Use of Open-Channel SSDs in Chinese Datacenters

Wei Xu
Shannon Systems

About Shannon Systems

- Founded in 2011, a subsidiary of SiliconMotion since 2015.
- Indigenous leading enterprise-grade SSD provider.
- 500+ customers, 100+PB shipment per year.
- From host-based PCIe SSD to Open-Channel SSD.

About Chinese Market

Companies have their own infrastructures.

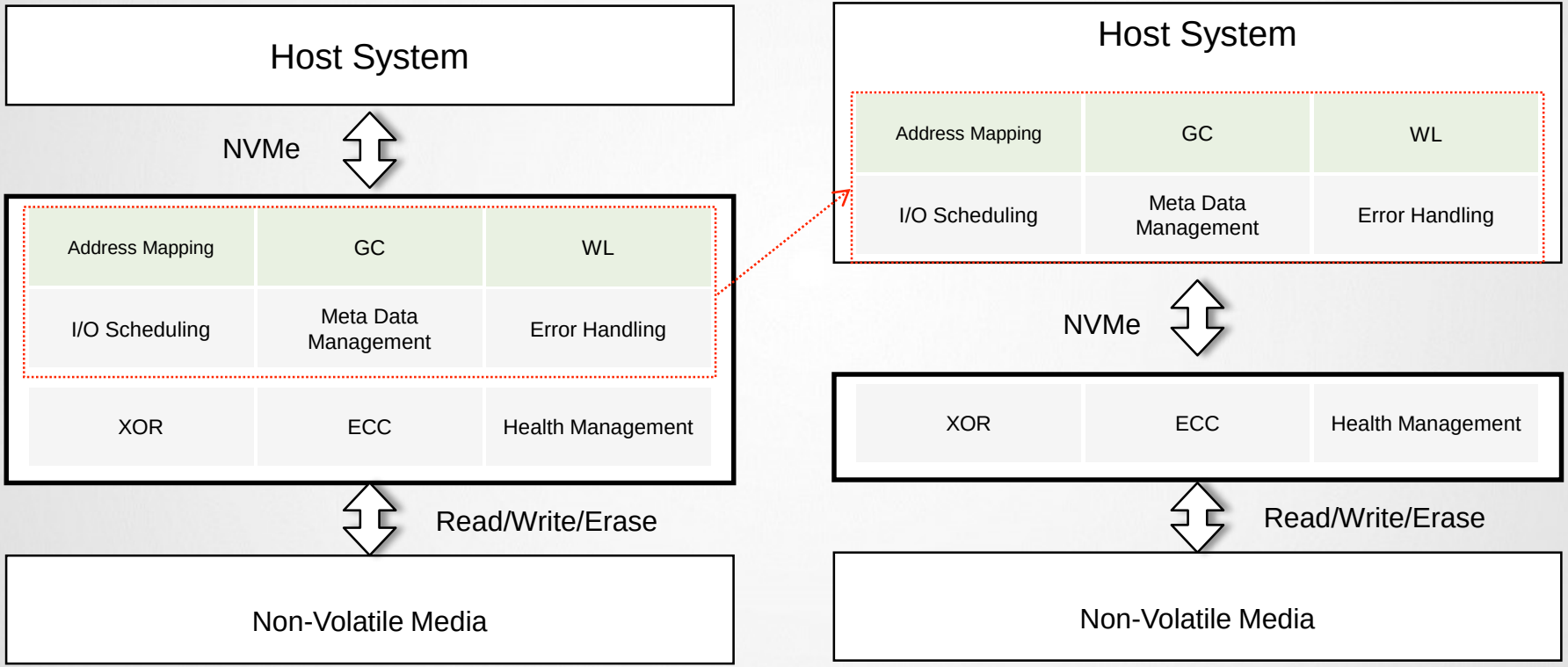
Huge demands come from internet giants continually.

Internet giants are looking for diversification.

Cloud services are growing rapidly.

Traditional companies are migrating to private clouds.

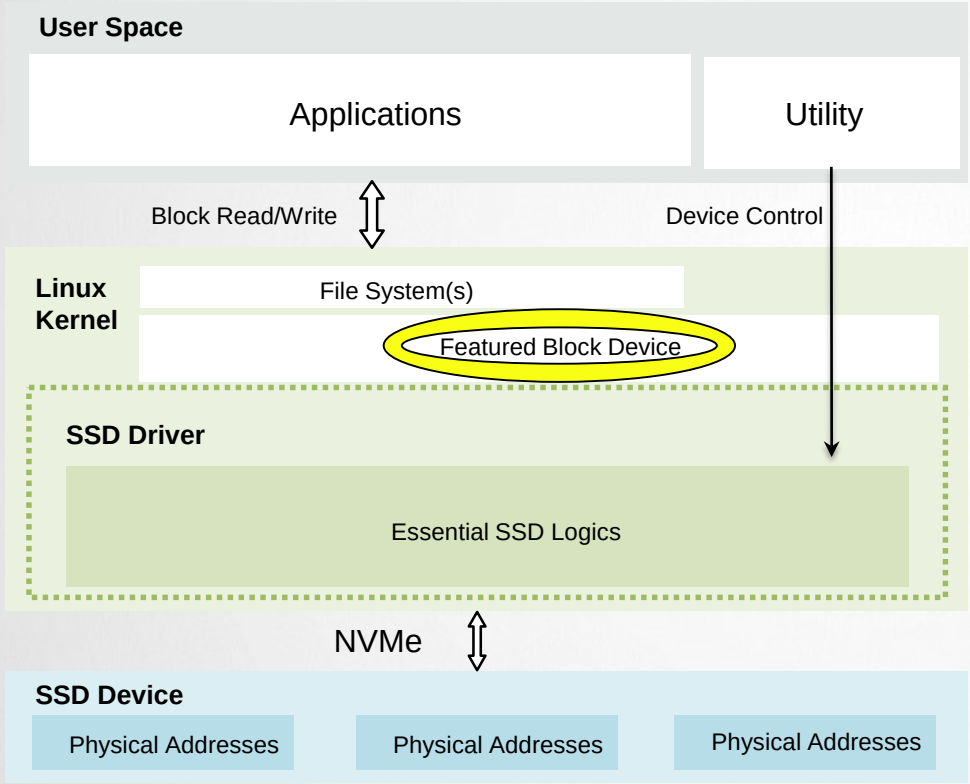
Open-Channel Architecture



Benefits from Open-Channel

- Better performance
- Improved QoS
- Better endurance
- Flexible resource management

Open-Channel in Kernel Mode



- Configurable SSD logics.
- Standard block I/O interface.
- No need of changing users' preferences.

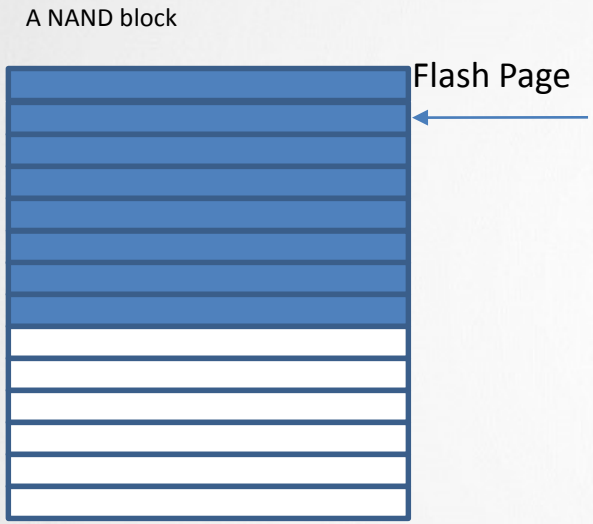
What We have Done.

- Atomic write for MySQL/MariaDB
- Advanced logical volume management
- Namespaces for I/O isolation
- Multi-stream

What We have Done.

- **Atomic write for MySQL/MariaDB**
- Advanced logical volume management
- Namespaces for I/O isolation
- Multi-stream

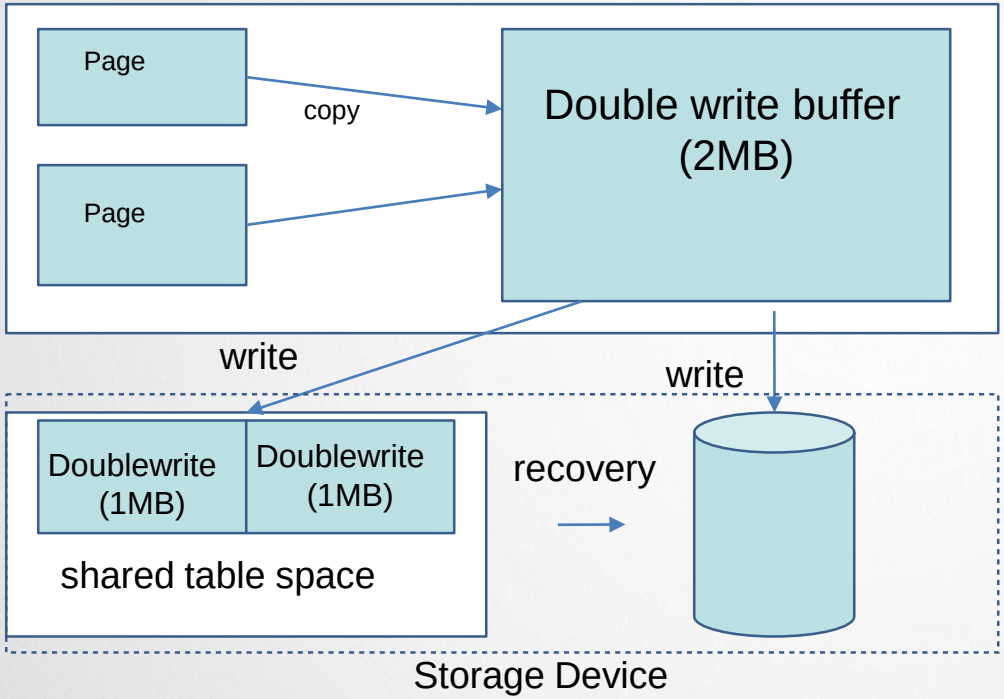
Atomic Write for MySQL/MariaDB



- Atomic operation - an operation that can't be divided, either succeed completely, or fail completely.
- A NAND page programing is an atomic operation.

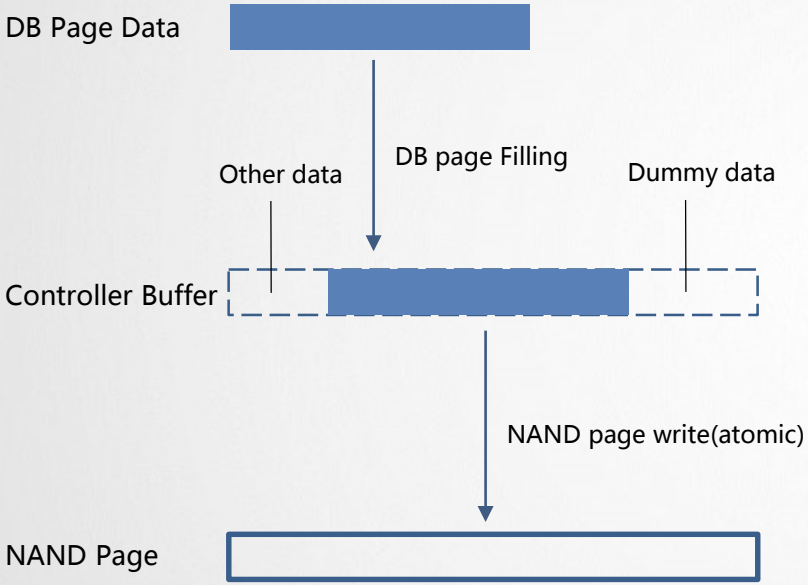
Atomic Write for MySQL/MariaDB

- Double-write mechanism in MySQL



- Conventional storage device cannot ensure the atomicity of “InnoDB page write”.
- “Partial write” of InnoDB page causes data corruption.
- Double Write brings:
  - Double amount of data written, resulting in reduced SSD life span.
  - Higher writing load, resulting in higher write latency.

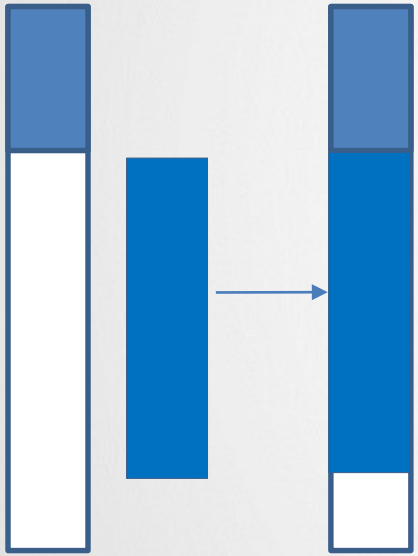
Atomic Write - Implementation



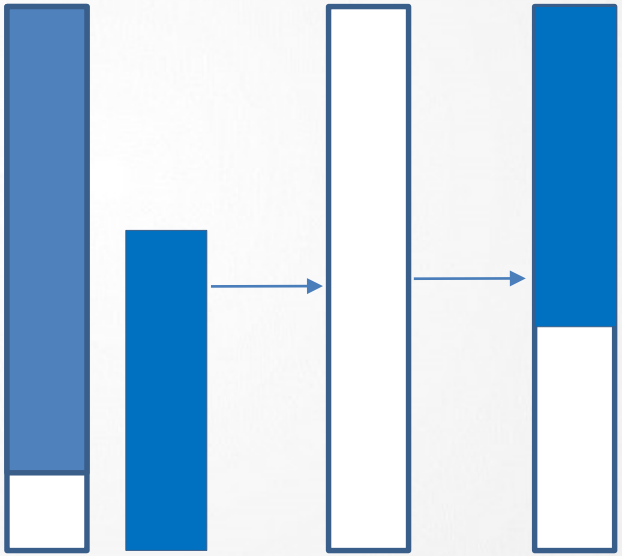
- "NAND page write" is an atomic operation.
- By controlling buffer, ensures every "InnoDB page data" isn't split.

Atomic Write – Buffer Controlling

Sufficient free buffer space



Insufficient free buffer space



Benefits from Atomic Write

- On random write tests
  - TPS increases by 15%
  - SQL write Latency reduces by 30% @99% percentile
  - SSD’s life expectancy doubles

What We have Done.

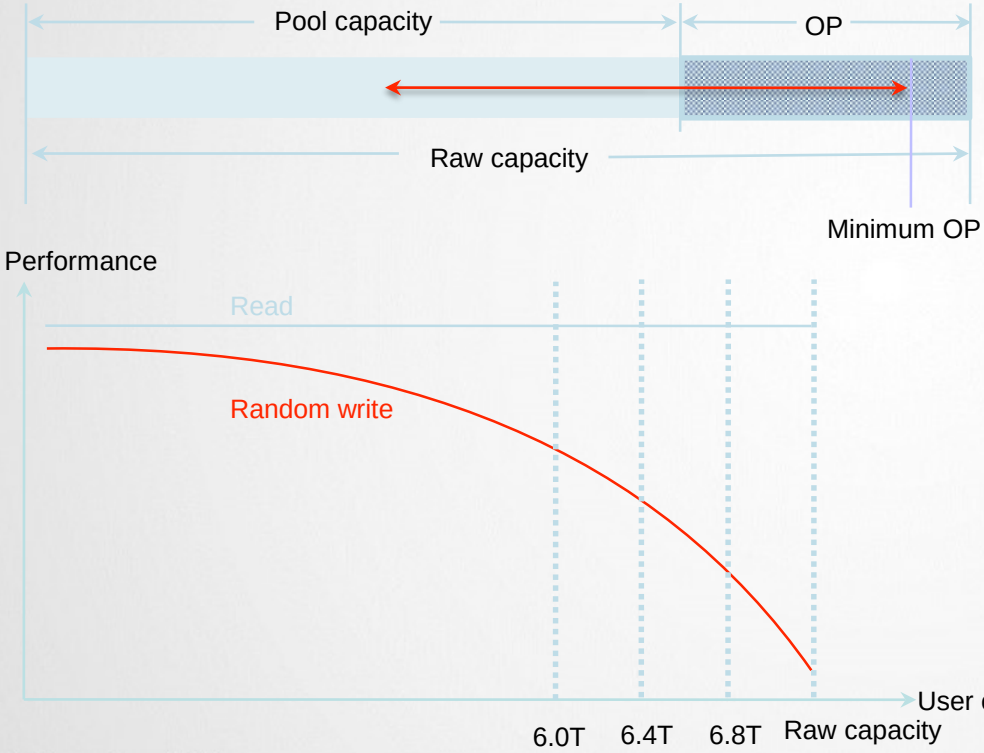
- Atomic write for MySQL/MariaDB
- **Advanced logical volume management**
- Namespaces for I/O isolation
- Multi-stream

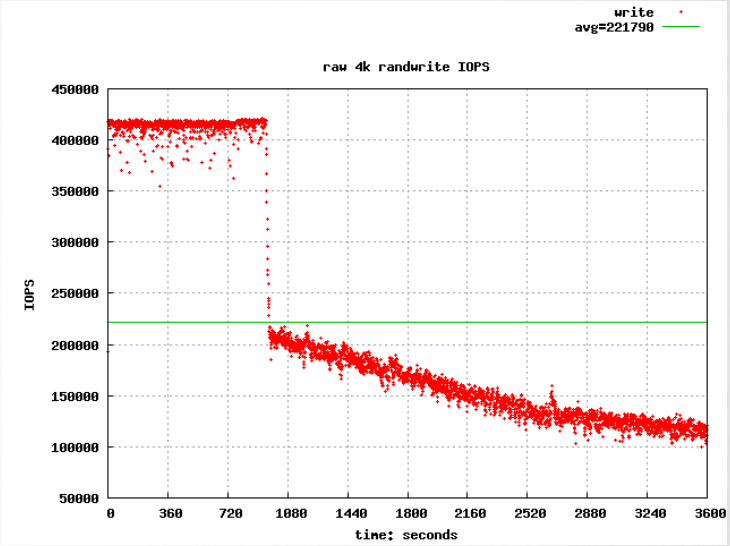
Advanced Logical Volume Management

A lighter and more effective way than LVM+Cgroup

- Merge multiple drives to form a pool(as vgcreate in LVM)
- Adjustable pool performance
- Create multiple logical volumes(as lvcreate in LVM)
- Extend logical volumes(as lvextend)
- Set and change every logical volume’s I/O limits
- Set one volume as high priority

Advanced Logical Volume Management

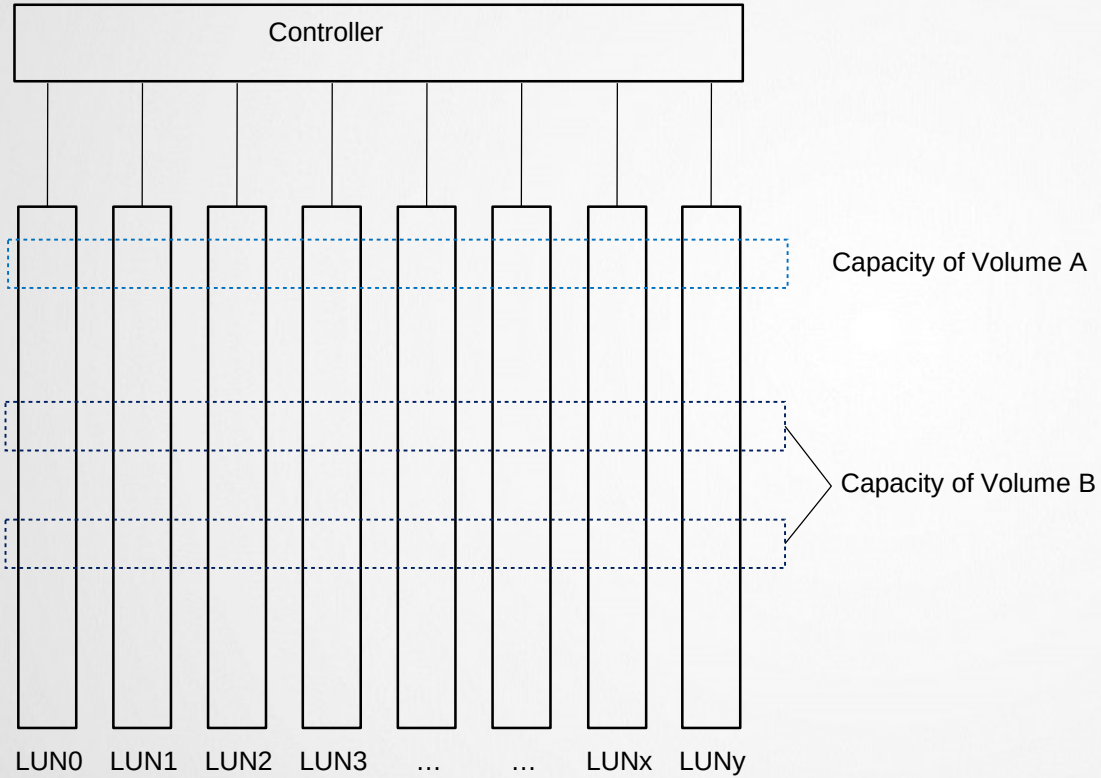




SSD's sustained random write performance depends on over-provision.

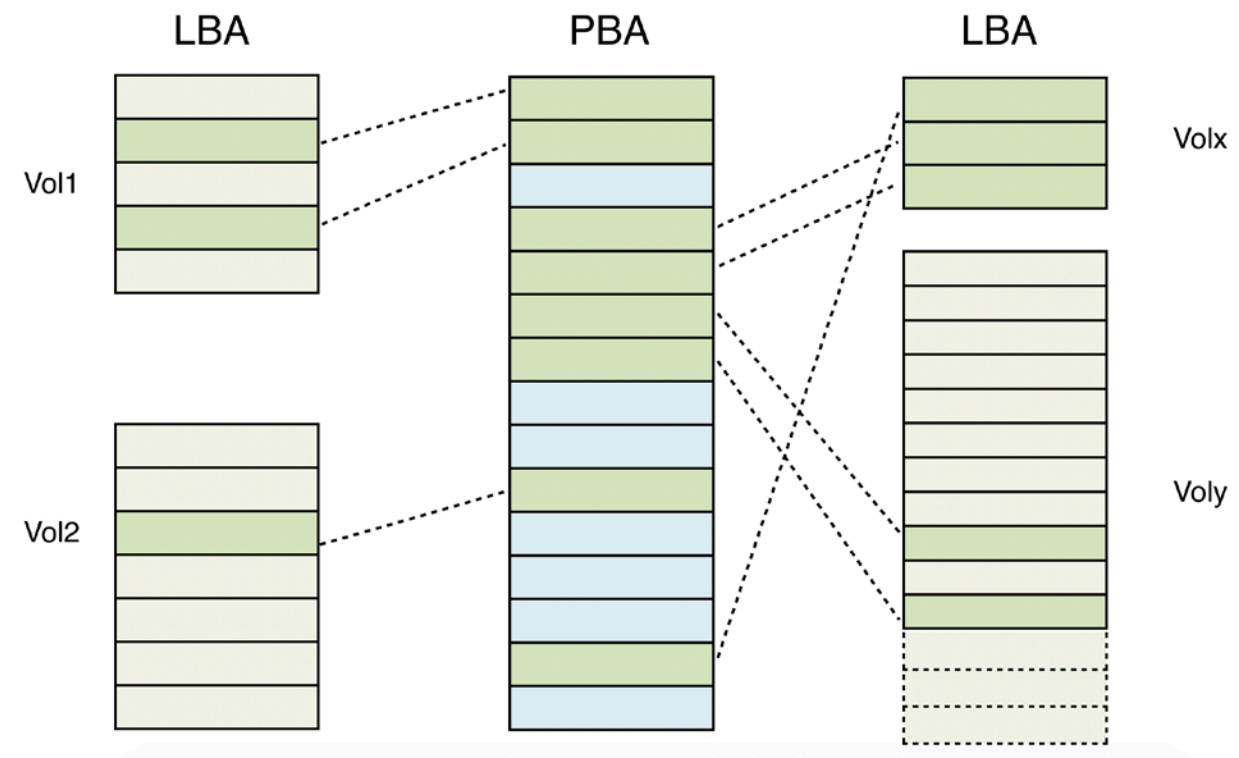
- Adjustable pool performance
  - Find your best cost-effective OP.

Advanced Logical Volume Management



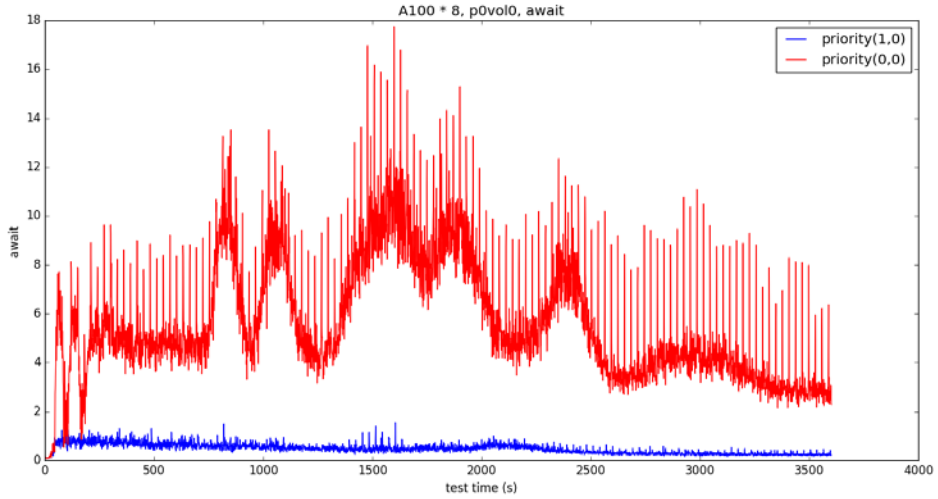
- Allocate capacity of logical volumes
  - Capacity doesn't represent fixed physical address.
  - Use internal counter to implement I/O limits.

Advanced Logical Volume Management



- Thin-provisioning
  - L2P mapping doesn't exist until data is written.
  - Extending volumes' capacity is simply adding some LBAs.

Advanced Logical Volume Management



Priority setting

- High prioritized volume always gets the lowest response time.

Advanced Logical Volume Management

Advanced logical volume management can be applied in:

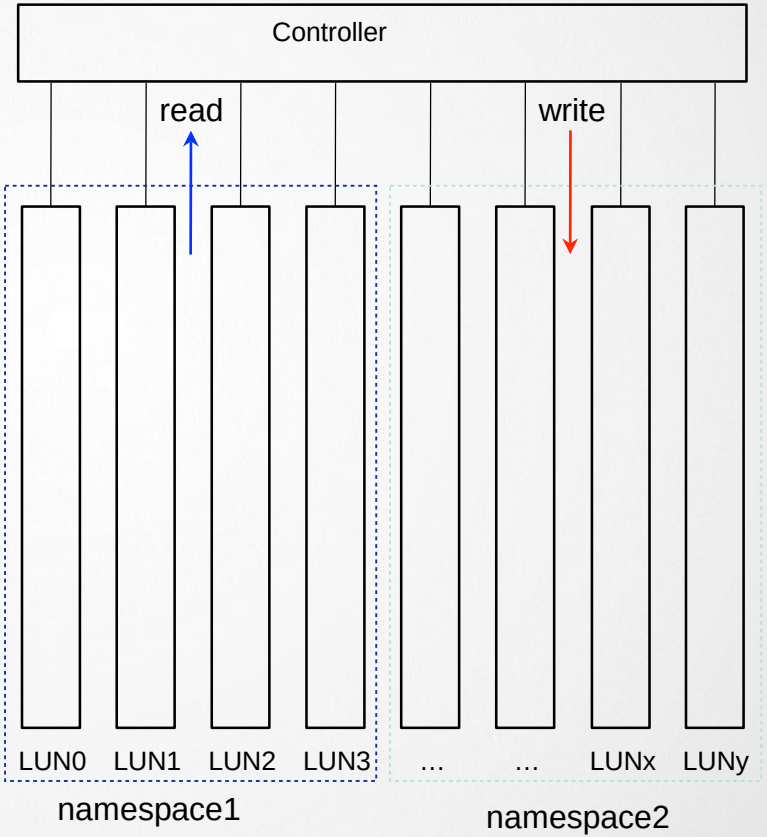
- RDS services:
  - Limit instances' I/O speed
  - Accelerate logs
- EBS services
  - Limit virtual drives' I/O speed
  - Quickly extend virtual drives' capacity

What We have Done.

- Atomic write for MySQL/MariaDB
- Advanced logical volume management
- **Namespaces for I/O isolation**
- Multi-stream

Namespaces for I/O Isolation

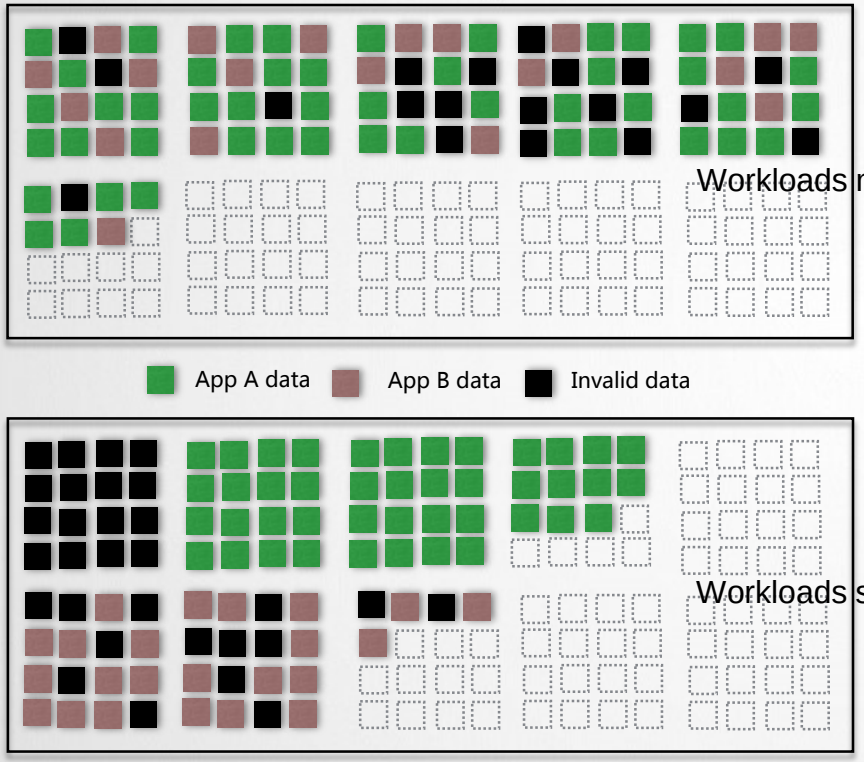
- Namespaces
  - Allocate whole LUNs to form a namespace.
  - Get predictable latency and better QoS.



What We have Done.

- Atomic write for MySQL/MariaDB
- Advanced logical volume management
- Namespaces for I/O isolation
- **Multi-stream**

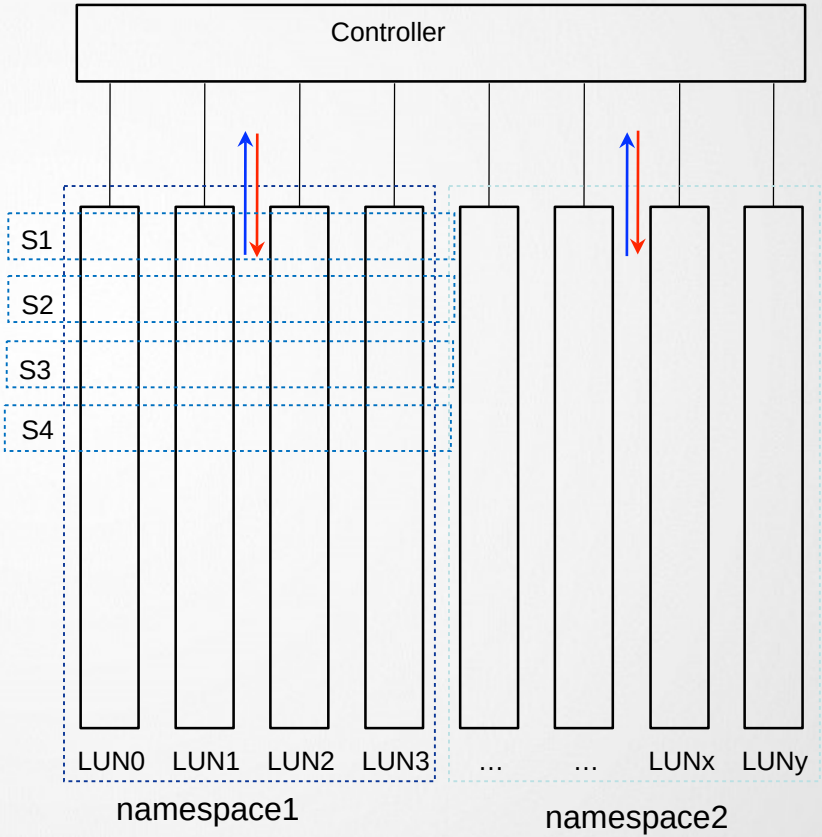
Multi-stream



- Multi-stream
  - Up to 4 streams support currently.
  - Reduce write amplification.
  - Better QoS.

Multi-stream in Namespaces

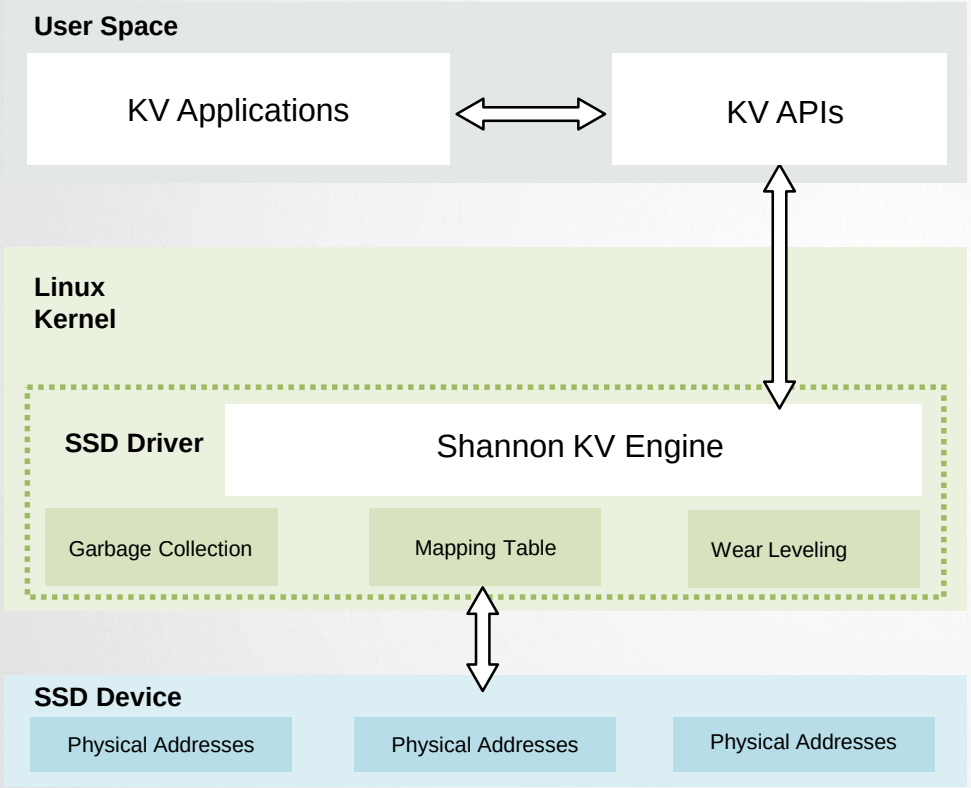
- Multi-stream and namespaces combined
  - Make full use of a single drive



What We have Done.

- Atomic write for MySQL/MariaDB
- Advanced logical volume management
- Namespaces for I/O isolation
- Multi-stream

To Be Continued.



- **Shannon Kernel KV Engine**
  - LevelDB compatible APIs.
  - 10X faster than LevelDB.
  - RocksDB compatible APIs under implementation.

THANK YOU!

Shannon Systems

9F Anlian Building, 168 Jingzhou Road, Yangpu, Shanghai

021-5558-0181

contact@shannon-sys.com

www.shannon-sys.com