NVMe Queue Depth Multipath

Red Hat has multiple storage partners and customers who are complaining that we are ending support for dm-multipath with NVMe-Over Fabrics and that the native nvme-multipath round-robin scheduler does not “work”. In an effort to get to the bottom of these complaints the RHEL storage team has been working on adding a queue depth scheduler to native nvme multipathing. We can now reproduce and demonstrate the round-robin multipathing problem with a number of different NVMe-OF platforms, and we have proposed patches to add a new Queue Depth scheduling algorithm to nvme/host/multipath.c.

These patches can be seen at:

https://lore.kernel.org/linux-nvme/20230925163123.16042-1-emilne@redhat.com/T/#u

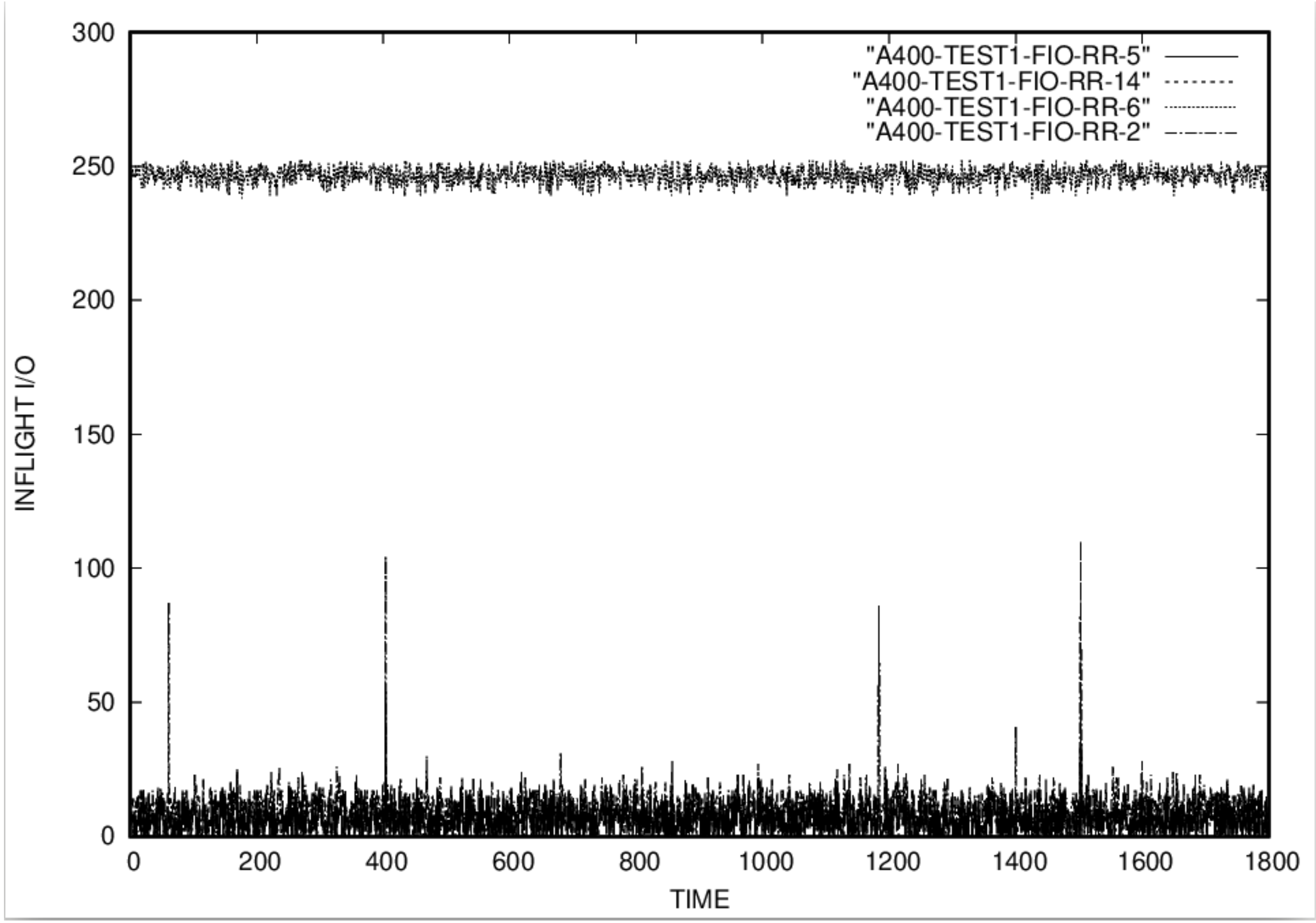
Testbed Configuration

A NetApp A400 with 8 controllers - 4 active optimize paths and 4 active non-optimized paths consiting of 4 32GB nvme-fc controllers and 4 100Gbps nvme-tcp controllers.

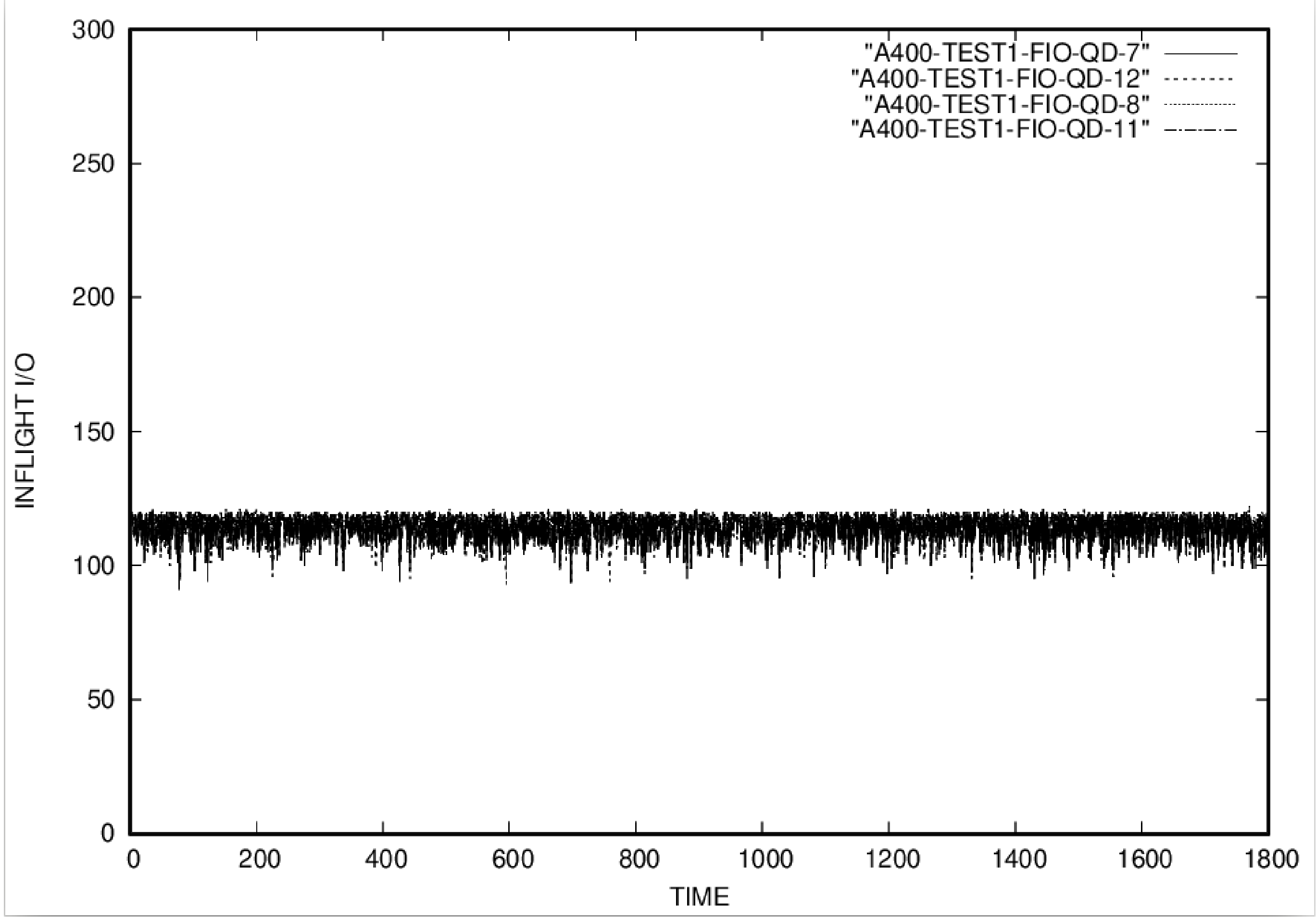
This provided a true multi-path test bed with 4 active-optimized paths with a mix of controllers paths that use different transports with inherently different latency characteristics.

```
 [root@rhel-storage-104 ~]# nvme list-subsys
nvme-subsys3 - NQN=nqn.1992-08.com.netapp:sn.2b82d9b13bb211ee8744d039ea989119:subsystem.SS104a
\
 +- nvme10 fc traddr=nn-0x2027d039ea98949e:pn-0x202cd039ea98949e,host_traddr=nn-0x200000109b9b7f0d:pn-0x100000109b9b7f0d live
 +- nvme11 fc traddr=nn-0x2027d039ea98949e:pn-0x2029d039ea98949e,host_traddr=nn-0x200000109b9b7f0c:pn-0x100000109b9b7f0c live
 +- nvme12 fc traddr=nn-0x2027d039ea98949e:pn-0x2028d039ea98949e,host_traddr=nn-0x200000109b9b7f0d:pn-0x100000109b9b7f0d live
 +- nvme13 tcp traddr=172.18.50.13,trsvcid=4420,src_addr=172.18.50.3 live
 +- nvme2 tcp traddr=172.18.60.16,trsvcid=4420,src_addr=172.18.60.4 live
 +- nvme3 fc traddr=nn-0x2027d039ea98949e:pn-0x202dd039ea98949e,host_traddr=nn-0x200000109b9b7f0c:pn-0x100000109b9b7f0c live
 +- nvme4 tcp traddr=172.18.50.15,trsvcid=4420,src_addr=172.18.50.3 live
 +- nvme9 tcp traddr=172.18.60.14,trsvcid=4420,src_addr=172.18.60.4 live
```

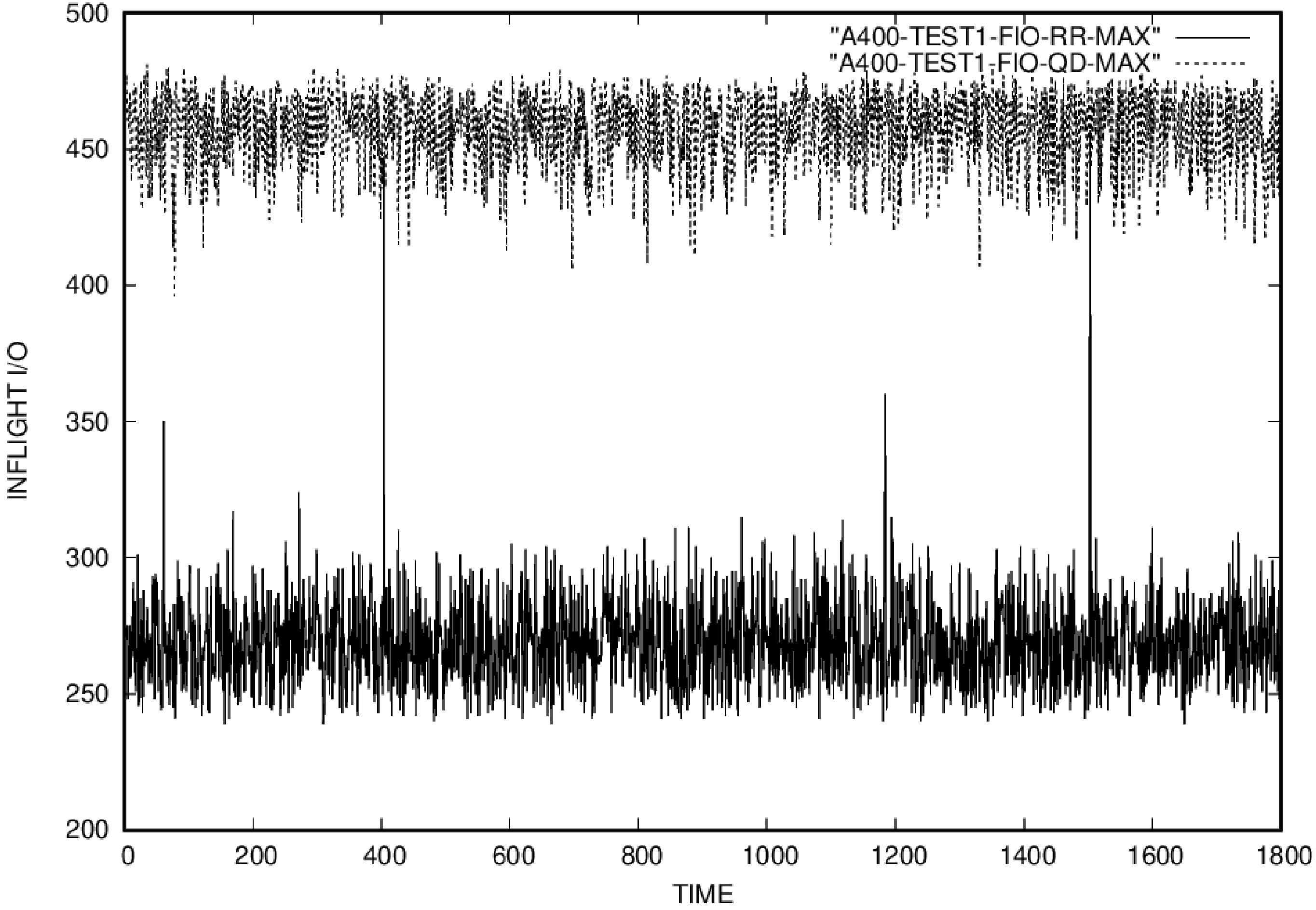
4 Optimized paths with round-robin



4 Optimized paths with queue-depth



Combined number I/Os, all paths, both RR and QD


"A400-TEST1-FIO-RR-MAX"
"A400-TEST1-FIO-QD-MAX"
INFLIGHT I/O
TIME