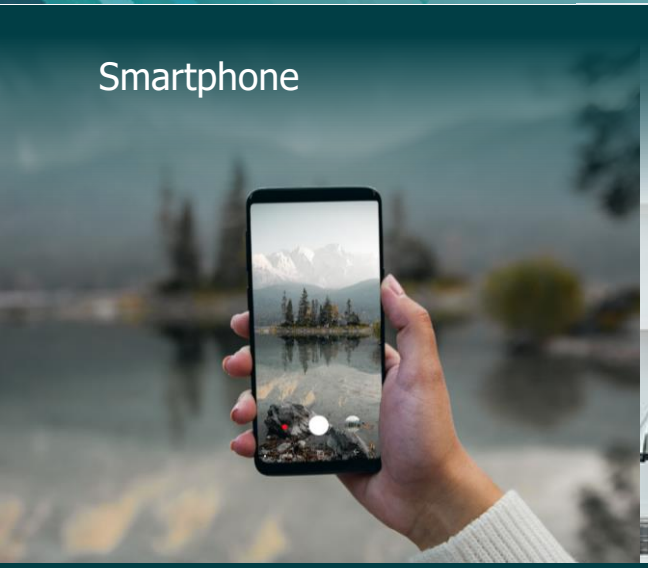
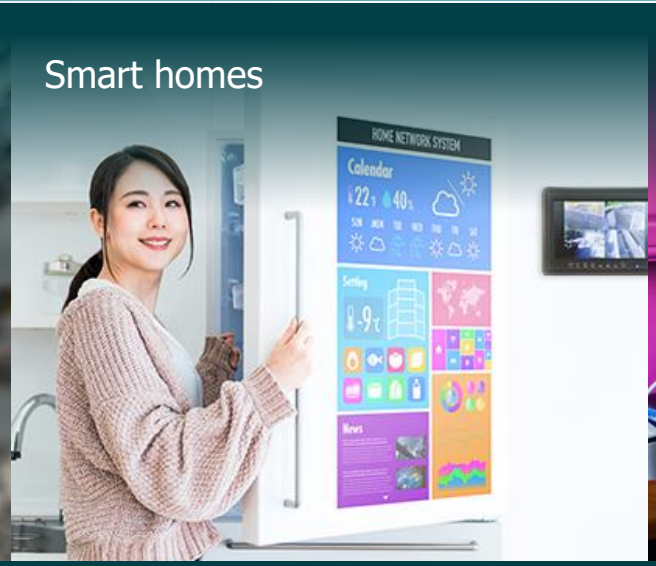
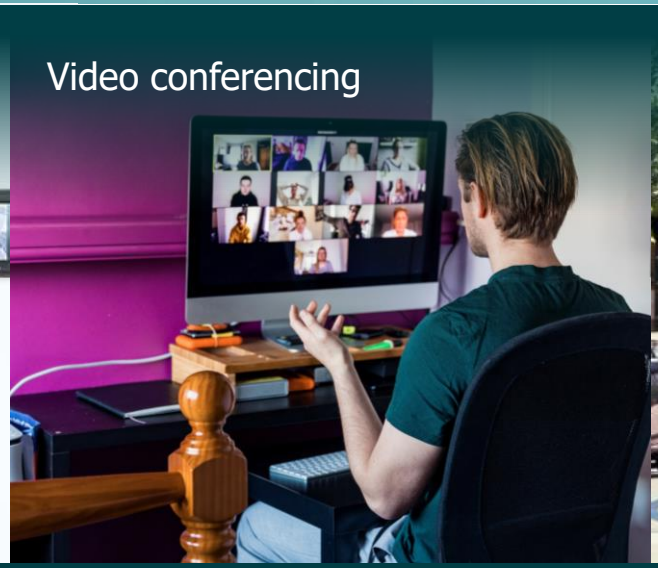
Autonomous Driving AI Workloads:
Technology Trends and Optimization Strategies

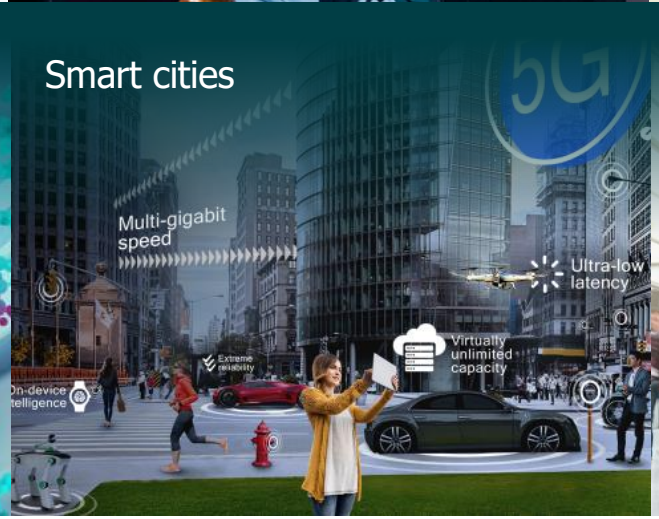
Ahmed K. Sadek
Senior Director of Engineering
Qualcomm Technologies, Inc.

The Need for Intelligent, Personalized Experiences Powered by AI is Ever-growing

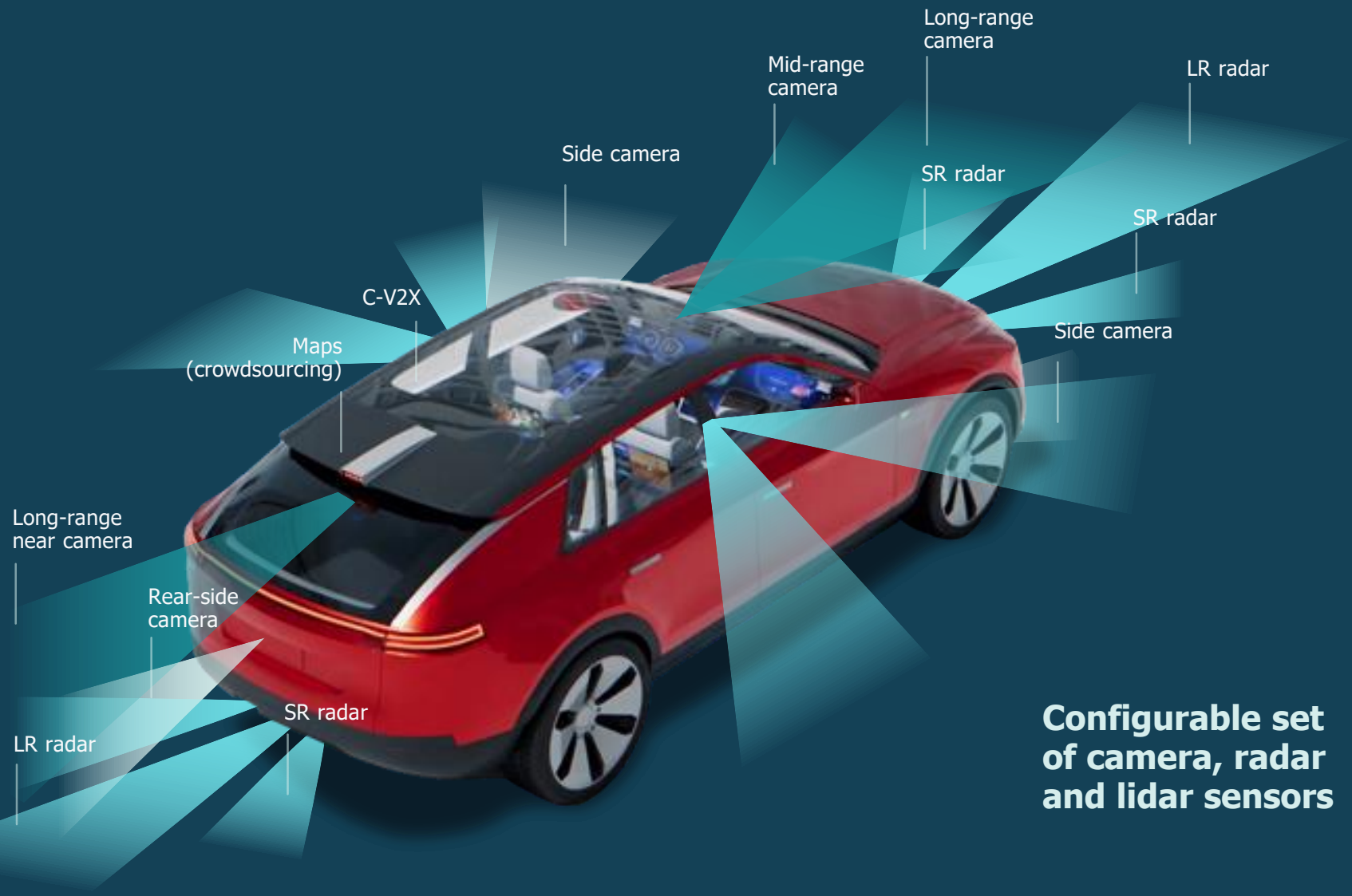








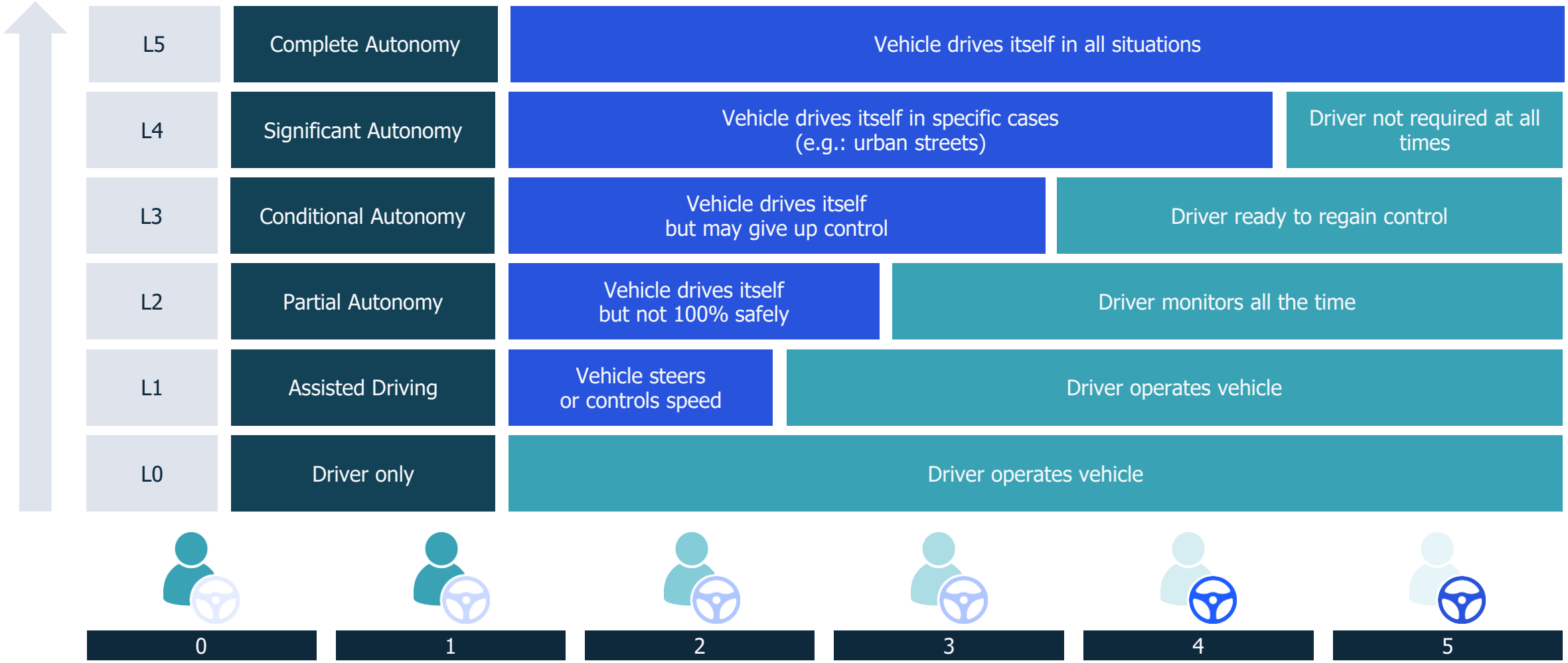
What Makes an Autonomous Vehicle (AV)?



Configurable set of camera, radar and lidar sensors

Qualcomm

SAE[1] Levels of Autonomy

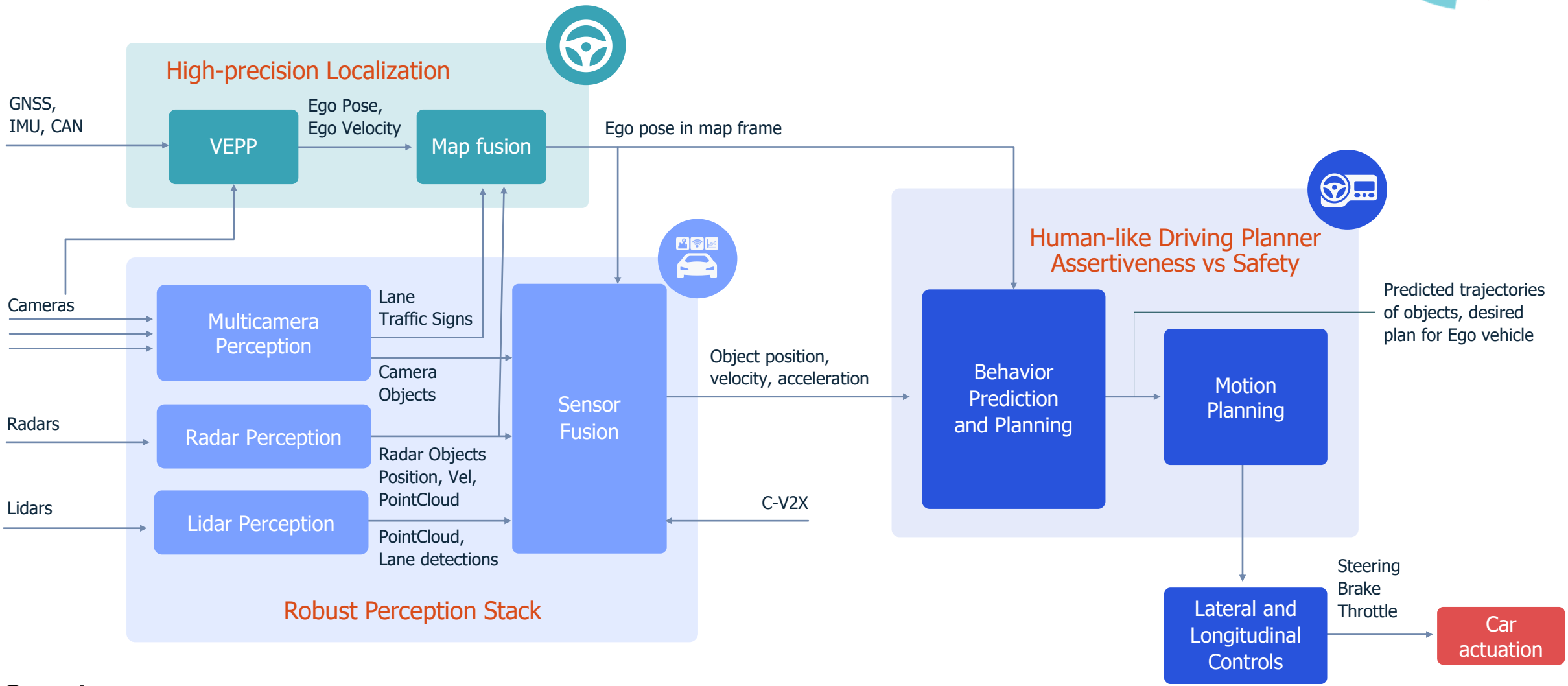


Qualcomm

1 SAE is Society of Automotive Engineers

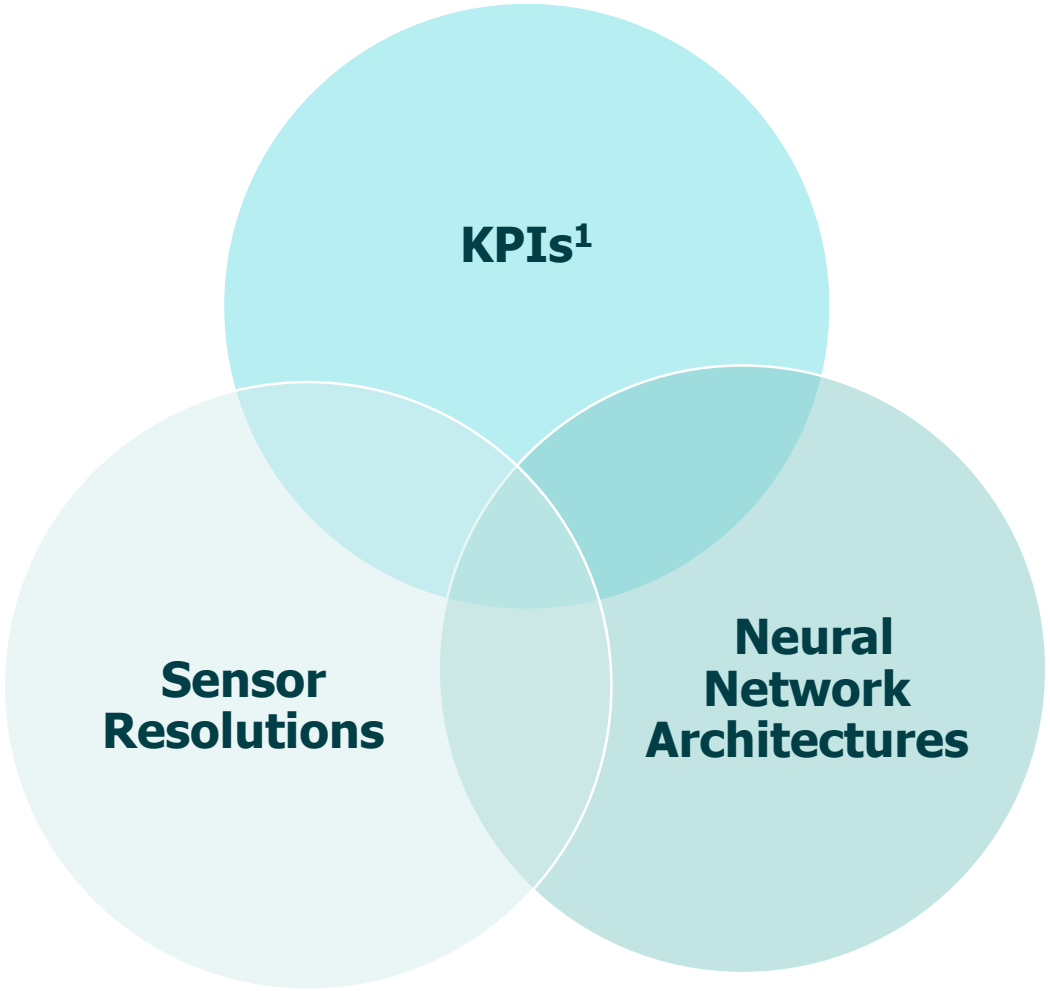
Autonomous Driving Stack

Solving the complex autonomous driving problem flow



Qualcomm

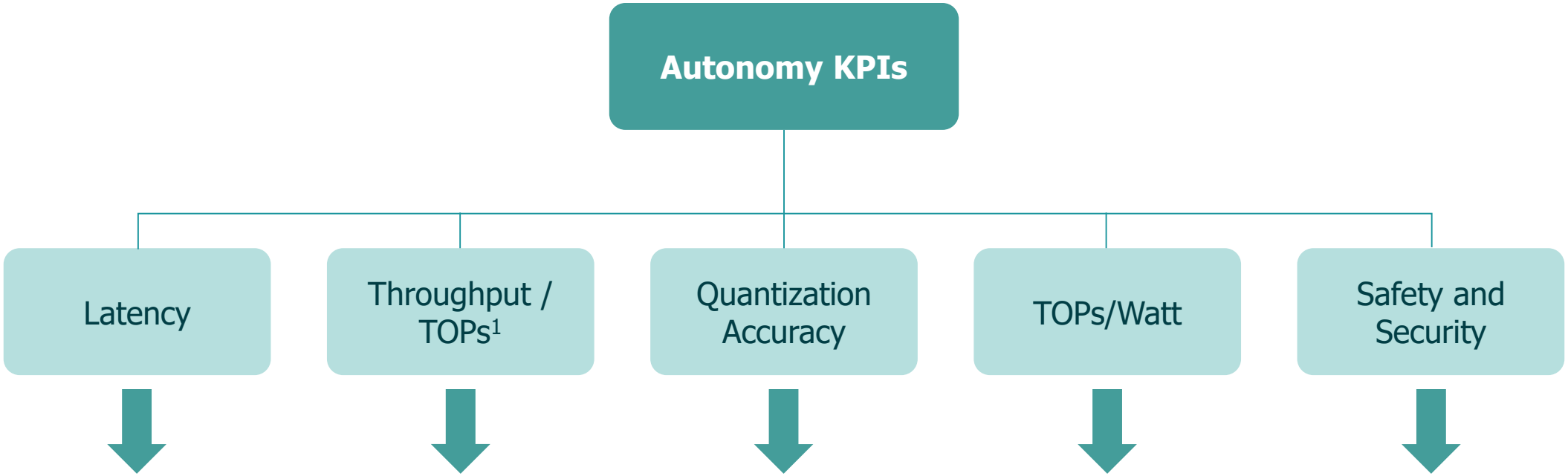
What Drives ADAS/Autonomy Workloads Complexity



Qualcomm

1 KPI is Key Performance Indicator

Peak TOPs is not Enough



Qualcomm

1 TOPs is Tera Operations Per Second 2 mAP is mean Average Precision

Robust Quantization Techniques

- mAP is a typical KPI used in studying accuracy before and after quantization
- **Low SNR scenarios (tail scenarios) are critical for autonomy**
- Low SNR in context of DL could mean the FP32 performance (e.g., decision boundaries) are barely meeting performance, and quantization noise results in failure for these scenarios, e.g.:
  - Objects represented by few pixels such as far away objects or small objects
  - Rare objects (e.g., non-conventional trailer trucks, animal on the road, ….)
  - Bad weather/lighting conditions
- **Significant improvement from quantization-aware training**

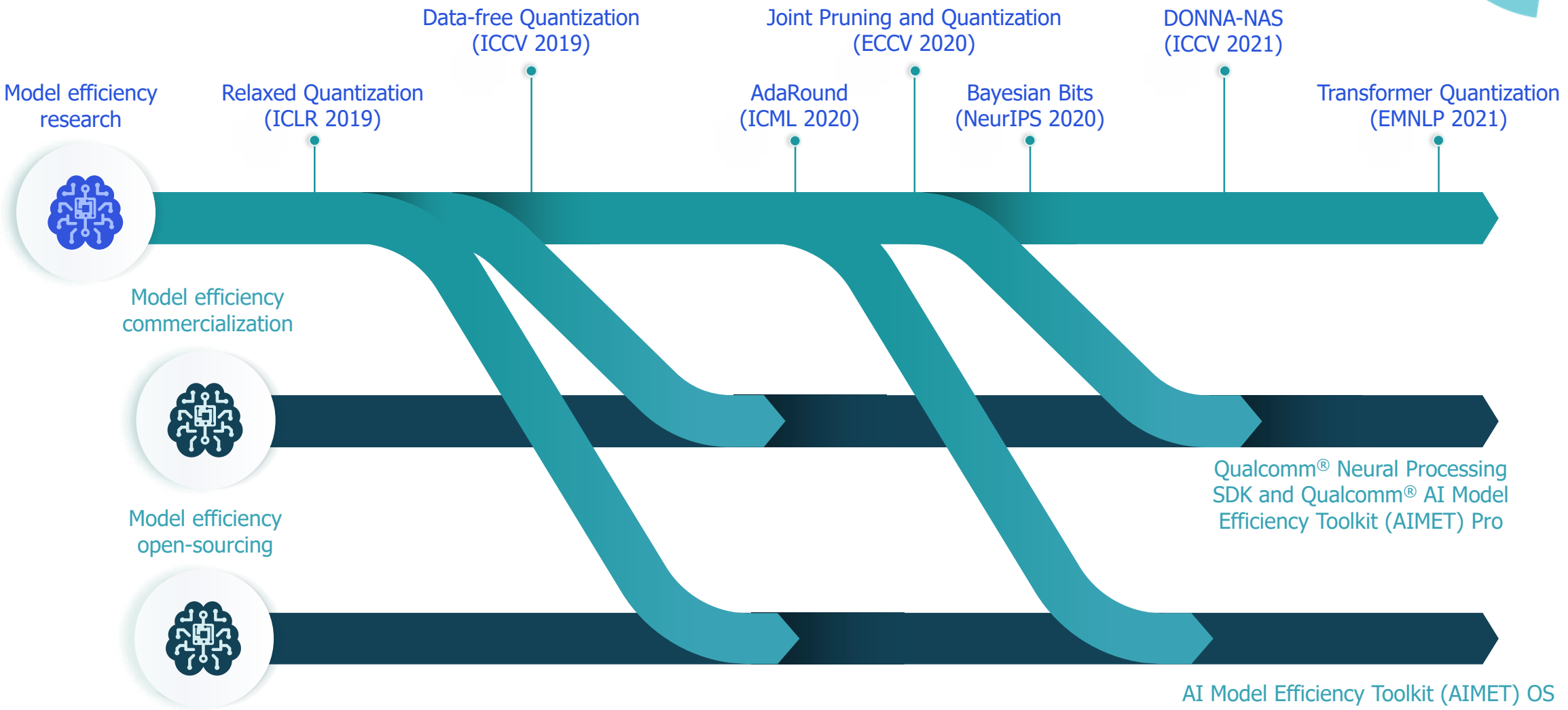
Example for Low SNR Scenario

Far object

Uncommon object

Driving the Industry Towards Integer Inference and Power-efficient AI

Leading model efficiency research and fast commercialization



Qualcomm Neural Processing SDK is a product of Qualcomm Technologies, Inc.
AIMET is a product of Qualcomm Innovation Center, Inc.

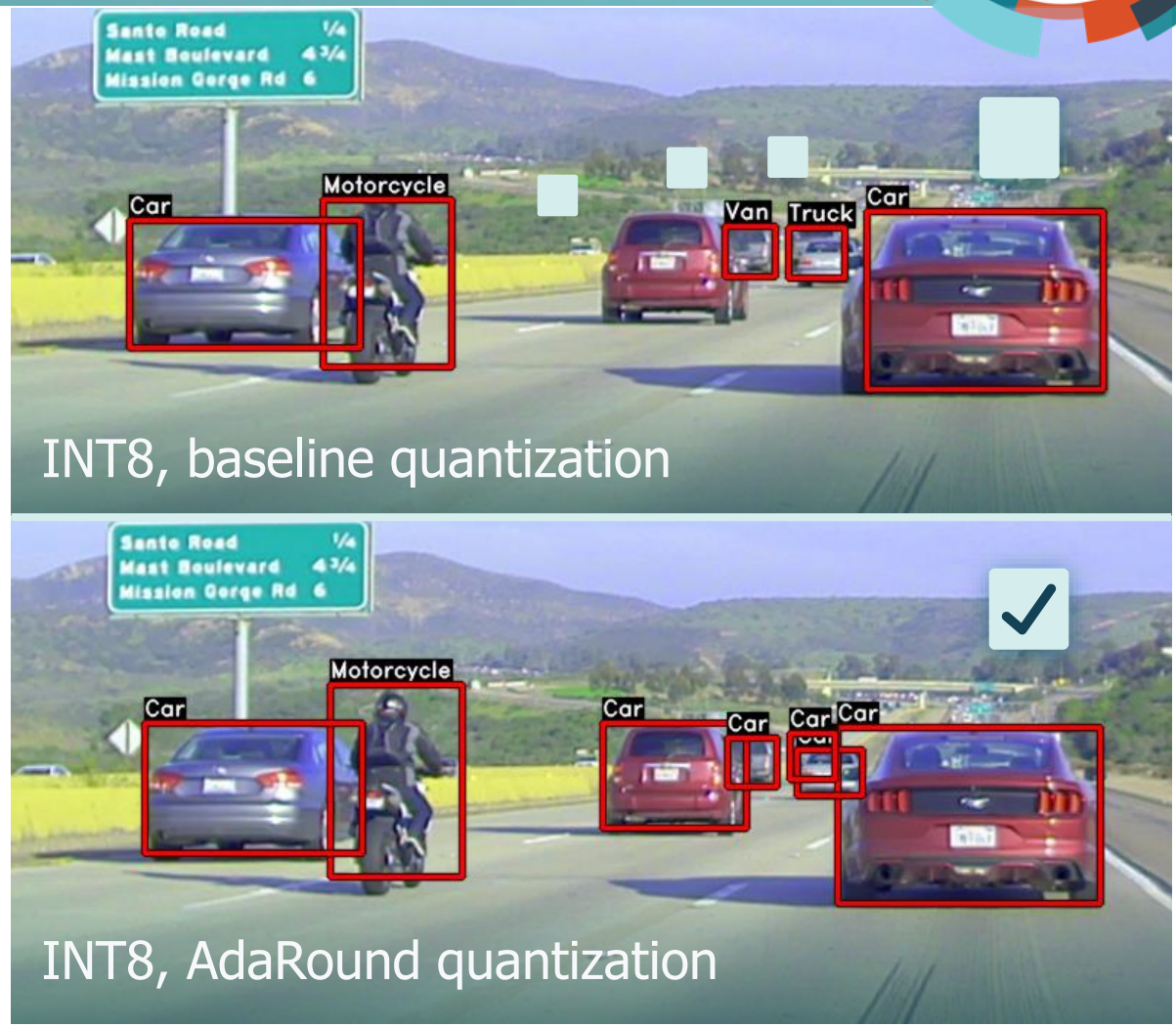
AdaRound Results

Post-training technique that makes INT8 quantization more accurate and INT4 quantization possible

Bit width	Mean AP (mAP)
FP32	**Baseline**
INT8 baseline quantization	**>10%**
INT8 AdaRound quantization	**<1%**

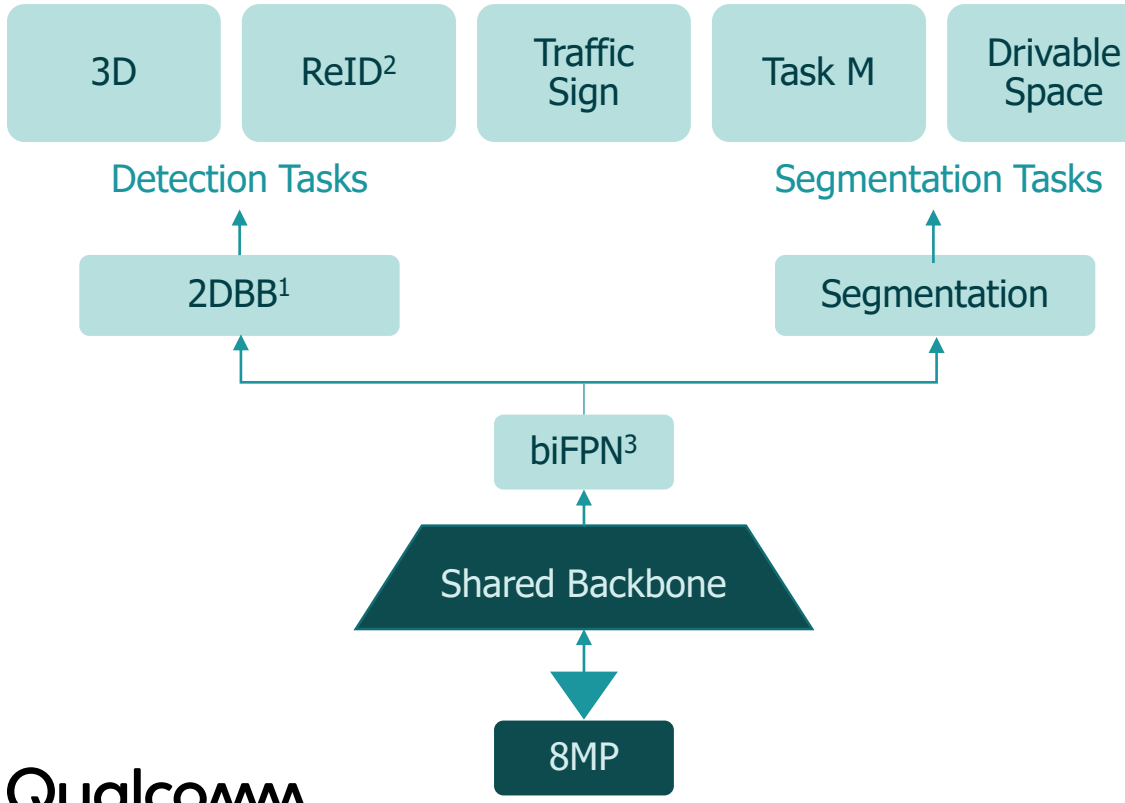
<1%

Reduction in accuracy between FP32 and INT8 AdaRound quantization



Qualcomm

Front Camera Multi-Task Architecture



Multi-task architecture with common shared backbone

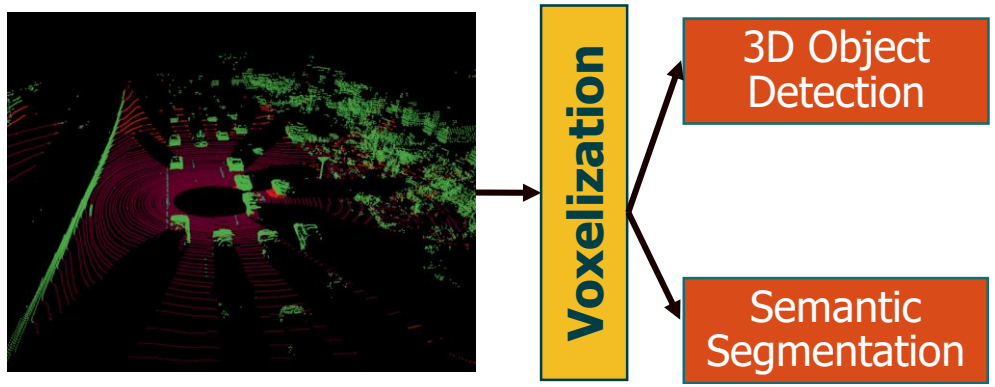
Architecture trends:

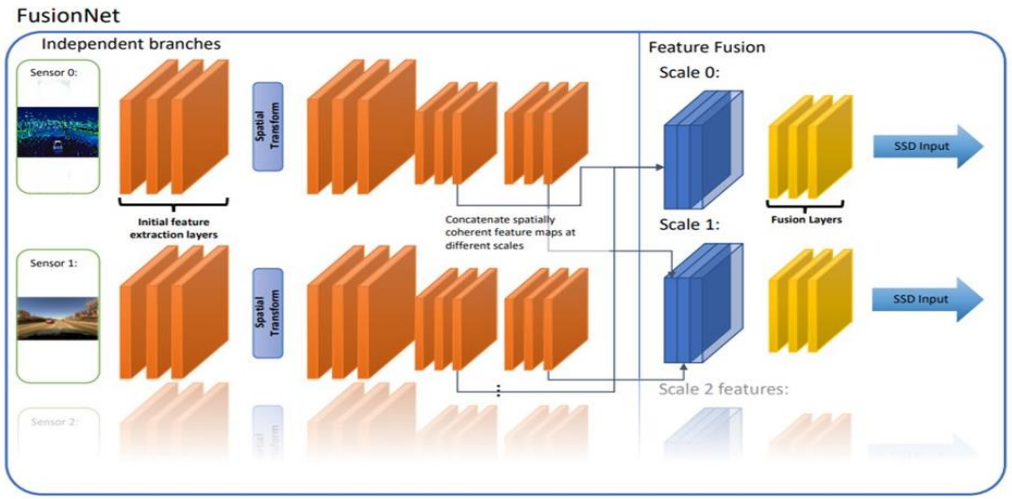
- Higher resolution → Pushing requirements on compute and memory
- High dilation factors
- Transformer heads
- Low level fusion across multiple cameras

Qualcomm

1 2DBB is 2D Bounding Box 2 ReID is Re-Identification
3 biFPN is Bi-directional Feature Pyramid Network

Low level fusion and sparse point cloud signals





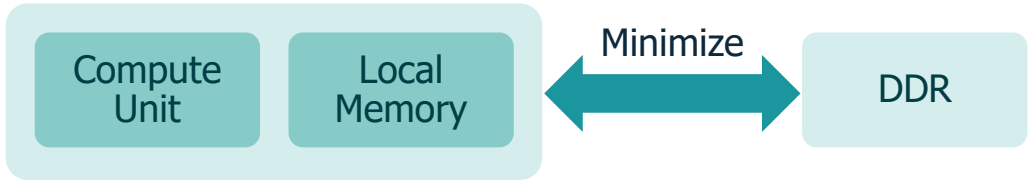
TY Lim, et.al, Radar and Camera Early Fusion for Vehicle Detection in Advanced Driver Assistance Systems,

- **3D Sparse Convolution** can reduce the computation and speedup the inference
- Efficient and novel approaches in both hardware and software architecture to **handle high sparsity from both data movement and compute**
- **Camera/radar low level fusion between high input resolution and high sparse multi-modal signals**

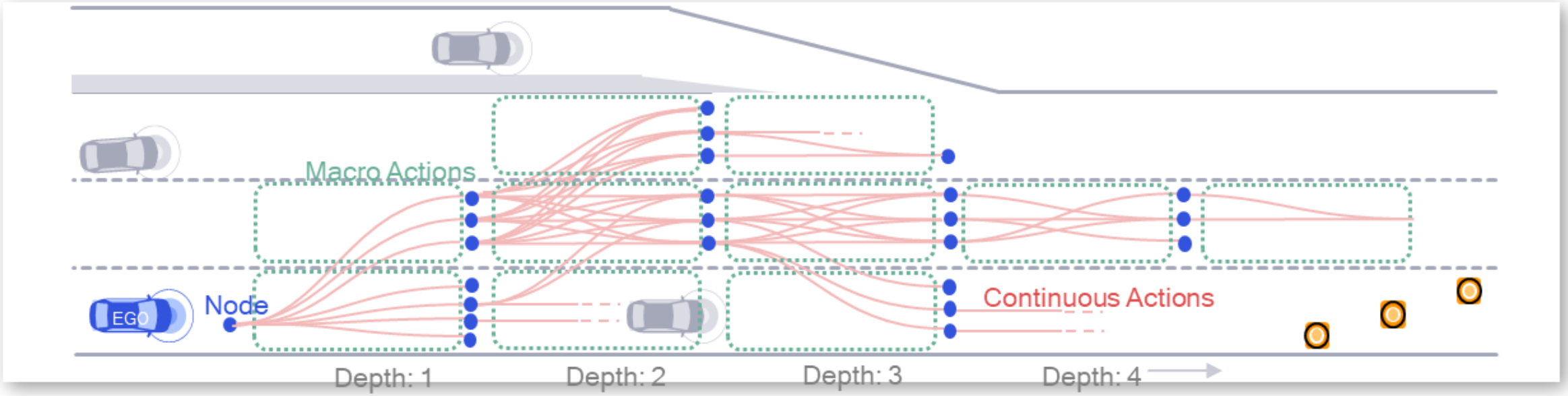
Qualcomm

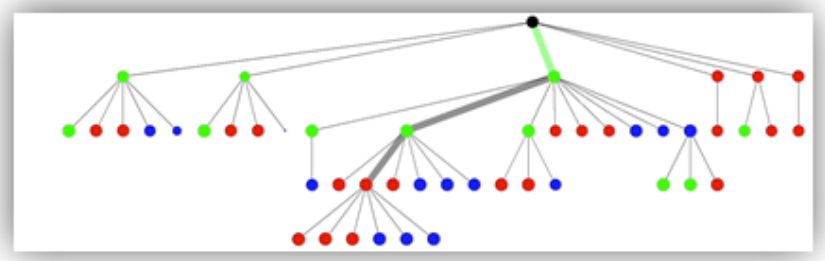
Scheduling / Graph Optimizations

- **Graph scheduler optimization to minimize spillage to DDR**
- **Reuse intermediate activations for next layer(s) processing on chip**
- Boost inferences/second
- Reduce MBytes/inference and preserve DDR BW for other applications



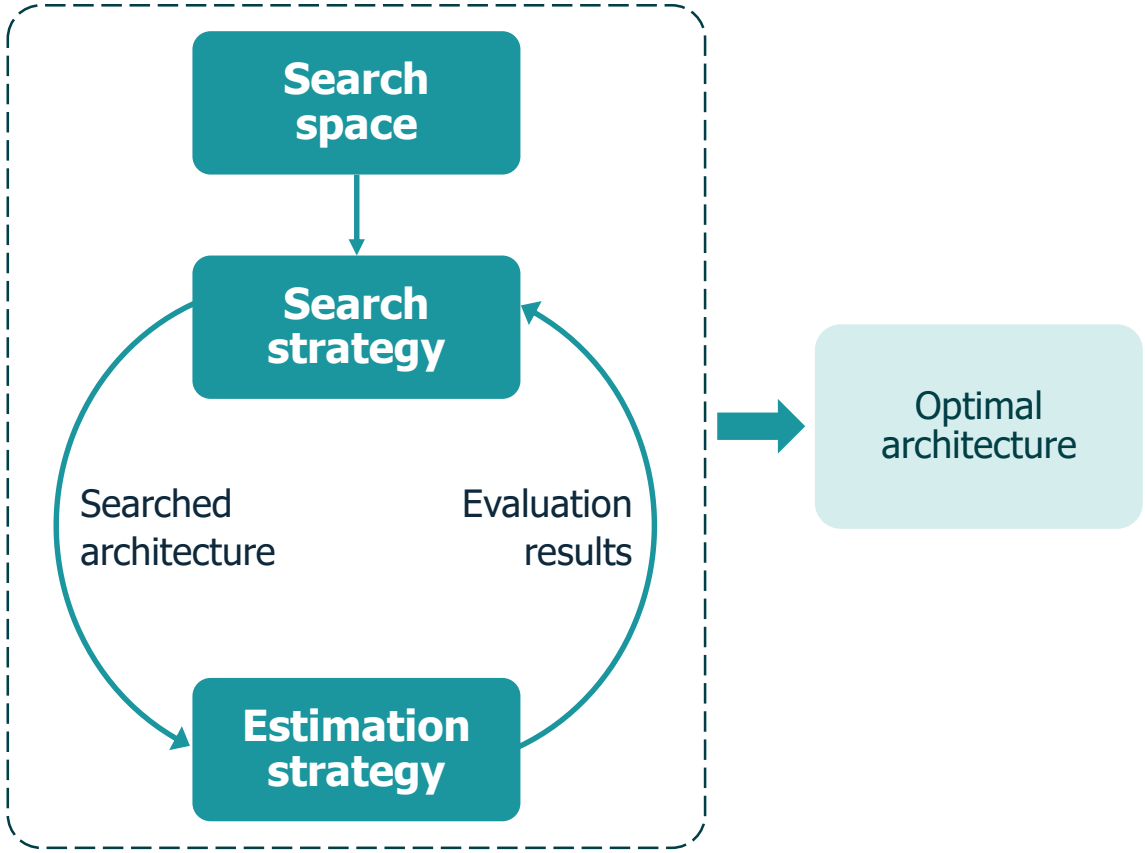
Behavior Planning: Model-Based RL
Highly dynamic dataflows





Neural Architecture Search (NAS)

- Exisiting techniques require lots of computational time and has large search space
  - e.g., RL and Evolutionary Algorithms
- The need for efficient techniques has emerged
- Weight-sharing
- Differentiable Architecture Search
- These methods often suffer from an instability issue, and in many cases require careful training methods or search space design

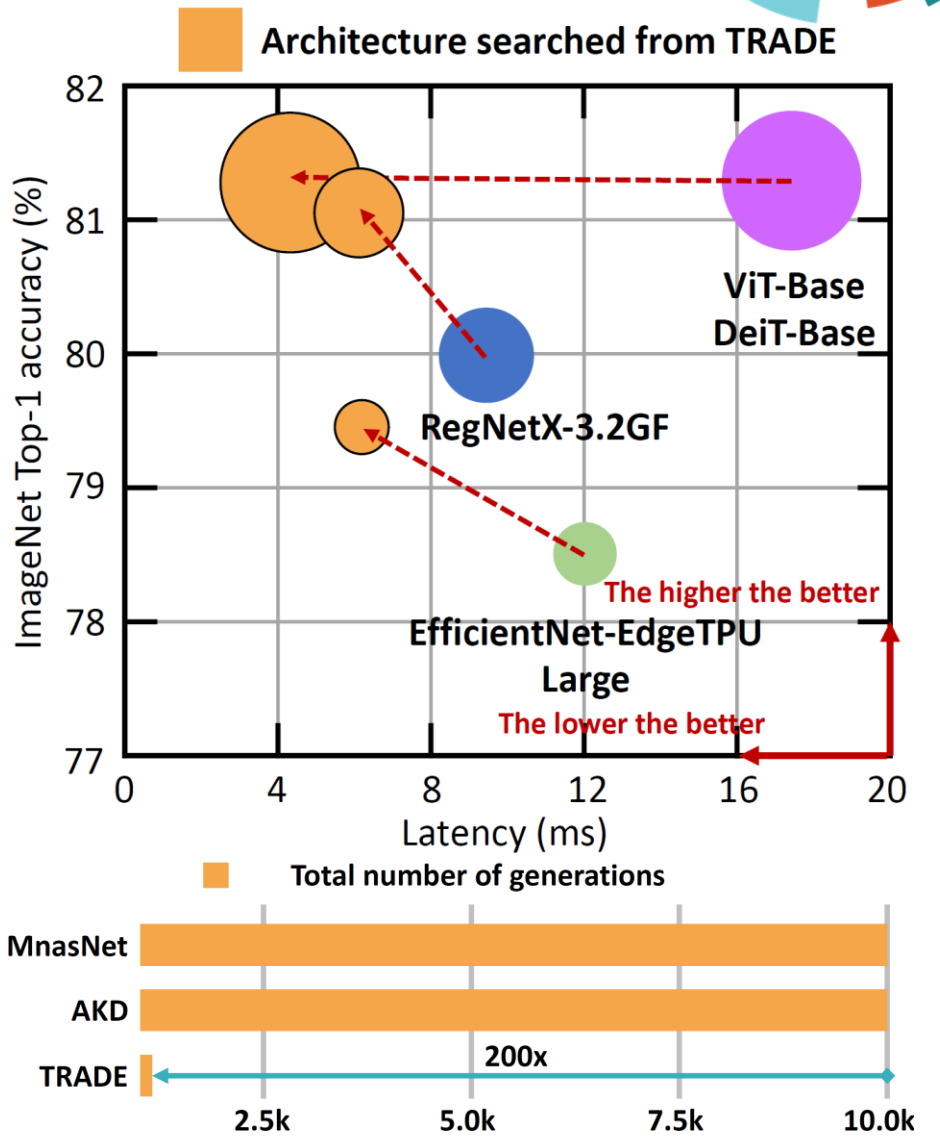


Improving NAS Efficiency

Trust Region Aware Sample-Efficient Architecture Search for Distillation

We have improved NAS in

- **Search phase**
  - Trust Region Bayesian Optimization for sample-efficient search
  - Knowledge Distillation-guided score to perform more efficient and teacher-aware search
- **Query phase**
  - Orthogonality regularization



Qualcomm

Snapdragon Ride SDK

- Middleware
- Optimized Libraries and Tools for AI, Math Libs, Vision Processing, Camera ISP, Multi-SoC comm.
- Production Ready Platform Support Package
- Safe Operating System with Hypervisors
- High Performance Multi-SoC Compute

Partners:

QNX

Red Hat

Infineon

VECTOR

Auto-Imaging System
Rich suite of camera support with multi-high-resolution cameras

Neural Processing Toolkit
AIMET, NAS, Compiler, Quantizer, Simulator and profiling tools for optimizing AI perception, planning

Embedded Vision and AD Libraries
Rich set of math and vision library functions optimized for DSP, CPU, GPU and Vision Accelerators

Middleware and Multi-SoC Infrastructure
Production Ready safety Platform, Automotive Multi-SoC middleware for seamless low latency high speed data movement

Tools
Profiler Tools to analyze processing blocks utilization, latency, memory bandwidth, power and thermal management

Development Platform
Reference hardware design with up to 16 cameras, radars, lidars, location for L2 to L3 system design
Multi-SoC architecture with Safety MCUs and Storage, with production ready thermal design

Qualcomm

Conclusion

- Autonomous driving AI workloads are increasing in complexity requiring SW-HW co-design for increased efficiency
  - Both for HW accelerator architecture and data flow optimizations
- New quantization methods presented that optimize for both average and low SNR regimes
- Further improvements required to increase NAS efficiency and speed innovation cycle
- Among topics not covered that will impact future network architecture design: causality and ability to reason

For More Information

Qualcomm AI

http://www.qualcomm.com/ai

Qualcomm ADAS

Qualcomm ADAS

Qualcomm Technology

Qualcomm YouTube

Qualcomm @ 2022 Embedded Vision Summit:

"A Practical Guide to Getting the DNN Accuracy You need and the Performance You Deserve" – Felix Baum – Wed, May 18, 2:40 PM

"Tools for Creating Next-Gen Computer Vision Apps on Snapdragon" – Judd Heape - Wed, May 18, 10:50 AM

"The Future of AI is Here Today: Deep Dive into Qualcomm's On-Device AI Offerings" – Vinesh Sukumar - Wed, May 18, 12:00 PM

"Seamless Deployment of Multimedia and Machine Learning Applications at the Edge" – Megha Daga - Tuesday, May 17, 2:40 PM

Qualcomm

Thank You

Qualcomm