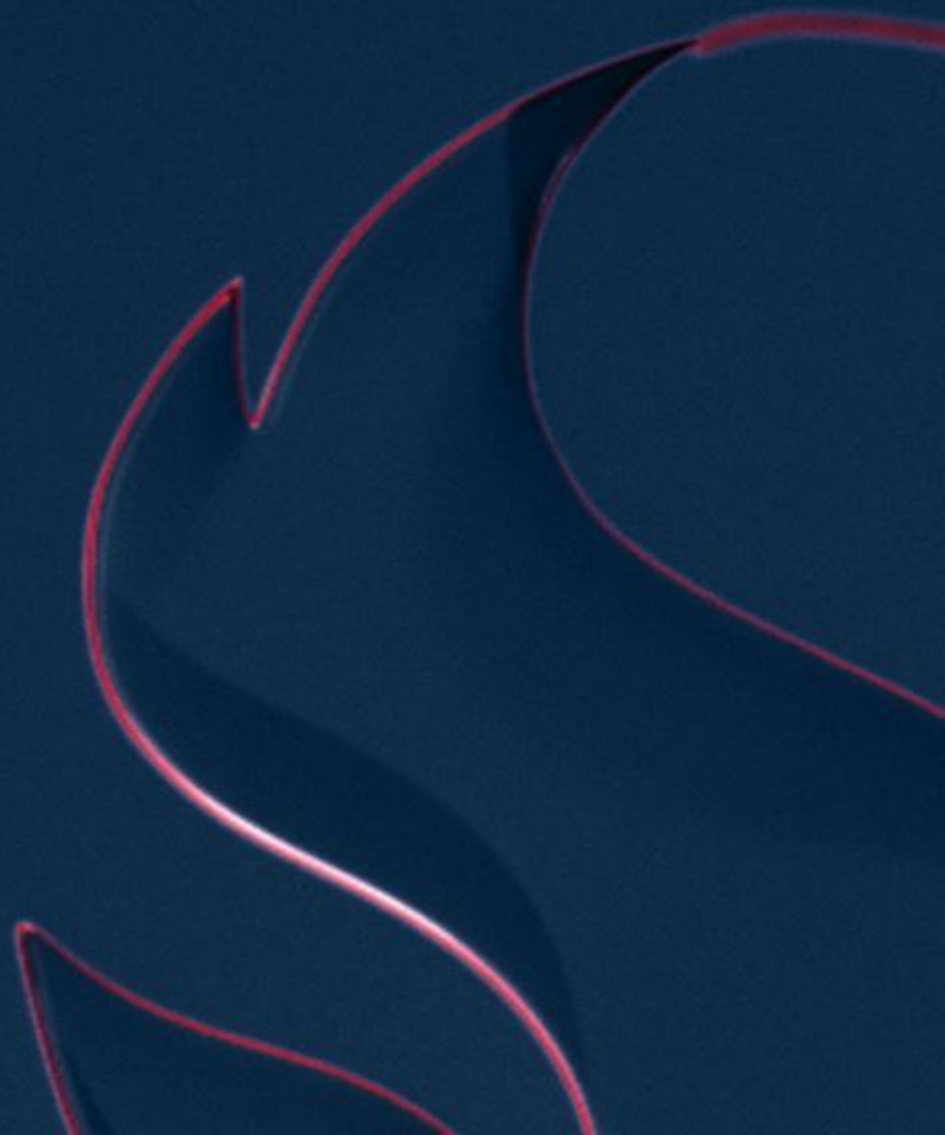
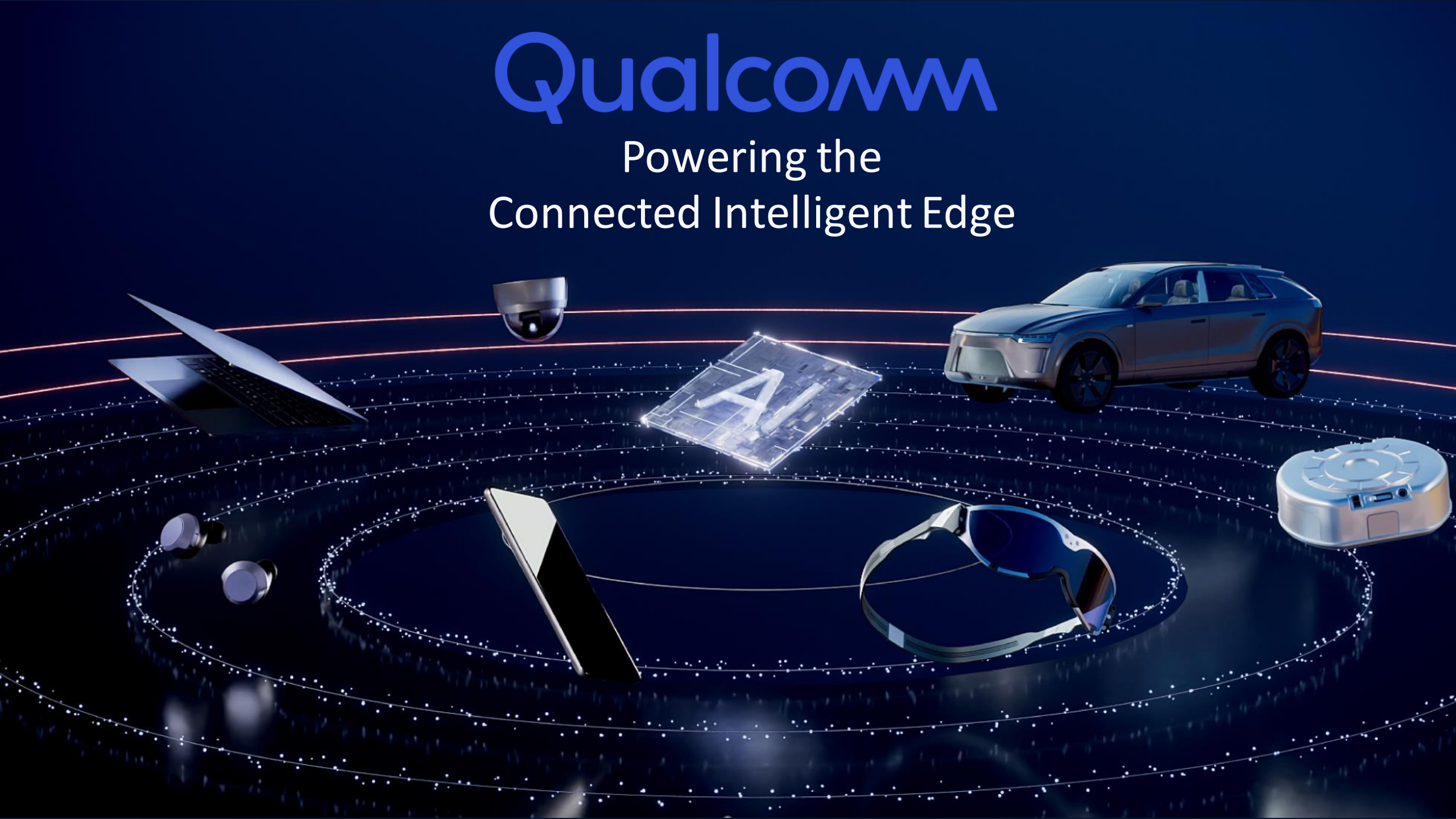
Felix Baum

Director,
AI Product Management
Qualcomm Technologies, Inc.



Qualcomm

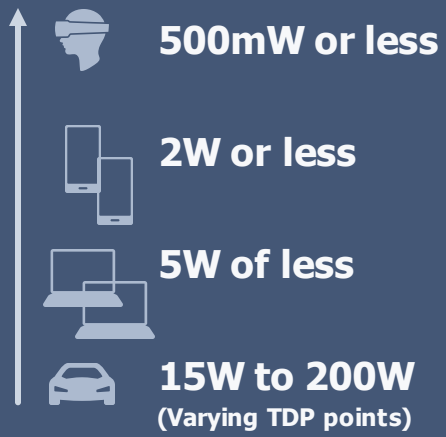
Powering the Connected Intelligent Edge



AI Challenges: Support structure across BU's

Performance optimization points

Innovative form factors are being constantly designed across many verticals and as such, one of the challenges is the ability to drive AI performance optimization (FPS or Latency or FPS/W) across multiple power envelopes



Support for different DL architectures and operators

AI based applications are quite widespread from image quality related (Mobile) to productivity (Compute) to assistance and monitoring (Auto ADAS) markets. This stretches Qualcomm's internal ecosystem to support:

Challenging DL architectures:

Generative models (Mobile) to transformer models (Compute) to Lidar models (Auto ADAS) which demands constant investment in compilers, tools, operators and other SW modules

Desired feature support across use-cases and business

Ability to drive innovation using AI has seen increased traction in the ecosystem but the need for various feature support varies by BU vertical

- Support for high concurrency (For Auto and XR verticals)

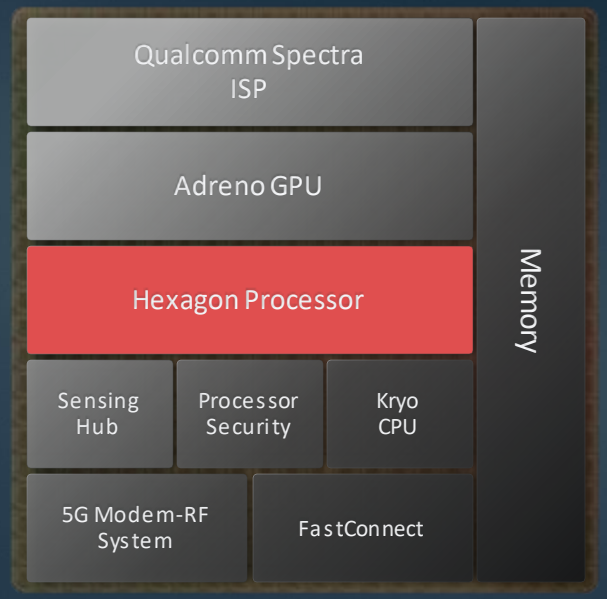
- Support for newer data types (For Data center verticals)

- Support for application scalability (For Mobile and Compute verticals)

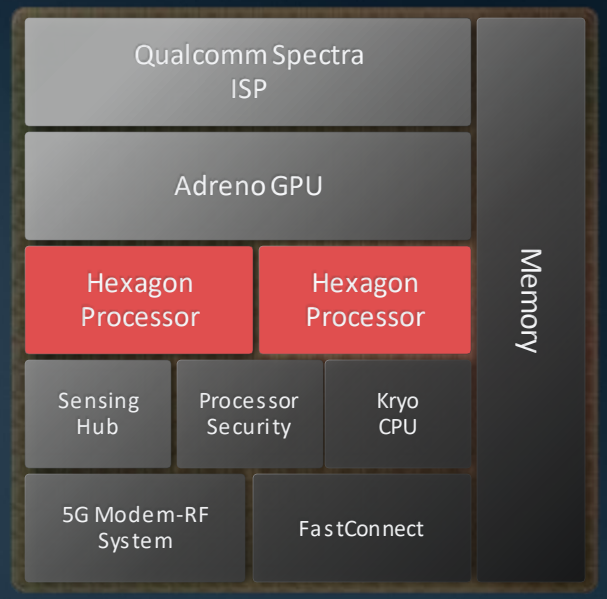
AI Challenge: Performance optimization points

Leveraging mobile and scaling to multiple markets

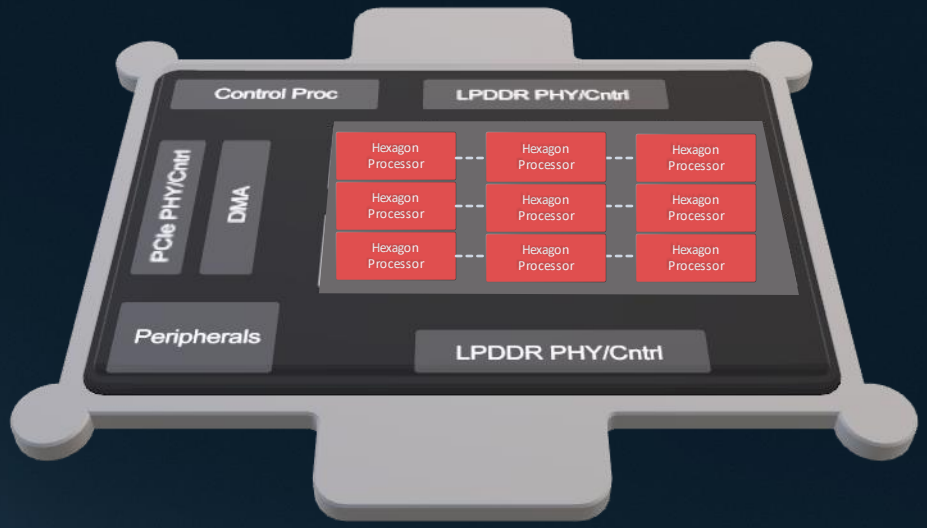
Anchor point at Mobile



Dual AI cores at Auto/Compute



Multiple AI cores at Cloud/Edge/Auto



Scalability and adaptability for different markets and AI needs

AI Challenge: Support for different DL architectures and operators

Leveraging mobile and scaling to multiple markets

- Customers need ability to drive innovation and many of them use ONNX models as an interchange format.
- By enabling direct conversion and preparation of the ONNX models into executable binaries we provide a seamless acceleration workflow from the training model to on device optimized execution
- Once our clients get to ONNX model they have ability to reuse the same model across the spectrum of performance of various Qualcomm SoCs.
- The benefit to Qualcomm is that it helps us to avoid building multiple converters
- To enable direct acceleration, we are collaborating with Microsoft to enable ONNX RT Execution Provider

AI Challenge: Desired feature support across use-cases and business

Unified Qualcomm AI Software Stack provides allows portability across the vertical markets

Qualcomm AI Stack

Tools:

AIMET

AIMET Model Zoo

NAS

Model analyzers

Infrastructure:

Prometheus

kubernetes

docker

AI Frameworks, Runtimes

Developer Libraries and Services

System Software

OS

AI Frameworks

TensorFlow PyTorch ONNX

AI Runtimes

Qualcomm® Neural Processing SDK | ONNX RUNTIME | TF Lite Micro | Direct ML | TF Lite

Qualcomm® AI Engine Direct

Math Libraries | Compilers | Virtual platforms

Profilers & Debuggers | Programming Languages | Core Libraries

System Interface | SoC, accelerator drivers | Emulation Support

android | Zephyr™ | ubuntu® | CentOS | QNX

Platforms

Smartphones | XR | ACPC | IoT | Robotics | Auto | Cloud







One technology roadmap that scales to address all growth vectors

Edge AI

Camera

GPU

Graphics

CPU

Processing

Connectivity

On-device intelligence

High performance, low power

Everything wireless

HIGH COMPLEXITY

LOW COMPLEXITY

Thank you

Follow us on:

For more information, visit us at:

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