Лекция
26.03.2019

DBA2 Theory. p3.

Тонкости установки MariaDB

Ильшат Каразбаев
руководитель группы DBA
АО ТК Центр

Немного обо мне

Вместе со своей командой администрирую:

СУБД MySQL, Mariadb, galeracluster, Postgres

Главный по базам в ТК Центр

Повестка дня:

1. Вводная
2. Настройки операционной системы
3. Настройки сервера
4. Литература

Вводная

Первичная конфигурация MariaDB для автоматизированной установки в лекции автоматизации

Конфигурация ОС

1. Диски
2. Файловая система
3. Параметры монтирования
4. swappiness
5. Точки монтирования и их размер
6. Выбор оптимального количества RAM
7. Выбор оптимального количества CPU

Диски

SSD/HDD

RAID + Battery backed up write cache

HDD: Последовательная запись

BBWC: Сокращает количество операций записи: кэш сбрасывает на диск большие порции данных.

SSD: random reads/writes, проигрывает HDD в последовательной записи

SSD NVMe: Еще больше производительности за счет ухода от SAS/SATA

Диски.

baremetal vs VMs
Если есть возможность установки MariaDB на baremetal, используйте ее

Сетевые диски (SAN) vs локальные
У локальных дисков нет сетевой составляющей в задержке при операциях ввода-вывода

Диски

Наилучшая производительность достигается, если использовать RAID HDD + BBWC и SSD NVMe:

SSD: Random i/o oriented:

- - Table files (*.ibd)
- - UNDO segments (ibdata)

HDD: Sequential write oriented:

- - REDO log files (ib_logfile*)
- - Binary log files (binlog.XXXXXX)
- - Doublewrite buffer (ibdata)
- - Insert buffer (ibdata)
- - Slow query logs, error logs, general query logs, etc

Диски. Пример конфигурации

my.cnf:
[mysqld]
basedir=/root/mysql5400
datadir=/ssd/mysql-data
innodb_data_file_path=/hdd1/ibdata1:500M:autoextend
innodb_file_per_table
innodb_log_group_home_dir=/hdd/log
innodb_log_files_in_group=2
innodb_flush_log_at_trx_commit=1
innodb_flush_method=O_DIRECT
log_bin=/hdd/binlog

Диски. Опции при использовании только SSD

```
innodb_io_capacity: 1000
innodb_flush_neighbors: 0
```

Диски. Amazon

storage optimized instances

https://docs.aws.amazon.com/en_us/AWSEC2/latest/UserGuide/storage-optimized-instances.html

Конфигурация ОС. Файловая система

Требования к файловой системе:

- Поддержка больших файлов
- Возможность расширить на лету
- Производительность
- Отказоустойчивость

Также хорошей практикой является использование LVM совместно с ФС

Конфигурация ОС. Файловая система

Подходят под условия:

- XFS
- EXT4
- XFS

Конфигурация ОС. Файловая система

XFS

```
/dev/mapper/data-data /var/lib/mysql xfs defaults,nobarrier     0 0
```

EXT4

```
/dev/mapper/data-data       /var/lib/mysql      ext4    noatime,data=writeback,barrier=0,nobh,errors=remount-ro
/dev/mapper/log-log         /var/log/mysql      ext4    noatime,data=writeback,barrier=0,nobh,errors=remount-ro
```

ZFS

https://www.percona.com/blog/2017/12/07/hands-look-zfs-with-mysql/

Конфигурация ОС. Файловая система

RAW InnoDB system tablespace

```
[mysqld]
...
innodb_data_file_path=/dev/sdc:10Gnewraw
```

REBOOT

```
[mysqld]
...
innodb_data_file_path=/dev/sdc:10Graw
```

Конфигурация ОС. IO scheduler

!!! Не для SSD NVMe

```
# View the I/O scheduler setting. The value in square brackets shows the running scheduler
cat /sys/block/sdb/queue/scheduler
noop deadline [cfq]
# Change the setting
sudo echo noop > /sys/block/sdb/queue/scheduler
```

GRUB:

```
# Change the line:
GRUB_CMDLINE_LINUX_DEFAULT="quiet splash"
# to:
GRUB_CMDLINE_LINUX_DEFAULT="quiet splash elevator=noop"
```

Конфигурация ОС. Swappiness

https://mariadb.com/kb/en/library/configuring-swappiness/

/etc/systctl.conf
vm.swappiness = 1

sysctl -w vm.swappiness=1

Совсем отключать swap не рекомендуется, так как при нехватке ОЗУ может сработать OOM Killer, лучше мониторить использование swap.

Точки монтирования и их размер

В простейшем виде:

/var/lib/mysql под данные

/var/log/mysql под бинарные логи

/var

/tmp

/backup

Точки монтирования и их размер. /var/lib/mysql

Вычислить примерное количество дискового пространства, до которого вырастет БД за 3 года. Для этого посчитать, на сколько растет БД за неделю и экстраполировать на 3 года.

Вычислить размер БД

```
SELECT ROUND(SUM(data_length + index_length) / 1024 / 1024, 2) "DB Size in MB" FROM
information_schema.tables;
```

Добавить размер innodb_log_file_size * 2 и gcache.size (galera)

Подробнее
https://severalnines.com/blog/capacity-planning-mysql-and-mariadb-dimensioning-storage-size

Точки монтирования и их размер. /var/log/mysql

Уменьшить размер бинарных логов с 1ГБ до 100 МБ

Посмотреть, сколько бинарных логов генерируется за день

Умножить на expire_logs_days

Взять 140% от полученного результата

Поставить на мониторинг

Точки монтирования и их размер. /var /tmp

Использовать небольшой размер

При использовании OPTIMIZE можно поменять tmpdir на /var/log/mysql, так как размер временных файлов при пересоздании таблицы растет до размера таблицы, а /tmp выделили небольшого размера (костыль, но этим экономится размер дисков)

Точки монтирования и их размер. /backup

Точка монтирования под бекапы

Посчитать, сколько необходимо места под заданное количество бекапов

Мы уже экстраполировали необходимое пространство на три года под данные. Во столько же раз и увеличить 120-140% от необходимого места под бекапы. Бекапы растут вместе с ростом данных

Выбор оптимального количества RAM

innodb_buffer_pool_size - размер, близкий к размеру данных в БД

max_connections - если количество коннектов 100%, то Max_used_connections не более 70%. Полезно ограничить количество коннектов под каждого пользователя max_user_connections

calc_mem.sql:

```
SELECT @@GLOBAL.KEY_BUFFER_SIZE + @@GLOBAL.INNODB_BUFFER_POOL_SIZE +
@@GLOBAL.INNODB_LOG_BUFFER_SIZE + @@GLOBAL.INNODB_ADDITIONAL_MEM_POOL_SIZE + @@GLOBAL.NET_BUFFER_LENGTH +
(@@GLOBAL.SORT_BUFFER_SIZE + @@GLOBAL.MYISAM_SORT_BUFFER_SIZE +
@@GLOBAL.READ_BUFFER_SIZE + @@GLOBAL.JOIN_BUFFER_SIZE + @@GLOBAL.READ_RND_BUFFER_SIZE) *
@@GLOBAL.MAX_CONNECTIONS AS TOTAL_MEMORY_SIZE;
```

Выбор оптимального количества RAM

innodb_buffer_pool_instances - количество инстансов должно быть не более размера innodb_buffer_pool_size в GB

Выбор оптимального количества ядер CPU

Для серверов с интенсивной записью нужно больше CPU

Нет готового решения по подсчету количества ядер, при выборе количества необходимо исходить из исторических данных мониторинга

Обычно формула для расчета количества CPU такая (для RAM < 48 GB):

Количество RAM в ГБ/2

Выбор оптимального количества ядер CPU

InnoDB

```
innodb_read_io_threads - количество  CPU
innodb_write_io_threads - количество  CPU
```

Выбор оптимального количества ядер CPU

Для серверов - реплик количество ядер может сыграть хорошую роль при использовании параллельной репликации

Асинхронная:

slave_parallel_threads - на 1-2 меньше, чем количество ядер

galera:

wsrep_slave_threads - как минимум, в два раза больше, чем количество ядер

Выбор оптимального количества ядер CPU

Для снятия бекапов

mariabackup --paralllel=N

N - на 1-2 меньше количества ядер

Конфигурация сервиса MariaDB

Правила:

1. Не верьте всему, что написано в интернете (даже мне)
2. Не занимайтесь тюнингом ради тюнинга
3. Меняйте по одному параметру за один раз, тестируйте каждый раз.
4. Не забудьте сохранить изменения в конфигурационном файле, если применили настройку динамически
5. Остерегайтесь повторений в опциях конфигурационных файлов, mariadb не будет ругаться, если одну переменную задать дважды
6. Не увеличивайте бездумно опции в два раза, если добавили в два раза больше ОЗУ
7. Используйте правильные секции в конфигурационном файле

Конфигурация сервиса MariaDB

Лимиты

```
/etc/systemd/system/mariadb.service.d/limits.conf
[Service]
LimitNOFILE=infinity
LimitMEMLOCK=infinity
```

Убрать из автозагрузки

```
chkconfig mysql off
systemctl disable mariadb
```

Конфигурация сервиса MariaDB

default_storage_engine = innodb

innodb_buffer_pool_size (зависит от количества ОЗУ и размера БД)

innodb_log_file_size (redo log, эдакий WAL для innodb, обычно размер 1-2 часа записи в БД)

innodb_log_buffer_size (особенно при использовании автокоммита)

innodb_flush_log_at_trx_commit (2 для галер, 1 для уверенности в согласованности данных у асинхронных мастеров)

Конфигурация сервиса MariaDB

sync_binlog (еще одна опция, которая сильно влияет на производительность, 1 - fsync каждого события бинарного логиирования)

innodb_flosh_method = O_DIRECT (для исключения кэша ОС)

innodb_buffer_pool_instances (количество ОЗУ в ГБ, макс 64, для уменьшения взаимных блокировок глобального мьютекса)

innodb_thread_concurrency (также необходимо поменять innodb_thread_sleep_delay, innodb_concurrency_tickets)

skip_name_resolve (убрать задержку на резолвинг имен)

Конфигурация сервиса MariaDB

innodb_io_capacity (в зависимости от способностей диска, количество IO операций в секунду, также поменять innodb_io_capacity_max)

innodb_stats_on_metadata (если выключить, запросы к I_S будут быстрее)

innodb_buffer_pool_dump_at_shutdown & innodb_buffer_pool_load_at_startup - (для согласованности данных при рестарте сервиса)

innodb_adaptive_hash_index_parts (если включен AHI)

query_cache_type (выключить, QC неэффективен)

Конфигурация сервиса MariaDB

innodb_checksum_algorithm = crc32 и full_crc32 для MariaDB > 10.4.3

table_open_cache_instances = 16

innodb_read_io_threads & innodb_write_io_threads (зависит от количества дисков в RAID и CPU)

max_connections (нужно контролировать этот параметр, не увеличивать без необходимости, так как это множитель при буфферов, которые аллоцирует каждый коннект)

Конфигурация вспомогательных сервисов

ntpd/chrony - нужно мониторить состояние сервиса (должен быть запущен), следить за тем, что установлен верный timezone

logrotate - все логи (кроме бинарных), которые генерирует сервис, нужно ротировать или отправлять в journald (slow log, general log, messages, audit log, etc)

selinux - mysqld_t в permissive

firewalld - открыть порты 3306 и для галеры 4567, 4444

Вопросы?

Все анонсы здесь:

- telegram чат: t.me/mariadb_course

Материалы курса:

- видео: https://www.youtube.com/channel/UCGsmu6YDpcR_kWcXzeQkWrA
- слайды лекций и примеры: git@github.com:barazbay/mariadb_course.git

Меня можно найти:

- vk, instagram: barazbay
- twitter: karazbay

Литература

1. https://severalnines.com/blog/capacity-planning-mysql-and-mariadb-dimensioning-storage-size
2. http://www.speedemy.com/17-key-mysql-config-file-settings-mysql-5-7-proof/
3. https://www.percona.com/blog/2018/07/03/linux-os-tuning-for-mysql-database-performance/
4. https://www.percona.com/blog/2017/12/07/hands-look-zfs-with-mysql/
5. https://www.percona.com/blog/2018/05/15/about-zfs-performance/

Литература

6. https://www.percona.com/blog/2018/05/15/about-zfs-performance/
7. https://en.wikipedia.org/wiki/ZFS
8. http://yoshinorimatsunobu.blogspot.com/2009/05/tables-on-ssd-redobinlogsystem.html
9. https://www.samsung.com/semiconductor/global.semi.static/best-practices-for-mysql-with-ssds-0.pdf
10. https://www.percona.com/blog/2018/07/03/linux-os-tuning-for-mysql-database-performance/
11. https://docs.aws.amazon.com/en_us/AWSEC2/latest/UserGuide/nvme-ebs-volumes.html

Литература

12. https://docs.aws.amazon.com/en_us/AWSEC2/latest/UserGuide/storage-optimized-instances.html
13. https://mariadb.com/kb/en/library/innodb-system-tablespaces/#using-raw-disk-partitions
14. https://mariadb.com/kb/en/library/configuring-swappiness/
15. https://www.percona.com/blog/2014/01/28/10-mysql-performance-tuning-settings-after-installation/
16. https://dba.stackexchange.com/questions/111988/how-do-you-calculate-how-much-hardware-resources-you-need-for-a-database