LightDB 麒麟 XFS mmp 风险解决方案

目录

风险问题

2023 年 12 月 6 日，通过充分的证据链证明了 xfs 在 mmap 操作下可能导致的数据不一致问题，影响操作系统版本：kylinV10 SP2 aarch64、kylinV10 SP2 x86_64。详细信息如下：

http://www.light-pg.com/warningPages/1.html

解决方案

将 LightDB 的实例级别 GUC 参数 shared_memory_type 由 mmap 改为 **sysv**（修改 $LTDATA/lightdb.conf 配置文件），重启 LightDB 集群。

性能测试

基于 shared_memory_type 参数改为 **sysv**，现通过 BenchmarkSQL 和 sysBench 测试工具来对 LightDB 进行 TPC-C 和 OLAP 性能测试。

服务器配置

配置	操作系统	机器属性	用途	部署服务	网络
96c Kunpeng aarch64,ky10sp1, NUMA0 756GB 内存 3T SSD 磁盘 nvme	Kylin Linux Advanced Server release V10 (Tercel)	物理机	1）部署数据库 2）测试工具	1）LightDB 2）Benchmarksql 3）sysBench	万兆

测试环境

安装单机实例，LightDB 版本为 V202303.04.000；

BenchmarkSQL 版本为 5.1。

参数配置

系统参数配置

kernel.core_pattern=core.%p

```
kernel.sysrq=0
net.ipv4.ip_forward=0
net.ipv4.conf.all.send_redirects=0
net.ipv4.conf.default.send_redirects=0
net.ipv4.conf.all.accept_source_route=0
net.ipv4.conf.default.accept_source_route=0
net.ipv4.conf.all.accept_redirects=0
net.ipv4.conf.default.accept_redirects=0
net.ipv4.conf.all.secure_redirects=0
net.ipv4.conf.default.secure_redirects=0
net.ipv4.icmp_echo_ignore_broadcasts=1
net.ipv4.icmp_ignore_bogus_error_responses=1
net.ipv4.conf.all.rp_filter=1
net.ipv4.conf.default.rp_filter=1
net.ipv4.tcp_syncookies=1
kernel.dmesg_restrict=1
net.ipv6.conf.all.accept_redirects=0
net.ipv6.conf.default.accept_redirects=0

#vm.nr_hugepages=360
vm.nr_hugepages=0
net.core.somaxconn = 2000
vm.overcommit_memory=2
fs.file-max=524288
kernel.sem=500 2048000 200 4096
kernel.shmmni=4096
fs.aio-max-nr=1048576
vm.swappiness=5
vm.overcommit_ratio=90
```

vm.dirty_background_ratio=5

vm.dirty_ratio=40

vm.dirty_expire_centisecs=500

vm.dirty_writeback_centisecs=250

net.ipv4.tcp_syn_retries=3

net.ipv4.tcp_retries2=5

net.ipv4.tcp_slow_start_after_idle=0

net.ipv4.tcp_tw_reuse=1

kernel.shmmni=4096

kernel.shmmax=711022721024

kernel.shmall=6271709

kernel.sem=500 2048000 200 4096

fs.aio-max-nr=1048576

fs.file-max=524288

vm.swappiness=5

vm.overcommit_memory=2

vm.overcommit_ratio=85

vm.dirty_background_ratio=5

vm.dirty_ratio=40

vm.dirty_expire_centisecs=500

vm.dirty_writeback_centisecs=250

net.core.somaxconn=2000

net.ipv4.tcp_max_syn_backlog=2000

net.ipv4.tcp_tw_reuse=1

net.ipv4.tcp_syn_retries=3

net.ipv4.tcp_retries2=5

net.ipv4.tcp_slow_start_after_idle=0

数据库参数配置（关键参数，详见附录）

cron.database_name = 'postgres'

enable_incremental_checkpoint=on

enable_double_write=on

pagewriter_sleep=4000

auto_explain.log_min_duration = '1s'

auto_explain.log_level=LOG

auto_explain.log_format=json

lt_show_plans.show_level = 'top'

max_wal_size=117085MB

lightdb_syntax_compatible_type='oracle'

parallel_setup_cost=10000

transform_null_equals=on

max_worker_processes=192

shared_buffers=240GB

lightdb_external_virtual_ip=''

min_parallel_table_scan_size=2GB

temp_buffers=64MB

lt_stat_statements.track_planning=on

bgwriter_lru_maxpages=15202

max_slot_wal_keep_size=353153MB

max_parallel_workers=192

wal_buffers=128MB

random_page_cost=1.0

logging_collector=on

min_wal_size=117718MB

log_min_messages=info

default_statistics_target=256

track_io_timing=on

shared_preload_libraries='lt_stat_statements,lt_stat_activity,lt_prewarm,lt_cro

n,lt_hint_plan,lt_show_plans,lt_sql_inspect'

```
enable_partitionwise_aggregate=on
min_parallel_index_scan_size=128MB
commit_siblings=10
effective_cache_size=549348MB
max_parallel_maintenance_workers=24
include_if_exists='lightdb.user.conf'
```

TPC-C 测试

数据准备

使用 BenchmarkSQL 准备数据。首先选择 TPC-C 和对应的数据库，然后配置数据库连接参数。本次测试使用 1000 个 Warehouses。

测试方法

每个数据库持续压力测试 30 分钟。客户端 200 个用户持续发送请求。

```
terminals=200
// To run specified transactions per terminal- runMins must equal zero
runTxnsPerTerminal=0
// To run for specified minutes- runTxnsPerTerminal must equal zero
runMins=30
// Number of total transactions per minute
limitTxnsPerMin=1000000
// Set to true to run in 4.x compatible mode. Set to false to use the
// entire configured database evenly.
terminalWarehouseFixed=false

// Set to true to use the stored procedure/function implementations. Not
// all of them exist for all databases and the use of stored procedures
// is strongly discouraged for comparing different database vendors as
// they may not have been implemented ideally for all of them. This is
```

```
// however useful to test how much network IO can be saved by using
// stored procedures.
useStoredProcedures=false
```

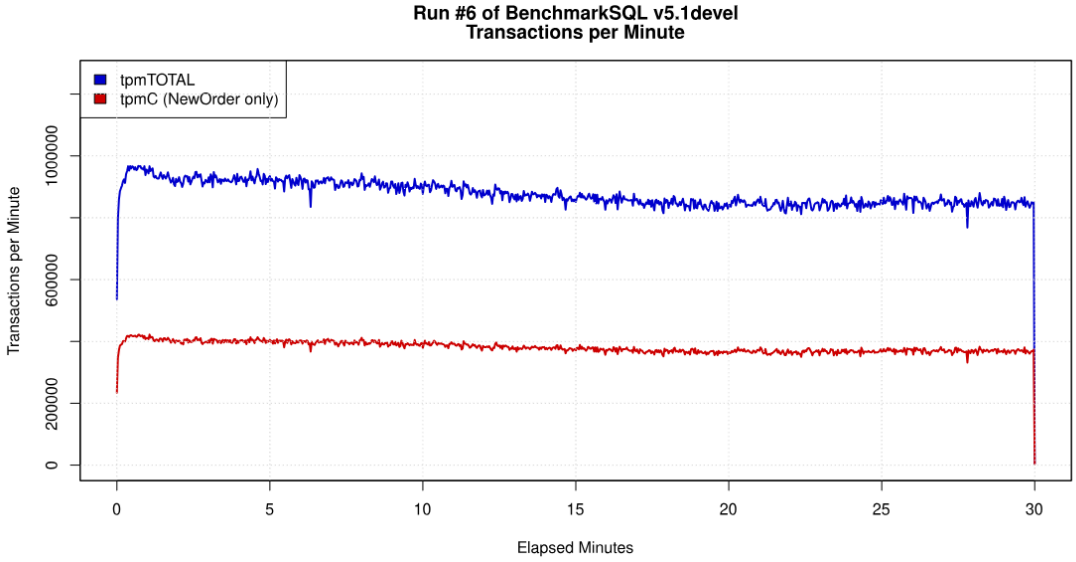
测试结果

采集每分钟吞吐量(tpm)和交易响应时间

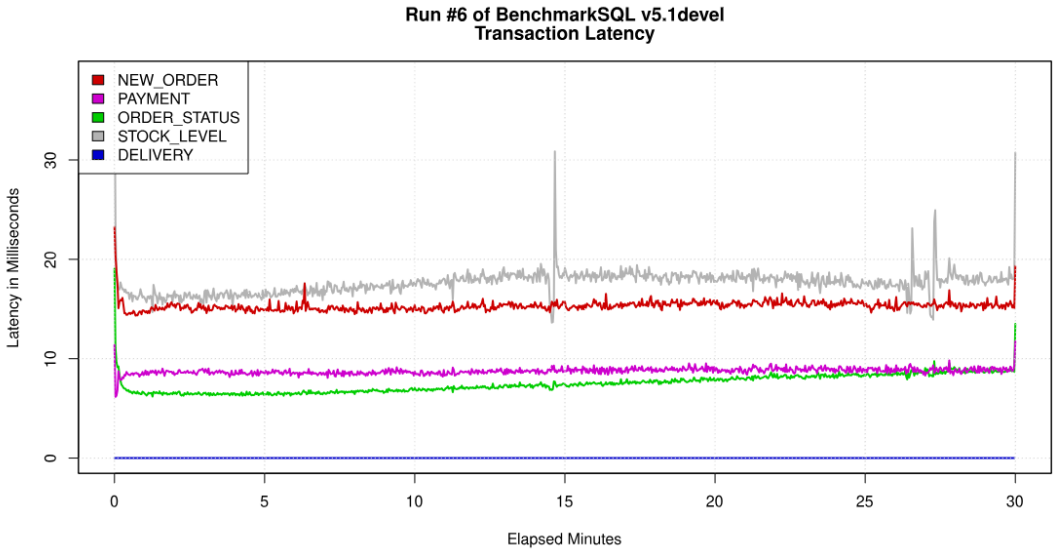
Transaction Type	Latency			Count	Percent	Rollback	Errors	Skipped Deliveries
	90th %	Avg	Max					
NEW_ORDER	0.023s	0.015s	0.238s	11435154	43.469%	0.999%	0	N/A
PAYMENT	0.015s	0.009s	0.241s	11437931	43.480%	N/A	0	N/A
ORDER_STATUS	0.011s	0.007s	0.212s	1145666	4.355%	N/A	0	N/A
STOCK_LEVEL	0.032s	0.018s	0.451s	1143552	4.347%	N/A	0	N/A
DELIVERY	0.000s	0.000s	0.011s	1144170	4.349%	N/A	0	N/A
DELIVERY_BG	0.065s	0.049s	0.459s	1144170	N/A	N/A	0	0

Overall tpmC: 381171.80
Overall tpmTotal: 876882.43

每分钟事务数



事务延时



CPU 资源利用率在 68%左右，每分钟事务量 **38.1w TpmC**。

OLTP 测试

数据准备

使用 sysBench 准备数据，共测试以下三种情况：**只读模式、只写模式、读写混合模式**。每个数据库持续压力测试 10 分钟。参数配置如下：

```
threads=100
time=600
percentile=99
rand-type=uniform
report-interval=10
tables=30
table_size=1000000
```

测试结果

1. **OLTP read_only:**

```
SQL statistics:
    queries performed:
        read:                            337785714
        write:                           0
        other:                           48255102
        total:                           386040816
    transactions:                        24127551 (40210.42 per sec.)
    queries:                             386040816 (643366.69 per sec.)
    ignored errors:                      0      (0.00 per sec.)
    reconnects:                          0      (0.00 per sec.)

General statistics:
    total time:                          600.0310s
    total number of events:              24127551

Latency (ms):
         min:                                    1.54
         avg:                                    2.49
         max:                                   52.75
         99th percentile:                        5.47
         sum:                                59958160.80

Threads fairness:
    events (avg/stddev):           241275.5100/32546.67
    execution time (avg/stddev):   599.5816/0.05
```

2. **OLTP read_write:**

```
SQL statistics:
    queries performed:
        read:                            175717122
        write:                           50202585
        other:                           25103613
        total:                           251023320
    transactions:                        12550653 (20915.51 per sec.)
    queries:                             251023320 (418327.34 per sec.)
    ignored errors:                      570    (0.95 per sec.)
    reconnects:                          0      (0.00 per sec.)

General statistics:
    total time:                          600.0631s
    total number of events:              12550653

Latency (ms):
         min:                                    2.36
         avg:                                    4.78
         max:                                   51.91
         95th percentile:                        6.55
         sum:                                59973130.30

Threads fairness:
    events (avg/stddev):           125506.5300/1241.08
    execution time (avg/stddev):   599.7313/0.01
```

3. **OLTP write_only:**

```
SQL statistics:
    queries performed:
        read:                            0
        write:                           137029414
        other:                           68530922
        total:                           205560336
    transactions:                        34257202 (57087.48 per sec.)
    queries:                             205560336 (342553.39 per sec.)
    ignored errors:                      4281   (7.13 per sec.)
    reconnects:                          0      (0.00 per sec.)

General statistics:
    total time:                          600.0813s
    total number of events:              34257202

Latency (ms):
         min:                                    0.67
         avg:                                    1.75
         max:                                  294.88
         95th percentile:                        2.91
         sum:                             59945629.31

Threads fairness:
    events (avg/stddev):           342572.0200/1836.54
    execution time (avg/stddev):   599.4563/0.01
```

测试结论

经测试，LightDB 的实例级别 GUC 参数改为 shared_memory_type=**sysv** 后，TPC-C 和 OLTP 性能与 shared_memory_type=mmap 基本一致，无性能下降、不稳定现象。