

Noise Suppression

Super Resolution

Night Shot

Face Recognition

Speech Recognition

Object Detection

Video Segmentation

Bokeh

Devices

April 9, 2019 @qualcomm San Francisco, CA

Qualcomm

AI for IoT

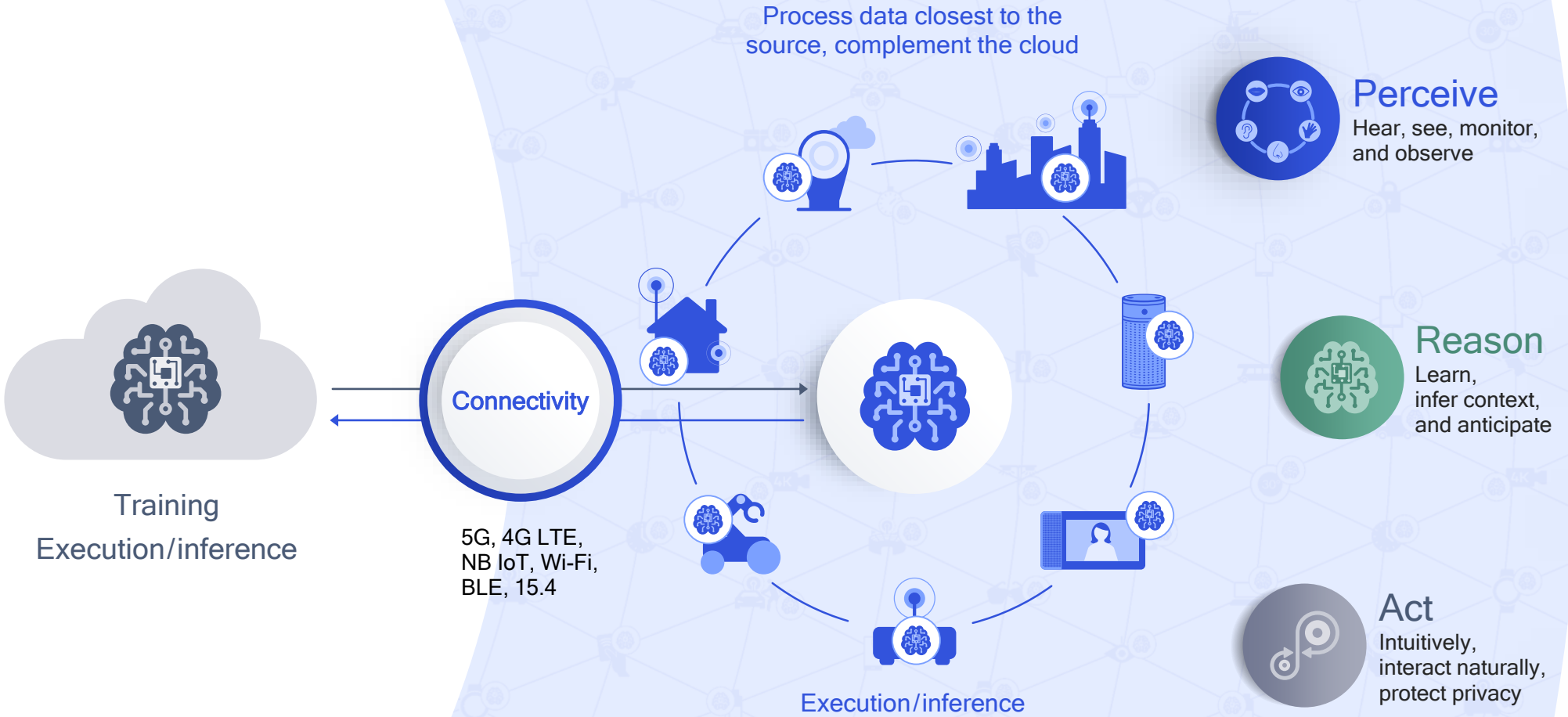
Sahil Bansal

Sr. Director, Product Management
Qualcomm Technologies, Inc.

Emerging growth and opportunities for AI in IoT

On-device AI and IoT

On-device computing is paramount



Perceive
Hear, see, monitor, and observe

Reason
Learn, infer context, and anticipate

Act
Intuitively, interact naturally, protect privacy

- Privacy
- Reliability
- Immediacy
- Efficiency
- Personalization


Realize low latency
New experiences
Customization
Extract local value
Content closer to user
Privacy
Immediacy
Efficiency
Reliability
Personalization
Edge cloud
On-device
To scale,
intelligence must be distributed
to the wireless edge, on the device
Billions of connected things
Massive amount of data
Security and privacy

Combination of 5G and compute is essential for IoT

Compute

Powerful processing
On-device CPU / DSP / GPU,
computer vision, audio, sensing,…

Artificial intelligence
Efficient machine learning,
on-device intelligence

Edge services/cloud
Data privacy, low latency,
local services, device management

Low power

Security

Connectivity

5G

Gigabit LTE | 5G NR | LTE IoT

Mesh | Wi-Fi | BLE

Cloud services

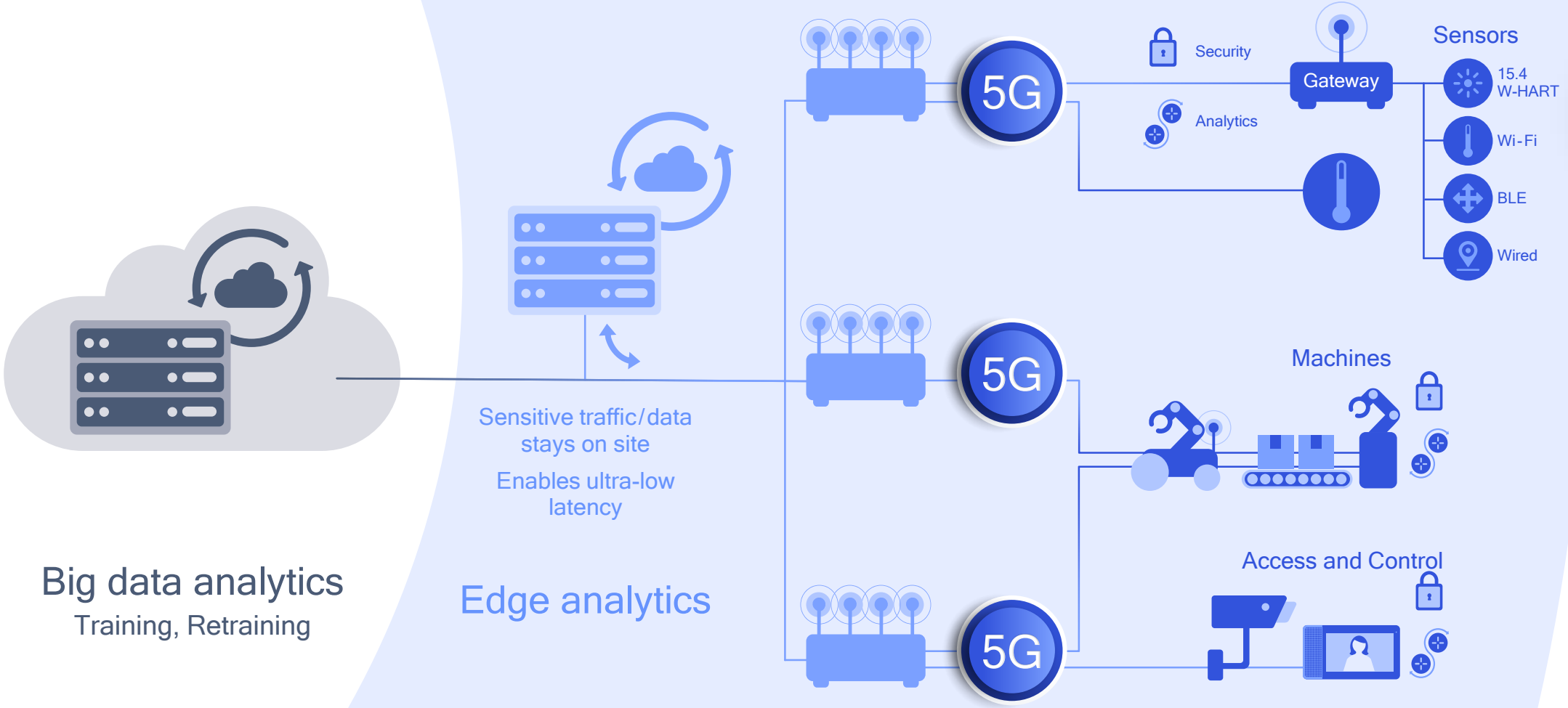
Cloud analytics and virtualized core network functions

Local network

Local gateway and/or local core network functions

On-device

Various degree of on-devices capabilities



Big data analytics

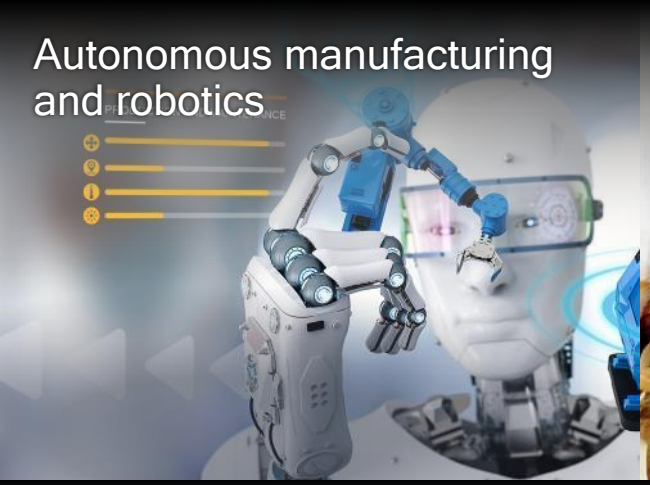
Training, Retraining

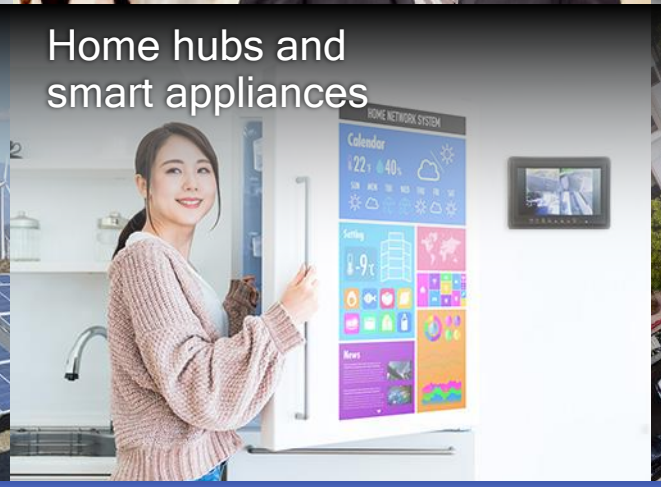
Sensor fusion

Computer vision

On-device intelligence

Low-power processing





AI for IoT across the home, industrial/enterprise, and Smart Cities

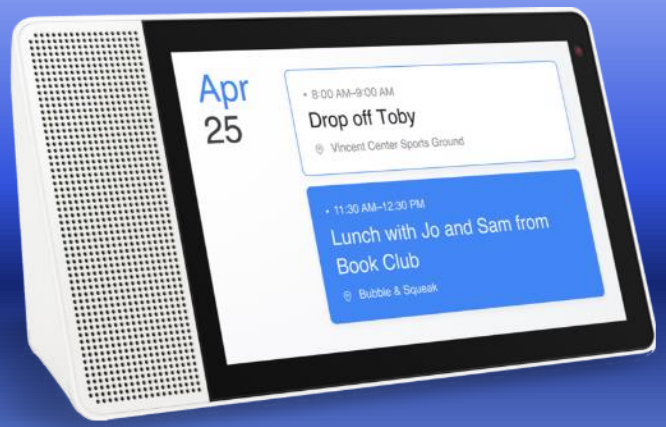
IoT devices are already available

Security & monitoring

Home security camera

Autonomous navigation for home vacuums

Voice recognition and identification



The challenge of AI workloads

- Compute intensive
- Large, complicated models
- Complex concurrencies
- Real-time
- Always-on

Constrained device environment

- Thermally efficient for sleek designs
- Requires long battery life for all-day use
- Storage/memory bandwidth limitations

Power and thermal efficiency

Critical to the promise of AI on a wide range of connected devices

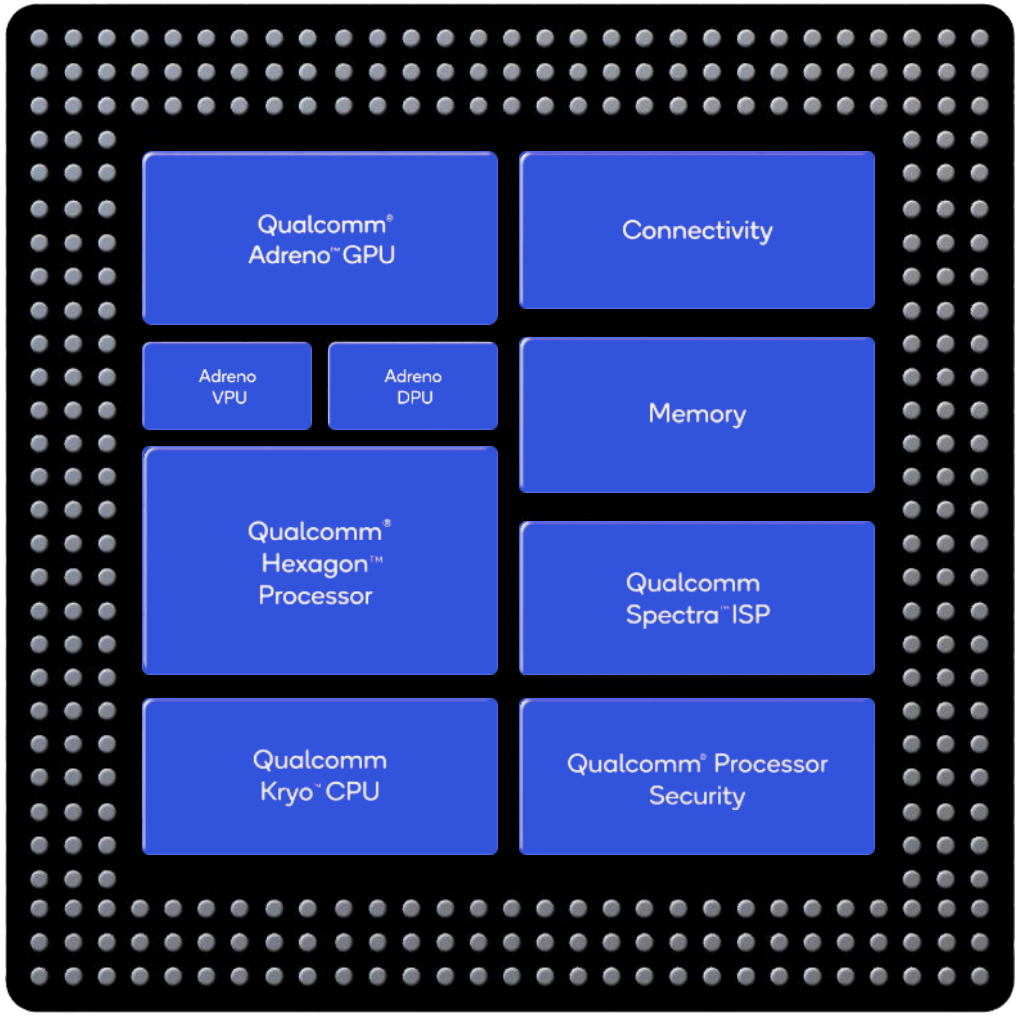
Qualcomm QCS605 is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.

Heterogeneous computing key for on-device intelligence

Designed to deliver performance and efficiency improvements

Broad portfolio of SoCs addresses different levels of performance and price points

Entire SoC is used

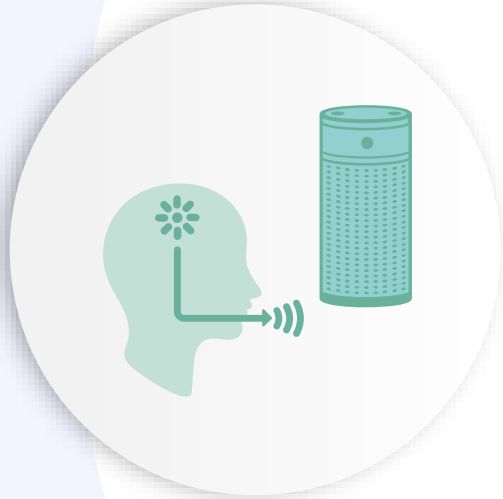


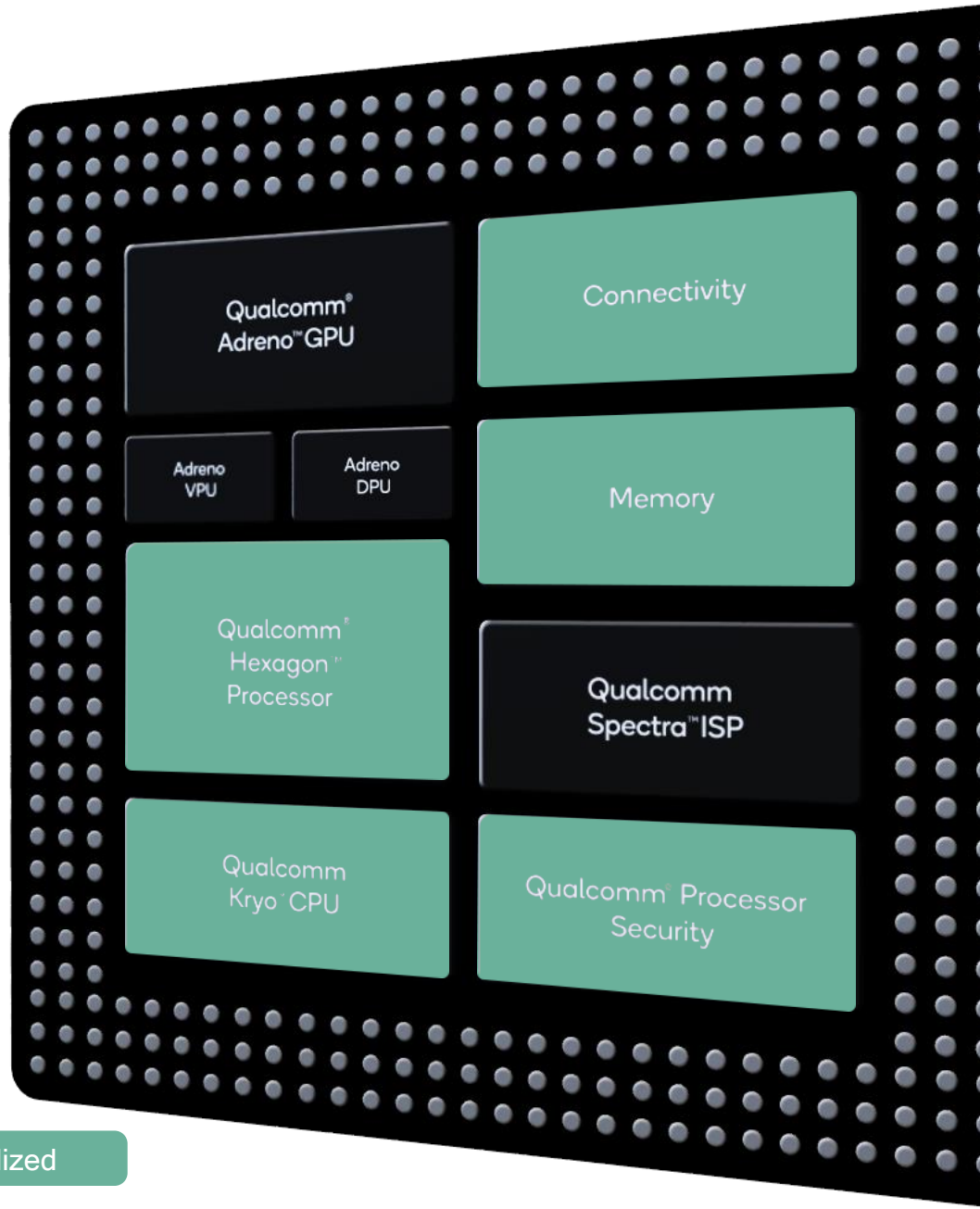
High-utilization

Graphic not to scale

Smart Speaker

- Voice Recognition and ID
- AI-based speech recognition
- Far-field voice, beam-forming and echo cancellation
- Support for cloud-based voice assistants
- Wi-Fi, BLE connectivity



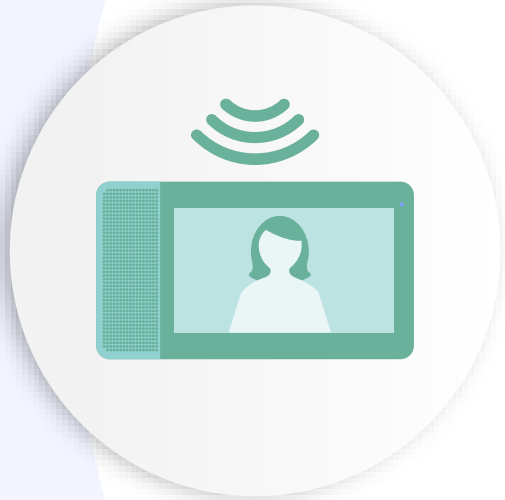


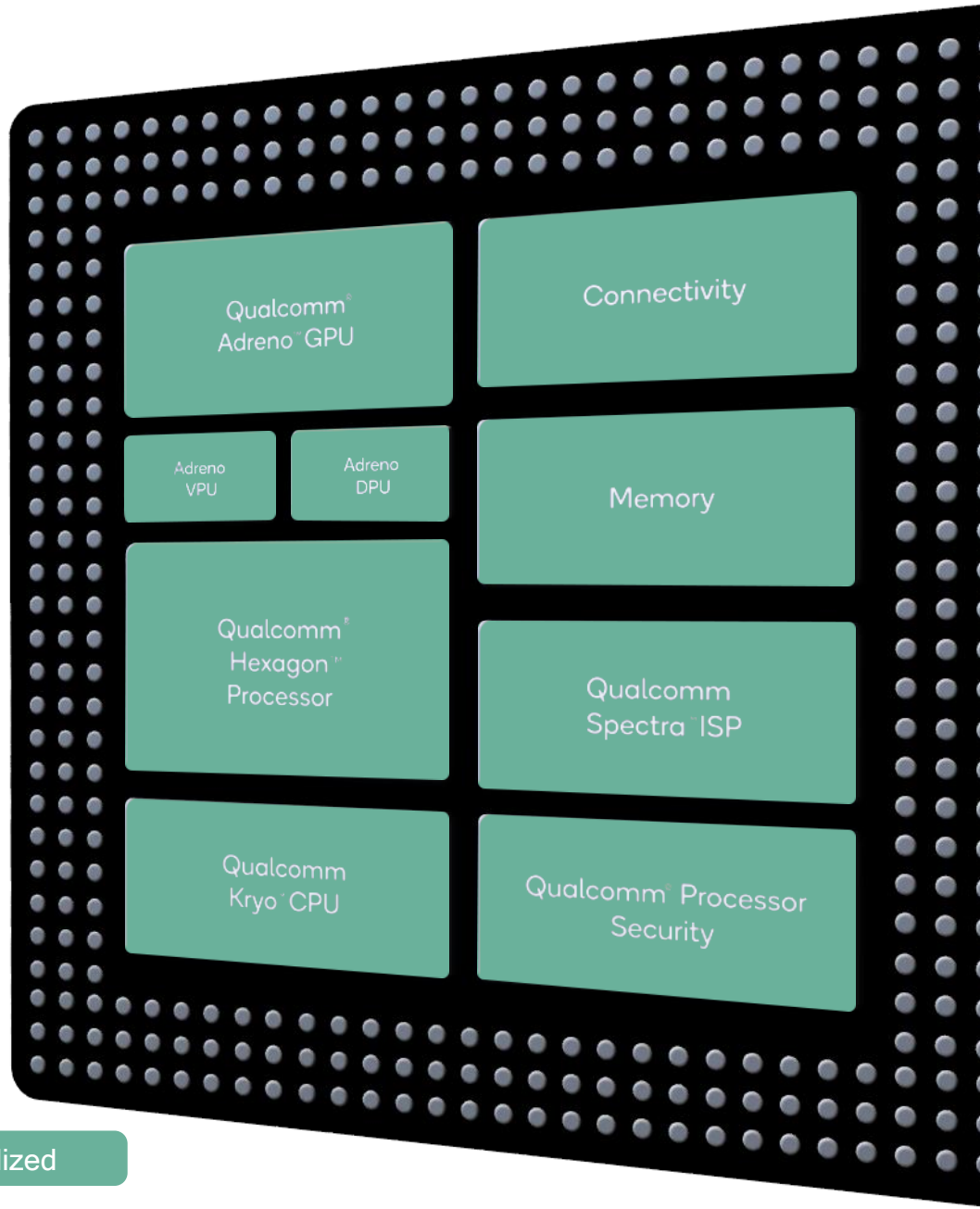
Blocks Utilized

Graphic not to scale

Smart Display

- Face/family recognition and audio sound recognition
- HD Display with 4K encode/ decode, video telephony
- Face detection object tracking
- HW-based acceleration support for popular neural networks



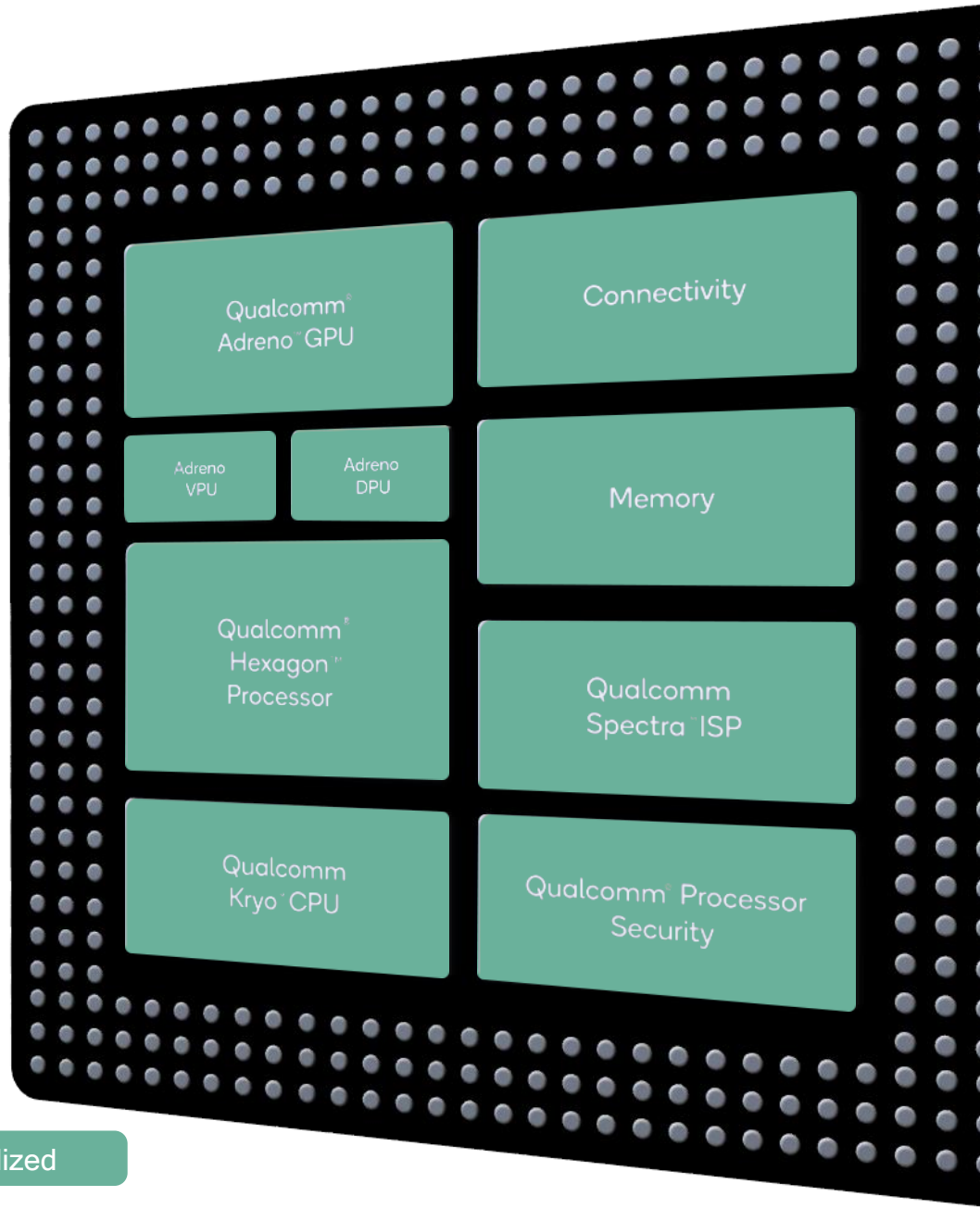


Blocks Utilized

Graphic not to scale

Enterprise IPC Camera

- People and object recognition
- Dual ISPs with advanced IQ - sHDR and TNR for low light
- Upto 4K encode/decode @60fps HEVC video
- Support for popular neural networks
- Wi-Fi, Ethernet, Cellular (4G, 5G) Connectivity



Blocks Utilized

Graphic not to scale

Qualcomm Vision Intelligence Platform

Ecosystem Partners

altek anyvision. A BETTER TOMORROW INTRINSYC LITEON® Microsoft PILOT.AI ACCELERATED INTELLIGENCE Qisda SAST. Thundercomm

Qualcomm Vision Intelligence Platform

+

Microsoft Azure IoT Edge

http://www.visionaidevkit.com

Run AI models on the edge with Qualcomm AI Engine or utilize the cloud

Create, deploy and manage your models in the cloud and the edge with Azure ML and Azure IoT Edge

Qualcomm® QCS400 series

for smarter audio

High-performance

Power optimized

Tightly integrated

Enhanced with AI

April 9, 2019 @qualcomm San Francisco, CA

Qualcomm

Transforming automotive with AI

Nakul Duggal

SVP, Product Management – Automotive
Qualcomm Technologies, Inc.

Strong asset: Market fit

Qualcomm's unique assets enable accelerated innovation and end-to-end system integration in automotive

3G/4G LTE
5G/C-V2X

RF
RF

CPU
CPU

GPU
GPU

DSP
DSP

Multimedia

Power
Management

Software/HLOS

Computer
Vision

Machine
Learning/AI

Position
Location

Wi-Fi /
DSRC

BT
Bluetooth

Ethernet

PLC

Security

EV Charging
PLC

Telematics • Connectivity • V2X • Digital Cockpit • Autonomous Driving

World’s leading automakers build with our solutions

ACURA · Audi · BMW · Buick · BYD · Cadillac · Chevrolet · Chrysler · Dodge · Ford · Geely · Honda · Hyundai · Infiniti · Jaguar · Jeep · Kia · Land Rover · Lexus · Lincoln · Mercedes-Benz · MINI · Nissan · Opel · Porsche · Peugeot · Renault · Rolls-Royce · smart · Subaru · Toyota · Volvo · Volkswagen

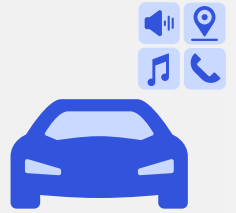
#1
In telematics and Bluetooth for automotive

A leader
In premium next-gen infotainment design-wins for production vehicles starting 2019-2020

18
Automakers have selected the Qualcomm® Snapdragon™ Automotive Infotainment Platform

$5.5B+
Design-win pipeline for telematics, infotainment, and in-car connectivity

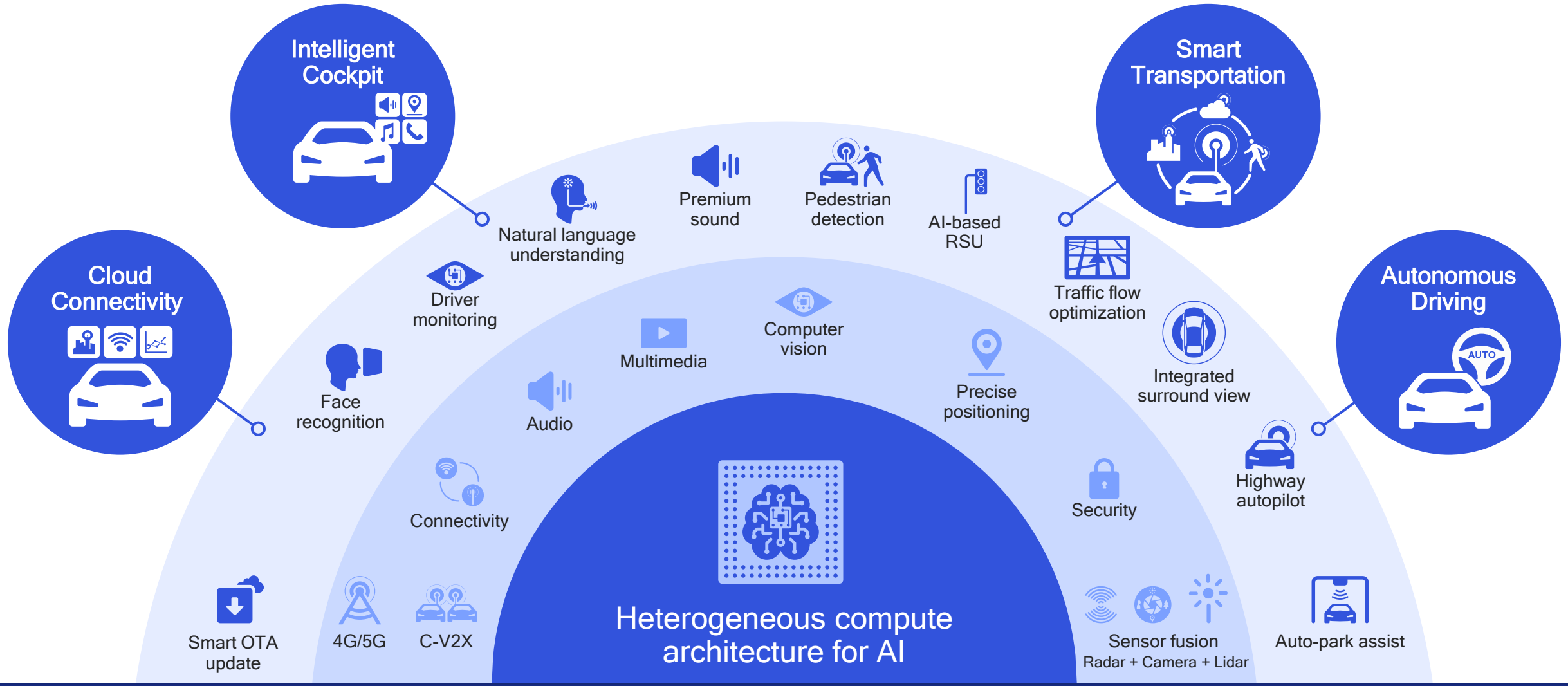
Telematics



Infotainment

In-car connectivity

Source: Company data
Qualcomm Automotive Infotainment Platform is a product of Qualcomm Technologies, Inc. and/or its subsidiaries.



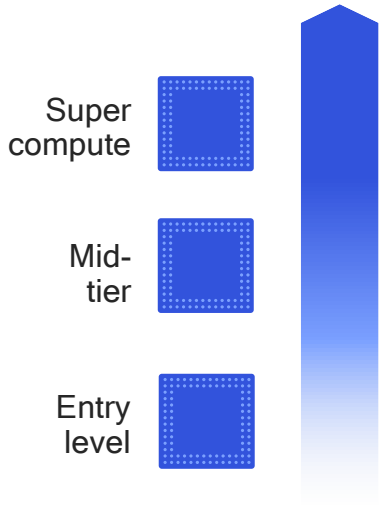
Artificial Intelligence is transforming automotive and the entire transportation industry

Our vision of scaling AI for different tiers and markets

Within the power and thermal constraints of different segments

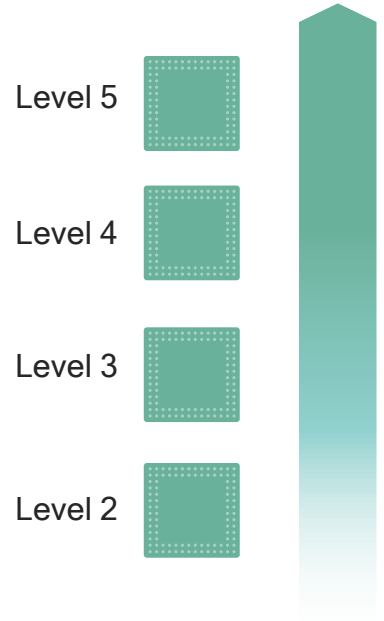
Automotive Cockpit

Scaling AI-based cockpit platforms across all vehicle classes



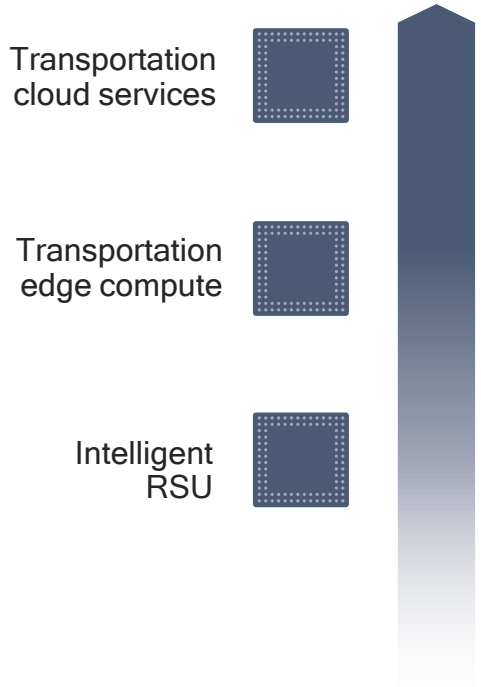
Autonomous Driving

Scaling AI for different autonomy levels



Smart Transportation

Scaling AI for smarter infrastructure and cloud services



Scaling AI for different automotive use cases

- Very compute intensive
- Large, complicated neural network models
- Complex concurrencies

Optimized for automotive use cases

- **Thermal efficiency**
  For cost efficiency (e.g. cooling systems)
- **Power efficiency**
  For longer EV battery range
- **Immediate response**
  For latency-sensitive, safety applications

Efficiently scaling AI for different use cases

Optimizing for maximum TFLOPS per Watt

Transforming in-vehicle experiences with on-board AI

Digital Cockpit – Scalable heterogeneous compute architecture

Qualcomm snapdragon

Mirror replacement display

AR/HUD display

Passenger display

Center display

Reconfigurable cluster display

Rear-seat display

Mirror replacement vision system

Foundational technologies

- Artificial intelligence
- Heterogeneous computing
- Computer vision
- Audio
- Multimedia
- Location
- Security
- Connectivity

Edge NLU based Voice Assistance

Natural language understanding

Face Detection Fingerprint Recognition

Driver Monitoring System

Virtual Assistant

Voice/speech recognition

Object classification

Camera Perception

Edge First AI Agent

Voice/noise cancellation

Scene understanding

AR based Navigation

Vision Enhanced Precise Positioning (VEPP)

Location Services

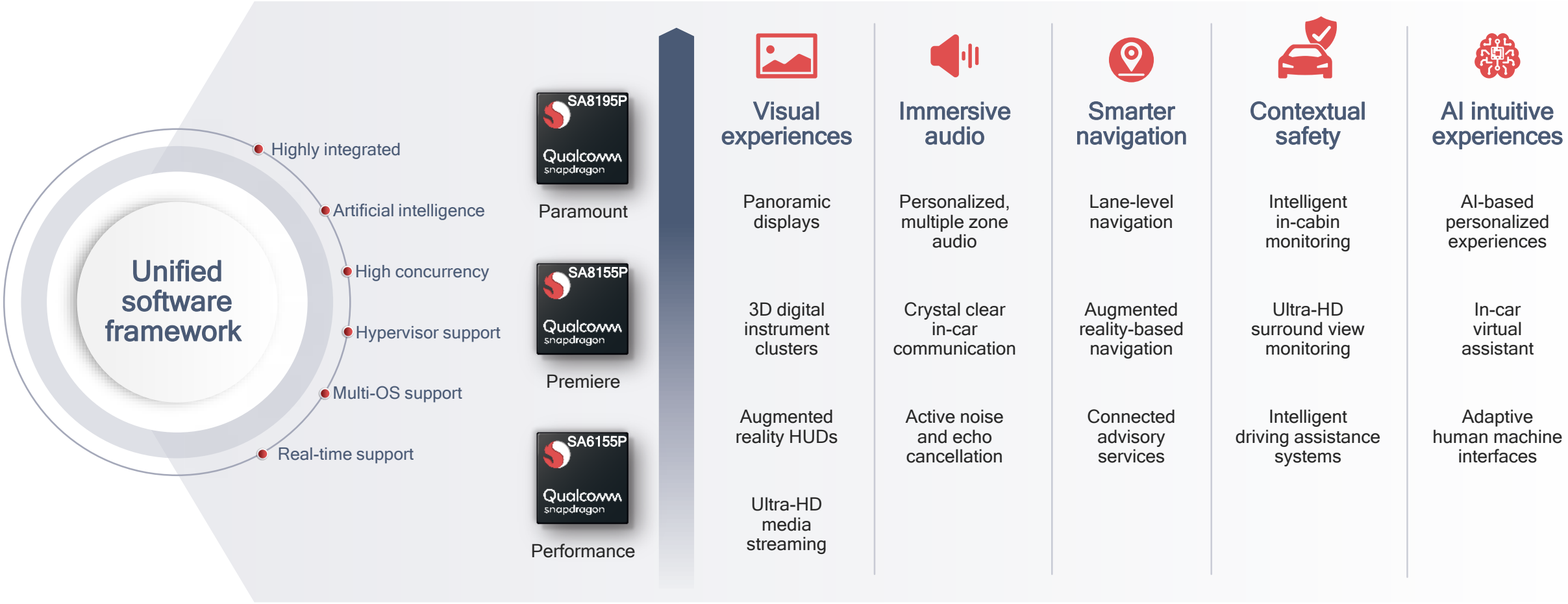
Sensor processing & fusion

Pedestrian/ Object Detection

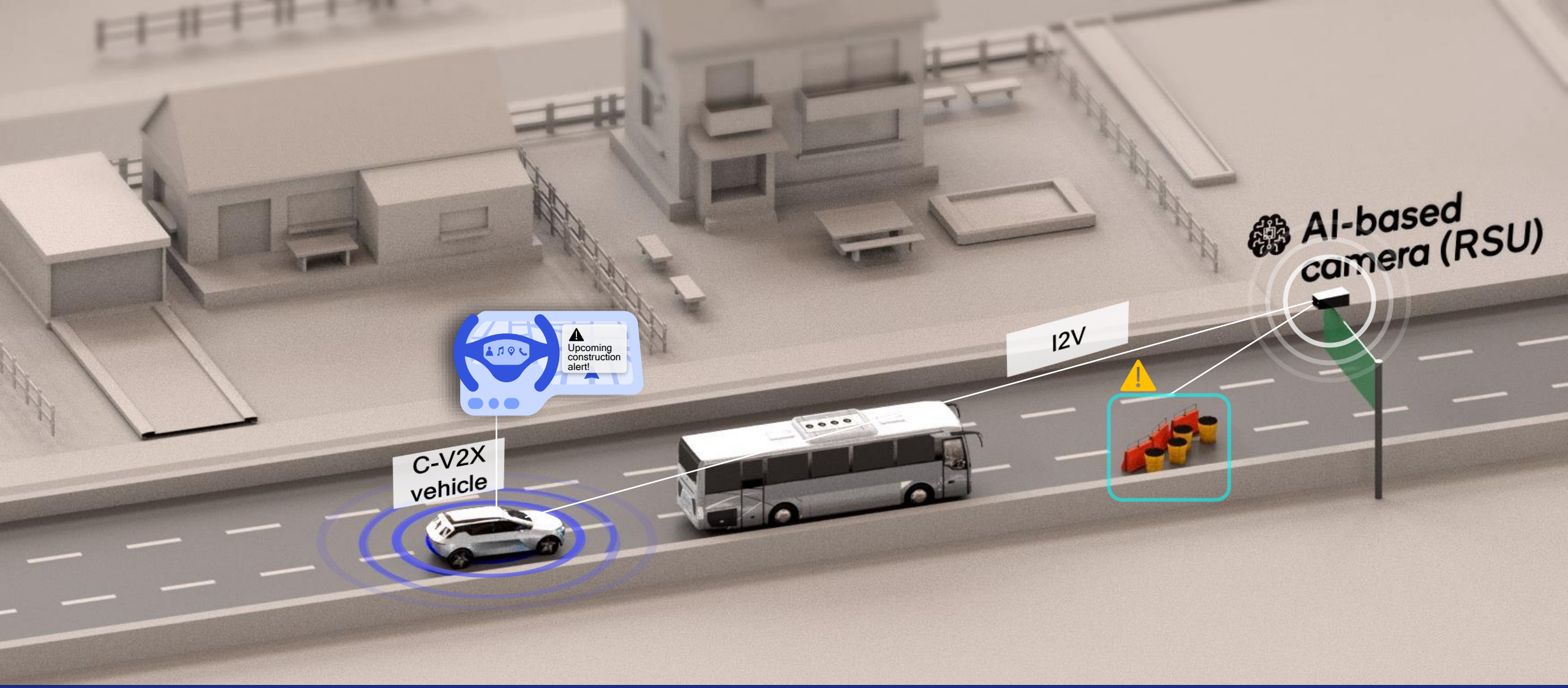
New on-board AI-based cockpit use cases will enrich the driving experience

Qualcomm Snapdragon automotive cockpit platforms, Gen 3

Introduced at CES 2019, bringing new levels of computing and intelligence to every tier



Shaping smart transportation



Shaping the Connected Freeway

Edge AI
AI-based RSU detects lane reconfiguration due to construction

C-V2X (I2V)
RSU send road world model to vehicle with lane reconfiguration

Localization
Vehicle determines its position and distance from road construction

Defining connected urban transport

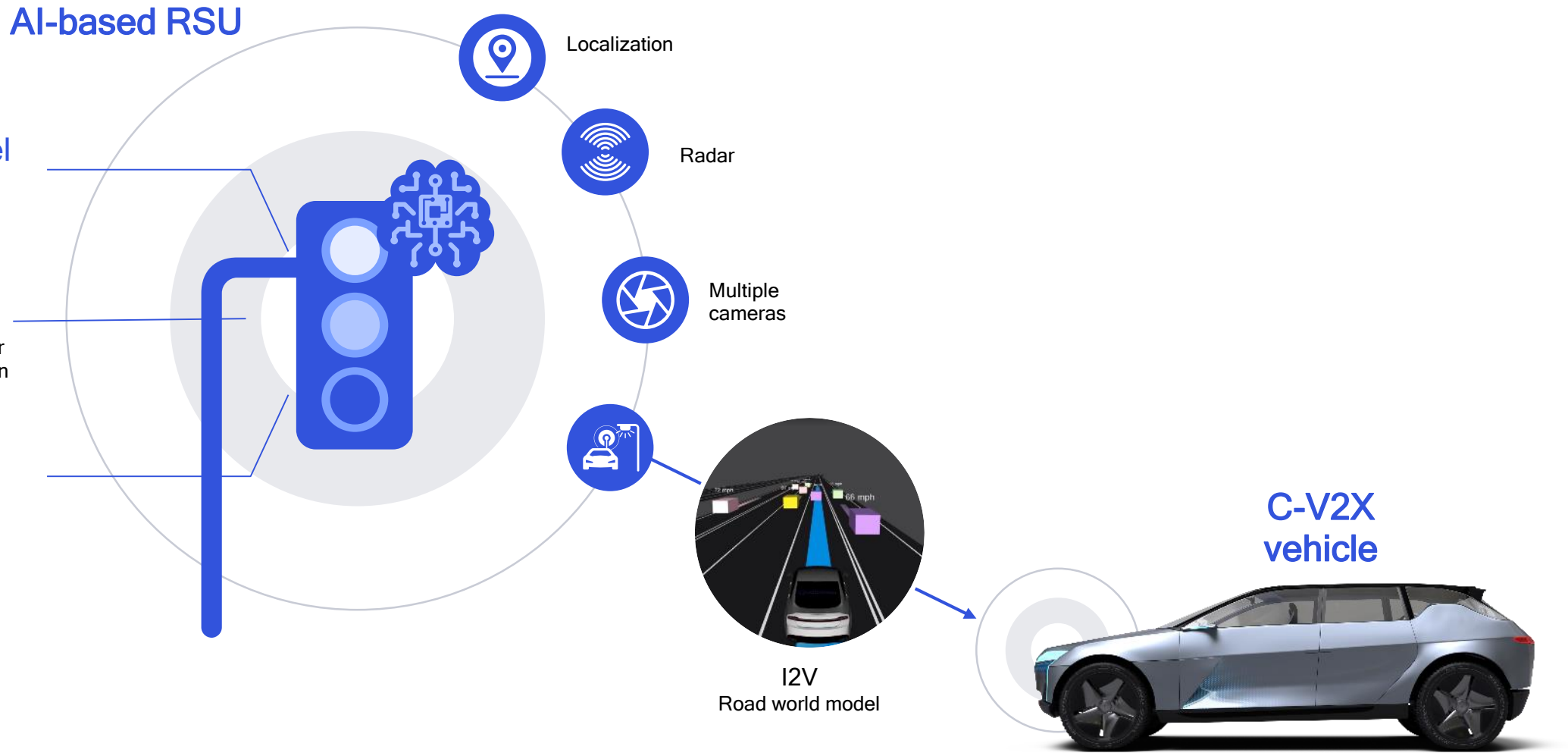
Edge AI
E.g. for detecting pedestrians or hazards

C-V2X (I2V)
E.g. send 3D HD map updates or hazard warning

Localization
E.g. for lane-level warning, and navigation

Smarter transportation infrastructure

Requires next level of compute and intelligence for perception and sensor fusion



Paving the road to autonomous driving

Perceive

Camera, radar sensors

CV2X, localization in maps extended horizon sensors

Low level sensor fusion

Plan

Behavior prediction

Behavior planning

Motion planning

Act

Actuation control

Drive-by-wire smooth maneuver

Connect

Tele-operations

Data analytics

Smart transportation

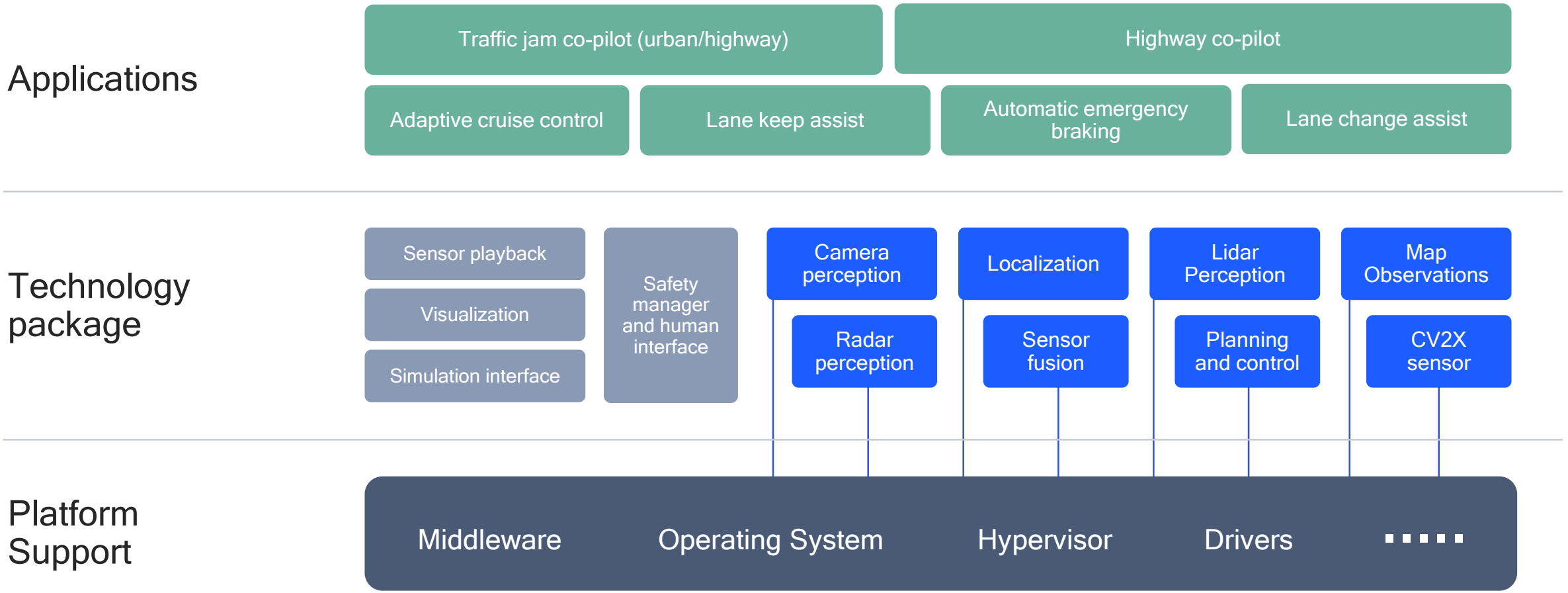
Simulator and tools

5G

Our system approach–autonomy stack

End-to-end system. Active sensing and extend horizon using connectivity and maps

Example of level 3 autonomy stack



- Detect and classify objects in radar signal and get their distance
- Using combination of CNN and LSTM

Radar Perception

Lidar Perception

- Detection for lane markers, drivable space vehicles, etc.
- Annotation for Deep Learning

- Detect and classify lane markers, drivable space, traffic sign, other vehicles, blinkers, etc.
- Using multiple CNN and LSTM-like (TAGM) networks on multiple cameras

Camera Perception

Behavior planning

- Reinforcement learning for safe and human-like driving behavior
- Prediction based on LSTM and convolutional networks

Applying different kinds of advanced AI techniques to handle autonomous driving-specific needs

Strong asset: Market fit

Qualcomm's unique assets enable accelerated innovation and end-to-end system integration in automotive

3G/4G LTE
5G/C-V2X

RF
RF

CPU
CPU

GPU
GPU

DSP
DSP

Multimedia

Power
Management

Software/HLOS

Computer
Vision

Machine
Learning/AI

Position
Location

Wi-Fi /
DSRC

BT
Bluetooth

Ethernet

PLC

Security

EV Charging
PLC

Telematics • Connectivity • V2X • Digital Cockpit • Autonomous Driving

April 9, 2019 @qualcomm_tech San Francisco, CA

Qualcomm

Leading research across the AI spectrum

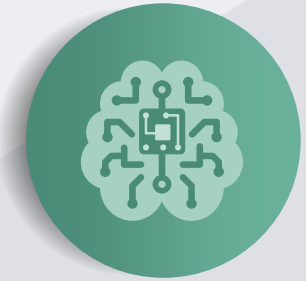
Rajesh Pankaj

SVP, Engineering
Qualcomm Technologies, Inc

Advancing research to make AI ubiquitous

Perception

Object detection, speech recognition, contextual fusion



Reasoning

Scene understanding, language understanding, behavior prediction

Action

Reinforcement learning for decision making

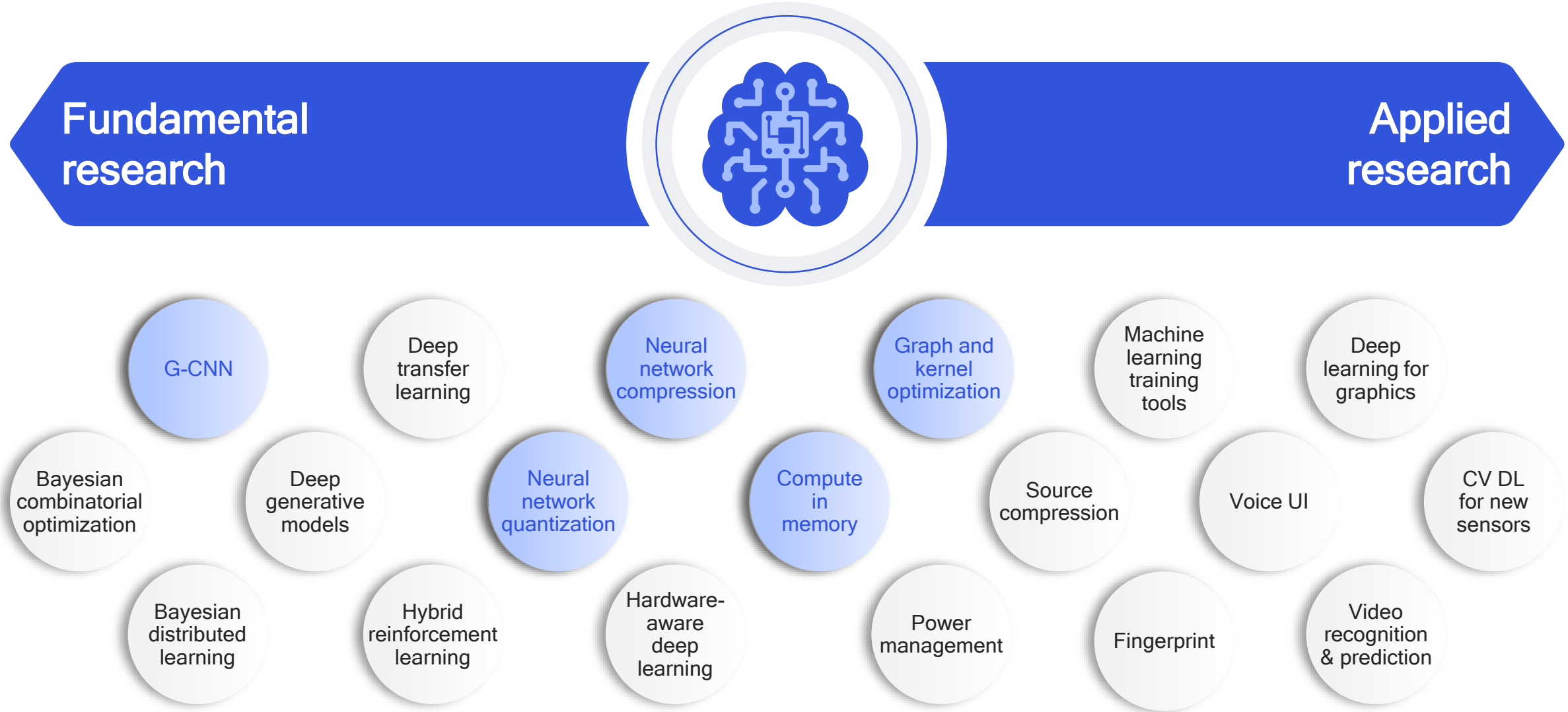
Power efficiency

Personalization

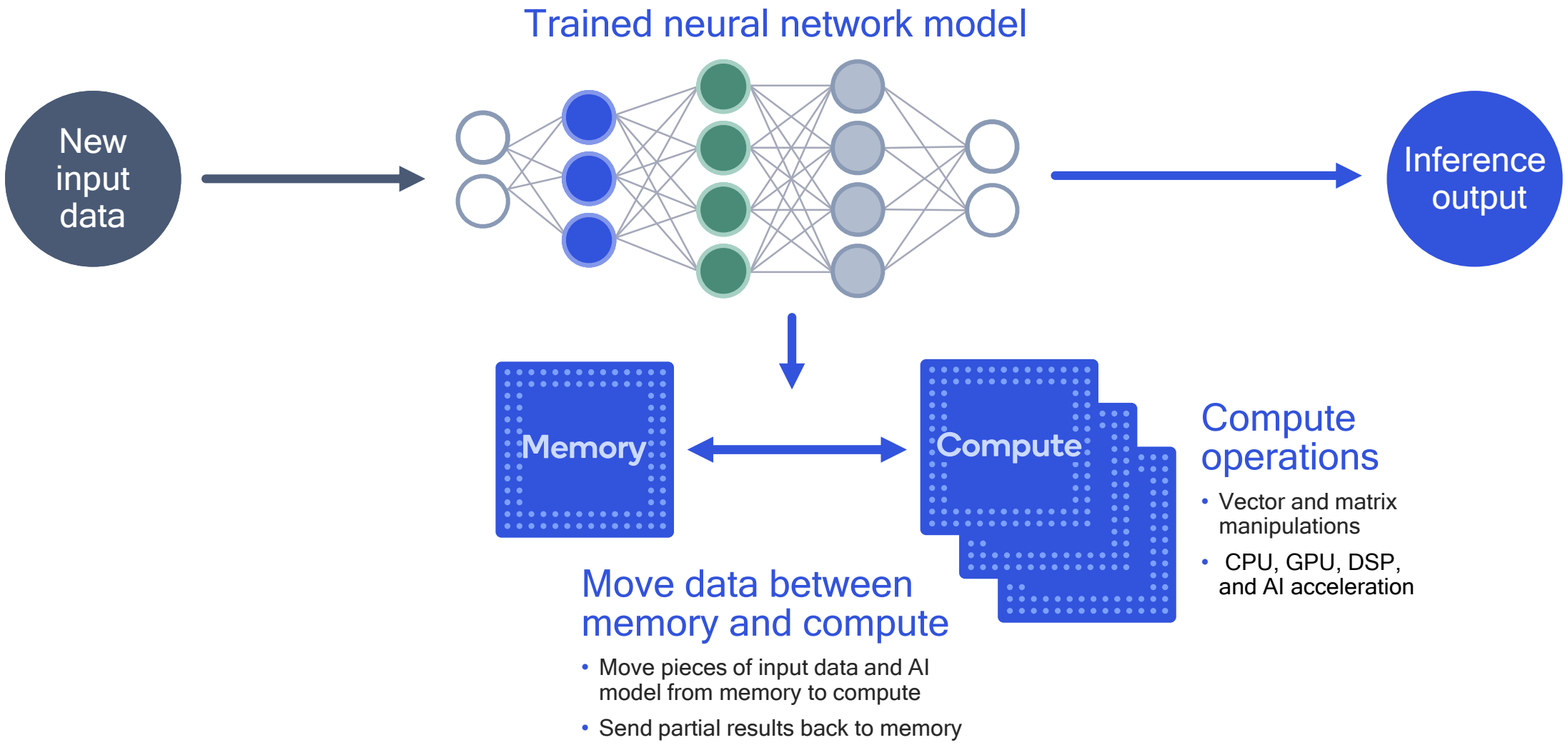
Efficient learning

We are creating platform innovations to scale AI across the industry

Leading research and development across the entire spectrum of AI



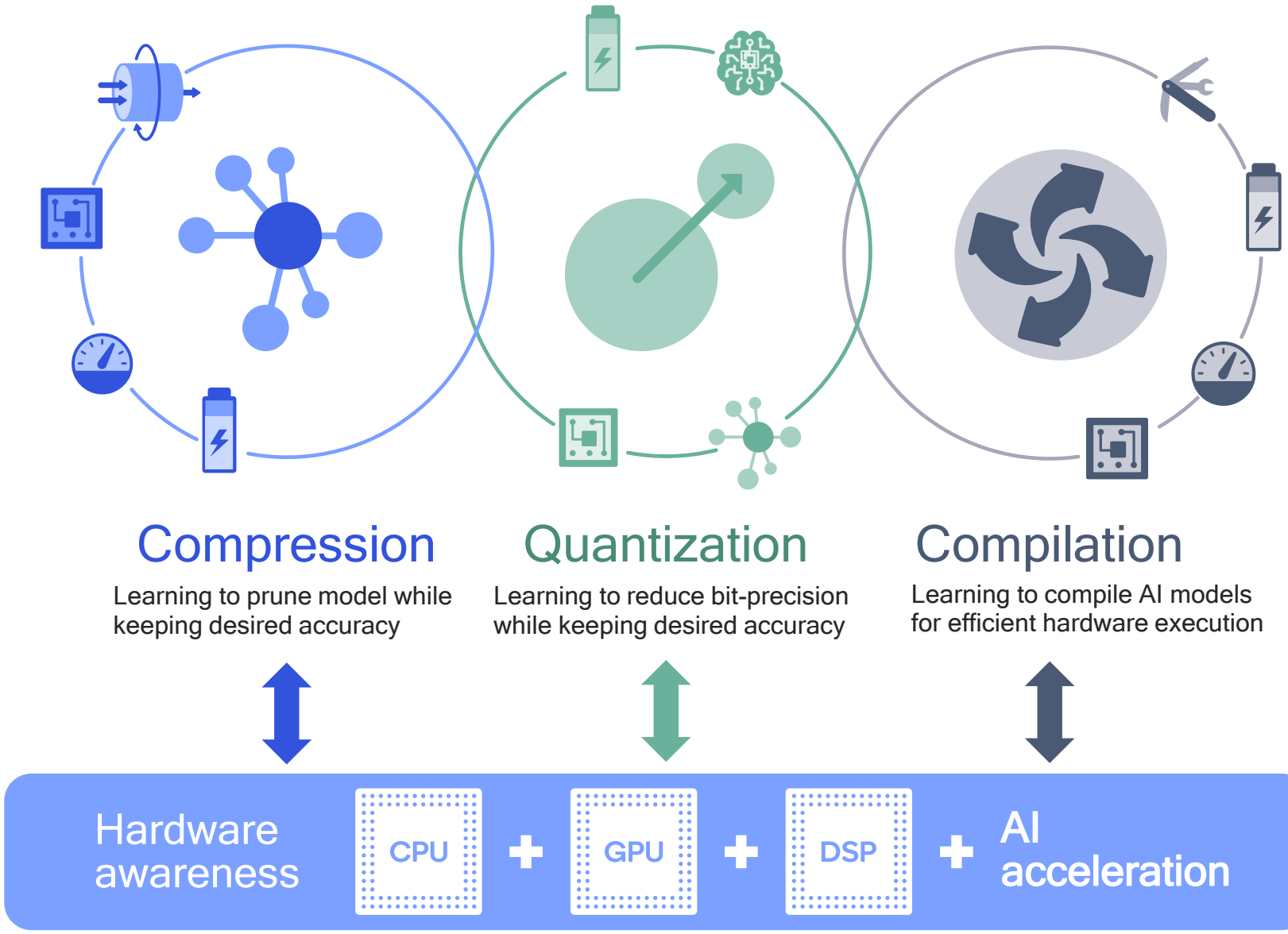
Advancing AI research to increase power efficiency



AI model optimization research for power efficiency

Applying AI to optimize AI models through automated techniques

Reduced time-to-market and engineering cost



Compression of AI model architectures

Automated removal of insignificant/redundant elements while maintaining accuracy

Tensor decomposition

Decomposing a single layer into two or more efficient layers

Spatial SVD

Channel pruning

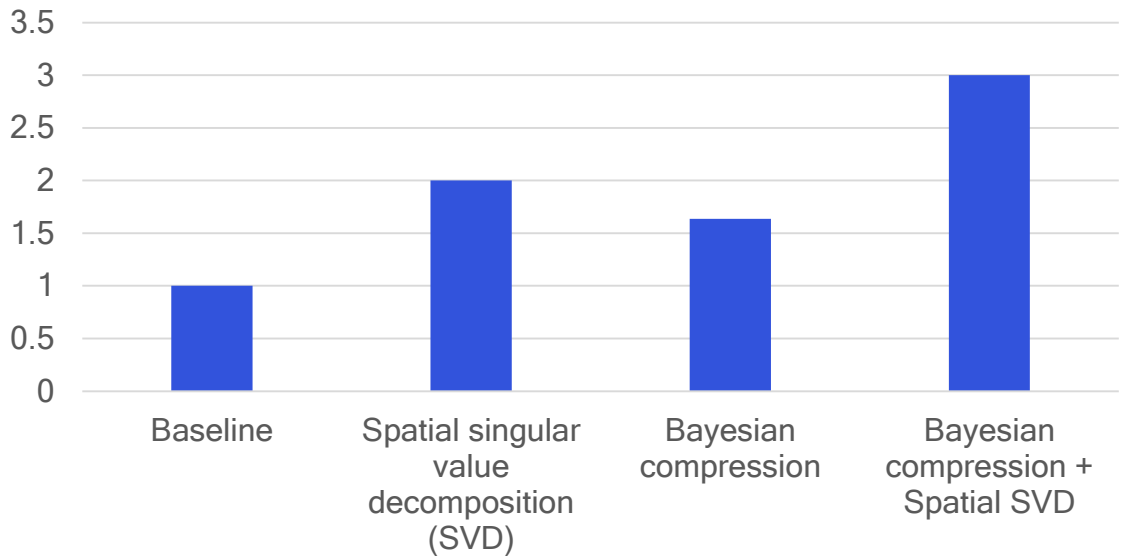
Removing channels from the network

L2 filter magnitude and Bayesian techniques

Hardware aware compression

3x

Compression with less than 1% loss in accuracy*

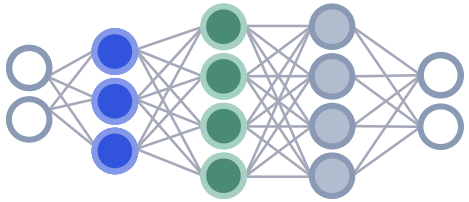


*: Comparison between baseline and compression with both Bayesian compression and spatial SVD. Example uses ResNet18 as baseline.

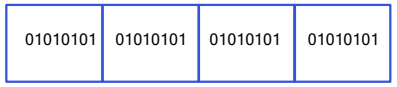
Quantization for power efficiency

Automated reduction in precision of weights and activations while maintaining accuracy

Models typically trained at high precision

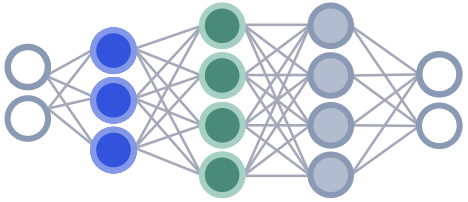


32-bit

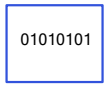


Floating point 3452.3194

Inference at lower precision



8-bit



Integer 3452

>4x

increase in perf. per watt from savings in memory and compute*

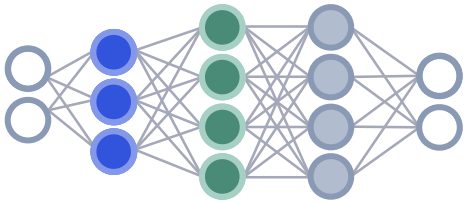
Promising results show that 8-bit AI models can become ubiquitous

Virtually same accuracy for FP32 and quantized INT8

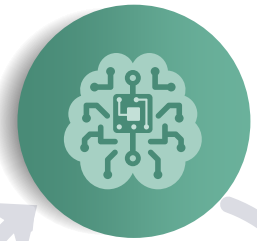
*: Compared to a FP32 model that is not quantized

Compiler research for efficient hardware usage

Reinforcement learning for automated HW compilation–as there are billions of potential configurations



Quantized and compressed AI model



AI optimization agent
Update model based on feedback

Schedule compilation[2]

Execute compiled code
On target hardware CPU, GPU, DSP, AI accelerators

Feedback: execution time

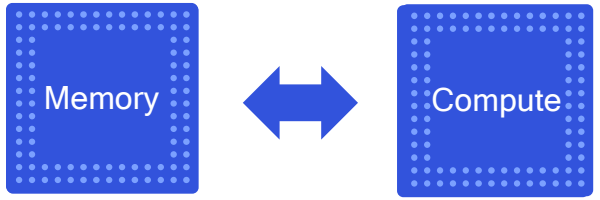
4x
Speedup improvement over TensorFlow Lite[1]

1) On average improvement of tested AI models
2) Schedule kernels and graphs, tile size, reorder, unroll, parallelize, vectorize,…

AI hardware acceleration research

Example: compute-in-memory AI research

- Analog compute
- New memory design
- Need low bit-width AI models



Traditional computer architecture

- Compute and memory are separate and data has to be shuffled back and forth
- Good for general purpose operations

Compute-in-memory

- Computations, like add and multiply, are done in memory
- Good for simple math operations and when memory becomes bottleneck

10-100x
Power efficiency improvement for 1-bit ops*

A paradigm shift from traditional computer architecture can bring orders of magnitude increase in power efficiency

* Compared to traditional Von Neumann architectures today

Can we apply foundational mathematics of physics, like quantum field theory, to deep learning?

Qualcomm

We are advancing AI research to make AI power efficient

We are conducting leading research and development across the entire spectrum of AI

We are creating AI platform innovations that are fundamental to scaling AI across the industry

Qualcomm

Thank you

Follow us on:

For more information, visit us at:

www.qualcomm.com & www.qualcomm.com/blog