SDC 20

BY Developers FOR Developers

Storage Developer Conference
September 22-23, 2020

A Low Latency and Scalable Key Value Store from Modern off the shelf Components

Dan Pollack
Data Storage Science LLC

Outline

- What am I talking about?
- Why do I think it’s worth doing?
- How does it work?
- How good is it?
- What’s next?

What am I talking about?

SDC 20

- A low latency Key-Value store
  - Small keys/values – operations not throughput
  - Two basic components
    - Software – Minio
    - Hardware – NVMeoF/TCP storage system

lightbits

What am I talking about?

- Minio
  - Simple – single command line operation
  - Scalable – single or many servers in a cluster
  - Performant – GB/s Throughput / >1ms Latency
  - Manageable – Multiple control and storage APIs supported
  - Open source

What am I talking about?

- NVMeoF/TCP storage system
  - Very high performance
    - 120us latency
    - 10GBps throughput
  - Common infrastructure – No new hardware
    - Standard server system
    - Standard ethernet networking
  - Disaggregated – No trapped resources
  - Open source software components

Why do I think it's worth doing?

- Disaggregated resources are more flexible – any resource for any workload
- TCP is everywhere and rebuilding/deploying uncommon infrastructure is unpopular – (FCOE, iSCSI, RoCE, other niche technologies)

Why do I think it's worth doing?

- Simple software is easier to understand and use
- Key/Value stores are a specific type of object storage
- DRAM Key/Value systems are common but -
  - NVMe storage 10X - 100X lower cost
  - NVMe storage 10X - 100X higher capacity

Why do I think it's worth doing?

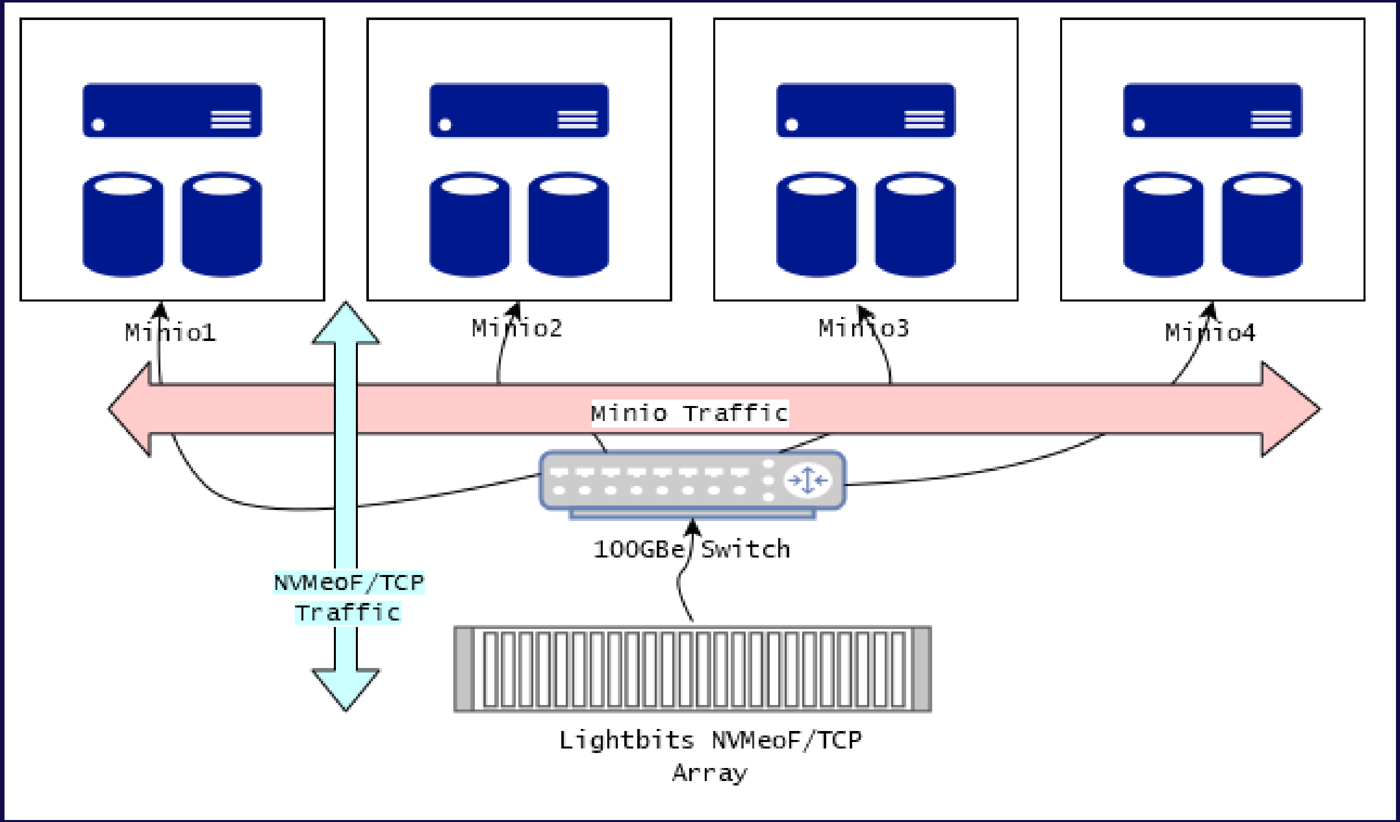
- Fulfills actual need people have with minimal effort and cost
- Rapid to evaluate

How does it work?

SDC 20

- NVMeoF/TCP storage system provides namespaces (LUNs/Volumes) to multiple host systems
- Minio cluster running on host systems exposes the NVMe devices as storage using object APIs
- Clients make Key/Value style requests to Minio

How does it work?



How does it work?

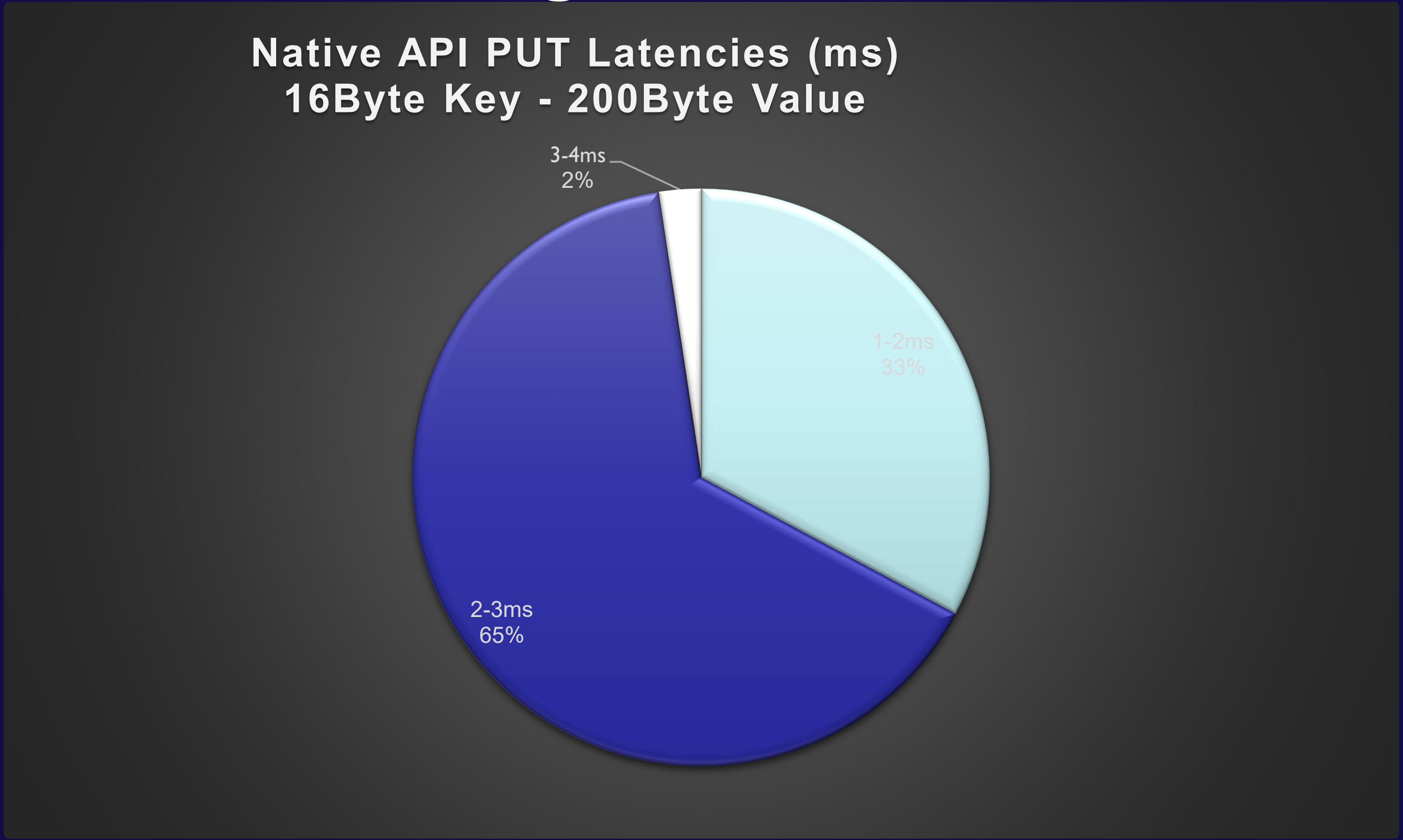
- NVMe storage discovery – once per new device
- NVMe storage access and format
- Minio command

  minio server http://host{1...4}/mount{1...2}

How does it work?

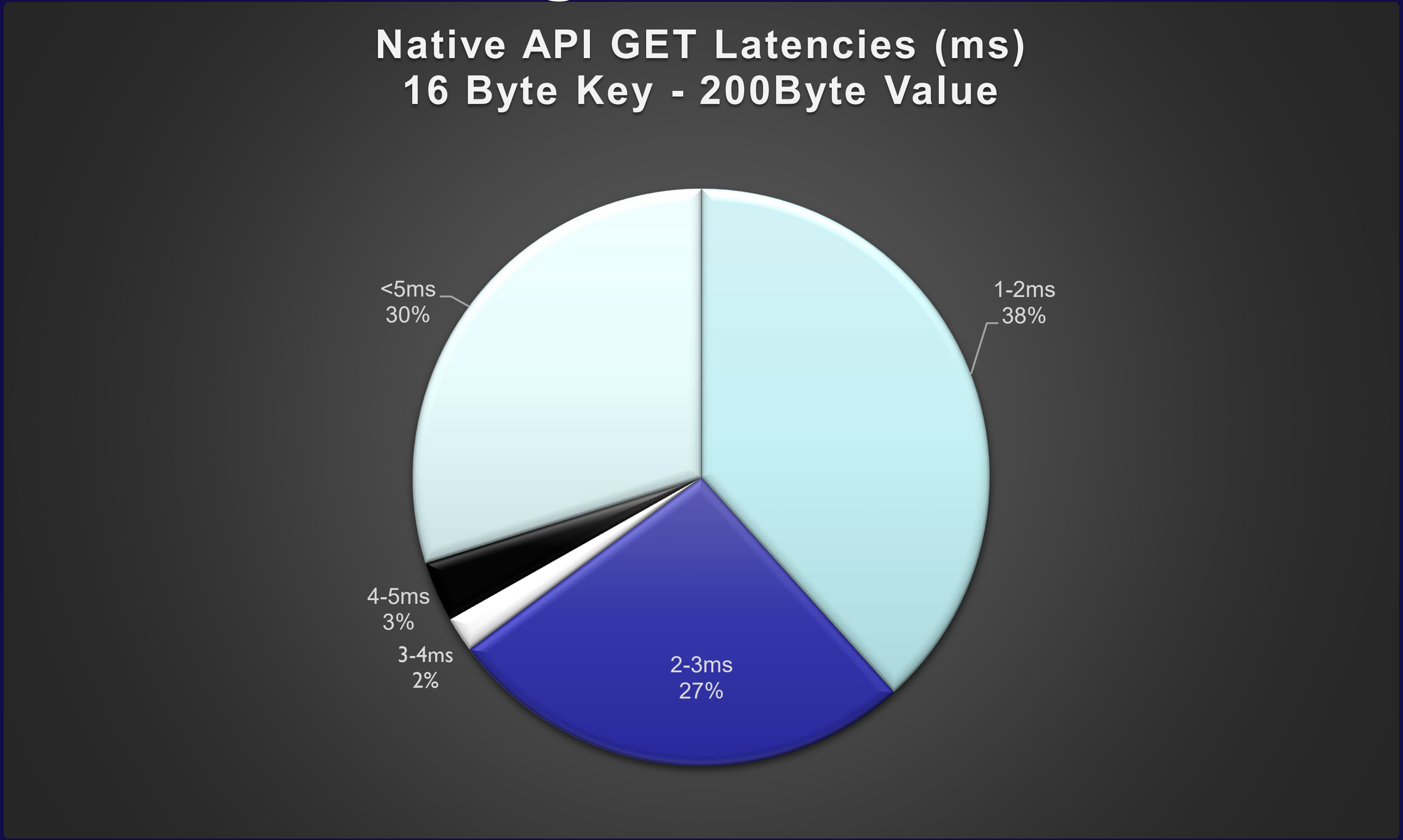
- YCSB for workload evaluation
- Native python API tools for workload evaluation

How good is it?

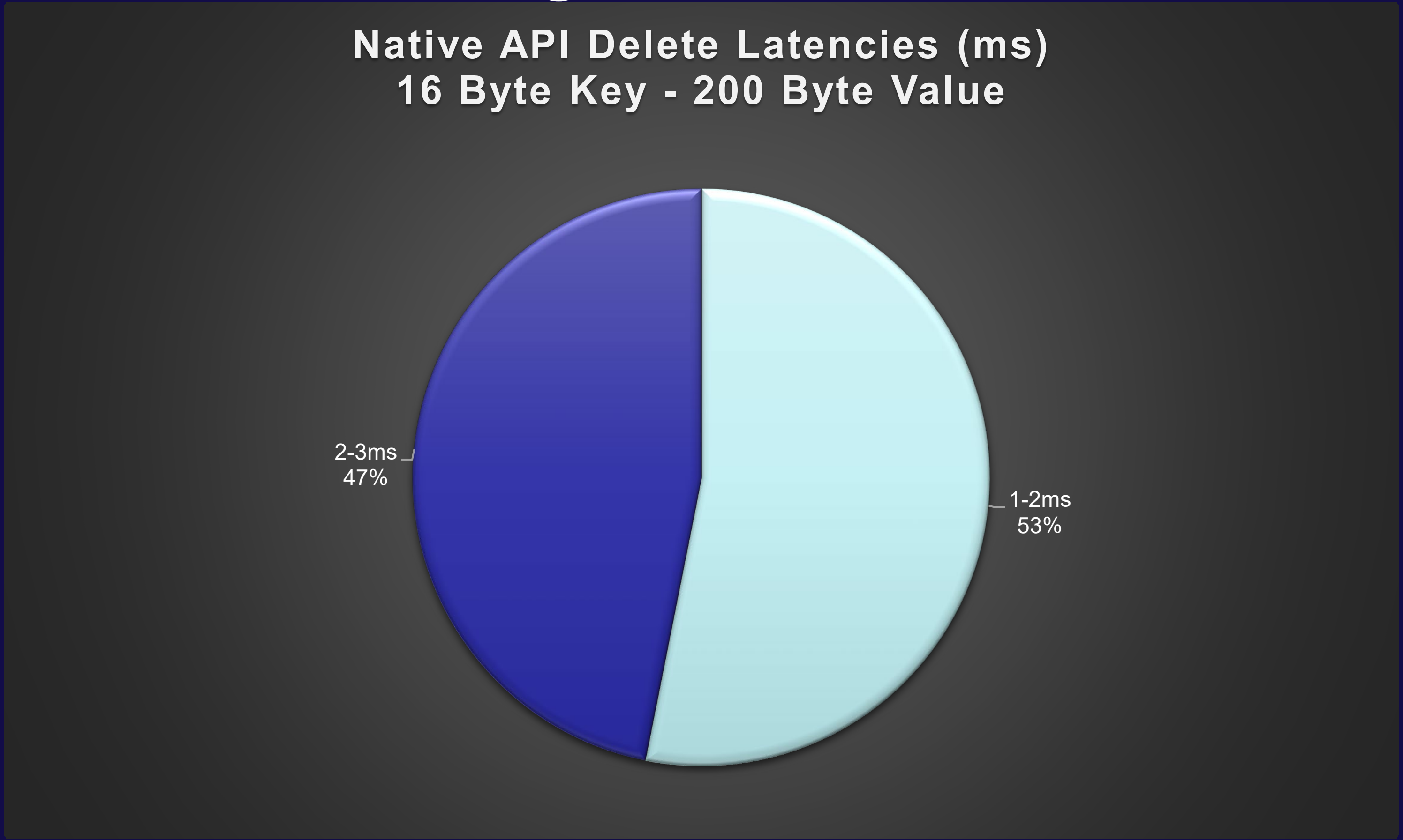


How good is it?

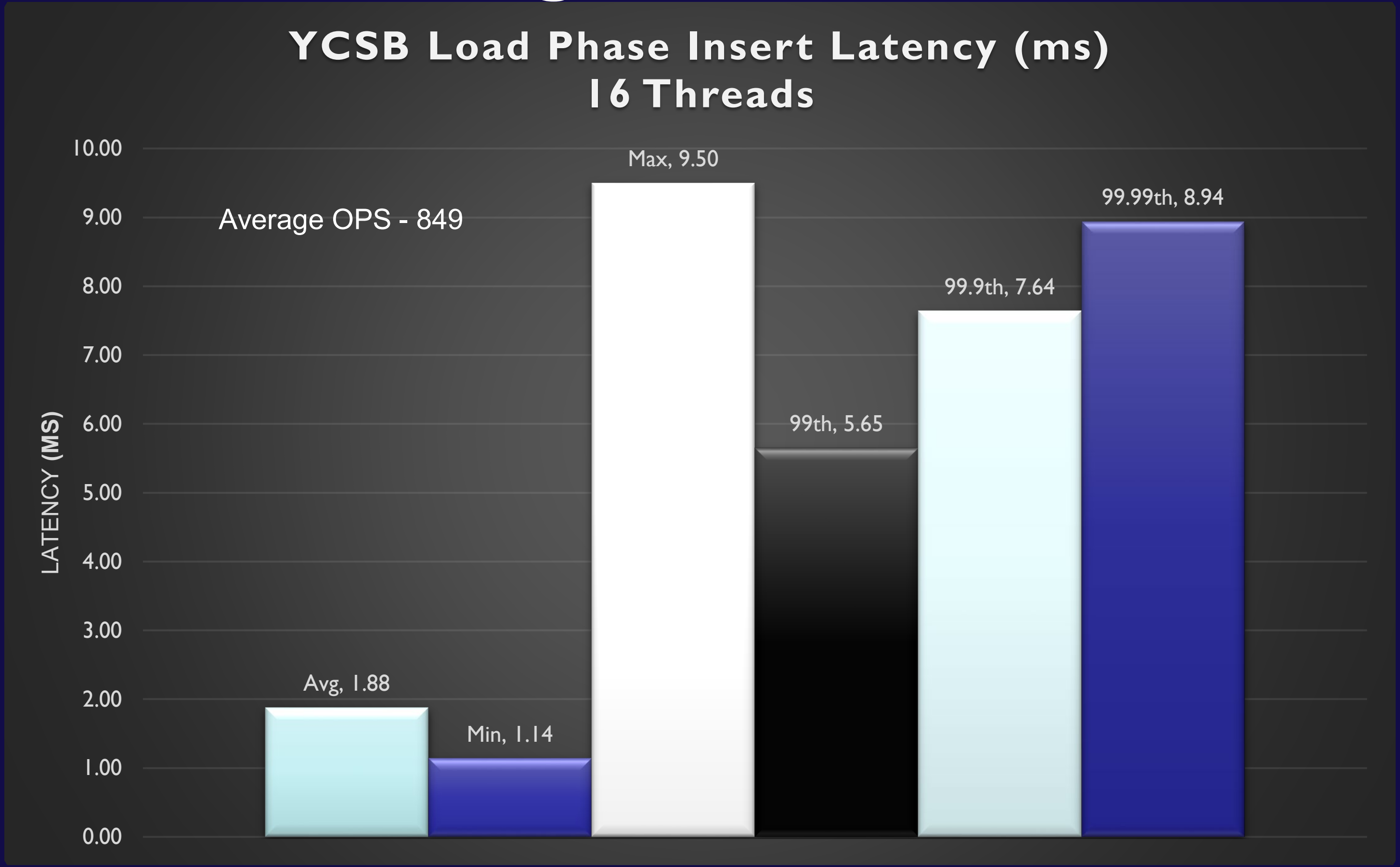
SDC 20



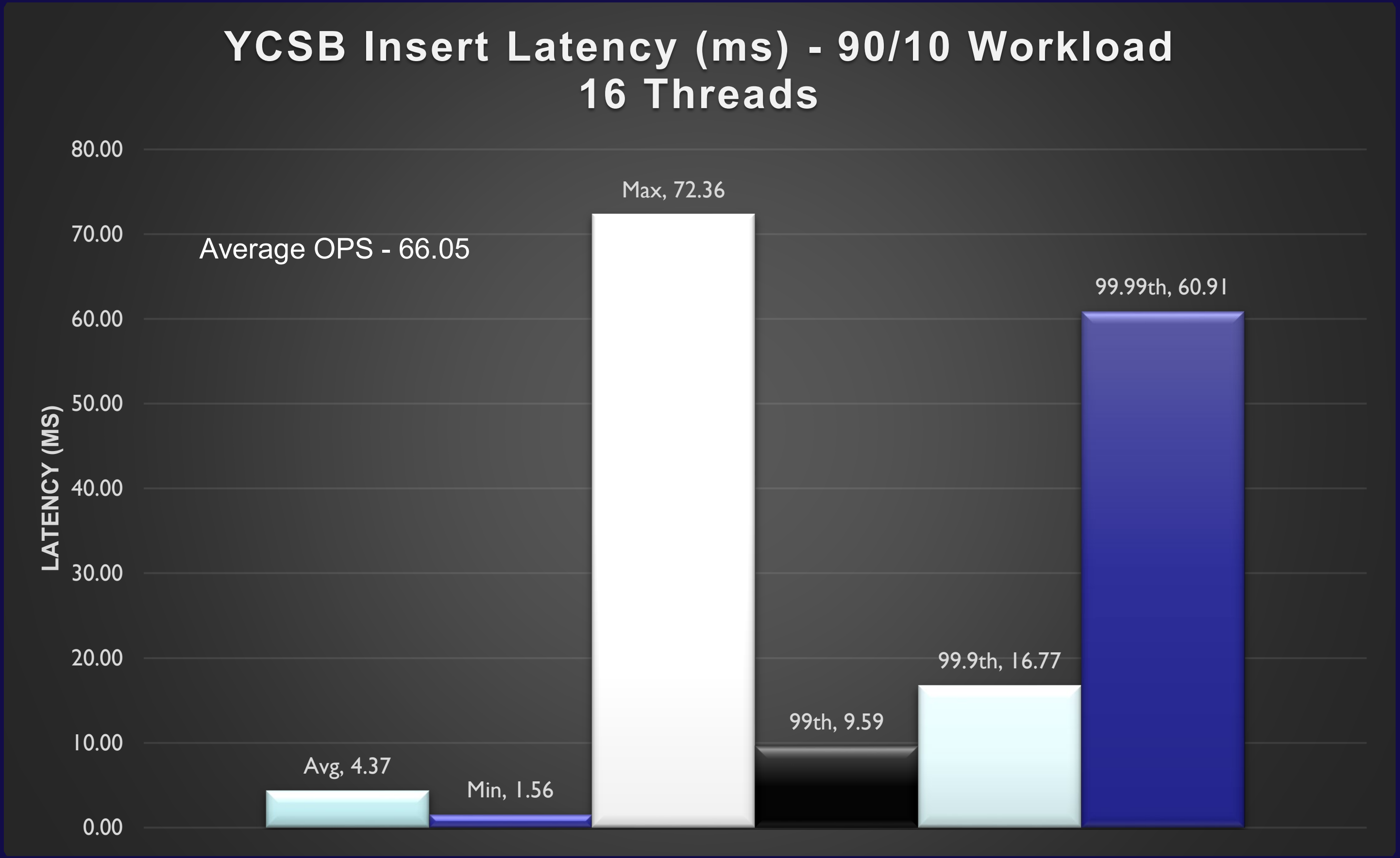
How good is it?



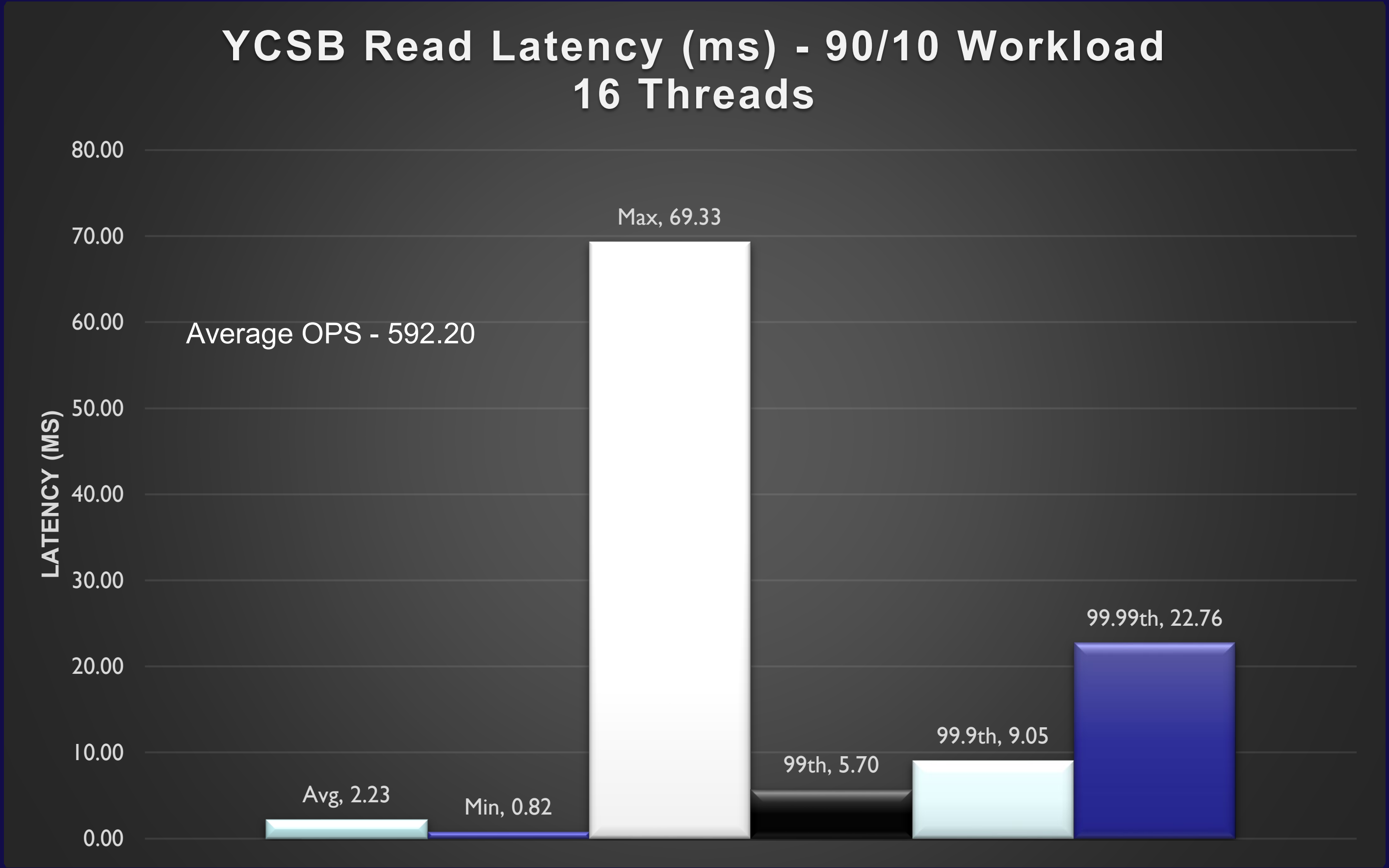
How good is it?



How good is it?

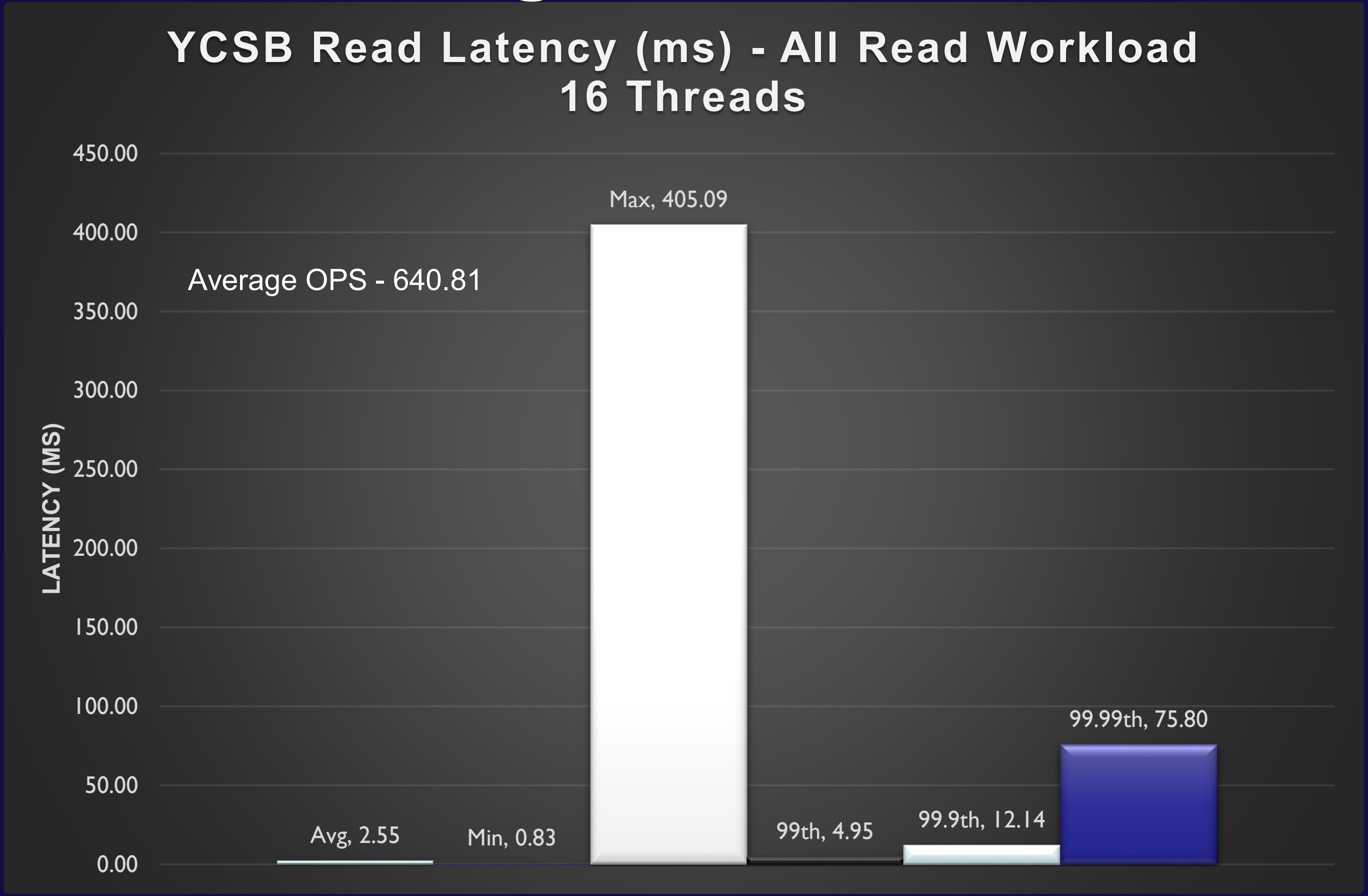


How good is it?



How good is it?

SDC 20



How good is it?

- Round trip network latency is .5ms
- 1ms seems to be a key target for read response time
- API reads are 38% within 2ms
- YCBS reads average slightly more than 2ms
- This is OK – Not great so tradeoffs are important

What's next?

- NVMeoF/TCP in common usage for most NVM block IO over ethernet
- Software enhancements for greater workload suitability
- Native NVMe Key/Value per namespace command set

What's next?

- Minio enhanced for data access
  - Versioning of stored data
  - Storage protection classes
  - Replication between instances
  - Data lifecycle management
  - Immutable data
  - SQL queries on data

Q & A

Please take a moment
to rate this session.

Your feedback matters to us.