# How many kIOPS on a NVMe PCIe Gen2 x4 device?

*(Jérôme Gaysse, IP-Maker, published on Chipestimate.com on Sept 17, 2013)*

The number of IO per second (IOPS) is certainly the most important parameter that characterizes the performance of a storage device. The NVM Express (NVMe) specification has been introduced in order to leverage the performances of PCI Express Solid-State Drives (PCIe SSDs). The first part of this Tech Talk explains the maximum theoretical performance reachable on such devices. In the second part, a NVMe hardware implementation is described and shows how a multi-channel DMA architecture is able to provide performances very close to the theory.

**Theoretical number of IOPS on a PCIe Gen2 x4 configuration.**

The IOPS is one of the key performance parameter for a storage system. The other parameters are the latency (in µs or ms), and the throughput (MB/s). The size of the IO is generally 4kB, and the results are provided in kIOPS. The typical measures listed in storage documentation are done for read and write, random and sequential accesses. In benchmarks studies, additional measures are done like read/write (70/30).

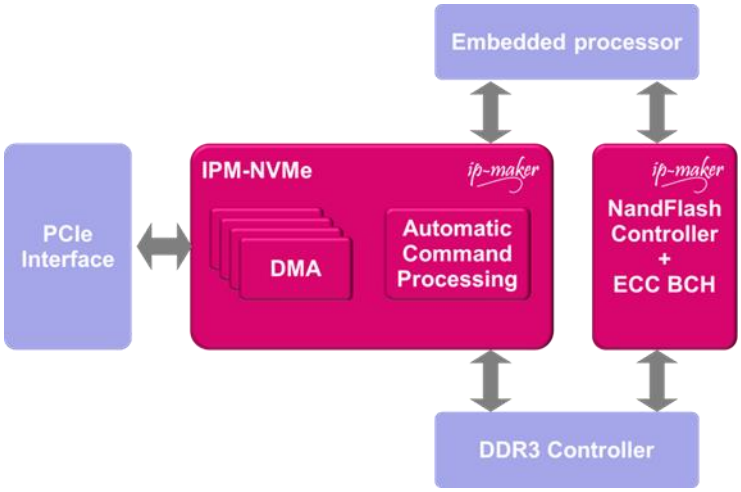
A Gen2 lane speed is 5Gb/s. With 4 lanes, the overall speed is 20Gb/s. It would be too easy to do the following calculation: 20Gb/s divided by 4096B, equal to 610k IOPS! The real value must take into account at least 3 parameters: the 8/10 bit coding, the PCIe overhead, and the fact that a bus can't be used 100% of the time. The first factor, the 8/10bit coding, basically results in 80% of real information transferred on the PCIe bus. The second factor is the PCIe overhead. In addition to a data packet, multiple added bytes are included in the data transfer, such as sequence number, header and CRC. That leads in a 20 or 24 bytes overhead. Let's do the calculation with the typical payload (data packet size) of 256B. 280B are used to transfer 256B, resulting in a 256/280=91% efficiency. Finally, the PCIe bus occupancy rate is estimated as 90%.

Therefore, the maximum IOPS is 610k x 80% (8/10b coding) x 91% (PCIe overhead) x 90% (PCIe bus occupancy rate) = 400kIOPS. This is the maximum that we can reach, assuming data transfer without any protocol or any command management. Unfortunately, it is impossible to send/receive data only on the bus of a PCIe SSD. A protocol is required in order to provide mandatory information in addition to the data (address, packet size, priority, system information…). The NVMe specification has been released in March 2011, and defines an optimized register interface, command set and feature for PCIe SSDs. The necessary commands for the data transfer will add traffic on the PCIe bus, resulting in a performance loss compared the theoretical maximum 400kIOPS. This loss is estimated as 5%, leading to 400 kIOPS x 95% = 380kIOPS. This is the maximum IOPS performance to be observed on an optimized storage system using a NVMe PCIe Gen 2 x4 interface.

**NVMe multi-channel DMA architecture**

IP-Maker has developed an IP which is NVMe compliant. It is mainly based on two full hardware blocks:

1) **Automatic command processing**, able to fetch the NVMe commands in few clock cycles only with a very low latency (see *Leveraging PCIe SSD performance with a full hardware NVMe* Tech Talk published on May 21, 2013).
2) **Multi-channel DMA**, allowing direct access between the host memory and the storage memory.

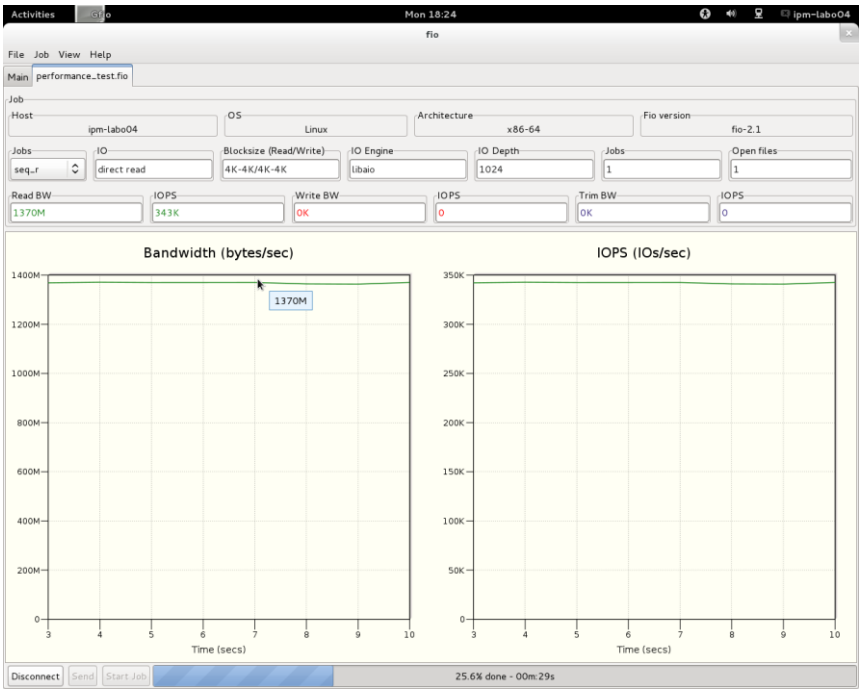



*IP-Maker NVMe IP integration*

The multi-channel DMA is configurable, up to 32 channels, for read and write. The DMA engines are activated as soon as a memory transfer command is fetched by the automatic command processing unit. Using multiple channels allow to perform data transfer all the time, avoiding transfer stop when the data are read or written in the memory. NVMe is based on a queue mechanism. Multiple commands can be used in one queue. For latency optimization, the best is to use one command per queue. But for data transfer, it is recommended to increase the number of commands in a single queue (e.g. 1024 commands per queue).

**PCIe Gen2 x4 NVMe reference design**

The NVMe IP has been integrated in a FPGA-based reference design. It is based on Virtex-6 FPGA from Xilinx with the ML605 evaluation kit. The NVMe IP is connected to the PCIe hard IP and a soft DDR3 controller IP. It is configured as Gen2 x4. The storage part of this NVMe reference design is based on DDR3 memory in order to demonstrate the NVMe IP performances.
On the host side, a PC platform is running Linux Fedora 17 with the NVMe driver (available at www.nvmexpress.org). When installed in the PCIe slot, the NVMe reference design is detected as a NVMe storage device. Since it is DDR-based, the file system needs to be mounted at each power up. The performances are measured with the standard FIO tool (GFIO version with graphical user interface).




*GFIO user interface*

The results of GFIO show a 343 kIOPS in sequential and random read, with 4k IO and 1024 commands per queue. This is very close to the maximum.

On an optimized system platform using a NVMe PCIe Gen2 x4 configuration, the multi-channel DMA used in the IP-Maker NVMe IP core is able to reach the maximum number of 380 KIOPS. The benefit of NVMe is clearly demonstrated by providing high performance IOPS while reducing the host CPU load since all the data transfer is managed by the NVMe device itself. Therefore the NVMe IP does not add any bottleneck on the data path. On a real PCIe SSD, the performance IOPS bottleneck may be the NandFlash memories, but with next generation NVM memories which are faster than NandFlash, such as MRAM, it is very important to avoid any performance loss in the NVMe protocol management. Since it is a scalable architecture, the million IOPS range is reachable on higher PCIe configuration, such as Gen3 x8.

**Contact information**

www.ip-maker.com
contact@ip-maker.com
+33 972 366 513
Domaine du Petit Arbois
Avenue Philibert BP50014
13545 AIX EN PROVENCE Cedex 4
France