RocksDB - Performance Benchmarks

Kết quả đo tải của RocksDB. Chi tiết: https://github.com/facebook/rocksdb/wiki/Performance-Benchmarks

Setup

All of the benchmarks are run on the same AWS instance. Here are the details of the test setup:

- **Instance type:** m5d.2xlarge 8 CPU, 32 GB Memory, 1 x 300 NVMe SSD.
- **Kernel version:** Linux 4.14.177-139.253.amzn2.x86_64
- **File System:** XFS with discard enabled

To understand the performance of the SSD card, we ran an fio test and observed 117K IOPS of 4KB reads (See Performance Benchmarks#fio test results for outputs).

All tests were executed against by executing benchmark.sh with the following parameters (unless otherwise specified):

- NUM_KEYS=900000000
- CACHE_SIZE=6442450944
- For long-running tests, the tests were executed with a duration of 5400 seconds (DURATION=5400)

Unless explicitly specified, the remaining tests used default parameters. DIO tests were executed with the options `--use_direct_io_for_flush_and_compaction --use_direct_reads`.

All other parameters used the default values, unless explicitly mentioned here. Tests were executed sequentially against the same database instance. The `db_bench` tool was generated via `make release`.

The following tests were executed in sequence:

Test 1. Bulk Load of keys in Random Order (benchmark.sh bulkload)

NUM_KEYS=900000000 CACHE_SIZE=6442450944 benchmark.sh bulkload

Measure performance to load 900 million keys into the database. The keys are inserted in random order. The database is empty at the beginning of this benchmark run and gradually fills up. No data is being read when the data load is in progress.

Version	Opts	Time	ops/sec	mb/sec	usec/op	p50	p75	p99	p99.9	p99.99	Stall-time	Stall %	du -sk
7.2.2	None	4021	1003732	402.0	1.0	0.5	0.8	2	7	22	00:00:52.558	6.3	101406408
7.2.2	DIO	3976	1021386	409.1	1.0	0.5	0.8	2	3	32	00:00:41.215	4.9	101404476
7.1.1	None	3951	1028135	411.8	1.0	0.5	0.8	2	3	21	00:00:42.580	5.1	101407124
7.1.1	DIO	3920	1046129	419.0	1.0	0.5	0.8	2	3	20	00:00:33.023	3.9	101407876
7.0.3	None	3934	1040089	416.6	1.0	0.5	0.8	2	3	22	00:01:02.307	7.4	101406288
7.0.3	DIO	3879	1060242	424.7	0.9	0.5	0.8	2	3	21	00:00:50.523	6.0	101405820
6.29.1	None	3898	1045486	418.8	1.0	0.5	0.8	2	3	55	00:01:17.876	9.3	101405948
6.29.1	DIO	3819	1065706	426.9	0.9	0.5	0.8	2	3	25	00:01:09.405	8.3	101404236

Version	Opts	Time	ops/sec	mb/sec	usec/op	p50	p75	p99	p99.9	p99.99	Stall-time	Stall %	du -sk
6.29.0	None	3899	1047693	419.6	1.0	0.5	0.8	2	3	108	00:01:25.637	10.2	10140703 2
6.29.0	DIO	3828	1061703	425.3	0.9	0.5	0.8	2	3	21	00:00:56.298	6.7	10140535 6
6.28.0	None	3924	1050028	420.6	1.0	0.5	0.8	2	3	60	00:01:17.288	9.2	10140626 0
6.28.0	DIO	3819	1072892	429.7	0.9	0.5	0.8	2	3	29	00:01:01.648	7.9	10140591 6
6.27.0	None	3898	1052489	421.6	0.9	0.5	0.8	2	3	22	00:01:07.776	8.1	10140679 6
6.27.0	DIO	3826	1066941	427.4	0.9	0.5	0.8	2	3	21	00:00:58.306	6.9	10140558 0
6.26.0	None	3892	1043630	418.0	1.0	0.5	0.8	2	3	54	00:01:17.288	9.2	10140752 8
6.26.0	DIO	3899	1060561	424.8	0.9	0.5	0.8	2	3	22	00:01:04.536	7.7	10140276 4
6.25.0	None	3989	1032155	413.4	1.0	0.5	0.8	2	3	102	00:01:23.783	10.0	10140714 0
6.25.0	DIO	3899	1048824	420.1	1.0	0.5	0.8	2	3	22	00:01:04.747	7.7	10140276 4
6.24.0	None	3983	1025562	410.8	1.0	0.5	0.8	2	3	32	00:01:12.296	8.6	10140652 4
6.24.0	DIO	3880	1052049	421.4	1.0	0.5	0.8	2	3	22	00:01:05.862	7.8	10140506 4
6.23.0	None	4175	1015722	406.8	1.0	0.5	0.8	2	3	69	00:01:17.541	9.2	10140529 2
6.23.0	DIO	3885	1055232	422.7	0.9	0.5	0.8	2	3	21	00:00:52.360	6.2	10140211 6

Version	Opts	Time	ops/sec	mb/sec	usec/op	p50	p75	p99	p99.9	p99.99	Stall-time	Stall %	du -sk
6.22.1	None	4143	1013002	405.8	1.0	0.5	0.8	2	3	224	00:01:26.032	10.2	101405804
6.22.1	DIO	4058	1031703	413.2	1.0	0.5	0.8	2	3	125	00:01:23.019	9.9	101403424
6.21.2	None	4141	1017259	407.5	0.9	0.5	0.8	2	3	556	00:01:32.279	11.0	101406320
6.15.5	None	4068	1045195	418.6	1.0	0.5	0.8	1	3	980	00:02:08.223	15.3	101401808
6.10.4	None	4002	1062310	425.5	0.9	0.5	0.8	1	3	1013	00:02:24.652	17.2	101402936

Test 2. Random Read (benchmark.sh readrandom)

NUM_KEYS=900000000 CACHE_SIZE=6442450944 DURATION=5400 benchmark.sh readrandom

Measure performance to randomly read existing keys. The database after bulkload was used as the starting point.

Version	Opts	ops/sec	mb/sec	usec/op	p50	p75	p99	p99.9	p99.99
7.2.2	None	136915	34.7	467.4	615.5	772.8	1270	1801	2840
7.2.2	DIO	189236	47.9	338.2	419.6	539.1	1022	1693	2297
7.1.1	None	145490	36.8	439.9	599.7	753.7	1252	1809	2813
7.1.1	DIO	189242	47.9	338.2	419.0	539.1	1037	1696	2294
7.0.3	None	145540	36.8	439.7	599.8	753.3	1251	1803	2803
7.0.3	DIO	189243	47.9	338.2	419.2	539.2	1029	1691	2246
6.29.1	None	145577	36.9	439.6	606.3	751.0	1204	1292	2091
6.29.1	DIO	189243	47.9	338.2	430.0	540.9	854	969	1291
6.29.0	None	145590	36.9	439.6	606.2	751.0	1204	1292	1936

Version	Opts	ops/sec	mb/sec	usec/op	p50	p75	p99	p99.9	p99.99
6.29.0	DIO	189241	47.9	338.2	430.0	540.8	854	932	1289
6.28.0	None	146980	37.2	435.4	604.3	748.9	1195	1291	1984
6.28.0	DIO	189232	47.9	338.2	430.0	540.9	854	991	1293
6.27.0	None	146921	37.2	435.6	604.4	748.8	1194	1291	1980
6.27.0	DIO	189250	47.9	338.2	430.1	540.8	854	902	1287
6.26.0	None	128341	32.5	498.7	639.6	805.7	1272	1298	2156
6.26.0	DIO	189244	47.9	338.2	430.1	540.8	854	894	1287
6.25.0	None	128517	32.5	498.0	639.0	804.6	1272	1298	2220
6.25.0	DIO	189245	47.9	338.2	430.1	540.8	854	897	1289
6.24.0	None	130852	33.1	489.1	632.6	791.4	1266	1297	2152
6.24.0	DIO	189240	47.9	338.2	430.0	540.7	854	930	1292
6.23.0	None	137664	34.9	464.9	618.4	766.5	1244	1295	2557
6.23.0	DIO	189252	47.9	338.2	430.0	540.7	854	926	1296
6.22.1	None	138623	35.1	461.7	616.8	763.9	1239	1295	2663
6.22.1	DIO	189237	47.9	338.2	430.0	540.7	854	960	1291
6.21.2	None	138633	35.1	461.6	616.8	764.1	1240	1295	2461
6.15.5	None	138513	35.1	462.0	616.9	764.2	1240	1295	3083
6.10.4	None	138496	35.1	462.1	617.1	764.3	1240	1295	2484

Test 3. Multi-Random Read (benchmark.sh multireadrandom)

NUM_KEYS=900000000 CACHE_SIZE=6442450944 DURATION=5400 benchmark.sh multireadrandom --multiread_batched

Measure performance to randomly multi-get existing keys. The database after bulkload was used as the starting point.

Version	Opts	ops/sec	p50	p75	p99	p99.9	p99.99
7.2.2	None	136928	4657.7	5774.9	9416	9873	18001
7.2.2	DIO	189216	3415.8	4064.2	6422	6586	8602
7.1.1	None	145548	4387.8	5568.6	8899	9831	16943
7.1.1	DIO	189213	3413.7	4064.2	6422	6586	8630
7.0.3	None	145587	4386.7	5567.1	8886	9829	16789
7.0.3	DIO	189230	3413.5	4063.8	6422	6586	8590
6.29.1	None	145652	4376.9	5549.2	8702	9813	15243
6.29.1	DIO	189233	3410.8	4048.6	6406	6583	7498
6.29.0	None	145660	4376.7	5549.1	8701	9811	15305
6.29.0	DIO	189231	3410.3	4048.0	6406	6583	7310
6.28.0	None	147022	4345.6	5523.9	8584	9804	14594
6.28.0	DIO	189228	3410.5	4048.4	6406	6583	7679
6.27.0	None	146989	4346.2	5524.0	8579	9806	15833
6.27.0	DIO	189227	3409.6	4047.2	6405	6583	7332
6.26.0	None	128366	4933.4	6093.6	9672	9884	13845
6.26.0	DIO	189229	3409.0	4046.8	6405	6583	7282
6.25.0	None	128523	4927.2	6087.8	9670	9883	13727
6.25.0	DIO	189241	3408.7	4046.6	6404	6583	7525
6.24.0	None	130859	4836.9	5995.1	9630	9880	14169
6.24.0	DIO	189234	3409.0	4047.2	6406	6584	7996
6.23.0	None	137638	4607.1	5736.4	9360	9869	17172
6.23.0	DIO	189237	3409.0	4047.0	6406	6584	8125
6.22.1	None	138660	461.6	4576.1	5706.3	9294	9867
6.22.1	DIO	189235	338.2	3410.7	4047.7	6406	6583
6.21.2	None	138623	461.7	4577.4	5707.2	9294	9866
6.15.5	None	138507	462.0	4582.3	5710.2	9299	9867
6.10.4	None	138476	462.1	4583.0	5710.9	9298	9864

Test 4. Range Scan (benchmark.sh fwdrange)

NUM_KEYS=900000000 CACHE_SIZE=6442450944 DURATION=5400 benchmark.sh fwdrange

Measure performance to randomly iterate over keys. The database after bulkload was used as the starting point.

Version	Opts	ops/sec	mb/sec	usec/op	p50	p75	p99	p99.9	p99.99
7.2.2	None	70097	280.8	913.0	791.9	1435.5	1892	2811	10210
7.2.2	DIO	78828	315.7	811.9	836.9	1093.2	1771	2601	2894
7.1.1	None	74491	298.4	859.1	775.3	1380.6	1889	1899	8592
7.1.1	DIO	78831	315.8	811.8	836.7	1093.1	1771	2598	2982
7.0.3	None	74510	298.4	858.9	775.2	1380.9	1889	2786	8384
7.0.3	DIO	78832	315.8	811.8	836.8	1093.1	1771	2603	2895
6.29.1	None	74530	298.5	858.7	775.8	1392.9	1881	1899	7807
6.29.1	DIO	78830	315.7	811.8	870.2	1090.8	1434	1862	2668
6.29.0	None	74535	298.5	858.6	775.7	1393.3	1881	1899	7553
6.29.0	DIO	78832	315.8	811.8	870.3	1090.3	1388	1858	2620
6.28.0	None	75231	301.3	850.7	773.7	1381.2	1880	1899	8224
6.28.0	DIO	78828	315.7	811.9	870.2	1090.8	1438	1862	2655
6.27.0	None	75246	301.4	850.5	773.2	1384.2	1880	1899	8360
6.27.0	DIO	78829	315.7	811.9	870.5	1090.2	1373	1855	2513
6.26.0	None	65717	263.2	973.8	808.2	1492.5	1884	1899	7217
6.26.0	DIO	78831	315.8	811.8	870.5	1090.2	1370	1855	2512
6.25.0	None	65813	263.6	972.4	807.7	1491.5	1884	1899	7338
6.25.0	DIO	78833	315.8	811.8	870.4	1090.2	1369	1856	2610
6.24.0	None	67004	268.4	955.1	802.8	1480.1	1884	1899	7216
6.24.0	DIO	78832	315.8	811.8	870.4	1090.2	1376	1856	2582
6.23.0	None	70459	282.2	908.3	789.6	1443.0	1883	1899	10273
6.23.0	DIO	78829	315.7	811.9	870.5	1090.1	1356	1855	2596

Version	Opts	ops/sec	mb/sec	usec/op	p50	p75	p99	p99.9	p99.99
6.22.1	None	70971	284.3	901.7	787.8	1437.2	1882	1899	10274
6.22.1	DIO	78829	315.7	811.9	870.3	1090.5	1411	1859	2618
6.21.2	None	70967	284.3	901.8	787.8	1437.3	1882	1899	10253
6.15.5	None	70978	284.3	901.7	787.7	1437.5	1882	1899	9890
6.10.4	None	70973	284.3	901.7	787.6	1438.0	1882	1899	9945

Test 4b. Reverse Range Scan (benchmark.sh revrange)

NUM_KEYS=900000000 CACHE_SIZE=6442450944 DURATION=5400 benchmark.sh revrange

Measure performance to randomly iterate over keys. The database after bulkload was used as the starting point.

Version	Opts	ops/sec	mb/sec	usec/op	p50	p75	p99	p99.9	p99.99
7.2.2	None	68785	275.5	930.4	806.0	1467.2	1892	2859	12052
7.2.2	DIO	76200	305.2	839.9	897.5	1114.2	1776	2617	2898
7.1.1	None	73116	292.9	875.3	788.1	1399.9	1889	2853	13338
7.1.1	DIO	76202	305.2	839.8	897.1	1114.1	1778	2631	3022
7.0.3	None	73149	293.0	874.9	788.0	1399.4	1889	2853	13524
7.0.3	DIO	76202	305.2	839.8	897.4	1114.0	1776	2632	3173
6.29.1	None	73167	293.1	874.7	788.9	1406.9	1882	1900	12818
6.29.1	DIO	76204	305.2	839.8	910.2	1112.1	1562	1874	2764
6.29.0	None	73170	293.1	874.6	788.6	1409.1	1882	1899	12688
6.29.0	DIO	76202	305.2	839.8	910.2	1111.5	1524	1870	2722
6.28.0	None	73839	295.8	866.7	786.5	1391.8	1881	1900	13492
6.28.0	DIO	76205	305.2	839.8	910.0	1112.1	1560	1873	2715
6.27.0	None	73861	295.8	866.5	786.0	1396.4	1881	1899	13488

Version	Opts	ops/sec	mb/sec	usec/op	p50	p75	p99	p99.9	p99.99
6.27.0	DIO	76204	305.2	839.8	910.3	1111.3	1510	1869	2718

Test 5. Overwrite (benchmark.sh overwrite)

NUM_KEYS=900000000 CACHE_SIZE=6442450944 DURATION=5400 benchmark.sh overwrite

Measure performance to randomly overwrite keys into the database. The database was first created by the previous benchmark.

Version	Opts	ops/sec	mb/sec	W-Amp	W-MB/s	usec/op	p50	p75	p99	p99.9	p99.99	Stall-time	Stall%	du -s -k
7.2.2	None	86617	34.7	9.5	149.7	738.9	449.7	777.6	10479	30005	58328	00:04:43.188	5.3	15854004 8
7.2.2	DIO	86839	34.8	9.4	154.6	737.0	460.7	775.4	9534	29149	54278	00:03:00.102	3.4	15913583 2
7.1.1	None	90203	36.1	9.4	155.4	709.5	418.5	746.0	10469	30015	58380	00:04:45.549	5.3	16099294 4
7.1.1	DIO	88590	35.5	9.6	154.6	722.4	440.2	754.0	9538	29313	55494	00:04:15.453	4.8	15837216 4
7.0.3	None	90985	36.4	9.4	155.8	703.4	418.2	743.9	10156	29987	57788	00:04:48.049	5.3	16111071 6
7.0.3	DIO	89686	35.9	9.5	154.1	713.6	439.3	752.3	9377	20921	53505	00:03:28.796	3.9	16035672 0
6.29.1	None	90711	36.3	9.4	155.1	705.5	418.2	740.0	10213	29779	57100	00:05:28.848	6.1	16109979 2
6.29.1	DIO	89469	35.8	9.5	154.4	715.3	431.6	748.0	9568	29143	54324	00:04:06.106	4.6	15966117 2

Version	Opts	ops/sec	mb/sec	W-Amp	W-MB/s	usec/op	p50	p75	p99	p99.9	p99.99	Stall-time	Stall%	du -s -k
6.29.0	None	89373	35.8	9.6	155.3	716.1	434.1	756.1	10447	29891	57952	00:04:21.622	4.9	15891285 6
6.29.0	DIO	88517	35.5	9.5	152.6	723.0	455.1	759.1	9219	28694	51750	00:03:31.235	3.9	16025877 2
6.28.0	None	89791	36.0	9.4	153.7	712.4	430.9	751.9	10292	29839	58737	00:04:15.276	4.8	16185985 6
6.28.0	DIO	88108	35.3	9.5	152.4	726.4	449.4	763.5	9508	28917	54122	00:03:25.719	3.8	15986544 0
6.27.0	None	89815	36.0	9.5	154.1	712.6	427.7	749.4	10533	29660	57615	00:04:30.399	5.0	16027377 2
6.27.0	DIO	88440	35.4	9.4	151.6	723.6	455.0	761.2	9383	28764	52844	00:03:20.977	3.7	15957248 4
6.26.0	None	90340	36.2	9.4	153.6	708.4	430.2	742.5	10198	29692	55635	00:04:54.193	5.5	16120243 2
6.26.0	DIO	88401	35.4	9.6	154.5	724.0	446.0	754.6	9418	28911	52526	00:03:50.428	4.3	15846967 2
6.25.0	None	89567	35.9	9.4	155.2	714.5	419.4	742.7	10327	29952	59957	00:05:52.335	6.5	16039224 4
6.25.0	DIO	88549	35.5	9.5	153.6	722.7	433.9	743.6	9483	29064	54109	00:05:00.728	5.6	15850048 8
6.24.0	None	90829	36.4	4.7	155.2	704.6	397.1	726.4	10359	29968	58160	00:07:01.849	7.9	16075704 8
6.24.0	DIO	90105	36.1	4.8	153.7	710.3	421.8	736.9	9344	28869	52676	00:05:22.128	6.0	16083357 2
6.23.0	None	89052	35.7	4.7	151.3	718.7	442.5	758.8	10263	29763	53874	00:04:40.429	5.2	16063319 6
6.23.0	DIO	88624	35.5	4.9	152.4	722.1	441.5	749.0	9319	28887	53792	00:04:53.783	5.5	15899450 8

Version	Opts	ops/sec	mb/sec	W-Amp	W-MB/s	usec/op	p50	p75	p99	p99.9	p99.99	Stall-time	Stall %	du -s -k
6.22.1	None	91586	36.7	4.7	155.0	698.8	380.5	709.4	10140	29887	58244	00:08:29.153	9.5	161321740
6.22.1	DIO	90310	36.2	4.8	154.7	708.7	419.0	730.1	9227	28816	55513	00:06:22.790	7.1	160400436
6.21.2	None	91776	36.8	4.7	155.6	697.3	379.9	708.7	10055	29782	55942	00:08:24.882	9.4	162082088
6.15.5	None	92911	37.2	4.7	158.4	688.8	351.9	697.7	10031	29894	58333	00:08:43.334	9.7	161156844
6.10.4	None	94539	37.9	4.7	161.9	676.9	328.4	700.4	10022	29843	56548	00:07:11.226	8.0	162965216

Test 6. Multi-threaded read and single-threaded write (benchmark.sh readwhilewriting)

NUM_KEYS=900000000 CACHE_SIZE=6442450944 DURATION=5400 MB_WRITE_PER_SEC=2 benchmark.sh readwhilewriting

Measure performance with one writer and multiple reader threads. The writes are rate limited.

Version	Opts	ops/sec	mb/sec	W-Amp	W-MB/s	usec/op	p50	p75	p99	p99.9	p99.99	du -s -k
7.2.2	None	98240	31.1	18.1	11.4	651.4	600.6	829.8	3963	6041	10139	140646588
7.2.2	DIO	143283	45.3	17.1	7.3	446.7	394.8	539.8	2820	4315	6393	140470436

Version	Opts	ops/sec	mb/sec	W-Amp	W-MB/s	usec/op	p50	p75	p99	p99.9	p99.99	du -sk
7.1.1	None	102056	32.5	16.9	10.6	627.1	584.8	803.2	3931	6031	9844	141627716
7.1.1	DIO	142958	45.3	17.9	7.6	447.7	395.6	540.1	2819	4316	6405	140849884
7.0.3	None	101948	32.5	17.0	10.7	627.7	585.6	803.6	3931	6028	9824	141767112
7.0.3	DIO	142923	45.4	18.2	7.8	447.8	393.0	539.4	2825	4322	6414	141164436
6.29.1	None	100445	31.8	28.3	18.2	637.1	593.0	810.3	3906	6524	18544	140795968
6.29.1	DIO	141799	44.8	31.5	14.2	451.3	397.1	541.4	2792	4744	9095	140017864
6.29.0	None	100853	31.9	27.7	17.7	634.6	592.9	810.8	3893	6480	17827	140416272
6.29.0	DIO	141947	44.8	32.6	14.4	450.9	397.4	542.1	2786	4791	9273	139972676
6.28.0	None	101233	32.0	28.3	18.3	632.2	591.6	807.0	3892	6530	19073	140616192
6.28.0	DIO	141854	44.7	35.0	15.2	451.2	394.1	541.4	2796	5000	9696	139803484
6.27.0	None	101375	32.1	27.7	17.8	631.3	587.9	805.0	3893	6477	18216	140673616
6.27.0	DIO	142460	44.9	31.2	14.1	449.2	394.7	539.6	2789	4685	9127	139867840
6.26.0	None	91879	29.1	27.7	19.0	696.5	630.8	904.6	4010	6424	13872	140615968
6.26.0	DIO	142148	44.8	31.8	14.4	450.2	394.7	540.0	2793	4697	8939	139826380
6.25.0	None	91736	29.0	28.7	20.0	697.6	630.8	906.2	4019	6418	13775	140615968
6.25.0	DIO	141618	44.7	33.0	14.9	451.9	394.4	540.8	2800	4825	9113	140031428
6.24.0	None	92974	29.5	27.6	19.0	688.3	624.8	869.7	4010	6436	14558	140384360
6.24.0	DIO	141491	44.7	32.7	15.0	452.3	395.8	540.8	2802	4867	9311	140255568
6.23.0	None	96811	30.6	29.1	18.9	661.1	607.3	835.3	3966	6433	13513	140384360

Version	Opts	ops/sec	mb/sec	W-Amp	W-MB/s	usec/op	p50	p75	p99	p99.9	p99.99	du -s-k
6.23.0	DIO	142410	44.9	29.6	13.5	449.4	394.0	539.3	2789	4598	8989	139961824
6.22.1	None	96812	30.7	28.4	18.5	661.1	606.5	832.8	3958	6500	15777	140972560
6.22.1	DIO	140635	44.5	32.5	14.9	455.1	400.4	543.4	2804	5051	9348	140465744
6.21.2	None	96891	30.7	29.1	18.9	660.5	607.1	833.4	3961	6465	13669	141208940
6.15.5	None	96223	30.6	28.0	18.7	665.1	609.4	835.1	3965	6475	15613	141339712
6.10.4	None	95649	30.5	30.2	19.7	669.1	608.1	834.5	3999	6597	17861	141636760

Test 7. Multi-threaded scan and single-threaded write (benchmark.sh fwdrangewhilewriting)

NUM_KEYS=900000000 CACHE_SIZE=6442450944 DURATION=5400 MB_WRITE_PER_SEC=2 benchmark.sh fwdrangewhilewriting

Measure performance with one writer and multiple iterator threads. The writes are rate limited.

Version	Opts	ops/sec	mb/sec	W-Amp	W-MB/s	usec/op	p50	p75	p99	p99.9	p99.99	du -s-k
7.2.2	None	40675	162.9	17.4	7.4	1573.4	1374.6	1855.1	6293	13434	24996	141346104
7.2.2	DIO	35619	142.7	18.3	7.5	1796.6	1540.7	2171.8	6533	9698	13325	140957044
7.1.1	None	42202	169.0	16.5	7.3	1516.4	1322.2	1821.4	6168	13098	25099	142336676

Version	Opts	ops/sec	mb/sec	W-Amp	W-MB/s	usec/op	p50	p75	p99	p99.9	p99.99	du -sk
7.1.1	DIO	35535	142.3	17.9	7.3	1800.8	1544.5	2175.7	6527	9691	13298	141591172
7.0.3	None	42436	170.0	17.3	7.8	1508.0	1310.0	1815.5	6198	13226	24937	142579812
7.0.3	DIO	35702	143.0	18.9	7.8	1792.5	1535.0	2165.4	6531	9702	13343	141636716
6.29.1	None	43138	172.8	17.6	7.8	1483.5	1294.3	1804.6	6089	13026	25065	141561940
6.29.1	DIO	36460	146.0	16.5	7.1	1755.2	1517.8	2128.9	6381	9572	12979	140761644
6.29.0	None	42806	171.5	17.0	7.6	1495.0	1311.0	1813.2	6101	13108	25308	140416272
6.29.0	DIO	36418	145.9	17.4	7.4	1757.2	1522.1	2124.5	6404	9624	13210	140619752
6.28.0	None	43564	174.5	17.3	7.7	1469.0	1282.2	1794.6	6055	12926	24865	141241492
6.28.0	DIO	36230	145.1	17.7	7.5	1766.3	1527.9	2142.4	6439	9653	13251	140537532
6.27.0	None	43229	173.2	18.3	8.3	1480.4	1290.3	1802.6	6123	13140	24987	141261580
6.27.0	DIO	35860	143.6	16.9	7.4	1784.5	1540.3	2181.7	6422	9603	13041	140542060
6.26.0	None	36960	148.0	17.2	8.2	1731.4	1534.2	2217.9	6477	13239	21533	141557396
6.26.0	DIO	35961	144.0	16.5	7.3	1779.5	1536.8	2174.2	6415	9603	13055	140627180
6.25.0	None	37344	149.6	17.9	8.4	1713.7	1513.5	2164.7	6489	13409	21654	141269488
6.25.0	DIO	36023	144.3	17.9	7.8	1776.5	1532.5	2162.0	6458	9672	13308	140685060
6.24.0	None	38940	156.0	17.7	8.1	1643.4	1445.8	1970.8	6411	13480	21757	141724476
6.24.0	DIO	35955	144.0	17.1	7.5	1779.8	1534.8	2173.0	6427	9615	13093	140989196
6.23.0	None	41322	165.5	16.7	7.6	1584.7	1359.9	1838.4	6225	13285	24338	141101776
6.23.0	DIO	35968	144.1	17.0	7.4	1779.2	1536.8	2167.6	6446	9648	13218	140731428

Version	Opts	ops/sec	mb/sec	W-Amp	W-MB/s	usec/op	p50	p75	p99	p99.9	p99.99	du -sk
6.22.1	None	41244	165.2	16.7	7.6	1551.6	1362.3	1845.8	6234	13346	24988	141716340
6.22.1	DIO	35962	144.0	17.0	7.4	1779.5	1538.8	2150.6	6455	9661	13264	141008104
6.21.2	None	41360	165.7	18.2	8.0	1547.2	1354.1	1840.5	6257	13434	25280	141820100
6.15.5	None	42197	169.0	17.5	8.0	1516.6	1315.5	1817.8	6185	13018	23081	142157224
6.10.4	None	41827	167.5	17.6	8.0	1530.0	1329.2	1826.2	6212	13129	23244	142497356

Test 7b. Multi-threaded scan and single-threaded write (benchmark.sh revrangewhilewriting)

NUM_KEYS=900000000 CACHE_SIZE=6442450944 DURATION=5400 MB_WRITE_PER_SEC=2 benchmark.sh revrangewhilewriting

Measure performance with one writer and multiple iterator threads. The writes are rate limited.

Version	Opts	ops/sec	mb/sec	W-Amp	W-MB/s	usec/op	p50	p75	p99	p99.9	p99.99	du -sk
7.2.2	None	33680	134.9	17.3	7.5	1900.1	1668.2	2417.1	7207	16605	29880	142066536
7.2.2	DIO	31215	125.0	16.4	6.9	2050.0	1755.3	2528.0	7637	10178	13981	141817860
7.1.1	None	34825	139.5	17.4	7.7	1837.6	1623.9	2360.8	6742	16569	30135	142975980
7.1.1	DIO	31259	125.2	17.4	7.3	2047.2	1744.1	2520.7	7673	10205	13980	142349268

Version	Opts	ops/sec	mb/sec	W-Amp	W-MB/s	usec/op	p50	p75	p99	p99.9	p99.99	du -sk
7.0.3	None	35015	140.3	17.5	7.8	1827.6	1614.4	2345.4	6695	16540	29947	143211480
7.0.3	DIO	31155	124.8	15.9	7.0	2054.0	1753.5	2529.0	7627	9970	13948	142568752
6.29.1	None	35535	142.3	17.6	7.8	1800.8	1598.9	2320.1	6572	16547	30132	142191812
6.29.1	DIO	31839	127.5	16.8	7.3	2009.9	1731.9	2489.0	7184	9882	13917	141520096
6.29.0	None	35676	142.9	17.3	7.8	1793.8	1592.1	2307.6	6569	16711	30812	141867556
6.29.0	DIO	31896	127.8	17.8	7.8	2006.3	1728.1	2483.7	7320	10017	13938	141162940
6.28.0	None	35882	143.7	16.6	7.5	1783.4	1588.0	2292.4	6534	16405	30385	142078076
6.28.0	DIO	31855	127.6	16.5	7.3	2008.9	1727.1	2485.2	7287	10012	14000	141298660
6.27.0	None	35594	142.6	17.0	7.7	1797.9	1596.8	2315.0	6566	16418	29917	142075232
6.27.0	DIO	32086	128.5	17.0	7.4	1994.4	1717.5	2470.2	7116	9865	13867	141261580

Appendix

fio test results

```
]$ fio --randrepeat=1 --ioengine=sync --direct=1 --gtod_reduce=1 --name=test --
filename=/data/test_file --bs=4k --iodepth=64 --size=4G --readwrite=randread --numjobs=32 --
group_reporting
test: (g=0): rw=randread, bs=4K-4K/4K-4K/4K-4K, ioengine=sync, iodepth=64
...
fio-2.14
Starting 32 processes
Jobs: 3 (f=3): [_(3),r(1),_(1),E(1),_(10),r(1),_(13),r(1),E(1)] [100.0% done] [445.3MB/0KB/0KB
/s] [114K/0/0 iops] [eta 00m:00s]
test: (groupid=0, jobs=32): err= 0: pid=28042: Fri Jul 24 01:36:19 2020
```

```
  read : io=131072MB, bw=469326KB/s, iops=117331, runt=285980msec
  cpu          : usr=1.29%, sys=3.26%, ctx=33585114, majf=0, minf=297
  IO depths    : 1=100.0%, 2=0.0%, 4=0.0%, 8=0.0%, 16=0.0%, 32=0.0%, >=64=0.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     issued    : total=r=33554432/w=0/d=0, short=r=0/w=0/d=0, drop=r=0/w=0/d=0
     latency   : target=0, window=0, percentile=100.00%, depth=64

Run status group 0 (all jobs):
   READ: io=131072MB, aggrb=469325KB/s, minb=469325KB/s, maxb=469325KB/s, mint=285980msec,
maxt=285980msec

Disk stats (read/write):
  nvme1n1: ios=33654742/61713, merge=0/40, ticks=8723764/89064, in_queue=8788592, util=100.00%
```

```
]$ fio --randrepeat=1 --ioengine=libaio --direct=1 --gtod_reduce=1 --name=test --
filename=/data/test_file --bs=4k --iodepth=64 --size=4G --readwrite=randread
test: (g=0): rw=randread, bs=4K-4K/4K-4K/4K-4K, ioengine=libaio, iodepth=64
fio-2.14
Starting 1 process
Jobs: 1 (f=1): [r(1)] [100.0% done] [456.3MB/0KB/0KB /s] [117K/0/0 iops] [eta 00m:00s]
test: (groupid=0, jobs=1): err= 0: pid=28385: Fri Jul 24 01:36:56 2020
  read : io=4096.0MB, bw=547416KB/s, iops=136854, runt=  7662msec
  cpu          : usr=22.20%, sys=48.81%, ctx=144112, majf=0, minf=73
  IO depths    : 1=0.1%, 2=0.1%, 4=0.1%, 8=0.1%, 16=0.1%, 32=0.1%, >=64=100.0%
     submit    : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.0%, >=64=0.0%
     complete  : 0=0.0%, 4=100.0%, 8=0.0%, 16=0.0%, 32=0.0%, 64=0.1%, >=64=0.0%
     issued    : total=r=1048576/w=0/d=0, short=r=0/w=0/d=0, drop=r=0/w=0/d=0
     latency   : target=0, window=0, percentile=100.00%, depth=64

Run status group 0 (all jobs):
   READ: io=4096.0MB, aggrb=547416KB/s, minb=547416KB/s, maxb=547416KB/s, mint=7662msec,
maxt=7662msec

Disk stats (read/write):
  nvme1n1: ios=1050868/1904, merge=0/1, ticks=374836/2900, in_queue=370532, util=98.70%
```

Revision #1
Created 23 September 2023 10:49:55 by Laptrinh.vn
Updated 23 September 2023 10:56:18 by Laptrinh.vn