Imagination

GPU | CPU | AI

White Paper

THE USE OF GPU COMPUTE IN AUTOMOTIVE

Gilberto Rodriguez, Director of Product Management, Imagination

November 2023

Intro

The pace of innovation in automotive is accelerating. Electrification, advanced driver assistance systems (ADAS) and vehicle connectivity are revolutionising the in-car experience, which is now largely determined by the capabilities of the car's software and electronic hardware.

When a vehicle can receive software upgrades while it is on the road, the electronic control units (ECUs) that are designed today need to have the performance and flexibility to run the workloads of the next decade. A well-designed control system therefore needs the adaptability to take advantage of the strengths of individual components and, depending on the workload mix, assign the right task to the right processor to optimise efficiency whilst delivering the right level of performance and safety required.

A single automotive ECU typically includes specialised AI accelerators, arrays of application and real-time CPUs as well as a GPU. The AI accelerators efficiently process the perception task of the ADAS – working out the car's surroundings - in particular object detection and semantic segmentation. The CPU addresses decision making and sequential control tasks as well as running the main application. The GPU has been commonplace in cars for over a decade thanks to its ability to deliver smooth, responsive user interfaces to the cockpit and infotainment systems.

But, as has been discovered in other industries, such as the data centre, GPUs are good for so much more than just graphics processing. Their general-purpose programmability and high performance for parallel dynamic compute tasks make them a valid processing option for a wide range of workloads that form the backbone of ADAS and autonomous vehicles.

So, when the GPU sits alongside the AI accelerator which might deliver over 100 TOPS of compute, as well as CPUs, what compute workloads could, and should the GPU be used for? This blog sets out why, where, and how GPU compute can be deployed in a vehicle.

GPU compute benefits for automotive

Flexibility and Programmability

The GPU is a flexible and programmable parallel accelerator that can be used for a diverse array of tasks. The primary task of the GPU in the vehicle is graphics and delivering a great in-car multimedia experience. But a GPU can also be used effectively in a vehicle's ADAS to process general purpose compute tasks. GPUs are present in multiple chips in the car adding a flexibility to the car architecture that enables a software-defined set of features. A single GPU can be dedicated specifically to a human-machine interface (HMI) function, with another GPU supporting ADAS; or the same GPU can be used to handle both graphics and compute-oriented applications.

For software developers looking to create high performance applications that run on the GPU, industry-standard APIs such as OpenCL™, Vulkan® and OpenGL® provide a stable programming interface. GPU hardware, software and libraries provide the mechanism to achieve maximum efficiency and tight control of scheduling and memory management.
The use of APIs reduces complexity and gives freedom from the hardware to the software developer. They make it easier to take advantage of the inherent parallelisms of GPUs and to port code to different platforms.

There is a growing ecosystem of frameworks and libraries with OpenCL back-ends that give a quick time to market as well as an opportunity for higher-level optimization and integration as part of a heterogeneous compute system. It includes AI deployment environments as well as computer vision and other general-purpose compute libraries.

A wide range of companies are supporting or developing applications for GPU compute because of the GPU's high density compute and flexibility. The ecosystem is active and collaborative, seeking to enhance the utility of GPUs to the benefit of the automotive sector as a whole. Porting from Digital Signal Processor code, CUDA, TensorFlow or Pytorch (among others) is becoming easier and better supported by the extended community of developers creating both open-source and in-house solutions.

The programmability and flexibility of the GPU is very different from the type of processing offered by an AI accelerator that sacrifices flexibility for extreme levels of performance for specific types of workloads. High-performance electronic control units will include a mix of both GPUs and AI accelerators, giving software developers the right balance between performance and flexibility to create differentiated in-car experiences that can evolve with the future.

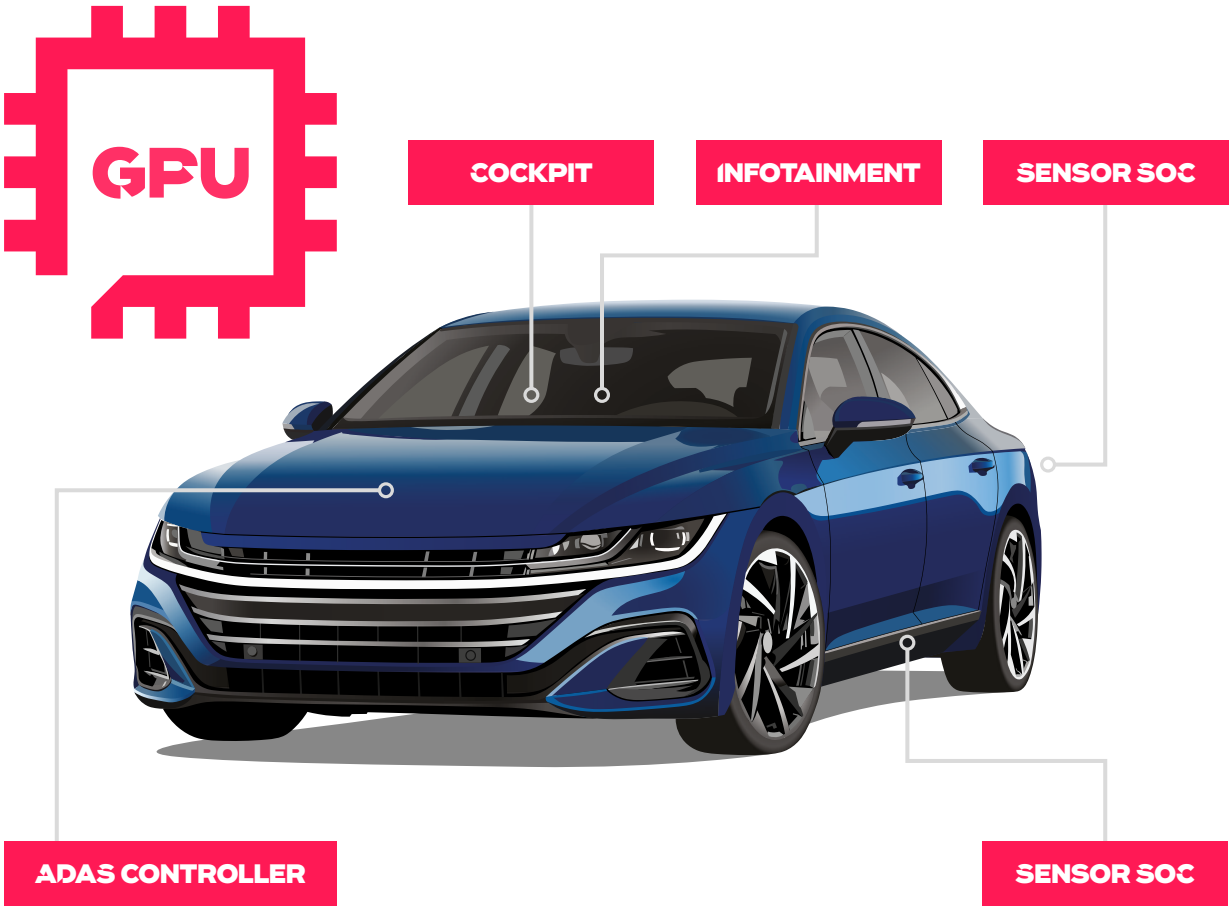
GPU compute benefits for automotive

Scalability

Within an individual vehicle, the same GPU architecture can scale to handle the increased range of tasks outlined above. Depending on the performance required for the target use case, hardware architects can choose between GPU configurations with different numbers of processing units or different numbers of cores.

For example, in a single car an architect could choose to deploy a small, fillrate focused GPU running the multimedia display in the infotainment unit; a dual-core GPU in the cockpit that handles the safety-critical display on one core and supports the driver monitoring functionality with the other core; and then a powerful multi-core GPU in the ADAS controller to provide high-performance, programmable compute. One GPU architecture from one supplier can scale to fit all of these applications, reducing complexity in hardware design, software development and purchasing.

OEMs can use the same principle of GPU scalability to offer different infotainment and ADAS options across the vehicle range without overly complicating the hardware design and software development and verification process.



Automotive tasks accelerated by the GPU

GPUs are parallel computation engines with hundreds of individual data lanes that are capable of processing independent, complex computations at the same time. This design, originally intended to meet the computational demands of the latest graphics trends, makes them a prime target for application programmers looking to accelerate non-graphic-related but still highly parallel tasks.

There is a wide range of tasks for which the parallelism of a GPU is well suited. In the case of ADAS functions, this includes AI pre-processing for data gathered by a vehicle's sensors (camera, radar, and lidar) which might include non-linear algebra as well as vector and data manipulation.

With the number of sensors in a vehicle only going up, other obvious workload choices for GPUs include the fusion of data from multiple sensors and AI post-processing tasks which demand high levels of compute and complex algorithm operations.

In addition to the typical graphics use cases, flexible GPU solutions can be efficiently deployed in vehicles to process more compute-oriented workloads such as:

- Neural network layers e.g. Region of Interest (ROI) Align, ROI Pooling, Non-Maximal Suppression (NMS)
- Floating point neural network processing
- Formatting of input/output data
- Dewarping
- Overlaying of metadata and source video/images
- Lidar display and ray tracing
- Lidar point cloud processing
- Image enhancement/camera stitching
- Perception and path-finding compute

Automotive tasks accelerated by the GPU

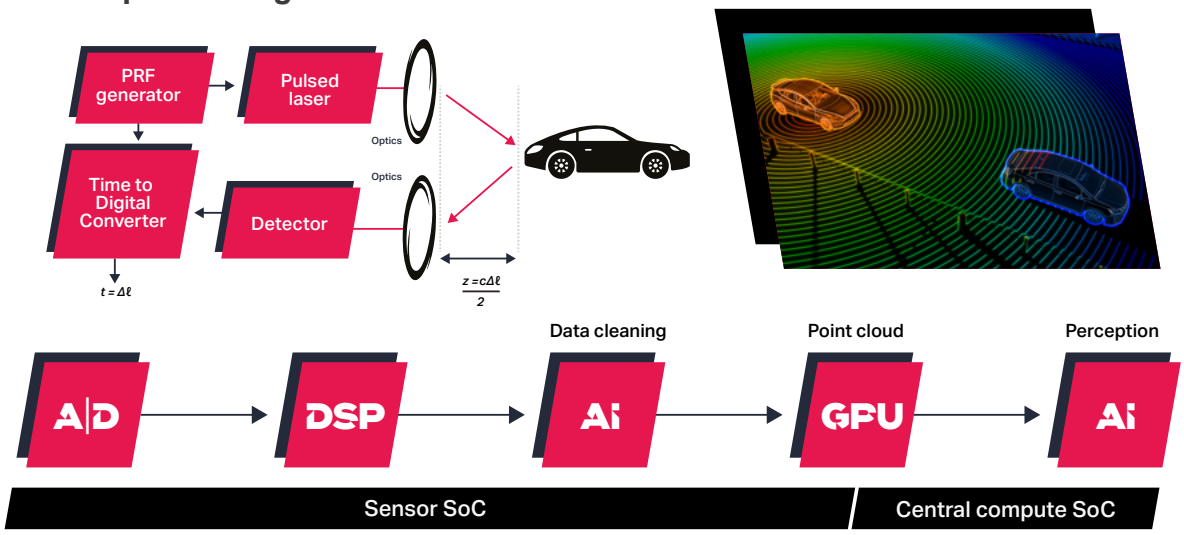
Use Case 1: Sensing compute

Compute in vehicles is centralising – but that's not to say that processing is no longer done next to the sensors. To get an understanding of the vehicle's surroundings, a vehicle on the road today could have a mixture of Lidar, Radar, and ultrasound sensors, as well as multiple cameras that combine to create a 360-degree view of the vehicle.

GPUs are well suited to performing video manipulation tasks efficiently, for example taking a camera stream, applying fish-eye correction with dewarping and 360 degree image stitching. They are also capable of processing the latest Lidar and Radar algorithms as a complement to the work done on the Digital Signal Processors (DSPs) prior to Perception tasks.

A typical GPU that can deal with this use case will require 1 TFLOPS of FP32 operations and 4 TOPS of DOT8 operations.

Sensor processing SoC



Use Case 2: Sensor fusion

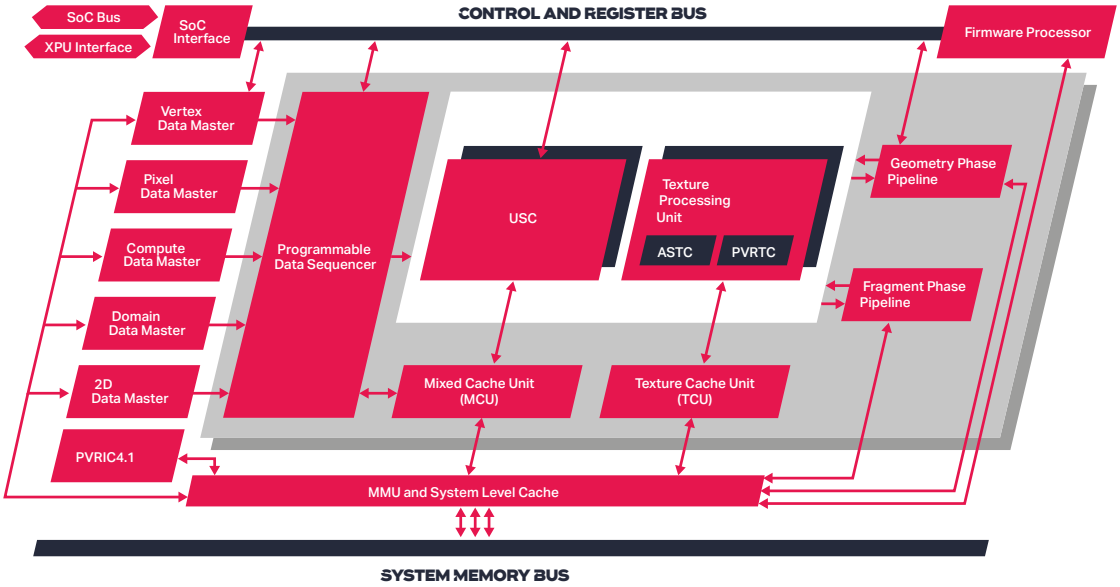
L4 ADAS, where the vehicle performs all driving tasks on pre-approved roads and in certain circumstances, requires a significant step-up: sensor fusion. This is where the data streams from all the sensors in the car are overlayed to build an accurate, digital representation of the surrounding world which the vehicle can then use for behaviour prediction and trajectory optimisation. These camera stitching tasks, along with the laying of metadata over source images, are workloads that can be accelerated with a GPU's high parallelism.

This use case will require a high-performance solution; typically over 10 TFLOPS FP32 and 40 int8 TOPS of GPGPU parallel programmable compute is recommended to add enough programmability to the system and be considered future proof.

How do IMG BXS GPUs do GPU compute?

IMG BXS is Imagination's range of functionally safe GPUs designed for the automotive market. IMG BXS builds on Imagination's leading low-power, high-performance GPU architecture that has been developed over the past thirty years and is deployed in billions of low power embedded devices while also delivering the performance required by high end L4 systems.

BXS-32-1024 Block Diagram



GPUs are not just for graphics …

Much of the silicon area of Imagination's GPUs is taken up by highly dense, optimised compute elements. Traditionally these compute parts of the GPU have targeted complex 3D user interface and game rendering, but they also form a very efficient general-purpose parallel compute engine.

Unified Shading Cluster (USC)

The USC is the compute heart of the GPU, a multi-threaded, programmable SIMT processor that can simultaneously process geometry, pixel, and compute data as well as 2D and data movement housekeeping tasks. More USCs in an Imagination GPU equate to higher compute performance for the GPU configuration. The highest-performance IMG BXS (single core) configuration, the IMG BXS-32-1024, has four USCs inside.

Ultra-wide, superscalar Arithmetic Logic Units (ALUs)

Each USC has two ultra-wide ALUs. These are based on a simpler, RISC-style form of ALU engine, and at 128 scalar data lanes wide they have more lanes packed into the same silicon area and power budget than earlier Imagination GPU architectures. High utilisation and efficiency are gained by embracing massive thread-level parallelism and simplifying algorithm optimisation efforts. Many ALU types (e.g Float, INT, Complex) and co-processors can operate concurrently based on data availability and instruction mix in the shaders/kernels.

Imagination's combination of a single instruction, multiple threads (SIMT) architecture with an ultra-wide ALU results in a very powerful compute engine with shorter inference times - ideal for automotive ADAS applications

How do IMG BXS GPUs do GPU compute?

Decentralised multi-core

Our advanced, decentralised multi-core architecture enables customers to scale GPU performance and unleash the full flexibility of a GPU. Each GPU in the IMG BXS range comes in a single, dual or quad-core configuration, delivering over 6 TFLOPS (and 24 TOPS) combined compute performance. The GPU cores can be programmed as a big GPU cluster or as independent cores to provide improved task isolation, concurrency, and more granular control of scheduling. This makes it easy to combine cockpit display with ADAS processing.

HyperLane Virtualisation

The other way to tap into the GPU's compute capabilities while also running graphics workloads is to take advantage of virtualisation. Each IMG BXS GPU can support up to eight virtual machines, each isolated in control registers and memory, with workload prioritisation and quality of service mechanisms applied through our HyperLane virtualisation solution.

Workload optimisation

IMG BXS GPUs achieve very high levels of utilisation on critical real-world workloads such as MatMul (Matrix Multiplication) or FFT (Fast Fourier Transforms) thanks to a combination of both their compute architecture and optimised software libraries.

Software support

We support all the main graphics and compute APIs with and without safety critical versions, Open GL® ES, OpenCL™, Vulkan® and OpenGL®. We have partnered with the safety-critical experts at CoreAVI to develop a world-class safety-critical driver for our automotive customers.

Safety and security

The BXS GPU range is ASIL-B certifiable, built according to ISO 26262 standards, and features hardware and software mechanisms that specifically improve functional safety and security.

IMG BXS GPUs are Imagination's safety-critical, ASIL B certifiable GPU range.

Click here to visit the product pages to find out more about Imagination's other GPU ranges.

Conclusion

The compute capabilities of a GPU are an asset to the automotive systems designer. Imagination has been delivering high-quality GPU IP into the automotive market for nearly two decades and over this time our IP has evolved to keep pace with the changing needs of increasingly functional vehicles. Our IMG BXS GPU range offers flexibility, scalability, functional safety, and efficient compute performance. To find out more, please contact our team today.

www.imaginationtech.com

Contact us now