Andrey Kudryavtsev

Changing data needs have exposed storage and memory gaps


Intel Optane technology and Intel QLC fill the gaps
- Intel Optane DC Persistent Memory - bigger, more affordable memory
- Intel Optane SSD DC P4800X - break through bottlenecks to increase value of storage data
- Intel QLC 3D NAND SSD - storage consolidation and acceleration

Recommended DC Configuration with Intel DC SSDs

2x vSAN performance for 1/2 the cost
DRAM-like performance for less in MySQL


optimization resources available now [photo]

EDSFF “Ruler” revolutionary form factor
Introducing Intel Memory Drive Technology [photo]
- Use Intel Optane SSD DC P4800X transparently as memory
- Grow beyond system DRM capacity, or replace high-capacity DIMMs for lower-cost alternative, with similar performance
- Leverage storage-class memory today
-- No change to software stack: Unmodified Linux OS, applications, and programming
-- No change to hardware: runs bare-metal, loaded before OS from BIOS or UEFI
- Aggregated single volatile memory pool

*Anu Rao
“Unleash non volatile media performance with SPDK”

Storage Architecture Through The Ages [photo]

Media latency is changing the game http://ieeexplore.ieee.org/document/8003284

SPDK
Scalable and Efficient Software Ingredients
- User space, lockless, polled-mode
- Up to millions of IOPS per core
- Minimise average and tail latencies
- Designed for non-volatile media

Storage Reference Software
- Optimised for latest generation CPUs and SSDs
- Provides future proofing
- Extends to storage virtualisation and networking

Open Source Community
- open source building blocks (BSD licensed)
- faster TTM, fewer resources required

Works with any NVMe SSDs

SPDK provides ~9x improvements in efficiency, allowing cores to do something more valuable

Alibaba use case - http://www.sohu.com/a/127452487_374240

Solarflare - https://apple.news/AAXBoN6fjS52fIMIhxigsEA

https://spdk.io/
Jim Harris
Storage virtualisation and disaggregation with SPDK

VirtIO
- paravirtualised driver specification
- common mechanisms and layouts for device discover, I/O queues, etc
- device types include […]

Vhost
- separate process for I/O processing
- vhost protocol for communicating guest VM parameters
  -- memory
  -- number of virtqueues
  -- virtqueue locations

Sharing SSDs in user space
Typically not 1:1 VM to local attached NVMe SSD
- otherwise just use direct PCI assignment
What about SR-IOV?
- SR-IOV SSDs not prevalent yet
- precludes features such as snapshots
What about LVM?
- LVM depends on Linux kernel block layer and storage drivers (i.e. nvme)
- SPDK wants to use user space polled mode drivers

SPDK Blobstore and Logical Volumes

Blobstore Design - design goals
- Minimalistic for targeted storage use cases like Logical Volumes and RocksDB
- Deliver only the basics to enable another class of application
- Design for fast storage media

Blobstore Design - high level
Application interacts with chunks of data called blobs
- Mutable array of pages of data, accessible via ID
Asynchronous
- No blocking, queuing or waiting
Fully parallel
- no locks in IO path
Atomic metadata operations
- Depends on SSD atomicity (i.e. NVMe)
- 1+ 4KB metadata pages per blob

Logical Volumes
Blobstore plus:
- UUID matter for lvolstore, lvols
- Friendly names
-- love name unique within lvolstore
-- lvolstore name unique within application
- Support for:
  -- thin provisioning
  -- clones / snapshots