Steering Towards Heterogeneous Compute for ADAS

Heterogeneous compute architectures optimize and scale advanced driver assistance systems across all your vehicle tiers.

Authors:

Ahmed Sadek, VP Engineering

Ajit Rao, Senior Director Engineering

Tom Toma, Director Product Marketing

Contents

Introduction

OEMs and Tier-1 suppliers face three main requirements as they design and build the vehicles of the future:

- Performance – They want a high-performance advanced driver assistance system (ADAS), with adequate compute power and data-movement capabilities for basic features, customer functions and artificial intelligence (AI) innovations.
- Power consumption – At the same time, they want that ADAS to operate within tight power and thermal budgets.
- Scalability – They want a single processor family versatile enough to meet ADAS needs at all vehicle tiers in their product line.

The requirements are simple, but they are not easy to fulfill. Common complaints from OEMs and Tier-1s center on underpowered units, high demands on electricity and purpose-built, single-model approaches that do not scale to other models.

To fulfill those requirements, Qualcomm has designed Snapdragon Ride™, a family of power-efficient system-on-chips (SoCs) based on a scalable, heterogeneous architecture. The SoCs feature multiple CPUs, graphics processing units (GPUs) and neural processing units (NPUs), along with dedicated hardware accelerators optimized for automotive applications.

This paper describes how Snapdragon Ride SoCs fulfill automakers' requirements for high performance, low power consumption and scalability. The SoCs integrate a heterogeneous set of blocks designed for different types of compute performance, ranging from basic driving tasks to time-sensitive deep learning algorithms. They form a reliable, end-to-end system capable of moving data among components on a low power budget. By mapping a wide variety of sensors and lower-level algorithms to the same set of hardware blocks in a single architecture, they can scale across multiple vehicle tiers.

OEMs and suppliers can use this paper to gauge the fit for Snapdragon Ride in their own design plans for current and future vehicles. Get more details on building scalable ADAS solutions.

The Difficulty in Scaling ADAS across Vehicle Tiers

ADAS revolutionizes the way vehicles perceive, interpret and respond to their surroundings. The development of ADAS features in passenger vehicles has been fueled not only by consumer demand but also by the test criteria of the Global New Car Assessment Programs (NCAP) and the European General Safety Regulations (GSR). The systems are defined on an autonomy scale from basic regulatory functionality Level 0 (L0) to fully autonomous Level 5 (L5).

Each level corresponds to a set of customer functions – the features that customers see when they buy and use a car equipped with ADAS. But from the perspective of the OEM or supplier, each level (and set of customer functions) corresponds to a tier – for example, entry-level, mid-level and advanced – within the product line. Each vehicle tier represents different design choices, different power requirements and separate processors. That makes it difficult to scale a single ADAS across tiers and compels automakers to adopt single-model approaches to ADAS.

Customer Functions – Emphasis on Interaction with Driver

A fresh look brings into focus the middle of the autonomy scale, where interaction with the driver is most important (see Figure 1).

Basic Levels

Basic customer functions include the following:

- Autonomous emergency braking (AEB)
- Lane departure warning (LDW)
- Lane keep assist (LKA)
- Lane centering system (LCS)
- Emergency lane keep assist (eLKA)
- Intelligent speed assist (ISA)

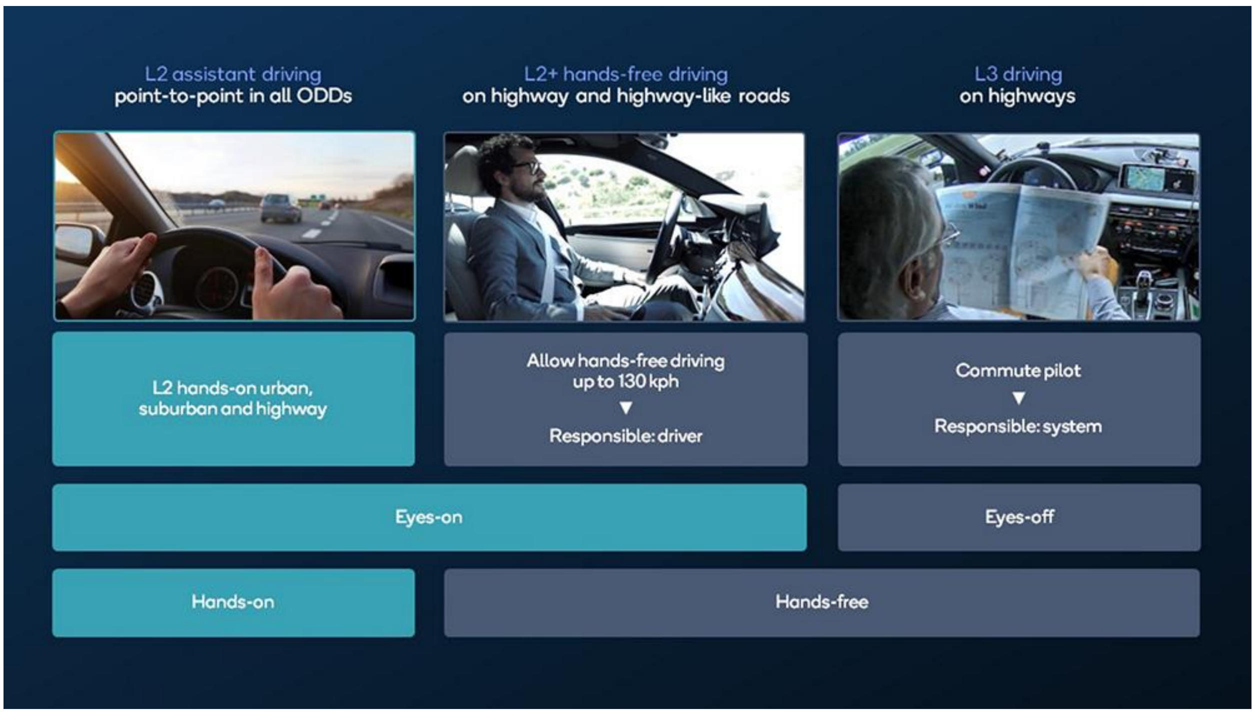

Figure 1: An overview of driving experiences, with emphasis on interaction with the driver

Level 2

Level 2 cars expand on and add to basic customer functions a variety of comfort functions, such as basic and intelligent adaptive cruise control (ACC and iACC), glare-free high beam (GFHB) control and adaptive suspension control. Note that levels up to and including L2 require the driver's hands to be on the steering wheel at all times.

To support basic and L2 customer functions, the following configuration of sensors is required:

- Three to five radars
- Up to five camera sensors
- One camera for the driver monitoring system and occupant monitoring system (DMS/OMS)

Level 2+

Level 2+ calls for partial automation under certain conditions. It expands on L2 with customer functions that include these:

- Traffic jam assist (TJA)
- Lane change assist (LCA)
- Highway navigate on autopilot (NOA)
- Urban NOA

In L2+, the driver must remain fully alert even when the vehicle handles such basic driving tasks as changing lanes, staying in flow with traffic, adhering to speed limits and handling on- and off-ramps.

Level 3

At Level 3, the vehicle can handle L2+ driving tasks independently; the driver is eyes-off but is ready to take over as the system requires. At this level, the main customer functions in current deployment are L3 highway pilot and L3 highway traffic jam assist.

The sensor configuration for L2+ and L3 includes the following:

- Five to nine radars
- Five to 11 cameras
- One or more DMS/OMS cameras
- Potential introduction of lidar (based on ODD)

Level 4

Level 4 introduces high automation and builds from L3 Operation Design Domain (ODD), enabling the vehicle to operate autonomously within defined scenarios and in a specific ODD.

Level 5

Level 5 represents full autonomy, with the vehicle capable of navigating all scenarios without human intervention.

Customer Functions – Parking and Non-Driving

Besides those functions oriented toward vehicle-driver interaction, several other customer functions place demands on ADAS and affect vehicle tiers.

Parking Functions

ADAS vehicles include customer functions to provide parking visualization and partial automation. Visualization functions provide real-time feedback on the vehicle's surroundings through in-car displays during parking, and automation assists in executing parking tasks. Advanced parking systems support L4 parking, which includes valet parking.

“Away” Functions

Advanced vehicles include customer functions that monitor and protect the vehicle when the driver is away. They include remote surveillance capability, which continuously records and monitors the surroundings. An advanced function detects break-in or tampering and responds by alerting the owner or deterring the intruder/vandal with a visible or audible warning. Away functions are usually integrated with a mobile application on the car owner's smartphone, which allows viewing of live footage and the ability to review recorded events.

Driver and Occupant Monitoring Functions

DMS and OMS rely on in-cabin sensors that provide views of the driver and occupants in both visible light and dark conditions with infrared (IR) illuminators. The alertness of the driver is monitored by tracking eye movements and facial expressions to detect signs of drowsiness, distraction or impairment. Passengers may also be monitored to check, for example, whether seatbelts are fastened. The ability to detect the number of passengers in the car helps determine eligibility for high-occupancy vehicle lanes.

Audio Processing

ADAS may use audio analytics to process emergency vehicle sirens and the horns of nearby vehicles to generate additional cues. Voice recognition and synthesis are extensively used for hands-free control and interaction. Examples include navigation (entering destinations and points of interest), climate control, activating driving modes and advanced commands like adjusting windows and mirrors.

OEM Functions

To advance research and development of ADAS, vehicle manufacturers often insert specific functions such as the following:

- Advanced data collection and analysis tools – For example, active learning and map crowdsourcing to gather real-world driving data for map building and improvement of ADAS algorithms.
- Additional sensors and shadow algorithms – Sensors and "shadow" algorithms (including machine learning algorithms) deployed for testing but not actuating the vehicle, in preparation for future use.
- Event recording – Algorithms that monitor unusual events in the environment of the vehicle and record sensor data for future enhancements.

Processing the Algorithms and AI Workloads for Customer Functions

Thus, the variety in customer functions across ADAS levels drives different sensor configurations. It also drives different algorithms to execute those functions.

Figure 2 depicts the ADAS architecture, emphasizing the types of algorithms executed and the hardware blocks where execution takes place:

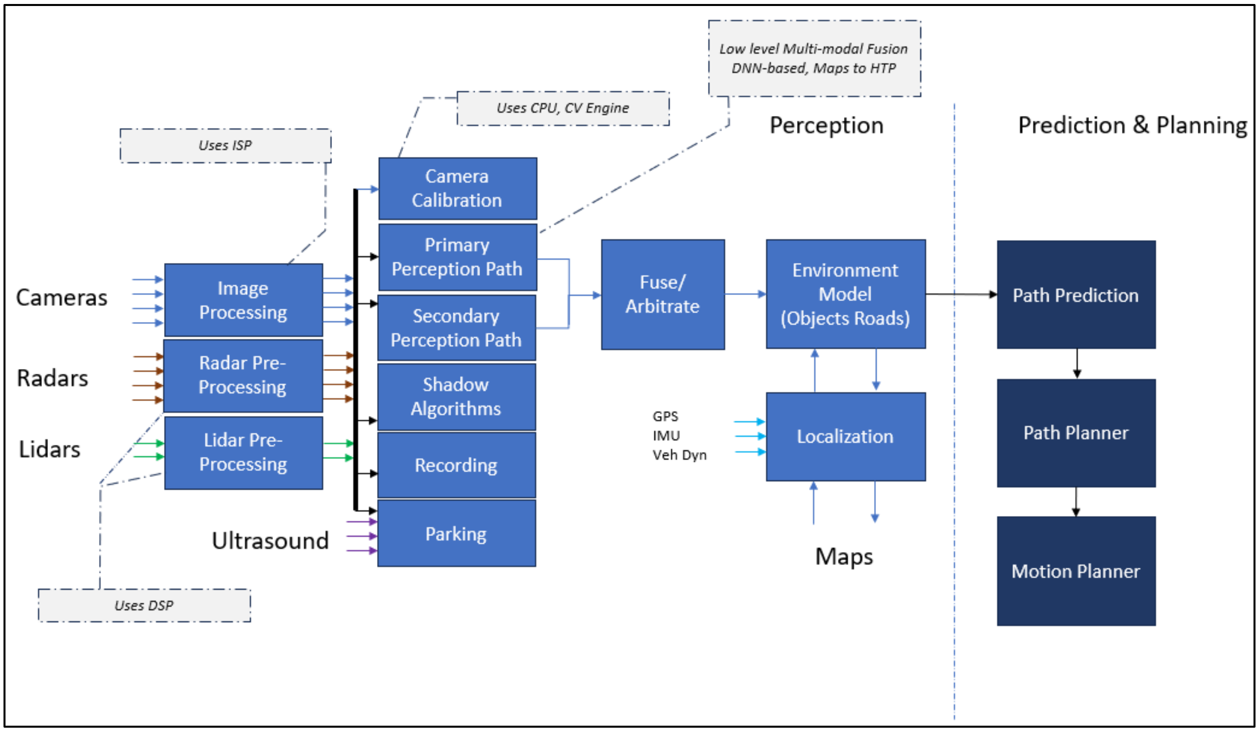

Figure 2: ADAS stack architecture

ADAS Sensors and Pre-processing Algorithms

As shown on the left side of Figure 2, an ADAS combines data from multiple sensors to comprehend the vehicle's surroundings. The common sensors used in ADAS include cameras, radars, lidars, GNSS receivers, inertial measurement units (IMU) and ultrasonic sensors (USS). (See Figure 3.) Time synchronization across sensor inputs may also be required.

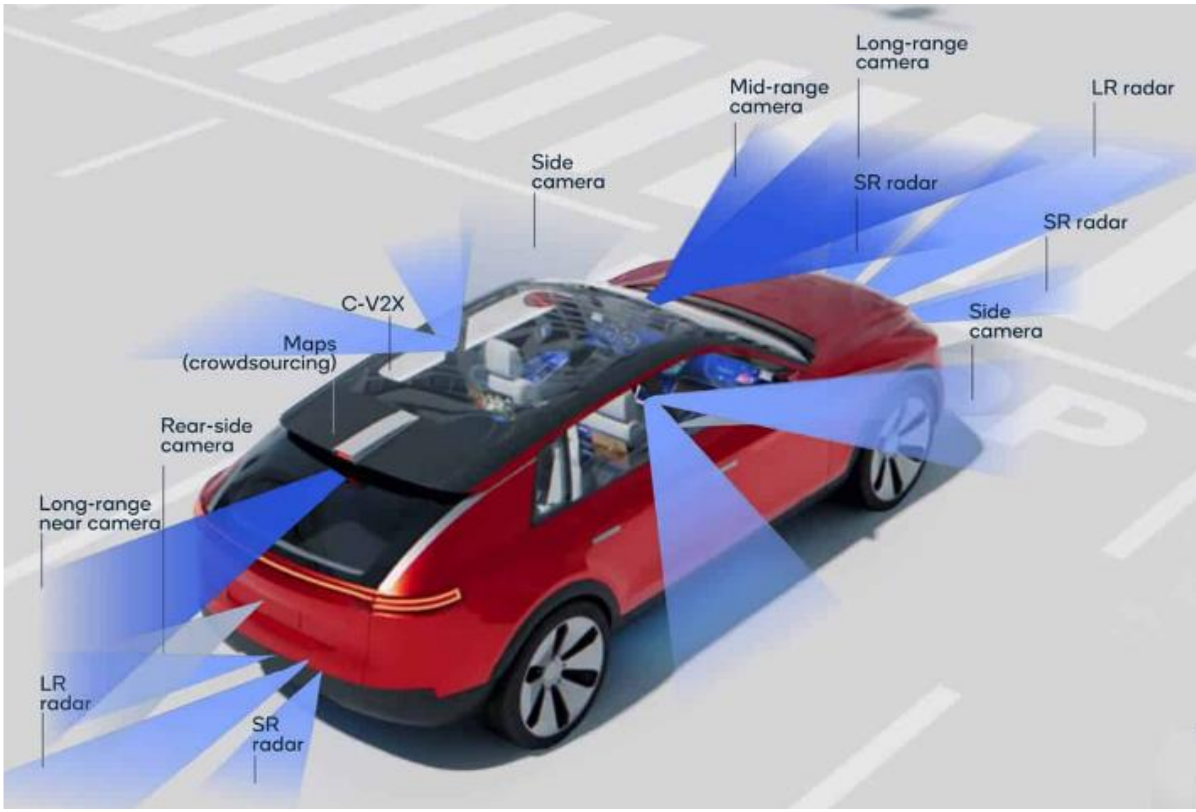

Figure 3: ADAS sensors

In the pre-processing step, camera images are enhanced, de-warped and scaled according to the requirements of the downstream algorithms. Radar detections are transformed into objects using filtering, clustering and tracking techniques. Lidar data is pre-processed using point cloud processing algorithms such as filtering, outlier removal and voxelization. Calibration algorithms monitor sensors to periodically re-estimate intrinsics like temperature-dependent optical properties, and extrinsics like relative changes in sensor orientations due to vibration.

Sensor Perception Algorithms

The middle of Figure 2 shows the role of the sensor perception algorithms.

The primary perception algorithm is typically a deep CNN / transformer network that fuses data from multiple sensors and produces perception outputs in a bird's-eye view (BEV) space. A secondary perception path relies on non-parametric algorithms such as depth-from-stereo and optical flow for object detection; that adds algorithmic diversity to improve reliability. Learn more about [ADAS perception algorithms].

Objects detected from multiple sensors may be fused using a grid fusion technique to produce a second representation of the vehicle's environment. An arbitrator fuses the models of the environment between the primary and secondary paths in L3 systems, providing higher levels of reliability.

Prediction and Planning Algorithms

The right side of Figure 2 depicts the behavioral layer and AI planner algorithms. They are responsible for predicting behavioral intentions of other dynamic agents (vehicles, pedestrians, animals) sharing the road with the ego-vehicle and determining a policy or sequence of policies for the ego-vehicle to follow.

The baseline architecture of the component consists of a separate intention prediction block that takes the environmental model as input. The block then generates a probabilistic intention as output, followed by a behavioral planner responsible for searching a sequence of driving policies. The behavior prediction block can use a transformer-like architecture with cross-attention among the multi-modal inputs; namely, agent tracks and a road model. The behavioral planning block can range from a rule-based system to a searcher or an AI-augmented searcher.

Steering Clear of Purpose-Built ADAS

Recall the main constraints that OEMs and suppliers face: performance, power consumption and scalability. If they want high-performing ADAS, they must accept high power consumption. If they want to stick to a power budget, they must sacrifice performance. And if they want ADAS that will scale across their product line, they must use a different, purpose-built processor for each vehicle tier.

Enter Snapdragon Ride.

Snapdragon Ride – A Heterogeneous Compute Architecture for ADAS

Snapdragon Ride is an SoC custom-designed for ADAS. Its heterogeneous processor cores, dedicated processing engines and memory and peripheral interfaces deliver the performance needed for responsive ADAS on a modest power budget. The SoC executes tens of AI and non-AI algorithms running in parallel and scales across entry-level, mid-level and advanced vehicle tiers, freeing OEMs and suppliers from single-model ADAS.

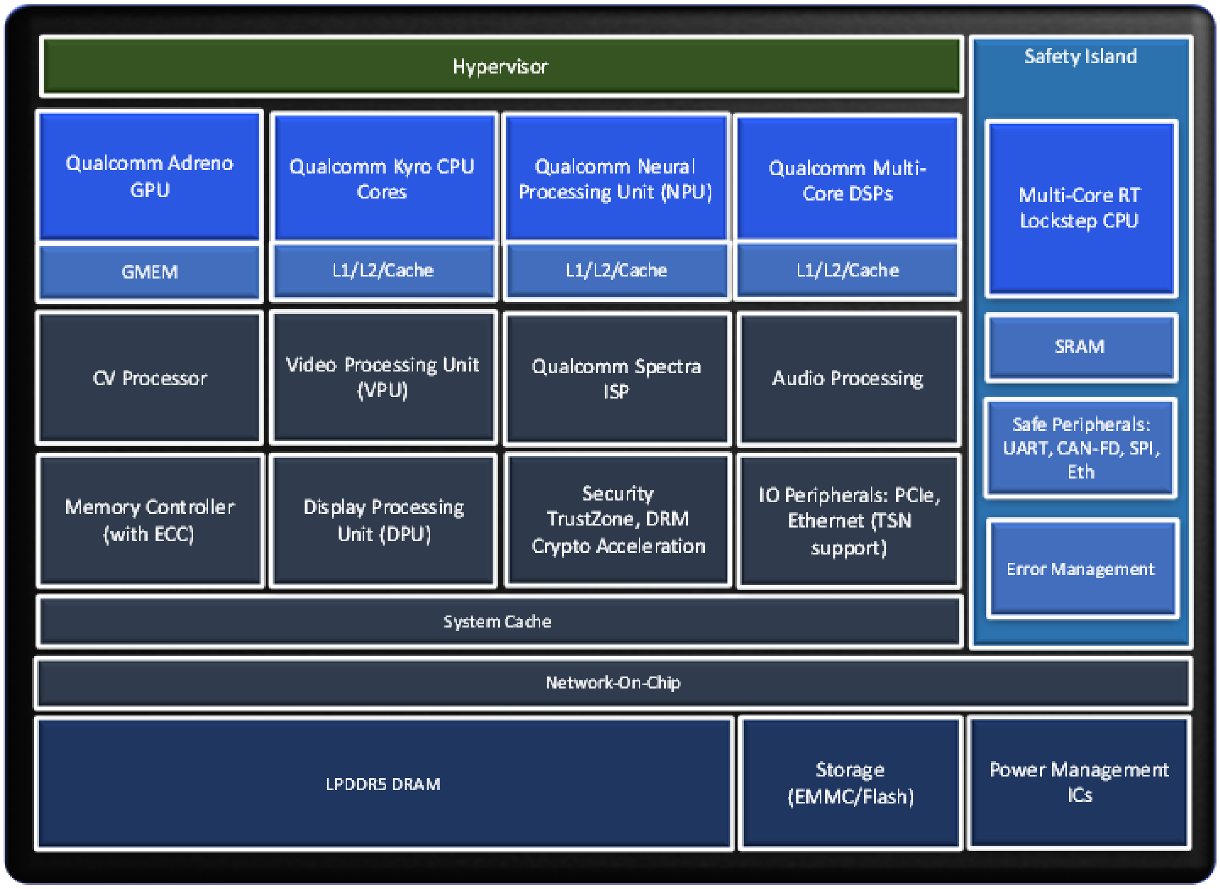

Figure 4: Snapdragon Ride architecture

Heterogeneous Processors

Snapdragon Ride includes Qualcomm® Kryo™ CPU cores, Qualcomm® Adreno™ GPU, Qualcomm® Hexagon™ neural processing unit (NPU) and multi-core DSPs. Its heterogeneity is suited to the variety of workloads and algorithms described above. (See bright blue row of Figure 4.)

Qualcomm Adreno GPU

GPUs excel at rendering parking visualizations from arbitrary viewpoints, providing realistic and immersive graphics for virtual environments. They perform well on compute algorithms such as point cloud filtering and clustering algorithms (used to process lidar data). They are well suited to grid fusion, which combines detections from sources like camera, radar and lidar to create a unified representation known as an occupancy grid. GPUs are ideal for Monte Carlo tree search algorithms used in behavioral planning and they enable quick, accurate decision making in complex scenarios.

Qualcomm Kryo CPU

Based on the Arm instruction set, the Kyro CPUs run the primary and secondary operating system and support functions including the following:

- Coordinating the control flows between algorithms running on various processing units
- Managing the IO and communication peripherals, including data transfer between the peripherals and compute blocks

- Supervising the states of the sensors, their calibration and synchronization
- Processing general-purpose compute tasks involving small data blocks, tracking, sensor fusion and decision-making algorithms that are rule-based and highly conditional
- Interpreting, manipulating and pre- and post-processing data structures exchanged between the compute blocks and accelerators
- Controlling the system, including monitoring system health and safety, and taking appropriate corrective actions
- Performing vehicle-level functions and diagnostics over an interface to the rest of the vehicle's electronics

Qualcomm Hexagon NPU

Snapdragon Ride contains one or more NPUs for executing multiple AI algorithms in parallel. Each NPU includes multiple Hexagon Vector Extension (HVX) accelerators and a Hexagon Matrix eXtension (HMX) accelerator custom-designed to execute AI workloads heavy on vector math and matrix math.

NPUs excel at handling large convolutional neural networks (CNNs) and multi-layer transformer networks, commonly used in perception and behavior prediction and planning. Their architecture is suited to the attention mechanisms and parallel computations of the transformer networks used for ADAS.

The optimized design of the NPU allows for fast and accurate processing of deep learning networks focused on image patches for semantic classification. That enables real-time analysis and classification of images.

The Behavior Prediction Networks and Path Planning networks at the heart of autonomous systems can quickly process and analyze sensor data on the NPU. In end-to-end AI architectures, the NPU can handle the entire AI pipeline efficiently, from data pre-processing to inference.

Although benchmark data comparing NPUs to GPUs may vary depending on architecture and implementation, NPUs generally offer advantages for AI workloads and performance (see Figure 5).

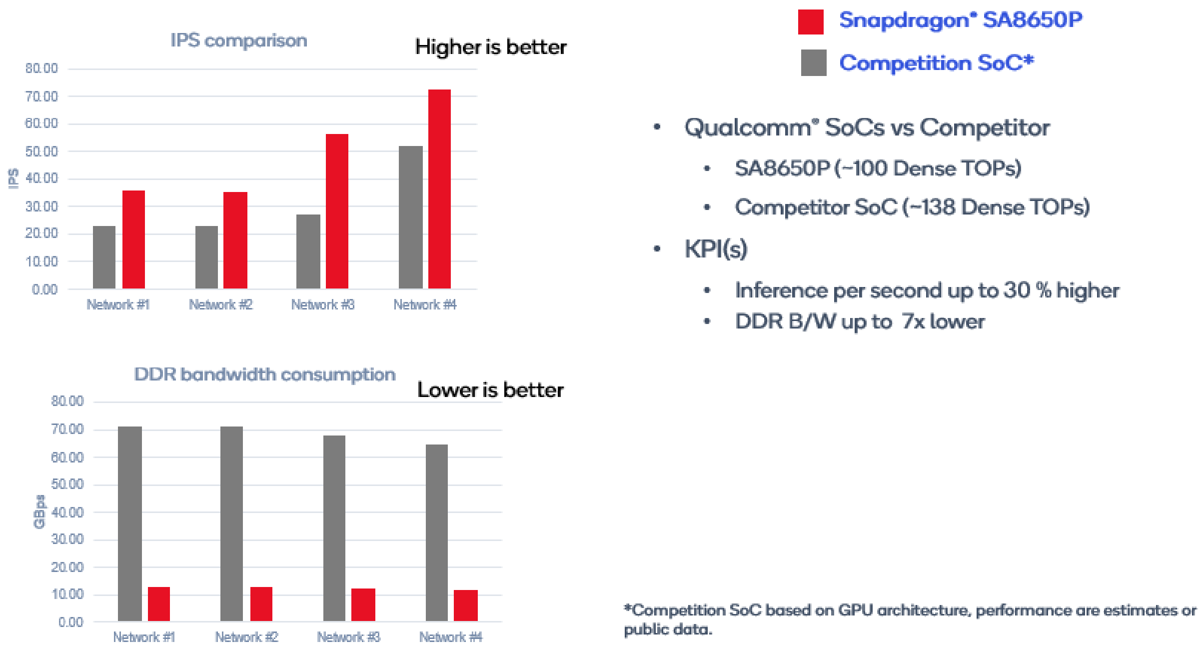

Figure 5: Network benchmarks - Snapdragon Ride vs. competitors

Some AI algorithms are more deadline-sensitive than others. For example, networks that determine the depth and speed of an object in the path of the ego-vehicle may need to execute with a tighter deadline than other networks. The compute requirements and frequency of execution of the networks may also vary widely, resulting in a wide range of execution times. Given the mixed priority requirement and wide range of execution times, it is often essential to have multiple NPUs executing simultaneously.

Qualcomm Multi-core DSPs

The multi-core digital signal processors (DSPs) in the ADAS engine are ideal for algorithms including the following:

- Radar pre-processing – Analyzing radar signals to detect and track objects in the environment
- Simultaneous localization and mapping (SLAM) – For autonomous navigation, utilizing sensor fusion and Kalman filtering
- Audio processing – For customer functions such as those requiring voice recognition and synthesis

Dedicated Hardware Engines

Supplementing those heterogeneous processors on Snapdragon Ride are a number of dedicated hardware engines (dark blue row in Figure 4) for computer vision and for processing images, audio and video.

Computer Vision Processor

The computer vision processor is optimized for real-time, non-parametric processing of visual data. Supported algorithms include the following:

- Real-time camera calibration – Hardware-based feature detection, descriptor generation and feature matching for intrinsic and extrinsic camera calibration. Necessary for accurate perception and understanding of the visual environment.
- Ego-motion estimation and localization – Also relying on feature detection, descriptors and matching
- Object tracking across frames – For object detection and tracking
- Dense and semi-dense optical flow – To provide information about the motion of objects in a scene for tracking and scene understanding. Also useful for eye tracking and head tracking in driver monitoring functions, and for detecting intruders and vandals when the vehicle is parked.
- Depth from stereo – Estimation from stereo images, allowing accurate reconstruction of the 3D structure of the scene in parking
- Depth from motion – For estimating road surface and object depth

Video Processing Unit (VPU)

The VPU is designed with high-quality video encoding and decoding capabilities, implemented in hardware. It supports multiple compression standards, including MPEG4, H.263, H.264 and H.265 (also known as HEVC). The engine is optimized to handle real-time encoding and decoding of high-resolution video at high frame rates.

Through triggered event recording, the video engine in the VPU allows the system to start recording upon specific events or triggers. That can be useful in surveillance systems, where video recording is triggered based on motion detection, sensor inputs or other predefined events.

Remote surveillance enables users to watch and access video from remote locations. It enables efficient encoding and streaming of feeds, with smooth, high-quality transmission over networks. This is particularly useful in surveillance systems and remote monitoring.

Qualcomm® Spectra™ Image Signal Processor (ISP)

In ADAS applications, the ISP must handle not only perception algorithms, where accuracy of the data is important, but also visualization applications, where naturalness of image is important. The Qualcomm Spectra ISP is designed to support multiple pipelines, including human and machine perception. Optimizations and applications include:

- Color conversion and de-mosaicing – For reconstructing the full-color image from captured raw data
- Exposure control and white balancing – To capture all objects, obstacles, pedestrians, traffic signs and traffic lights around the vehicle accurately, irrespective of the time of the day, lighting conditions and vehicle speeds
- Tone mapping – For adjusting the dynamic range of the image, making sure that details in both bright and dark areas are preserved. That enhances the visibility of objects in high-contrast scenes such as roads with both dimly lit and well-lit conditions.
- Aberration correction – To compensate for chromatic aberrations and distortions introduced by lenses; for example, when wide-field-of-view cameras are used around the vehicle
- Algorithms such as gamma correction, image sharpening, noise reduction and color enhancement – For higher-quality display and human viewing
- Image cropping and down-scaling – To aid in focusing perception algorithms on specific areas of interest, and for storage

- Warping/de-warping – For geometric transformations such as perspective correction, image rectification and distortion correction
- Image format conversion – For easy integration and compatibility with other systems and applications

Audio Processing

Snapdragon Ride supports a dedicated audio processing block designed to bridge mixed-criticality workloads – ADAS and infotainment domains – with a single integrated SoC platform.

The audio processing block offers these optimizations for ADAS:

- Zonal audio processing – For independent control and processing of audio in different zones or areas. This is useful in applications where audio needs to be directed or isolated to specific regions, such as in car audio systems or smart home setups.
- Audio chimes generation – For notifications, alerts or user feedback. Chimes can be customized and played in response to specific events or user interactions.
- Keyword detection – Continuous listening for specific keywords or wake words. Devices can then activate or respond to specific commands or triggers, even in low-power or standby modes.

As functions become more capable and ODDs further expand, additional audio processing use cases may arise, such as listening for emergency vehicles and flagging the system to follow local rules in a given traffic situation.

Additional Processor Characteristics

Besides the heterogeneous processors and dedicated hardware engines, other characteristics of Snapdragon Ride play a role in performance, data movement and power consumption.

Memory Architecture

The system memory architecture plays a critical role in ensuring that each processing block gets adequate data bandwidth.

Snapdragon Ride relies on multi-tiered memory: at the lowest level, all processing blocks and peripheral interfaces have access to shared LPDDR5 DRAM. However, the processors and hardware accelerators store their frequently accessed data in a locally accessible memory bank commonly referred to as Level 1 (L1) or tightly coupled memory (TCM). The size of L1/TCMs is individually optimized for each processor or accelerator. Processors such as the CPU and NPUs also have access to a slower, larger Level 2 (L2) cache/SRAM and a large system cache that is shared across multiple processors and hardware blocks. Error correction techniques are used across the architecture to preserve data integrity and reliability, which is vital for ADAS. Powerful compression engines located on the primary data processing blocks ensure that image/video and AI (activation) data is compressed for higher efficiency when stored in DRAM, pushing effective memory throughputs even higher.

Interconnect

An optimized, high-performance interconnect architecture provides efficient, high-bandwidth data movement among data producers, consumers and processors on the SoC. That includes CPUs, GPU, NPUs, dedicated engines, sensors, memory controllers and peripherals.

The interconnect can transfer critical data at lower latency while also ensuring that no sensor data is lost and all the compute blocks are sufficiently fed with the data required to meet their compute timelines. As it moves sensitive data among sensors and compute blocks, the interconnect imposes safety checks like data ECC protection and safety features like time-outs.

Power Efficiency

The power ICs and power grid ensure that the entire SoC can scale from low-power parking and sleeping states to the compute-heavy driving state. Away functions like camera-based security monitoring require always-on operation, forcing automakers to contend with parasitic power drain and functional power requirements. In electric vehicles, there is a reserve for these features, but every watt consumed affects range. Leaving vehicles sitting unplugged may present vehicle owners with an unpleasant surprise after a long vacation away. In internal combustion engine (ICE) vehicles, the risk of the battery losing its starting power is greater than the risk to driving range, but that type of power loss affects all vehicles.

Learn More

Snapdragon Ride is a family of SoCs designed for the heterogeneous computing and parallel processing demanded by the wide mixture of algorithms used in ADAS.

Many AI algorithms map efficiently to the collection of Qualcomm Hexagon NPUs, which are custom-designed for sustained, high-performance inference. The NPU is designed to be extensible as new neural architectures emerge. The CPUs are based on the Qualcomm IP/architectures, and the GPU is based on the Qualcomm Adreno architecture.

A variety of dedicated hardware accelerators optimized for image tasks, video and computer vision share the heavy lifting needed for high-resolution pixel processing. An efficient interconnect and a specialized memory architecture ensure efficient data movement to and from the peripherals and across the SoC. The integration of high-performance communication modules further facilitates seamless connectivity, serving as a communication bridge to the world.

The combination of heterogeneous processors and dedicated hardware engines in Snapdragon Ride offers automotive OEMs and Tier-1 suppliers competitive advantages in performance, power efficiency and scalability.

Learn more about Snapdragon Ride.