Designing SSD Storage Systems for Low Latency Without Large Outliers

Sebastien Jean, Phison Electronics

Imran Hirani, Everspin Technologies

Improving High Reliability Storage

- Advanced storage appliances use two key techniques to improve the reliability of a multi-drive array
  - Journaling (ex: Transaction logging, checksums, data logging)
  - Physical Redundancy (ex: RAID-6, Redundant Power)

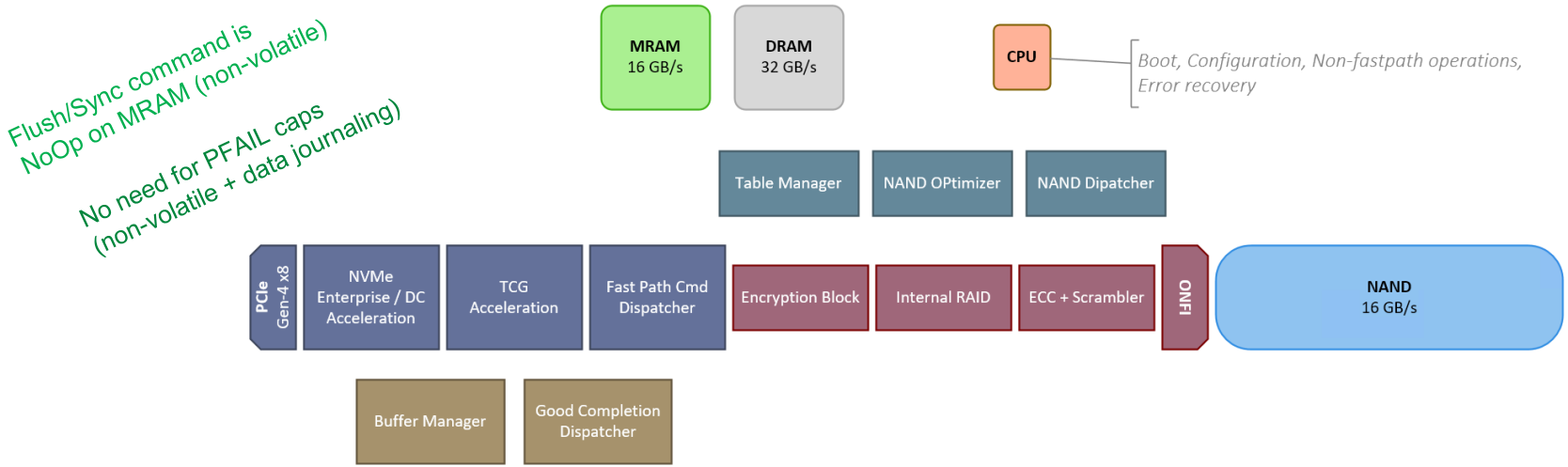
- Journaling generally provides three levels of protection with increasingly slower throughput
  - **Writeback mode** – Only the metadata is journaled describing the transaction and filesystem structural changes. User data and metadata are written in parallel and can fall out of sync during a power failure.
  - **Ordered mode** – Forces the data to be written first, serializing the data and journal operations
  - **Journal mode** – Will write both the metadata and user data to the journal, then write the user data to the data drives

Improving High Reliability Storage

- Journal data is very short lived, but must be persistent
  - It has a significant impact on write amplification and is a major contributor to high SSD TBW requirements
  - This forces the use of high Over Provisioning (OP) or high cycling NAND which are both expensive
  - Moving the journal to dedicated redundant drives consumes slots, power and cooling
- Organizations that must use high reliability storage configurations pay a very high operational cost
- The most effective way of improving the performance of a fully journaling file system is to move the journal off the user data path onto a dedicated redundant storage solution
  - Journal size tends to be relatively small
  - Still consumes 2-5 slots depending on the type of redundancy that is needed
  - Placing multiple journals from different volume onto one drive-set splits the efficiency improvement and pushes out the latency on every volume in the new “meta set”

Improving High Reliability Storage

- Ideal configuration: a redundant high-speed side-band solution that does not require any additional drive slots
  - MRAM has very high write bandwidth and program cycles; perfect for data with low tenure like journals
  - Does not require an FTL, can be used like DRAM, naturally PFAIL capable
  - Very tight latency distribution that ensures minimal degradation in a RAID environment
  - Cost is much lower than high P/E NAND or high OP SSD and there is no additional slot overhead



Proof of Concept

- We set up a Linux server with EXT4 set to mode=journal (full journaling)
  - PS5012-DC SSD as RAID-6
  - PS5012-DC SSD as RAID-6 + PS5012 Journal
  - PS5012-DC SSD as RAID-6 + NVNitro MRAM Journal
- Enterprise deployments tend to be optimized for specific tasks
  - Transaction Workloads: 4-8K IO, short tenure data, 50/50 read/write
  - Data Workloads: 64-256K IO, 70/30 read/write
- Experiment 1 – Transaction WL
  - FIO 3.15
  - Precondition drive with random write until steady-state
  - Measure Transaction Workload for 1 hour
- Experiment 2 – Data WL
  - FIO 3.15
  - Precondition drive with random write until steady-state
  - Measure Data Workload for 1 hour

Transaction WL

	Read	Write
4K	34%	48%
8K	26%	12%
16K	7%	12%
32K	7%	6%
64K	10%	7%
128K	12%	12%
256K	14%	3%

Data WL

	Read	Write
4K	9%	14%
8K	6%	4%
16K	3%	2%
32K	2%	1%
64K	30%	27%
128K	38%	43%
256K	13%	11%

Proof of Concept

- The data for Experiment 1 & 2 were accidentally run in parallel
  - This resulted in a bimodal distribution (4-8K; 64-256K)
  - CPU loading was only 8% despite have 64 tasks
  - Storage write bandwidth was only pushed to 10%

- Initially this data appeared to be unusable
  - Upon further consideration we realized this represented lightly loaded virtualized servers
  - As the workload increased, these virtual machines would be moved to dedicated hardware

Transaction WL

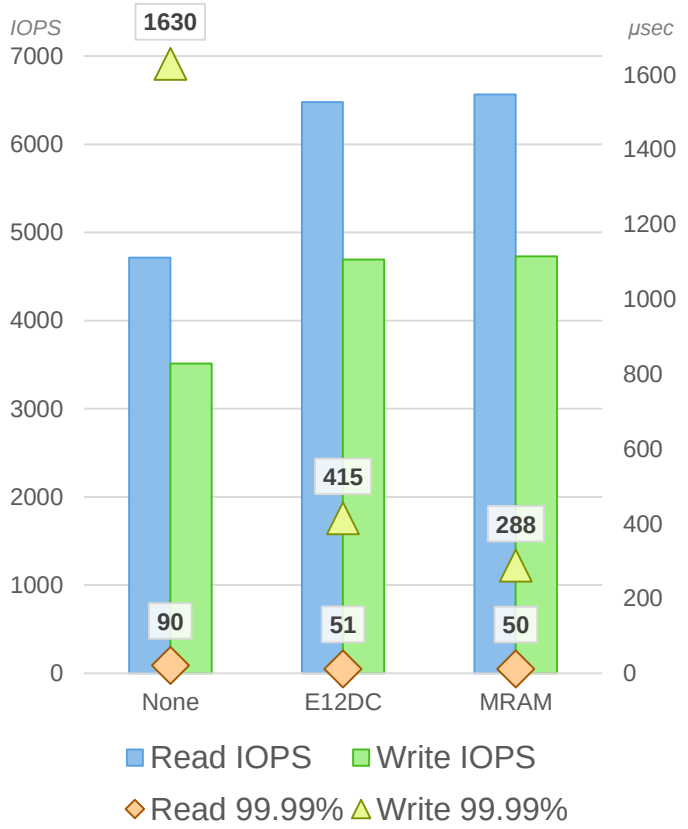
	Read	Write
4K	34%	48%
8K	26%	12%
16K	7%	12%
32K	7%	6%
64K	10%	7%
128K	12%	12%
256K	14%	3%

Data WL

	Read	Write
4K	9%	14%
8K	6%	4%
16K	3%	2%
32K	2%	1%
64K	30%	27%
128K	38%	43%
256K	13%	11%

Study Results

- Assuming an organization requires RAID-6 with full journaling
  - Despite only using ~10% of the system resources
  - Moving the journal off the data drive increases the R/W IOPS 36%
  - Read 99.99% latency is reduced by 44%
  - Write 99.99% latency is reduced by 75% (E12DC) and 82% (MRAM)
- Higher system loading will amplify the differences between all three configuration
  - Moving the journal off the data media allows the NAND to be used for user IO
  - MRAM provides PFAIL and can be integrated into an SSD
  - No additional slots required each SSD has its own journal MRAM pool
  - Journal MRAM can also be configured as RAID
- Next steps
  1. Split workloads and increase system stress to 100%
  2. Add Flush/Sync operations and Move Data RAID bitmap to journal
  3. Produce detailed latency (99.9999%) and Write Amplification analysis
  4. Review power study (SSD Journal vs MRAM Journal)



PHISON

Knows What You Need

"PUSHING BOUNDARIES"

YOUR PCIe GEN-4 SSD LEADER

VISIT US AT BOOTH #219!