Benefits of NVMe over SATA in Defense Systems Data Recording

By Dr. Robert Maskasky

To understand the benefits of NVMe over SATA we need to understand the history of the SATA specification and the PCIe bus standard from which the NVMe standard was derived. I will discuss the history of the SATA standard, PCIe bus standard and how NVMe built upon the strengths of those standards to create a fast bus standard specifically for Solid State hard Drives.

History

In the 1980's IBM wrote a specification for attaching Hard Disk Drives (HDD) to support the IBM PC, this was titled Advanced Technology for Attachment (ATA). This was a parallel data transfer specification that required an 80 conductor ribbon cable. The ATA specification was designed around a spinning magnetic storage media Hard Disk Drive that stored data in pie-shaped sectors, tracks (concentric circles) and platter number. Computers think in binary numbers, rows & columns (Linear-**B**inary Addressing). The ATA driver must have hardware and software to convert data addresses from linear-binary to pie-shaped sectors / cylinders.

As computers became smaller and faster, the bulky 80 conductor ATA ribbon cable became too bulky and too slow for the current generation of computer processors. In the year 2000, the computer manufacturers developed a new set of specifications to enhance the old ATA specification. These new specifications were called Serial Advanced Technology for Attachment (SATA) and Advanced Host Controller Interface specification (AHCI). Together these specifications define the connections and the command structure protocol to support the SATA bus. AHCI protocol was optimized for the magnetic spinning hard disk, where data is addressed by sectors, tracks and platters.

Usually serial data transmissions are associated with slower speeds, by using high quality cables and sending the data like a digital Radio Frequency transmission line they were able improve transfer speed to a theoretical 6 Gigabits per second). At that speed, data errors are frequent, so when the SATA controller detects data errors, it will automatically repeat the transmission and slow the communication speed to reduce the number of errors. In practice most SATA devices achieve a maximum speed of about 4.8Gigabits/second which is about 600MegaBytes/second (a Byte is 8 bits).

Peripheral Component Interconnect Express (PCIe)

In the 1980s Industry Standard Architecture (ISA) was the bus standard for attaching peripheral cards to the IBM-PC. In the 1990s computers had become much faster and outgrown the ISA bus. Intel introduced the Peripheral Component Interconnect bus (PCI) as a faster version of the ISA bus. By the turn of the century (late 1990's), computer processors became faster and were introducing 32-bit processors, the PCI bus was becoming a speed bottleneck. Computer manufacturers required faster data transfer from memory to the 32-bit CPUs. In the early 2000's the computer manufacturers developed the PCI-Express bus (PCIe). Their purpose was not to reinvent the PCI bus, they just needed a faster PCI bus to support the latest generation of faster 32-bit and 64-bit CPUs.

The PCIe specification introduced the concept of Data Lanes. A DATA LANE is a serial transmission wire pair that is similar to a lane on a highway. If you need a highway to handle more traffic you can either increase the traffic speed or increase the number of lanes of traffic. A single "Data Lane" is two pair of wires capable of simultaneously transferring data at a rate up to 2 GigaBytes per second in both directions (equivalent to a 2-lane highway). Peripheral cards that didn't require a lot of speed (a sound card for example) could be made into a very small size and only utilize a fraction of the bus (two data lanes referred to as PCIx2). On the same bus, a memory card or video card that requires a lot of high speed data could utilize 16 or 32 data lanes. Additionally, the PCI-Express specification allows for the computer manufacturers to expand the bus to support 64, 128 or 256 data lanes to support future generations of faster computers.

Solid State Drives (SSD)

By the year 2020 computer chip manufacturers had made significant improvements in solid state memory technology, and finally it was practical to utilize Solid State Drives (SSD) to replace the spinning magnetic Hard Disk Drive (HDD). Solid State Drives have significantly improved Access Time and are capable of much faster data transfer than the traditional spinning magnetic HDD. For marketing reasons, SSD devices were initially designed to replace a SATA HDD in existing computer designs; using the SATA cabling and AHCI interface, which significantly constrained the speed capabilities of the non-volatile memory used in the SSD. The non-volatile memory chips which comprise the SSD are closely related in design and architecture to computer RAM memory, so the PCIe bus is almost ideal for communication between the CPU processor and the SSD.

Non-Volatile Memory Express (NVMe)

Legacy computers were designed to utilize SATA and ACHI to communicate with the Hard Disk. For a SSD to communicate on a SATA bus requires a data format converter on either end of the SATA bus to convert the PCIe communications to ACHI, transmit the data over the SATA bus then convert back to PCIe on the other end.

In 2021, the computer manufacturers working with the memory manufacturers developed a new communication bus specifically non-volatile SSD memory, as a replacement for SATA. This new specification was called Non-Volatile Memory Express (or NVMe). The "Express" was included in the name to indicate it is an extension of the PCI-Express specification specifically customized for Non-Volatile Memory. Like PCIe, the NVMe specification allows for the bus to be expanded by adding more data lanes to support future generations of higher-speed memory chips. The current generation of non-volatile memory chips only requires 4 data lanes to support maximum speed data transfer of 8 GigaBytes/second[1] (13 times the speed of SATA[2]). This doesn't include the speed advantage since the NVMe doesn't have to convert PCIe to SATA and back. The latest SSD memory chips are just now maturing to 7 GigaBytes/second read/write speeds. At the current rate of growth, within a year SSD

[1] PCIe Gen 4 Specification allows 2 Gigabytes/second per data lane
[2] SATA Gen 3 can achieve 600 MegaBytes/second.

memory chips will exceed 8 GigaBytes/second read/write speeds. Fortunately, PCIe Generation 5 will double that speed to 15.75 GigaBytes/second[3] for a standard NVMe 4-lane M.2 size memory card.

In 2022, 99% the new computer designs utilize NVMe with solid-state computer memory. With the dramatic increase in demand, the price of NVMe devices has fallen below the price of their slower SATA cousins. As the price of NVMe devices continues to fall, SATA devices will be obsolete in the next 5 years.

Form Factor… NGFF (a.k.a. "M.2")

Non-volatile memory chips are only a fraction of the size of the equivalent spinning Hard Disk Drive (HDD). The computer manufacturers wanted to take advantage of the smaller size to build smaller computers, so they developed the Next Generation Form Factor (NGFF) specification, also known as "M.2" (Miniature circuit card version 2). The NGFF (or M.2) specification has the memory chips mounted on a circuit board that is generally 22 millimeters wide by 80 millimeters long (about the size of a stick of gum). The NGFF specification also allows for other widths and other lengths, but the most common sizes are 22mm wide by either 42mm, 80mm or 110 mm long.

NGFF (M.2) and NVMe are frequently confused. The NGFF (M.2) is only a circuit card size specification while the NVMe is a communication specification that takes advantage of the NGFF circuit card sizes.

Initially, computer manufacturers had not yet re-designed their motherboards to take advantage of NVMe. Aftermarket consumers wanted a path to retrofit the smaller NGFF (M.2) devices into legacy computers designed to use SATA HDDs. The NGFF specification allowed for either an NVMe electrical edge connector or a variant of a SATA connector (known as mSATA). This allowed for easy installation of NGFF cards into legacy SATA computers (only a connector adapter was required). Initially the mSATA memory devices had a surge of popularity which drove the prices down, but as the computer manufacturer designs caught-up, the demand for NVMe has increased, diminishing the demand for SATA (mSATA) devices.

This doesn't mean the older 2-1/2" spinning hard drive chassis form factor is obsolete (70mm wide x 100.5mm long x 7mm high). Large corporate servers need very high volume of storage capacity. Many of these servers have converted to SSD devices, but there can be cooling issues associated with multi-terabyte SSDs in the smaller NGFF form factor. By mounting NVMe SSDs in the legacy 2-1/2" chassis (a.k.a "U.2" form factor) the SSD designer could take advantage of NVMe data speeds while providing more surface area for cooling the SSD. In the foreseeable future the U.2 form factor will be popular for extremely high capacity SSD drives.

Real-world defense scenarios where faster data transfer rates can provide a tactical advantage.

Advances in technology significantly increases in the amount of data collected and the rate of that data acquisition. The data can be processed / reduced using on-board processing, it can be transmitted to

[3] Gen 5 PCIe supports 3.94 GigaBytes/second per data lane

another vehicle / ground site, or it can be recorded for later analysis. Each has its own advantages and disadvantages.

On-board processing requires extensive data processing capability. Advances in computer processing capability can be a significant help, but this also requires the software to be preloaded into the onboard computers to perform this analysis. Frequently we don't precisely what type analysis is required, or what data is significant before the data is collected and the analysis has begun. Even with on-board data processing / data reduction, improvements in sensor technology have increased the amount of data by orders of magnitude.

Transmitting of the data to a ground site, via satellite or to another aircraft presents significant bandwidth problems. Line of Sight communications provide good communications bandwidth, but recent development in unmanned vehicles and drones makes Line of Sight impractical. Satellite communications do not provide the bandwidth and can be very expensive.

Fortunately, recent developments in solid state memory technology provides economical storage for terabytes of data. This data can be downloaded and analyzed after the mission is over.

NVMe Durability and Reliability in Harsh Defense Environments

NVMe SSDs utilize the same memory cell technologies as the mSATA SSDs so you would not expect NVMe SSDs to have a significant reliability advantage over their mSATA brothers. However, memory manufacturers have already begun converting their production lines to NVMe. Newer technology is implemented in NVMe and usually not reflected back to the older mSATA production lines.

SSD memory cell technology has a limited life, each memory cell can only be rewritten a limited number of times. Each rewrite consists of a block of data (a group of memory cells) requires erasing that entire memory block and re-writing the appropriate contents back to the memory block. This was a significant factor in SSD reliability as frequently used memory blocks tend to wear-out quickly. To improve the reliability of the memory devices, manufacturers have developed "Wear Leveling Technology" (a.k.a. "Load Leveling") inside the SSD controller. The Wear Leveling algorithms are designed to spread the write cycles to memory blocks that have less wear history. Thus you would expect an SSD that has more free space will demonstrate longer overall life than SSDs that are almost full of data. Recent advances in Wear Leveling algorithms can compensate by relocating data that is not rewritten very often to other memory cells, providing years of service before the SSD experiences a significant degradation in storage capacity.

This wear factor measured in terms of "Terabytes Written" or "TBW". SSD manufacturers wear varies from about 300TBW to over 1,500TBW depending on the technology involved and the block size (assuming a 1TB SSD). This can be misleading since a manufacturer's specification for a 3TB SSD will appear to have three times the TBW of a 1TB SSD using the same technology. Recently more sophisticated Wear Leveling algorithms will track the wear cycles on the entire SSD and will automatically relocate static data to some of the older memory blocks, providing the ability to spread

the wear to younger memory cells. The most sophisticated Wear Leveling algorithms also include error detection and error correction features, further improving the overall life and the reliability of the SSD.

Since the newer SSD designs are aimed at NVMe platforms, the more sophisticated Wear Leveling algorithms and the higher reliability memory cell technology (technology that has higher TBW) will be found in the newer NVMe SSDs. This gives the NVMe SSDs a slight edge in reliability over their mSATA brethren and that reliability edge will become more significant as the memory cell technology and the sophistication of the Wear Leveling algorithms advances.

Security Implications, Data Encryption and Protection against Threats.

Government and defense applications frequently require the data to be encrypted (Data at Rest Encryption). If your system requires NSA Suite B cryptography or FIPS 140-2 Data at Rest encryption, your options are limited. NSA Suite B cryptography requires either a Type-1 Encryptor card or CSfC encryption.

Type-1 Encryptor cards are very expensive to design and obtain NSA certification. Additionally, NSA requires a Type-1 Encryptor design to have a government program sponsor. As a result the Type-1 Encryptor cards typically lag ten years behind current computer technology and the manufacturers need a government program to fund the development costs.

The Type-1 Encryptor card inserts between the computer CPU and the storage media. Most current Type-1 Encryptors are designed specifically to support the SATA / AHCI specifications and cannot be used with NVMe type SSDs, so they cannot take advantage of NVMe transfer speeds. As SATA devices become obsolete, these SATA Type-1 Encryptor cards will also become obsolete. A few manufacturers of Type 1 Encryptor cards are working to develop NVMe Type-1 Encryptor cards, but limited funding has resulted in very high schedule and budget risk.

As an alternative to the Type-1 Encryptor, NSA has recently developed the Commercial Solutions for Classified storage (CSfC) specification, intended to reduce the development costs and make encrypted solutions more widely available. CSfC requires two completely independent NSA approved encryption algorithms (developed by separate manufacturers). Thus if one encryption algorithm has an unknown vulnerability, the other algorithm presumably would not have the same vulnerability. Due to limited demand and very high certification cost, software developers are reluctant to invest in NSA testing and certification. Programs wanting to utilize CSfC face a significant risk being able to find the two NSA certified algorithms required for CSfC.

Some CSfC solutions have resorted to a hybrid approach, using one software layer and one hardware layer. Fortunately, a few SSD manufacturers are offering an SSD with a CSfC certified hardware layer of encryption and there is at least one software algorithm that has been CSfC certified, providing a viable low-risk path for NSA CSfC system certification. But this does limit those systems to a particular model SSD with the CSfC encryption certification (which presents another type risk).

Power Loss Protection

One factor that is too frequently overlooked is what happens in the event of a momentary power interruption. Most NSA approved Data at Rest Encryption technologies utilize sophisticated rolling-code algorithms. Any lost or corrupted data can result in all the data on the storage media to be unrecoverable.

For years hard drive manufacturers have utilized a high-speed cache memory to temporarily store the data before writing it to the long-term storage media, effectively increasing the overall speed of the data transfer. A power interruption will cause data in the cache memory to be lost before the long-term storage update is complete. This is exacerbated by Wear Leveling algorithms that operate in the background, utilizing the cache to relocate data on the SSD in order to increase the SSD reliability and durability[4]. A few SSD manufacturers have implemented power loss protection in their designs; others don't consider it significant or leave it to the system designer[5]. Power Loss Protection is a factor that must be considered in any system that utilizes Data at Rest Encryption. While this applies equally to SATA and NVMe devices, it is important to remind system designers not to overlook this important requirement.

Conclusions

The SATA/AHCI bus standards were designed as an improvement over the ATA specification for attaching spinning magnetic hard disk drives (HDD) to a computer. With the development of significantly faster Solid State Drives (SSD), the SATA bus was constraining the speed of the Non-Volatile Memory chips used in the SSD. The Non-Volatile Memory Express communication standard (NVMe) is an extension of the PCIe bus (used in most modern computers) specifically to support the faster SSD drives. The NVMe standard is currently about 5x faster data transfer speed than the SATA standard and the NVMe can be expanded to meet requirements of future generations of faster SSD drives. Most computers being built today utilize NVMe and do not have provisions for attaching SATA devices. Over the next few years, SATA will become obsolete and a maintainability cost risk.

If your system requires NSA data encryption, your options are limited to a handful of SATA Type-1 Encryptors, NVMe Type-1 Encryptor or CSfC. Each encryption solution has some advantages, but the SATA Type-1 Encryptors will soon become obsolete and a DMS/MS risk (Diminishing Manufacturing Source / Material Shortage). In today's economic environment, many customers are risk adverse (cost, schedule, and obsolescence risk). Those programs should look at NVMe Type-1 Encryptors or a Hybrid CSfC solution.

---

[4] See "NVMe Durability and Reliability in Harsh Defense Environments" in this document

[5] Most server systems with data reliability requirements utilize RAID (Redundant Array of Independent Disks) to provide a back-up of critical data. Data Encryption makes RAID significantly more complicated and impractical.