Optimizing SSD use for power and endurance

Sampath Ratnam, Anthony Constantine
Distinguished Member of Technical Staff, Micron Technology

Intelligence Accelerated™

SSD capacity trends, impacts

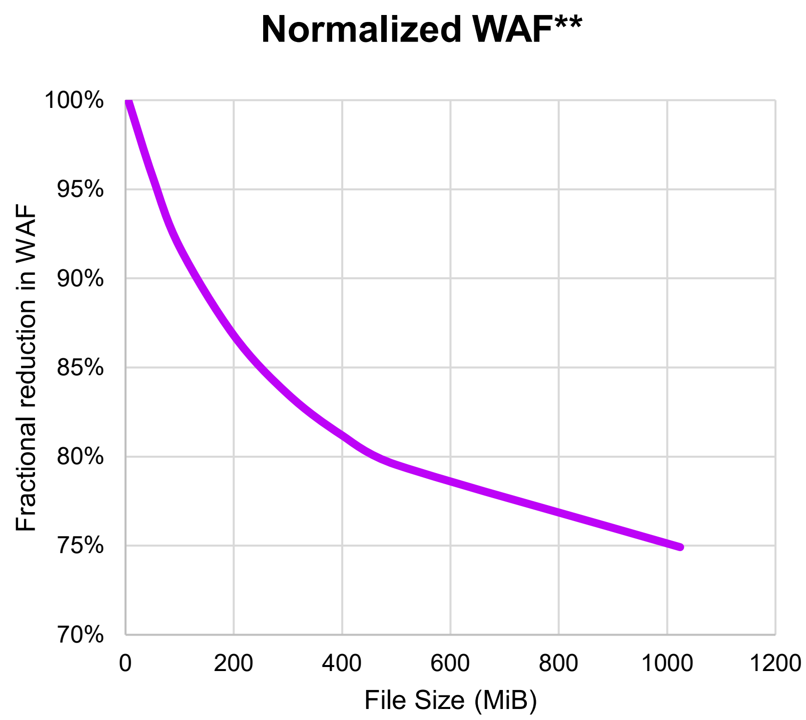
- High-capacity SSDs are rapidly doubling capacity
  - NAND bit density growth ~30% per year
  - Increased die stacking
  - Increased number of NAND packages per SSD
- Most enterprise SSDs use DRAM to manage Indirection Units (IU)
  - IUs translate the host’s address to the physical address
  - Improves performance and endurance
- DRAM capacity cannot keep up with the capacity growth of high-capacity SSDs
  - Scaling the IU becomes cost/space prohibitive

Larger IU inevitable with increasing capacity points

SSD endurance

- SSD endurance is governed by 2 parameters – media capability and system Write Amp Factor (WAF)
  - Media capability – number of program/erase cycles flash media can support
  - Write Amp Factor – total number of media writes performed for every host write, on average
- Reducing or managing WAF towards ideal target (WAF ~= 1) gets most of the drives on several fronts
  - Improves RW IOPs/Watts (due to reduced fragmentation), improved DWPD for a given NAND endurance, cost
- Typical WAF for aligned (host transfer size and IU) Random Writes (on a 7% OP drive) is ~4*
  - E.g., WAF for a 4k aligned write to a 16K IU would be 4x
- Increasing write block size will gradually improve WA, asymptotically approaching a WA =~1

Accumulating writes into large chunks helps (Storage, snapshots)



* Will vary with Workload/System/Media over provision budgets. **Plot for illustrative purposes only

Impact of larger IU size (16k instead of 4k)

Real life data of WAF_{IU} from benchmarks (by Volume)

Flash Memory Summit

$WAF_{Total} = WAF_{App} * WAF_{SSD} * WAF_{IU}$

$1 \leq WAF_{IU} \leq 4$ for 16KB IU – The lower the better

Application	4096	8192	16384	32768	65536	131072	262144	524288	1048576	Avg Size Wr (KB)	Worst Case 16K WAF_{TU}	Measured 16K WAF_{TU}
	Bucketized Write Size (by Volume)											
Expected based on 4KB RW	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4	4.00	4.00
1350-02 TPCH/XFS 8-Streams, Low Mem	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	64	1.25	1.0001
1350-03 TPCH/XFS Single Stream, Hi Mem	0.0%	0.0%	0.0%	0.6%	99.4%	0.0%	0.0%	0.0%	0.0%	64	1.25	1.0028
1350-04 TPCH/XFS Single Stream, Low Mem	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	64	1.25	1.0032
1363-A: YCSB on RocksDB - Workload A	0.0%	0.0%	0.1%	0.2%	0.3%	0.6%	14.8%	64.2%	19.8%	570	1.04	1.0066
1363-B: YCSB on RocksDB - Workload B	0.5%	0.0%	0.0%	0.0%	0.0%	0.0%	16.6%	82.9%	0.0%	467	1.05	1.0064
1363-F: YCSB on RocksDB - Workload F	0.0%	0.0%	0.1%	0.2%	0.0%	0.6%	14.2%	64.3%	20.6%	577	1.04	1.0066
1413-00: Cassandra/XFS YCSB 512GB Load	0.0%	1.4%	0.0%	0.0%	0.0%	0.0%	90.7%	1.6%	6.2%	304	1.09	1.0343
1413-01: Cassandra/XFS YCSB 512GB Workload A	0.0%	0.9%	0.0%	0.0%	0.1%	0.2%	49.2%	17.2%	32.4%	546	1.06	1.0305
1413-02: Cassandra/XFS YCSB 512GB Workload B	0.1%	1.0%	0.1%	0.1%	0.2%	0.6%	57.6%	18.4%	22.0%	468	1.07	1.0311
1413-04: Cassandra/XFS YCSB 512 GB Workload F	0.4%	1.0%	0.1%	0.1%	0.2%	0.5%	57.6%	18.6%	21.5%	463	1.08	1.0318
1413-EXT4-01: Cassandra YCSB/ EXT4 128 GB Workload A - nvmet	0.0%	0.4%	0.3%	0.1%	0.2%	0.6%	18.3%	48.5%	31.6%	620	1.04	1.019
1413-EXT4-01: Cassandra YCSB/ EXT4 128 GB Workload A	1.1%	0.4%	0.3%	0.1%	0.2%	0.6%	17.7%	48.1%	31.4%	614	1.08	1.019
1413-EXT4-04: Cassandra YCSB/ EXT4 128 GB Workload F - nvmet	0.0%	0.7%	0.3%	0.1%	0.1%	0.4%	43.2%	29.8%	25.5%	524	1.06	1.0236
1413-EXT4-04: Cassandra YCSB/ EXT4 128 GB Workload F	3.7%	0.8%	0.3%	0.2%	0.2%	0.9%	40.5%	31.3%	22.1%	491	1.17	1.0236
1413-XFS-01: Cassandra YCSB/ XFS 128 GB Workload A	0.1%	0.6%	0.3%	0.0%	0.1%	0.4%	26.7%	10.7%	61.2%	750	1.05	1.0256
1453-02dc Ceph RadosBench- Both data and metadata	4.5%	3.1%	0.9%	0.2%	89.8%	0.0%	1.5%	0.0%	0.0%	62	1.43	1.18
1453-b7ca Ceph RadosBench- Data nvme0n1	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	64	1.25	1.12
1453-b7ca Ceph RadosBench- Metadata nvme7n1	50.6%	36.9%	10.7%	1.7%	0.0%	0.0%	0.0%	0.0%	0.0%	7	3.37	2.53

"Worst Case": assumes all Writes are sized as bucket and misaligned to IU start

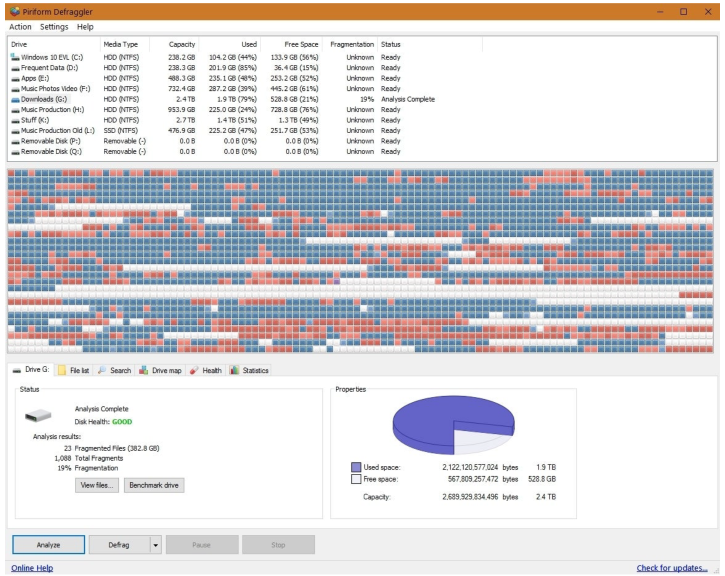
Green= low %; Red= high %; Rest is gradient colors 6

Several applications already optimized to larger writes

Details: https://www.micron.com/about/blog/storage/innovations/real-life-workloads-more-efficient-large-ssd-capacities

Best practices for “good” host write behavior for SSDs

- Convert Random to Sequential traffic where possible
  - Use industry approved append file systems or techniques to co-locate LBAs spatially
  - E.g.1GB write in 128MB chunks, Each chunk writes 1MB aggregated in ~256K extents
- Exercise Deallocates - Create a “runway” for writes.
  - Large Deallocates are better than smaller deallocates
  - Avoid multiple write streams that have different lifetimes and will be deallocated at different times
- Gather writes up to NVMe MDTS (Max Data Transfer Size) ~1MB
  - Larger the writes the better, multiples of MDTS sized IOs preferred
  - Writes are aggregated internally and flushed to media
- Avoid in-place fragmentation →
  - writing large and rewriting smaller chunks
- Use OP Techniques
  - Track logical saturation and deallocate ahead of writes

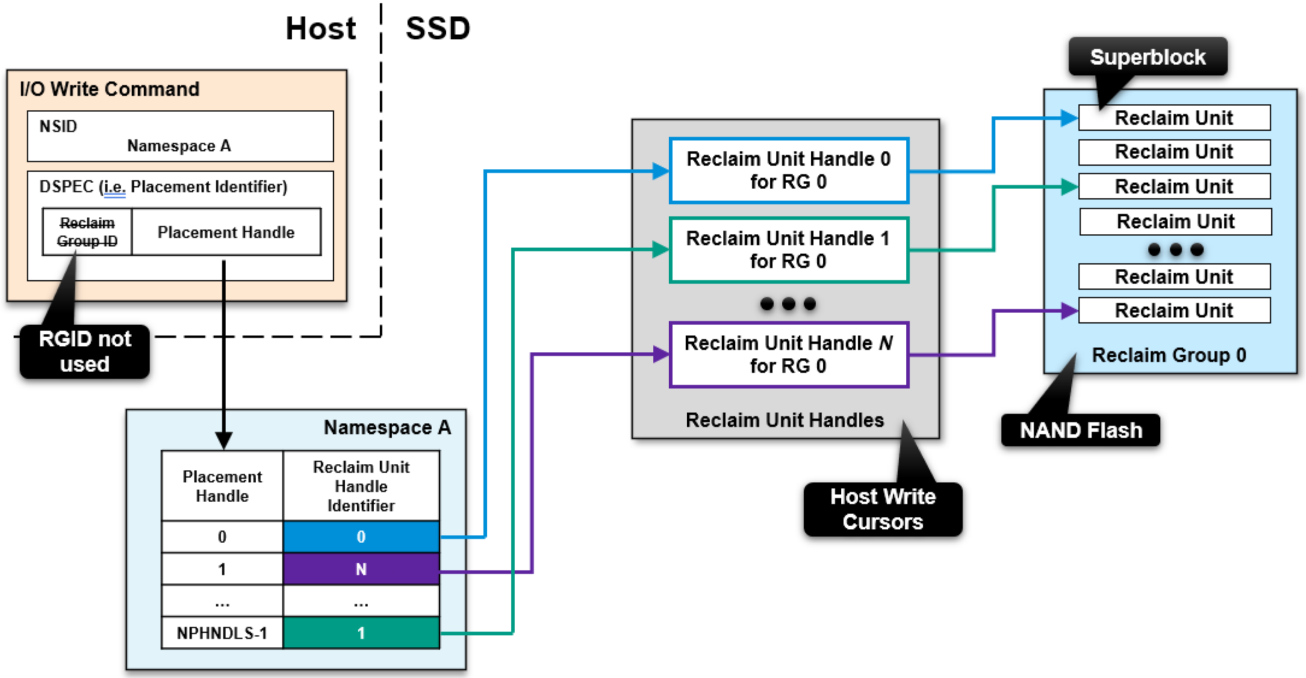


Micron's Gen5 Flexible Data Placement (FDP)

Flexible Data Placement (TP4146) is a feature designed to **reduce Write Amplification (WA)** by aligning LBA usage with physical media

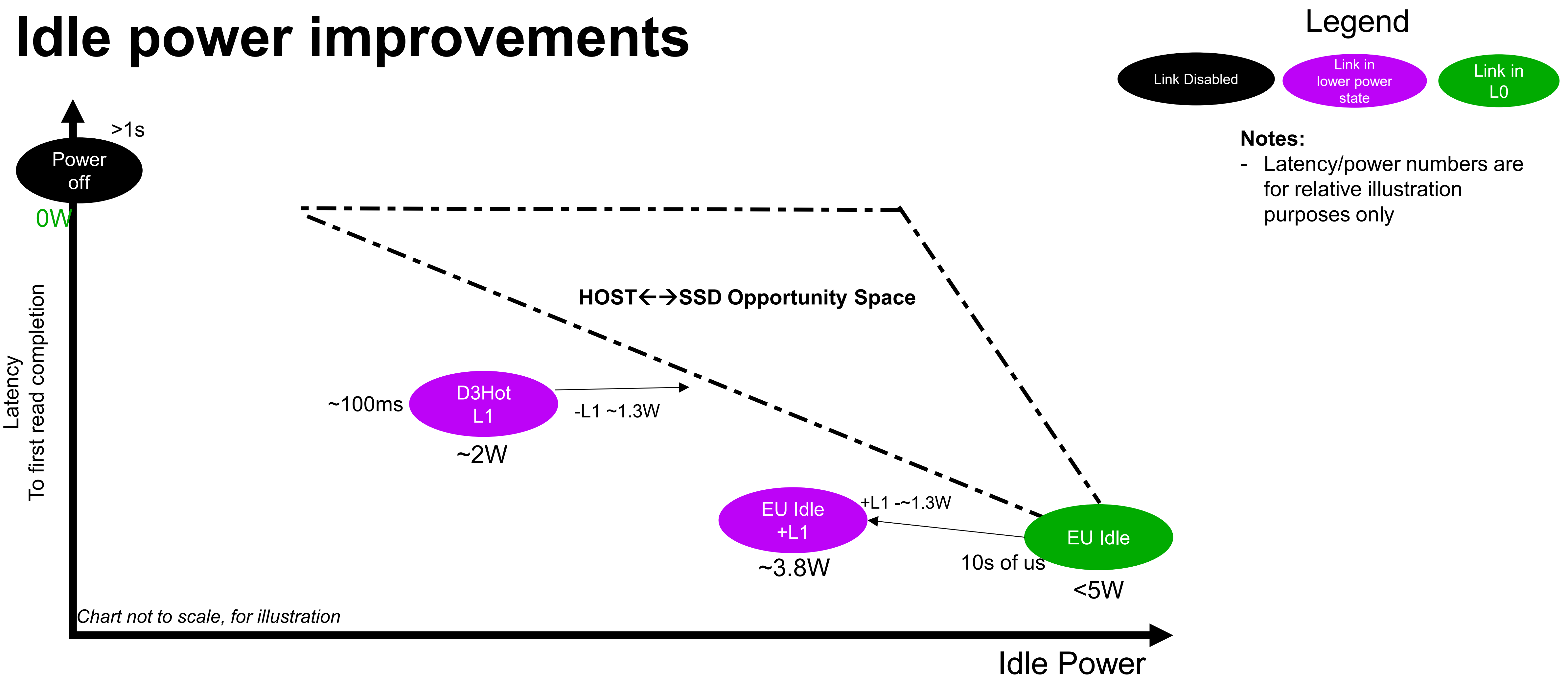
- Host Writes use Directives to place data into SSD Write Cursors
- **SSD reports physical media characteristics** so Host can deallocate data on NAND block granularity

	FDP Features*
FDP Configurations	2
Endurance Groups	1
Sector Sizes Supported	512, 4096, 4160
Reclaim Unit Handles by configuration	8 initially isolated
Reclaim Groups	1
Reclaim Unit Nominal Size *Subject to change*	Superblock: • **1.92TB:** ~5GB • **3.84TB:** ~10GB • **7.68TB:** ~19GB • **15.3TB:** ~39GB
Max Number of Namespaces *(when FDP is enabled)*	16



FDP Write Data Path for 1 Reclaim Group (RG)

Idle power improvements



Notes:
- Latency/power numbers are for relative illustration purposes only

Most of the power in idle comes from the ASIC High-Capacity Challenges for Idle
- Large number of NAND die, may be worthwhile to power NAND off completely but will increase power on latency
- Media management such as scans and folds impact expected to get worse with technology scaling

HOST can proactively manage SSDs to improve power consumption

Call to action

- With Higher Capacity SSDs, care needs to be taken to ensure the SSD does not prematurely wear out.
- Implementing some best practices including keeping the minimum transfer size greater than the IU size, large deallocates, write combining to MDTS, and avoiding fragmentation will extend the life of the SSD.
- FDP can also help SSD endurance if implemented correctly.
- Implementing idle power features like L1 and D3hot will help reduce power per SSD

Intelligence Accelerated™