Fine-tuning RocksDB for NVMe SSD

Praveen Krishnamoorthy (p.krishna@samsung.com)
Changho Choi (changho.c@samsung.com)

Memory Solutions Lab @ Samsung

SAMSUNG

PERCONA LIVE · USA
SANTA CLARA

DATA PERFORMANCE CONFERENCE · MySQL · NoSQL · DATA IN THE CLOUD

Motivation & Goals

- ## What is our motivation?
- Implement features like multi-stream for RocksDB
- Characterize the performance & lifetime improvement with multi-stream
- Deliver highest ops/sec using Samsung PM1725 NVMe SSD

- ## What we observed?
- Numerous parameters/knobs (~120)
- Default values non-optimal for High Performance SSD
- Good starting points from RocksDB developer forums and benchmarking articles

- ## What we wanted to achieve?
- Arrive at a methodology (using greedy algorithmic approach) to identify optimal values for RocksDB configuration parameters that deliver higher performance for specific workloads in our usage scenario

System, Dataset & Workload Details

Processor/Memory Details	Operating System	SSD Details
Processor : - 2 x Intel(R) Xeon(R) CPU E5-2640 v3 @ 2.60GHz. **Total Physical CPU :** 16 **Total Logical CPU (HT):** 32 **Total Memory** : 64 GB	**Distro :** Ubuntu 14.04.1 LTS **Kernel :** 3.19.0-11-generic **Arch :** x86_64	**SSD :** 1 x Samsung NVMe PM1725 400GB Samsung SSD activated

- Filesystem Used : XFS Filesystem mounted with discard (FS Block Size : 4K)
- Benchmark Tool : DB_bench

- Dataset : 253GB (240 [uncompressed] /400 million [snappy] records)
  - Dataset size is kept greater than main memory size. ie, 1 : 4 ratio
  - Fills 68% of disk capacity
- Key_size : 16 Bytes
- Value_size : 800 Bytes
- Total Operation Count : 500 Million
- Number of DB_bench Threads : 16
- Compression : None/Snappy
- Workloads
  - 80% Read - 20% Write
  - 50% Read - 50% Write

Optimization

1.Parameter Assessment

- Identified ~25 rocksdb configuration parameters eg, block_size, write_buffer_size, etc
- Identified 4 or 5 possible values for each parameter eg, block_size (2K, 4K, 8K, 16K, 32K)

2.Parameter Isolation

- Selected one parameter to study for different values

3.Benchmarking

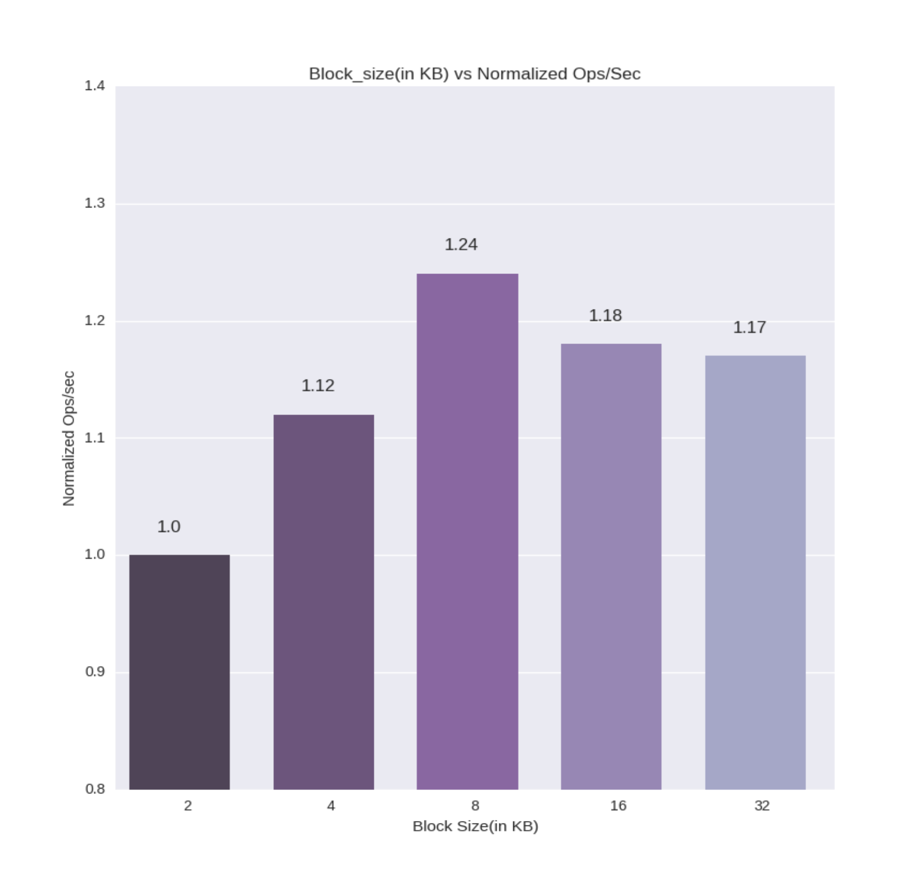
- Performed the benchmark with different values for that isolated parameter

4.Post-Processing

- Selected the value that gave highest ops/sec & use it
- Repeated the process from #2 until all parameters are covered

Optimal Value Selection

Parameter	Value1	Value2	Value3	Value4	Value5
Block_size	2K	4K	8K	16K	32K
Write_buffer_size	128	256	512	1024	2048
Max_write_buffer_number	4	8	16	32	64
Min_write_buffer_number_to_merge	1	2	4	-	-
Max_bytes_for_level_base	256	512	1024	2048	4096
Max_bytes_for_level_multiplier	5	10	20	-	-
Target_file_size_base	32	64	128	256	-
Target_file_size_multiplier	1	2	5	10	-
Max_background_flushes	1	2	4	-	-
Max_background_compactions	16	32	48	64	128
Level0_file_num_compaction_trigger	1	2	4	8	16
Level0_slowdown_writes_trigger	16	24	32	40	48
Level0_stop_writes_trigger	48	56	64	-	-



* Green blocks indicate the value that gave the best performance

Optimal Values for Different Workloads - 1 /2

Parameter	Baseline	80/20 snappy	80/20 uncompressed	50/50 snappy	50/50 uncompressed
Block_size	4K	8K	4K	8K	16K
Cache_size	1GB	4GB	0GB	2GB	0GB
Write_buffer_size	256 MB	512 MB	1024 MB	512 MB	512 MB
Max_write_buffer_number	8	8	4	16	16
Min_write_buffer_number_to_merge	1	4	3	4	1
Max_bytes_for_level_base	512 MB	2048 MB	4096 MB	2048 MB	512 MB
Max_bytes_for_level_multiplier	10	5	5	10	10
Target_file_size_base	64 MB	128 MB	64 MB	256 MB	128 MB
Target_file_size_multiplier	1	1	1	1	1
Max_background_flushes	2	2	1	2	1
Max_background_compactions	32	32	16	48	32
Level0_file_num_compaction_trigger	1	2	2	1	8
Level0_slowdown_writes_trigger	32	24	16	48	40
Level0_stop_writes_trigger	46	56	56	56	46

SAMSUNG

* Green blocks indicate the optimized value deviated from the initial baseline value

PERCONA LIVE

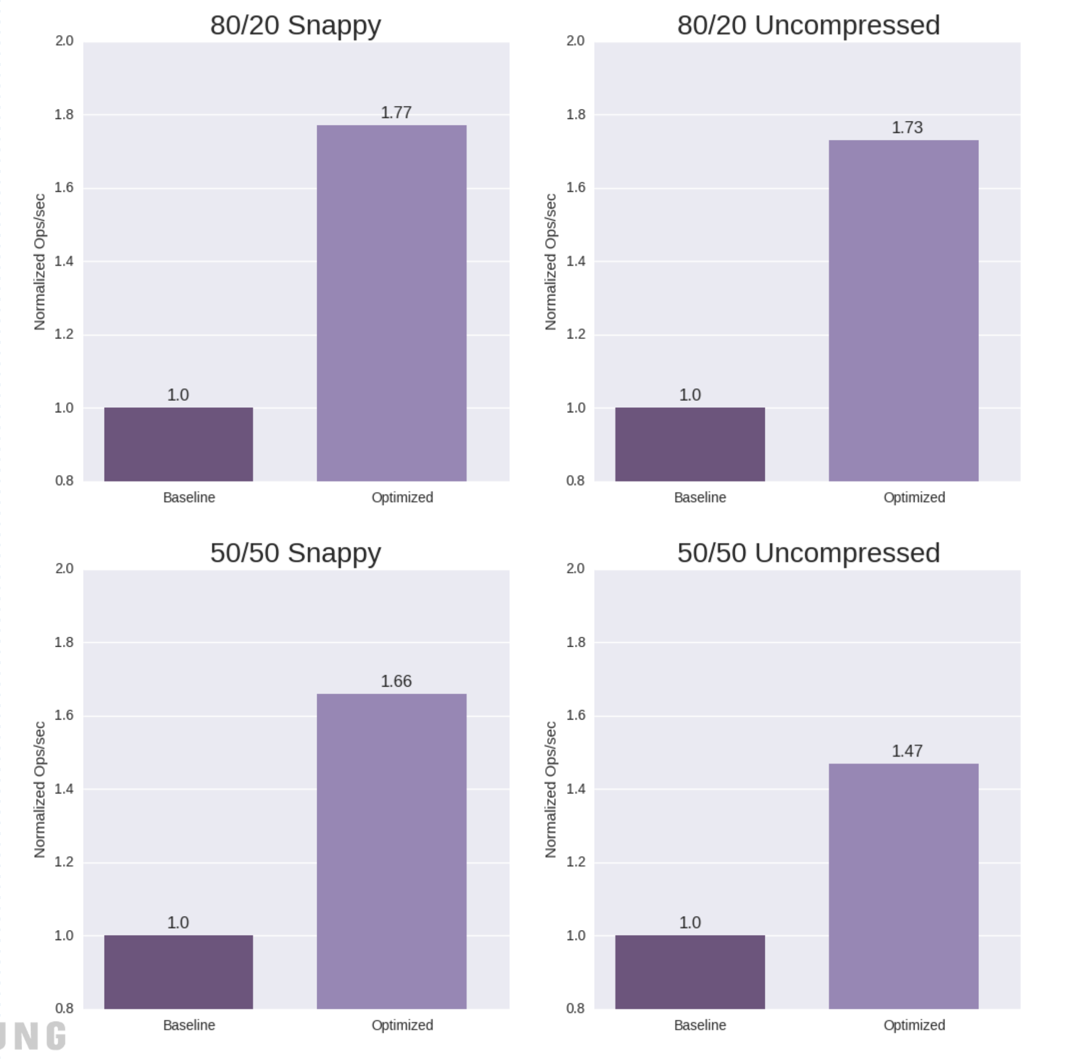
Optimal Values for Different Workloads - 2 /2

Parameter	Baseline Value	80/20 snappy	80/20 uncompressed	50/50 snappy	50/50 uncompressed
Cache_numshardbits	Not Defined – use default (4)	6	-	4	-
Table_cache_numshardbits	Not Defined – use default (4)	4	-	6	-
Mmap_read	Not Defined – use default (0)	1	1	1	1
Mmap_write	Not Defined – use default (0)	0	0	0	1
Use_fsync	Not Defined – use default (0)	0	0	0	1
Use_adaptive_mutex	Not Defined – use default (0)	0	0	0	1
Wal_dir	Not Defined – use default	Default	WAL on different SSD	WAL on different SSD	WAL on different SSD
Bytes_per_sync	Not Defined – use default	2048 KB	512 KB	512 KB	1024 KB
Source_compaction_factor	1	1	1	2	2
Max_grandparent_overlap_factor	10	5	10	10	5

SAMSUNG

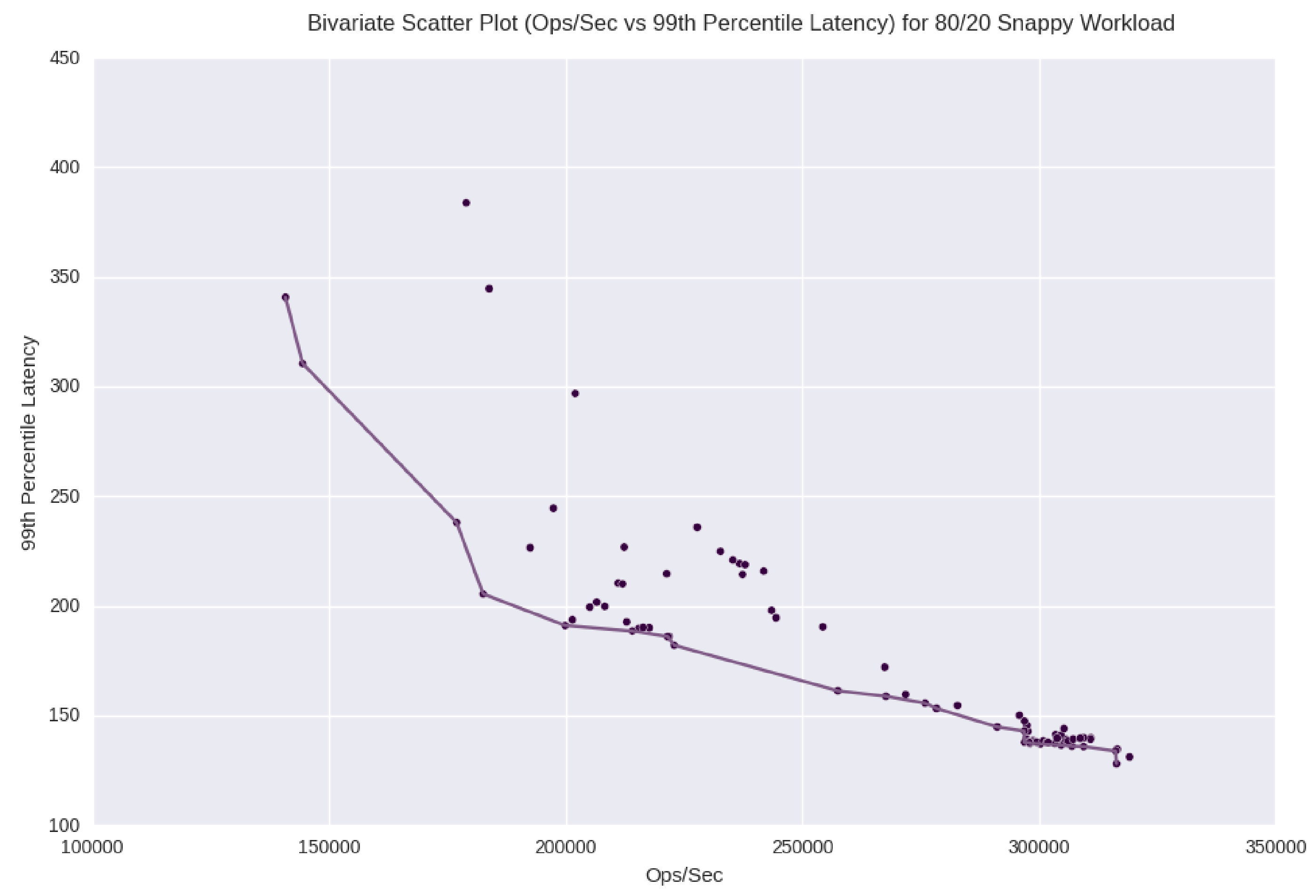
* Green blocks indicate the optimized value deviated from the initial baseline value

Performance Comparison



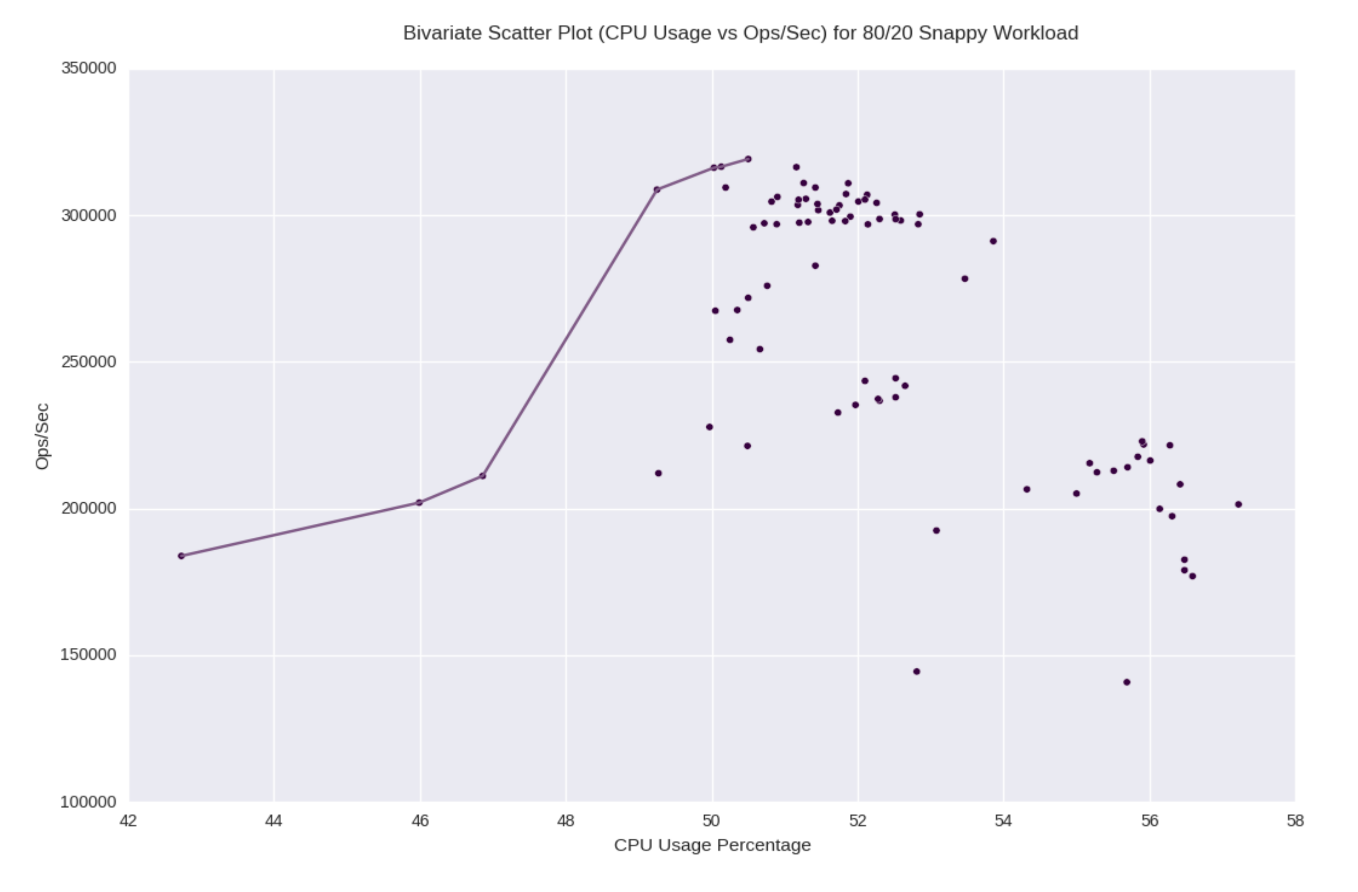
Workload	Baseline (Ops/Sec)	Optimized (Ops/Sec)	Percentage Increase
80% Read/20% Write (Snappy compression)	179 K	316 K	**76.79%**
80% Read/20% Write (Uncompressed)	174 K	301 K	**72.79%**
50% Read/50% Write (Snappy compression)	79 K	132 K	**66.25%**
50% Read/50% Write (Uncompressed)	88 K	129 K	**46.72%**

Ops/sec vs 99th Percentile Latency



- Pareto front – provides the most optimal selection
- Start point (140 K Ops/Sec & 340 µs 99^{th} percentile)
- End Point (316 K Ops/Sec &128 µs 99^{th} percentile)
- Ability to select points based on performance SLA

CPU Usage Percentage vs Ops/Sec



- Pareto start point (42.7 % CPU usage & 184 K Ops/sec)
- End Point (50.5 % CPU usage & 319 K Ops/sec)
- Points to the right of pareto end point doesn't result in higher Ops/sec (with higher CPU usage)

Conclusion

- Take-Away:
  - ~77% increased performance seen with this fine-tuning model
  - Can be replicated for the desired workload/environment
  - Used as a good start-off point and later tuned based on consistent performance monitoring

- Summary:
  - Top 6 contributors for increased performance (80/20 Snappy)
    - max_bytes_for_level_base (11.2%)
    - block_size (10.2%)
    - max_bytes_for_level_multiplier (8.8%)
    - cache_size (8.8%)
    - target_file_size_base (7.8%)
    - write_buffer_size (6.2%)

Questions?

Praveen Krishnamoorthy (p.krishna@samsung.com)
Changho Choi (changho.c@samsung.com)

Memory Solutions Lab @ Samsung

SAMSUNG

DATA PERFORMANCE CONFERENCE · MySQL · NoSQL · DATA IN THE CLOUD

PERCONA
LIVE · USA
SANTA CLARA