May 22, 2019 @qualcomm_tech Berlin, Germany

Qualcomm

Combining AI and 5G for new user experiences

Thomas Stockhammer

Director Technical Standards
Qualcomm Technologies, Inc.

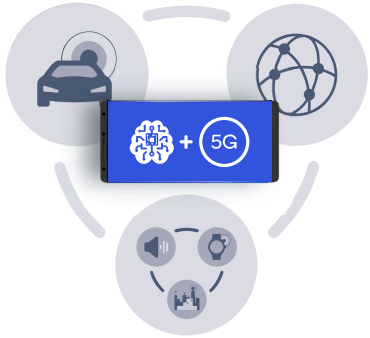
Leading mobile innovation for over 30 years

Digitized mobile communications

Analog
to digital

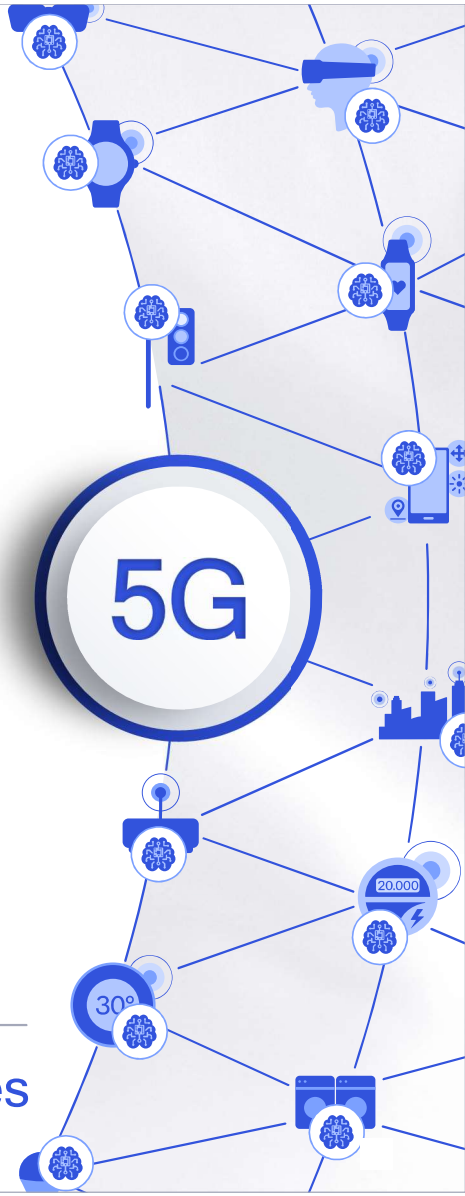
Redefined computing

Desktop to
smartphones

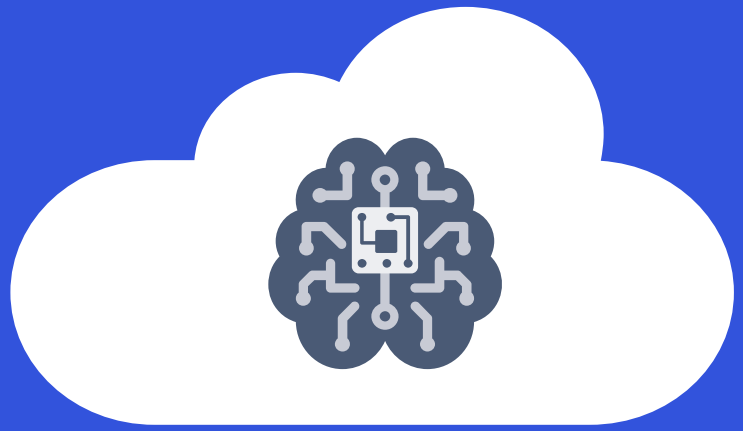


Transforming industries

Connecting virtually everything
at the wireless edge



Transforming how the world connects, computes and communicates



Today,

intelligence is primarily associated with the cloud

Mobile is becoming the pervasive AI platform

~7.8 Billion

Cumulative smartphone unit shipments forecast between 2018-2022

Source: IDC Aug. ‘18

Mobile computing

Smart homes

Smartphones

Industrial IoT

Wearables

Mobile scale
changes everything

Bringing AI to the masses

Qualcomm

Smart cities

Automotive

Healthcare

Networking

Extended reality

Rapid replacement cycles | Superior scale | Integrated/optimized technologies


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
Trillions of connected things
Massive amount of data
Security and privacy

On-device intelligence is paramount

Perceive
Hear, see, monitor, and observe

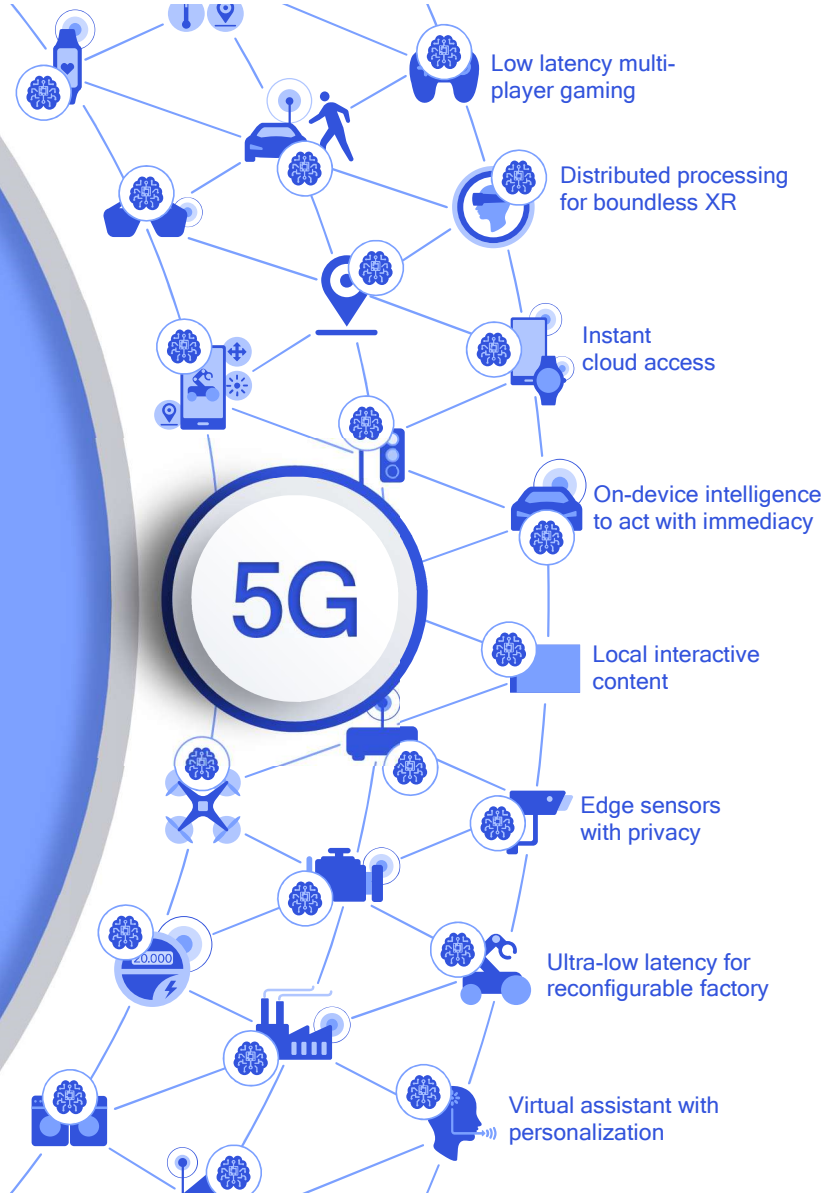
Reason
Learn, infer context, and anticipate

Act
Intuitively, interact naturally, protect privacy

- Privacy
- Reliability
- Immediacy
- Efficiency
- Personalization

The wireless edge transformation realizes the full potential of 5G

- New experiences with new levels of immersion, immediacy, personalization and privacy
- Creating new industries and transforming existing industries in the new era of distributed autonomy
- Essential on-device capabilities augmented with processing/compute, content, control,… at edge cloud



Enhancing services, creating new services, new deployment models with our end-to-end expertise

Cloud

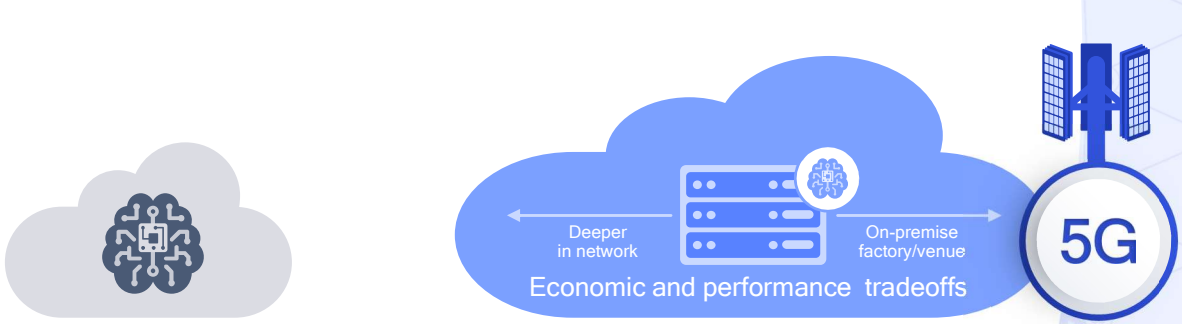
Big data, AI training, less delay sensitive content, storage,…

Augmented by edge cloud

Compute/processing, content, control, storage,.. closer to user[1]

Driving the best possible on-device capabilities

Sensing, processing, security, intelligence



Latency could be over 100s ms today

Latency as low as 1 ms

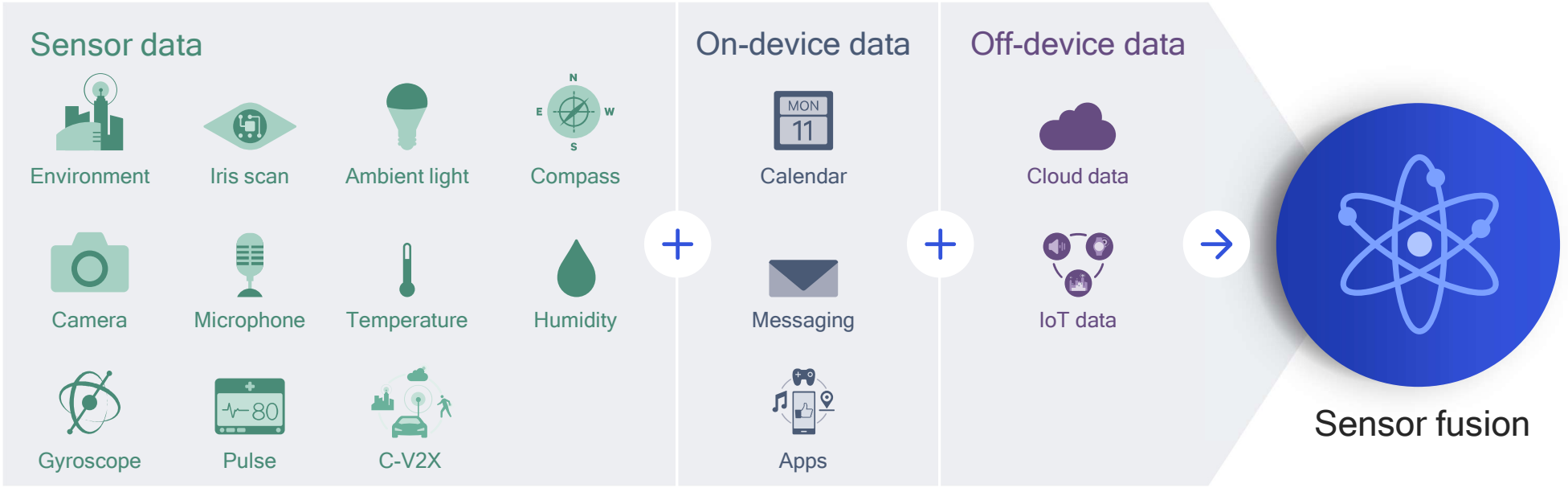
Cooperation between operators, infra vendors, OEMs, content providers, cloud providers,…

- Realize 5G's low latency
- Distributed/split functions over 5G
- Extract value – vs. 'dumb' data pipe
- Customization, local value
- New deployments, like private networks

1. Such as distributed/virtualized core, distributed packet gateway functionality for low latency, mobile edge compute, related to MEC Multi Access Edge Computing as defined by ETSI

Achieving personalization through contextual intelligence

The fusion of many types of sensors and personal information



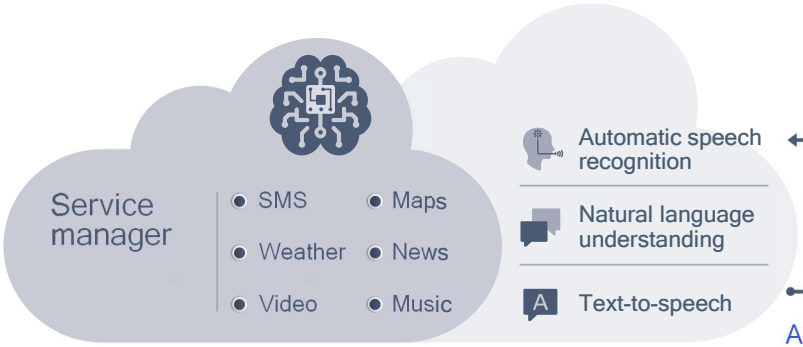
Low power sensing, processing, and connectivity

- Efficient, heterogeneous architectures
- Sensor fusion and machine learning
- Integrated, always-on data capturing
- Low-energy wireless technologies (e.g. BT-LE, 5G NR IoT)

On-device AI creates a true personal virtual assistant

Cloud-centric (today)

Less privacy and personalization

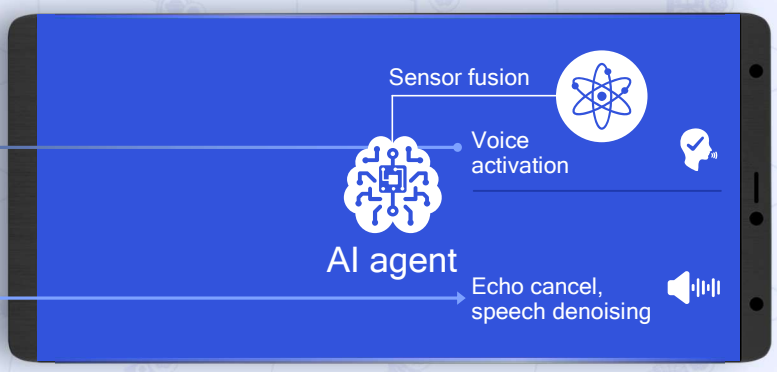


5G

Access to rich content

On-device (future)

Continuously learns personal knowledge and acts intuitively with immediacy



Both ends are needed – but AI functions gradually moving on to the device

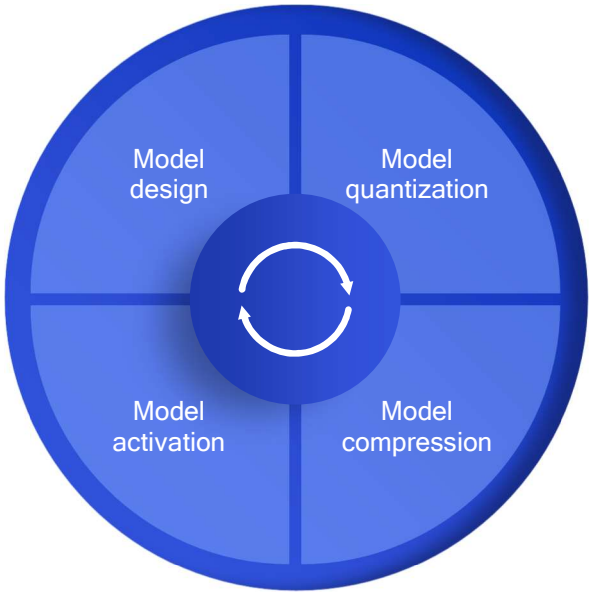
Qualcomm Technologies is making on-device AI ubiquitous

Efficient hardware

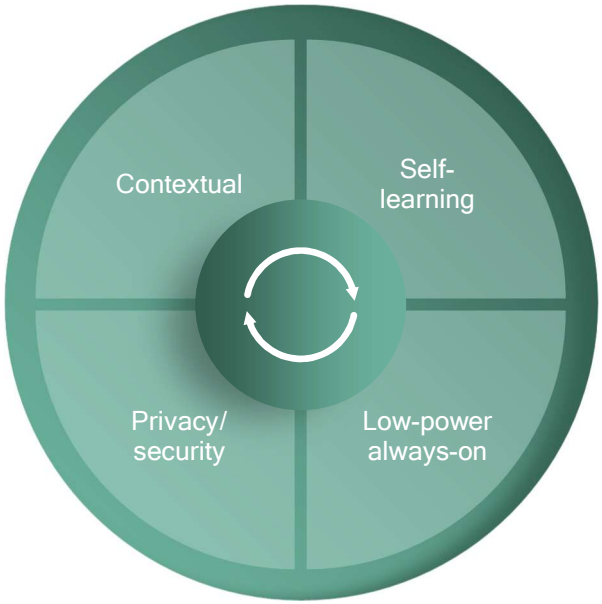
Algorithmic advancements

Software tools

Power efficient AI



Personalized AI



The challenge of AI workloads

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

Constrained mobile environment

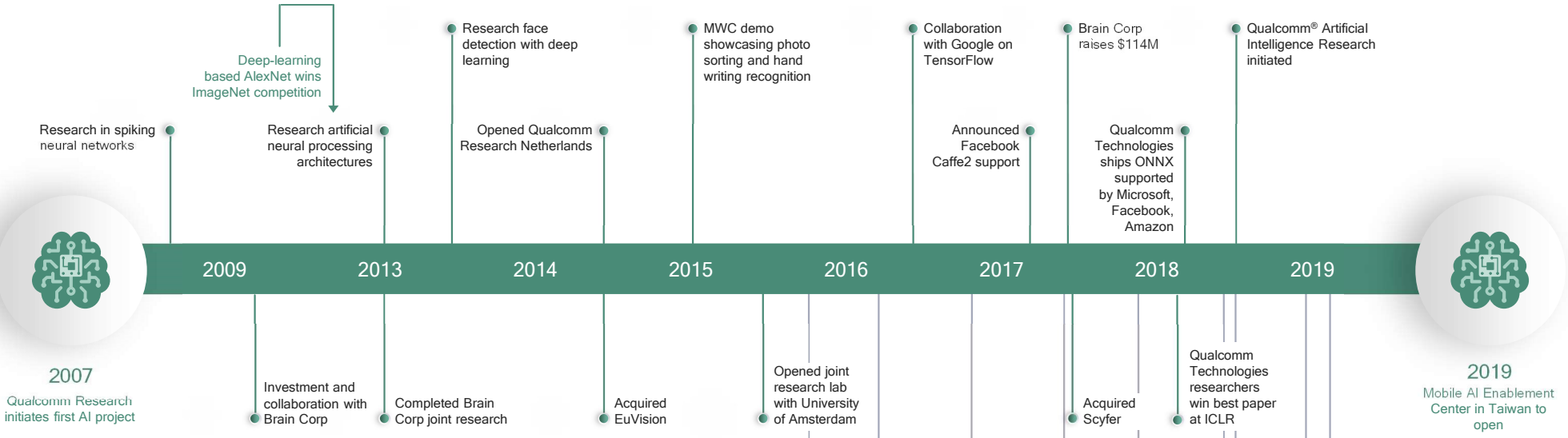
- Thermally efficient for sleek designs
- Requires long battery life for all-day use
- Storage/memory bandwidth limitations

Power and thermal efficiency

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

Our AI leadership

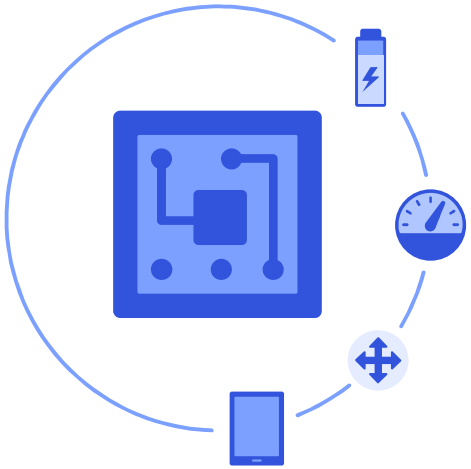
Over a decade of cutting-edge AI R&D, speeding up commercialization and enabling scale



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

Making on-device intelligence power efficient

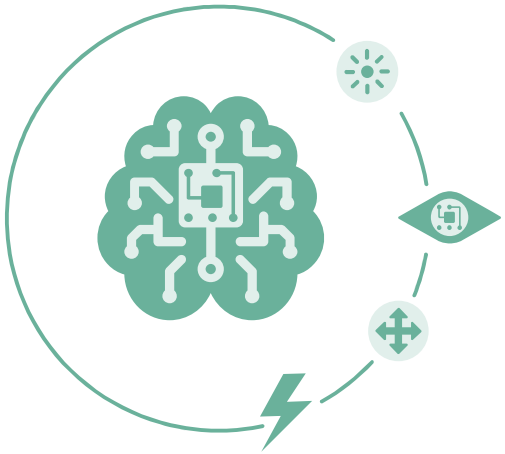
Focusing on high performance HW/SW and optimized network design



Efficient hardware

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

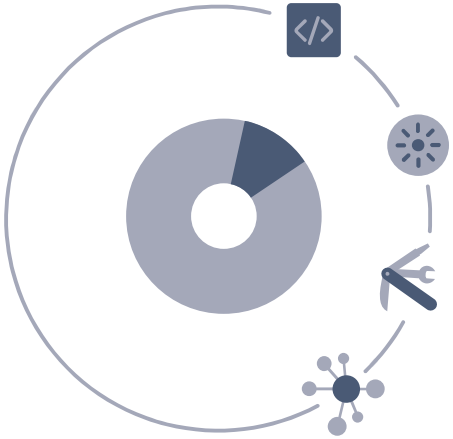
Selecting the right compute block for the right task



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 855
AI highlights

4th Gen AI Engine

7+ trillion operations per second

3x[1] performance improvement

Experiences built for the future

1: Compared to Snapdragon 845

50%[1] More ALUs
FP32 & FP16

New Tensor Accelerator

4x Vector eXtensions

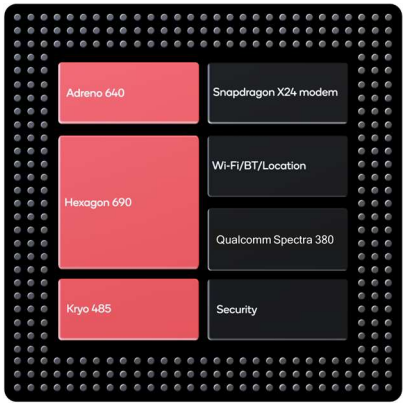
Optimized Scalar

Voice Assistant

INT16, INT8 & Mixed

New dot product instructions
FP32 & INT8

1: Compared to Snapdragon 845



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 | 听得见的智能

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

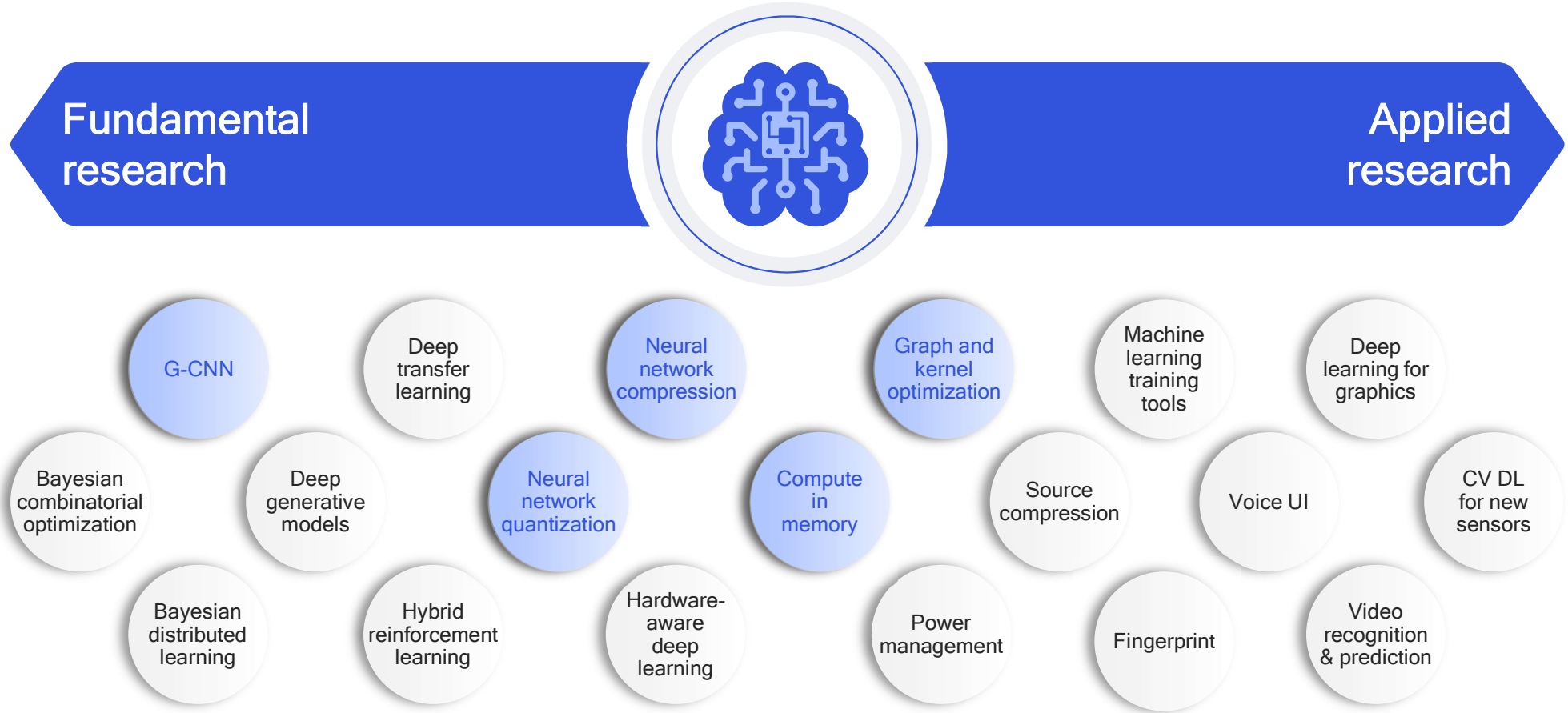
Video Segmentation

Bokeh

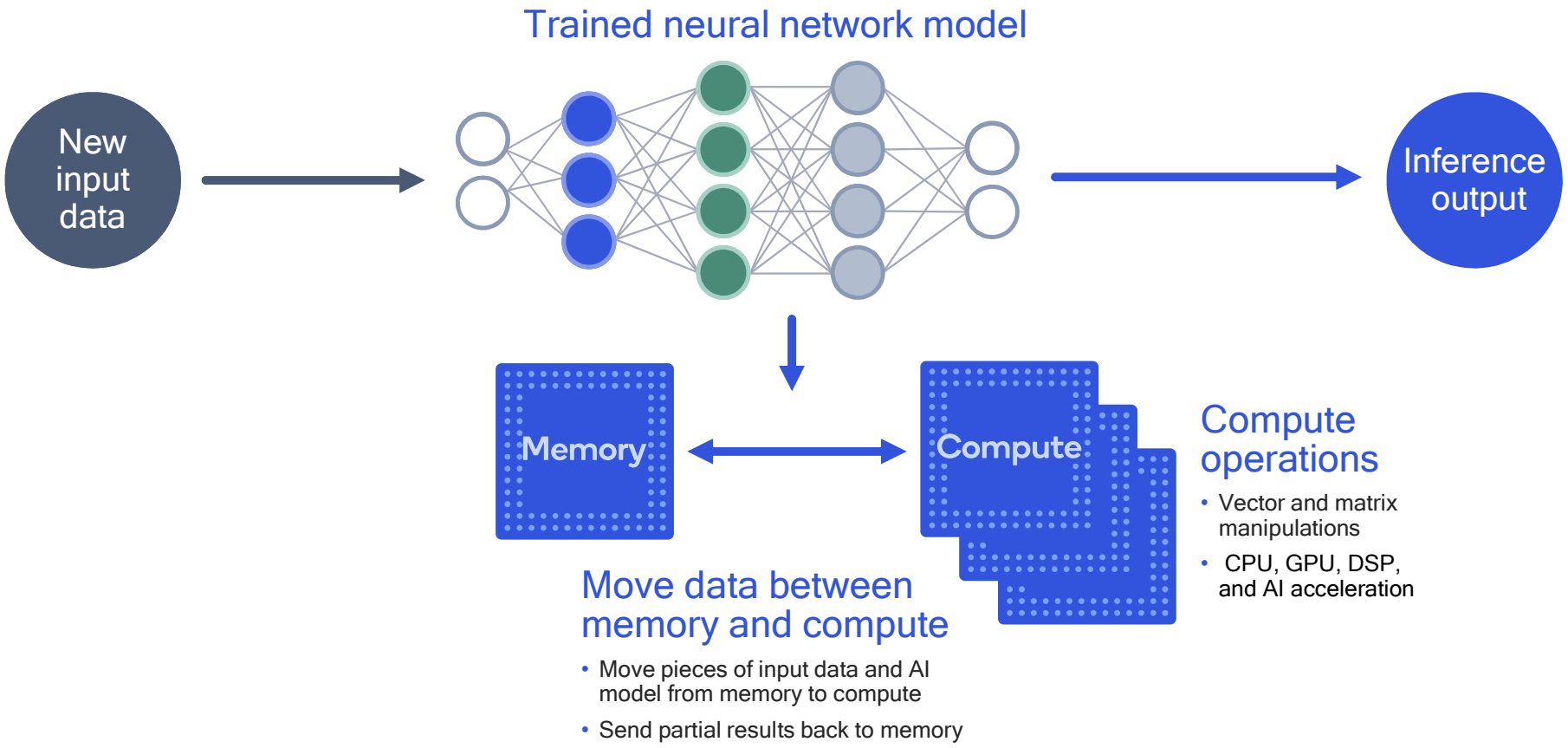
Devices

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

Leading research and development across the entire spectrum of AI



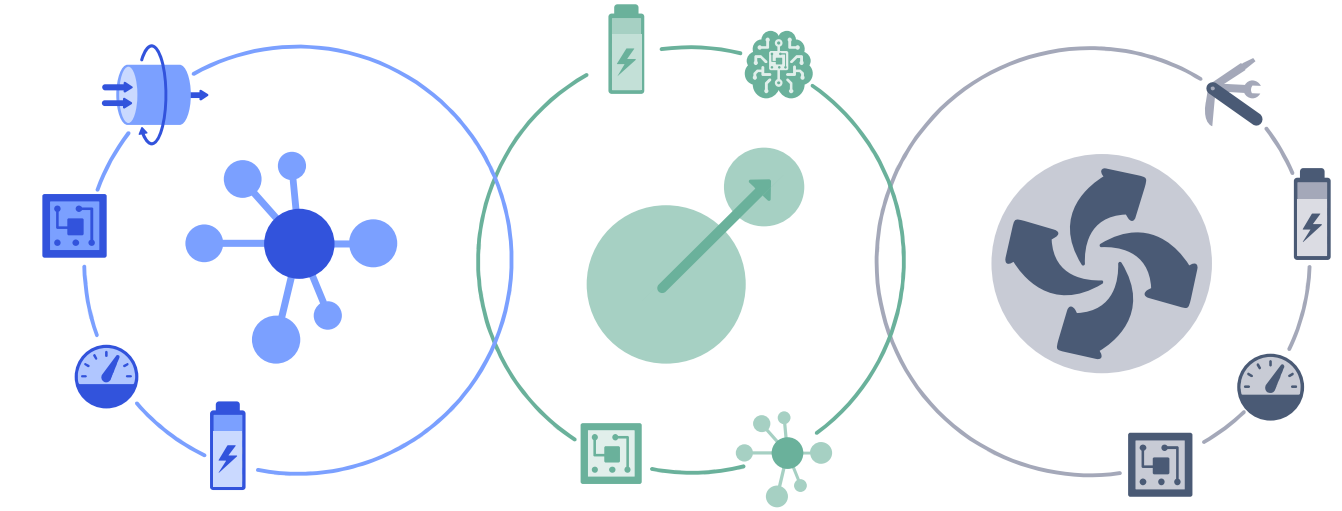
Advancing AI research to increase power efficiency



AI model optimization research for power efficiency

Applying AI to optimize AI models through automated techniques

Reduced time-to-market and engineering cost



Compression

Learning to prune model while keeping desired accuracy

Quantization

Learning to reduce bit-precision while keeping desired accuracy

Compilation

Learning to compile AI models for efficient hardware execution

Hardware awareness + CPU + GPU + DSP + AI acceleration

Final Thoughts on AI and Media Consumption Experiences

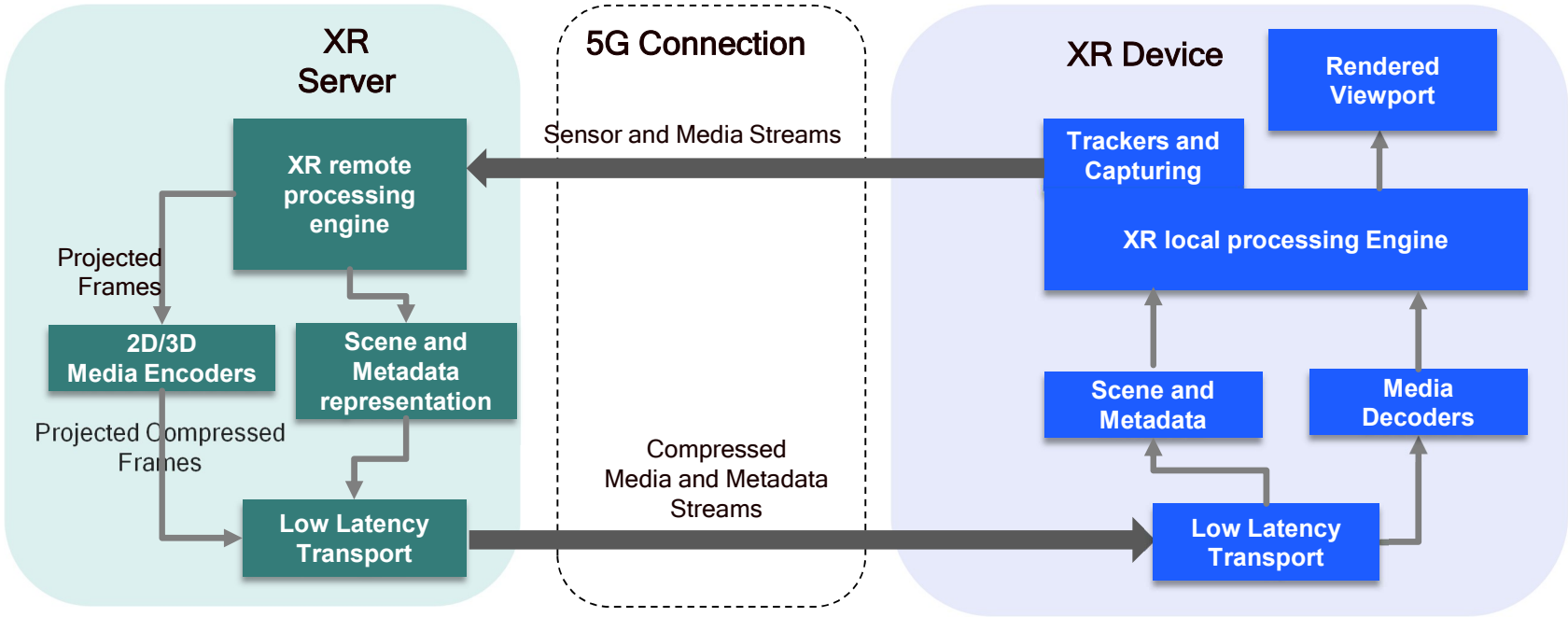
(not compression or production)

New Media Experiences

Example: Emotional Streaming

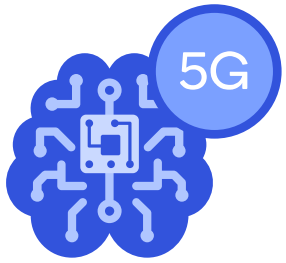
- See 3GPP XR Study and https://www.cnet.com/news/with-5g-you-wont-just-be-watching-video-itll-be-watching-you-too/
- Bob is watching a horror movie using an HMD. He is fascinated, but his body reaction, eye rolling, and other attributes are collected and are used to create a personalized story line. Movie effects are adjusted for personal preferences while reactions are collected when watching the movie. Bob's emotional reactions determine the story-line.
- Remember the last time you felt terrified during a horror movie? Take that moment, and all the suspense leading up to it, and imagine it individually calibrated for you. It's a terror plot morphing in real time, adjusting the story to your level of attention to lull you into a comfort zone before unleashing a personally timed jumpscare.
- Bandersnatch ++

Social
Immersive
Personalized
Interactive
Gaming-like
Cognitive
Short and long
Live and On-Demand
And and and …



The 5G XR architectures includes AI based media consumption

The future of AI and Media



5G + AI

New yet to be defined experiences

Wireless edge

Integrated and mobile

Broad industry collaboration

Collaborations, Platforms, Technologies, Enablers for the Creatives

Qualcomm

Thank you!

Follow us on:

For more information, visit us at:

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