NetApp Knowledge Base

Why is the reported LUN latency higher than the volume latency in Data ONTAP 7-Mode?

https://kb-stage.netapp.com/on-prem/ontap/OHW/OHW-KBs/Why_is_the_reported_LUN_latency_highe…
Updated: Wed, 24 Jun 2026 06:32:15 GMT

Applies to

- DATA ONTAP 7-Mode
- Logical Unit (LUN)
  - ISCSI
  - FCP
    - R2T (Ready to Transfer)
    - XFER_RDY (Transfer Ready)

Answer

- It is often observed that the latency measured for iSCSI (and FCP) is significantly higher than that for the underlying volume, and the operation count at the volume level is higher than that measured on the contained LUNs.
- This is since the Volume latency(WAFL latency) is only a subset of the LUN latency, so it's the expected behavior to see a lower Volume latency compared to the LUN latency.
- This article focuses on the scenarios where LUN latency is significantly higher than Volume latency.
    - The major reason for this significant latency difference is that at the WAFL/Volume layer, operation size is limited to 64KB. If the client has to send an operation with a payload larger than 64KB, the payload has to be broken down into multiple Volume operations. Because of this WAFL side limit, each iSCSI session or FCP login will negotiate settings, specifying the amount of data that can be sent in one single PDU(Protocol Data Units). So the Volume latency only stands for the time it takes to handle one single 64KB PDU, but the LUN latency might be measuring the total handling time of several 64KB PDUs.
    - LUN latency(iSCSI latency or FCP latency) is measured from when the first PDU of the command is fully received, until the time the last PDU of the response is sent to the output queue of ONTAP.
    - In 7-mode ONTAP, this works in a strictly **serial**ized way.

Example with a 256KB op size iSCSI write:

1. The client sends out the first 64KB PDU
2. ONTAP receives the PDU, then after handling the first 64KB write, it sends out an R2T back to the client
3. The client receives the R2T and then send out the second 64KB PDU
4. ONTAP receives the PDU, then after handling the second 64KB write, it sends out another R2T back to the client
5. The client receives the R2T and then send out the third 64KB PDU
6. ONTAP receives the PDU, then after handling the third 64KB write, it sends out another R2T back to the client
7. The client receives the R2T and then send out the fourth 64KB PDU
8. ONTAP receives the PDU, then after handling the fourth 64KB write, it sends out the iSCSI write response to the client, that's where ONTAP ends the measuring of the LUN latency

- Volume latency is only for one single 64KB PDU handling, such as the time spent in step **2**, **4**, **6** and **8**
- The time spent between each 64KB PDU handling is the Network Round Trip time, which also includes the client handling time, the time that it takes for the client to send out the next 64KB PDU, such as the time between **2** and **4**, between **4** and **6**, and between **6** and **8**
- LUN latency is for the whole 256KB iSCSI write handling, including 4 64KB PDU handling(Volume latency) plus 3 Network Round Trips Time (RTT)
- FCP has a similar term as iSCSI R2T, which is called XFER_RDY

- In 7-Mode ONTAP,
    - if the LUN latency is significantly higher than Volume Latency * ROUNDUP(OperationSize / 64KB), it means the R2T(or XFER_RDY for FCP) takes a long time to reach the client, or the client takes a long time to send out the next PDU.
      It indicates that the data path between the clients and the storage is slow.
      this could be:
        - Congestion
        - Low Bandwidth
        - Packet Losses
        - And so on

Example:

- A comparison of the Ops and Latency from the Volume layer and LUN layer for a client having 256KB reads
- iSCSI protocol as an example, but the same applies to FCP as well

To distinguish between these two scenarios:

- **Volume Latency**

```
System1> stats show volume:myvol
volume:myvol:total_ops:132/s
volume:myvol:avg_latency:5ms
volume:myvol:read_ops:5/s
volume:myvol:read_data:1923b/s
volume:myvol:read_latency:3ms
volume:myvol:write_ops:186/s
volume:myvol:write_data:1876b/s
volume:myvol:write_latency:6ms
volume:myvol:other_ops:0/s
volume:myvol:other_latency:0ms
```

- **LUN latency**

```
system1>lun stats -o -i 1
Read  Write Other QFull Read  Write Average  Queue  Partner   Lun
 Ops   Ops   Ops       kB   kB   Latency  Length  Ops kB
  0   351    0    0    0  44992   11.35    3.00    0   0  /vol/tpcc/myvol
  0   233    0    0    0  29888   14.85    2.05    0   0  /vol/tpcc/myvol
  0   411    0    0    0  52672    8.93    2.08    0   0  /vol/tpcc/myvol
```