

Журналирование Контрольная точка



Авторские права

© Postgres Professional, 2016–2022.

Авторы: Егор Рогов, Павел Лузанов, Илья Баштанов

Использование материалов курса

Некоммерческое использование материалов курса (презентации, демонстрации) разрешается без ограничений. Коммерческое использование возможно только с письменного разрешения компании Postgres Professional. Запрещается внесение изменений в материалы курса.

Обратная связь

Отзывы, замечания и предложения направляйте по адресу:

edu@postgrespro.ru

Отказ от ответственности

Компания Postgres Professional не несет никакой ответственности за любые повреждения и убытки, включая потерю дохода, нанесенные прямым или непрямым, специальным или случайным использованием материалов курса. Компания Postgres Professional не предоставляет каких-либо гарантий на материалы курса. Материалы курса предоставляются на основе принципа «как есть» и компания Postgres Professional не обязана предоставлять сопровождение, поддержку, обновления, расширения и изменения.

Процесс контрольной точки

Процесс фоновой записи

Мониторинг

Необходимость контрольной точки

Процесс выполнения контрольной точки

Алгоритм восстановления после сбоя

Настройка

Размер хранимых журнальных записей

с какого сегмента начинать применять журнальные записи при восстановлении?

активно используемая страница может не вытесняться из кеша

Время восстановления

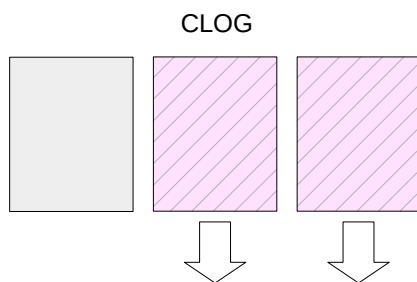
сколько времени займет восстановление после сбоя?

Если не принять специальных мер, то активно используемая страница, попав в буферный кеш, может никогда не вытесняться. Это означает, что при восстановлении после сбоя нам пришлось бы просматривать *все* журнальные записи, созданные с момента запуска сервера.

На практике это, конечно, недопустимо. Во-первых, файлы занимают много места — их все придется хранить на сервере. Во-вторых, время восстановления будет запредельно большим — придется просмотреть множество журнальных записей.

Поэтому и существует специальный фоновый процесс контрольной точки (checkpoint), который периодически сбрасывает все грязные страницы на диск (но не вытесняет их из кеша). После того как контрольная точка завершена, журнальные записи, предшествующие ее началу, больше не нужны для восстановления.

записываются грязные буферы CLOG

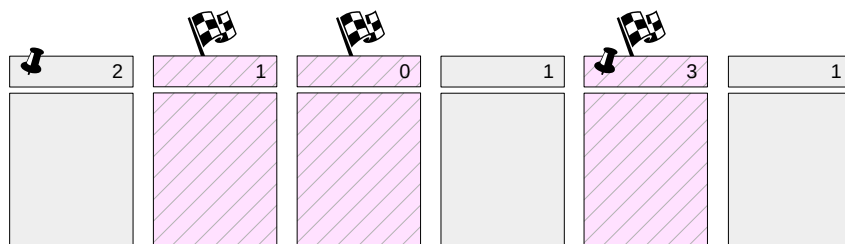


Рассмотрим подробнее, что происходит при выполнении контрольной точки.

Во-первых, помимо буферного кеша, в оперативной памяти располагаются и другие структуры, содержимое которых нужно сохранять на диске.

В частности, контрольная точка сбрасывает на диск буферы статуса транзакций (CLOG). Поскольку количество таких буферов невелико (их 128 штук), они записываются сразу же.

помечаются измененные страницы в буферном кеше



Во-вторых, основная работа контрольной точки — сбросить на диск все страницы из буферного кеша, которые были грязными на момент начала контрольной точки.

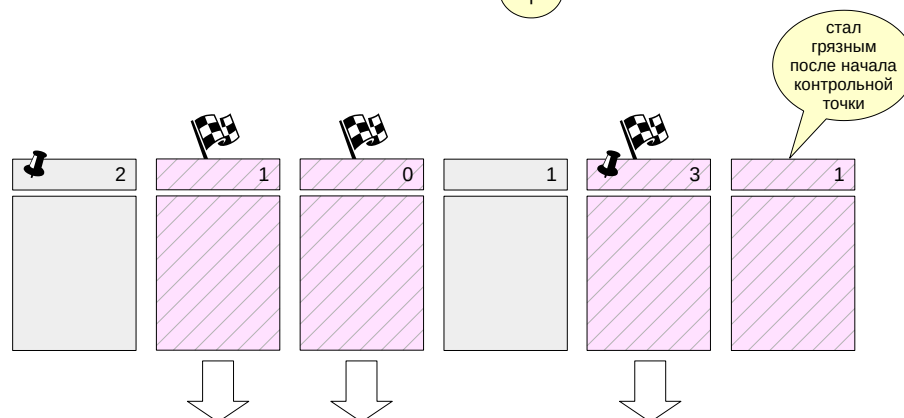
Поскольку размер буферного кеша может быть очень велик, сбрасывать сразу все страницы плохо — это приостановит нормальную работу сервера и создаст большую нагрузку на дисковую подсистему. Поэтому контрольная точка растягивается во времени и фактически превращается в отрезок.

Сначала все измененные на текущий момент страницы помечаются в заголовке буфера специальным флагом...

Процесс контрольной точки

помеченные страницы постепенно записываются,
пометка убирается из заголовка буфера

`checkpoint_completion_target = 0.5`



7

...а затем процесс контрольной точки *постепенно* проходит по всем буферам и сбрасывает помеченные на диск. Отметим, что страницы не вытесняются из кеша, а только записываются. Поэтому контрольная точка не обращает внимания на число обращений к буферу и признак закрепленности.

Помеченные буферы могут также быть записаны и обслуживающими процессами — смотря кто доберется до буфера первым. В любом случае при записи снимается установленный ранее флаг, так что буфер будет записан только один раз.

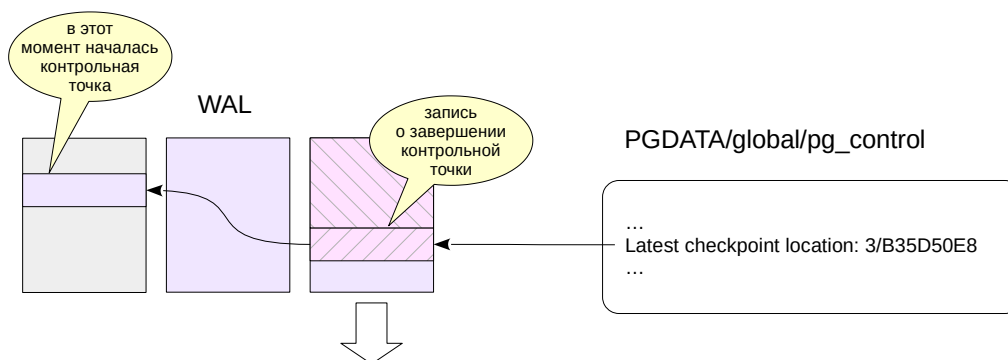
В процессе работы контрольной точки страницы продолжают изменяться в буферном кеше. Но новые грязные буферы уже не рассматриваются процессом контрольной точки, так как на момент начала работы они не были грязными.

Активность записи грязных буферов определяется значением параметра `checkpoint_completion_target`. Он показывает, какую часть времени между двумя соседними контрольными точками следует использовать для записи страниц. Значение по умолчанию равно 0.5, то есть запись занимает примерно половину времени между контрольными точками. Обычно значение увеличивают, например до 0.9, чтобы нагрузка была более равномерной. Значения выше 0.9 использовать не рекомендуется, поскольку фактически процесс может занять несколько больше времени, чем указано в параметре.

Процесс контрольной точки

в журнале создается запись о завершении контрольной точки
с указанием момента ее начала

в файл pg_control записывается LSN контрольной точки



8

В конце работы процесс создает журнальную запись об окончании контрольной точки. В этой записи содержится LSN момента начала работы контрольной точки. Поскольку контрольная точка ничего не записывает в журнал в начале своей работы, по этому LSN может находиться любая журнальная запись.

Кроме того, в специальный файл PGDATA/global/pg_control записывается указание на созданную журнальную запись. Таким образом можно быстро выяснить последнюю пройденную контрольную точку.

Контрольная точка

Заглянем в управляющий файл /var/lib/postgresql/13/main/global/pg_control. Это можно сделать с помощью утилиты pg_controldata.

```
postgres$ /usr/lib/postgresql/13/bin/pg_controldata -D /var/lib/postgresql/13/main
```

```
pg_control version number:      1300
Catalog version number:        202007201
Database system identifier:     7162043816462880038
Database cluster state:        in production
pg_control last modified:       Чт 19 янв 2023 14:31:00
Latest checkpoint location:     1/52A5E970
Latest checkpoint's REDO location: 1/52A5E938
Latest checkpoint's REDO WAL file: 000000010000000100000052
Latest checkpoint's TimelineID: 1
Latest checkpoint's PrevTimelineID: 1
Latest checkpoint's full_page_writes: on
Latest checkpoint's NextXID:    0:1204415
Latest checkpoint's NextOID:    103309
Latest checkpoint's NextMultiXactId: 11
Latest checkpoint's NextMultiOffset: 21
Latest checkpoint's oldestXID:  1103000
Latest checkpoint's oldestXID's DB: 13488
Latest checkpoint's oldestActiveXID: 1204414
Latest checkpoint's oldestMultiXid: 11
Latest checkpoint's oldestMulti's DB: 102843
Latest checkpoint's oldestCommitTsXid: 0
Latest checkpoint's newestCommitTsXid: 0
Time of latest checkpoint:      Чт 19 янв 2023 14:31:00
Fake LSN counter for unlogged rels: 0/3E8
Minimum recovery ending location: 0/0
Min recovery ending loc's timeline: 0
Backup start location:          0/0
Backup end location:            0/0
End-of-backup record required:  no
wal_level setting:              replica
wal_log_hints setting:          off
max_connections setting:        100
max_worker_processes setting:   8
max_wal_senders setting:        10
max_prepared_xacts setting:     0
max_locks_per_xact setting:     64
track_commit_timestamp setting: off
Maximum data alignment:         8
Database block size:            8192
Blocks per segment of large relation: 131072
WAL block size:                 8192
Bytes per WAL segment:          16777216
Maximum length of identifiers:  64
Maximum columns in an index:    32
Maximum size of a TOAST chunk:  1996
Size of a large-object chunk:   2048
Date/time type storage:         64-bit integers
Float8 argument passing:        by value
Data page checksum version:     1
Mock authentication nonce:       9e2a724b92922e7c0dc53158396bf20016f48a0f0775e2daade43872e55d2738
```

Видим много справочной информации, из которой особый интерес представляют данные о последней контрольной точке и статус кластера: «in production».

Выполним ручную контрольную точку и посмотрим, как это отражается в журнале и в управляющем файле.

```
=> SELECT pg_current_wal_insert_lsn();
```

```
pg_current_wal_insert_lsn
-----
1/52A5EA68
(1 row)
```

```
=> SELECT pg_walfile_name('1/52A5EA68');
```

```
pg_walfile_name
-----
000000010000000100000052
(1 row)
```

```
=> CHECKPOINT;
```

CHECKPOINT

```
=> SELECT pg_current_wal_insert_lsn();
```

```
pg_current_wal_insert_lsn
-----
1/52A5EB18
(1 row)
```

В журнал попадает запись о том, что контрольная точка пройдена (CHECKPOINT_ONLINE):

```
postgres$ /usr/lib/postgresql/13/bin/pg_waldump -p /var/lib/postgresql/13/main/pg_wal -s 1/52A5EA68 -e 1/52A5EB18 000000010000000100000052
```

```
rmgr: Standby      len (rec/tot):  50/   50, tx:      0, lsn: 1/52A5EA68, prev 1/52A5EA10, desc: RUNNING XACTS nextXid 1204415 latestCompletedXid 1204414 oldestRunningXid 120441
rmgr: XLOG          len (rec/tot): 114/  114, tx:      0, lsn: 1/52A5EAA0, prev 1/52A5EA68, desc: CHECKPOINT_ONLINE redo 1/52A5EA68; tli 1; prev tli 1; fpw true; xid 0:1204415; o
```

В описании записи указан LSN начала контрольной точки (redo).

Сравним с данными управляющего файла:

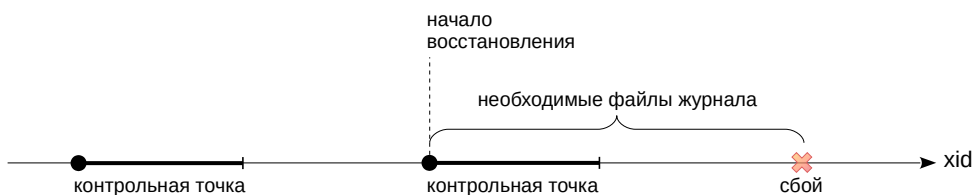
```
postgres$ /usr/lib/postgresql/13/bin/pg_controldata -D /var/lib/postgresql/13/main | egrep 'Latest.*location'
```

```
Latest checkpoint location:      1/52A5EAA0
Latest checkpoint's REDO location: 1/52A5EA68
```

Информация об LSN, очевидно, совпадает.

При старте сервера после сбоя

1. найти LSN_0 начала последней завершенной контрольной точки
2. применить каждую запись журнала, начиная с LSN_0 , если LSN записи больше, чем LSN страницы
3. перезаписать нежурналируемые таблицы init-файлами
4. выполнить контрольную точку



10

Если в работе сервера произошел сбой, при последующем запуске процесс startup обнаруживает это (в файле `pg_control` статус отличен от «shut down») и выполняет автоматическое восстановление.

Сначала процесс читает из того же файла LSN записи о последней контрольной точке. Из этой записи процесс узнает позицию LSN начала контрольной точки.

(Для полноты картины заметим: если присутствует файл `backup_label`, запись о контрольной точке читается из него — это нужно для восстановления из резервных копий. Подробнее см. курс DBA3.)

Далее процесс startup читает журнал от найденной позиции, последовательно применяя записи к страницам, если в этом есть необходимость (что можно проверить, сравнив LSN страницы на диске с LSN журнальной записи). Изменение страниц происходит в буферном кеше, как при обычной работе.

Записи, относящиеся к страницам CLOG, восстанавливают статус транзакций. Транзакции, не зафиксированные к концу восстановления, считаются оборванными; их изменения не видны в снимках данных.

Аналогично записи применяются и к файлам: например, если запись говорит, что файл должен быть создан, а его нет — файл создается.

В конце процесса все нежурналируемые таблицы перезаписываются с помощью образов в init-файлах. На этом процесс startup завершает работу, после чего процесс checkpointer выполняет контрольную точку, чтобы зафиксировать восстановленное состояние.

Восстановление

Теперь симитируем сбой, принудительно выключив сервер.

```
=> CREATE DATABASE wal_checkpoint;
```

```
CREATE DATABASE
```

```
=> \c wal_checkpoint
```

```
You are now connected to database "wal_checkpoint" as user "student".
```

```
=> CREATE TABLE test(t text);
```

```
CREATE TABLE
```

```
=> INSERT INTO test VALUES ('Перед сбоем');
```

```
INSERT 0 1
```

```
student$ sudo head -n 1 /var/lib/postgresql/13/main/postmaster.pid
```

```
132036
```

```
student$ sudo kill -9 132036
```

Сейчас на диске находятся журнальные записи, но табличные страницы не были сброшены на диск.

Проверим состояние кластера:

```
postgres$ /usr/lib/postgresql/13/bin/pg_controldata -D /var/lib/postgresql/13/main | grep state
```

```
Database cluster state:          in production
```

Состояние не изменилось. При запуске PostgreSQL поймет, что произошел сбой и требуется восстановление.

```
student$ sudo pg_ctlcluster 13 main start
```

```
postgres$ tail -n 6 /var/log/postgresql/postgresql-13-main.log
```

```
2023-01-19 14:31:02.037 MSK [132745] LOG:  database system was interrupted; last known up at 2023-01-19 14:31:01 MSK
2023-01-19 14:31:02.742 MSK [132745] LOG:  database system was not properly shut down; automatic recovery in progress
2023-01-19 14:31:02.746 MSK [132745] LOG:  redo starts at 1/52A60A90
2023-01-19 14:31:02.747 MSK [132745] LOG:  invalid record length at 1/52A7CB88: wanted 24, got 0
2023-01-19 14:31:02.747 MSK [132745] LOG:  redo done at 1/52A7CB60
2023-01-19 14:31:02.787 MSK [132744] LOG:  database system is ready to accept connections
```

```
student$ psql wal_checkpoint
```

```
=> SELECT * FROM test;
```

```
      t
-----
Перед сбоем
(1 row)
```

Как видим, таблица и данные восстановлены.

Теперь остановим экземпляр корректно. При такой остановке PostgreSQL выполняет контрольную точку, чтобы сбросить на диск все данные.

```
=> \q
```

```
student$ sudo pg_ctlcluster 13 main stop
```

Проверим состояние кластера:

```
postgres$ /usr/lib/postgresql/13/bin/pg_controldata -D /var/lib/postgresql/13/main | grep state
```

```
Database cluster state:          shut down
```

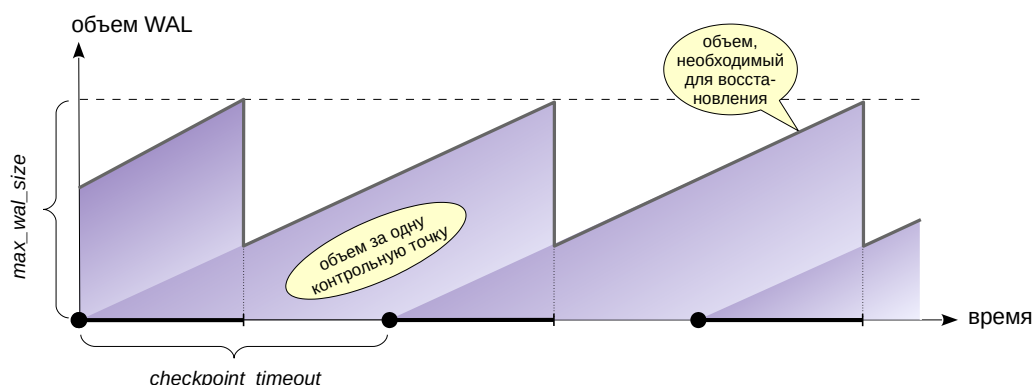
Теперь состояние — «shut down», что соответствует корректной остановке.

Частота контрольных точек

Настройки

`checkpoint_timeout` = 5min

`max_wal_size` = 1GB



12

Обычно контрольная точка настраивается из следующих соображений.

Сначала надо определиться, какая частота срабатываний нас устраивает (исходя из допустимого времени восстановления и объема журнальных файлов за это время при стандартной нагрузке). Чем реже можно позволить себе контрольные точки, тем лучше — это сокращает накладные расходы.

Устраивающая частота записывается в параметр `checkpoint_timeout` (значение по умолчанию — 5 минут — слишком мало, обычно время увеличивают до получаса).

Однако возможна ситуация, когда нагрузка станет выше расчетной и за указанное время будет сгенерирован слишком большой объем журнальных записей. Для этого в параметре `max_wal_size` указывают общий допустимый объем журнальных записей.

Для восстановления после сбоя сервер должен хранить файлы с момента последней контрольной точки, плюс файлы, накопившиеся во время работы текущей контрольной точки. Поэтому общий объем можно оценить как

$(1 + \text{checkpoint_completion_target}) * \text{объем-между-контр-точками}$.

Таким образом большая часть контрольных точек происходит по расписанию, раз в `checkpoint_timeout` единиц времени. Но при повышенной нагрузке контрольная точка вызывается чаще, чтобы постараться уложиться в объем `max_wal_size`.

Сервер хранит журнальные файлы

необходимые для восстановления (обычно $< max_wal_size$)
еще не прочитанные через слоты репликации
еще не записанные в архив, если настроена непрерывная архивация
не превышающие по объему wal_keep_size
не превышающие по объему min_wal_size (при переиспользовании)

Настройки

| | |
|-------------------|--------|
| max_wal_size | = 1GB |
| wal_keep_size | = 0 |
| $wal_recycle$ | = on |
| min_wal_size | = 80MB |

13

Надо понимать, что объем, указанный в параметре max_wal_size , может быть превышен. Это не жесткое ограничение, а ориентир для процесса checkpoint, влияющий на активность записи грязных буферов.

Кроме того, сервер не имеет права стереть журнальные файлы, еще не переданные через слоты репликации, и файлы, еще не записанные в архив при непрерывном архивировании. Это может привести к перерасходу места, так что если этот функционал используется, необходим постоянный мониторинг.

Можно на случай отставания реплики параметром wal_keep_size установить минимальный объем файлов журнала, остающихся после контрольной точки. Это не гарантирует, что журнальная запись сохранится до момента, когда она понадобится реплике, но позволяет работать без слота репликации.

По умолчанию журнальные файлы могут не удаляться, а просто переименовываться и использоваться заново. Параметр min_wal_size задает минимальный неудаляемый объем. Это позволяет сэкономить на постоянном создании и удалении файлов. Однако для файловых систем с сору-он-write быстрее создать новый файл, поэтому для них рекомендуется отключить переиспользование, установив $wal_recycle=off$.

<https://postgrespro.ru/docs/postgresql/13/wal-configuration>

<https://postgrespro.ru/docs/postgresql/13/runtime-config-wal#RUNTIME-CONFIG-WAL-CHECKPOINTS>

Объем журнала

Снова запустим экземпляр.

```
student$ sudo pg_ctlcluster 13 main start
```

Установим минимальное значение min_wal_size и отключим переиспользование, чтобы после контрольной точки оставался только один сегмент.

```
student$ psql wal_checkpoint
```

```
=> ALTER SYSTEM SET min_wal_size = '16MB';
```

```
ALTER SYSTEM
```

```
=> ALTER SYSTEM SET wal_recycle = off;
```

```
ALTER SYSTEM
```

```
=> SELECT pg_reload_conf();
```

```
pg_reload_conf
-----
t
(1 row)
```

Теперь запишем что-нибудь в таблицу.

```
student$ psql wal_checkpoint
```

```
=> INSERT INTO test
SELECT g.id::text FROM generate_series(1,500000) AS g(id);
```

```
INSERT 0 500000
```

Список файлов журнала:

```
=> SELECT * FROM pg_ls_waldir() ORDER BY modification;
```

| name | size | modification |
|---------------------------|----------|------------------------|
| 0000000100000000100000059 | 16777216 | 2023-01-19 13:43:14+03 |
| 0000000100000000100000055 | 16777216 | 2023-01-19 13:43:18+03 |
| 0000000100000000100000058 | 16777216 | 2023-01-19 13:43:19+03 |
| 0000000100000000100000057 | 16777216 | 2023-01-19 13:43:22+03 |
| 0000000100000000100000056 | 16777216 | 2023-01-19 13:43:26+03 |
| 000000010000000010000005A | 16777216 | 2023-01-19 14:30:30+03 |
| 0000000100000000100000054 | 16777216 | 2023-01-19 14:31:08+03 |
| 0000000100000000100000053 | 16777216 | 2023-01-19 14:31:08+03 |
| 0000000100000000100000052 | 16777216 | 2023-01-19 14:31:08+03 |

(9 rows)

Выполним ручную контрольную точку и опять посмотрим на журнал:

```
=> CHECKPOINT;
```

```
CHECKPOINT
```

```
=> SELECT * FROM pg_ls_waldir() ORDER BY modification;
```

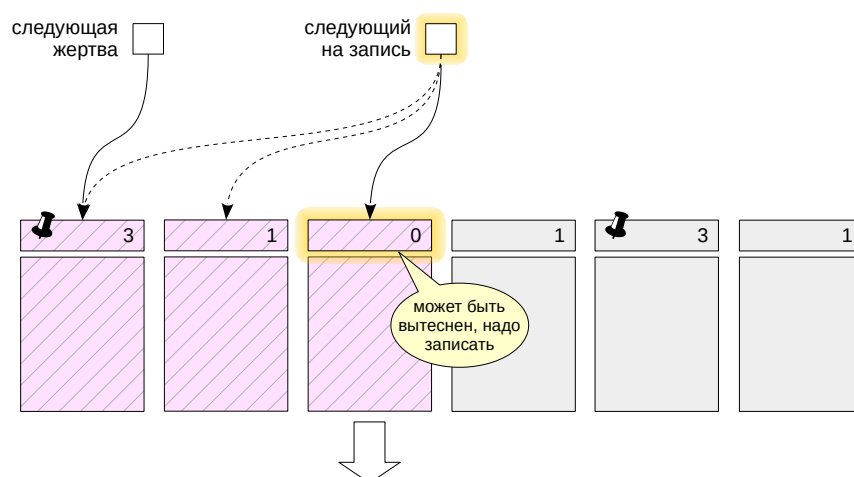
| name | size | modification |
|---------------------------|----------|------------------------|
| 0000000100000000100000059 | 16777216 | 2023-01-19 13:43:14+03 |
| 0000000100000000100000055 | 16777216 | 2023-01-19 13:43:18+03 |
| 0000000100000000100000058 | 16777216 | 2023-01-19 13:43:19+03 |
| 0000000100000000100000057 | 16777216 | 2023-01-19 13:43:22+03 |
| 0000000100000000100000056 | 16777216 | 2023-01-19 13:43:26+03 |
| 000000010000000010000005A | 16777216 | 2023-01-19 14:30:30+03 |
| 0000000100000000100000054 | 16777216 | 2023-01-19 14:31:08+03 |

(7 rows)

Все сегменты, кроме текущего, удалены.

Процесс фоновой записи

Настройка



Когда обслуживающий процесс собирается вытеснить страницу из буфера, он может обнаружить, что буфер грязный, и ему придется записать страницу на диск. Чтобы этого не происходило, в дополнение к процессу контрольной точки (checkpoint) существует также процесс фоновой записи (background writer, bgwriter или просто writer).

Процесс фоновой записи использует тот же самый алгоритм поиска буферов для вытеснения, что и обслуживающие процессы, только использует свой указатель. Он может опережать указатель на «жертву», но никогда не отстает от него.

Записываются буферы, которые одновременно:

- содержат измененные данные (грязные),
- не закреплены (pin count = 0),
- имеют нулевое число обращений (usage count = 0).

Таким образом фоновый процесс записи находит те буферы, которые с большой вероятностью вскоре потребуются вытеснить. За счет этого обслуживающий процесс скорее всего обнаружит, что выбранный им буфер не является грязным.

Алгоритм

уснуть на *bgwriter_delay*

если в среднем за цикл запрашивается N буферов, то записать
 $N * bgwriter_lru_multiplier \leq bgwriter_lru_maxpages$
грязных буферов

Настройки

| | |
|--------------------------------|---------|
| <i>bgwriter_delay</i> | = 200ms |
| <i>bgwriter_lru_maxpages</i> | = 100 |
| <i>bgwriter_lru_multiplier</i> | = 2.0 |

Процесс фоновой записи работает циклами максимум по *bgwriter_lru_maxpages* страниц, засыпая между циклами на *bgwriter_delay*.

(Таким образом, если установить параметр *bgwriter_lru_maxpages* в ноль, процесс фактически не будет работать.)

Точное число буферов для записи определяется по среднему количеству буферов, которые запрашивались обслуживающими процессами с прошлого цикла (используется скользящее среднее, чтобы сгладить неравномерность между циклами, но при этом не зависеть от давней истории). Вычисленное количество буферов умножается на коэффициент *bgwriter_lru_multiplier*.

Если процесс совсем не обнаружил грязных буферов (то есть в системе ничего не происходит), он «впадает в спячку», из которой его выводит обращение серверного процесса за буфером. После этого процесс просыпается и опять работает обычным образом.

Процесс фоновой записи имеет смысл настраивать после того, как настроена контрольная точка. Совместно с контрольной точкой эти процессы должны успевать записывать грязные буферы до того, как они потребуются обслуживающим процессам.

Контрольные точки

checkpoint_warning = 30s

log_checkpoints = off

Запись грязных буферов всеми процессами

представление *pg_stat_bgwriter*

Параметр *checkpoint_warning* позволяет выводить в журнал предупреждения о слишком частом выполнении контрольных точек. Если это происходит регулярно, следует подумать об увеличении *max_wal_size* или уменьшении интервала между контрольными точками *checkpoint_timeout*.

Включение параметра *log_checkpoints* выводит в журнал подробную информацию о каждой выполненной контрольной точке.

Статистику работы процессов, записывающих грязные буферы (контрольной точки, фоновой записи и обслуживающих процессов), показывает представление *pg_stat_bgwriter*.

Мониторинг

Параметр `checkpoint_warning` выводит предупреждение, если контрольные точки, вызванные переполнением размера журнальных файлов, выполняются слишком часто. Его значение по умолчанию:

```
=> SHOW checkpoint_warning;

checkpoint_warning
-----
30s
(1 row)
```

Его следует привести в соответствие со значением `checkpoint_timeout`.

Параметр `log_checkpoints` позволяет получать в журнале сообщений сервера информацию о выполняемых контрольных точках. По умолчанию параметр выключен:

```
=> SHOW log_checkpoints;

log_checkpoints
-----
off
(1 row)
```

Включим его.

```
=> ALTER SYSTEM SET log_checkpoints = on;
```

ALTER SYSTEM

```
=> SELECT pg_reload_conf();
```

```
pg_reload_conf
-----
t
(1 row)
```

Теперь запишем что-нибудь в таблицу и выполним контрольную точку.

```
=> INSERT INTO test
SELECT g.id::text FROM generate_series(1,100000) AS g(id);
```

```
INSERT 0 100000
```

```
=> CHECKPOINT;
```

CHECKPOINT

Вот какую информацию можно будет узнать из журнала сообщений:

```
postgres$ tail -n 2 /var/log/postgresql/postgresql-13-main.log
```

```
2023-01-19 14:31:09.154 MSK [132967] LOG:  checkpoint starting: immediate force wait
2023-01-19 14:31:09.184 MSK [132967] LOG:  checkpoint complete: wrote 443 buffers (2.7%); 0 WAL file(s) added, 0 removed, 0 recycled; write=0.013 s, sync=0.006 s, total=0.031 s; sync f
представление осталось).
```

Статистика работы процессов контрольной точки и фоновой записи отражается в одном общем представлении (раньше обе задачи решались одним процессом; затем их функции разделили, но представление осталось).

```
=> SELECT * FROM pg_stat_bgwriter \gx
```

```
-[ RECORD 1 ]-----+-----
checkpoints_timed    | 0
checkpoints_req      | 2
checkpoint_write_time | 52
checkpoint_sync_time | 14
buffers_checkpoint   | 2656
buffers_clean        | 0
maxwritten_clean     | 0
buffers_backend      | 2657
buffers_backend_fsync | 0
buffers_alloc        | 3038
stats_reset          | 2023-01-19 14:31:02.790821+03
```

* `checkpoints_timed` — контрольные точки по расписанию (`checkpoint_timeout`);

* `checkpoints_req` — контрольные точки по требованию (`max_wal_size`);

* `buffers_checkpoint` — страницы, сброшенные при контрольных точках;

* `buffers_backend` — страницы, сброшенные обслуживающими процессами;

* `buffers_clean` — страницы, сброшенные процессом фоновой записи.

В хорошо настроенной системе значение `buffers_backend` должно быть существенно меньше, чем сумма `buffers_checkpoint` и `buffers_clean`.

Большое значение `checkpoint_req` (по сравнению с `checkpoints_timed`) говорит о том, что контрольные точки происходят чаще, чем предполагалось.

Процесс контрольной точки ограничивает размер хранимых журнальных файлов и сокращает время восстановления

Контрольная точка и фоновая запись сбрасывают на диск грязные буферы

Обслуживающие процессы могут сбрасывать грязные буферы, но не должны этим заниматься

1. Настройте выполнение контрольной точки раз в 30 секунд. Установите параметры *min_wal_size* и *max_wal_size* в 16 МБ.
2. Несколько минут с помощью утилиты *pgbench* подавайте нагрузку 100 транзакций/сек.
3. Измерьте, какой объем журнальных файлов был сгенерирован за это время. Оцените, какой объем приходится в среднем на одну контрольную точку.
4. Проверьте данные статистики: все ли контрольные точки выполнялись по расписанию? Как можно объяснить полученный результат?
5. Сбросьте настройки к значениям по умолчанию.

1. Настройка контрольной точки

```
=> ALTER SYSTEM SET checkpoint_timeout = '30s';
```

```
ALTER SYSTEM
```

```
=> ALTER SYSTEM SET min_wal_size = '16MB';
```

```
ALTER SYSTEM
```

```
=> ALTER SYSTEM SET max_wal_size = '16MB';
```

```
ALTER SYSTEM
```

```
=> SELECT pg_reload_conf();
```

```
pg_reload_conf
-----
t
(1 row)
```

2. Нагрузка

Инициализируем таблицы.

```
=> CREATE DATABASE wal_checkpoint;
```

```
CREATE DATABASE
```

```
student$ pgbench -i wal_checkpoint
```

```
dropping old tables...
```

```
NOTICE: table "pgbench_accounts" does not exist, skipping
```

```
NOTICE: table "pgbench_branches" does not exist, skipping
```

```
NOTICE: table "pgbench_history" does not exist, skipping
```

```
NOTICE: table "pgbench_tellers" does not exist, skipping
```

```
creating tables...
```

```
generating data (client-side)...
```

```
100000 of 100000 tuples (100%) done (elapsed 0.11 s, remaining 0.00 s)
```

```
vacuuming...
```

```
creating primary keys...
```

```
done in 0.37 s (drop tables 0.00 s, create tables 0.01 s, client-side generate 0.14 s, vacuum 0.15 s, primary keys 0.07 s).
```

Сбросим статистику.

```
=> SELECT pg_stat_reset_shared('bgwriter');
```

```
pg_stat_reset_shared
-----
(1 row)
```

Запускаем pgbench, предварительно запомнив позицию в журнале.

```
=> SELECT pg_current_wal_insert_lsn();
```

```
pg_current_wal_insert_lsn
-----
1/7F25BA18
(1 row)
```

```
student$ pgbench -T 180 -R 100 wal_checkpoint
```

```
starting vacuum...end.
```

```
transaction type: <builtin: TPC-B (sort of)>
```

```
scaling factor: 1
```

```
query mode: simple
```

```
number of clients: 1
```

```
number of threads: 1
```

```
duration: 180 s
```

```
number of transactions actually processed: 17938
```

```
latency average = 3.543 ms
```

```
latency stddev = 1.136 ms
```

```
rate limit schedule lag: avg 1.182 (max 11.008) ms
```

```
tps = 99.660523 (including connections establishing)
```

```
tps = 99.662047 (excluding connections establishing)
```

3. Объем журнальных файлов

```
=> SELECT pg_current_wal_insert_lsn();
```

```
pg_current_wal_insert_lsn
-----
1/83657028
(1 row)
```

```
=> SELECT pg_size_pretty('1/83657028'::pg_lsn - '1/7F25BA18'::pg_lsn);

pg_size_pretty
-----
68 MB
(1 row)
```

4. Статистика

```
=> SELECT checkpoints_timed, checkpoints_req FROM pg_stat_bgwriter;

checkpoints_timed | checkpoints_req
-----+-----
4 | 4
(1 row)
```

Несмотря на то что в среднем объем журнальных записей за контрольную точку не превосходит установленного предела, часть контрольных точек выполнялась не по расписанию. Это говорит о неравномерности потока журнальных записей (вызванной выполнением автоочистки и другими причинами).

Поэтому в реальной системе замеры лучше выполнять на достаточно больших интервалах времени.

5. Настройки по умолчанию

```
=> ALTER SYSTEM RESET ALL;

ALTER SYSTEM

=> SELECT pg_reload_conf();

pg_reload_conf
-----
t
(1 row)
```