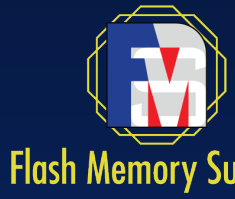
TOMA – a Cluster Manager for a Datapath-Blind Distributed Storage

Ronen Hod

Excelero Storage



Excelero's Technology
(a narrowed-down perspective)

- Software only
- Extremely fast and scalable when run on top of RDMA & NVMe
- Underlying data protection: RAID1 and RAID6
- No caching. Client accesses remote disks directly, bypassing remote CPU

Performance highlights

- I/O goes all the way to the remote disk
- A single NVMesh remote drive vs. a local NVMe drive:
  - Same IOPS, same Bandwidth. +5us Latency
- 128 servers @NASA:
  - More than 140 GB/s write throughput (bounded by NVMe devices)
- Using NVMesh protocol over TCP instead of over RDMA
  - Drawback: The server's CPU is busy
  - Slowdown: We measured 1x-4x (depends)

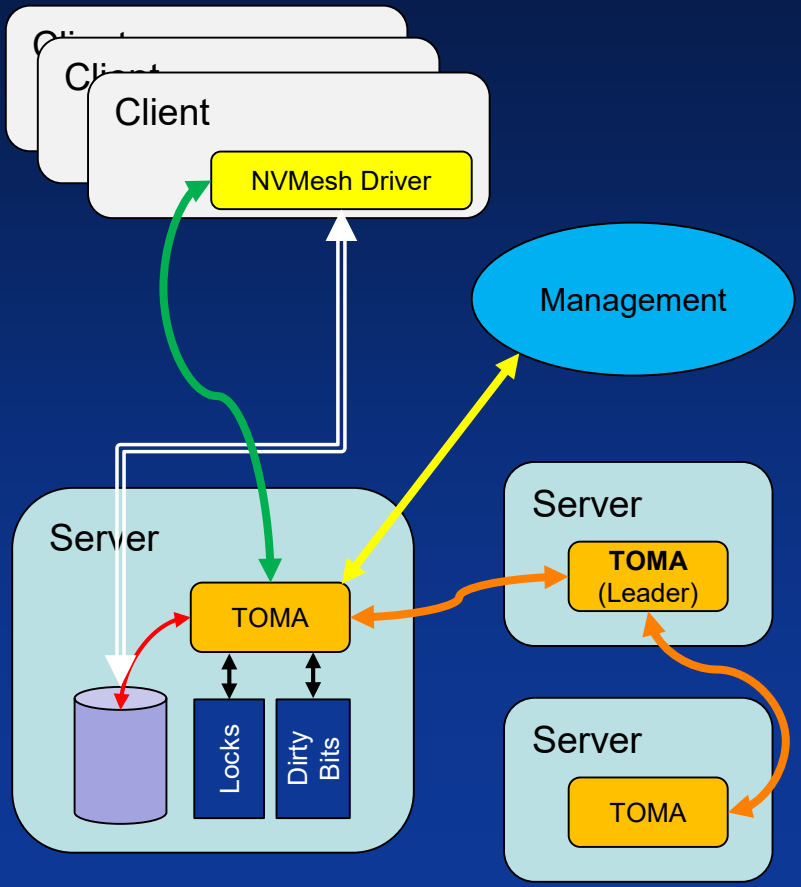
Design constraints

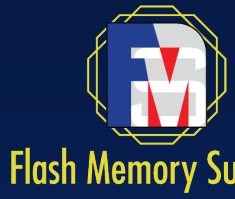
- Disks can be unexpectedly hot-unplugged and then hot-plugged on a different server
- Persistence is stored only on the data disks
- No clocks sync (time & speed)
  - Between TOMAs, their clients, and the management
- Any communication line can break or become unidirectional
- Any TOMA might die
  - specifically the TOMA leader might change
- The server is blind to datapath activity
  - Cannot intercept, redirect or modify a specific client's I/O

Flash Memory Summit

Design principles

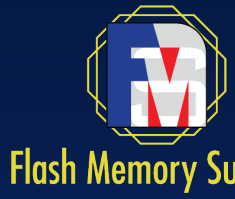
- The TOMA ⇔ client relationships are private per TOMA
- Locks and recovery of stale-locks are private per TOMA
- A client registers on all the RAID's disks before I/O
- Only clients do I/O. TOMA uses the local client to rebuild.
- Soft transitions (no I/O interruptions)





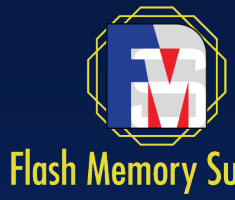
Primary topologies

- *Normal* - (R/W/L, R/W/L) interleaved
  - R: Read
  - W: Write
  - L: Lock
- *Degraded* - (R/W/L, DEAD)
- *Rebuild* - (R/W/L, W)



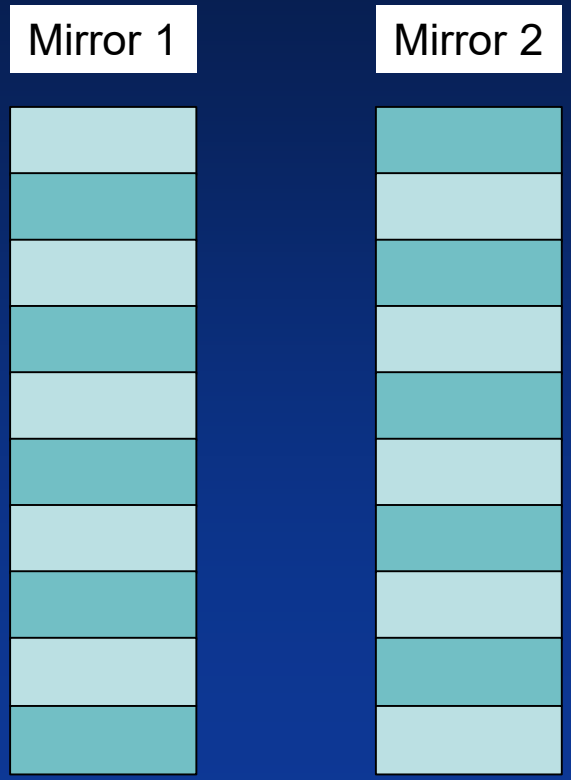
No synchronization of topology changes

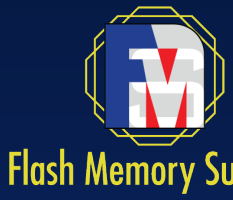
- Not synchronized in time, and no I/O interruption
  - Different TOMAs apply new topologies at different times
  - Clients apply the topologies that they receive from the TOMAs at different times
- Usually, two adjacent topologies can co-exist
- The leader will calculate the next topology only after all the TOMAs finished synchronizing their clients



Locks (RAID1 - Mirroring)

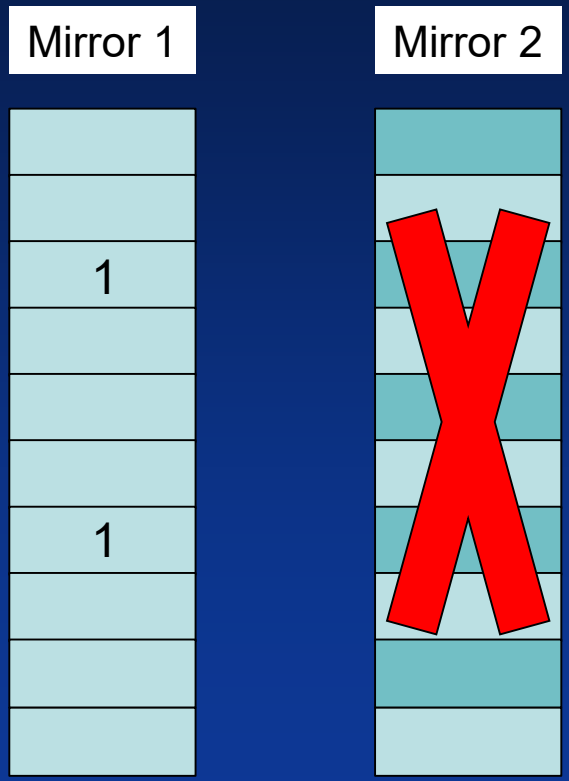
- Per disk, on the server
- In the server's RAM
- Interleaved
  - for performance
- Lock before write
- Read and Lock on the same disk
  - test after read, piggyback





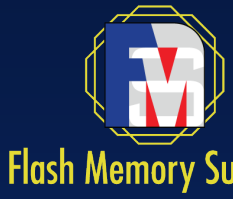
Dirty-Bits (RAID1 - Mirroring)

- Per disk, on the server
- In the server's RAM
- Turned on when writing in degraded mode
- Turned off when writing to both



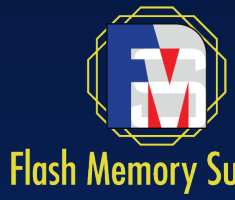
(R/W/L, DEAD) → (R/W/L, W) Init

- Make sure that no I/O is in the pipeline for the DEAD disk
- Clean the locks and dirty-bits on the DEAD disk (in RAM)
- If the DEAD disk is actually NEW, then turn-on all the dirty-bits (in the R/W/L's memory)
- (and I/O doesn't stop)



(R/W/L, DEAD) → (R/W/L, W)

- As always the leader distributes the topology
- The R/W/L disk (TOMA) sends a SWITCH_TOPO message to all its clients
- Once no client is using older topologies, notify the leader
- Once the leader is certain that all the I/O is performed according to the new topology, it sends a “rebuild” topology



(R/W/L, W) → (R/W/L, W) Recovered

- Run one cycle of dirty-bits rebuild
  - Every dirty block is recovered (disk → disk)
  - Every stale-lock is copied (RAM → RAM)
- Note that every new (client's) I/O is already aware of the current topology (R/W/L, W)
- Ordinary client reads/writes also recover blocks, and clear dirty-bit

Where do we stand?

- Data-wise, all is good, and both sides hold the same data (or a lock is held)
- Now we need to transition all the clients to read and lock on the recovered disk
- Again, I/O keeps happening

Recovered → Dual Lock

- The leader distributes a topology of (R/W/L, W/L), no recovery
  - Lock on the left-hand disk, and then lock on the right-hand disk
- The TOMAs ask all their clients to switch to the new topology
- Once the switch is done, report to the leader

Dual Lock → Normal

- (R/W/L, R/W/L)
  - Back to the interleaved fashion
- The TOMAs ask their clients to switch to the new topology
- Once the switch is done, report to the leader