Fusion Engine
Next generation storage engine for Flash-SSD and 3D XPoint storage system

Fei Liu, Sheng Qiu, Jianjian Huo, Shu Li

Alibaba Group

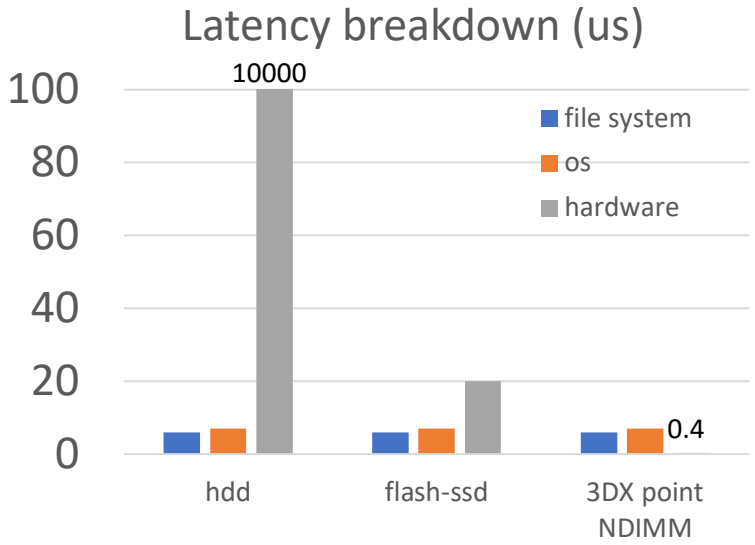
Software overhead become critical

- Legacy file system and OS is heavy !

New storage trend

	Latency
NVM (3DX Point)	400 ns (expected)
Flash-SSD	20 us
HDD	10 ms

Note: numbers come from online third-party websites



Our Solution: Customize storage engine

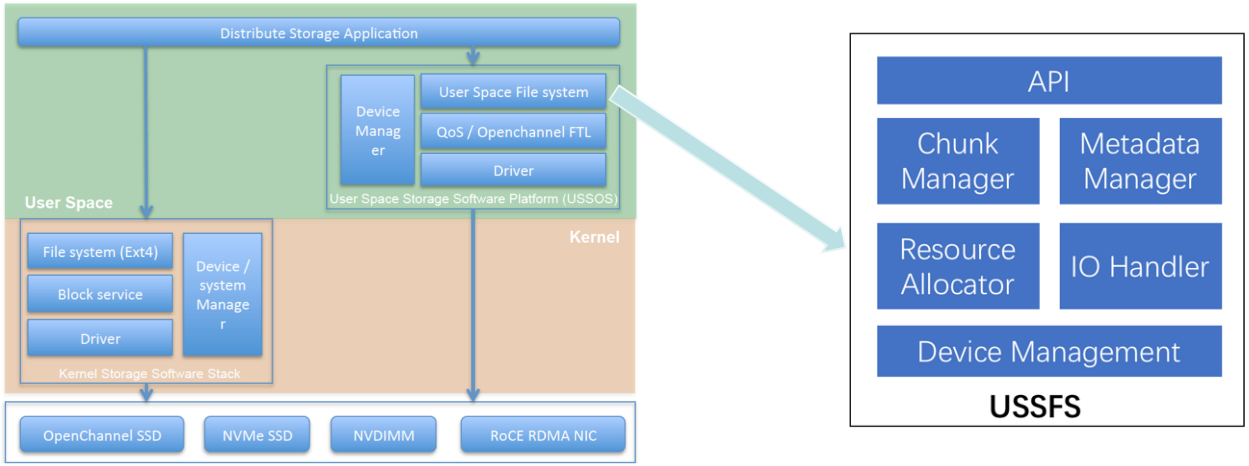
- FusionEngine
  - Customized file-system
  - Customized device management
  - Customized FTL
- Bypass OS kernel

Light-weight FS for NVMe SSD

- Why a light-weight user space file system
  - POSIX semantic incurs unnecessary overhead
  - Kernel bug fixes/upgrades are painful
  - Easy to customize and add new feature
  - Adopt to new hardware

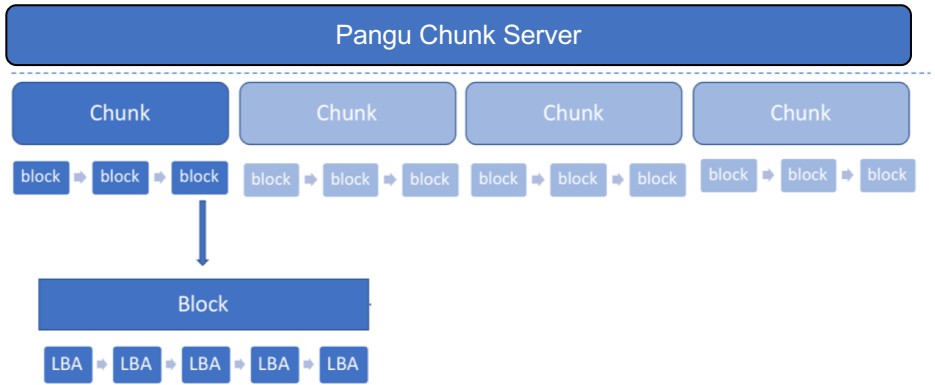
User Space Storage File System (USSFS)

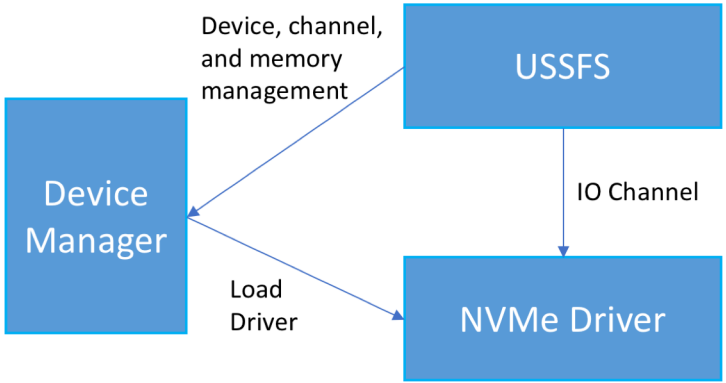
- Light-weight, allocate/read/write/remove chunks
- User space, much easier to upgrade and customize
- Very low latency, make best of modern SSDs
- Enable run-to-completion usage model.



How USSFS works

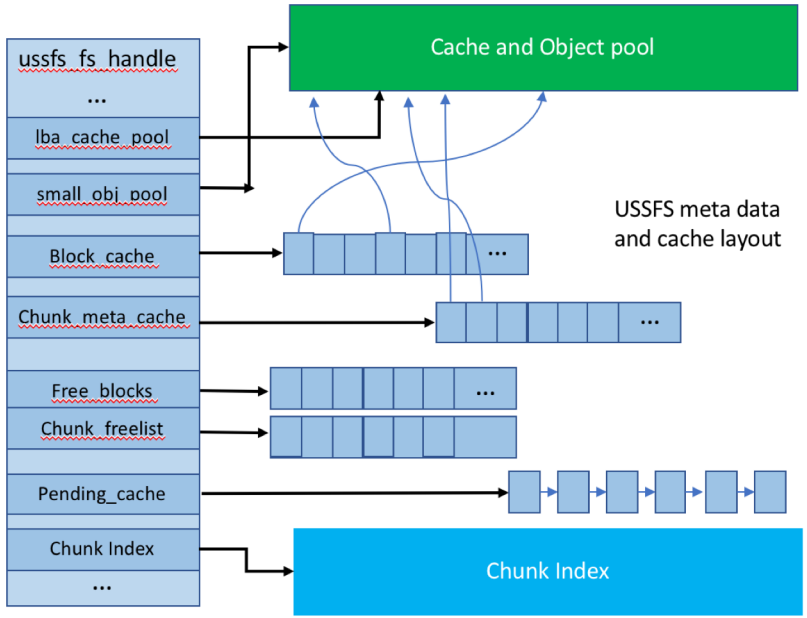
- Separation of control path and data path.
- SPDK user space NVMe driver
- USSOS and device manager





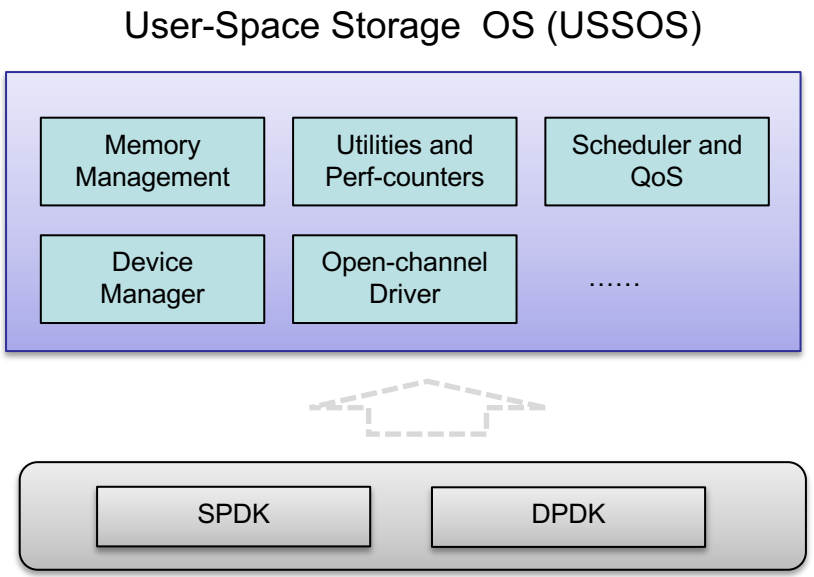
USSFS memory and meta data management

- On-disk:
  - Block table and chunk table
- In-memory:
  - Block cache, chunk cache
  - Chunk index
- DPDK user space memory library based on physical huge pages.



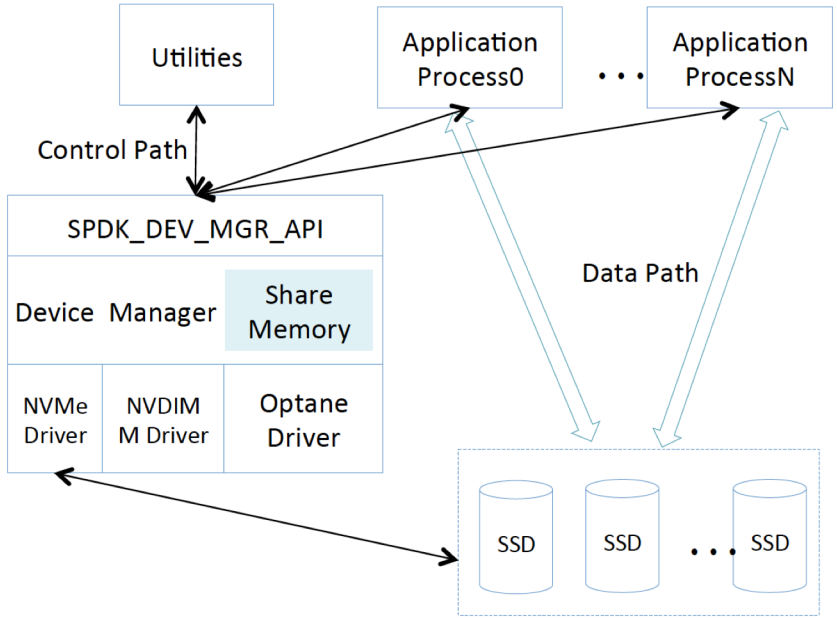
User-space Storage OS

- USSOS
  - Build on top of Intel SPDK/DPDK
  - BUT add tons of new components and features



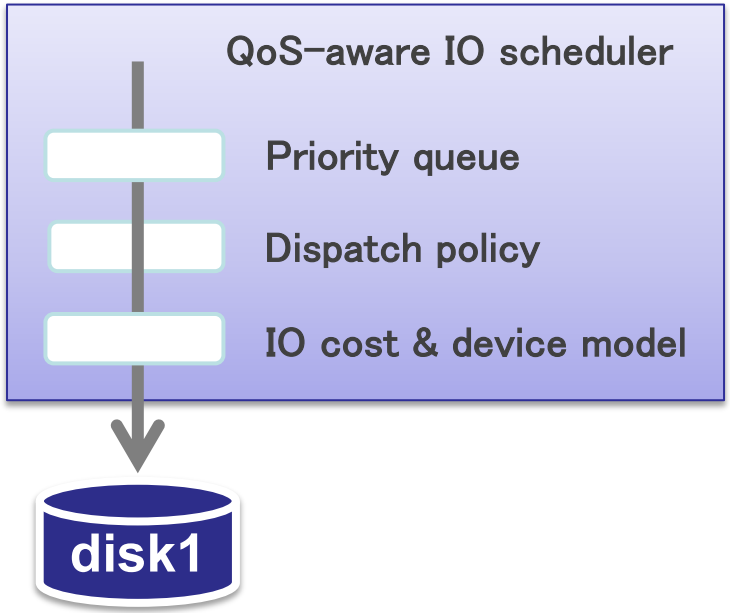
USSOS Device Management

- Manage user space device drivers, NVMe, Optane and etc.
- Manage device, qpair and resources.
- Handle hot plug events, bad disks and etc.



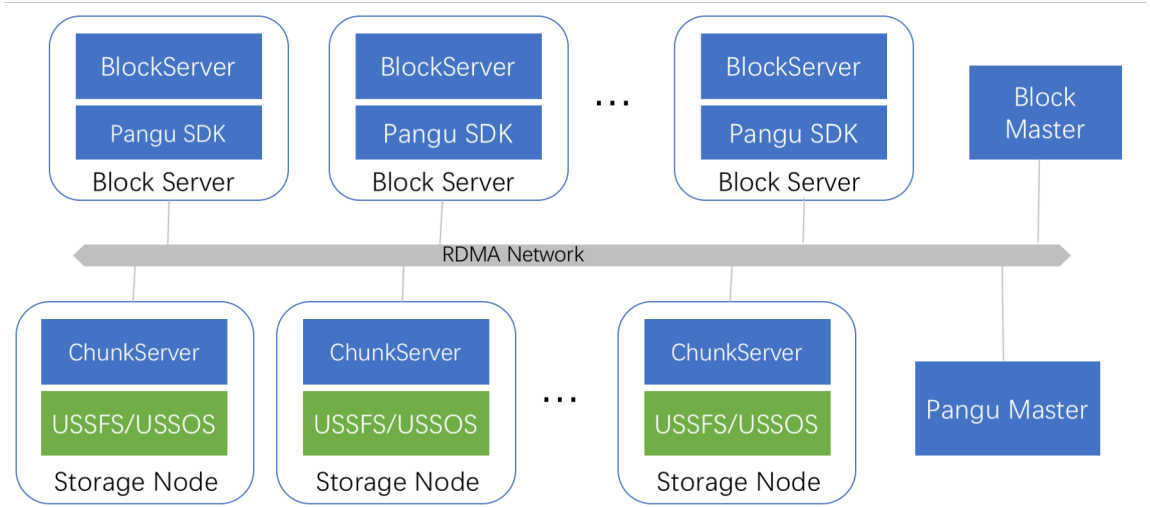
USSOS QoS-aware IO Scheduler

- Priority-aware
  - Dispatch order
  - Bandwidth usage
    - align IO cost with the assigned resource quota
    - Normalize io cost and device capability



Fusion Engine Application in Alibaba Cloud

- Deployed widely inside Alibaba cloud infrastructure
- ESSD cloud disk reaches 1M IOPS, latency at μs scale
- 50% IOPS increase, 30% CPU utilization decrease.

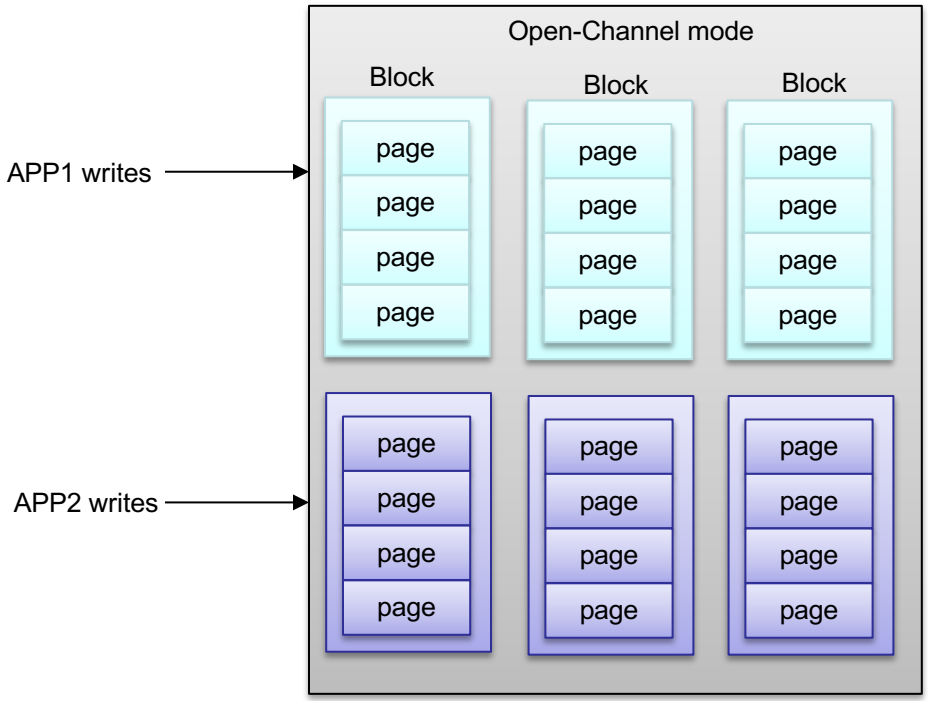


Use due-mode SSD

- Open-channel and legacy mode
- Opportunities
  - Co-optimization
    - Application and storage
  - Flexibility
    - Diversified and fast-changing workloads
  - More…

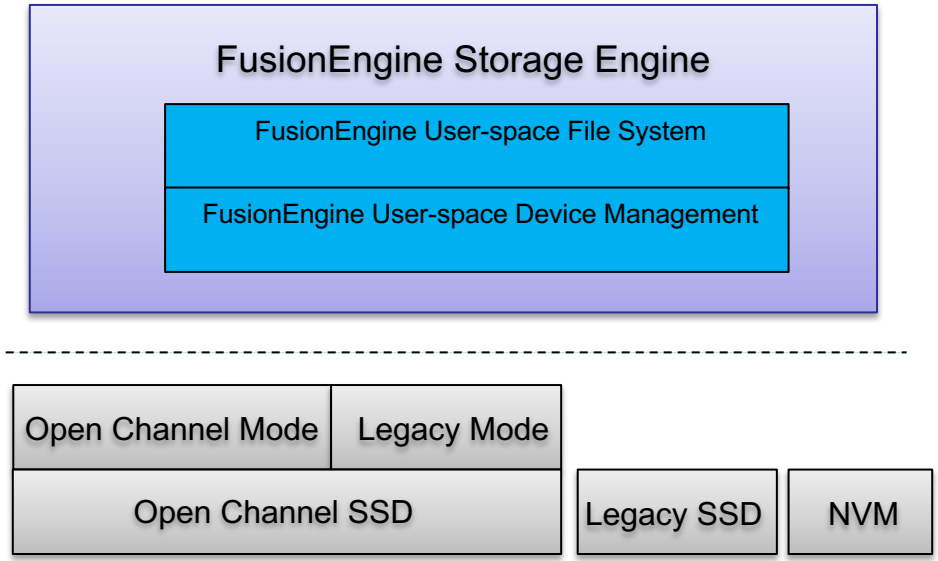
Beyond block interface

- Control data placement
- Control GC



SW/HW integrated storage architecture

- Customized FS
- Customized device management
- Customized FTL
  - Open-channel mode



Conclusion

- Customized SW/HW co-optimization is required
- FusionEngine solves many challenges BUT more are waiting !

Thank You!

- Questions?