Imagination

THE RISE OF CLOUD GAMING

Changing the Dynamic for Data Centre Gaming

Updated November 2023

Introduction

Gaming has always had a high barrier to entry – it requires an investment in relatively powerful hardware and then purchasing games on top of that. Cloud gaming changes that dynamic. It is simple, yet revolutionary. With cloud gaming users typically pay for a subscription service to access a game running remotely on a server in a data centre. This means that users do not need to download the game to their device, and they do not need expensive, high-end equipment to enjoy them – instead, games can be brought to the user on virtually any screen. While this seemed a farfetched concept only a decade ago, thanks to the prevalence of streaming TV and music services, it is almost a surprise that cloud gaming subscription services are not already the mainstream method used to access and play games.

That said, it is still early days for the cloud gaming industry, and some high-profile players have already entered and left the market. Even so, cloud gaming has moved from a niche gaming activity into the open, with a wider range of users and commercial options emerging. Cloud technologies are democratising gaming and the use of server performance to deliver superb graphics to lower-end user devices is bringing the joy of gaming to an even wider audience.

The expansion of 5G network coverage and the growing number of compatible handsets is the main driver to making cloud gaming an effective option. By combining cloud delivery through high-speed networks along with more performant and affordable 5G handsets, mobile users are poised to enjoy a revolution in the gaming experience.

In this white paper, we will explore the opportunities and challenges of this new world and show how Imagination's approach is ideally suited to the next generation of cloud gaming.

$5.13 BILLION
Cloud Gaming by 2023

2.8 BILLION
Active gamers in 2021

685 MILLION
Gamers in China
+135% CAGR growth
between 2020-2023*

*Growth areas driving cloud momentum.
(Source: Newzoo.com, Global Cloud Gaming Report, March 2021)

Introduction

Gaming ready

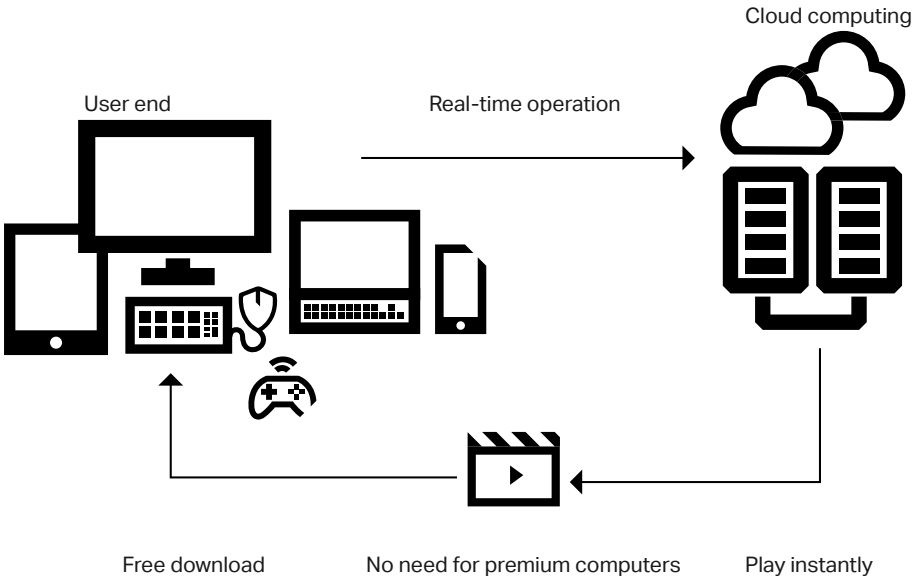
Imagination has decades of experience in creating market-leading immersive graphics for the world's most popular mobile games. Imagination's GPU architecture has been designed for outstanding graphics and with gamers in mind, so, logically, we understand the gamer as well as the infrastructure required. We offer Android compatibility through support for OpenGL® ES and Vulkan®, with performance and power efficiency at the heart of our approach. The addition of hardware-based DirectX® support into IMG DXD* means we also support popular Windows titles.

Security is delivered through our HyperLane Technology, providing hardware virtualization for complete application robustness. This makes it easier for cloud gaming providers to deploy across containers whatever their cloud platforms deployment method, while our growing ecosystem of partners have all the expertise required to help our customers get to market smoothly. Our GPUs also integrate with popular game engines such as Unreal Engine and Unity.

*IMG DXD is a GPU IP based on the published DirectX and Khronos API specifications. Hardware designed with this IP is expected to pass conformance testing.

What is cloud gaming?

Cloud gaming concept



Tech specs

- Framerate: 60fps
- Delay: <40ms
- Resolution: 1080P

Cloud gaming enables games that would normally need a high-powered, expensive PC to be played on much more accessible low-cost devices. In a cloud gaming scenario, the game is run on a server, utilising its high performance, and is then live-streamed to the remote user's handset. This enables a wider variety of endpoint devices to be used to game and removes the need for expensive consoles or desktop computers to experience a premium gaming experience. It also opens up the range of games that can be played, while enabling communities of players to play together. With this model, state-of-the-art graphics can be delivered to end-users efficiently and without the up-front cost of high-performance hardware.

As cloud gaming grows in popularity, companies offering cloud-gaming services are provided with new opportunities to monetise their offerings. For them to be successful, however, they must reduce the costs involved in providing the service, particularly in relation to server power consumption, heat management and cooling. To do this, they require technology that will maximise their server rack density and ultimately, deliver a more competitively priced, and profitable service.

Cloud gaming use cases

As the number of cloud gaming users has expanded, so has the range of use cases. Typically, the gaming market has been synonymous with dedicated, or “hardcore”, gamers. However, the relatively easy set-up and technical requirements for cloud gaming means this has diversified to include casual gamers looking to play a wider range of games across different game types and interaction levels.

Usage

Cloud gaming platforms

Cloud gaming cafe

Home devices (TV, tablet)

Ads - playable

Cloud-based game demo

Live streaming

Personas and the need for scaling

Cloud gamers can therefore be split into several personas. These are:

- **casual gamers** — playing less graphically intensive games for shorter periods;
- **premium gamers** — extensive use of team gaming, and visually immersive games. They have high visual expectations and therefore are demanding on graphics hardware;
- **hybrid gamers** — playing a mix of casual games and more graphically demanding team gaming.

The cloud gaming services offer different options dependent on the gamer's requirements and willingness to pay.

Lower service levels enable hybrid and casual gamers to access shared cloud gaming machines. This is a more affordable option for playing a wider range of games, but the disadvantage is that at times of peak demand it can be affected by connection glitching and performance dips.

For a premium service, gamers are typically buying access to a dedicated cloud-based gaming machine which will guarantee them performance levels, independent of the volume of users. Premium services can also bring gamers earlier access to new "AAA" titles. Premium gamers expect maximum quality on a variety of devices. As such, they are ready to pay for access to the highest-quality games, with the best visuals and immersive experiences.

This, in conjunction with the growth of users, has put additional demand on cloud providers to scale the number of users they can support, while offering a diversified range of gaming experiences. This flexibility requires a different approach to assigning workloads across the available server resources.

A day in the life of a cloud gaming server

Enabling dynamic demand over time

Gaming patterns of demand will vary across the day. As such, to satisfy specific customers' needs, the cloud infrastructure will need to manage these changing demand patterns and bring resources on-stream rapidly and correctly configured. The service will therefore have to cater for:

- **intense demand** — offering flexibility to provide for intensive gaming, even at peak hours;
- **low demand** — where at quieter periods of the day such as early morning (pre-rush hour) and mid-morning there is less strain on the servers;
- **dynamic allocation and cost-savings** — generating dynamic savings by powering down when peak gaming is not required, but maintaining the ability to bring server resources online when needed, rapidly and efficiently in a manner that is seamless and transparent to the end user.

Varying demand per application mix

The tasks required could involve handling casual gamers playing less graphically demanding games, right the way through the continuum of graphics intensity to premium gamers playing titles such as Honor of Kings™, Call of Duty Mobile™ and PUBG™. This requires the ability to:

- dynamically expand and contract with render reconfiguration;
- rapidly iterate;
- offer high resolution and immense scale;
- prepare for the deployment of ray tracing to offer exceptional graphics for premium tiers.

Key for diagrams appearing below

Casual — shared resources for lower quality of service expectations

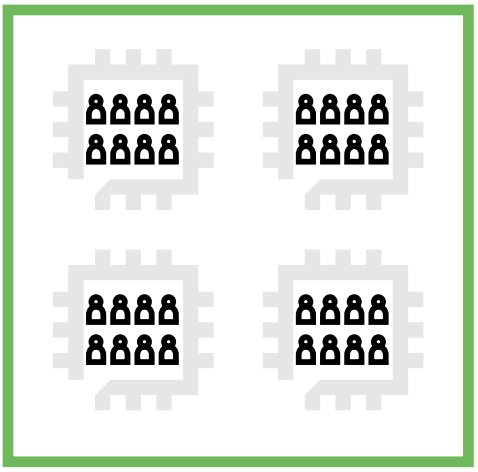
Premium — dedicated hardware per user

Hybrid — multiple users with shared resources

A day in the life of a cloud gaming server

Casual gamers

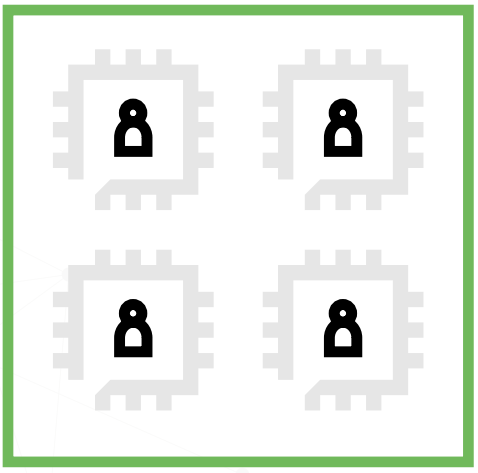
Where a typical desktop GPU caters for only a single user, the cloud enables multiple users to access multiple GPUs. Crucially, a multi-core GPU could be configured to act as several independent cores. In the example below, each core can support up to eight casual gamers, enabling up to a total of 32 gamers on a four-core GPU. This would suit lower-end, casual games and would be ideal for those looking to try out the cloud-gaming experience.

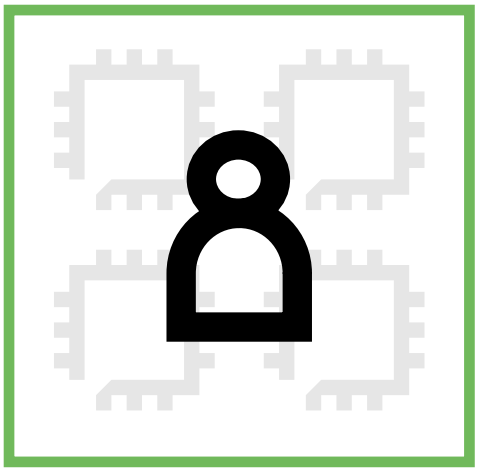


As gamers roll on and off the system, the multi-core approach can accommodate the changes dynamically, without needing to reboot or power down.

Premium gamers

For premium gamers, performance and low latency are critical, especially for team gaming, first-person shooters, and battle royales. Premium users can be assigned a single core, multiple cores, and even a full multi-core GPU to maximise their experience at all times.

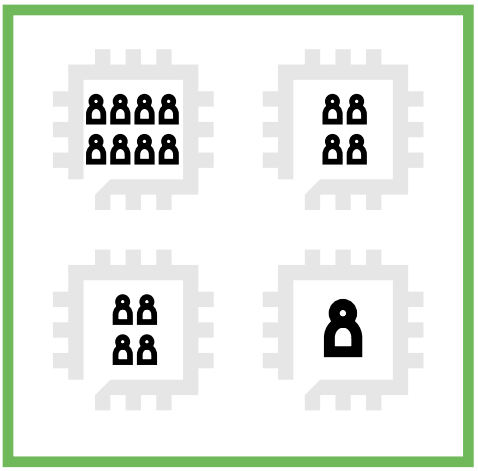




Hybrid gamers

In this instance, cores are assigned to different audiences and varied throughout the day to address demand and services. Cores can either be used for casual, standard gamers or reserved for premium gamers. These can also be mixed and matched, with some single cores and some higher service levels taking advantage of multiple cores assigned to their gaming experience.

It also enables extra service offerings such as reduced latency, and access to new titles, all of which can be dynamically managed, and monetised.



This range of user options and potential service offerings opens up new opportunities for the monetisation of cloud gaming, enabling service providers to offer services tailored to the needs of a set of end users who are increasingly diverse, but all demanding an enhanced gaming experience.

Cloud gaming infrastructure challenges

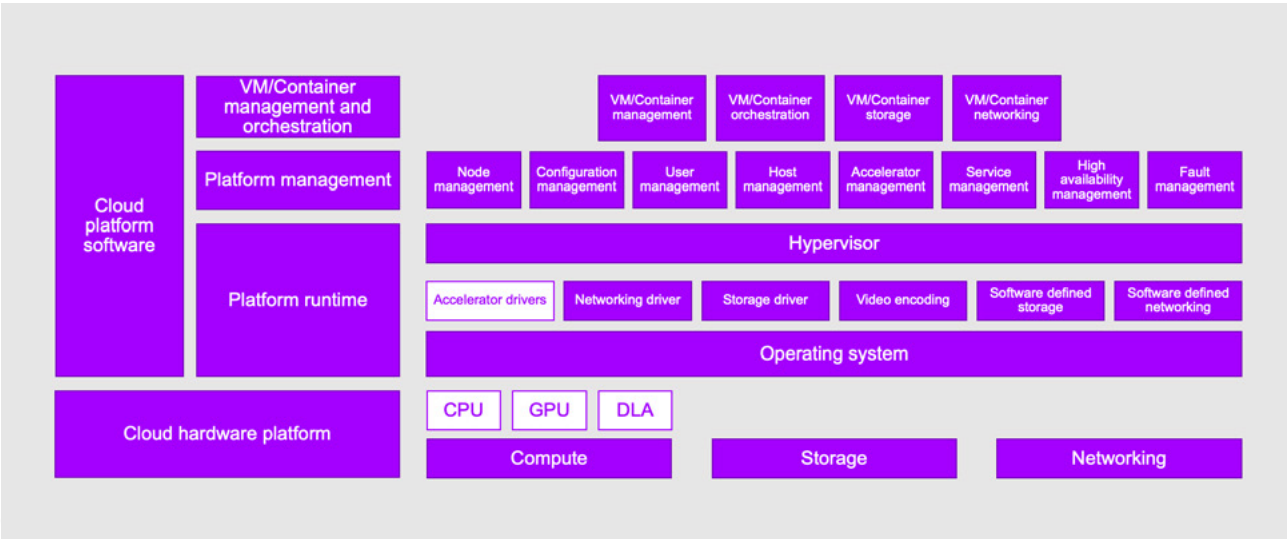
Cloud gaming is a unique computational challenge. As described above, demand will dynamically grow and shrink, with gamers coming online to play at different times of the day or when specific new games become available. Therefore, a different approach is needed to the traditional single PCI board-level approach, which only supports a single user at a time. Additional flexibility and granularity are required to avoid the need for large amounts of expensive and power-hungry graphics boards, which at quiet times could be unused.

The ability to dynamically expand capacity is vital, but as the footprint and power consumption of individual racks is limited, performance improvements must be delivered within the constraints of the existing facilities.

The aim is to increase the number of instances that can be handled within the same rack and to increase the dynamic scalability while maintaining end-customer satisfaction and providing the potential to boost service monetisation.

Typically, data centre structures are predominately based on Linux/x86 and Android servers. Imagination's GPUs have excellent support for both, in a platform-independent approach for maximum flexibility.

Cloud technology stack



The diagram above shows how a data centre is currently architected. It is possible to be modified to provide standard and ray-traced graphics and AI as needed while being equally able to work with x86 and Android servers.

Flexible GPU solutions are needed to enable the dynamic changes in render reconfiguration. In other words, this means taking a GPU or a cluster of GPU instances and re-adjusting them during the day to meet the needs of the gamer profiles mentioned above.

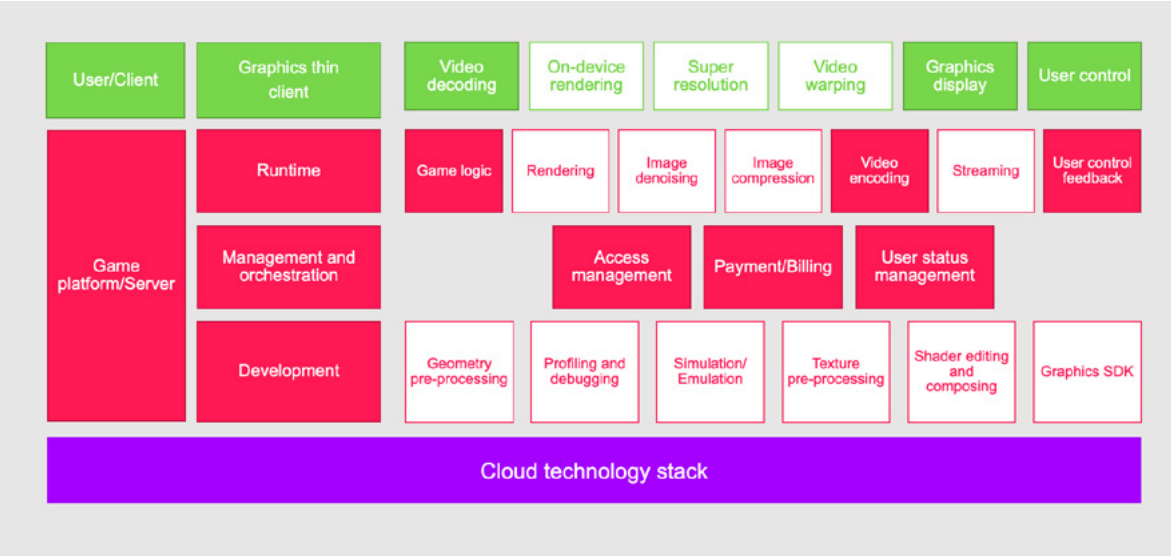
Enabling the casual gamer to share a GPU with other gamers, and at other times allowing multiple cores to be allocated to fewer users to ensure a premium experience for the intensive gamer, requires an architecture that offers virtualization.

The problems with existing GPUs

While cloud gaming is growing as a global opportunity, existing GPU technologies are not ideal for meeting growing demand. Conventional GPU desktop architectures consume too much power and generate too much heat, limiting the expansion opportunities within rack space and power envelopes. As such, a different approach is needed to address this growing demand and to effectively monetise this rapidly growing market.

Cloud gaming server and client architecture

Cloud gaming technology stack



Challenges for maximising the rack

Typical gaming cards have extensive power and fan requirements, and due to heat constraints, it may not be possible to pack them closely together. To address the changing demands on these racks – a different approach is required. As such, energy-efficient, yet high-performance GPUs that were originally designed for the mobile market are now being considered to address the scalable and dynamic requirements of cloud gaming.

Designed for area-efficiency

Traditional data centre designs come from the general-purpose compute space, where neither power draw nor the cost of heat dissipation and cooling is a factor. However, area- and power-efficient designs are now changing the dynamic for data-centre computing. These designs are specifically created to maximise concentration, allowing for greater performance density with minimal power consumption, and data centre companies around the globe are exploring this approach as the way forward.

By their nature, these embedded GPUs from a mobile heritage are low-energy and low-heat designs and require minimal cooling. These GPUs can be configured for fanless, low-power operation, as this is the environment for which they were intended: a thermal design point which had no physical space for fans and no airflow.

Designing for efficiency and scale

Enabling cloud gaming at scale

Imagination has designed its mobile GPUs to work across a wide range of performance points. With a range of multi-core solutions, it can provide graphics solutions that deliver high performance per Watt that can dynamically scale to meet the needs of gamers, without having to reduce the user experience due to power constraints.

The high-performance density and energy efficiency of Imagination's IP designs also means there is a reduced need for air or liquid cooling. Imagination's solution can contain many more cards, and many more gamers and therefore, deliver more paying gamers per rack unit than our competitors.

Efficiency meets visual quality

Imagination has built its graphics expertise in the power-sensitive mobile space, but we also focus on visual quality and optimal solutions for developers. As such, our latest GPUs have a range of features that boost performance and efficiency for gaming, such as Fragment Shading Rate, 2D Dual-Rate Texturing, Pipelined Data Masters and a performant RISC-V firmware processor.

If ray tracing is needed, our Photon architecture is more advanced than anything else in the market. The combination of our energy-efficient hardware and high-performance dedicated ray-tracing functionality is bringing a more immersive and realistic visual experience to mobile devices, while also simplifying the workflow for developers. This energy-efficient next-gen graphics experience can now also be harnessed to address the needs of cloud gaming. Thanks to the power of our Photon architecture, ray tracing can be delivered without significantly hampering performance levels, maximising visual quality for premium gaming experiences.

Dynamic cloud gaming architecture

Here we will look at how Imagination's GPUs can be used to provide a dynamic, power-efficient solution to address the needs of each gaming type.

The Imagination solution for next-generation cloud gaming

Cloud games demand high-performance networks, especially in terms of latency. Imagination IP designs:

- are designed for a distributed world of gamers;
- allow for process isolation through virtualization (via HyperLane Technology);
- can be quickly deployed and reconfigured, with a variety of specialised cores which can be deployed in GPU matrix primary-primary mode.

Some existing approaches can be difficult to reconfigure and slow to deploy. They can be difficult to scale and require extensive resources, both computational and human (architects and engineers), which increases running costs and lowers utilisation. Our technology, however, allows cloud gaming providers to quickly deploy, reconfigure and automatically allocate and de-allocate resources. As we explained earlier, the growth of 5G is enabling a revolution in how games can be delivered. As such, there is an increased need for supporting distributed data streaming to cater for super-high-demand gamers. The ultra-flexible range of IP provided by Imagination can again address these challenges.

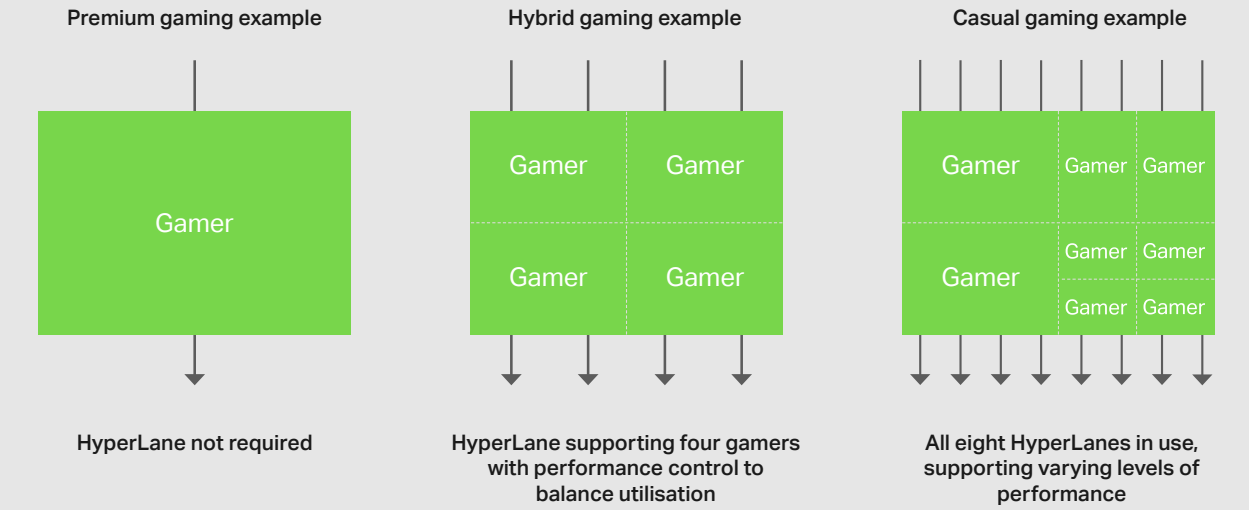
Dynamic cloud gaming architecture

HyperLane Technology

HyperLane Technology inside all Imagination GPUs is fundamental to enabling the dynamic performance sharing discussed earlier in this paper. HyperLane provides multiple, individual hardware control lanes, each isolated in memory, enabling different tasks to be submitted to the GPU simultaneously, for fully secure GPU multitasking. HyperLane supports different levels of priority for these tasks.

It is this, in conjunction with built-in dynamic performance control, that enables gaming workloads to be efficiently balanced, so a single core can support up to eight users at a time. This also ensures power savings, by enabling multiple tasks to be run on a single GPU core while maximising utilisation for increased efficiencies.

HyperLane Technology in use



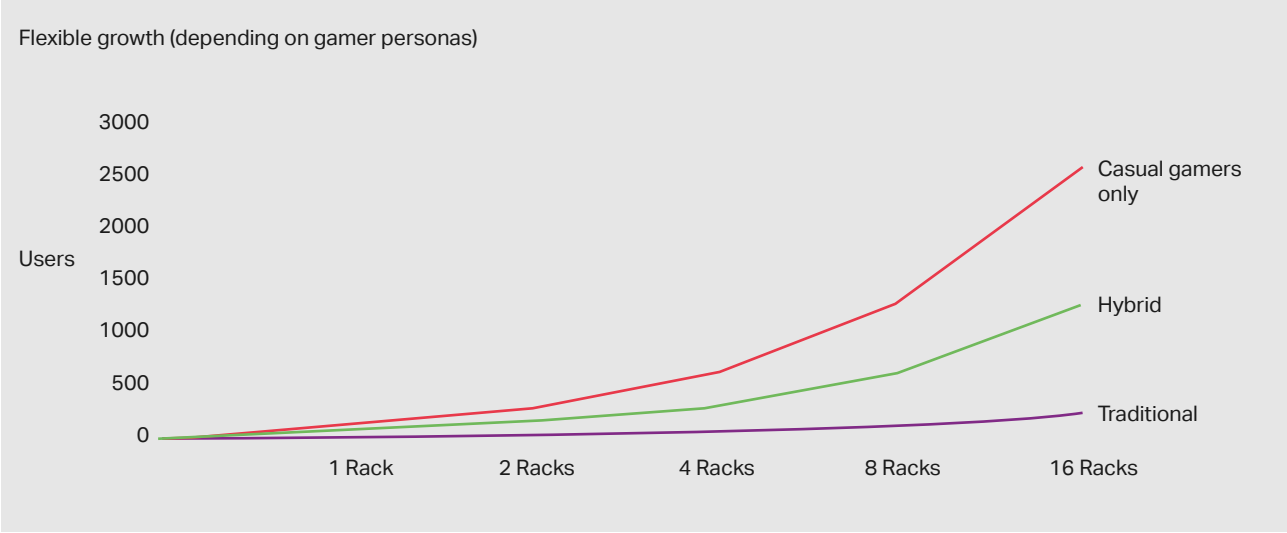
Every Imagination GPU core contain eight HyperLanes, enabling gaming workloads to be efficiently balanced to support multiple users.

Dynamic cloud gaming architecture

Graphics Density = Rack Density

Imagination's IP is designed to maximise the compute functionality available and be area-efficient at incredibly low-energy requirements. Its highly efficient performance per area and power efficiency has been proven in the mobile market and, when scaled up, our GPUs offer much greater graphics density than traditional desktop/data centre cards. Compared to desktop solutions, Imagination GPUs offer a "triple win": more compute performance, less power consumption and a smaller silicon area.

With the flexibility to meet scalable bandwidth requirements, by selecting Imagination GPUs cloud gaming providers can offer more gamers per chiplet and more cards per rack. This means more monetisable gamers, improving the business return for cloud-gaming services.



Creating a multi-user optimised architecture

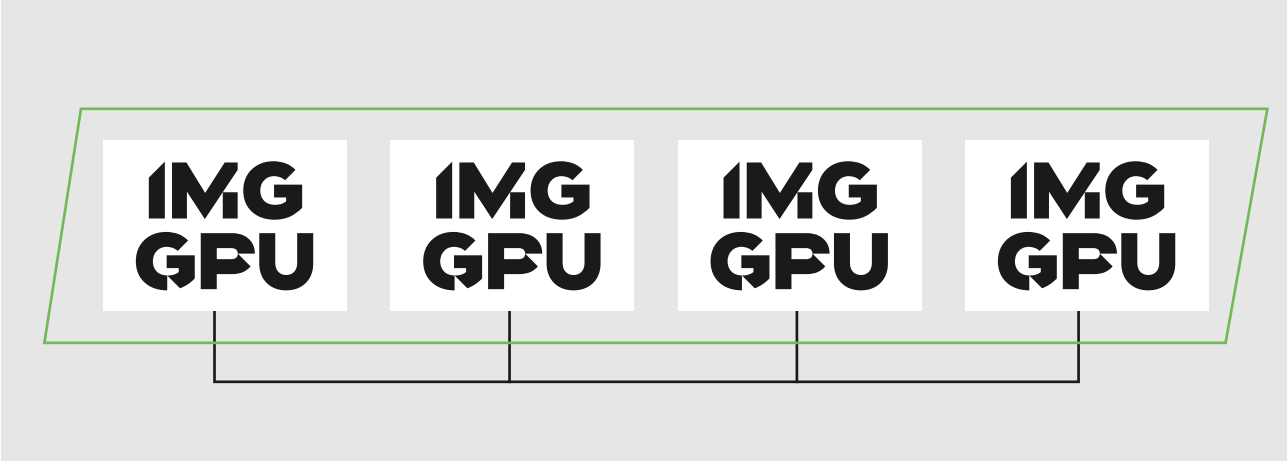
Imagination's approach to creating a multi-user optimised architecture for data centres is based on our multi-core GPU designs. However, the typical complex star-shaped design relies on a central block that includes the Job Manager, Tiling and Central Cache, which if you wish to move to multi-core, offers up several issues. It will require the central layout block to change and will lead to the internal Network on Chip and Tiler/Job Manager having to be redesigned.

This traditional star-shaped design is also incompatible with chiplet architectures, which are now commonplace, and places dependency on complex, latency-sensitive signalling. It also requires all decisions to be made at silicon time, rather than dynamically via software during operation runtime.

Dynamic cloud gaming architecture

Imagination's approach to multi-core addresses these issues by instantiating a flexible number of GPU cores without a direct dependency on a connection to a central unit. This more loosely-coupled design offers increased scalability, and as it can be used within chiplet designs it offers several additional benefits. These include cores that are identical and sit on standard NoC, enhanced physical design, reduced time-to-market, fewer complex signals and allowing for many usage modes, such as virtualization, security and ASYNC operation.

In its simplest form, our approach to multi-core effectively provides multiple GPUs in an SoC, but with the ability for cores to work on compute and graphics processing either separately, or jointly. This adaptive multi-core approach is a close match to the dynamic needs required by cloud gaming, as discussed above.



An example of how the GPU cores appear to the system in a four-core configuration.

As each core is designed as a fully 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 running 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) within each GPU core, our bandwidth efficiency is maintained. Each core continues to work on a coherent region of the screen, ensuring maximal cache-hit efficiency. A similar approach applies to other processing types, including geometry and compute processing, which can each be assigned to different GPU cores.

As Imagination's GPUs are based on tile-based deferred rendering it is easy to understand how multiple GPUs can work together. Each core works on a different set of tiles, which make up the total render target. In effect, the cores are, relatively, "loosely coupled", only sharing command and tile buffer lists in memory and sharing the workload between them.

Dynamic cloud gaming architecture

Chiplets and 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, therefore avoiding unnecessary data movement and duplication between cores, increasing efficiency.

By using an industry-standard bus fabric between the cores (e.g., AXI) we 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 and opportunities to be addressed simply through packaging.

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 enables 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. Imagination offers server-grade cores with ECC RAM and fault-detection mechanisms where GPUs can check each other's operation, for high-criticality data centre use.

Maximum performance, maximum flexibility

Servers require flexible performance scaling, and this is particularly relevant to cloud gaming where Imagination's flexible multi-core approach combined with HyperLane technology can be used to dynamically reconfigure GPU matrixes.

IMG DXD delivers 2.25x the per-core performance of IMG BXT. Its dual-core configuration has 144 GTexel/s pixel processing power and nearly 5 TFLOPS compute. Generational architectural improvements and extra ALU capacity mean that even when compared to its closest equivalent IMG BXT configuration, IMG DXD can offer 40-60% higher performance efficiency.

Imagination's investment in APIs

For gamers, API support is the difference between the ability to enjoy a popular title or missing out. It is a key factor in the purchasing decision of a new graphics card, second only to its performance statistics (or maybe price!). Through both hardware development and regular DDK releases, Imagination offers support for updated and new APIs, and enhances standard API functionality by integrating support for popular API extensions into our GPU drivers. This means that our customers are able to offer gamers detailed graphics with smooth frame rates on the titles of their choice.

DirectX

The DirectX API solved the hardware fragmentation problem that software developers for Windows platforms faced. By adding in DirectX as a new layer on top of the hardware, Windows developers could programme for DirectX and trust that the software will run on any Windows-based hardware. It opened up the gaming market and made game development quicker.

In 2023 Imagination launched a new line of GPU IP with hardware-based DirectX support, starting with IMG DXD with DirectX FL11_0. Popular titles based on DirectX 11 include Genshin Impact, PUBG Battlegrounds and Dota 2.

IMG DXD represents the first step in Imagination's DirectX journey. We are continuing to invest in our desktop API coverage to ensure that our IP meets the expectations of all customers in the desktop graphics card market.

Vulkan

Vulkan is a cross-platform industry standard for developers looking to develop once and deploy to a variety of hardware, whether desktop, cloud, console or mobile. One of its key goals is to reduce the level of developer fragmentation between the mobile and desktop gaming markets. Its latest iteration, Vulkan 1.3, was released in 2022 and is fully supported by Imagination GPU IP. This version absorbed select extensions into the core standard such as dynamic rendering and improved synchronisation. The API remains under constant development and its roadmap is available via the Khronos website.

OpenCL

OpenCL™ is an open standard that simplifies the parallel processing of compute intensive workloads in multi-processor, heterogeneous systems. It is an API for more than just the GPU: by running these complex workloads in parallel across all available processors, whether that's the CPU, GPU, DSP or a tensor processor, OpenCL can greatly accelerate the speed and responsiveness of compute based applications. Imagination GPU IP supports the latest version, OpenCL 3.0.

Imagination's investment in APIs

OpenGL

According to the Khronos® Group, OpenGL® is the most widely adopted 2D and 3D graphics API. Since its launch in 1992 it has been used extensively by software developers for PCs and workstations to create high-performance, visually compelling graphics applications for markets such as CAD, content creation, entertainment, game development and virtual reality. Although development of the API stopped several years ago, there is a vast range of legacy content that still depends on OpenGL support.

Imagination GPU IP offers support for OpenGL 4.6. This has been achieved through Zink, a layered OpenGL implementation, part of the open-source Mesa project, that allows OpenGL 4.6 content to run on top of a native Vulkan driver. It is fully Khronos conformant.

OpenGL ES

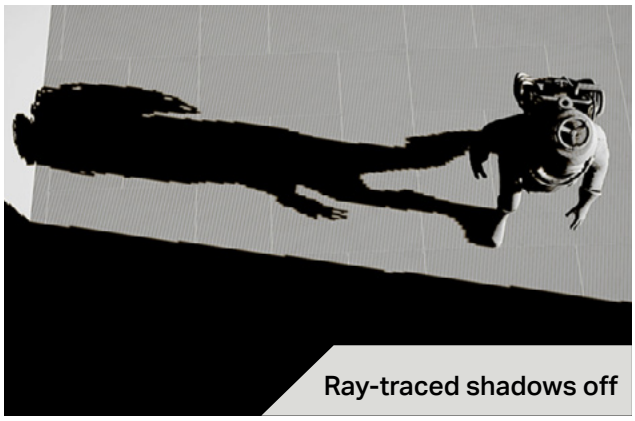
OpenGL ES is targeted at developers looking to design engaging 2D and 3D hardware accelerated graphics for devices with a restricted power budget, such as smartphones. It has been the bedrock of the mobile gaming industry for the past decade. Not only does it enable developers to run graphics workloads on the GPU, it also supports general purpose compute tasks.

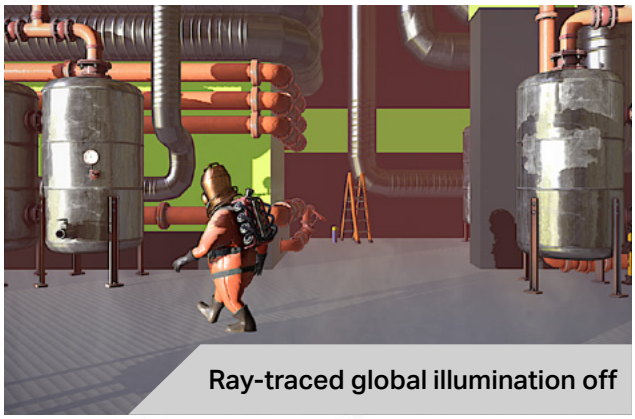
OpenGL ES is no longer under active development by the Khronos Group having been superseded by the more modern Vulkan. All Imagination GPUs support its last version, OpenGL ES 3.2.

Ray Tracing

Capturing the power of light

Today's gamers want the best, most immersive experience possible and ray tracing technology delivers "Hollywood-quality" graphics. Imagination's Photon Architecture provides a fully immersive experience for gamers in the most energy-efficient way. Imagination can enable a new tier of ultra-premium graphics that could be monetised by cloud gaming companies.



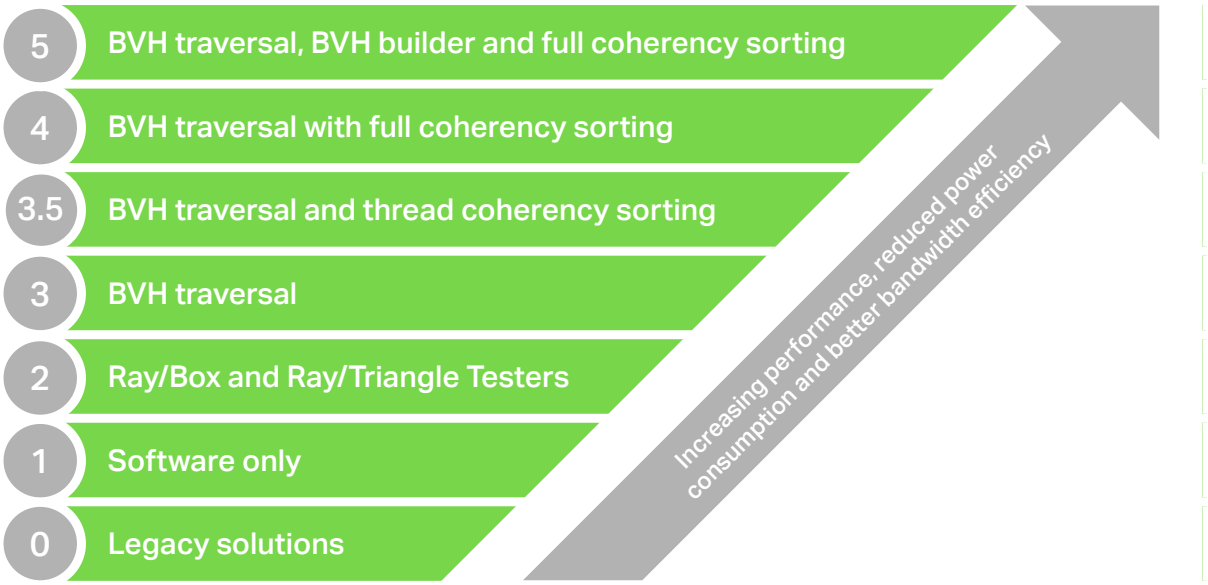


Ray Tracing

Imagination ray tracing – our unique approach

Ray tracing is computationally expensive, making it challenging to achieve in real time. To make it a reality, our Photon Architecture contains patented specialist hardware blocks that enable either faster full ray tracing or an efficient hybrid rendering approach that combines traditional rasterisation techniques with ray tracing for realistic lighting, reflections, and shadows.

Through this, Imagination has created a solution that offers a balance between image quality and performance that can meet gaming needs. While other companies are beginning to offer ray tracing solutions, many do not realise that not all solutions are created equal. To help the industry understand this we have set out the Ray Tracing Levels System. It defines how ray tracing solutions have evolved in the recent past and explain how efficiency increases with each level, which is especially critical for power-sensitive use cases, such as mobile.



IMG CXT and IMG DXT GPUs contain this unique technology, enabling Imagination to sit at Level 4 of the Ray Tracing Levels System.

Cloud gaming: a huge opportunity

Cloud gaming is a market that is growing exponentially. Casual gamers on mobile are beginning to see the advantages of immersive graphics and gameplay from screen to screen that can be enabled by 5G and the cloud. Importantly, this enables companies to move gamers from just being "mobile players" to being "mobile payers", and ones that can benefit from the high-performance graphics and outstanding visuals from ray tracing.

However, the challenges noted above require a completely new approach if companies are to achieve maximum success. Cloud gaming companies require extensive flexibility and scalability, and new business models for achieving outstanding results in a competitive world.

Imagination has the technologies required to enable game companies and data centres to offer consumers the next generation of gaming and achieve significant business success. Bringing low-power, performance-dense GPUs from the mobile space into the data centre is a breakthrough moment in the world of cloud gaming. Those that act now, will be best placed to gain a competitive edge.

Is Cloud Gaming the Future? | Top Cloud Gaming Trends 2022 (datamation.com)

www.imaginationtech.com