SDC 20

BY Developers FOR Developers

Storage Developer Conference
September 22-23, 2020

High-performance SMR Drives with dm-zoned

Dr. Hannes Reinecke
SUSE Software Solutions GmbH

SMR drives and dm-zoned

- SMR drives have two type of zones
  - Random access
  - Sequential write
- Number of sequential write zones are substantially higher than the number of random access zones

SMR drives and filesystems

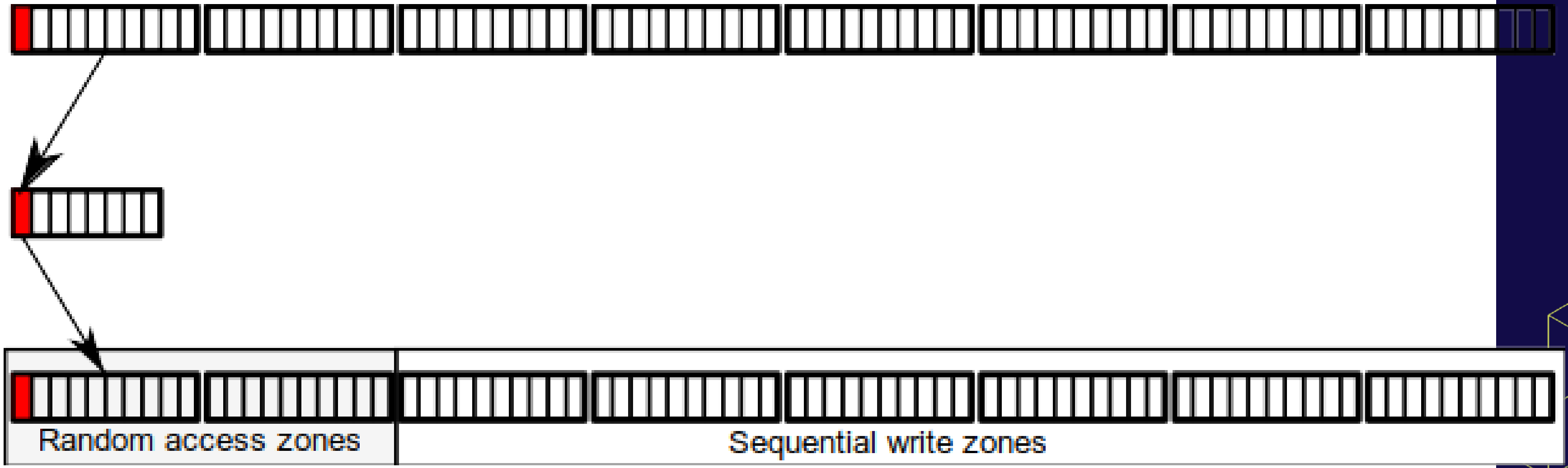
- Regular filesystems assume random access devices
- Modifications required to work natively on SMR drives
- Modifications on filesystems take a long time to be deployed in the field

DM-zoned operation

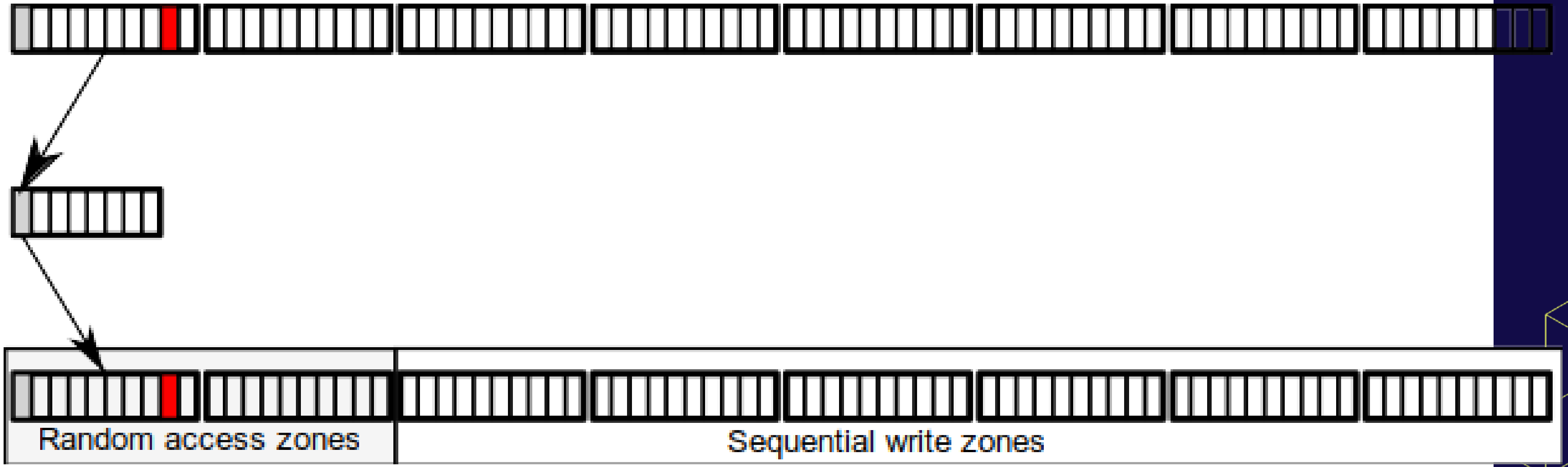
SMR drives and dm-zoned

- Dm-zoned design idea:
  - Use random-access zones to cache data
  - Copy assembled data from random-access zones to sequential write zones
  - Use internal remapping for zones

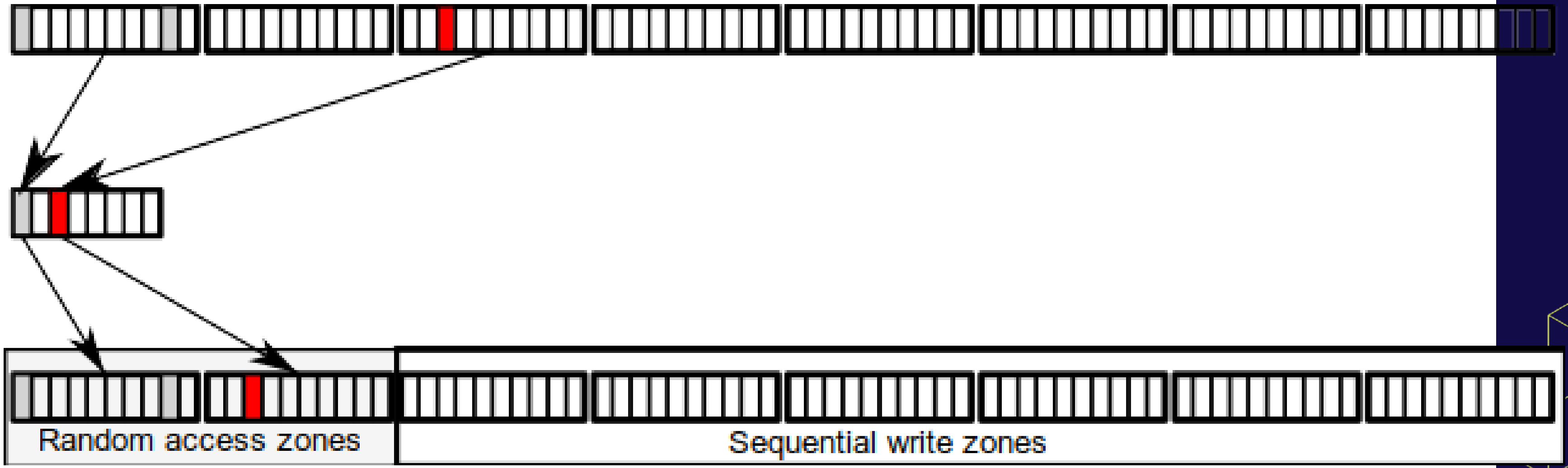
Dm-zoned: map zones



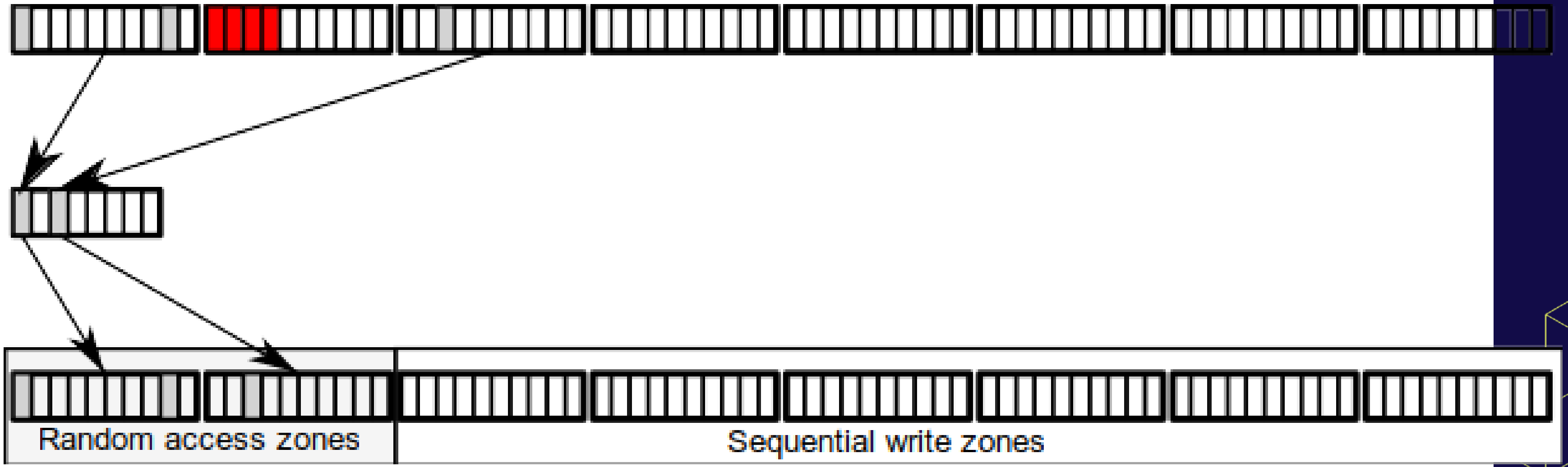
Dm-zoned: map zones



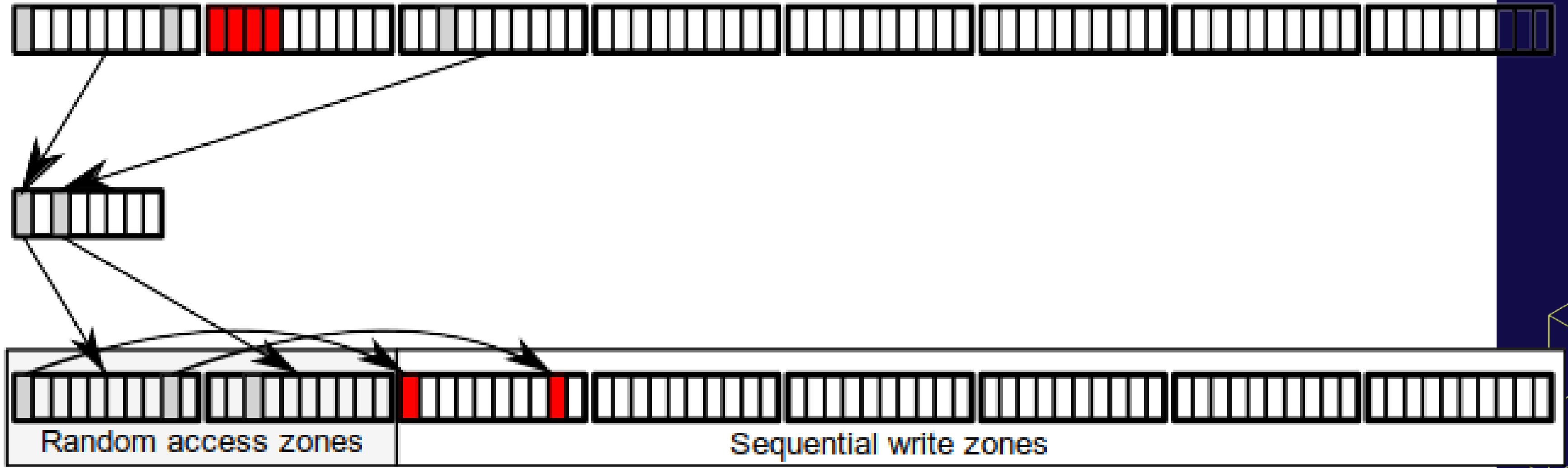
Dm-zoned: map zones



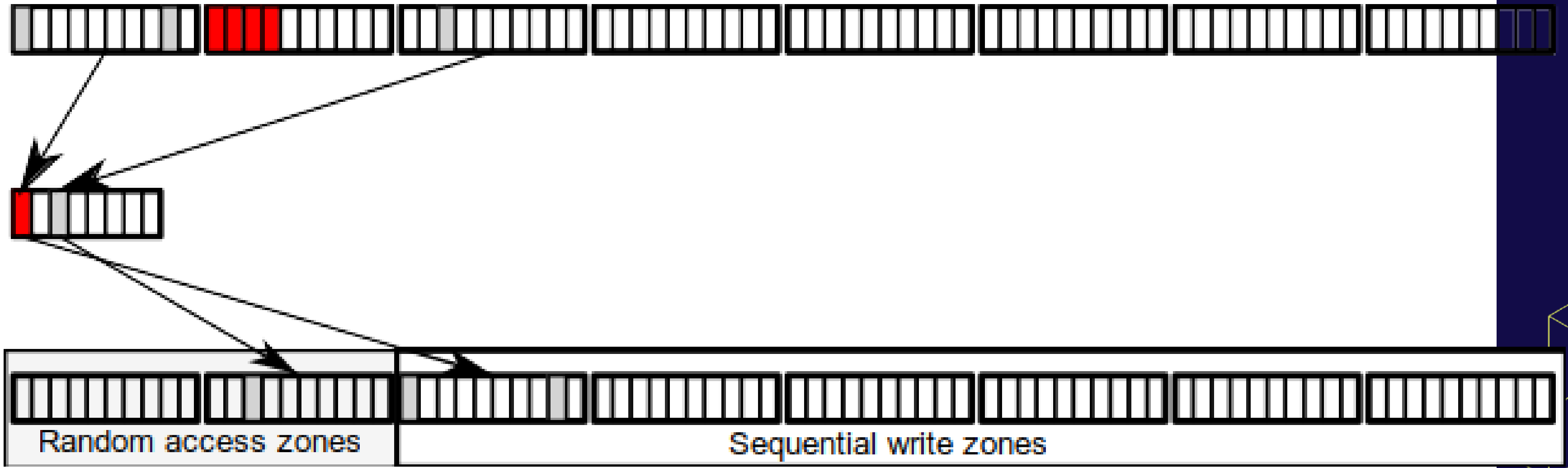
Dm-zoned: copy zones



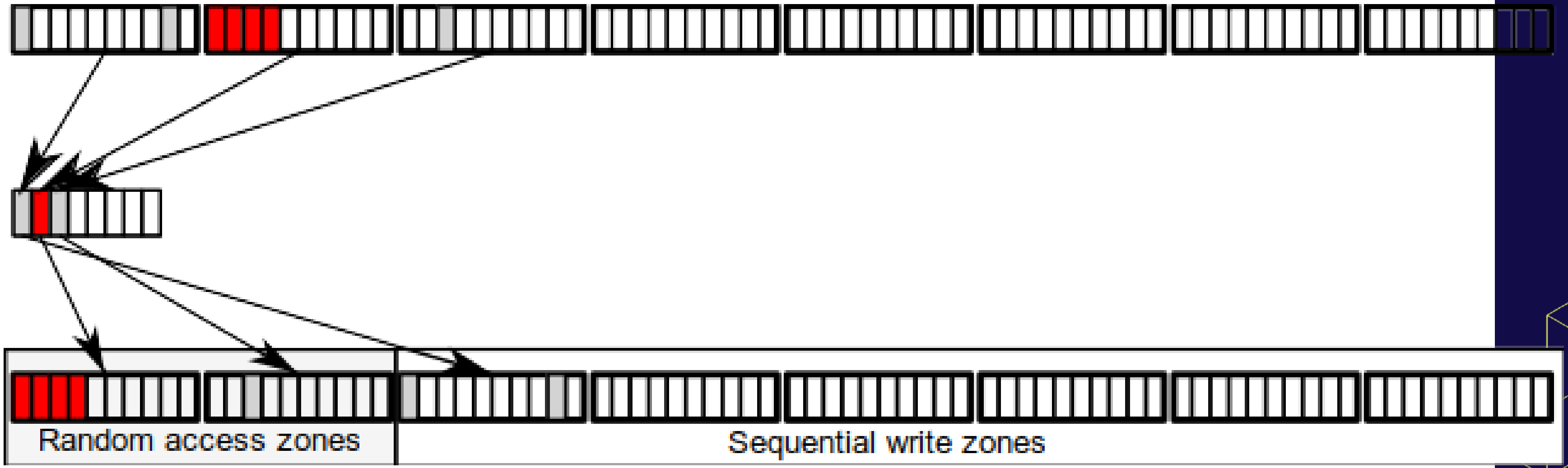
Dm-zoned: copy zones



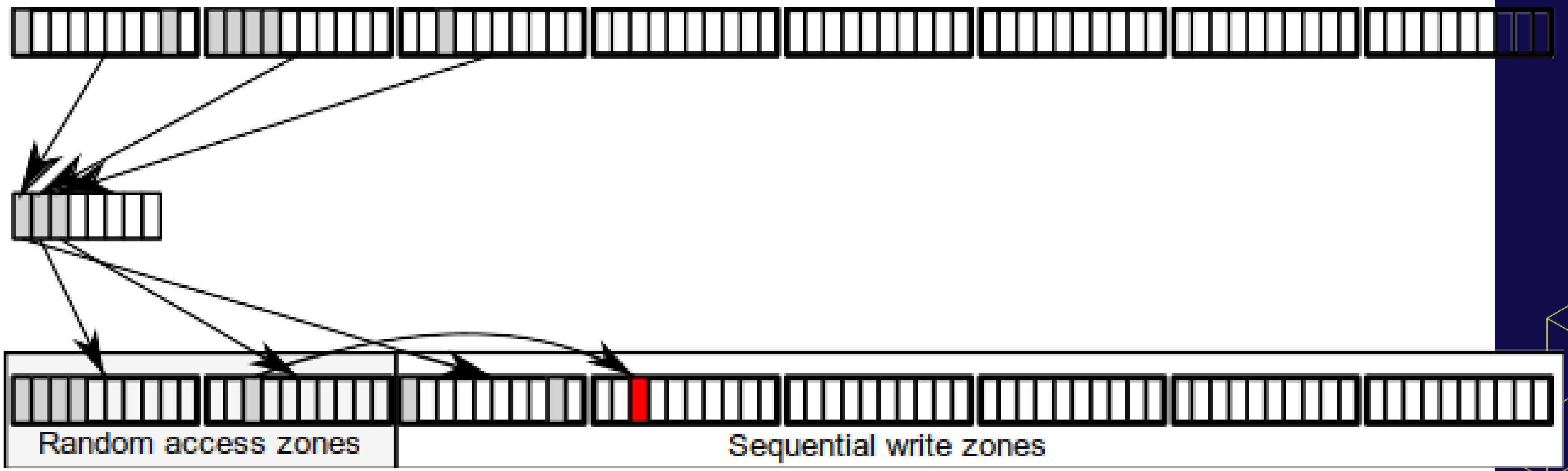
Dm-zoned: copy zones



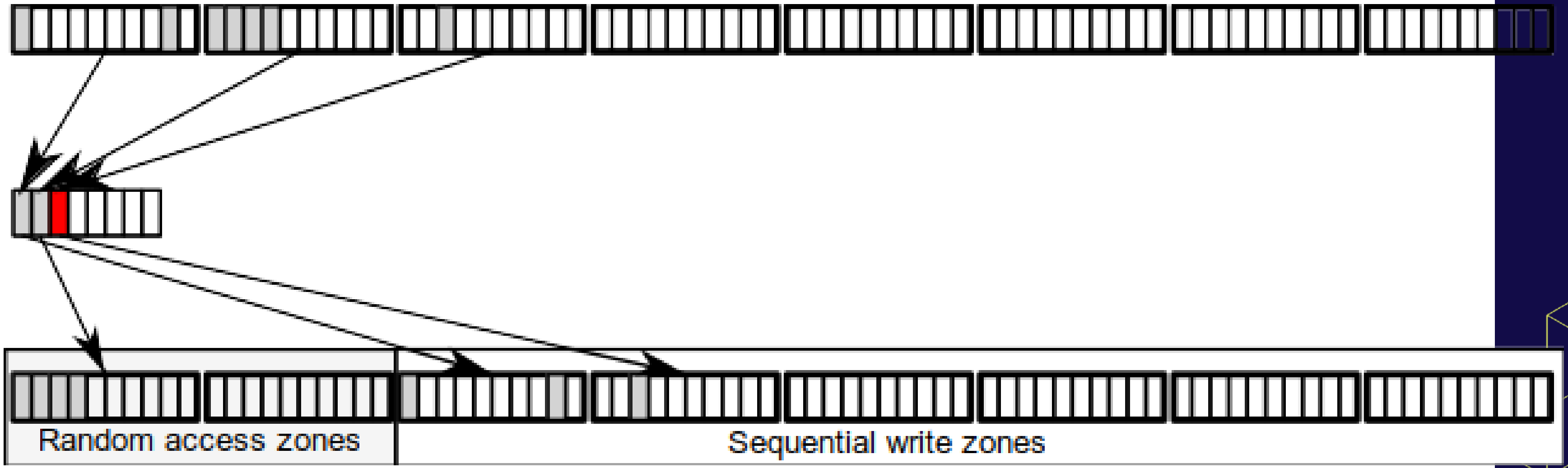
Dm-zoned: copy zones



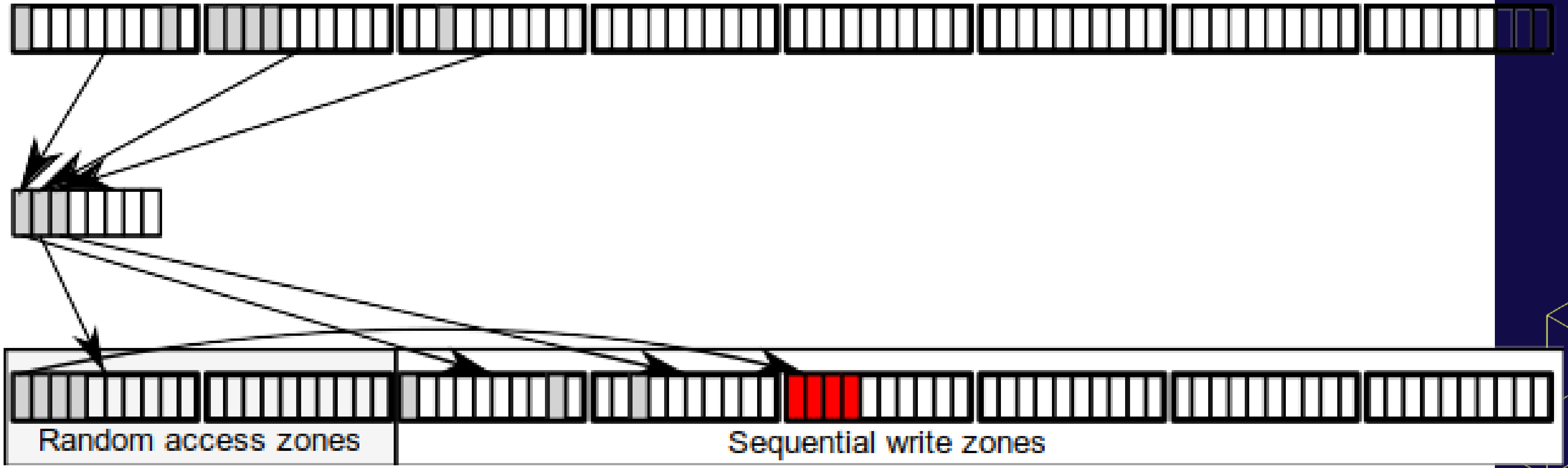
Dm-zoned: reclaim zones



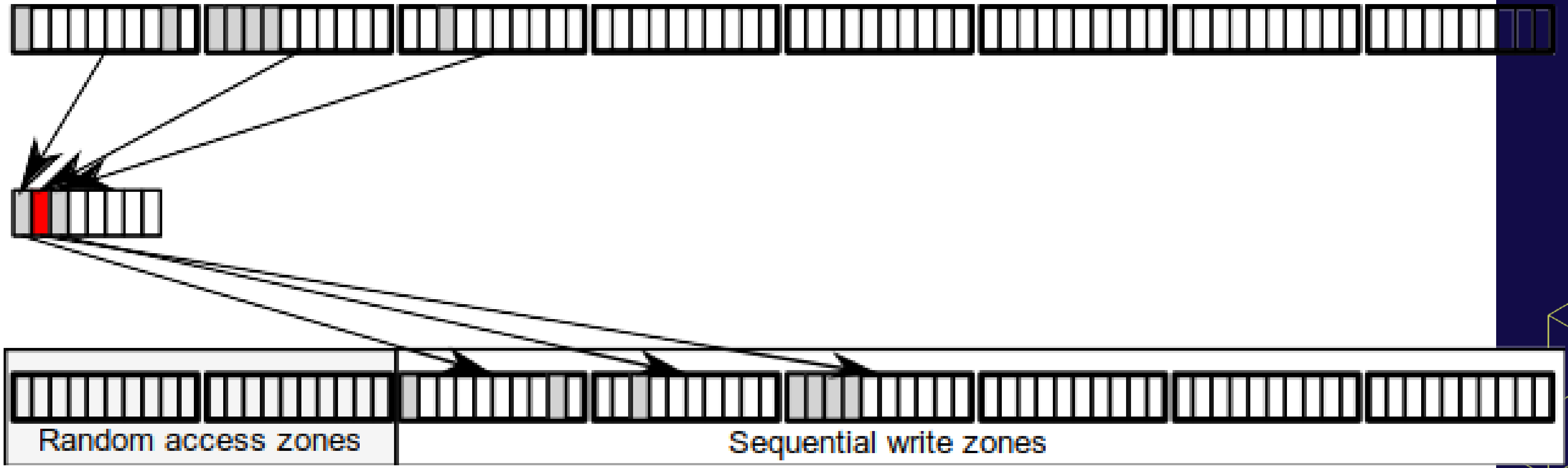
Dm-zoned: reclaim zones



Dm-zoned: reclaim zones



Dm-zoned: reclaim zones



Dm-zoned: cache control

- High watermark:
  - Start reclaim
  - Throttle reclaim
- Low watermark:
  - Always reclaim even if busy
  - Remove throttle on reclaim

Dm-zoned limitations

- Random-access zones have a lower performance than sequential-write zones
- Degrading disk performance during copying zones between random-access and sequential-write zones

Scaling DM-zoned

Design ideas

- Random zones act like a cache, and can live on a separate device
- Sequential-write zones are linear, so we can combine several SMR devices to form a large device
- Device-mapper already provides the infrastructure for such a setup
- Zone mapping can direct I/O to unused disks, thereby improving performance

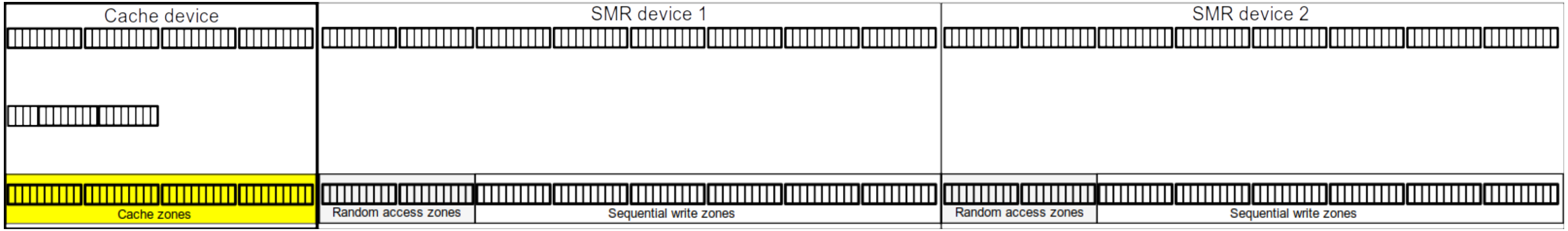
Benefits

- I/O can be directed to unused/less loaded drives
- Cache can be a fast device (NVMe, NVDIMM) to increase burst performance
- Should scale reasonably well as all disks are independent.

Implementation

- Update on-disk metadata to allow several devices
- Update metadata handling:
  - Primary and secondary metadata on cache device
  - ‘Tertiary’ metadata on SMR devices
- Only primary and secondary metadata is updated during I/O, tertiary metadata is just for assembling the device-mapper device.
- Implement cache zones as emulated zones on regular device.

Scaling dm-zoned



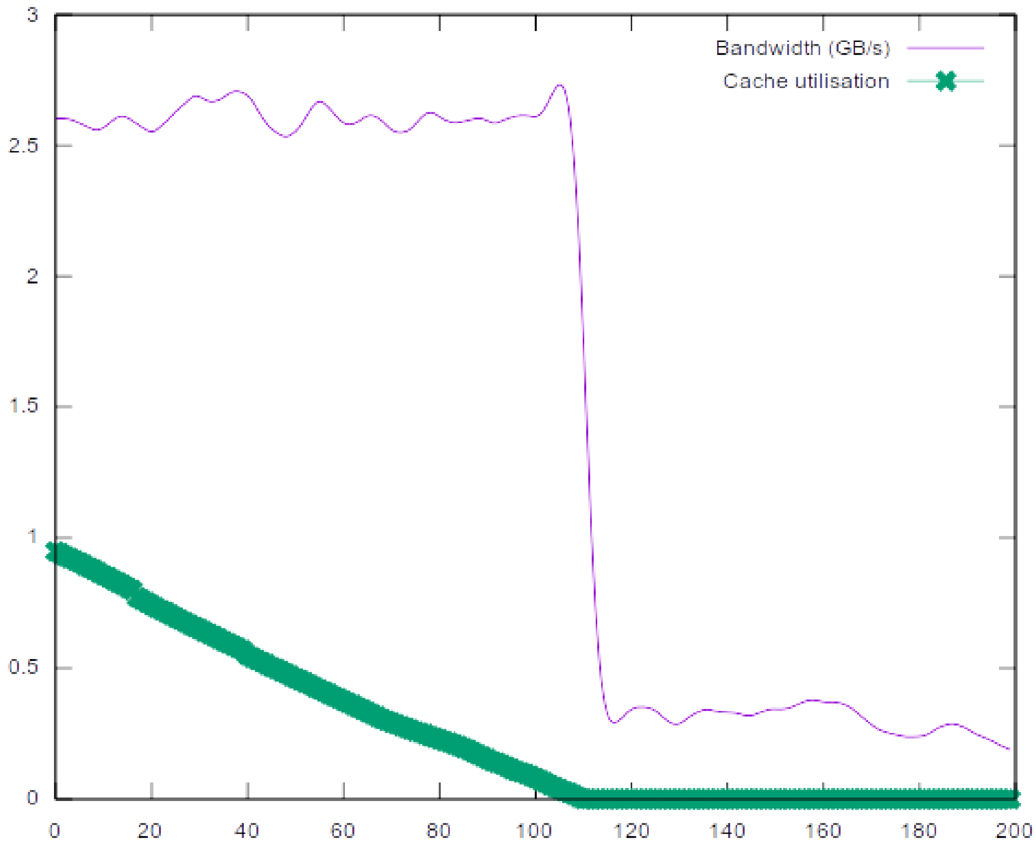
Testing limitations

- SMR support on RAID HBAs very limited
- Broadcom sole remaining vendor of SAS HBAs
- Standard 2U servers can fit up to 12 3.5” HDDs
- Higher disk count require dedicated enclosure
- Limitations due to enclosure connection (6G SAS)

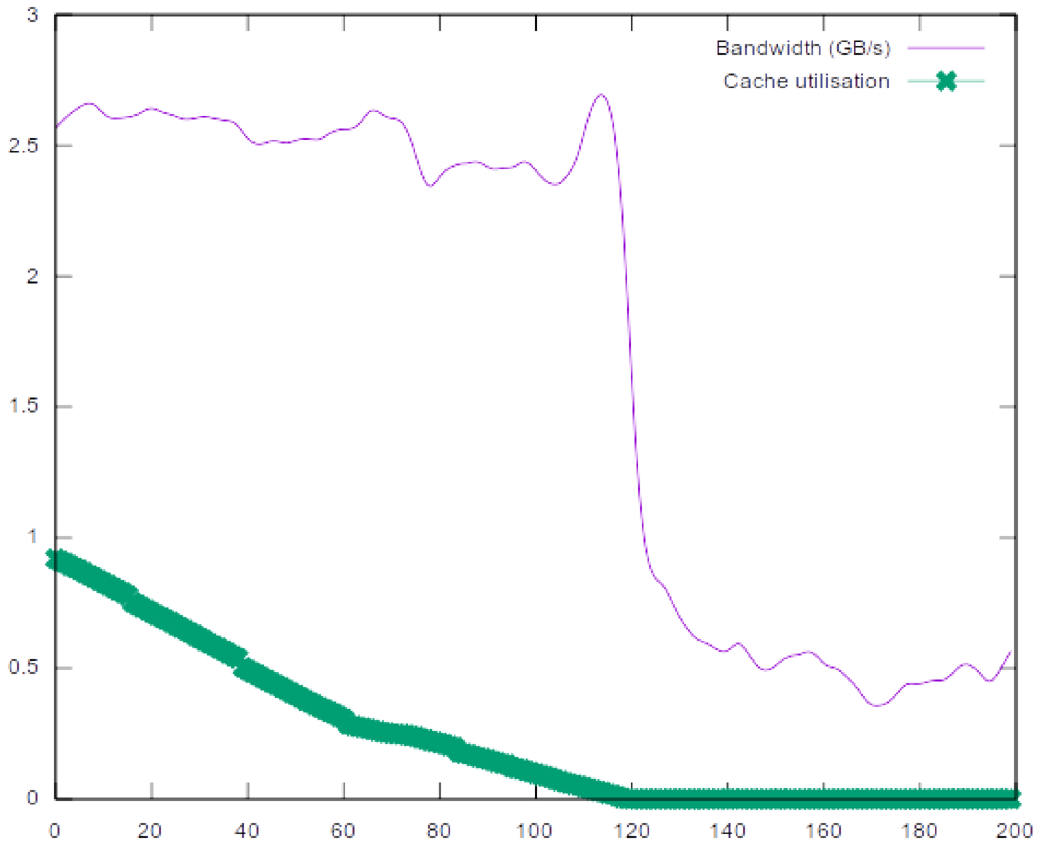
Performance testing

- 20-core dual-socket Intel Xeon
- 128GB RAM
- 256GB NVDIMM as cache
- Broadcom SAS9300-8e connected to JBOD
- 12x WD 14TB SMR drives

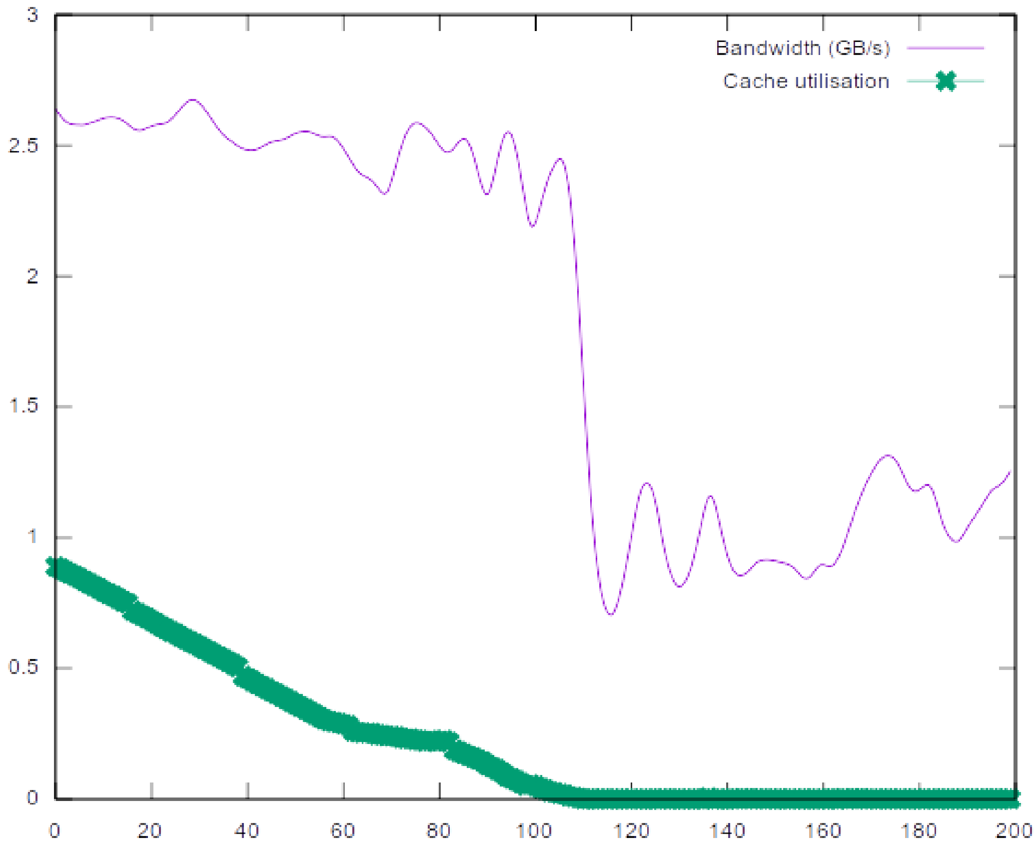
Performance: 2 disks



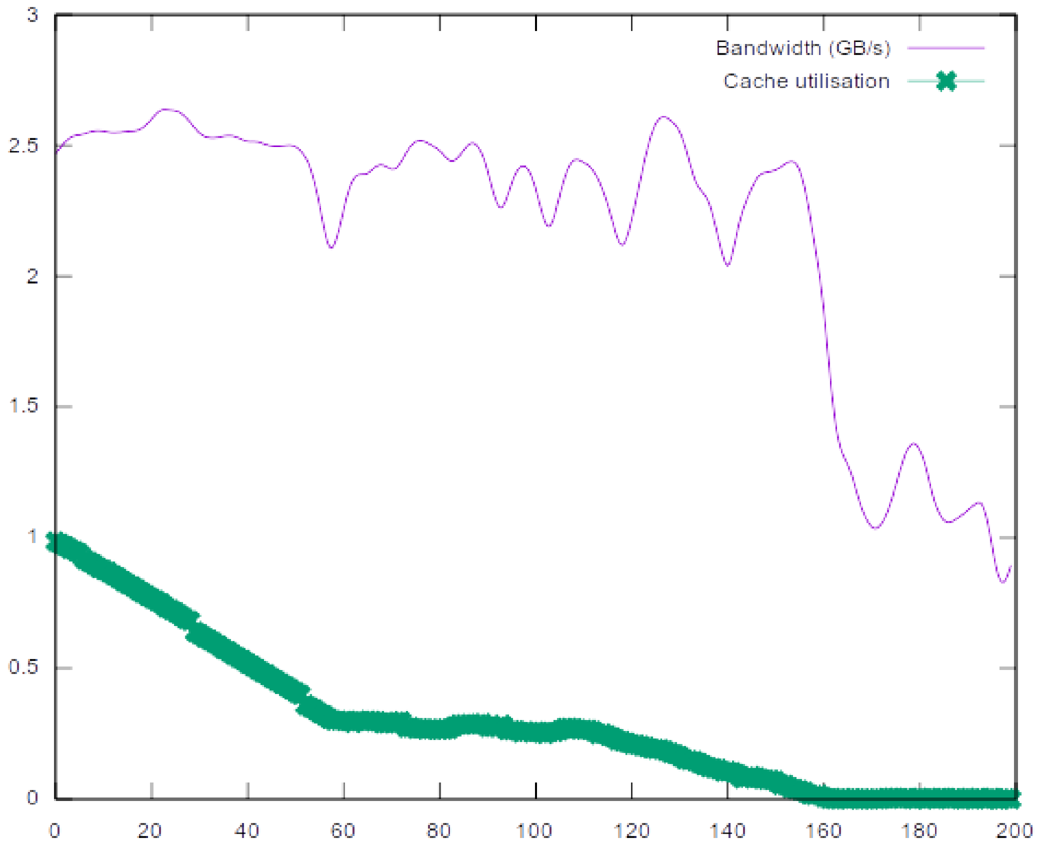
Performance: 4 disks



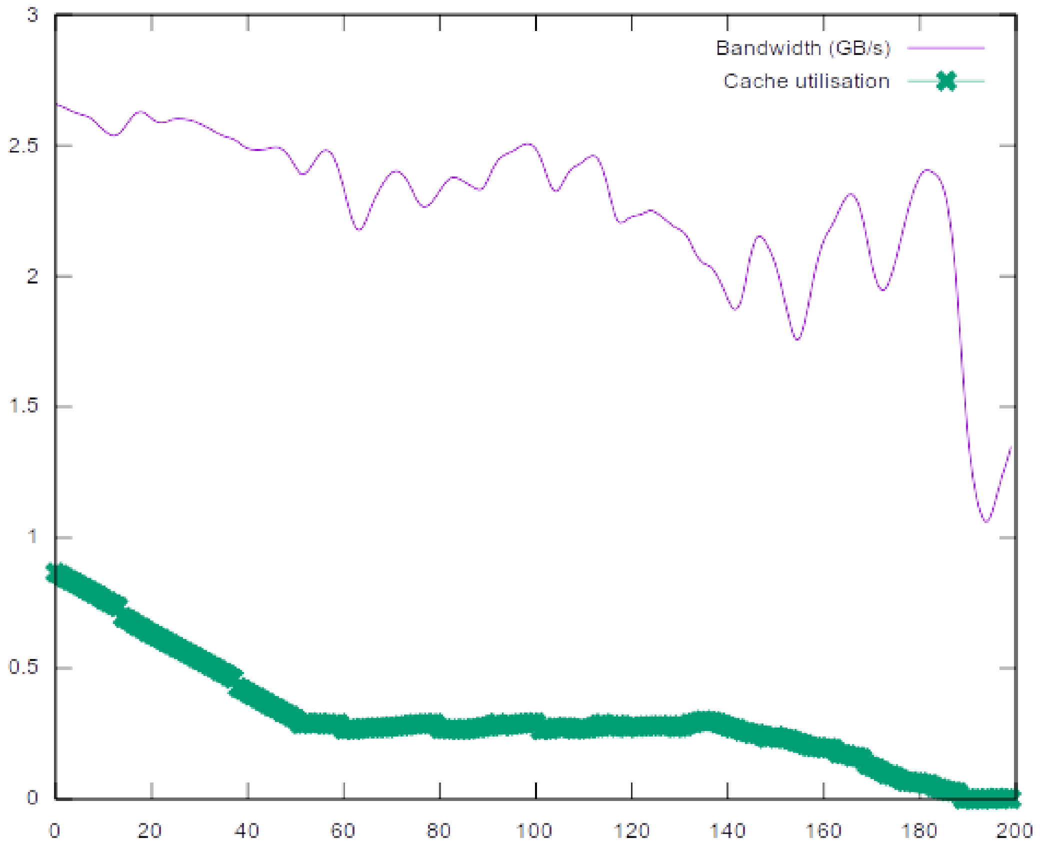
Performance: 6 disks



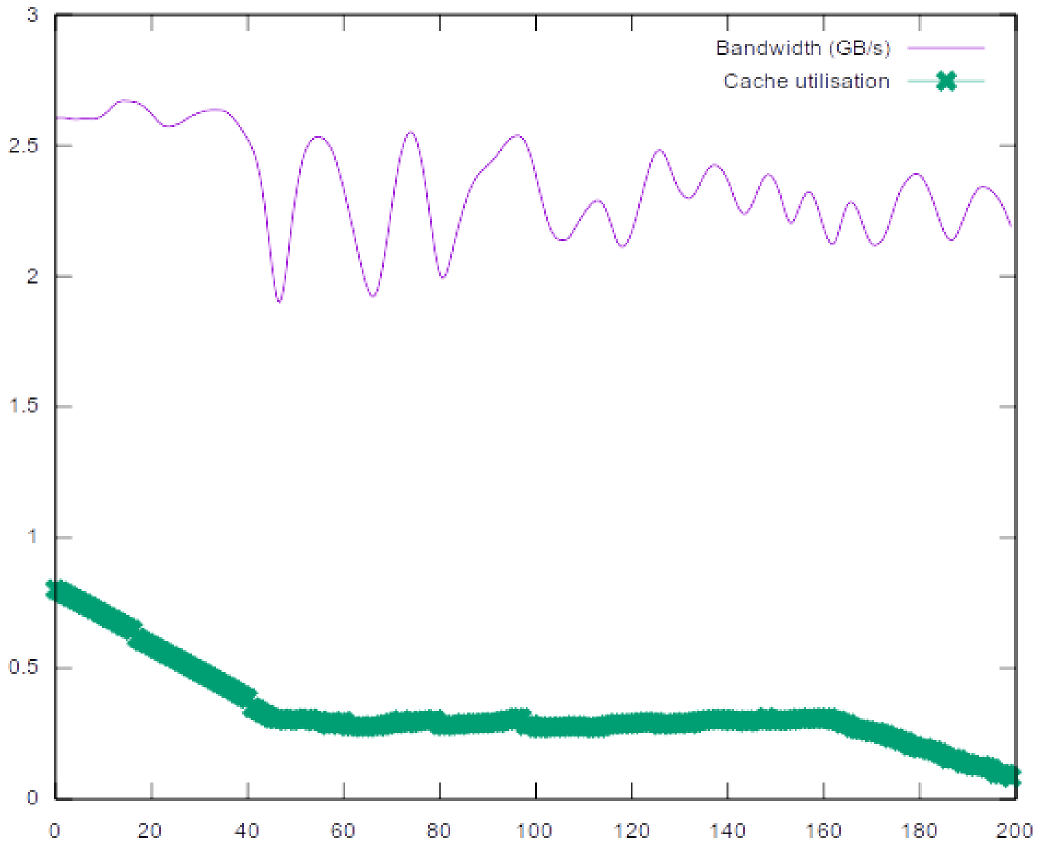
Performance: 8 disks



Performance: 10 disks



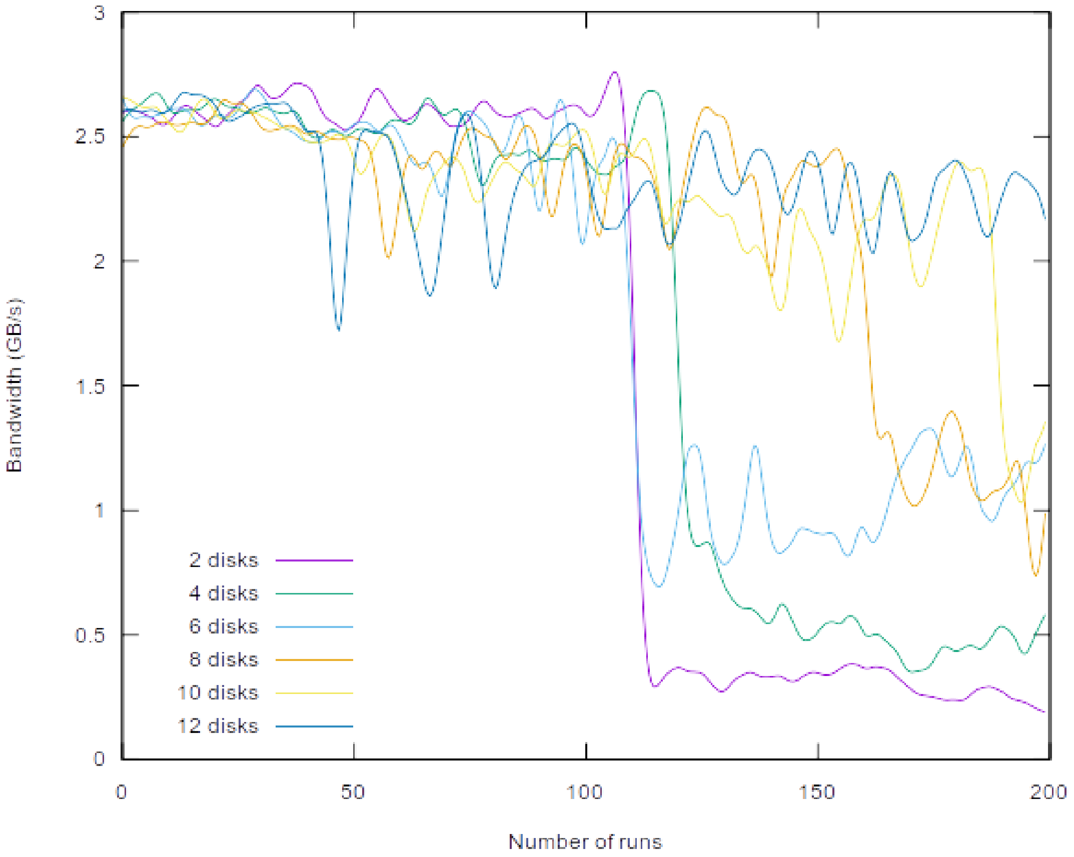
Performance: 12 disks



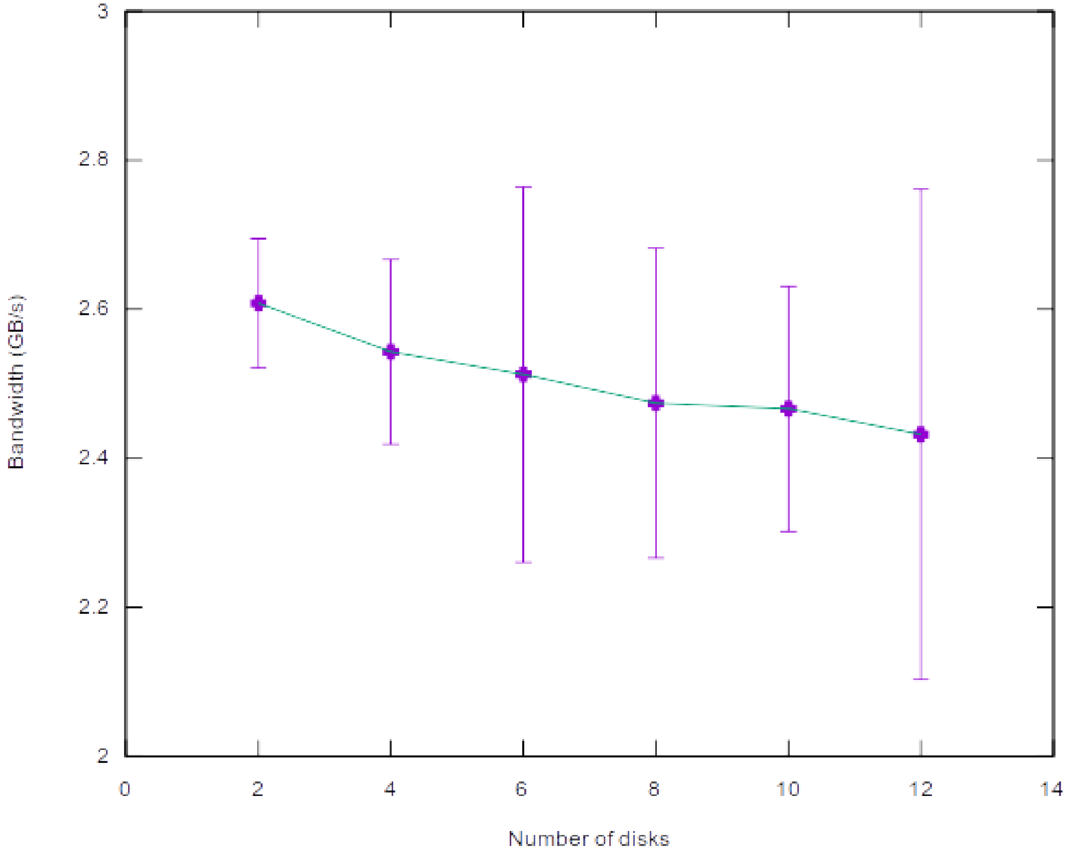
Performance results

- Cache performance around 2.5 GB/s
- Drop in performance once all cache zones are in use
- Performance drop less noticeable with number of disks
- Higher disk counts incur higher performance fluctuation

Scalability effects



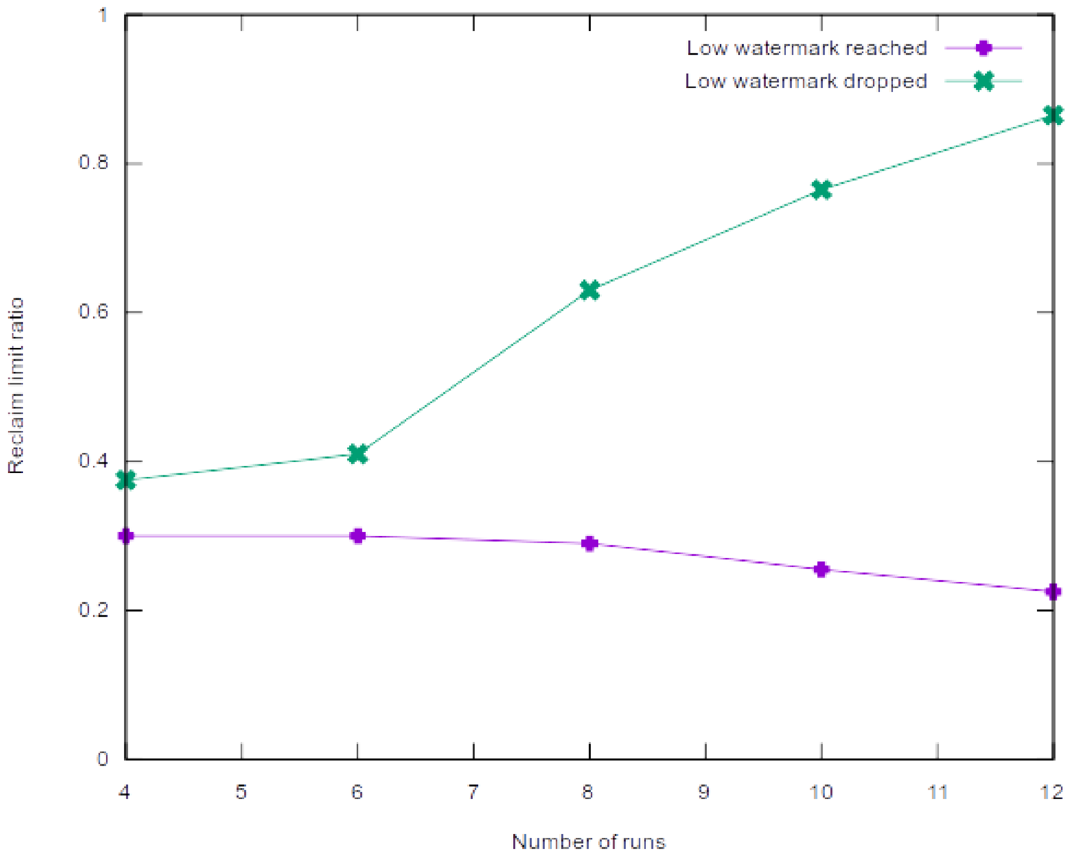
Cache scalability



Cache scalability

- Slight performance degradation (approx. 1.5%) in cache-only performance with number of disks
- Possible interaction with reclaim
- Still very good scalability

Reclaim scalability



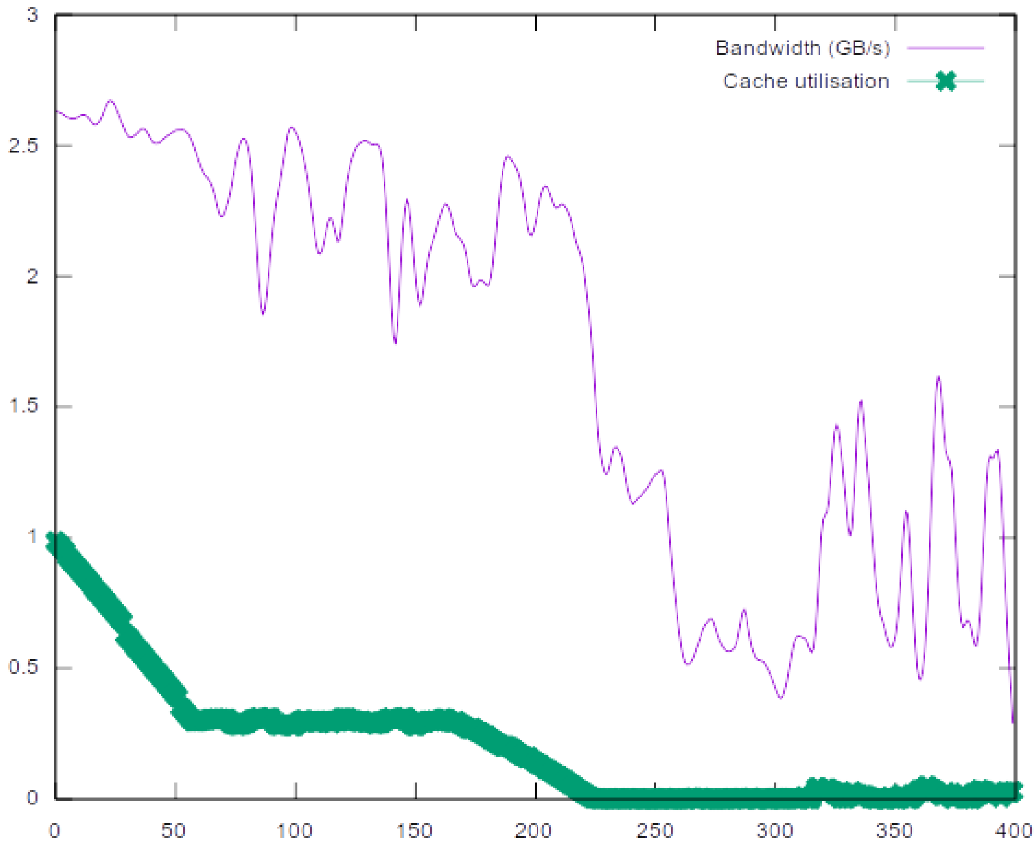
Reclaim scalability

- Reclaim is scaling with number of disks
- Start earlier with higher number of disks
- Longer period at low watermark with higher number of disks
- Becomes more ‘aggressive’ with higher number of disks

Performance on high disk counts

- Drop on performance barely noticeable on higher disk counts
- Not all cache zones have been used with 12 disks
- Retest with larger number of runs to get comparable results

Performance: 12 disks



Performance on cache saturation

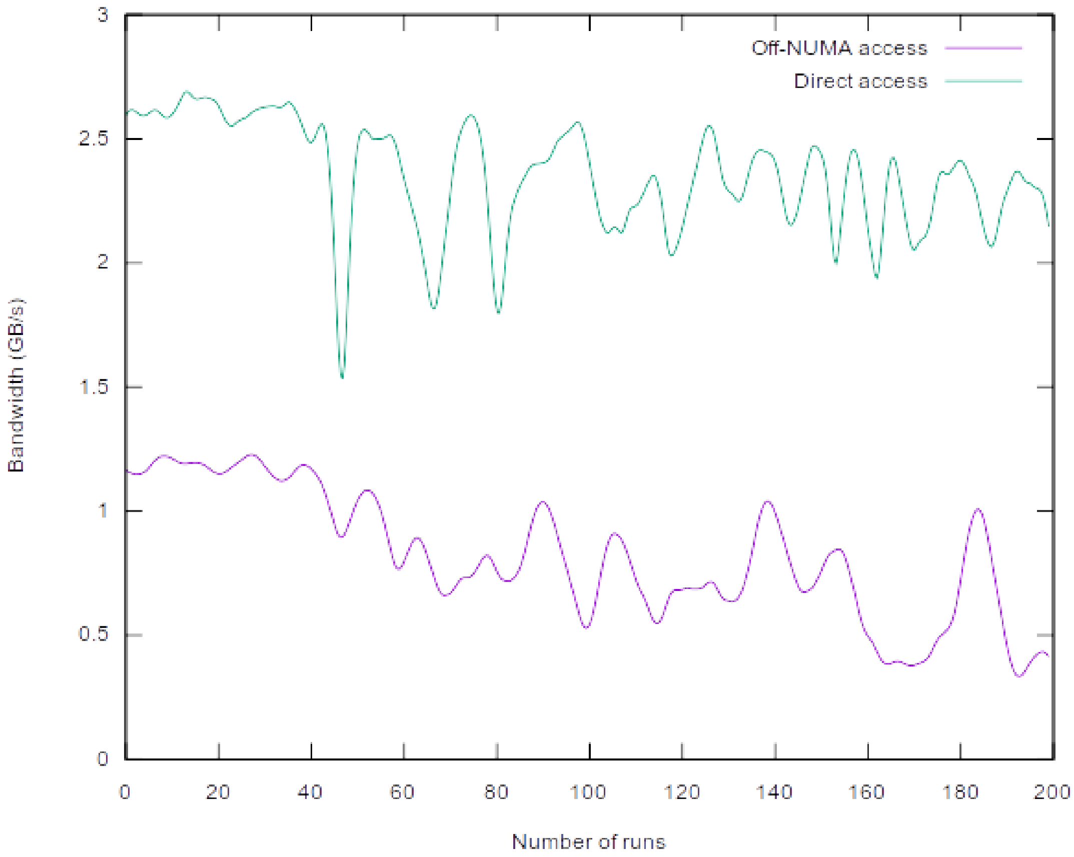
- Performance increases with number of disks
- Fluctuation increases with number of disks
- Resource contention on the HBA:
  - All drives are behind a 6G SAS HBA
  - Limited number of tags available

NUMA Effects

NUMA effects on NVDIMM

- Single namespace attached to one socket
- NUMA access from the other socket
- Performance degradation when accessing namespace from other socket:

Performance: 12 disks



NUMA effects on NVDIMM

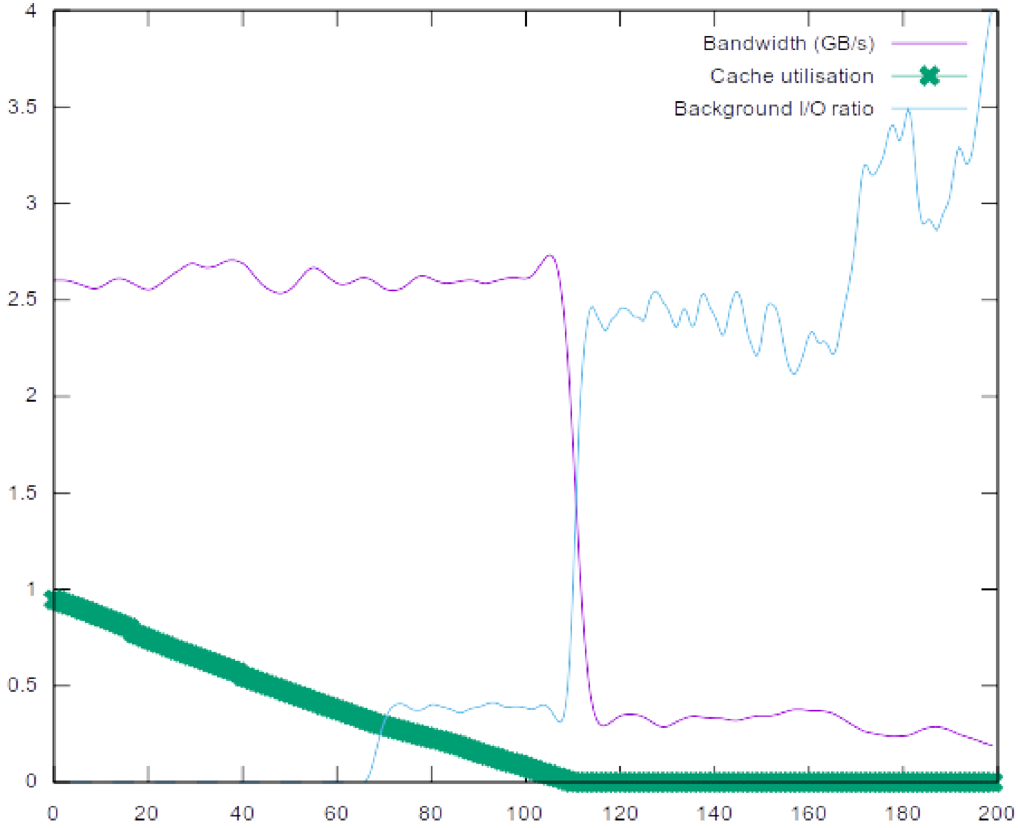
- Performance drop of 50% for Off-socket NUMA access
- Noticeable lower variance for Off-Socket NUMA access
- Might be explained by reclaim running on an off-socket CPU

Write amplification

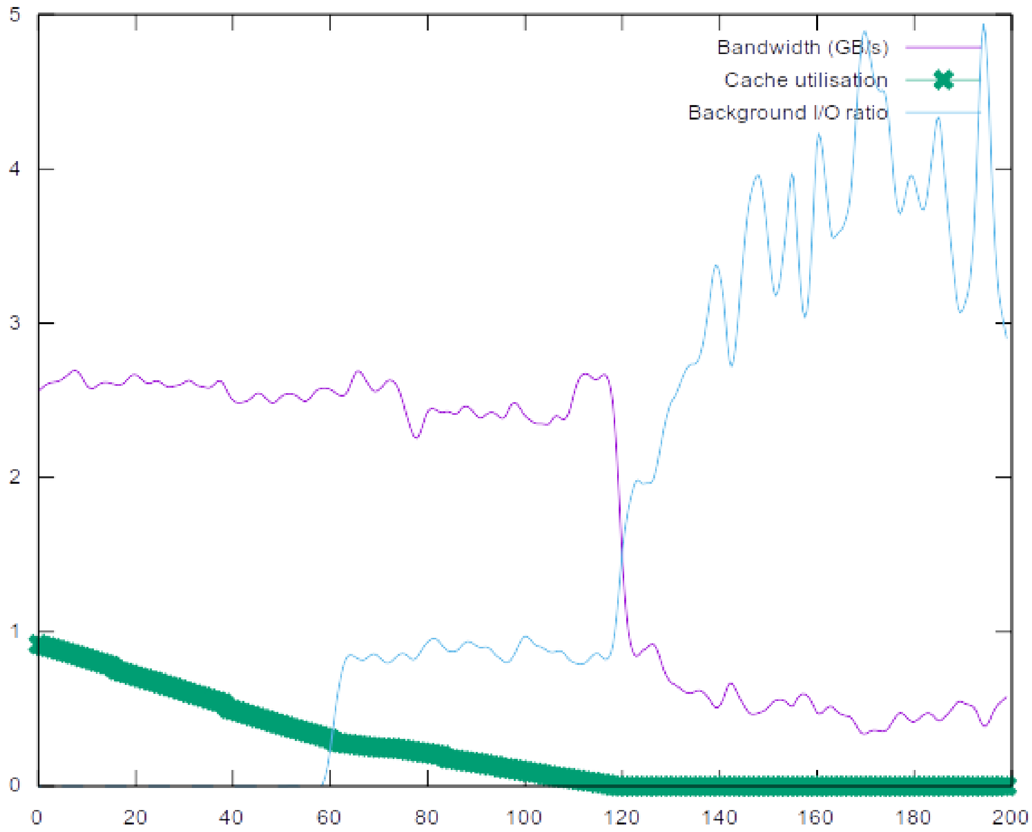
Write amplification

- Cache algorithm induces write amplification
  - Copy contents from to sequential zones on ‘copy’ or ‘reclaim’ operation
  - Read-in required for modification
  - 1:3 worst-case behaviour (write out old contents, read in new contents, write back new contents)

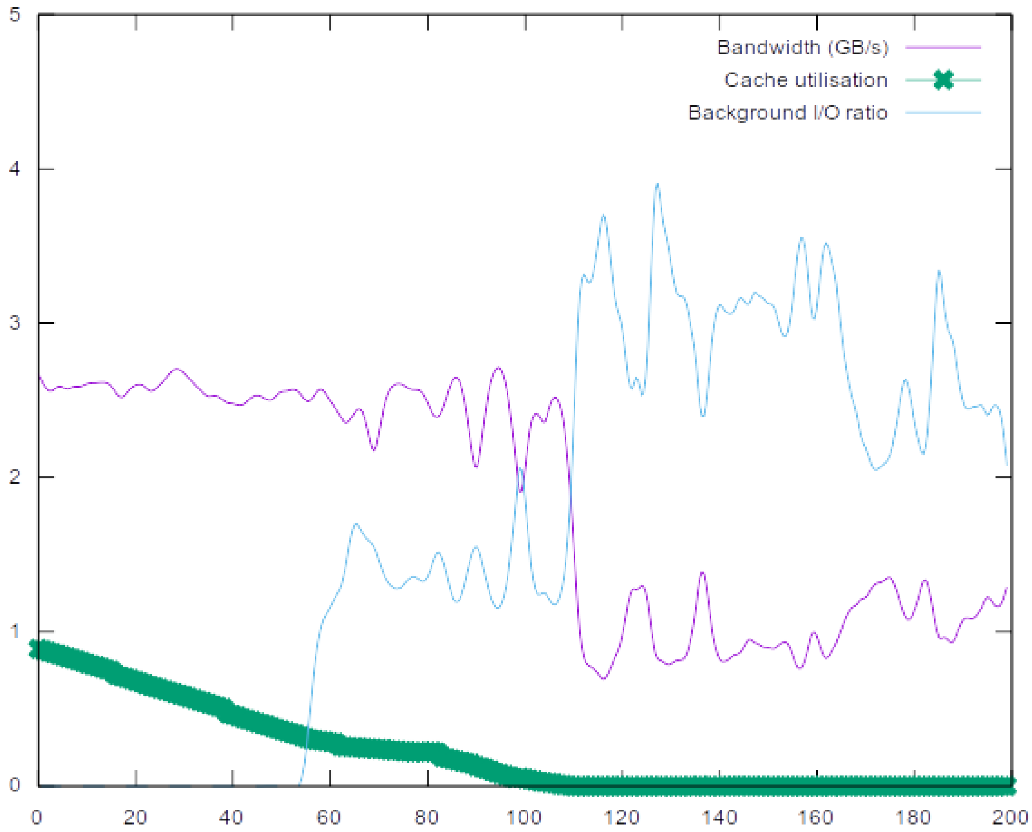
Write amplification: 2 disks



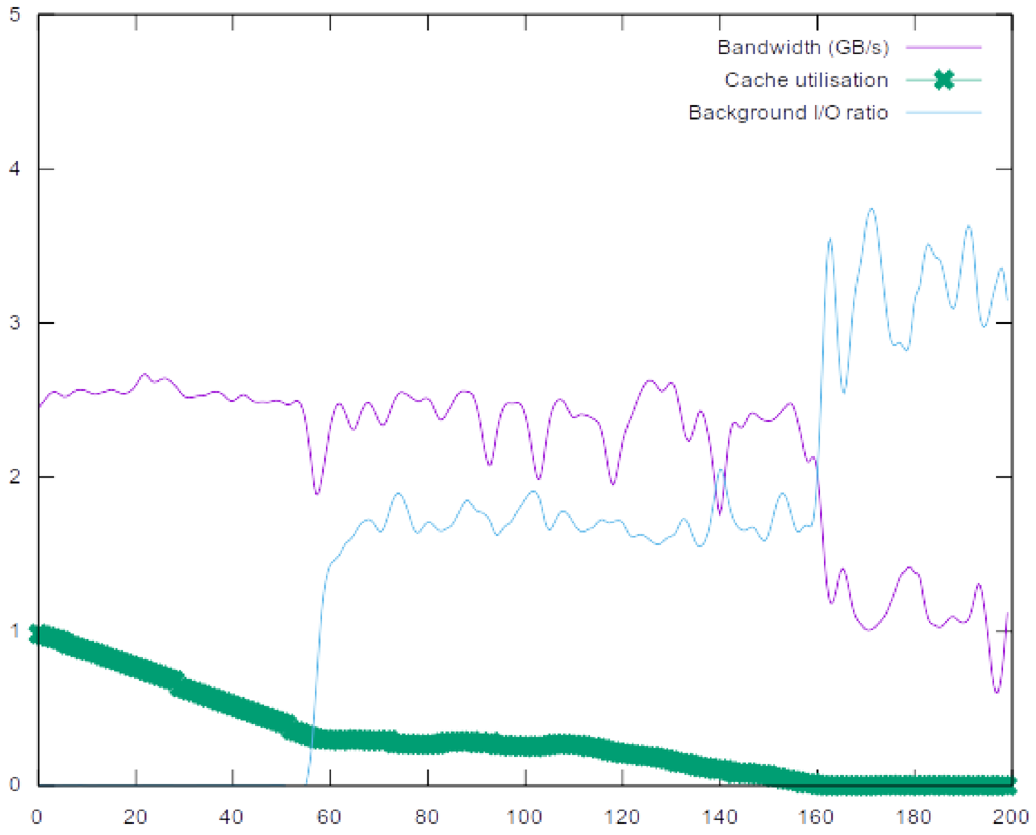
Write amplification: 4 disks



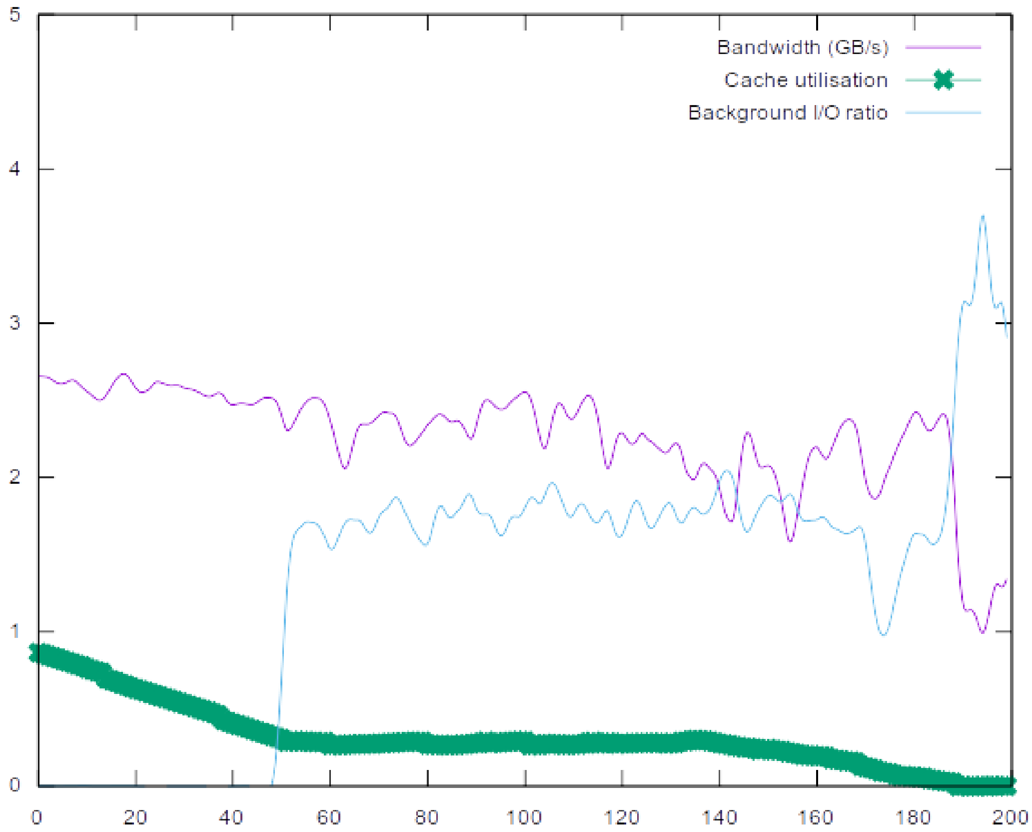
Write amplification: 6 disks



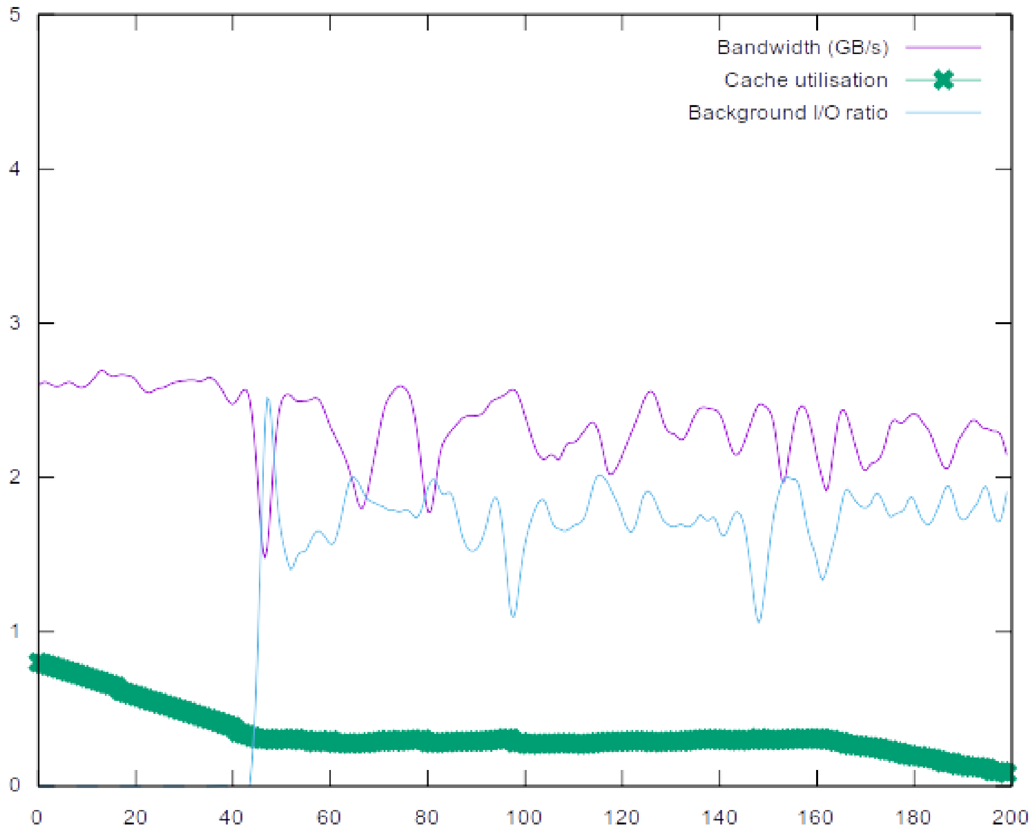
Write amplification: 8 disks



Write amplification: 10 disks



Write amplification: 12 disks



Write amplification

- Direct correlation between performance degradation and write amplification
- Inverse correlation between cache utilisation and write amplification
- Reclaim tries to run with constant speed per disk; higher fluctuations once it drops below low watermark.

Future work

NVDIMM tuning

- Implement DAX for metadata
- Avoid NUMA effects
  - Restrict reclaim to on-socket CPUs
  - Analyse smp_call vs cache bouncing

Cache-parameter tuning

- Scale caching size with number of disks
  - Currently limited by NVDIMM size
- Improve reclaim throttling
- Establish best parameters for high/low watermarks
- Tests with even more disks

Redundancy

- Currently no redundancy
- Mirror-like functionality possible by duplicating the setup
- Declustered RAID; zone mapping on both sides does not need to be identical

Please take a moment
to rate this session.

Your feedback matters to us.

Thank you!