GP104 inside GTX 1070/1080

**SECTION I:**

GTX cards have been around for quite a long time starting with the 8800GTX, the first gaming card that has the GTX designation. The newest edition is packed with the GP104 graphic processor, a GPU designed and produced by Nvidia.

The GeForce GP104 is a performance-segment GPU by NVIDIA, released in June 2016. It is built on the 16nm process, with its GP104-200-A1 variant, the GPU comes and supports the latest DirectX121, which ensures every modern game has the compatibility to run on it. The GP104 us a large chip that contains 7,200 transistors. Packed with 8GB GDDR5 memory, GP104 could run any demanding graphics programs which its previous predecessor could not. The base model of GP104 operates at a frequency of 1506 MHz, which can be boosted to the range of 1800-1900 depending on where the retail boards from (ASUS ROG STRIX GTX 1070 O8G has a base clock of 1633Hz and can be boosted to 1835 Hz thanks to its already overclocked board from ASUS).

The GP104’s power is supplied by 8-pin power connector (or sometimes 2x4 pins) and usually draws 150W maximum. Its display output includes a HDMI port, a DVI port, and 3 Display ports and it’s connected to the mainboard using a PCI-Express 3.0 x16 interface.

NVIDIA is also famous for their CUDA technology, which is a parallel computing platform and programming model developed on its own for general graphical processing units. Therefore, the GP104 comes with 1920 CUDA Cores and could potentially outperform TITAN X Pascal’s processor (one of the most powerful GPUs in years) in both performance gamming and graphics designing.

The most interesting about this GPU is that it seems Nvidia released this one in as a budget friendly graphic card. Its starting price is at $379 or around the range of $400-$500 which performance can widely vary as different manufacture has different specs for their boards. The GP104 is likely a smaller sibling of GP104-200-A1 at that time. The most significant differences between these two models are their memory. While 1080 is packed with GDDR5X, its younger sibling is designed with GDDR5, with lower memory speed but respectively speaking, still a rarely case for any graphic cards of that time designed with this type of memory. Due to that nature, GP104 would have a lower power draw than its big sibling which is around 20W. Furthermore, as for the GP104, its design does not use a system of vapor chamber cooler but instead an integrated heat-pipe cooler similar to what the older cards used.

Figure Performance vs Power source NVIDIA.com

The second most interesting about graphics card of GP104 lineup is their power consumption paring with their actual performance. The power draws of these two cards is a really big step up from older models of the Maxwell family. Performance-wise, the new generation of Pascal GTX packs 25% more cores and 25% more texture units than the GTX 970/980’s processor with relative much higher base clock and significant faster RAM while still maintain a considerable power consumption.

**SECTION II:**

The GP104 line up runs on Pascal, a GPU microarchitecture developed by Nvidia, as the successor to the Maxwell family (GTX 970/980). Pascal was manufactured using TSMC’s 16nm FinFET (fin field-effect transistor) process and later Samsung’s 14nm FinFET process. In this microarchitecture, there is a streaming multiprocessor that consists the total of 128 CUDA cores (parallel computing platform). The card uses a high bandwidth memory 2 which yields the total of 4096-bit bus with a memory bandwidth of 720 GB per second. It also built on a platform of integrated architecture where the CPU and GPU can access both main system memory and memory on the graphics card with the help of a technology called “Page Migration Engine”.

In the GP104, there are twenty of Pascal Streaming multiprocessors, eight memory controllers, which also contains 256 KB of register file capacity ( almost double the previous line) , a 96 KB shared memory unit (between base system and GPU), 48 KB of total L1 cache storage, and eight texture units. GTX 1070/1080 optimizes on its highly parallel multiprocessor that can schedule warps (groups of 32 threads) to CUDA cores and other execution units within. It is also the most important processor hardware unit inside of the GPU, almost all flow of operations has to come through it at some point in the rendering pipeline.

**GDDR5X Memory and Enhanced Memory Compression**

GP104 has its speed from utilizing GDDR5 memory which first introduced in 2009. Since then, this type of memory is often used on RAM as it is the culmination of the fastest and most advanced interface standard in history. It is also capable of achieving 10 Gbps transfer rate or nearly 100 picoseconds between data bits. Like the previous GPUS, the memory subsystem of the GeForce GTX uses lossless memory compression techniques to reduce DRAM bandwidth demands. The bandwidth reduction provided by memory compression reduces the amount of data written out to memory, data transferred from memory to L2 cache and between clients such as the Texture Unit and the frame buffer.

**Asynchronous Compute**

Modern gaming software requires an insane amount of complex workload with multiple independent workloads that can ultimately work together to render the final image to display such as the processing of rendered frames (v-sync, g-sync), GPU based physics and audio processing or maybe the newest technology VR (virtual reality). These workloads create two scenarios for the GPU architecture to consider.

The first one that involves overlapping the workloads in which certain type of workloads do not fill the GPU by themselves. Therefore, in this case, it is easier to optimize the system by running two workloads at the same time, sharing the GPU and to run it more efficiently.

Time critical workloads are the second compute scenario. A time-wrap operation must be completed before scan outs starts or a frame will be drops. In this case, the GPU needs to support fast and low latency to move not critical workload off the GPU so that the more critical and important one can run as soon as possible.

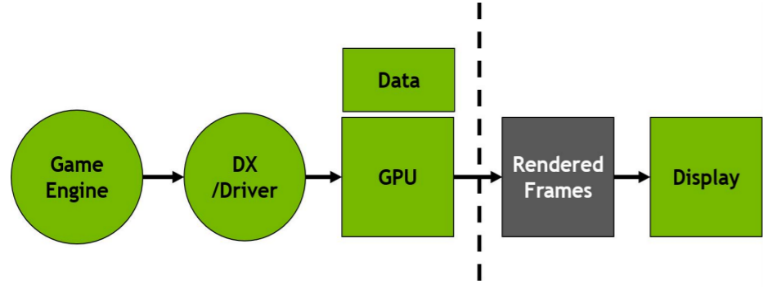
To fix these 2 cases of scenarios, Pascal was made to implement Pixel Level Preemption. The unit that takes care of the graphics have been improved fast enough to keep track of the progress on rendering the workloads, so that as soon as the preemption is requested, they stop and save off context information about where to start up again later.

**Enhanced SLI Interface**

In any cases, one GPU is always enough for any types of workloads. However, some just prefers to push the limits of their setups and SLI was introduced to deliver the very best experience at the highest resolutions and framerate settings. The main ingredient to any NVIDIA’s SLI technology is the SLI Bridge, a small piece of hardware that could enable the transferring display data between multiple graphics card.

The fantastic thing about the technology is that each GPU has to do their own part, that means rendering their own frame and then together send them to the display connected through the mainboard. For instance, any two interfaces of GPU could be linked together to improve the bandwidth between GPUS. With the releasing of model 1070/1080, a new SLI HB Bridge is also needed. Capable of running at 650MHz, this SLI subsystem provides more than double the bandwidth between each GPU running individually. This is particularly important to achieve high resolutions like 4K and 5K.

**Fast Sync**

Fast Sync is a new latency-conscious alternative to Vertical Sync (capping framerate at the refresh rate of the monitor) that can prevent image from tearing, allow GPU to render unrestrained to reduce input latency integrated in GP104.

The game engine generates the frames which then are sent to DirectX, the encoding inside the frame get rendered and pushed to be converted inside the GPU then it spits out the rendered frames to GPU to buffer and output displays to monitor.

With Fast Sync, there is no flow control. The game engine works as

V-sync is off. And because there is no backpressure, input latency is almost as low as with V-Sync off. Best thing is there is no visual tearing because Fast Sync automatically chooses which of the already rendered frames to scan to display. Therefore, it allows the front of the pipeline run as fast as it can, simultaneously preserve the frames so the output display produces no tearing.

**SECTION III:**

The performance for price of the line of GP104 processor integrated graphics card is fantastic at MSPR $375. At 1440p, it dominates the market with the DisplayPort 1.4 it is also a future proofing for any graphics for 4K monitors at 120Hz. It is a very well-balanced card and worth investing if you can find it at the right place and price.

Back when it was first released, it was also considered overkill generally for around 90-95% of the games on the market. Furthermore, if you have yourself a SLI setups, this could be also ideal for saving up for a new generation of GPU in 6-7 years as they are packed with enough performance to outdo even the GTX Titan’s processor. However, with the release of new generation of RTX cards’ processor, these GP104s at the same price seems less and less appealing to users at the same price tag.

**SECTION IV:**

**Glossary**

DirectX: a collection of application programming interfaces for handling tasks related to multimedia, especially game programming and video, on Microsoft platforms.

GDDR5 memory:  an abbreviation for graphics double data rate type five synchronous dynamic random-access memory, is a modern type of synchronous graphics random-access memory with a high bandwidth interface designed for use in graphics cards, game consoles, and high-performance computing.

FinFET: A fin field-effect transistor is a multigate device, a MOSFET built on a substrate where the gate is placed on two, three, or four sides of the channel or wrapped around the channel, forming a double gate structure.

Cache: a hardware or software component that stores data so that future requests for that data can be served faster; the data stored in a cache might be the result of an earlier computation or a copy of data stored elsewhere.

Bandwidth: the maximum rate of data transfer across a given path.

V-sync: a graphics technology that synchronizes the frame rate of a game and the refresh rate of a gaming monitor

G-sync: a proprietary adaptive sync technology developed by Nvidia aimed primarily at eliminating screen tearing and the need for software alternatives such as Vsync.

DisplayPort: a digital display interface developed by a consortium of PC and chip manufacturers and standardized by the Video Electronics Standards Association.

HDMI: a proprietary audio/video interface for transmitting uncompressed video data and compressed or uncompressed digital audio data from an HDMI-compliant source device, such as a display controller, to a compatible computer monitor, video projector, digital television, or digital audio device.

Reference:

NVIDIA Whitepapers GTX 1070/1080 <http://international.download.nvidia.com/geforce-com/international/pdfs/GeForce_GTX_1080_Whitepaper_FINAL.pdf>

Storey, Zak. “MSI GeForce GTX 1070 Gaming X Review.” TechRadar, 25 Apr. 2019, www.techradar.com/reviews/pc-mac/pc-components/graphics-cards/msi-geforce-gtx-1070-gaming-x-1323106/review.

“NVIDIA GeForce GTX 1070 Review.” *NVIDIA GeForce GTX 1070 Review*, 26 Feb. 2020, benchmarks.ul.com/hardware/gpu/NVIDIA GeForce GTX 1070 review.

Hagedoorn, Hilbert. “Nvidia GeForce GTX 1070 Review.” *Guru3D.Com*, www.guru3d.com/articles-pages/nvidia-geforce-gtx-1070-review,1.html.