CONSUMER ELECTRONICS

Measuring GPU compute performance

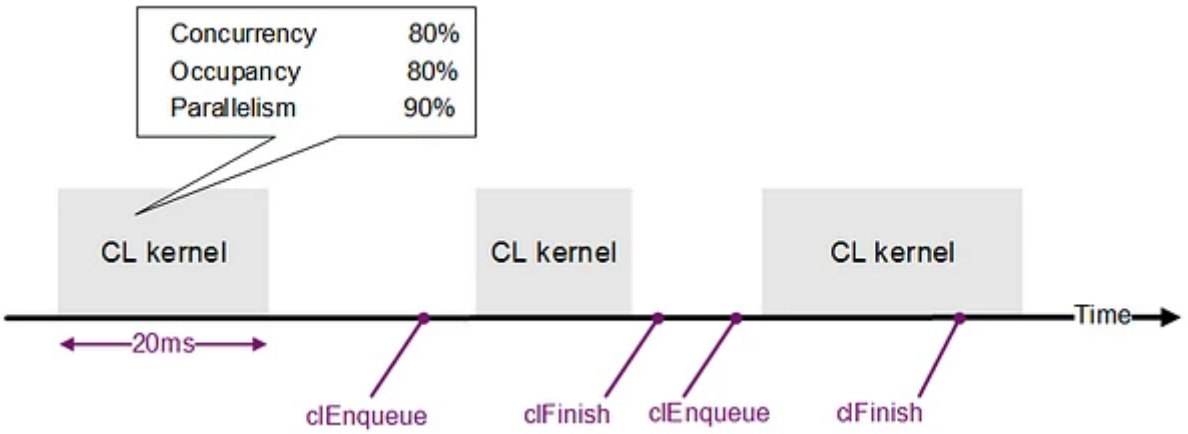
08 DECEMBER 2015 | SALVATORE | NO COMMENTS

After exploring a quick guide to writing OpenCL kernels for PowerVR Rogue GPUs and analyzing a heterogeneous compute case study focused on image convolution filtering, I am going to spend some time looking at how developers can measure the performance of their OpenCL kernels on PowerVR Rogue GPUs.

factors such as the choice of datatypes (relating to ALU capabilities) and compiler flags (for example, loop unrolling).

The performance of vector code running on a GPU is more difficult to quantify. As explained in this article, Rogue GPUs comprise a number of *concurrent*, *multi-threaded* processors. In this context, each work-item is executed by a single thread and has a scalar efficiency that can be defined similarly to code running on a scalar processor such as a CPU. However, in addition, there are also performance metrics related to *utilization* (how well memory latency is hidden as a result of the concurrent scheduling of multiple warps), *occupancy* (how easy it is for the multiprocessor to hide latency) and *parallelism* (to what extent threads in a warp execute in lock-step without diverging).

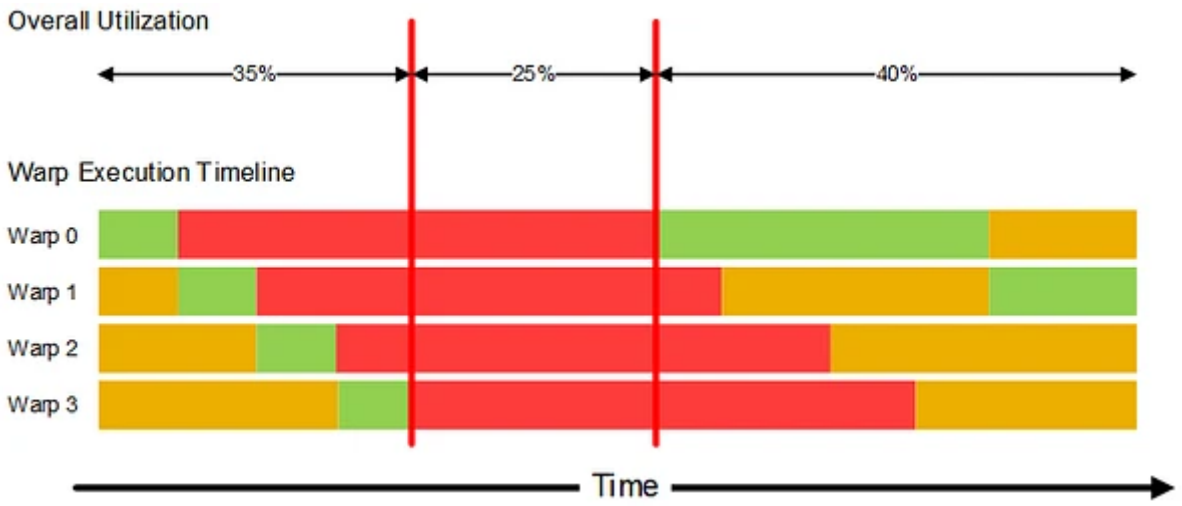
The figure below shows an example of three kernels executing on the GPU over time. Each kernel has an absolute execution time and, within a larger system, there may be delays between multiple executions of a kernel, for example representing the time taken for a CPU to prepare the next batch of data for processing. In addition to these absolute times, each kernel has the three efficiency metrics as mentioned above, which are discussed in more detail in the following subsections.



Rogue GPUs contain hardware counters that can be used to measure these performance metrics. These hardware counters are read by Imagination's OpenCL development tools, allowing you to 'see inside' a kernel's execution and gain a better understanding of any performance bottlenecks that can be eliminated. Once you have created the first implementation of your application, you should profile its performance to understand its performance and determine whether to invest more time in improving its performance. These tools include PVRTune, an OpenCL Occupancy Calculator and PVRShaderEditor.

Utilization

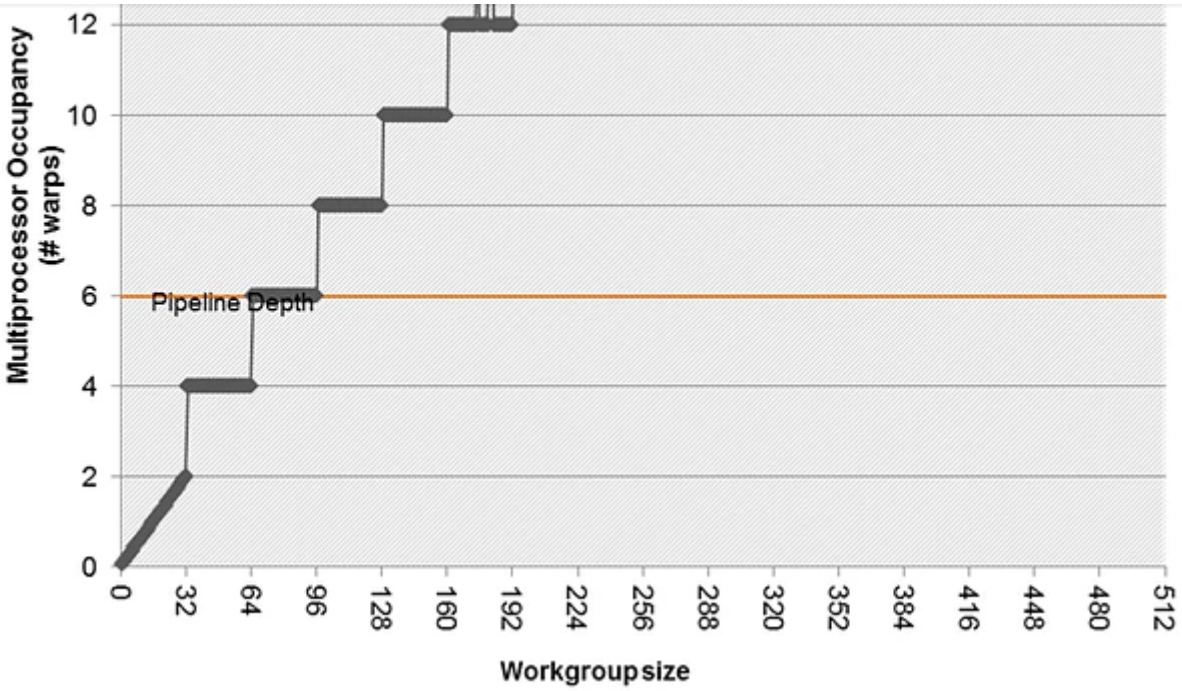
are blocked on memory operations for 25% of the execution time, the kernel's utilization is 75%.



Occupancy

Multiprocessor occupancy is the ratio of resident warps to the total number of available residency slots. As discussed in full inside our OpenCL programming guidelines, the total number of available residency slots may be limited due to a kernel's private and local memory requirements. Of these available slots, the total number actually used may be further limited by the speed at which the GPU can issue warps to the multiprocessors. The former metric can usually be calculated at compile-time, with the latter being determined at run-time.

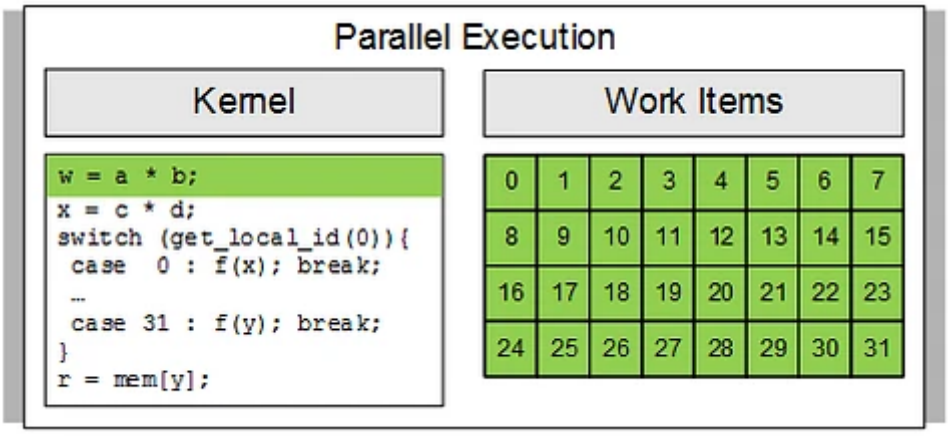
The figure below shows an occupancy graph for a sample kernel, produced by Imagination's OpenCL occupancy calculator tool. The purple triangle represents a specified workgroup size of 256, which the graph shows has a best-case occupancy of 16 warps (100%). The graph also shows the impact of varying the workgroup size, for example, reducing the workgroup size to 128 reduces occupancy to 8 warps (50%). This could be related to a workgroup's memory requirements. For example, if the workgroup of size 256 allocates 2048 words from the common-store memory, which has a total capacity of 4096 words, then two workgroups can be held on a multiprocessor occupying 16 slots. If the workgroup size is reduced to 128, and assuming the same memory requirements, then two workgroup will allocate all of the available local memory thus occupying only 8 slots, and preventing the multiprocessor from accepting further warps for the other 8 slots.

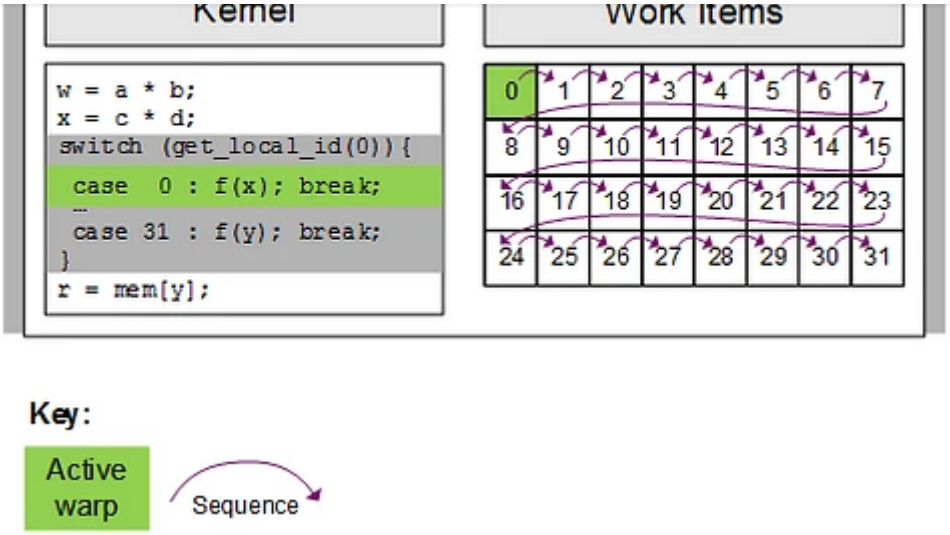


Note that occupancy is not a direct measure of performance: a kernel that achieves 100% utilization with 50% occupancy is as efficient as a kernel that achieves 100% utilization with 100% occupancy. In the former situation it might even be desirable to double the amount of private memory available to each work-item, to further improve the scalar performance of the work-items.

Parallelism

Parallelism is the ratio of lock-step to serialized operations performed by the work-items; work-items usually execute in lock-step parallelism. If work-items in a warp diverge via a conditional branch, the hardware serializes execution of each divergent branch, disabling work-items not on that path, and when all paths complete the work-items converge back to the same execution path. These two types of execution are shown below:





In the first case, all threads execute the statement in lock-step (100% parallel efficiency) but in the second statement all threads take turns executing the statement in sequence (0% parallel efficiency).

With the above metrics in mind, our OpenCL programming guidelines will give you more detail on how the Rogue GPU executes OpenCL programs, enabling you to apply even more advanced tuning techniques to improve performance.

Further reading

Here is a menu to help you navigate through every article published in this heterogeneous compute series:

- A primer on mobile systems used for heterogeneous computing
- A quick guide to writing OpenCL kernels for PowerVR Rogue GPUs
- Increasing performance and power efficiency in heterogeneous software
- The PowerVR Imaging Framework for Android
- Heterogeneous compute case study: image convolution filtering
- Deep dive: Implementing computer vision with PowerVR
  - Part 1: Computer vision algorithms
  - Part 2: Hardware IP for computer vision
  - Part 3: OpenCL face detection on PowerVR
- The PowerVR Imaging Framework camera demo
- Supported zero-copy flows inside the PowerVR Imaging Framework
- Measuring GPU compute performance
- Imagination's smart, efficient approach to mobile compute
- The complete glossary to heterogeneous compute on PowerVR

(@ImaginationTech, @GPUCompute and @PowerVRInsider) for more news and announcements from Imagination.

GPU Compute | Heterogeneous Compute | performance analysis | PowerVR | Rogue

Salvatore

Salvatore De Dominicis is a Leading Technical Specialist on OpenCL and GPU compute. Salvatore joined Imagination in 2013 and has more than eight years of combined experience in the industry, including various roles at STMicroelectronics. He brings to Imagination extensive experience in supporting customers and helping them create cutting-edge products in the market on a variety of operating systems (Linux, Android) and computing architectures (MIPS, Intel, ARM). Salvatore holds a MSC. in Computer Engineering from Politecnico di Milano.

Please leave a comment below

Comment policy: We love comments and appreciate the time that readers spend to share ideas and give feedback. However, all comments are manually moderated and those deemed to be spam or solely promotional will be deleted. We respect your privacy and will not publish your personal details.

Name *

Email *

Save my name and email in this browser for the next time I comment

Blog Contact

If you have any enquiries regarding any of our blog posts, please contact:

United Kingdom

benny.har-even@imgtec.com

Tel: +44 (0)1923 260 511

Search by Tag

Search for posts by tag.

Select a Tag

Search by Author

Search for posts by one of our authors.

Select an Author

Featured Posts

How AI is conducting the future of music technology
05 FEB 2019 PAUL WEIR

Separating the wheat from the chaff in embedded AI with PowerVR Series3NX
24 JAN 2019 BENNY HAR-EVEN

The ultimate embedded GPUs for the latest applications
06 DEC 2018 BENNY HAR-EVEN

Imagination Technologies: the ray tracing pioneers
10 OCT 2018 BENNY HAR-EVEN

Amazon Lights up its Fire TV Stick 4K with PowerVR
05 OCT 2018 BENNY HAR-EVEN

Neural networks – a guide for my mom
25 SEP 2017 JEN BERNIER

Face detection and identification using OpenCL on PowerVR GPUs
12 SEP 2017 ASHLEY SMITH

Popular Posts

Unreal Engine and the ray tracing revelation

PowerVR GR6500: ray tracing is the future… and the future is now

Related blog articles

The Android Invasion: Imagination GPU IP buddies up with Google-powered devices

December 14, 2020

Google Android continues to have the lion share of the mobile market, powering around 75% of all smartphones and tablets, making it the most used operating system in the world. Imagination's PowerVR architecture-based IP and the Android OS are bedfellows, with a host of devices based on Android coming to market all the time. Here we list a few that have appeared in Q4 2020.

Read More »

Back in the high-performance game

October 13, 2020

My first encounter with the PowerVR GPU was helping the then VideoLogic launch boards for Matrox in Europe. Not long after I joined the company, working on the rebrand to Imagination Technologies and promoting both our own VideoLogic-branded boards and those of our partners using ST's Kyro processors. There were tens of board partners but only for one brief moment did we have two partners in the desktop space: NEC and ST.

Read More »

Connect

Sign up to receive the latest news and product updates from Imagination straight to your inbox.

First name *

Last name

Email *

Enter your email address

Region *

Please Select

Imagination Technologies is committed to protecting and respecting your privacy, and we'll only use your personal information to administer your account and to provide the products and services you requested from us. From time to time, we would like to contact you about our products and services, as well as other content that may be of interest to you. If you consent to us contacting you for this purpose, please tick below to say how you would like us to contact you:

I agree to receive other communications from Imagination Technologies. *

In order to provide you the content requested, we need to store and process your personal data. If you consent to us storing your personal data for this purpose, please tick the checkbox below.

I agree to allow Imagination Technologies to store and process my personal data. *

You can unsubscribe from these communications at any time. For more information on how to unsubscribe, our privacy practices, and how we are committed to protecting and respecting your privacy, please review our Privacy Policy.

Sign Up