Designing an enterprise grade NVMe SSD

What is the definition of an enterprise grade SSD: high reliability or high performance? Both. Reliability is one of the most important criteria for an enterprise storage system. Losing data is not allowed. The main potential issues are the memory corruption, a controller failure and unstable performances which may lead in an availability issue. The expected reliability from the IT manager is defined with a 99.999% availability. The other important criteria for an enterprise SSD is the performance (latency and IOPS). For big data analytics, data base or virtualization, the return on investment of the IT capex depends on the performances.

This article describes how a full hardware NVMe architecture can be used to design a NVMe PCIe SSD with an enterprise grade, providing 10µs latency range and five nines quality of service.

Reliability

Below is a short list of features to integrate in a NVMe SSD in order to reach the five nines.

Quality of service: this is often related to the latency. SSD manufacturers communicate often with the lowest latency number (e.g 20µs) but IT manager will look at the 99.999 percentile latency (e.g. few hundred of µs). Therefore, the internal SSD management, including NVMe protocol processing, must be designed in order to ensure a low latency in any cases.

Redudancy: all flash arrays are typically based on a dual controller architecture. There is an active and a standby controller. That allows the access to the SSDs even if a controller has failed. On the SSD side, that means that each SSD is accessible by the two controllers. A switch mechanism is required at the front of the SSD, or as a better solution, a dual port interface can be integrated inside of the SSD. That can be done by using 2 PCIe interfaces.

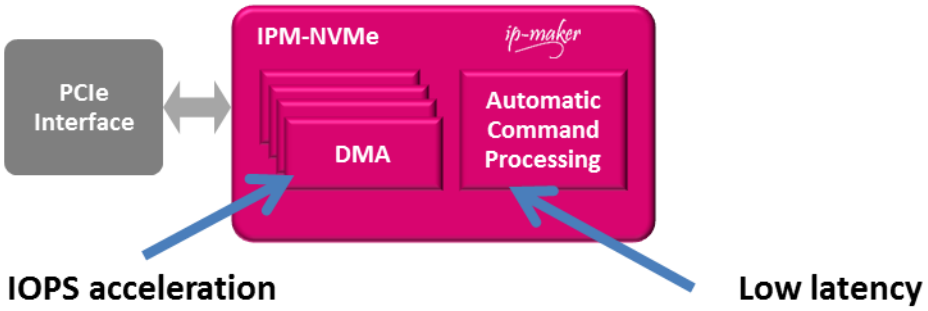
Data protection**:** the data coming from the host are going through many interfaces and memory buffer: PCIe interface, internal RAM, CPU, external DRAM, Flash memory… Then at each step, there is an opportunity to get corrupted data due to power spikes for example. In order to ensure end to end protection, some specific mechanism must be integrated. It is well defined by the T10 organization.

ECC: NandFlash memory is well known to have limited endurance (few hundreds to few thousands cycles). When reaching this limitation, bits may be corrupted in the memory. Then, an error correction code (ECC) is required in order to detect and correct the corrupted data. This is included in the SSD controller, in software or hardware. Common ECC are based on a BCH technology. LDPC are better for coming NandFlash memories.

Performances

IP-Maker has developed its own NVMe IP, from the ground-up, to be integrated in SSD controllers. Using pre-validated NVMe IP core, allows to greatly reduce time-to-tarket for storages OEM which want to benefit from a powerful NVMe compliant solution. The IP-Maker NVMe IP core is full featured, easy to use into both FPGA and ASIC designs.

Below is the architecture of the NVMe IP from IP-Maker. All the different part required by the NVMe specification have been designed through multiple hardware blocks, including configuration space, queue context management, queue arbitration and read/write engines.



Each of the hardware blocks takes only few clock cycles to be processed, therefore reducing dramatically the NVMe processing latency. So the impact of the NVMe processing on the system latency is very low compared to the other latency parameters.

The NVMe commands are processed by an automatic command processing unit. The data transfer rate is accelerated with the multi DMA channels integrated in the read and write engines. So, the PCIe bus is always used by NVMe accesses. The maximum throughput is defined by the PCIe configuration: number of lanes and speed generation.

This full hardware NVMe architecture is ideal for persistent memories, such as 3DXP, NVRAM, MRAM or RRAM. The IP-Maker NVMe IP latency offers a latency in the range of hundreds of nanoseconds, which is in range of read and write timings od the new memories. That would make no sense to use so fast storage memories with a NVMe controller providing a 10µs range latency.

IP-Maker solution

The way to manage and process the NVMe protocol may have an important impact on the quality of service. When based on a full software implementation, the processing time may change according to IRQ management for example. If based on a full hardware architecture, the processing time is deterministic and will provide an accurate system latency QoS.

Using a full hardware NVMe implementation is more easy to use with a dual PCIe interface. A NVMe IP can be used on the back end of each PCIe controller, or only one can be used, shared by the PCIe controllers. The second case seems to be easier to manage. The use of a tag system will help in identifying which PCIe controller is accessing the NVMe IP.

IP-Maker also developed a fully configurable BCH ECC, easy to integrate with the NVMe IP. This ECC block is able to manage 74 errors per 1kB block.

Conclusion

This NVMe hardware implementation provides the benefits of an ultra-low latency and reliability. In addition, because of its reduced gate count footprint, that consumes less power than software implementation and reduces the silicon cost. Therefore it is an ideal solution for high performance data storage systems. Industry leaders will benefit from the enterprise grade without adding cost. In addition, it is ready to support next generation of NVM, which comes with better performances in term latency, density and power consumption.

Contact information

Les Néréides
55 rue Pythagore
13290 Aix en Provence
France

www.ip-maker.com
contact@ip-maker.com
+33 972 366 513