SDC | StorageAI™

A SNIA Event

April 29, 2026 • Denver, Colorado

Addressing QoS issues through I/O Command Prioritization in NVMe® Technology

Anthony Constantine

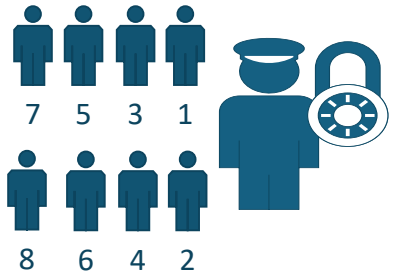
Question?

Who has been to a theme park lately?

Typical Theme Park (Ideal conditions)

- 3 (or more) arbitration points
  - Each with multiple queues

Line for Security



Line for Entry

Line for Ride
"NVM Express® Technology"

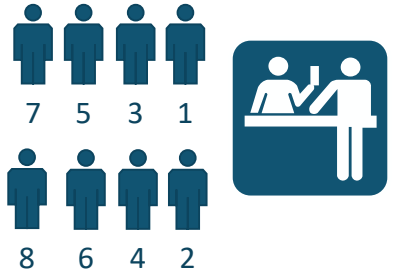
SDC | StorageAI™

Typical Theme Park (Ideal conditions)

- 3 (or more) arbitration points
  - Each with multiple queues
- Every person takes the same amount of time to get through each queue

Line for Security

Line for Entry



Line for Ride
"NVM Express® Technology"

SDC | StorageAI™

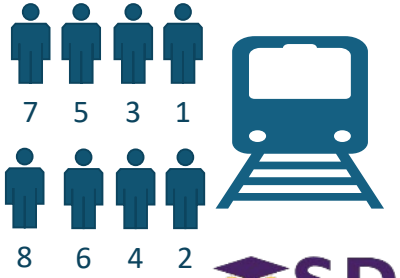
Typical Theme Park (Ideal conditions)

- 3 (or more) arbitration points
  - Each with multiple queues
- Every person takes the same amount of time to get through each queue

Line for Security

Line for Entry

Line for Ride
"NVM Express® Technology"



Typical Theme Park (Ideal conditions)

- 3 (or more) arbitration points
  - Each with multiple queues
- Every person takes the same amount of time to get through each queue
- This simple view looks a lot like Round Robin

Line for Security

Line for Entry

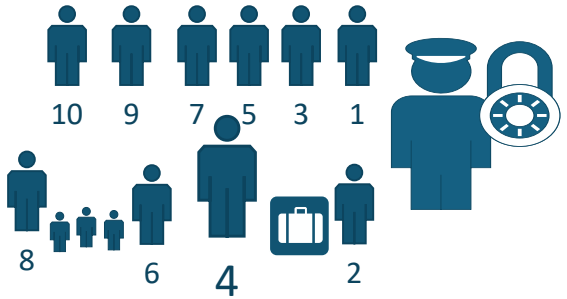
Line for Ride
"NVM Express® Technology"

SDC | StorageAI™

Typical Theme Park (Reality)

- There are some deltas introduced
  - Person with a backpack

Line for Security



Line for Entry

Line for Ride
"NVM Express® Technology"

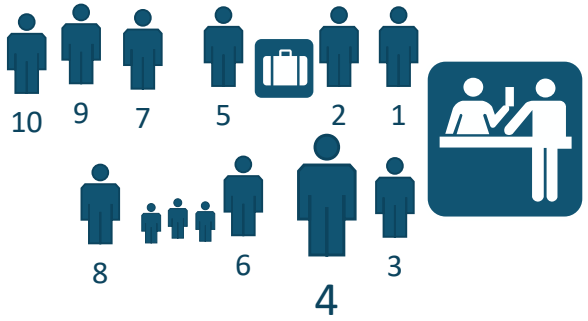
SDC | StorageAI™

Typical Theme Park (Reality)

- There are some deltas introduced
  - Family of 4

Line for Security

Line for Entry



Line for Ride
"NVM Express® Technology"

SDC | StorageAI™

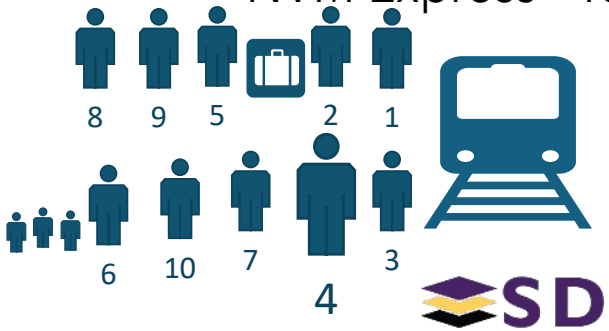
Typical Theme Park (Reality)

- There are some deltas introduced
  - Really tall person

Line for Security

Line for Entry

Line for Ride
"NVM Express® Technology"



SDC | StorageAI™

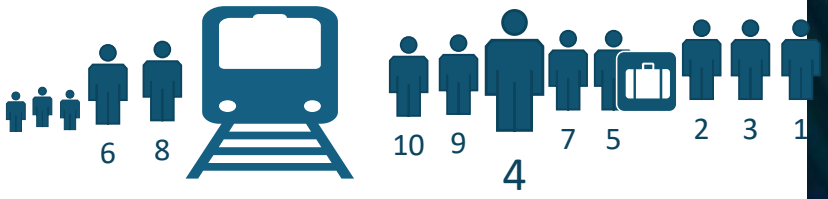
Typical Theme Park (Reality)

- There are some deltas introduced
- These deltas induce delays
- Delays impact experience

Line for Security

Line for Entry

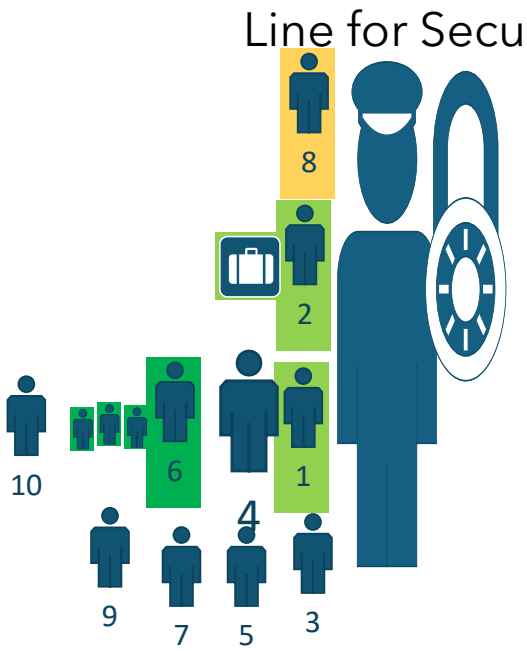
Line for Ride
“NVM Express® Technology”

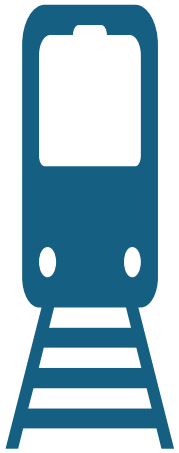


SDC | StorageAI™

How were these deltas resolved?

- Three types of security lines
  - VIPs
  - People with bags
  - People without bags

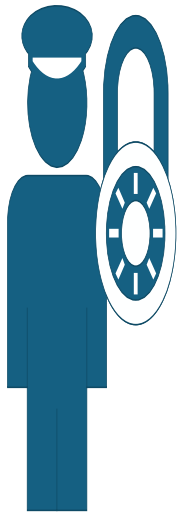




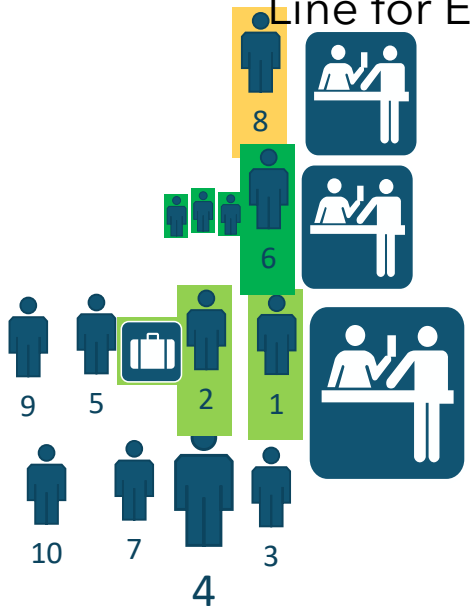
How were these deltas resolved?

- Three types of security lines
- Priority entry points (multiple places)
  - VIPs
  - Early entry
  - Everyone else.

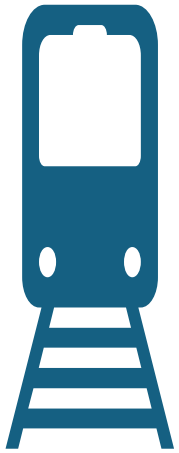
Line for Security



Line for Entry



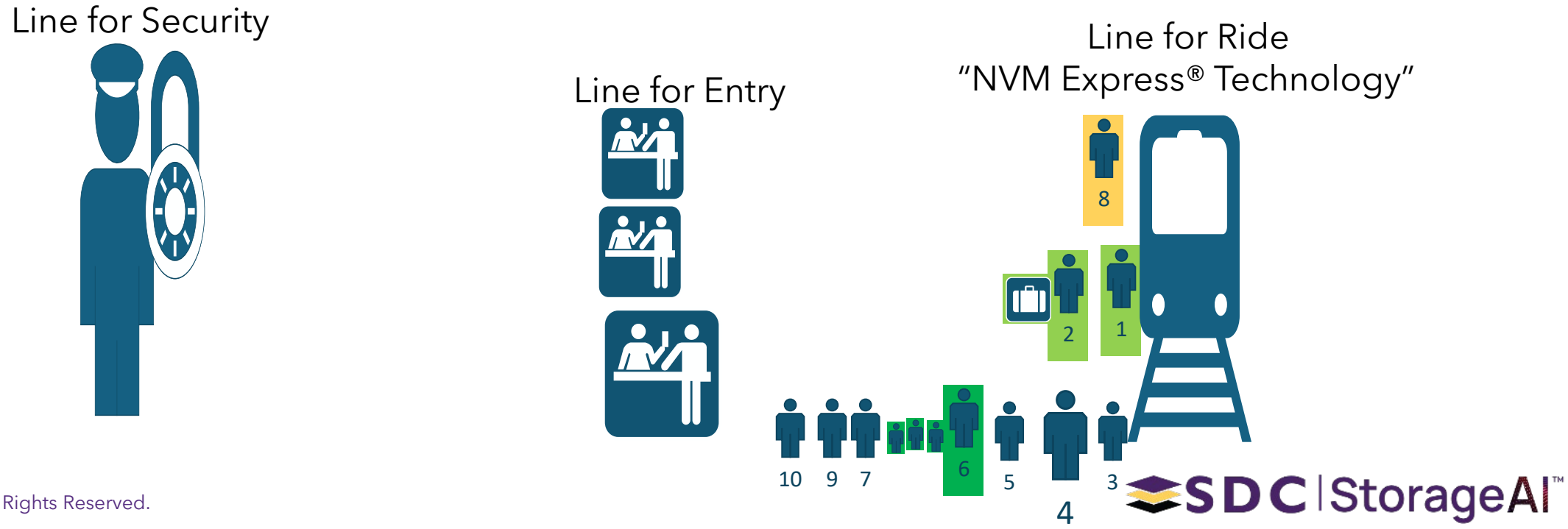
Line for Ride
"NVM Express® Technology"



SDC | StorageAI™

How were these deltas resolved?

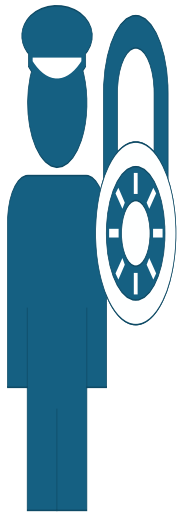
- Three types of security lines
- Priority entry points (multiple places)
  - VIP
  - "Skip the line"
  - Everyone else
- Some ride cars got bigger



How were these deltas resolved?

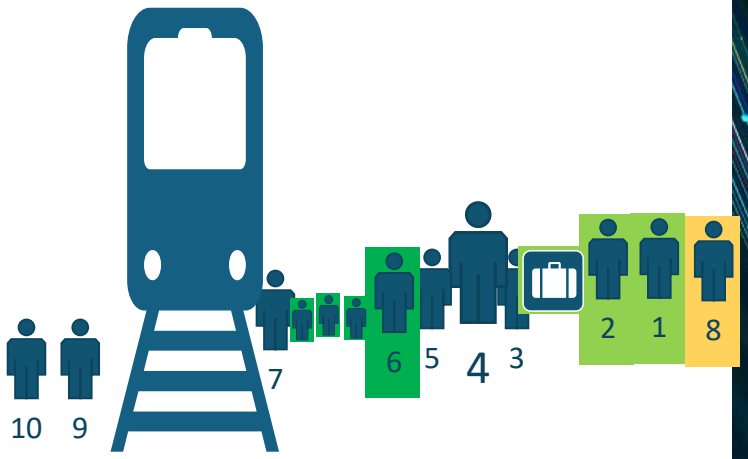
- Two types of security lines
- Priority entry points (multiple places)
- Some ride cars got bigger
- Conclusion: Prioritization helped the experience for some at a minor impact of others

Line for Security



Line for Entry

Line for Ride
"NVM Express® Technology"

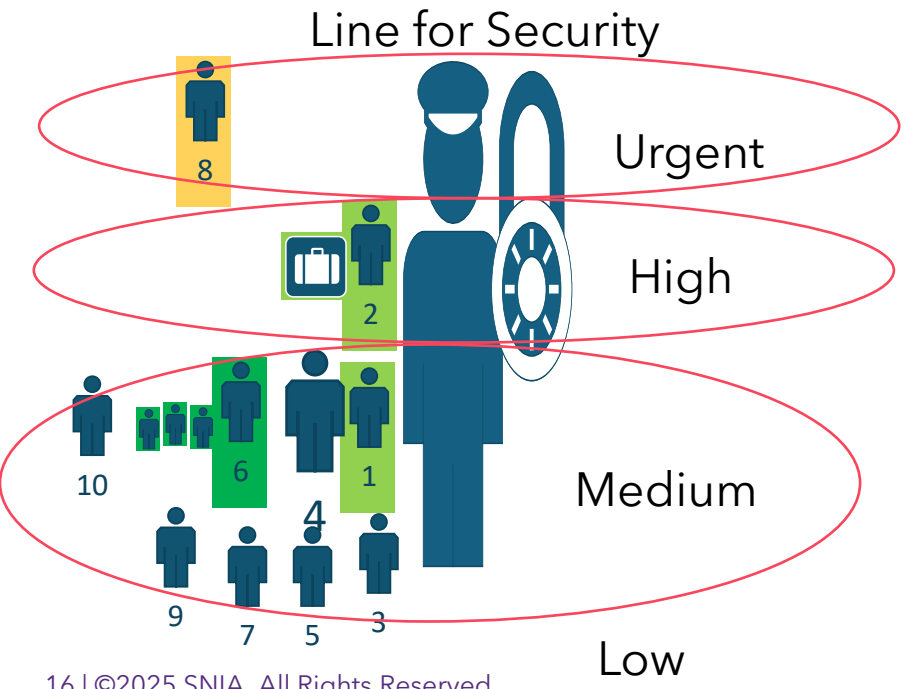


The Moral of the Story

- Theme parks queues are not about being fair
- Theme park queues are about extracting value
  - Important Caveat: Without ticking too many people off

Get on with It!!!

- We do some of this with NVMe® Technology today
  - Weighted Round Robin (WRR) w/ Urgent priority allows hosts to prioritize queues based on 4 levels



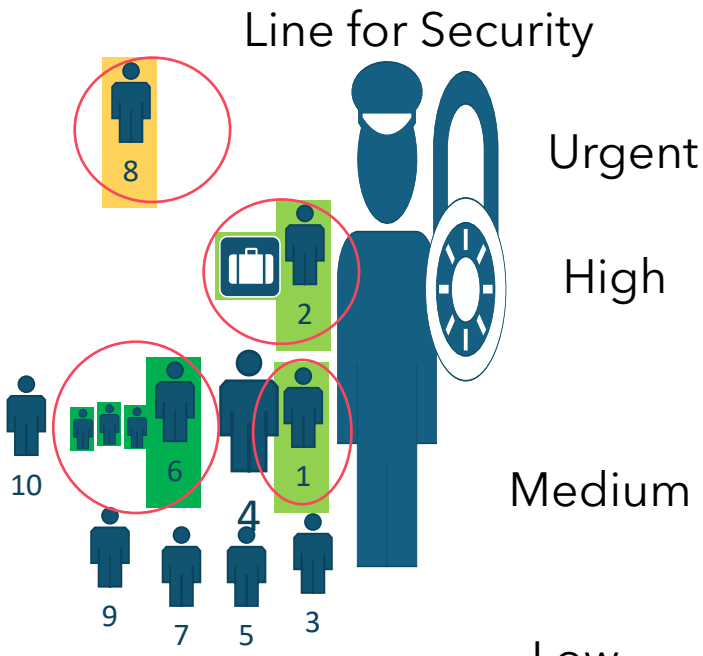
Line for Entry

Line for Ride
"NVM Express® Technology"

SDC | StorageAI™

The problem with NVMe® Technology arbitration

- Some queues need their priority to flow through all stages
- In a 24/7/365 environment, some higher priority queues need a better delivery contract vs. the lower priority queues.



SDC | StorageAI™

I/O Prioritization

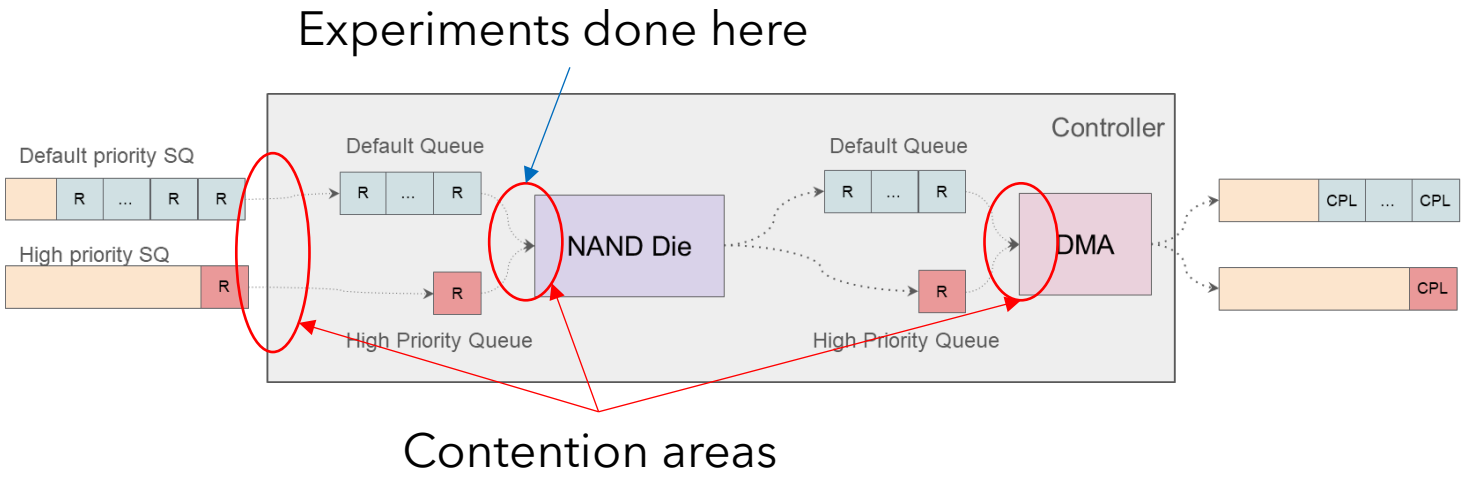
- Two key areas proposed to address NVMe® Technology arbitration issues:
  1. Maintaining priority end to end (from fetch to completion)
  2. Drive the lowest latencies on highly prioritized I/O commands even in presence of other I/O commands (e.g., a time-based arbitration option)

- This is proposed in NVM Express® as TP4221 IO Prioritization

Proposal, subject to change

Maintaining priority End to End ("skip the line")

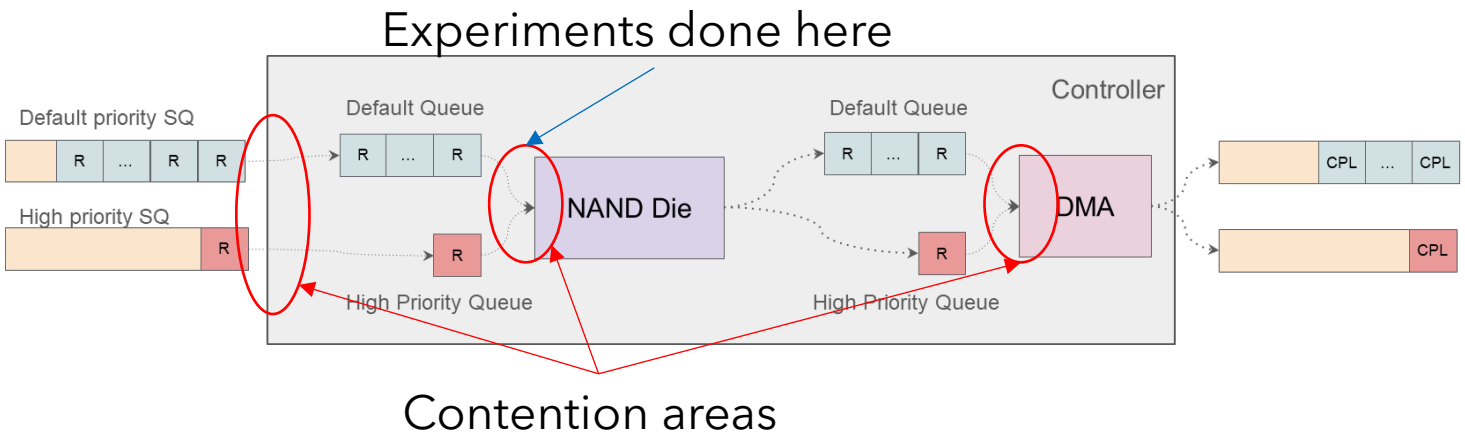
- For queues with a certain priority, a hint or tag will follow the command to indicate that its priority shall be maintained.
- This allows the various contention/arbitration areas in the device to know that certain commands need to be prioritized



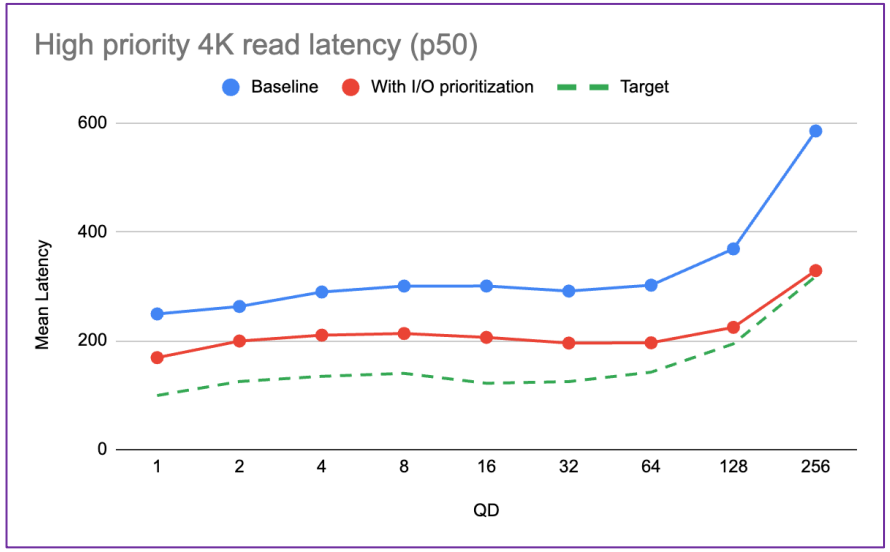
Proposal, subject to change

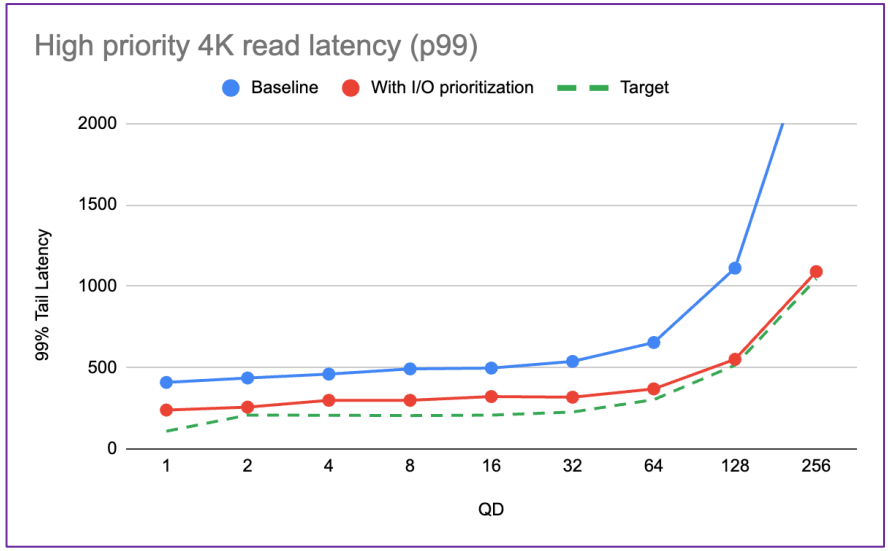
Maintaining Priority E2E Benefits

- Experiments on 2, 3, and 4 queues showed ~20% benefit to p50 and ~50% benefit to P99 for high priority queue.
  - Experiments ran in area with no NVMe® Technology host-controlled arbitration.



Credit Yuliya Tarnikova, Google





Proposal, subject to change

SDC | StorageAI™

Maintaining Priority E2E Details

- Incremental changes for tracking priority levels E2E
  - Device tracking which priority levels need to be carried E2E
  - Host allocating priorities will need to make sure they do not cause starvation of lower priorities
- Corner cases to handle
  - If an internal transaction must occur, it may need to take higher priority.
- NVMe® Technology does not specify performance requirements
  - WL's proving compliance may need to be managed outside of NVMe.

New Arbitration

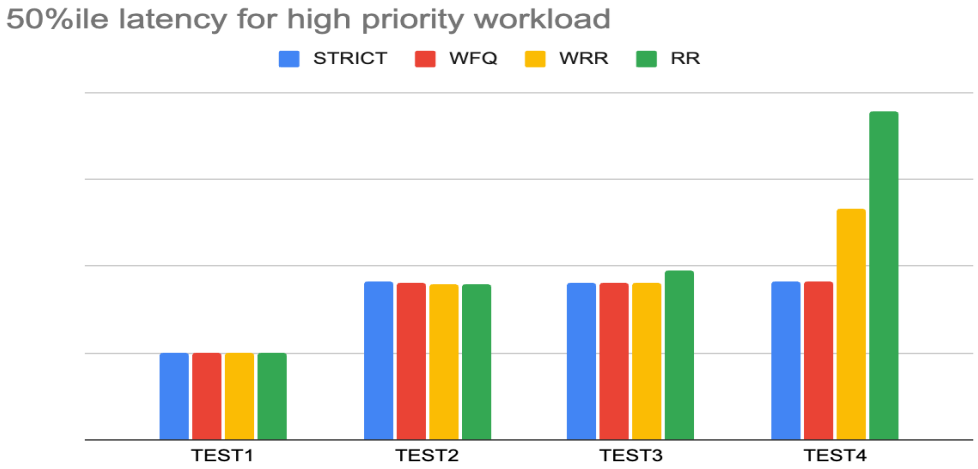
- NVMe® Technology todays lists 2 arbitration methods
  - Round Robin
  - Weighted Round Robin w/ Urgent Priority
- These methods are simple but with drawbacks
  - Each command may take varying times to execute (e.g., 4KiB read vs. 64KiB write)
  - These execution times can mean certain queues and commands may get more share of execution than desired
- A better arbitration method can help alleviate this bottleneck.

Proposal, subject to change

Example arbitration benefits

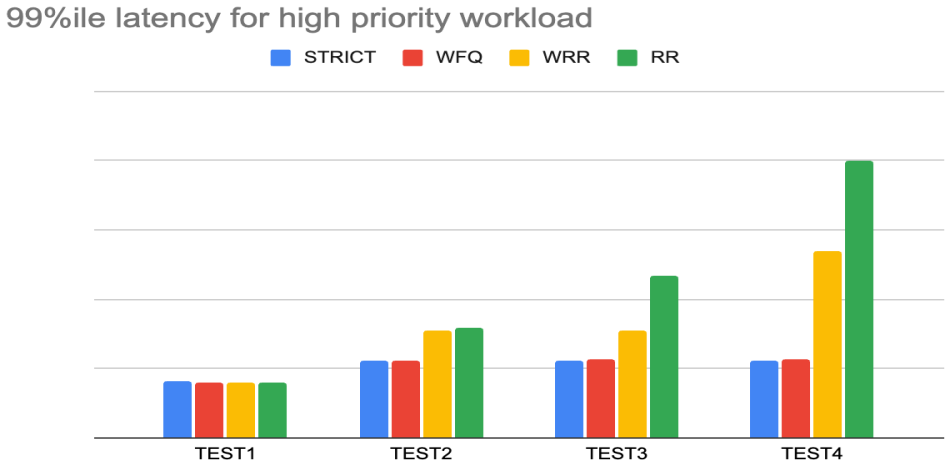
- Using Weighted Fair Queuing (WFQ) for arbitration, improvements seen over WRR especially in P99
  - WFQ does arbitration based on time allocation.
  - Note: WFQ was used as an example arbitration and not necessarily the only method

Credit Yuliya Tarnikova, Google



Workloads:
Chips: 16
High priority workload: QD=4, W=85
Aggressor workloads:

- TEST1:
  - None
- TEST2:
  - QD=1024, W=15
- TEST3:
  - QD=16, W=14
  - QD=1024, W=1
- TEST4:
  - QD=16, W=10
  - QD=1024, W=3
  - QD=1024, W=2
  - QD=1024, W=1

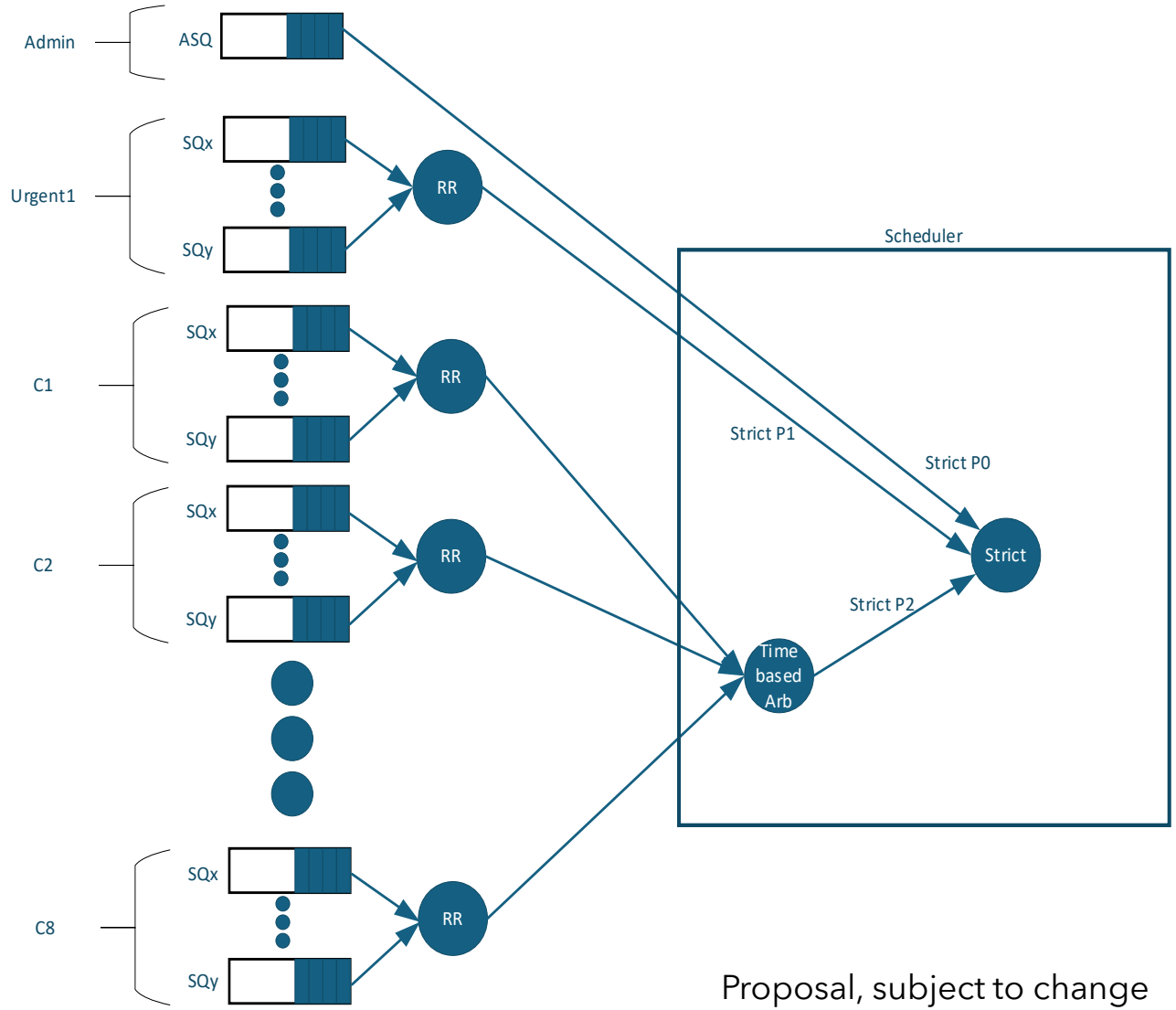


Proposal, subject to change

SDC | StorageAI™

One possible implementation

- Implementation for Fetch
  - Could be implemented as well at multiple levels.



Proposal, subject to change

Summary

- NVMe® Technology arbitration needs to evolve to cover prioritizing certain queues end to end.
- Concepts like skipping the line and new arbitration can help prioritize the queues that need better servicing.
- Work has started in NVM Express® on this (TP4221 IO Prioritization)
- Participate in NVM Express if you are interested in this topic.


SDC | StorageAI™
A SNIA Event
Thank You