Flash Memory Summit

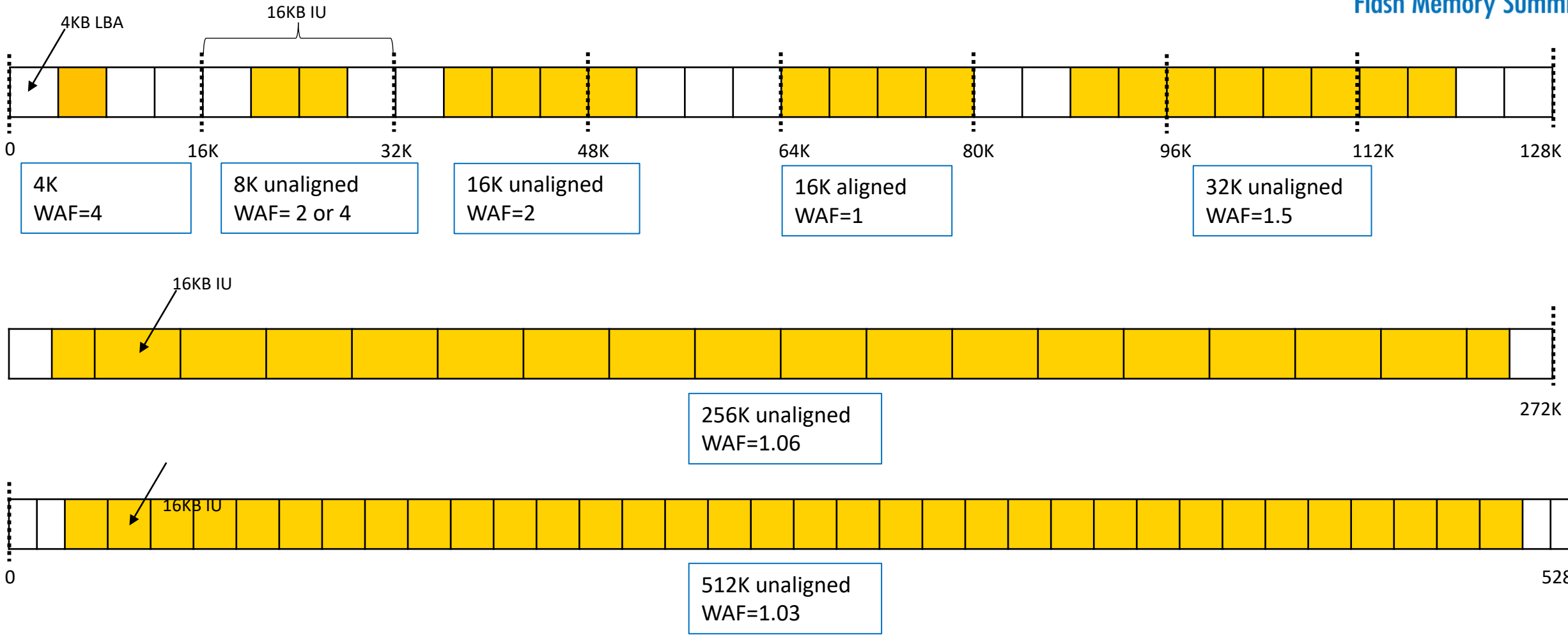
Impact of 16KB Indirection Units on Real Life workloads

Luca Bert, Micron Technology, Inc.

Support for large Indirection Units (IU): Problem statement

- General term to define a FTL mapping unit larger than LBA size
- Large IU is necessary to support large capacity SSD
  - DRAM size to keep 4K IU maps becoming prohibitive
  - 16KB is the most promising size for large SSD but others can be considered
- Main concerns around induced Write Amplification (WAF) due to unsized/ unaligned Writes
  - $WAF_{Total} = WAF_{App} * WAF_{SSD} * WAF_{IU}$
  - WAF_{IU} Is the multiplicative factor induced by Large IU
  - $1 \leq WAF_{IU} \leq 4$ for 16KB IU
  - Perception is that 16K IU will result in $WAF_{IU} = 4$ and thus 4x worse endurance
- We need real life data to support/ challenge above statement

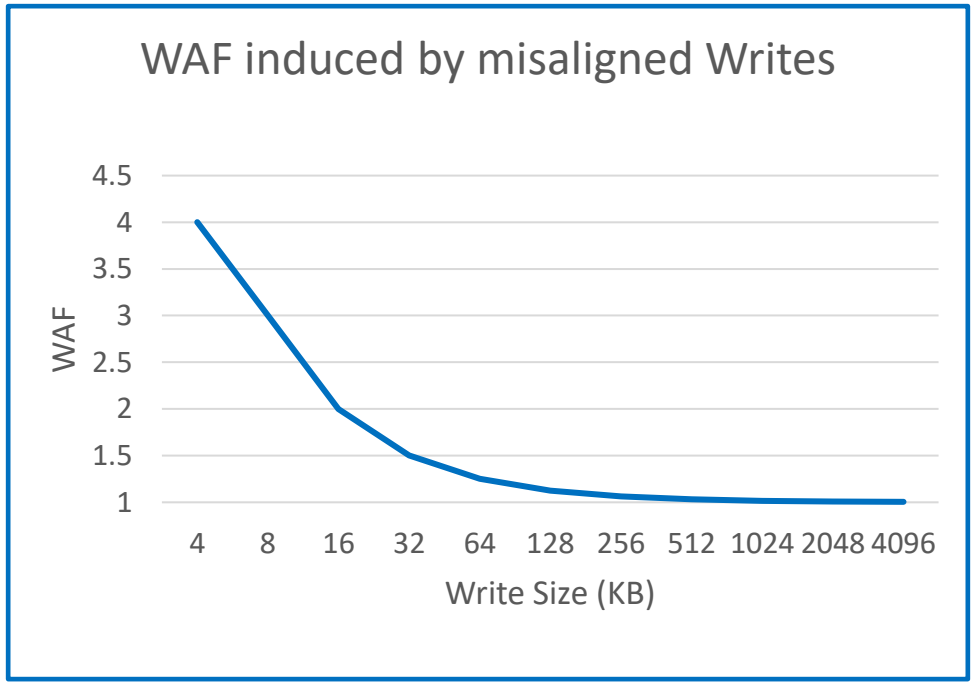
16K IU induced WAF: larger IOs have smaller impact



Head and Tail IU are the only one affected. All others are “aligned” by definition

16K IU induced WAF: larger IOs have smaller impact

- Impacts only writes
  - Reads are unaffected
- Impacts only 16KB-boundary misaligned writes
  - 16KB-boundary aligned Writes do not introduce any WAF
- Impacts decreases exponentially with increased Write Size
- Most File System tends to aggregate writes and issue large Writes
- Large IU induced WAF based on Write size:
  - At 4KB = 4 (e.g. Legacy)
  - At 64KB = 1.25 (e.g. TPC-H)
  - At 1MB = 1.015 (e.g. Kernel Block Layer max)
  - At 2MB = 1.007 (e.g. RocksDB)
  - At 4MB = 1.003 (e.g. several SPDK based solutions)



At some point, induced WAF becomes viable for the capacity/ cost advantage it provides

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

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

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

	Bucketized Writes (by IO count)											
Application	4096	8192	16384	32768	65536	131072	262144	524288	1048576	Avg Size Wr (KB)	Worst Case 16K WAF_{TU}	Measured 16K WAF_{TU}
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%	1.1%	98.9%	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	2.1%	2.1%	2.1%	2.1%	2.1%	2.1%	25.0%	54.2%	8.3%	**432**	1.18	1.0066
1363-B: YCSB on RocksDB - Workload B	36.4%	0.0%	0.0%	0.0%	0.0%	0.0%	18.2%	45.5%	0.0%	**281**	2.12	1.0064
1363-F: YCSB on RocksDB - Workload F	2.2%	2.2%	2.2%	2.2%	0.0%	2.2%	24.4%	55.6%	8.9%	**442**	1.18	1.0066
1413-00: Cassandra/XFS YCSB 512GB Load	0.2%	32.9%	0.1%	0.0%	0.0%	0.0%	64.7%	0.6%	1.1%	**183**	1.70	1.0343
1413-01: Cassandra/XFS YCSB 512GB Workload A	1.2%	27.6%	0.2%	0.2%	0.2%	0.5%	49.4%	8.6%	8.2%	**257**	1.59	1.0305
1413-02: Cassandra/XFS YCSB 512GB Workload B	4.3%	26.9%	1.2%	0.6%	0.6%	0.9%	48.5%	7.7%	4.6%	**215**	1.67	1.0311
1413-04: Cassandra/XFS YCSB 512 GB Workload F	19.8%	23.2%	0.8%	0.5%	0.5%	0.8%	41.0%	6.6%	3.8%	**182**	2.07	1.0318
1413-EXT4-01: Cassandra YCSB/ EXT4 128 GB Workload A - nvmetr	1.3%	17.4%	6.3%	1.2%	1.0%	1.6%	25.9%	34.2%	11.2%	**361**	1.49	1.019
1413-EXT4-01: Cassandra YCSB/ EXT4 128 GB Workload A	50.4%	9.1%	3.2%	0.8%	0.6%	0.9%	12.2%	16.7%	5.4%	177	2.74	1.019
1413-EXT4-04: Cassandra YCSB/ EXT4 128 GB Workload F - nvmetra	3.2%	23.3%	4.3%	0.7%	0.4%	0.8%	44.4%	15.3%	6.5%	263	1.64	1.0236
1413-EXT4-04: Cassandra YCSB/ EXT4 128 GB Workload F	70.2%	7.9%	1.4%	0.5%	0.3%	0.5%	12.0%	4.6%	1.6%	76	3.28	1.0236
1413-XFS-01: Cassandra YCSB/ XFS 128 GB Workload A	8.9%	21.8%	4.8%	0.4%	0.4%	0.8%	30.2%	6.0%	17.3%	**290**	1.68	1.0256
1453-02dc Ceph RadosBench- Both data and metadata	37.5%	13.1%	1.8%	0.2%	46.9%	0.0%	0.2%	0.0%	0.0%	**33**	2.52	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	69.9%	25.5%	3.7%	0.3%	0.0%	0.0%	0.0%	0.0%	0.0%	**6**	3.64	2.53

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

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

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

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

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

	Bucketized Write Size (by Volume)											
Application	4096	8192	16384	32768	65536	131072	262144	524288	1048576	Avg Size Wr (KB)	Worst Case 16K WAF_{TU}	Measured 16K WAF_{TU}
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 - nvmetr	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

Takeaways

- 16K IU comes with great benefits on memory footprint (75% DRAM size savings) but may contribute to WAF
- On 4 corner analysis 16KB SSD WAF may be as high as 4x than 4KB IU SSD
- On Real life profiles additional WAF has shown to be much lower,
  - Definitely <2x
  - Sometimes, close to 1x
- Some workloads will be more suitable than others
  - Metadata are not a good choice but, when blended with data, are not disrupting them in any significant way
- Moving to 16KB IU will be less impactful to performance than assumed

Questions?