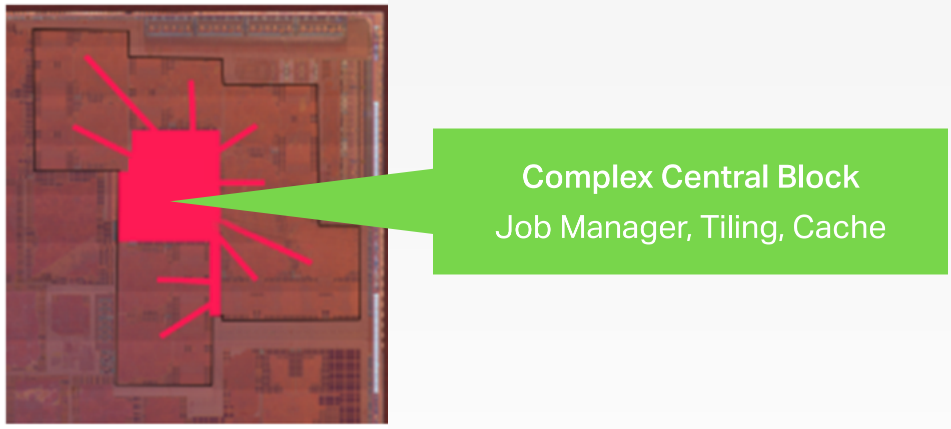
IMG B-Series

MULTI-CORE TECHNOLOGY

IMG B-Series introduces a decentralised, loosely-coupled approach to multi-core performance scalability.

Traditional approaches to GPU scalability

The traditional approach to GPU scalability is limited by the connections between centralised shared blocks and the shader cores. Typically, the shared logic includes a centralised memory data path, job manager and geometry tiling engines. The central dependency generates a star-network-style structure where all cores need to be connected to this single centralised entity. However, this causes issues with congestion and layout flexibility, as illustrated below.



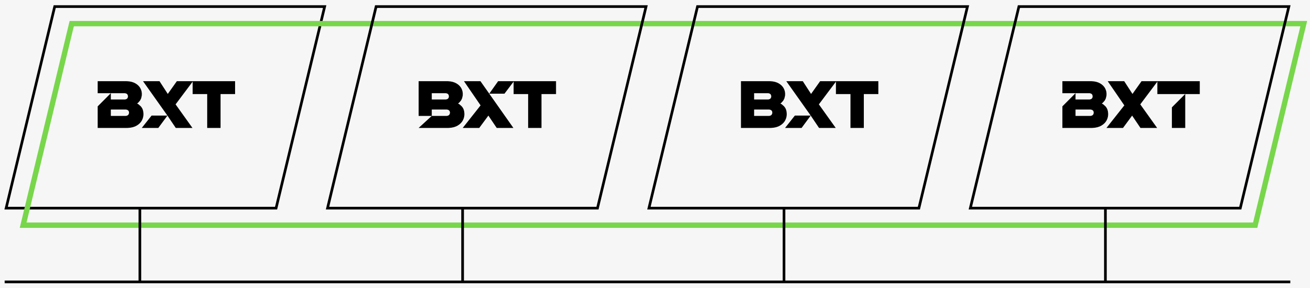
Traditional multi-shader core design with central block and star network

Imagination’s Decentralised Multi-Core

Design Concept

Imagination’s novel approach to multi-core instantiates a flexible number of GPU cores without a direct dependency on a connection to a central unit. In its simplest form this can be seen as multiple GPUs, which are present in an SoC design, but with the ability for cores to jointly work on compute and graphics processing.

As Imagination’s GPUs are based on tile-based deferred rendering it’s easy to understand how multiple GPUs can work together by having each core work on a different set of tiles which make up the total render target.



Imagination’s decentralised multi-core in a four-core configuration

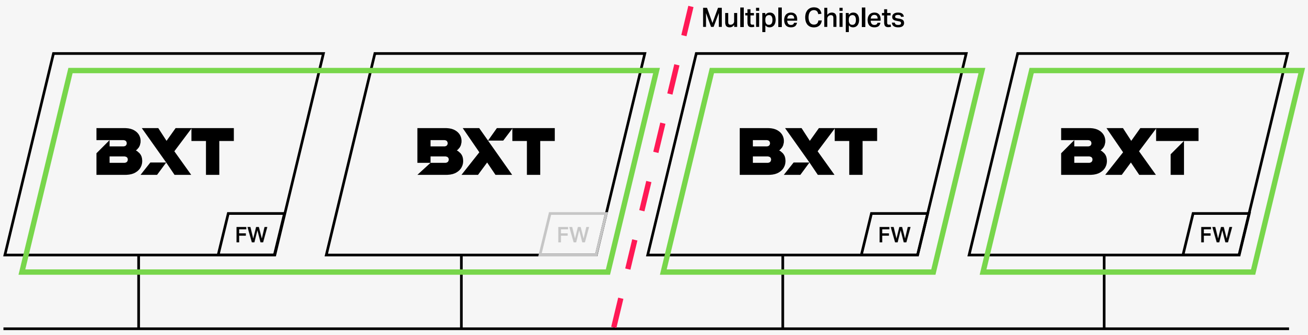
In effect, the cores are loosely coupled, only sharing command and tile buffer lists in memory and sharing the workload between them. As each core is designed as a full independent GPU it contains all the functionality required to self-manage and execute workloads based on priority. The fundamental change to previous layouts is that instead of a single core directly working on a workload we now have multiple GPU instances sharing command streams and jointly finishing the work as quickly as possible. By working on distinct regions of the work (the render target) in each GPU core our bandwidth efficiency is maintained, as each core continues to work on a coherent region of the screen, thus ensuring maximal cache hit efficiency. A similar approach also applies to other processing types, including geometry and compute processing, which can be assigned to different GPU cores for processing.

Physical Design Flexibility

Each GPU core sits directly on the SoC fabric, providing greater layout and design flexibility. Data is cached specific to the work processed by each GPU core, thus avoiding unnecessary data movement and duplication between cores, increasing efficiency.

By using an industry-standard bus fabric between the cores (e.g. AXI) we automatically benefit from latency tolerance. This means that the fabric can also be used to distribute the GPU cores across multiple chiplets, chips or even boards. This provides our customers with more flexibility in their design choices and allows for significant cost savings, as a customer can design a single chiplet (or chip) which can then be used to build up multiple performance points, simply by packaging one or more chiplets to scale GPU performance.

The cost saving is that only a single chiplet is made, enabling multiple markets simply through packaging. Historically, this would have meant taping out multiple chips at a significantly higher cost.



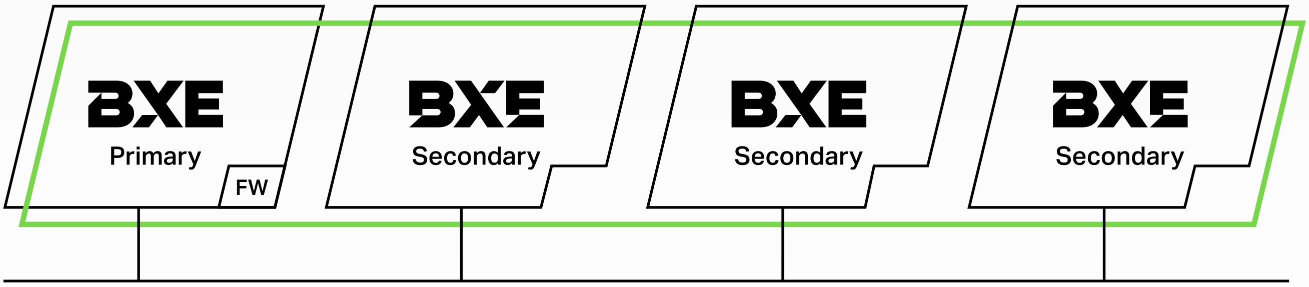
Imagination's decentralised multi-core showing flexible configuration of MC2+MC1+MC1 across two chiplets

Usage Flexibility

As each core is a full GPU, capable of working on its own or together, it allows for maximal usage flexibility. Cores can be used as separate GPU instances directly controlled by a driver and executing a specific workload, or cores can work together to form a higher-performance GPU. This capability can be used in combination with Imagination's HyperLane Technology to enable virtualization using priority-based scheduling mechanisms (single GPU mode) or also enabling simpler isolation approaches where a specific core can be assigned to a specific virtual machine (multiple individual GPUs). Of course, this also inherently equates to redundancy and in our BXS series, we use this to ensure ISO 26262 functional safety where GPUs can check each other's operation.

Cost Optimisation

For lower-cost market segments, GPU designs with replicated functionality are undesirable and unnecessary and this can be resolved by creating a single primary GPU core which includes all functionality and a secondary GPU core which omits unnecessary functionality, such as the firmware processor and geometry processing pipelines. This means that a primary/secondary configuration is significantly smaller than the primary multi-core configuration, while still having more than enough performance to process and prioritise geometry processing as well as handling firmware-based event handling for the multiple GPU cores. Effectively, this is a form of heterogeneous scaling where GPUs with different feature sets and capabilities work together efficiently.



Imagination's decentralised multi-core in a heterogeneous primary/secondary configuration
for optimal performance density (silicon area)

How would I benefit from the Decentralised Multi-Core technology?

Smartphone/tablet

The decentralised approach allows for higher block re-use and easier layouts by avoiding complex data paths to a single central point inside the GPU design, which otherwise must be redesigned for each performance point. Multiple performance points can also be enabled from a single chiplet design where a single chiplet targets a high-end mobile market, dual chiplets target premium mobile market and quad chiplets could target a beyond-mobile ultra-portable/tablet design.

Automotive

The inherent redundancy in the multi-core approach allows for low-cost ISO 26262 compliance where cores can check each other's correct operation. But where functional safety is not required the cores can be dynamically reconfigured to allow for usage cases which fully maximise the performance of the multiple GPU cores. Of course, performance scaling from an entry-level to a premium automotive system can be enabled at a much lower cost by packaging chiplets as required.

Digital Television (DTV) and Set-Top Box (STB)

For cost-sensitive markets, performance scaling can be optimised through the usage of a single primary GPU core with multiple secondary GPU cores. This can enable a premium 1080p GUI rendering solution, which can be combined using chiplet technology to deliver 4K GUI rendering solution simply by packaging multiple chiplets to scale up the performance.

Server

Servers require flexible performance scaling and this is particularly relevant to cloud gaming, where Imagination's flexible multi-core approach can be used to dynamically reconfigure GPU grids using HyperLane technology. A BXT MC4 design, running at 1.5GHz in 7nm, provides up to 6 TFLOPS of performance, which exceeds the capabilities of next-generation, mid-range consoles. Multiple B-Series GPU instances could be employed as a single cloud gaming or remote desktop chipset when combined with multiple video encoders. Using HyperLane technology, this allows dynamic performance budgets to be assigned to a variable number of users. For example, an MC4 can service up to 36 users with flexible performance allocation in a highly-efficient way with full privacy and security using hardware virtualization and even isolation.