Latency Analysis for NVMe/TSN

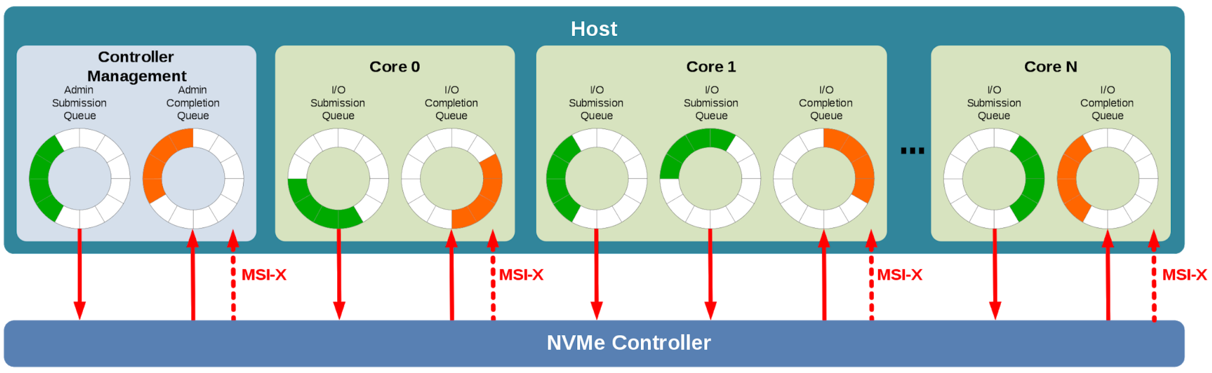
Version 1.0

Table of Content

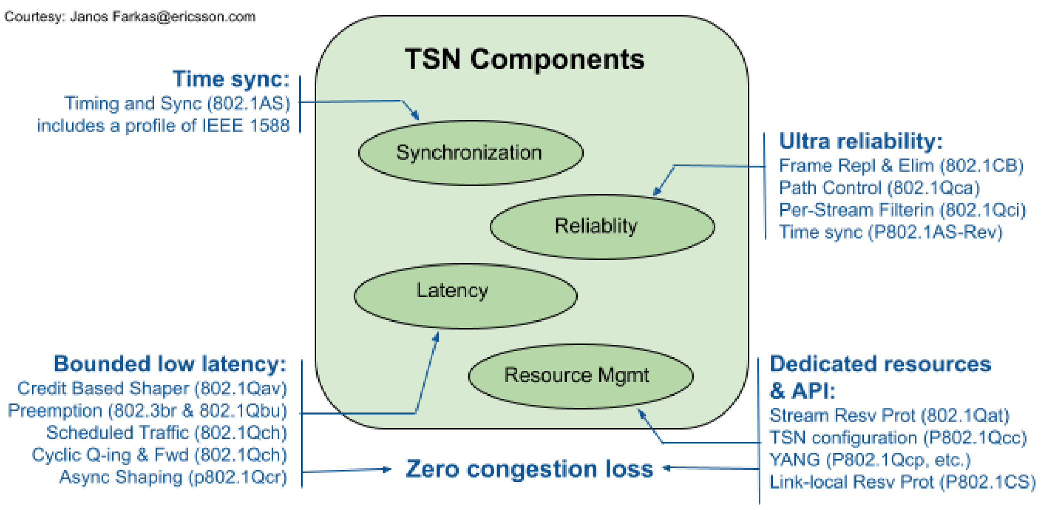
Executive Summary

NVMe/TSN is a range extension for NVM Express (NVMe) done by tunneling PCI Express (PCIe) over TCP/IP over Time Sensitive Networking (TSN). This MLE Technical Brief gives a quantitative analysis of the latency when tunneling NVMe over TCP/IP over TSN.

NVMe has proven many advantages for Solid-State Drive (SSD) storage compared to legacy standards such as Serial ATA (SATA) or Serial Attached SCSI (SAS): Most importantly, it scales in terms of bandwidth with newer PCIe Standards while reducing the protocol overhead and processing latencies.

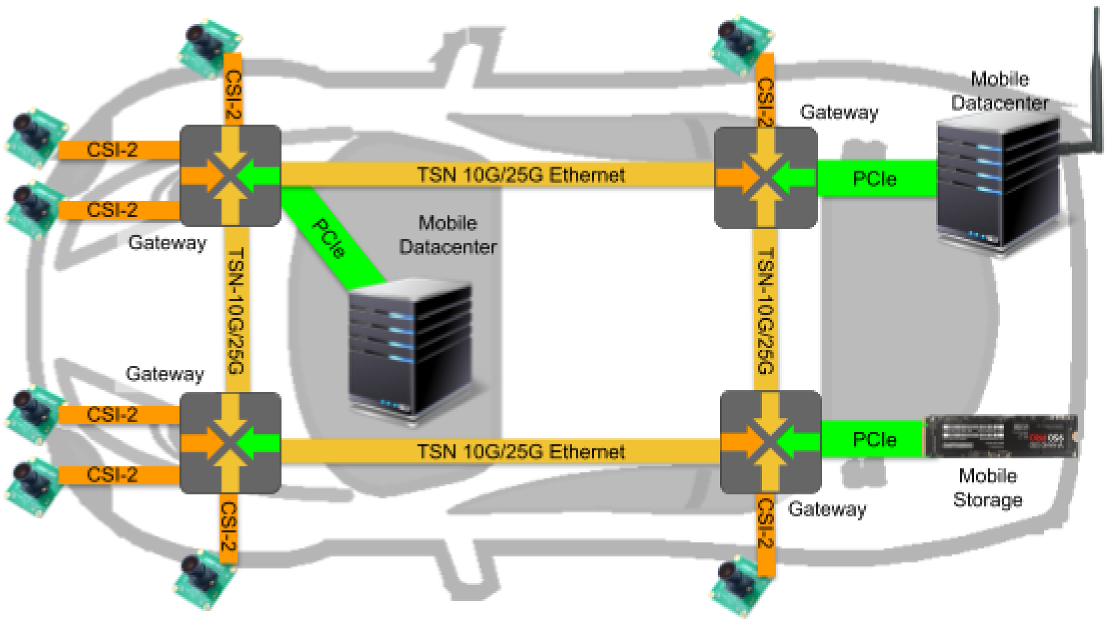


TSN is a collection of IEEE Standards providing real-time behavior in systems utilizing Ethernet. This collection combines Time Synchronization, Traffic Shaping and Traffic Scheduling, reliability and resource management:


Courtesy: Janos Farkas@ericsson.com
TSN Components
Time sync:
Timing and Sync (802.1AS)
includes a profile of IEEE 1588
Synchronization
Ultra reliability:
Frame Repl & Elim (802.1CB)
Path Control (802.1Qca)
Per-Stream Filterin (802.1Qci)
Time sync (P802.1AS-Rev)
Reliablity
Latency
Bounded low latency:
Credit Based Shaper (802.1Qav)
Preemption (802.3br & 802.1Qbu)
Scheduled Traffic (802.1Qch)
Cyclic Q-ing & Fwd (802.1Qch)
Async Shaping (p802.1Qcr)
Resource Mgmt
Dedicated resources & API:
Stream Resv Prot (802.1Qat)
TSN configuration (P802.1Qcc)
YANG (P802.1Qcp, etc.)
Link-local Resv Prot (P802.1CS)
Zero congestion loss

TSN can be combined with TCP/IP to provide so-called Deterministic Networking (DetNet) which is, for example, described in MLE Technical Brief 20201203 “Deterministic Networking with TCP-TSN-Cores for 10/25/50/100 Gigabit Ethernet” [https://www.missinglinkelectronics.com/devzone/index.php/deterministic-networking-with-tcp-tsn-cores-for-10-25-50-100-gigabit-ethernet-2].

NVMe/TSN addresses the need for disaggregated and centralized storage in certain embedded systems. For example distributed industrial systems, test & measurement systems, as well as those next generation automotive / in-vehicle networks, nowadays called “Zone-Based Architectures”. These Zone-Based Architectures aim at reducing the costs of wire harnesses in modern vehicles while providing more flexibility for changes during a vehicle’s product life cycle (mostly driven by Advanced Driver Assist Systems - ADAS). Automotive Ethernet has a great potential to become such a “backbone” for connecting those so-called Zone Computers via Zonal Gateways:

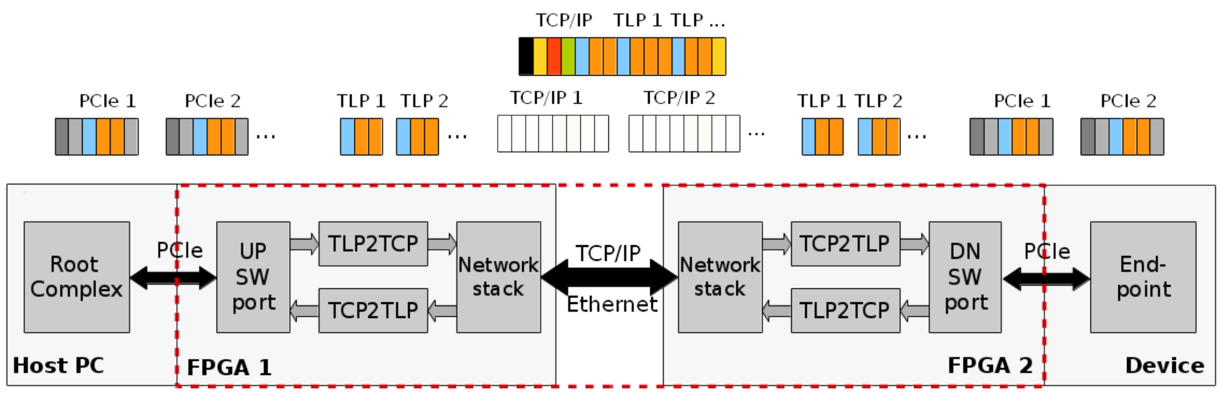


This Technical Brief starts by explaining the concept of how to tunnel PCIe using MLE’s patented Heterogeneous Packet-Based Transport technology and briefly explaining NVMe from a higher-level protocol’s perspective. We then present an experimental setup utilizing a PCIe Protocol Analyzer with time-stamping for analyzing the added latency when tunneling NVMe, including a comparison of tunneling over plain Ethernet vs tunneling over TSN over Ethernet.

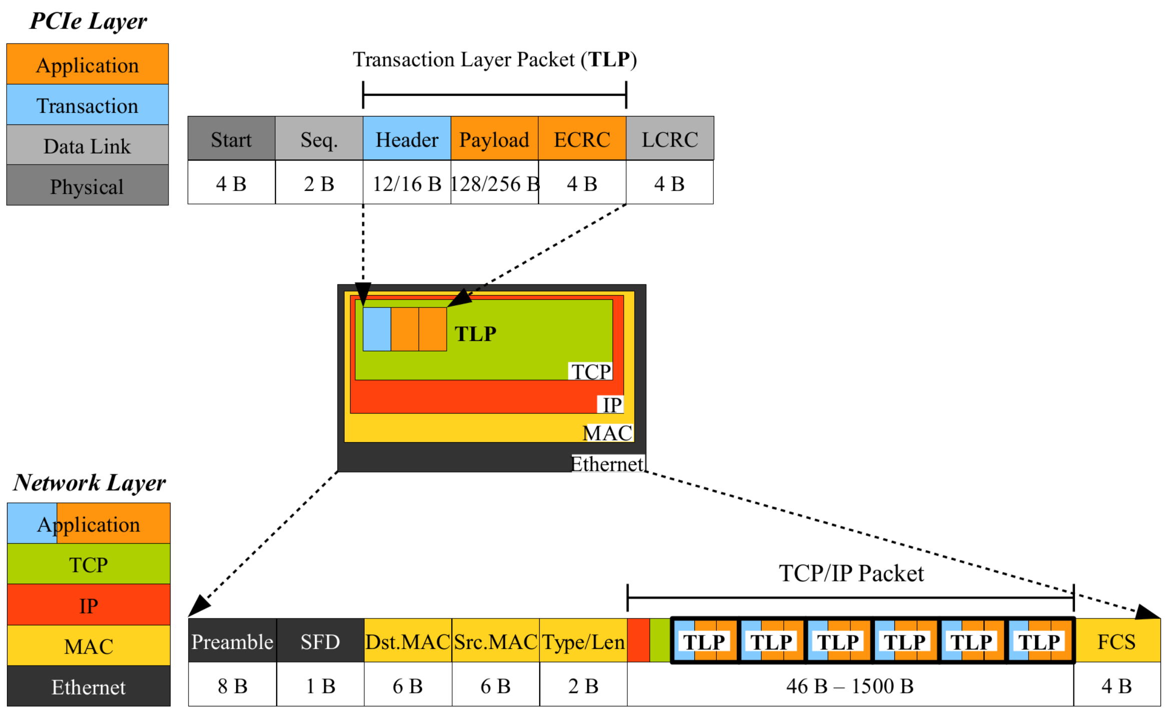
1. Concept Behind NVMe/TSN and PCIe Range Extension

The fundamental concept behind NVMe/TSN is PCIe Range Extension by implementing a “distributed” PCIe Switch. This PCIe Switch is in accordance to PCI-SIG PCI Express Base Specification 3.0 and implements one PCIe Up-Stream Switch Port connected via a reliable transport mechanism to multiple PCIe Down-Stream Switch Ports. Like all other PCIe Switches our PCIe Switch operates at so-called PCIe Transaction Layer where the PCIe data is transported via so-called Transaction-Layer Packets (TLP).

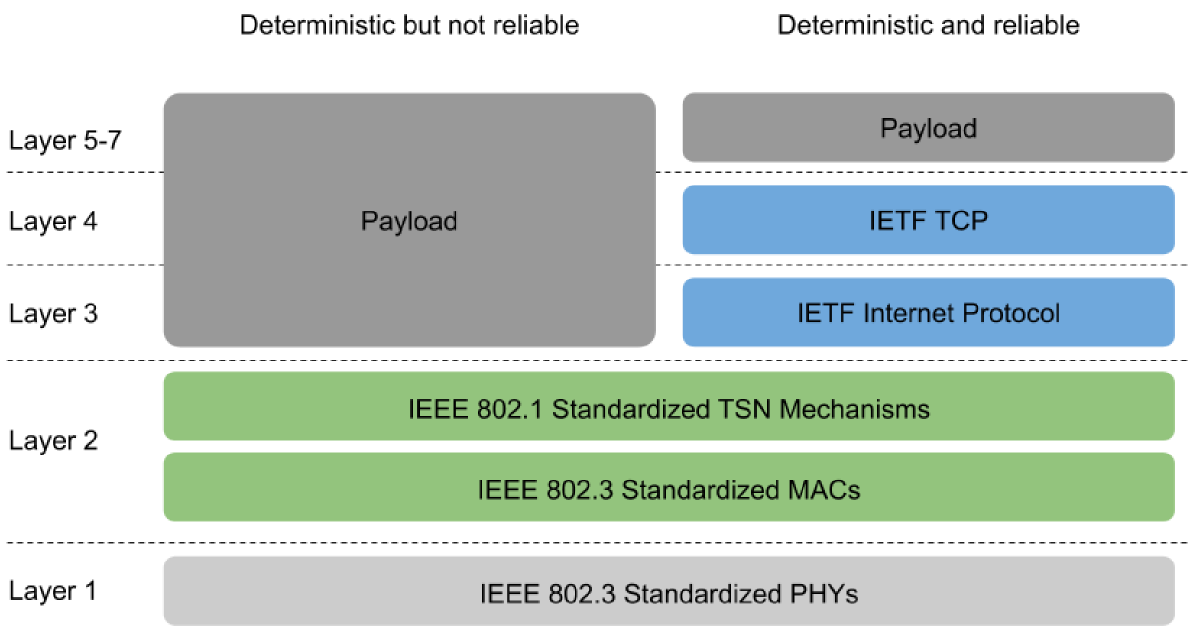
As a reliable transport mechanism we chose TCP/IP because it is widely used and widely understood, and can be transported over many different physical Ethernet standards for versatility. For low deterministic latency this tunnel is implemented as a digital circuit which builds on top of MLE’s Network Protocol Accelerator Platform (NPAP), a TCP/IP Full Accelerator from Fraunhofer HHI and which encapsulates/decapsulates the PCIe TLP into TCP/IP packets. The following figure shows an exemplary block diagram of that PCIe Range Extension: A first FPGA implements the PCIe Up-Stream Switch Port which connects to the PCIe Root (which again connects to a Host CPU), a second FPGA implements a PCIe Down-Stream Switch Port which connects to a PCIe Device (a.k.a. PCIe Endpoint). Obviously, this can also be within a deeper PCIe hierarchy built from multiple PCIe Switches. In this example, PCIe Up-Stream Switch Port and PCIe Down-Stream Switch Port are connected via 10G Ethernet to form the distributed PCIe Switch.



Color coding visualizes the encapsulation and decapsulation of the PCIe TLP within the TCP/IP packets, within the Ethernet frames:



Because NVMe is a protocol that operates on top of the PCIe protocol, the PCIe Range Extension can also be used to tunnel NVMe. And, because the implementation follows the OSI layered network architecture, the Ethernet Media Access Controller can be TSN, or not. And, it can be over any reasonable physical (Ethernet) link.

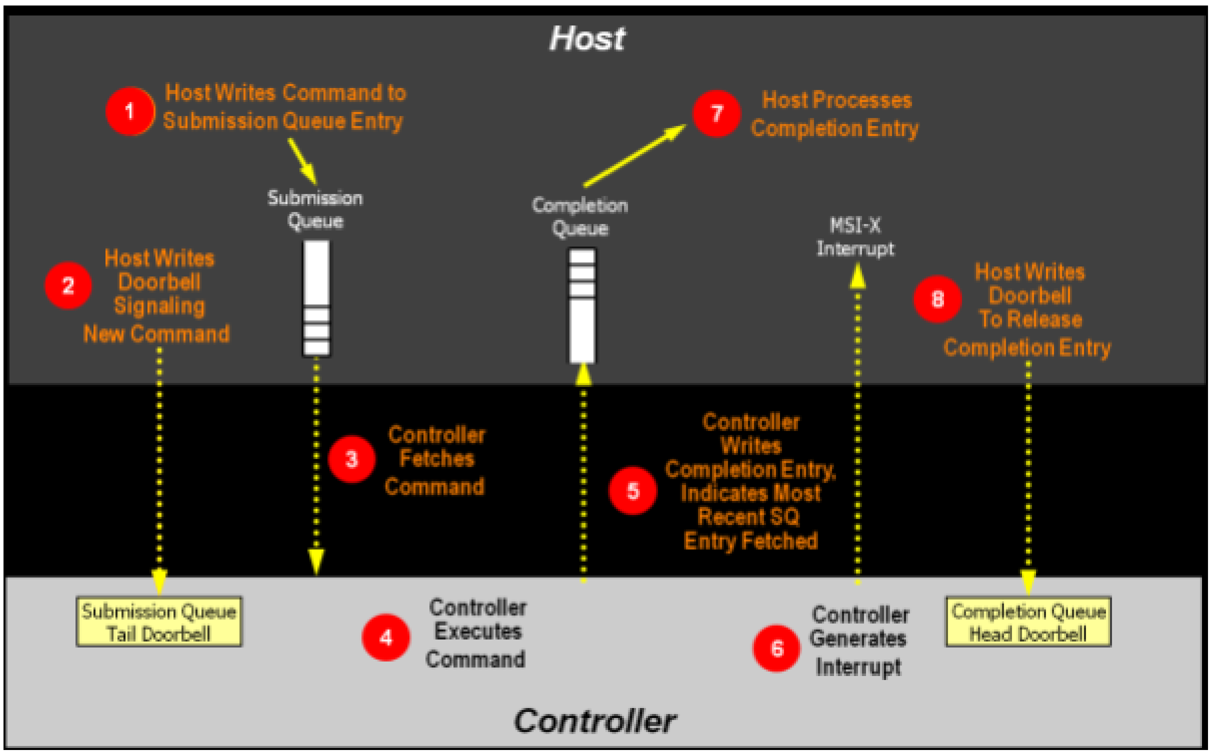

Deterministic but not reliable
Deterministic and reliable
Layer 5-7
Layer 4
Layer 3
Layer 2
Layer 1
Payload
Payload
IETF TCP
IETF Internet Protocol
IEEE 802.1 Standardized TSN Mechanisms
IEEE 802.3 Standardized MACs
IEEE 802.3 Standardized PHYs

2. Fundamentals of NVMe

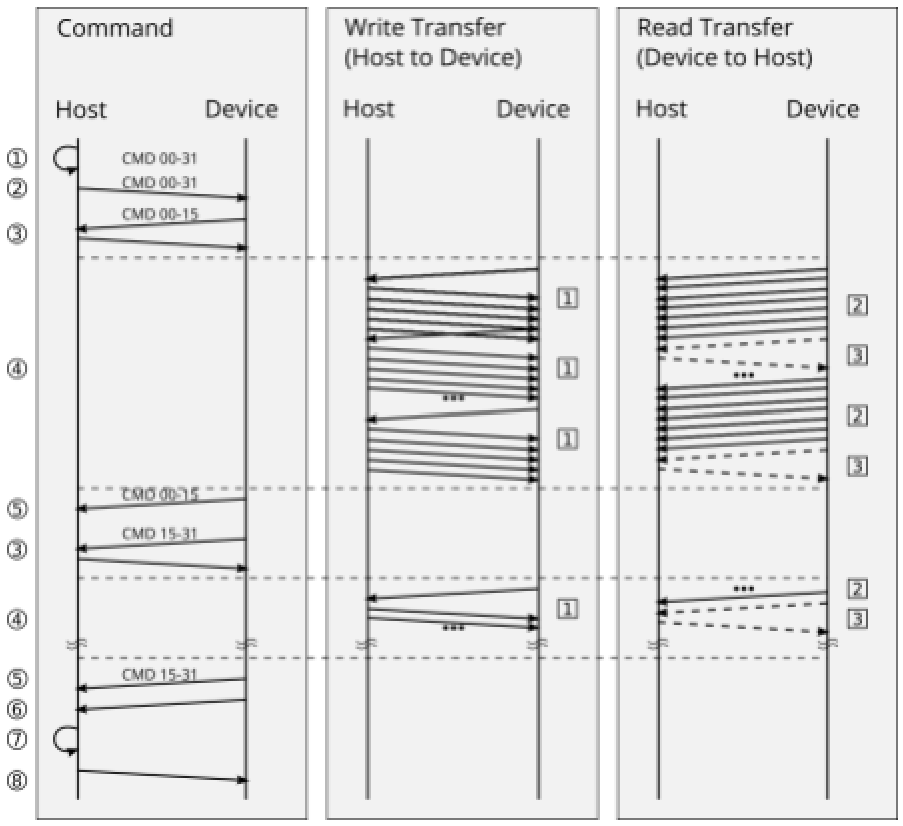
Because of the request/response scheme in NVMe, application-level latency is defined by multiple round-trips between the host and the SSD, according to the TLP sequence of the NVMe protocol. Therefore, a short backgrounder on NVMe.

It is important to know that the NVMe command interfaces are built around queues that reside in host memory and are shared between host and device (let's ignore special optimizations with queues in the device). A doorbell mechanism synchronizes the device with the host and MSI-X interrupts synchronize the host with the device. For CPU processing efficiency, the number of interrupts are kept as low as possible by interrupt aggregation (coalescing). This technique is well known from network device interfaces, which also deal with a large number of transfers per second.

According to the NVMe specification the basic command communication pattern looks like this:


Host
1 Host Writes Command to Submission Queue Entry
7 Host Processes Completion Entry
Submission Queue
Completion Queue
MSI-X Interrupt
2 Host Writes Doorbell Signaling New Command
8 Host Writes Doorbell To Release Completion Entry
3 Controller Fetches Command
5 Controller Writes Completion Entry, Indicates Most Recent SQ Entry Fetched
Submission Queue Tail Doorbell
4 Controller Executes Command
6 Controller Generates Interrupt
Completion Queue Head Doorbell
Controller

To better align this with the communication pattern, we show a sequence diagram with the respective TLP flow shown on the PCIe level. Further down in this document we will observe this communication pattern using a PCIe protocol analyzer and actual hardware/software.



Only communication related to data transfers is shown in this sequence diagram, additional device management commands, e.g. power state management, etc, are not taken into account. On the left-hand side, the command interface and its communication pattern is shown. On the right-hand side, read transfers are shown, meaning data transfers from device to the host (read from disk), are outlined. In the middle write transfers, meaning data transfers from host to device (write to disk), are shown. Please note that the commands are executed by the device with the help of its DMA controller(s) which leads to reverse access patterns in comparison to the command, e.g. a read command from the host to the device results in the device's DMA controller writing to the host memory.

The numbering scheme between both figures is the same. In addition to the circled numbers marking command associated action, squared numbers are introduced marking data transfer (command execution) associated actions.

1. At first, commands may be en-queued at any time (1). The doorbell writes signaling that any number of new commands have been en-queued (2).
2. Afterwards the device can choose to fetch any number of commands from the queue (3).
3. Then the commands are executed (4).
4. When the commands are finished the corresponding completion is en-queued in the completion queue (5).
5. To communicate to the host that the completion queue has been updated, the device issues an interrupt (6).
   To lower the interrupt rate, which otherwise may easily overload the Host CPUs processing capabilities, interrupts typically are aggregated (so-called Interrupt Coalescing). The interrupts are typically signaled via PCIe Message Signaled Interrupts (MSI / MSI-X), which are in-band interrupts signaled with regular posted write request TLPs.
6. On reception of the interrupt, the host processes the completion queue entries (7).
7. When processing the completion queue entries has finished the host signals this via another doorbell write (8), freeing up the respective entries in the completion queue.

Within the data transfer phase (4) the device is in charge to schedule memory accesses in a performance optimized manner.

As previously mentioned a host write transfer to the device actually results in a PCIe device DMA read access [1] fetching the data from host memory. Due to requirements by the PCIe specification, a read request may never cross a 4 KiB boundary. So, the maximum memory portion which could be read during one read access is 4 KiB. On the other hand the TLP payload size is usually much lower, e.g. typically 128 Bytes or 256 Bytes, seldom 512 Bytes, theoretically up to 4 KiB. The actual payload size of the end to end link is defined by system and device parameters and is discovered during device enumeration phase. Additionally the maximum read request size is the maximum memory size allowed to be requested by one TLP. This parameter is set per device and Linux typically sets it to 512 Bytes. Drivers may override the default setting for specific devices to optimize for read accesses. First, this means that one read request TLP most likely results in some number of completion TLPs carrying the data, as shown in [1]. Second, the setting of the maximum read request size is typically not tuned for distributed systems which tunnel NVMe over TSN.

A read transfer from device to host is actually a write access [2] executed by the device DMA controller. PCIe memory writes are posted, which means they are neither nacked nor acked on the transport layer, because the DLLP takes care about data integrity. This way the device may just issue writes until a command is completed.

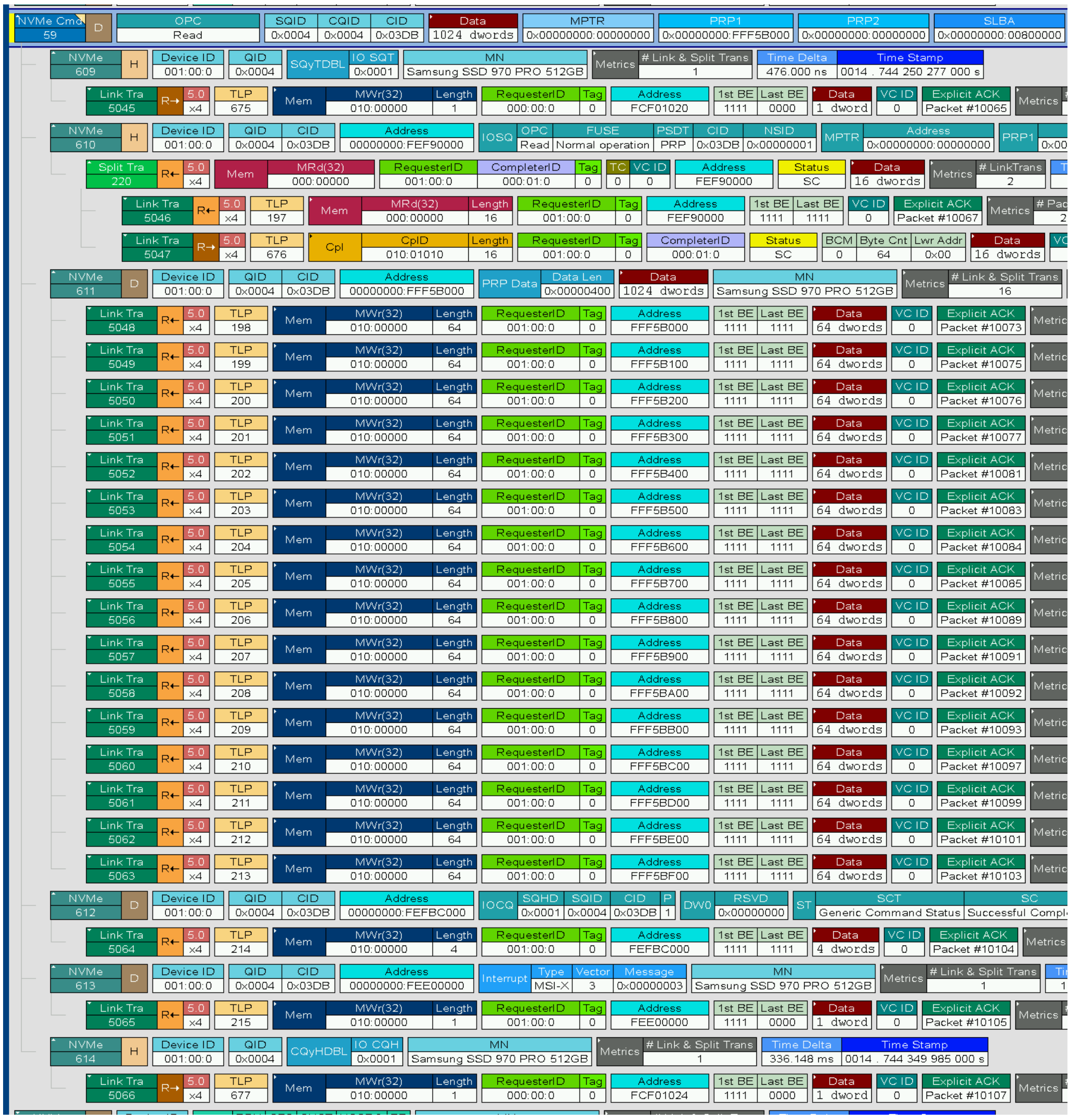
The following explains how NVMe behaves for SSD Read and SSD Write operations. For visualization purposes we have added screenshots of a PCIe Protocol Analyzer which captured those NVMe / PCIe TLPs.

2.1. Anatomy of an NVMe Read Request

Below we show some practical examples of NVMe communication using a Samsung 970 Pro SSD as the DUT to read and write a single 4 KiB block. The PCIe protocol analyzer with matching software was used to observe PCIe communications and decode/summarize it on different levels.

The following figure shows a 4 KiB read (device to host data transfer) in NVMe protocol view with low level TLPs expanded. NVMe Transaction 609 is executed as Link Transaction 5045, which is the NVMe submission queue tail doorbell update (2) in the form of a Memory Write TLP from host to device.

The device then fetches the new submission queue entry in a 64 Byte read (3). This is seen as NVMe Transaction 610 in the form of Link Transactions 5046 and 5047. Next, the device writes the requested data from flash to host memory (4) [2] in NVMe Transaction 611, which is executed as sixteen 256 Byte payload (equals the maximum payload setting) write TLPs (Link Transaction 5048 through 5063).



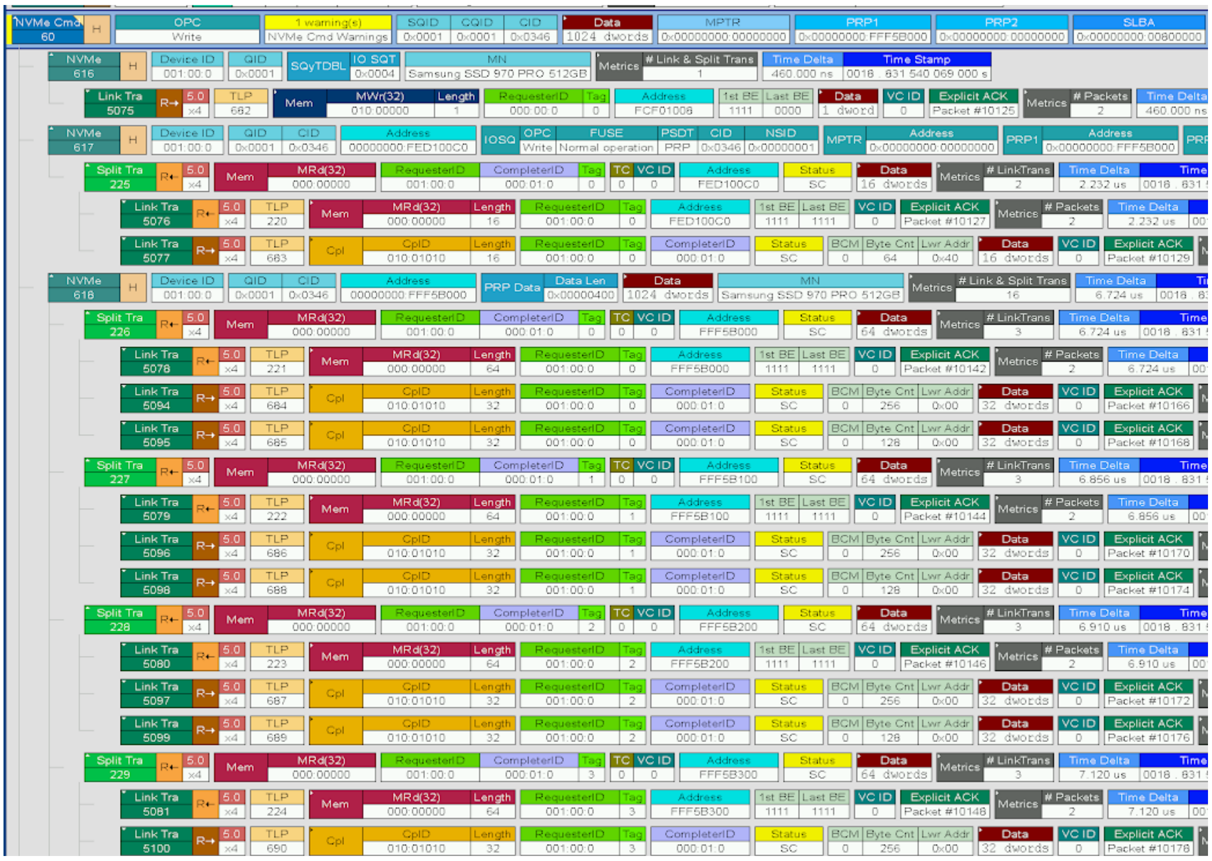
```
NVMe Cmd 59 | D | OPC: Read | SQID: 0x0004 | CQID: 0x0004 | CID: 0x03DB | Data: 1024 dwords | MPTR: 0x00000000:00000000 | PRP1: 0x00000000:FFF5B000 | PRP2: 0x00000000:00000000 | SLBA: 0x00000000:00800000
  NVMe 609 | H | Device ID: 001:00:0 | QID: 0x0004 | SQyTDBL | IO SQT: 0x0001 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 476.000 ns | Time Stamp: 0014 . 744 250 277 000 s
    Link Tra 5045 | R→ | 5.0 x4 | TLP 675 | Mem | MWr(32) 010:00000 | Length 1 | RequesterID 000:00:0 | Tag 0 | Address FCF01020 | 1st BE 1111 | Last BE 0000 | Data 1 dword | VC ID 0 | Explicit ACK Packet #10065 | Metrics
  NVMe 610 | H | Device ID: 001:00:0 | QID: 0x0004 | CID: 0x03DB | Address: 00000000:FEF90000 | IOSQ | OPC: Read | FUSE: Normal operation | PSDT: PRP | CID: 0x03DB | NSID: 0x00000001 | MPTR | Address: 0x00000000:00000000 | PRP1 | 0x00
    Split Tra 220 | R← | 5.0 x4 | Mem | MRd(32) 000:00000 | RequesterID 001:00:0 | CompleterID 000:01:0 | Tag 0 | TC 0 | VC ID 0 | Address FEF90000 | Status SC | Data 16 dwords | Metrics | # LinkTrans 2
      Link Tra 5046 | R← | 5.0 x4 | TLP 197 | Mem | MRd(32) 000:00000 | Length 16 | RequesterID 001:00:0 | Tag 0 | Address FEF90000 | 1st BE 1111 | Last BE 1111 | VC ID 0 | Explicit ACK Packet #10067 | Metrics | # Pac 2
      Link Tra 5047 | R→ | 5.0 x4 | TLP 676 | Cpl | CplD 010:01010 | Length 16 | RequesterID 001:00:0 | Tag 0 | CompleterID 000:01:0 | Status SC | BCM 0 | Byte Cnt 64 | Lwr Addr 0x00 | Data 16 dwords
  NVMe 611 | D | Device ID: 001:00:0 | QID: 0x0004 | CID: 0x03DB | Address: 00000000:FFF5B000 | PRP Data | Data Len: 0x00000400 | Data: 1024 dwords | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 16
    Link Tra 5048 | R← | 5.0 x4 | TLP 198 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5B000 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10073 | Metric
    Link Tra 5049 | R← | 5.0 x4 | TLP 199 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5B100 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10075 | Metric
    Link Tra 5050 | R← | 5.0 x4 | TLP 200 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5B200 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10076 | Metric
    Link Tra 5051 | R← | 5.0 x4 | TLP 201 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5B300 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10077 | Metric
    Link Tra 5052 | R← | 5.0 x4 | TLP 202 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5B400 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10081 | Metric
    Link Tra 5053 | R← | 5.0 x4 | TLP 203 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5B500 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10083 | Metric
    Link Tra 5054 | R← | 5.0 x4 | TLP 204 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5B600 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10084 | Metric
    Link Tra 5055 | R← | 5.0 x4 | TLP 205 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5B700 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10085 | Metric
    Link Tra 5056 | R← | 5.0 x4 | TLP 206 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5B800 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10089 | Metric
    Link Tra 5057 | R← | 5.0 x4 | TLP 207 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5B900 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10091 | Metric
    Link Tra 5058 | R← | 5.0 x4 | TLP 208 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5BA00 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10092 | Metric
    Link Tra 5059 | R← | 5.0 x4 | TLP 209 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5BB00 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10093 | Metric
    Link Tra 5060 | R← | 5.0 x4 | TLP 210 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5BC00 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10097 | Metric
    Link Tra 5061 | R← | 5.0 x4 | TLP 211 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5BD00 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10099 | Metric
    Link Tra 5062 | R← | 5.0 x4 | TLP 212 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5BE00 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10101 | Metric
    Link Tra 5063 | R← | 5.0 x4 | TLP 213 | Mem | MWr(32) 010:00000 | Length 64 | RequesterID 001:00:0 | Tag 0 | Address FFF5BF00 | 1st BE 1111 | Last BE 1111 | Data 64 dwords | VC ID 0 | Explicit ACK Packet #10103 | Metric
  NVMe 612 | D | Device ID: 001:00:0 | QID: 0x0004 | CID: 0x03DB | Address: 00000000:FEFBC000 | IOCQ | SQHD: 0x0001 | SQID: 0x0004 | CID: 0x03DB | P: 1 | DW0 | RSVD: 0x00000000 | ST | SCT: Generic Command Status | SC: Successful Compl
    Link Tra 5064 | R← | 5.0 x4 | TLP 214 | Mem | MWr(32) 010:00000 | Length 4 | RequesterID 001:00:0 | Tag 0 | Address FEFBC000 | 1st BE 1111 | Last BE 1111 | Data 4 dwords | VC ID 0 | Explicit ACK Packet #10104 | Metrics
  NVMe 613 | D | Device ID: 001:00:0 | QID: 0x0004 | CID: 0x03DB | Address: 00000000:FEE00000 | Interrupt | Type: MSI-X | Vector: 3 | Message: 0x00000003 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Ti 1
    Link Tra 5065 | R← | 5.0 x4 | TLP 215 | Mem | MWr(32) 010:00000 | Length 1 | RequesterID 001:00:0 | Tag 0 | Address FEE00000 | 1st BE 1111 | Last BE 0000 | Data 1 dword | VC ID 0 | Explicit ACK Packet #10105 | Metrics
  NVMe 614 | H | Device ID: 001:00:0 | QID: 0x0004 | CQyHDBL | IO CQH: 0x0001 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 336.148 ms | Time Stamp: 0014 . 744 349 985 000 s
    Link Tra 5066 | R→ | 5.0 x4 | TLP 677 | Mem | MWr(32) 010:00000 | Length 1 | RequesterID 000:00:0 | Tag 0 | Address FCF01024 | 1st BE 1111 | Last BE 0000 | Data 1 dword | VC ID 0 | Explicit ACK Packet #10107 | Metrics
```

NVMe Transaction 612 updates the completion queue in host memory (5) by writing the 16 Byte completion queue entry via Link Transaction 5064. This is followed by an MSI-X interrupt (6) in NVMe Transaction 613, notifying the host of the new completion queue entry.

To conclude the IO command the host informs the device via NVMe Transaction 614 that it has consumed the completion queue entry by updating the completion queue head doorbell (8).

2.2. Anatomy of an NVMe Write Request

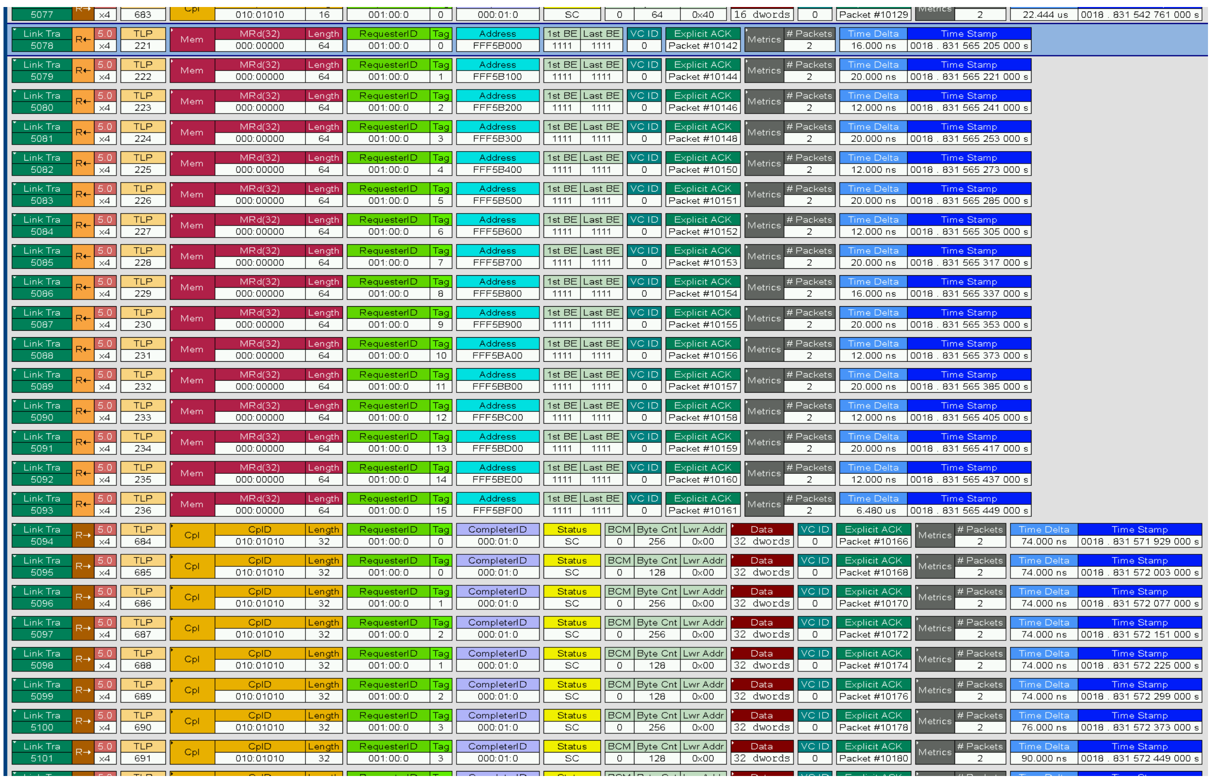
Writes to the device, as seen in the figure below, work in exactly the same way, except that of course the actual data transfer is performed by the device reading from host memory and writing data to flash.



```
NVMe Cmd 60 | H | OPC: Write | 1 warning(s): NVMe Cmd Warnings | SQID: 0x0001 | CQID: 0x0001 | CID: 0x0346 | Data: 1024 dwords | MPTR: 0x00000000:00000000 | PRP1: 0x00000000:FFF5B000 | PRP2: 0x00000000:00000000 | SLBA: 0x00000000:00800000
  NVMe 616 | H | Device ID: 001:00:0 | QID: 0x0001 | SQyTDBL | IO SQT: 0x0004 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 460.000 ns | Time Stamp: 0018 . 831 540 069 000 s
    Link Tra 5075 | R→ | 5.0 x4 | TLP: 682 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF01008 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | VC ID: 0 | Explicit ACK: Packet #10125 | Metrics | # Packets: 2 | Time Delta: 460.000 ns
  NVMe 617 | H | Device ID: 001:00:0 | QID: 0x0001 | CID: 0x0346 | Address: 00000000:FED100C0 | IOSQ | OPC: Write | FUSE: Normal operation | PSDT: PRP | CID: 0x0346 | NSID: 0x00000001 | MPTR | Address: 0x00000000:00000000 | PRP1 | Address: 0x00000000:FFF5B000 | PRP
    Split Tra 225 | R← | 5.0 x4 | Mem | MRd(32): 000:00000 | RequesterID: 001:00:0 | CompleterID: 000:01:0 | Tag: 0 | TC: 0 | VC ID: 0 | Address: FED100C0 | Status: SC | Data: 16 dwords | Metrics | # LinkTrans: 2 | Time Delta: 2.232 us | Time: 0018 . 831
      Link Tra 5076 | R← | 5.0 x4 | TLP: 220 | Mem | MRd(32): 000:00000 | Length: 16 | RequesterID: 001:00:0 | Tag: 0 | Address: FED100C0 | 1st BE: 1111 | Last BE: 1111 | VC ID: 0 | Explicit ACK: Packet #10127 | Metrics | # Packets: 2 | Time Delta: 2.232 us
      Link Tra 5077 | R→ | 5.0 x4 | TLP: 683 | Cpl | CplD: 010:01010 | Length: 16 | RequesterID: 001:00:0 | Tag: 0 | CompleterID: 000:01:0 | Status: SC | BCM: 0 | Byte Cnt: 64 | Lwr Addr: 0x40 | Data: 16 dwords | VC ID: 0 | Explicit ACK: Packet #10129
  NVMe 618 | H | Device ID: 001:00:0 | QID: 0x0001 | CID: 0x0346 | Address: 00000000:FFF5B000 | PRP Data | Data Len: 0x00000400 | Data: 1024 dwords | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 16 | Time Delta: 6.724 us | Ti: 0018 . 83
    Split Tra 226 | R← | 5.0 x4 | Mem | MRd(32): 000:00000 | RequesterID: 001:00:0 | CompleterID: 000:01:0 | Tag: 0 | TC: 0 | VC ID: 0 | Address: FFF5B000 | Status: SC | Data: 64 dwords | Metrics | # LinkTrans: 3 | Time Delta: 6.724 us | Time: 0018 . 831
      Link Tra 5078 | R← | 5.0 x4 | TLP: 221 | Mem | MRd(32): 000:00000 | Length: 64 | RequesterID: 001:00:0 | Tag: 0 | Address: FFF5B000 | 1st BE: 1111 | Last BE: 1111 | VC ID: 0 | Explicit ACK: Packet #10142 | Metrics | # Packets: 2 | Time Delta: 6.724 us
      Link Tra 5094 | R→ | 5.0 x4 | TLP: 684 | Cpl | CplD: 010:01010 | Length: 32 | RequesterID: 001:00:0 | Tag: 0 | CompleterID: 000:01:0 | Status: SC | BCM: 0 | Byte Cnt: 256 | Lwr Addr: 0x00 | Data: 32 dwords | VC ID: 0 | Explicit ACK: Packet #10166
      Link Tra 5095 | R→ | 5.0 x4 | TLP: 685 | Cpl | CplD: 010:01010 | Length: 32 | RequesterID: 001:00:0 | Tag: 0 | CompleterID: 000:01:0 | Status: SC | BCM: 0 | Byte Cnt: 128 | Lwr Addr: 0x00 | Data: 32 dwords | VC ID: 0 | Explicit ACK: Packet #10168
    Split Tra 227 | R← | 5.0 x4 | Mem | MRd(32): 000:00000 | RequesterID: 001:00:0 | CompleterID: 000:01:0 | Tag: 1 | TC: 0 | VC ID: 0 | Address: FFF5B100 | Status: SC | Data: 64 dwords | Metrics | # LinkTrans: 3 | Time Delta: 6.856 us | Time: 0018 . 831
      Link Tra 5079 | R← | 5.0 x4 | TLP: 222 | Mem | MRd(32): 000:00000 | Length: 64 | RequesterID: 001:00:0 | Tag: 1 | Address: FFF5B100 | 1st BE: 1111 | Last BE: 1111 | VC ID: 0 | Explicit ACK: Packet #10144 | Metrics | # Packets: 2 | Time Delta: 6.856 us
      Link Tra 5096 | R→ | 5.0 x4 | TLP: 686 | Cpl | CplD: 010:01010 | Length: 32 | RequesterID: 001:00:0 | Tag: 1 | CompleterID: 000:01:0 | Status: SC | BCM: 0 | Byte Cnt: 256 | Lwr Addr: 0x00 | Data: 32 dwords | VC ID: 0 | Explicit ACK: Packet #10170
      Link Tra 5098 | R→ | 5.0 x4 | TLP: 688 | Cpl | CplD: 010:01010 | Length: 32 | RequesterID: 001:00:0 | Tag: 1 | CompleterID: 000:01:0 | Status: SC | BCM: 0 | Byte Cnt: 128 | Lwr Addr: 0x00 | Data: 32 dwords | VC ID: 0 | Explicit ACK: Packet #10174
    Split Tra 228 | R← | 5.0 x4 | Mem | MRd(32): 000:00000 | RequesterID: 001:00:0 | CompleterID: 000:01:0 | Tag: 2 | TC: 0 | VC ID: 0 | Address: FFF5B200 | Status: SC | Data: 64 dwords | Metrics | # LinkTrans: 3 | Time Delta: 6.910 us | Time: 0018 . 831
      Link Tra 5080 | R← | 5.0 x4 | TLP: 223 | Mem | MRd(32): 000:00000 | Length: 64 | RequesterID: 001:00:0 | Tag: 2 | Address: FFF5B200 | 1st BE: 1111 | Last BE: 1111 | VC ID: 0 | Explicit ACK: Packet #10146 | Metrics | # Packets: 2 | Time Delta: 6.910 us
      Link Tra 5097 | R→ | 5.0 x4 | TLP: 687 | Cpl | CplD: 010:01010 | Length: 32 | RequesterID: 001:00:0 | Tag: 2 | CompleterID: 000:01:0 | Status: SC | BCM: 0 | Byte Cnt: 256 | Lwr Addr: 0x00 | Data: 32 dwords | VC ID: 0 | Explicit ACK: Packet #10172
      Link Tra 5099 | R→ | 5.0 x4 | TLP: 689 | Cpl | CplD: 010:01010 | Length: 32 | RequesterID: 001:00:0 | Tag: 2 | CompleterID: 000:01:0 | Status: SC | BCM: 0 | Byte Cnt: 128 | Lwr Addr: 0x00 | Data: 32 dwords | VC ID: 0 | Explicit ACK: Packet #10176
    Split Tra 229 | R← | 5.0 x4 | Mem | MRd(32): 000:00000 | RequesterID: 001:00:0 | CompleterID: 000:01:0 | Tag: 3 | TC: 0 | VC ID: 0 | Address: FFF5B300 | Status: SC | Data: 64 dwords | Metrics | # LinkTrans: 3 | Time Delta: 7.120 us | Time: 0018 . 831
      Link Tra 5081 | R← | 5.0 x4 | TLP: 224 | Mem | MRd(32): 000:00000 | Length: 64 | RequesterID: 001:00:0 | Tag: 3 | Address: FFF5B300 | 1st BE: 1111 | Last BE: 1111 | VC ID: 0 | Explicit ACK: Packet #10148 | Metrics | # Packets: 2 | Time Delta: 7.120 us
      Link Tra 5100 | R→ | 5.0 x4 | TLP: 690 | Cpl | CplD: 010:01010 | Length: 32 | RequesterID: 001:00:0 | Tag: 3 | CompleterID: 000:01:0 | Status: SC | BCM: 0 | Byte Cnt: 256 | Lwr Addr: 0x00 | Data: 32 dwords | VC ID: 0 | Explicit ACK: Packet #10178
```

The device's DMA engine reads (4) [1] can be seen in NVMe Transaction 618. It is interesting to see that the DMA engine requests only 256 Bytes per read request TLP, while a 512 Bytes maximum read request size is configured, and that the host responds with two 128 Byte completions per read request even though a maximum payload of 256 Bytes would have been allowed. This is implementation specific behavior and a design decision by the implementers (in this case Samsung and AMD).

Note that the high level NVMe Transaction view does not necessarily reflect the actual order of TLPs, but arranges them into a logical NVMe Transaction view. Below you can see the Link Transaction view providing more insight into the actual ordering and showing how multiple read requests are issued (with incrementing tags) in rapid sequence before, much later, the first completions return to the device. This is an important latency mitigation strategy, as discussed below.



Link Tra		5.0	TLP		Cpl	CplD	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	VC ID	Explicit ACK	Metrics	# Packets	Time Delta	Time Stamp
5077	R→	x4	683		Cpl	010:01010	16	001:00:0	0	000:01:0	SC	0	64	0x40	16 dwords	0	Packet #10129	Metrics	2	22.444 us	0018 . 831 542 761 000 s

Link Tra		5.0	TLP	Mem	MRd(32)	Length	RequesterID	Tag	Address	1st BE	Last BE	VC ID	Explicit ACK	Metrics	# Packets	Time Delta	Time Stamp
5078	R←	x4	221	Mem	000:00000	64	001:00:0	0	FFF5B000	1111	1111	0	Packet #10142	Metrics	2	16.000 ns	0018 . 831 565 205 000 s
5079	R←	x4	222	Mem	000:00000	64	001:00:0	1	FFF5B100	1111	1111	0	Packet #10144	Metrics	2	20.000 ns	0018 . 831 565 221 000 s
5080	R←	x4	223	Mem	000:00000	64	001:00:0	2	FFF5B200	1111	1111	0	Packet #10146	Metrics	2	12.000 ns	0018 . 831 565 241 000 s
5081	R←	x4	224	Mem	000:00000	64	001:00:0	3	FFF5B300	1111	1111	0	Packet #10148	Metrics	2	20.000 ns	0018 . 831 565 253 000 s
5082	R←	x4	225	Mem	000:00000	64	001:00:0	4	FFF5B400	1111	1111	0	Packet #10150	Metrics	2	12.000 ns	0018 . 831 565 273 000 s
5083	R←	x4	226	Mem	000:00000	64	001:00:0	5	FFF5B500	1111	1111	0	Packet #10151	Metrics	2	20.000 ns	0018 . 831 565 285 000 s
5084	R←	x4	227	Mem	000:00000	64	001:00:0	6	FFF5B600	1111	1111	0	Packet #10152	Metrics	2	12.000 ns	0018 . 831 565 305 000 s
5085	R←	x4	228	Mem	000:00000	64	001:00:0	7	FFF5B700	1111	1111	0	Packet #10153	Metrics	2	20.000 ns	0018 . 831 565 317 000 s
5086	R←	x4	229	Mem	000:00000	64	001:00:0	8	FFF5B800	1111	1111	0	Packet #10154	Metrics	2	16.000 ns	0018 . 831 565 337 000 s
5087	R←	x4	230	Mem	000:00000	64	001:00:0	9	FFF5B900	1111	1111	0	Packet #10155	Metrics	2	20.000 ns	0018 . 831 565 353 000 s
5088	R←	x4	231	Mem	000:00000	64	001:00:0	10	FFF5BA00	1111	1111	0	Packet #10156	Metrics	2	12.000 ns	0018 . 831 565 373 000 s
5089	R←	x4	232	Mem	000:00000	64	001:00:0	11	FFF5BB00	1111	1111	0	Packet #10157	Metrics	2	20.000 ns	0018 . 831 565 385 000 s
5090	R←	x4	233	Mem	000:00000	64	001:00:0	12	FFF5BC00	1111	1111	0	Packet #10158	Metrics	2	12.000 ns	0018 . 831 565 405 000 s
5091	R←	x4	234	Mem	000:00000	64	001:00:0	13	FFF5BD00	1111	1111	0	Packet #10159	Metrics	2	20.000 ns	0018 . 831 565 417 000 s
5092	R←	x4	235	Mem	000:00000	64	001:00:0	14	FFF5BE00	1111	1111	0	Packet #10160	Metrics	2	12.000 ns	0018 . 831 565 437 000 s
5093	R←	x4	236	Mem	000:00000	64	001:00:0	15	FFF5BF00	1111	1111	0	Packet #10161	Metrics	2	6.480 us	0018 . 831 565 449 000 s

Link Tra		5.0	TLP	Cpl	CplD	Length	RequesterID	Tag	CompleterID	Status	BCM	Byte Cnt	Lwr Addr	Data	VC ID	Explicit ACK	Metrics	# Packets	Time Delta	Time Stamp
5094	R→	x4	684	Cpl	010:01010	32	001:00:0	0	000:01:0	SC	0	256	0x00	32 dwords	0	Packet #10166	Metrics	2	74.000 ns	0018 . 831 571 929 000 s
5095	R→	x4	685	Cpl	010:01010	32	001:00:0	0	000:01:0	SC	0	128	0x00	32 dwords	0	Packet #10168	Metrics	2	74.000 ns	0018 . 831 572 003 000 s
5096	R→	x4	686	Cpl	010:01010	32	001:00:0	1	000:01:0	SC	0	256	0x00	32 dwords	0	Packet #10170	Metrics	2	74.000 ns	0018 . 831 572 077 000 s
5097	R→	x4	687	Cpl	010:01010	32	001:00:0	2	000:01:0	SC	0	256	0x00	32 dwords	0	Packet #10172	Metrics	2	74.000 ns	0018 . 831 572 151 000 s
5098	R→	x4	688	Cpl	010:01010	32	001:00:0	1	000:01:0	SC	0	128	0x00	32 dwords	0	Packet #10174	Metrics	2	74.000 ns	0018 . 831 572 225 000 s
5099	R→	x4	689	Cpl	010:01010	32	001:00:0	2	000:01:0	SC	0	128	0x00	32 dwords	0	Packet #10176	Metrics	2	74.000 ns	0018 . 831 572 299 000 s
5100	R→	x4	690	Cpl	010:01010	32	001:00:0	3	000:01:0	SC	0	256	0x00	32 dwords	0	Packet #10178	Metrics	2	76.000 ns	0018 . 831 572 373 000 s
5101	R→	x4	691	Cpl	010:01010	32	001:00:0	3	000:01:0	SC	0	128	0x00	32 dwords	0	Packet #10180	Metrics	2	90.000 ns	0018 . 831 572 449 000 s

3. Experimental Test Setup

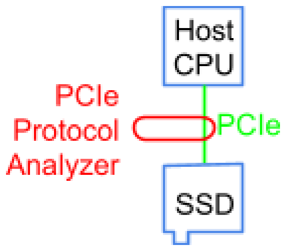
The following describes the experimental test setup used for analyzing the latency of NVMe/TSN. Tests were done for PCIe Gen2 with 5GT/s and when running TSN over 1 Gigabit Ethernet, and, hence, reflect somewhat a worst-case scenario (obviously, when using 10 GigE or 25 GigE latencies will be shorter which is better).

The LeCroy Summit T28 PCIe Protocol Analyzer was used for taking PCIe latency (and protocol) measurements. Restricted by the LeCroy Summit T28 which (only) supports PCIe Gen2 with 5 GT/s data rates, all measurements were taken for PCIe Gen2. Hence, we can assume that latency numbers, in reality for PCIe Gen3 / Gen4, will be slightly better than measured. To compute the latency we used the time stamping capabilities of the LeCroy Summit T28, for which LeCroy reports a precision of +/- 8 ns.

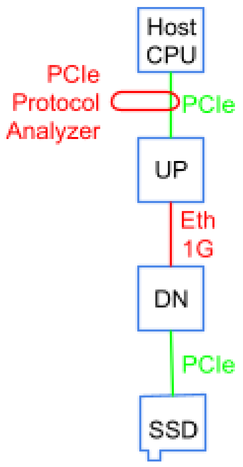
Our experimental setup uses a standard mini-ITX PC with an AMD Ryzen 3200G CPU running Debian 9 64bit Linux and a Samsung 970 Pro 512 GB SSD. The FPGAs used for implementing the PCIe Upstream Switch Port and the PCIe Down-Stream Switch Port are Xilinx ZU19EG Zynq UltraScale+ MPSoC instantiating Xilinx "PG047 - 1G/2.5G Ethernet PCS/PMA or SGMII (v16.2)" and the Xilinx "PG213 - UltraScale+ Devices Integrated Block for PCI Express v1.3" for PCIe connectivity.

For comparison we analyze and compare the latency of the following three different topologies:

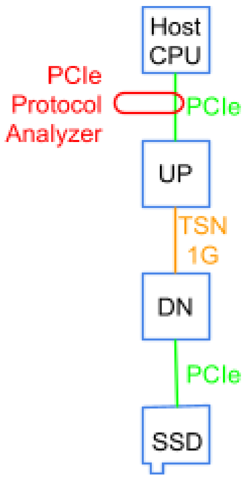
- Direct attached NVMe - this measurement is the latency from the Host CPU to the NVMe SSD (or, to be precise, from the PCIe protocol analyzer's test access point to the NVMe SSD DUT and back) and serves as a baseline which can be subtracted from the other measurements taken. The following block diagram shows this topology:



- NVMe tunneled over TCP/IP - this measurement is the latency from the Host CPU via the FPGA-based PCIe Upstream Switch Port via TCP/IP over 1 Gigabit Ethernet (without TSN) via the FPGA-based PCIe Down-Stream Switch Port to the NVMe SSD.



- NVMe/TSN - this measurement is the latency from the Host CPU via the FPGA-based PCIe Upstream Switch Port via TCP/IP and TSN over 1 Gigabit Ethernet, via the FPGA-based PCIe Down-Stream Switch Port to the NVMe SSD. By comparing this measurement against the second measurement we can identify the impact of the TSN IP Core on overall latency.

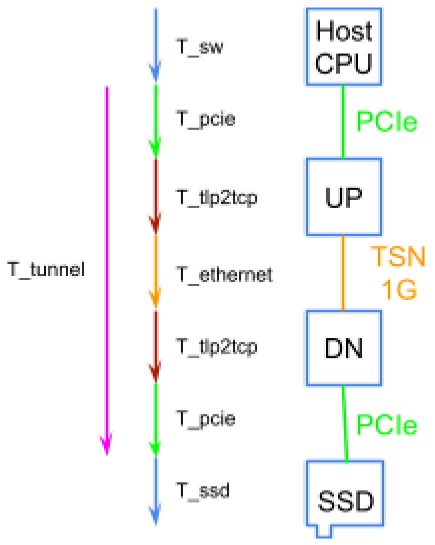


- Given by the concept of NVMe/TSN latency can be attributed to the different segments of connectivity. Thus, we can define the latency *T_tunnel* as the latency introduced by
- *T_pcie* = Transport delays inherent to PCIe

- *T_tlp2tcp* = Transport delays caused by tunneling (on-chip encapsulation and decapsulation and buffering)
- *T_ethernet* = Transport delays inherent to Ethernet

For *T_tunnel* we can ignore

- *T_sw* = userspace software and operating system processing times within the Host CPU as well as host memory access delays
- *T_ssd* = Storage data processing times within the NVMe SSD



Hence, *T_tunnel* = *T_sw* + *2* * *T_pcie* + *2* * *T_tlp2tcp* + *T_ethernet*.

Please note that *T_tunnel* is not the round-trip time but the “one-way” time from PCIe through encapsulation, network transport and decapsulation.

4. Latency of NVMe/TSN

To analyze the latency for NVMe/TSN we investigate two cases: The first case is more of a theoretical nature as it looks at the latency added to a typical PCIe TLP when tunneling over TCP/IP with and without TSN Ethernet. The second case is more of practical relevance as it shows the latency effects of tunneling for SSD read and write - which comprises many TLPs going back and forth between the Host CPU and the SSD. In either case we perform tests on all the three topologies described above.

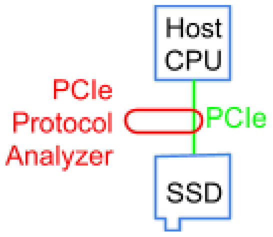
4.1. Latency for Config TLPs (T_tunnel)

Our first experiment performs Linux PCIe Configuration Space access. Because these TLPs do not involve the SSDs flash translation layer, this is a good indication for measuring the tunneling latency.

We perform this first experiment by capturing PCIe TLPs with the PCIe protocol analyzer and using the associated PCIe protocol analyzer software to measure latency differences in different setups. The precision of the measurement is specified as +/- 8 ns (nanoseconds) by the protocol analyzer manufacturer. All measurements are taken at the root port link of the Host CPU, i.e. between CPU and SSD or Host CPU and PCIe Upstream Switch Port respectively.

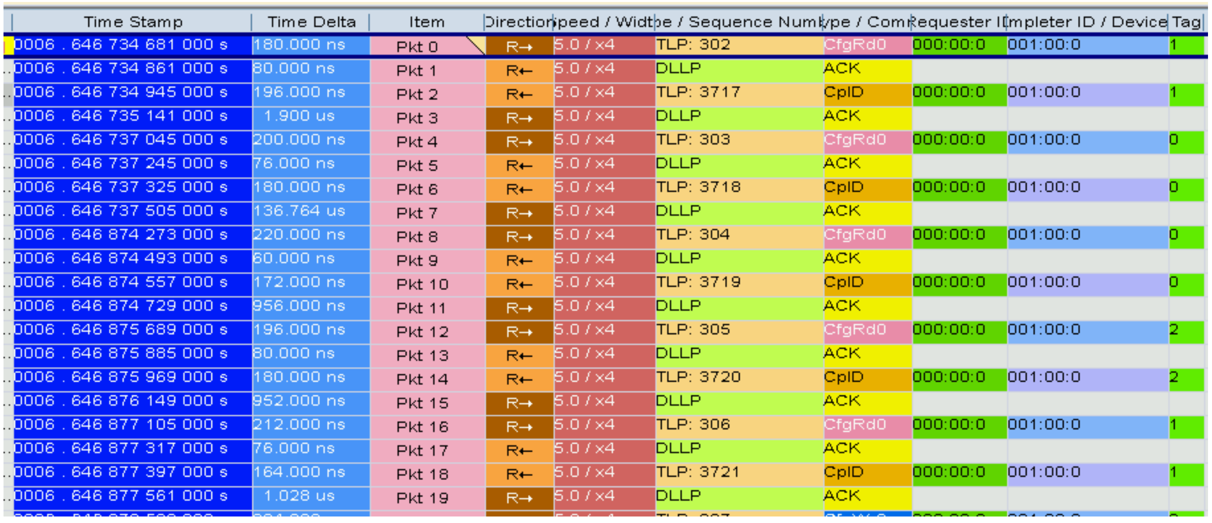
4.1.1. Latency for Direct Attached NVMe

This measurement is the latency from the Host CPU to the SSD (or, to be precise, from the PCIe protocol analyzer's test access point to the SSD and back) and serves as a baseline which can be subtracted from the other measurements taken.



Configuration TLPs are responded to in a quite deterministic way by the SSD and thus are useful for comparing latencies in different PCIe topologies. All PCIe links and components are in an idle condition when performing this measurement so other transfers don't influence the results.

We take the first five configuration read requests to get a small sample size average latency. The timestamps shown in the PCIe analyzer trace software reference the start of each TLP and the time delta specifies the distance from the end of TLP/DLLP to the start of the next TLP/DLLP. We will calculate the time from start of each configuration read request to the start of the corresponding completion with data. The time for the final ACK is outside of the tunneled path and not relevant for this discussion.

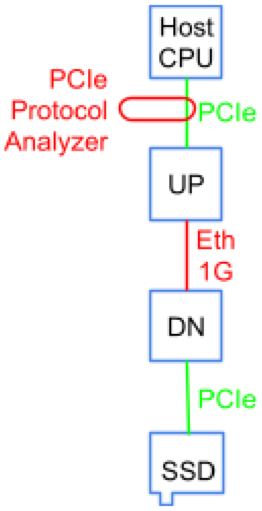


Time Stamp	Time Delta	Item	Direction	peed / Widt	pe / Sequence Num	ype / Com	equester I	mpleter ID / Device	Tag
0006 . 646 734 681 000 s	180.000 ns	Pkt 0	R→	5.0 / x4	TLP: 302	CfgRd0	000:00:0	001:00:0	1
0006 . 646 734 861 000 s	80.000 ns	Pkt 1	R←	5.0 / x4	DLLP	ACK			
0006 . 646 734 945 000 s	196.000 ns	Pkt 2	R←	5.0 / x4	TLP: 3717	CplD	000:00:0	001:00:0	1
0006 . 646 735 141 000 s	1.900 us	Pkt 3	R→	5.0 / x4	DLLP	ACK			
0006 . 646 737 045 000 s	200.000 ns	Pkt 4	R→	5.0 / x4	TLP: 303	CfgRd0	000:00:0	001:00:0	0
0006 . 646 737 245 000 s	76.000 ns	Pkt 5	R←	5.0 / x4	DLLP	ACK			
0006 . 646 737 325 000 s	180.000 ns	Pkt 6	R←	5.0 / x4	TLP: 3718	CplD	000:00:0	001:00:0	0
0006 . 646 737 505 000 s	136.764 us	Pkt 7	R→	5.0 / x4	DLLP	ACK			
0006 . 646 874 273 000 s	220.000 ns	Pkt 8	R→	5.0 / x4	TLP: 304	CfgRd0	000:00:0	001:00:0	0
0006 . 646 874 493 000 s	60.000 ns	Pkt 9	R←	5.0 / x4	DLLP	ACK			
0006 . 646 874 557 000 s	172.000 ns	Pkt 10	R←	5.0 / x4	TLP: 3719	CplD	000:00:0	001:00:0	0
0006 . 646 874 729 000 s	956.000 ns	Pkt 11	R→	5.0 / x4	DLLP	ACK			
0006 . 646 875 689 000 s	196.000 ns	Pkt 12	R→	5.0 / x4	TLP: 305	CfgRd0	000:00:0	001:00:0	2
0006 . 646 875 885 000 s	80.000 ns	Pkt 13	R←	5.0 / x4	DLLP	ACK			
0006 . 646 875 969 000 s	180.000 ns	Pkt 14	R←	5.0 / x4	TLP: 3720	CplD	000:00:0	001:00:0	2
0006 . 646 876 149 000 s	952.000 ns	Pkt 15	R→	5.0 / x4	DLLP	ACK			
0006 . 646 877 105 000 s	212.000 ns	Pkt 16	R→	5.0 / x4	TLP: 306	CfgRd0	000:00:0	001:00:0	1
0006 . 646 877 317 000 s	76.000 ns	Pkt 17	R←	5.0 / x4	DLLP	ACK			
0006 . 646 877 397 000 s	164.000 ns	Pkt 18	R←	5.0 / x4	TLP: 3721	CplD	000:00:0	001:00:0	1
0006 . 646 877 561 000 s	1.028 us	Pkt 19	R→	5.0 / x4	DLLP	ACK			

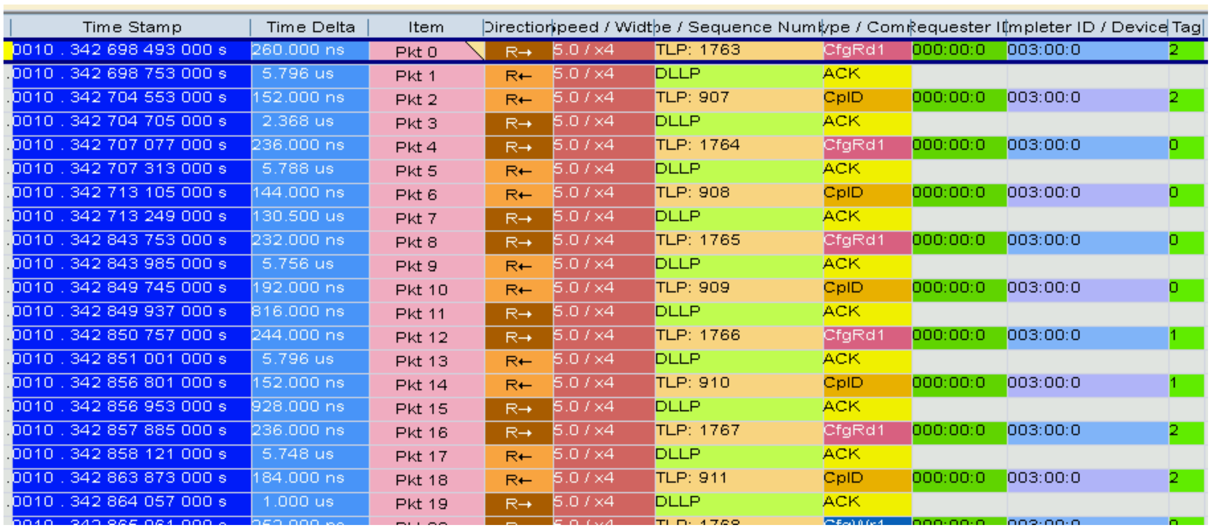
The above screenshot from the PCIe Protocol Analyzer shows the mentioned first configuration read requests and matching competitions with data for the directly connected SSD. By subtracting the two timestamps it can be seen that the latency measured from start of request to start of completion (as discussed above) is 264 ns, 280 ns, 284 ns, 280 ns, and 292 ns.

4.1.2. Latency for NVMe Tunneled over TCP/IP

This measurement is the latency from the Host CPU via the FPGA-based PCIe Upstream Switch Port via TCP/IP over 1 Gigabit Ethernet (without TSN) via the FPGA-based PCIe Down-Stream Switch Port to the NVMe SSD.



Below you find the screenshot of the PCIe Protocol Analyzer capturing this traffic:



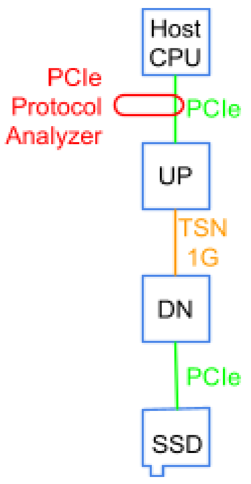
Time Stamp	Time Delta	Item	Direction	Speed / Width	Type / Sequence Number	Type / Command	Requester ID	Completer ID / Device	Tag
0010 . 342 698 493 000 s	260.000 ns	Pkt 0	R→	5.0 / x4	TLP: 1763	CfgRd1	000:00:0	003:00:0	2
0010 . 342 698 753 000 s	5.796 us	Pkt 1	R←	5.0 / x4	DLLP	ACK			
0010 . 342 704 553 000 s	152.000 ns	Pkt 2	R←	5.0 / x4	TLP: 907	CplD	000:00:0	003:00:0	2
0010 . 342 704 705 000 s	2.368 us	Pkt 3	R→	5.0 / x4	DLLP	ACK			
0010 . 342 707 077 000 s	236.000 ns	Pkt 4	R→	5.0 / x4	TLP: 1764	CfgRd1	000:00:0	003:00:0	0
0010 . 342 707 313 000 s	5.788 us	Pkt 5	R←	5.0 / x4	DLLP	ACK			
0010 . 342 713 105 000 s	144.000 ns	Pkt 6	R←	5.0 / x4	TLP: 908	CplD	000:00:0	003:00:0	0
0010 . 342 713 249 000 s	130.500 us	Pkt 7	R→	5.0 / x4	DLLP	ACK			
0010 . 342 843 753 000 s	232.000 ns	Pkt 8	R→	5.0 / x4	TLP: 1765	CfgRd1	000:00:0	003:00:0	0
0010 . 342 843 985 000 s	5.756 us	Pkt 9	R←	5.0 / x4	DLLP	ACK			
0010 . 342 849 745 000 s	192.000 ns	Pkt 10	R←	5.0 / x4	TLP: 909	CplD	000:00:0	003:00:0	0
0010 . 342 849 937 000 s	816.000 ns	Pkt 11	R→	5.0 / x4	DLLP	ACK			
0010 . 342 850 757 000 s	244.000 ns	Pkt 12	R→	5.0 / x4	TLP: 1766	CfgRd1	000:00:0	003:00:0	1
0010 . 342 851 001 000 s	5.796 us	Pkt 13	R←	5.0 / x4	DLLP	ACK			
0010 . 342 856 801 000 s	152.000 ns	Pkt 14	R←	5.0 / x4	TLP: 910	CplD	000:00:0	003:00:0	1
0010 . 342 856 953 000 s	928.000 ns	Pkt 15	R→	5.0 / x4	DLLP	ACK			
0010 . 342 857 885 000 s	236.000 ns	Pkt 16	R→	5.0 / x4	TLP: 1767	CfgRd1	000:00:0	003:00:0	2
0010 . 342 858 121 000 s	5.748 us	Pkt 17	R←	5.0 / x4	DLLP	ACK			
0010 . 342 863 873 000 s	184.000 ns	Pkt 18	R←	5.0 / x4	TLP: 911	CplD	000:00:0	003:00:0	2
0010 . 342 864 057 000 s	1.000 us	Pkt 19	R→	5.0 / x4	DLLP	ACK			

Now looking at the SSD connected to the TCP tunneled PCIe switch without TSN, the latencies increase to 6060 ns, 6030 ns, 5990 ns, 6040 ns, and 5990 ns, with an average of 6020 ns. While this transport back and forth is not symmetrical, a sufficiently precise estimate for T_tunnel is half of this round trip time. The time taken by the SSD to process the request is from the first measurement (Labcar 6.pre3), which results in

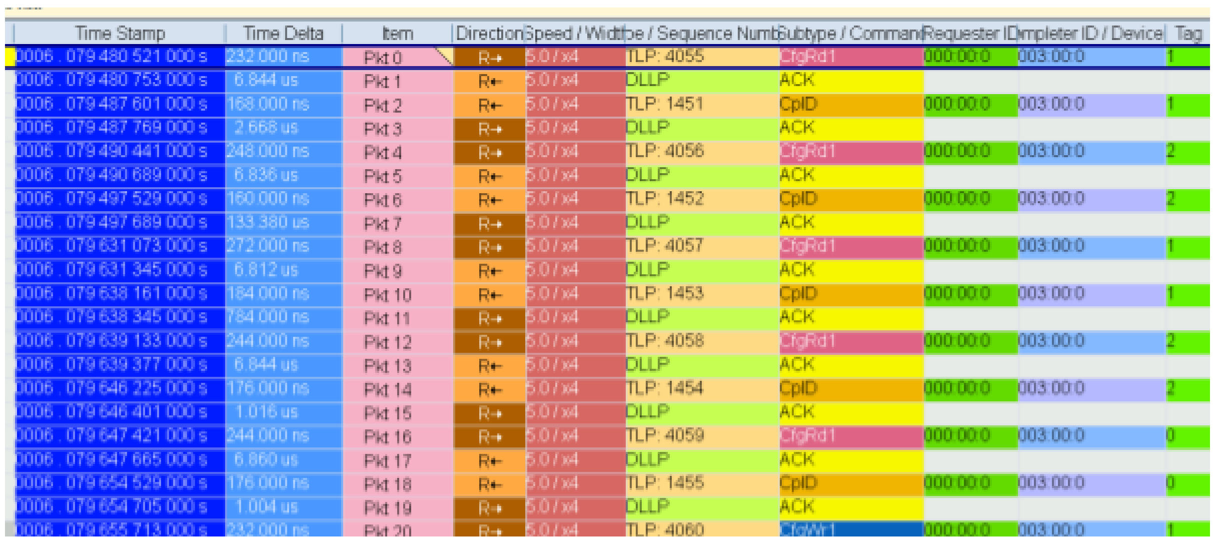
- *T_tunnel* = (6020 ns - 280 ns) / 2 = 2870 ns

4.1.3. Latency of NVMe/TSN

This measurement is the latency from the Host CPU via the FPGA-based PCIe Upstream Switch Port via TCP/IP and TSN over 1 Gigabit Ethernet, via the FPGA-based PCIe Down-Stream Switch Port to the NVMe SSD. By comparing this measurement against the second measurement we can identify the impact of TSN.



Below you find the screenshot of the PCIe Protocol Analyzer capturing this traffic:



Time Stamp	Time Delta	Item	Direction	Speed / Width	pe / Sequence Numb	Subtype / Comman	Requester ID	ompleter ID / Device	Tag
0006 . 079 480 521 000 s	232.000 ns	Pkt 0	R→	5.0 / x4	TLP: 4055	CfgRd1	000:00:0	003:00:0	1
0006 . 079 480 753 000 s	6.844 us	Pkt 1	R←	5.0 / x4	DLLP	ACK			
0006 . 079 487 601 000 s	168.000 ns	Pkt 2	R←	5.0 / x4	TLP: 1451	CplD	000:00:0	003:00:0	1
0006 . 079 487 769 000 s	2.668 us	Pkt 3	R→	5.0 / x4	DLLP	ACK			
0006 . 079 490 441 000 s	248.000 ns	Pkt 4	R→	5.0 / x4	TLP: 4056	CfgRd1	000:00:0	003:00:0	2
0006 . 079 490 689 000 s	6.836 us	Pkt 5	R←	5.0 / x4	DLLP	ACK			
0006 . 079 497 529 000 s	160.000 ns	Pkt 6	R←	5.0 / x4	TLP: 1452	CplD	000:00:0	003:00:0	2
0006 . 079 497 689 000 s	133.380 us	Pkt 7	R→	5.0 / x4	DLLP	ACK			
0006 . 079 631 073 000 s	272.000 ns	Pkt 8	R→	5.0 / x4	TLP: 4057	CfgRd1	000:00:0	003:00:0	1
0006 . 079 631 345 000 s	6.812 us	Pkt 9	R←	5.0 / x4	DLLP	ACK			
0006 . 079 638 161 000 s	184.000 ns	Pkt 10	R←	5.0 / x4	TLP: 1453	CplD	000:00:0	003:00:0	1
0006 . 079 638 345 000 s	784.000 ns	Pkt 11	R→	5.0 / x4	DLLP	ACK			
0006 . 079 639 133 000 s	244.000 ns	Pkt 12	R→	5.0 / x4	TLP: 4058	CfgRd1	000:00:0	003:00:0	2
0006 . 079 639 377 000 s	6.844 us	Pkt 13	R←	5.0 / x4	DLLP	ACK			
0006 . 079 646 225 000 s	176.000 ns	Pkt 14	R←	5.0 / x4	TLP: 1454	CplD	000:00:0	003:00:0	2
0006 . 079 646 401 000 s	1.016 us	Pkt 15	R→	5.0 / x4	DLLP	ACK			
0006 . 079 647 421 000 s	244.000 ns	Pkt 16	R→	5.0 / x4	TLP: 4059	CfgRd1	000:00:0	003:00:0	0
0006 . 079 647 665 000 s	6.860 us	Pkt 17	R←	5.0 / x4	DLLP	ACK			
0006 . 079 654 529 000 s	176.000 ns	Pkt 18	R←	5.0 / x4	TLP: 1455	CplD	000:00:0	003:00:0	0
0006 . 079 654 705 000 s	1.004 us	Pkt 19	R→	5.0 / x4	DLLP	ACK			
0006 . 079 655 713 000 s	232.000 ns	Pkt 20	R→	5.0 / x4	TLP: 4060	CfgWr1	000:00:0	003:00:0	1

Latency slightly increases to 7080 ns, 7088 ns, 7088 ns, 7092 ns and 7108 ns, which averages to 7091 ns.

Again, while this transport back and forth is not symmetrical, a sufficiently precise estimate for *T_tunnel* is half of this round trip time. The time taken by the SSD to process the request is from the first measurement which results in

- *T_tunnel* = (7091 ns - 280 ns) / 2 = 3405 ns

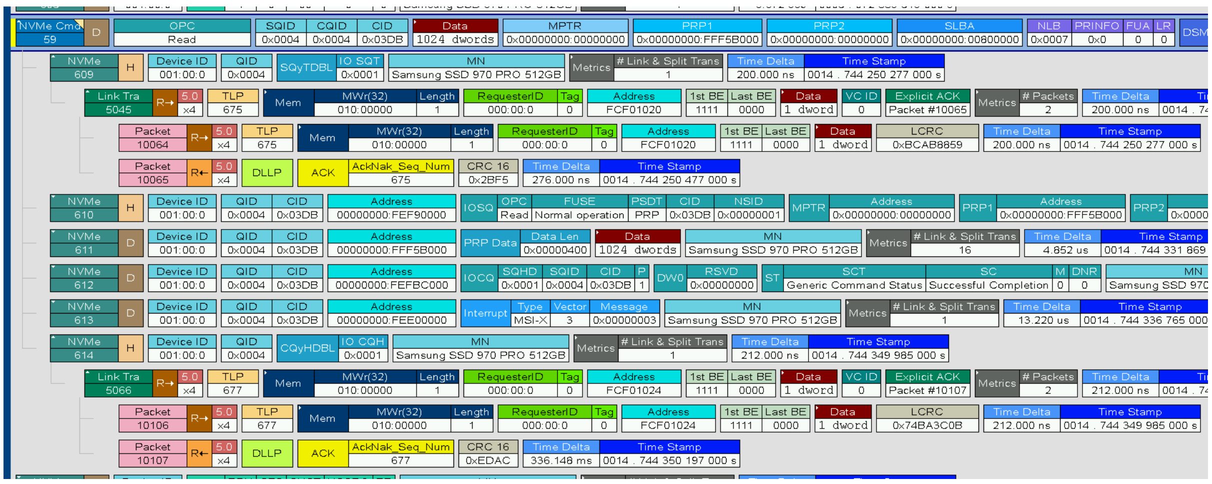
We contribute this additional latency when tunneling via TSN to additional buffer (FIFO) stages within the Fraunhofer IPMS TSN IP Core.

4.2. NVMe Application Level Latency - SSD Read

To measure application-level latencies, the Linux “dd” program is used to read single 4 KiB blocks in an otherwise idle system. The test setup is identical to the previous PCIe Configuration Access tests. Please note that a large portion of the execution time is associated with SSD internal flash access and processing and thus may vary largely regardless of the connection latency. This consumer grade SSD does not make latency guarantees.

4.2.1. Latency for Direct Attached SSD

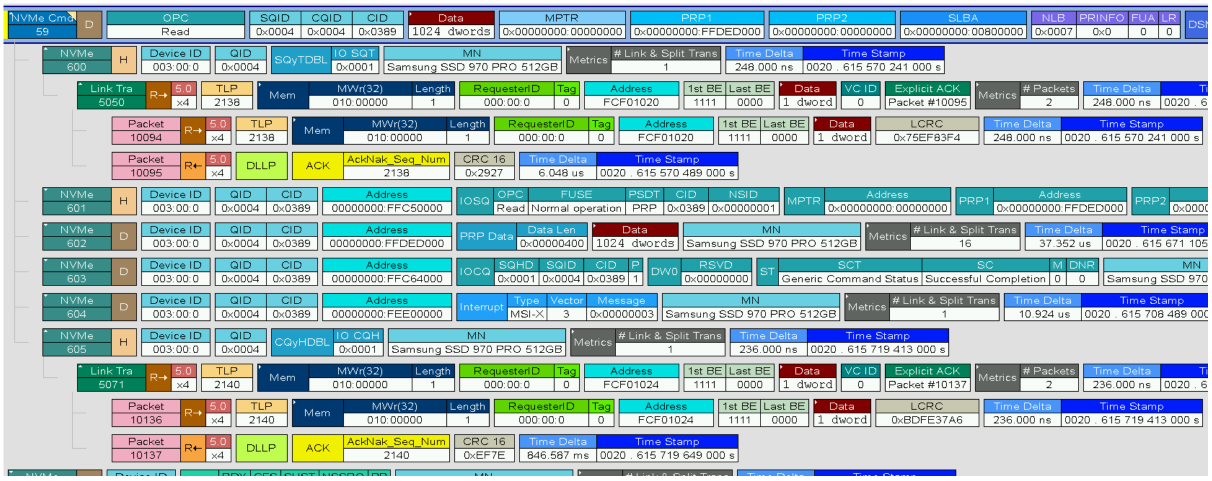
The following screenshot shows the total execution time for a 4 KiB read from the SSD device directly attached to the Host CPU. The time it takes for the completion queue head doorbell write can not be measured, but since writes are posted the host assumes the IO command is complete as soon as this write TLP is issued. The difference from the start of the submission queue tail doorbell update to the DLLP ACK of the completion queue head doorbell update, and thus the IO command execution time, is 99.92 us.



```
NVMe Cmd 59 | D | OPC: Read | SQID: 0x0004 | CQID: 0x0004 | CID: 0x03DB | Data: 1024 dwords | MPTR: 0x00000000:00000000 | PRP1: 0x00000000:FFF5B000 | PRP2: 0x00000000:00000000 | SLBA: 0x00000000:00800000 | NLB: 0x0007 | PRINFO: 0x0 | FUA: 0 | LR: 0 | DSM
  NVMe 609 | H | Device ID: 001:00:0 | QID: 0x0004 | SQyTDBL | IO SQT: 0x0001 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 200.000 ns | Time Stamp: 0014 . 744 250 277 000 s
    Link Tra 5045 | R→ | 5.0 x4 | TLP: 675 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF01020 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | VC ID: 0 | Explicit ACK: Packet #10065 | Metrics | # Packets: 2 | Time Delta: 200.000 ns | Ti 0014 . 74
      Packet 10064 | R→ | 5.0 x4 | TLP: 675 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF01020 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | LCRC: 0xBCAB8859 | Time Delta: 200.000 ns | Time Stamp: 0014 . 744 250 277 000 s
      Packet 10065 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num: 675 | CRC 16: 0x2BF5 | Time Delta: 276.000 ns | Time Stamp: 0014 . 744 250 477 000 s
  NVMe 610 | H | Device ID: 001:00:0 | QID: 0x0004 | CID: 0x03DB | Address: 00000000:FEF90000 | IOSQ | OPC: Read | FUSE: Normal operation | PSDT: PRP | CID: 0x03DB | NSID: 0x00000001 | MPTR | Address: 0x00000000:00000000 | PRP1 | Address: 0x00000000:FFF5B000 | PRP2 | 0x0000
  NVMe 611 | D | Device ID: 001:00:0 | QID: 0x0004 | CID: 0x03DB | Address: 00000000:FFF5B000 | PRP Data | Data Len: 0x00000400 | Data: 1024 dwords | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 16 | Time Delta: 4.852 us | Time Stamp: 0014 . 744 331 869
  NVMe 612 | D | Device ID: 001:00:0 | QID: 0x0004 | CID: 0x03DB | Address: 00000000:FEFBC000 | IOCQ | SQHD: 0x0001 | SQID: 0x0004 | CID: 0x03DB | P: 1 | DW0 | RSVD: 0x00000000 | ST | SCT: Generic Command Status | SC: Successful Completion | M: 0 | DNR: 0 | MN: Samsung SSD 970
  NVMe 613 | D | Device ID: 001:00:0 | QID: 0x0004 | CID: 0x03DB | Address: 00000000:FEE00000 | Interrupt | Type: MSI-X | Vector: 3 | Message: 0x00000003 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 13.220 us | Time Stamp: 0014 . 744 336 765 000
  NVMe 614 | H | Device ID: 001:00:0 | QID: 0x0004 | CQyHDBL | IO CQH: 0x0001 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 212.000 ns | Time Stamp: 0014 . 744 349 985 000 s
    Link Tra 5066 | R→ | 5.0 x4 | TLP: 677 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF01024 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | VC ID: 0 | Explicit ACK: Packet #10107 | Metrics | # Packets: 2 | Time Delta: 212.000 ns | Ti 0014 . 74
      Packet 10106 | R→ | 5.0 x4 | TLP: 677 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF01024 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | LCRC: 0x74BA3C0B | Time Delta: 212.000 ns | Time Stamp: 0014 . 744 349 985 000 s
      Packet 10107 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num: 677 | CRC 16: 0xEDAC | Time Delta: 336.148 ms | Time Stamp: 0014 . 744 350 197 000 s
```

4.2.2. Latency for NVMe Tunneled over TCP/IP

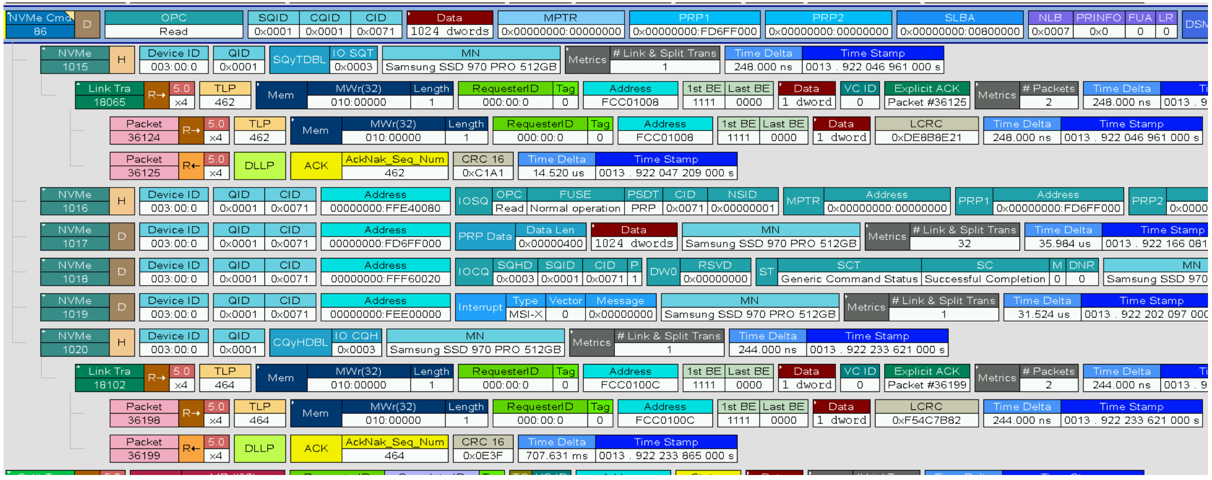
The following screenshot shows the latency for the NVMe tunneled over TCP/IP (without TSN). The measured latency for the 4 KiB read is 149.41 us.



```
NVMe Cmd 59 | D | OPC Read | SQID 0x0004 | CQID 0x0004 | CID 0x0389 | Data 1024 dwords | MPTR 0x00000000:00000000 | PRP1 0x00000000:FFDED000 | PRP2 0x00000000:00000000 | SLBA 0x00000000:00800000 | NLB 0x0007 | PRINFO 0x0 | FUA 0 | LR 0 | DSM
  NVMe 600 | H | Device ID 003:00:0 | QID 0x0004 | SQyTDBL | IO SQT 0x0001 | MN Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans 1 | Time Delta 248.000 ns | Time Stamp 0020 . 615 570 241 000 s
    Link Tra 5050 | R→ | 5.0 x4 | TLP 2138 | Mem | MWr(32) 010:00000 | Length 1 | RequesterID 000:00:0 | Tag 0 | Address FCF01020 | 1st BE 1111 | Last BE 0000 | Data 1 dword | VC ID 0 | Explicit ACK Packet #10095 | Metrics | # Packets 2 | Time Delta 248.000 ns | Ti 0020 . 6
      Packet 10094 | R→ | 5.0 x4 | TLP 2138 | Mem | MWr(32) 010:00000 | Length 1 | RequesterID 000:00:0 | Tag 0 | Address FCF01020 | 1st BE 1111 | Last BE 0000 | Data 1 dword | LCRC 0x75EF83F4 | Time Delta 248.000 ns | Time Stamp 0020 . 615 570 241 000 s
      Packet 10095 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num 2138 | CRC 16 0x2927 | Time Delta 6.048 us | Time Stamp 0020 . 615 570 489 000 s
  NVMe 601 | H | Device ID 003:00:0 | QID 0x0004 | CID 0x0389 | Address 00000000:FFC50000 | IOSQ | OPC Read | FUSE Normal operation | PSDT PRP | CID 0x0389 | NSID 0x00000001 | MPTR | Address 0x00000000:00000000 | PRP1 | Address 0x00000000:FFDED000 | PRP2 | 0x000
  NVMe 602 | D | Device ID 003:00:0 | QID 0x0004 | CID 0x0389 | Address 00000000:FFDED000 | PRP Data | Data Len 0x00000400 | Data 1024 dwords | MN Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans 16 | Time Delta 37.352 us | Time Stamp 0020 . 615 671 105
  NVMe 603 | D | Device ID 003:00:0 | QID 0x0004 | CID 0x0389 | Address 00000000:FFC64000 | IOCQ | SQHD 0x0001 | SQID 0x0004 | CID 0x0389 | P 1 | DW0 | RSVD 0x00000000 | ST | SCT Generic Command Status | SC Successful Completion | M 0 | DNR 0 | MN Samsung SSD 970
  NVMe 604 | D | Device ID 003:00:0 | QID 0x0004 | CID 0x0389 | Address 00000000:FEE00000 | Interrupt | Type MSI-X | Vector 3 | Message 0x00000003 | MN Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans 1 | Time Delta 10.924 us | Time Stamp 0020 . 615 708 489 000
  NVMe 605 | H | Device ID 003:00:0 | QID 0x0004 | CQyHDBL | IO CQH 0x0001 | MN Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans 1 | Time Delta 236.000 ns | Time Stamp 0020 . 615 719 413 000 s
    Link Tra 5071 | R→ | 5.0 x4 | TLP 2140 | Mem | MWr(32) 010:00000 | Length 1 | RequesterID 000:00:0 | Tag 0 | Address FCF01024 | 1st BE 1111 | Last BE 0000 | Data 1 dword | VC ID 0 | Explicit ACK Packet #10137 | Metrics | # Packets 2 | Time Delta 236.000 ns | Ti 0020 . 6
      Packet 10136 | R→ | 5.0 x4 | TLP 2140 | Mem | MWr(32) 010:00000 | Length 1 | RequesterID 000:00:0 | Tag 0 | Address FCF01024 | 1st BE 1111 | Last BE 0000 | Data 1 dword | LCRC 0xBDFE37A6 | Time Delta 236.000 ns | Time Stamp 0020 . 615 719 413 000 s
      Packet 10137 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num 2140 | CRC 16 0xEF7E | Time Delta 846.587 ms | Time Stamp 0020 . 615 719 649 000 s
```

4.2.3. Latency for NVMe/TSN

The following screenshot shows the latency for NVMe/TSN. Adding TSN to the tunneled PCIe switch the latency increases to 186.90 us:



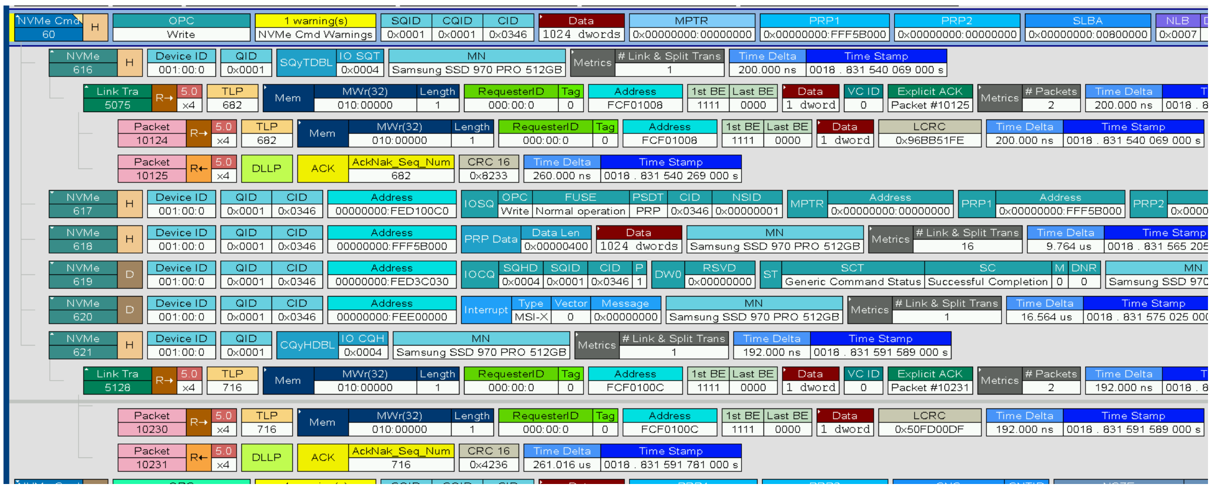
```
NVMe Cmd 86 | D | OPC Read | SQID 0x0001 | CQID 0x0001 | CID 0x0071 | Data 1024 dwords | MPTR 0x00000000:00000000 | PRP1 0x00000000:FD6FF000 | PRP2 0x00000000:00000000 | SLBA 0x00000000:00800000 | NLB 0x0007 | PRINFO 0x0 | FUA 0 | LR 0 | DSM
  NVMe 1015 | H | Device ID 003:00:0 | QID 0x0001 | SQyTDBL | IO SQT 0x0003 | MN Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans 1 | Time Delta 248.000 ns | Time Stamp 0013 . 922 046 961 000 s
    Link Tra 18065 | R→ | 5.0 x4 | TLP 462 | Mem | MWr(32) 010:00000 | Length 1 | RequesterID 000:00:0 | Tag 0 | Address FCC01008 | 1st BE 1111 | Last BE 0000 | Data 1 dword | VC ID 0 | Explicit ACK Packet #36125 | Metrics | # Packets 2 | Time Delta 248.000 ns | Ti 0013 . 9
      Packet 36124 | R→ | 5.0 x4 | TLP 462 | Mem | MWr(32) 010:00000 | Length 1 | RequesterID 000:00:0 | Tag 0 | Address FCC01008 | 1st BE 1111 | Last BE 0000 | Data 1 dword | LCRC 0xDE8B8E21 | Time Delta 248.000 ns | Time Stamp 0013 . 922 046 961 000 s
      Packet 36125 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num 462 | CRC 16 0xC1A1 | Time Delta 14.520 us | Time Stamp 0013 . 922 047 209 000 s
  NVMe 1016 | H | Device ID 003:00:0 | QID 0x0001 | CID 0x0071 | Address 00000000:FFE40080 | IOSQ | OPC Read | FUSE Normal operation | PSDT PRP | CID 0x0071 | NSID 0x00000001 | MPTR | Address 0x00000000:00000000 | PRP1 | Address 0x00000000:FD6FF000 | PRP2 | 0x000
  NVMe 1017 | D | Device ID 003:00:0 | QID 0x0001 | CID 0x0071 | Address 00000000:FD6FF000 | PRP Data | Data Len 0x00000400 | Data 1024 dwords | MN Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans 32 | Time Delta 35.984 us | Time Stamp 0013 . 922 166 081
  NVMe 1018 | D | Device ID 003:00:0 | QID 0x0001 | CID 0x0071 | Address 00000000:FFF60020 | IOCQ | SQHD 0x0003 | SQID 0x0001 | CID 0x0071 | P 1 | DW0 | RSVD 0x00000000 | ST | SCT Generic Command Status | SC Successful Completion | M 0 | DNR 0 | MN Samsung SSD 970
  NVMe 1019 | D | Device ID 003:00:0 | QID 0x0001 | CID 0x0071 | Address 00000000:FEE00000 | Interrupt | Type MSI-X | Vector 0 | Message 0x00000000 | MN Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans 1 | Time Delta 31.524 us | Time Stamp 0013 . 922 202 097 000
  NVMe 1020 | H | Device ID 003:00:0 | QID 0x0001 | CQyHDBL | IO CQH 0x0003 | MN Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans 1 | Time Delta 244.000 ns | Time Stamp 0013 . 922 233 621 000 s
    Link Tra 18102 | R→ | 5.0 x4 | TLP 464 | Mem | MWr(32) 010:00000 | Length 1 | RequesterID 000:00:0 | Tag 0 | Address FCC0100C | 1st BE 1111 | Last BE 0000 | Data 1 dword | VC ID 0 | Explicit ACK Packet #36199 | Metrics | # Packets 2 | Time Delta 244.000 ns | Ti 0013 . 9
      Packet 36198 | R→ | 5.0 x4 | TLP 464 | Mem | MWr(32) 010:00000 | Length 1 | RequesterID 000:00:0 | Tag 0 | Address FCC0100C | 1st BE 1111 | Last BE 0000 | Data 1 dword | LCRC 0xF54C7B82 | Time Delta 244.000 ns | Time Stamp 0013 . 922 233 621 000 s
      Packet 36199 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num 464 | CRC 16 0x0E3F | Time Delta 707.631 ms | Time Stamp 0013 . 922 233 865 000 s
```

4.3. NVMe Application Level Latency - SSD Write

Next we look at write latencies for a 4 KiB write from host to device. Latencies inside the SSD are generally lower because the SSD does not need to access flash and can store this write in its cache. The PCIe communication is expected to have slightly higher latency because the necessary DMA read requests are non-posted.

4.3.1. Latency for Direct Attached SSD

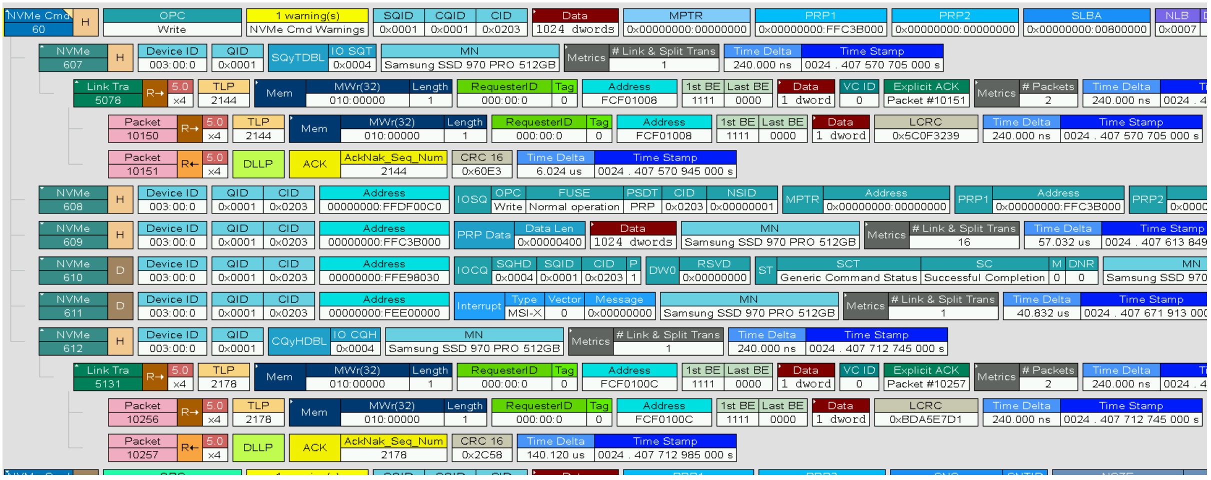
The following screenshot is for the directly-attached SSD. The latency is 51.71 us.



```
NVMe Cmd 60 | H | OPC: Write | 1 warning(s): NVMe Cmd Warnings | SQID: 0x0001 | CQID: 0x0001 | CID: 0x0346 | Data: 1024 dwords | MPTR: 0x00000000:00000000 | PRP1: 0x00000000:FFF5B000 | PRP2: 0x00000000:00000000 | SLBA: 0x00000000:00800000 | NLB: 0x0007
  NVMe 616 | H | Device ID: 001:00:0 | QID: 0x0001 | SQyTDBL | IO SQT: 0x0004 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 200.000 ns | Time Stamp: 0018 . 831 540 069 000 s
    Link Tra 5075 | R→ | 5.0 x4 | TLP: 682 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF01008 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | VC ID: 0 | Explicit ACK: Packet #10125 | Metrics | # Packets: 2 | Time Delta: 200.000 ns | T: 0018 . 8
      Packet 10124 | R→ | 5.0 x4 | TLP: 682 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF01008 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | LCRC: 0x96BB51FE | Time Delta: 200.000 ns | Time Stamp: 0018 . 831 540 069 000 s
      Packet 10125 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num: 682 | CRC 16: 0x8233 | Time Delta: 260.000 ns | Time Stamp: 0018 . 831 540 269 000 s
  NVMe 617 | H | Device ID: 001:00:0 | QID: 0x0001 | CID: 0x0346 | Address: 00000000:FED100C0 | IOSQ | OPC: Write | FUSE: Normal operation | PSDT: PRP | CID: 0x0346 | NSID: 0x00000001 | MPTR | Address: 0x00000000:00000000 | PRP1 | Address: 0x00000000:FFF5B000 | PRP2 | 0x0000
  NVMe 618 | H | Device ID: 001:00:0 | QID: 0x0001 | CID: 0x0346 | Address: 00000000:FFF5B000 | PRP Data | Data Len: 0x00000400 | Data: 1024 dwords | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 16 | Time Delta: 9.764 us | Time Stamp: 0018 . 831 565 205
  NVMe 619 | D | Device ID: 001:00:0 | QID: 0x0001 | CID: 0x0346 | Address: 00000000:FED3C030 | IOCQ | SQHD: 0x0004 | SQID: 0x0001 | CID: 0x0346 | P: 1 | DW0 | RSVD: 0x00000000 | ST | SCT: Generic Command Status | SC: Successful Completion | M: 0 | DNR: 0 | MN: Samsung SSD 970
  NVMe 620 | D | Device ID: 001:00:0 | QID: 0x0001 | CID: 0x0346 | Address: 00000000:FEE00000 | Interrupt | Type: MSI-X | Vector: 0 | Message: 0x00000000 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 16.564 us | Time Stamp: 0018 . 831 575 025 000
  NVMe 621 | H | Device ID: 001:00:0 | QID: 0x0001 | CQyHDBL | IO CQH: 0x0004 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 192.000 ns | Time Stamp: 0018 . 831 591 589 000 s
    Link Tra 5128 | R→ | 5.0 x4 | TLP: 716 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF0100C | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | VC ID: 0 | Explicit ACK: Packet #10231 | Metrics | # Packets: 2 | Time Delta: 192.000 ns | T: 0018 . 8
      Packet 10230 | R→ | 5.0 x4 | TLP: 716 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF0100C | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | LCRC: 0x50FD00DF | Time Delta: 192.000 ns | Time Stamp: 0018 . 831 591 589 000 s
      Packet 10231 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num: 716 | CRC 16: 0x4236 | Time Delta: 261.016 us | Time Stamp: 0018 . 831 591 781 000 s
```

4.3.2. Latency for NVMe Tunneled over TCP/IP

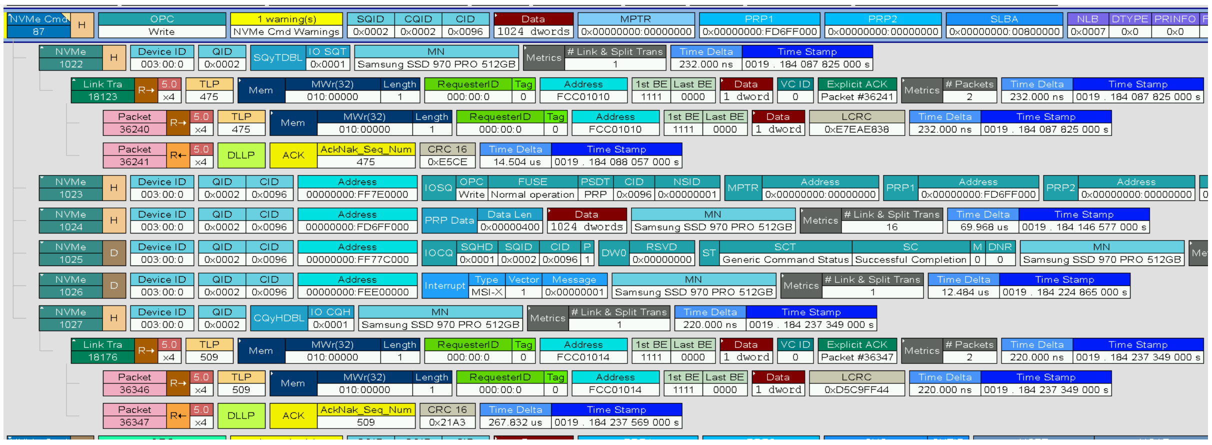
The following screenshot is for the NVMe tunneled over TCP/IP (without TSN). The latency is 142.28 us.



```
NVMe Cmd 60 | H | OPC: Write | 1 warning(s): NVMe Cmd Warnings | SQID: 0x0001 | CQID: 0x0001 | CID: 0x0203 | Data: 1024 dwords | MPTR: 0x00000000:00000000 | PRP1: 0x00000000:FFC3B000 | PRP2: 0x00000000:00000000 | SLBA: 0x00000000:00800000 | NLB: 0x0007
  NVMe 607 | H | Device ID: 003:00:0 | QID: 0x0001 | SQyTDBL | IO SQT: 0x0004 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 240.000 ns | Time Stamp: 0024 . 407 570 705 000 s
    Link Tra 5078 | R→ | 5.0 x4 | TLP: 2144 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF01008 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | VC ID: 0 | Explicit ACK: Packet #10151 | Metrics | # Packets: 2 | Time Delta: 240.000 ns | T: 0024 . 4
      Packet 10150 | R→ | 5.0 x4 | TLP: 2144 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF01008 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | LCRC: 0x5C0F3239 | Time Delta: 240.000 ns | Time Stamp: 0024 . 407 570 705 000 s
      Packet 10151 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num: 2144 | CRC 16: 0x60E3 | Time Delta: 6.024 us | Time Stamp: 0024 . 407 570 945 000 s
  NVMe 608 | H | Device ID: 003:00:0 | QID: 0x0001 | CID: 0x0203 | Address: 00000000:FFDF00C0 | IOSQ | OPC: Write | FUSE: Normal operation | PSDT: PRP | CID: 0x0203 | NSID: 0x00000001 | MPTR | Address: 0x00000000:00000000 | PRP1 | Address: 0x00000000:FFC3B000 | PRP2 | 0x0000
  NVMe 609 | H | Device ID: 003:00:0 | QID: 0x0001 | CID: 0x0203 | Address: 00000000:FFC3B000 | PRP Data | Data Len: 0x00000400 | Data: 1024 dwords | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 16 | Time Delta: 57.032 us | Time Stamp: 0024 . 407 613 849
  NVMe 610 | D | Device ID: 003:00:0 | QID: 0x0001 | CID: 0x0203 | Address: 00000000:FFE98030 | IOCQ | SQHD: 0x0004 | SQID: 0x0001 | CID: 0x0203 | P: 1 | DW0 | RSVD: 0x00000000 | ST | SCT: Generic Command Status | SC: Successful Completion | M: 0 | DNR: 0 | MN: Samsung SSD 970
  NVMe 611 | D | Device ID: 003:00:0 | QID: 0x0001 | CID: 0x0203 | Address: 00000000:FEE00000 | Interrupt | Type: MSI-X | Vector: 0 | Message: 0x00000000 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 40.832 us | Time Stamp: 0024 . 407 671 913 000
  NVMe 612 | H | Device ID: 003:00:0 | QID: 0x0001 | CQyHDBL | IO CQH: 0x0004 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 240.000 ns | Time Stamp: 0024 . 407 712 745 000 s
    Link Tra 5131 | R→ | 5.0 x4 | TLP: 2178 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF0100C | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | VC ID: 0 | Explicit ACK: Packet #10257 | Metrics | # Packets: 2 | Time Delta: 240.000 ns | T: 0024 . 4
      Packet 10256 | R→ | 5.0 x4 | TLP: 2178 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCF0100C | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | LCRC: 0xBDA5E7D1 | Time Delta: 240.000 ns | Time Stamp: 0024 . 407 712 745 000 s
      Packet 10257 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num: 2178 | CRC 16: 0x2C58 | Time Delta: 140.120 us | Time Stamp: 0024 . 407 712 985 000 s
```

4.3.3. Latency for NVMe/TSN

The following screenshot is for the NVMe/TSN. The latency is 149.74 us.



```
NVMe Cmd 87 | H | OPC: Write | 1 warning(s): NVMe Cmd Warnings | SQID: 0x0002 | CQID: 0x0002 | CID: 0x0096 | Data: 1024 dwords | MPTR: 0x00000000:00000000 | PRP1: 0x00000000:FD6FF000 | PRP2: 0x00000000:00000000 | SLBA: 0x00000000:00800000 | NLB: 0x0007 | DTYPE: 0x0 | PRINFO: 0x0
  NVMe 1022 | H | Device ID: 003:00:0 | QID: 0x0002 | SQyTDBL | IO SQT: 0x0001 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 232.000 ns | Time Stamp: 0019 . 184 087 825 000 s
    Link Tra 18123 | R→ | 5.0 x4 | TLP: 475 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCC01010 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | VC ID: 0 | Explicit ACK: Packet #36241 | Metrics | # Packets: 2 | Time Delta: 232.000 ns | Time Stamp: 0019 . 184 087 825 000 s
      Packet 36240 | R→ | 5.0 x4 | TLP: 475 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCC01010 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | LCRC: 0xE7EAE838 | Time Delta: 232.000 ns | Time Stamp: 0019 . 184 087 825 000 s
      Packet 36241 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num: 475 | CRC 16: 0xE5CE | Time Delta: 14.504 us | Time Stamp: 0019 . 184 088 057 000 s
  NVMe 1023 | H | Device ID: 003:00:0 | QID: 0x0002 | CID: 0x0096 | Address: 00000000:FF7E0000 | IOSQ | OPC: Write | FUSE: Normal operation | PSDT: PRP | CID: 0x0096 | NSID: 0x00000001 | MPTR | Address: 0x00000000:00000000 | PRP1 | Address: 0x00000000:FD6FF000 | PRP2 | Address: 0x00000000:00000000
  NVMe 1024 | H | Device ID: 003:00:0 | QID: 0x0002 | CID: 0x0096 | Address: 00000000:FD6FF000 | PRP Data | Data Len: 0x00000400 | Data: 1024 dwords | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 16 | Time Delta: 69.968 us | Time Stamp: 0019 . 184 146 577 000 s
  NVMe 1025 | D | Device ID: 003:00:0 | QID: 0x0002 | CID: 0x0096 | Address: 00000000:FF77C000 | IOCQ | SQHD: 0x0001 | SQID: 0x0002 | CID: 0x0096 | P: 1 | DW0 | RSVD: 0x00000000 | ST | SCT: Generic Command Status | SC: Successful Completion | M: 0 | DNR: 0 | MN: Samsung SSD 970 PRO 512GB
  NVMe 1026 | D | Device ID: 003:00:0 | QID: 0x0002 | CID: 0x0096 | Address: 00000000:FEE00000 | Interrupt | Type: MSI-X | Vector: 1 | Message: 0x00000001 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 12.484 us | Time Stamp: 0019 . 184 224 865 000 s
  NVMe 1027 | H | Device ID: 003:00:0 | QID: 0x0002 | CQyHDBL | IO CQH: 0x0001 | MN: Samsung SSD 970 PRO 512GB | Metrics | # Link & Split Trans: 1 | Time Delta: 220.000 ns | Time Stamp: 0019 . 184 237 349 000 s
    Link Tra 18176 | R→ | 5.0 x4 | TLP: 509 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCC01014 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | VC ID: 0 | Explicit ACK: Packet #36347 | Metrics | # Packets: 2 | Time Delta: 220.000 ns | Time Stamp: 0019 . 184 237 349 000 s
      Packet 36346 | R→ | 5.0 x4 | TLP: 509 | Mem | MWr(32): 010:00000 | Length: 1 | RequesterID: 000:00:0 | Tag: 0 | Address: FCC01014 | 1st BE: 1111 | Last BE: 0000 | Data: 1 dword | LCRC: 0xD5C9FF44 | Time Delta: 220.000 ns | Time Stamp: 0019 . 184 237 349 000 s
      Packet 36347 | R← | 5.0 x4 | DLLP | ACK | AckNak_Seq_Num: 509 | CRC 16: 0x21A3 | Time Delta: 267.832 us | Time Stamp: 0019 . 184 237 569 000 s
```

4.4. Implications of Application-Level Latency

The following table reviews the results of the latency comparison between Config TLPs for a) the direct attached SSD, b) a networked SSD with NVMe tunneled over TCP/IP, and c) a networked SSD with NVMe tunneled over TCP/IP and over TSN:

Measurement	**NVMe with Direct Attached SSD**	**SSD with NVMe Tunneled over TCP/IP**	**SSD with NVMe/TSN**
1	280 ns	6060 ns	7080 ns
2	284 ns	6030 ns	7088 ns
3	280 ns	5990 ns	7092 ns
4	290 ns	6040 ns	7108 ns

This shows that the extra round trip latency introduced by TSN is approximately 1 microsecond. We believe that this is very acceptable given the many benefits of TSN, such as traffic shaping and the real-time system capabilities for a distributed system.

For the application-level NVMe read and write test of 4 KiB of data the increase in latency is much lower than for a single TLP, as the NVMe SSD can exploit parallelism in TLP transfers.

Measurement	NVMe with Direct Attached SSD	SSD with NVMe Tunneled over TCP/IP	SSD with NVMe/TSN
4 KiB NVMe Read	100 us	149 us	187 us
4 KiB NVMe Write	52 us	142 us	150 us

Also, please keep in mind that due to the queuing nature of NVMe multiple Read Requests and/or multiple Write Requests typically overlap in time. That means that at application-level the extra latency introduced by TSN will be much less visible for typical read and write loads.

5. Conclusion & Backgrounder

Our experimental analysis shows that NVMe/TSN comes with reasonable additional TLP latency, 3.4 microseconds, which is good enough for many real-time systems, including Automotive Zone-Based Architectures.

It demonstrates that NPAP, the TCP/IP Full Accelerator technology from Fraunhofer HHI is very low latency (something you would expect from a technology once designed and optimized for low-latency financial trading).

And, a comparison between tunneling without and with TSN also shows that extra latency is around 0.5 microseconds, which seems well worth given the many advantages of TSN.

For further information regarding PCIe Range Extension over TCP/IP over TSN over 1/10/25/50/100G Ethernet please refer to our public technology presentations including:

- Jim Peek, Dir. of Technology MLE: "PCIe Range Extension via Robust, Long Reach Protocol Tunnels", PCI-SIG Developers Conference 2018 https://www.missinglinkelectronics.com/files/papers/04_10_PCIe_Range_Extension_via_Robust_Long_Reach_Protocol_Tunnels_UL.pdf
- Endric Schubert, CTO MLE: "Sensor Fusion and Data-in-Motion Processing for Autonomous Vehicles", PCI-SIG Developers Conference https://www.missinglinkelectronics.com/files/papers/04_04_Sensor-Fusion-and-Data-in-Motion-Processing-for-Autonomous-Vehicle.pdf
- Endric Schubert, CTO MLE: "PCIe-over-TCP-over-TSN-over10/25GigE", 4th Workshop Programmable Processing for the Autonomous / Connected Vehicle 2020 https://www.missinglinkelectronics.com/files/papers/MLE-FPGA4ADAS-20200924.pdf
- Endric Schubert, CTO MLE: "Zone-Based Automotive Backbones Tunneling PCIe®", PCI-SIG Virtual Developers Conference 2021 https://www.missinglinkelectronics.com/files/papers/Track3_Session9_Zone_Based_Automotive_Backbones_Tunneling_PCIe_EndricSchubert_Frozen.pdf

Authors and Contact Information

David Epping, Sr. Engineer, Missing Link Electronics GmbH
Ulrich Langenbach, Dir. Engineering, Missing Link Electronics GmbH
Endric Schubert, PhD, CTO, Missing Link Electronics, Inc.

Missing Link Electronics, Inc.
2880 Zanker Road, Suite 203
San Jose, CA 95134, USA
+1-408-475-1490

Missing Link Electronics GmbH
Industriestrasse 10
89231 Neu-Ulm
Germany

www.missinglinkelectronics.com

MLE (Missing Link Electronics) is offering technologies and solutions for Domain-Specific Architectures, which focus on heterogeneous computing using FPGAs. MLE is headquartered in Silicon Valley with offices in Neu-Ulm, Germany.

NPAP is the Network Protocol Accelerator Platform, an IP Core for FPGA/ASIC from Fraunhofer Heinrich-Hertz Institute (HHI). Fraunhofer HHI focuses on 10 to 100 Gbit transmission in the field of high-performance telecom components and on mobile broadband systems. Fraunhofer HHI is located in Berlin, Germany. More information can be found at http://MLEcorp.com/NPAP

TSN is technology from Fraunhofer IPMS. More information can be found at https://www.ipms.fraunhofer.de/en/Components-and-Systems/Components-and-Systems-Data-Communication/ip-cores/IP-Core-Modules/time-sensitive-networking.html