

Hardware RAID vs Linux Software RAID в схватке за симпатии PostgreSQL

Артём Носов <chip@unixstyle.ru>

Rev: 1.1 от 17 Декабря 2009

Возникла необходимость исследовать производительность нескольких решений:

- linux md: RAID 5 Stripe Size 128 KB;
- linux md: RAID10 Stripe Size 128 KB;
- linux md: RAID 10 Stripe Size 256 KB;
- 3ware: RAID10 Stripe Size 256 KB.

Прикладная задача: использование системы базы данных PostgreSQL с размещением PGDATA на одном из описанных выше массивов.

Средство тестирования: pgbench

Параметры аппаратной платформы:

```
~$ cat /proc/cpuinfo | grep -E '(processor|model name) '
processor       : 0
model name     : Intel(R) Xeon(R) CPU           E5450  @ 3.00GHz
processor       : 1
model name     : Intel(R) Xeon(R) CPU           E5450  @ 3.00GHz
processor       : 2
model name     : Intel(R) Xeon(R) CPU           E5450  @ 3.00GHz
processor       : 3
model name     : Intel(R) Xeon(R) CPU           E5450  @ 3.00GHz
processor       : 4
model name     : Intel(R) Xeon(R) CPU           E5450  @ 3.00GHz
processor       : 5
model name     : Intel(R) Xeon(R) CPU           E5450  @ 3.00GHz
processor       : 6
model name     : Intel(R) Xeon(R) CPU           E5450  @ 3.00GHz
processor       : 7
model name     : Intel(R) Xeon(R) CPU           E5450  @ 3.00GHz
~$ head -1 /proc/meminfo
MemTotal:      32962584 kB
~$ tw_cli /c0 show all
/c0 Driver Version = 2.26.08.003-2.6.18RH
/c0 Model = 9690SA-8I
/c0 Available Memory = 448MB
/c0 Firmware Version = FH9X 4.08.00.006
/c0 Bios Version = BE9X 4.08.00.001
/c0 Boot Loader Version = BL9X 3.08.00.001
...
/c0 PCB Version = Rev 041
/c0 PCHIP Version = 2.00
/c0 ACHIP Version = 1.31A4
...
/c0 Drives = 8 of 128
...
/c0 Controller Bus Type = PCIe
...
~$
```

Во всех рассматриваемых случаях использовались восемь дисков SAS номенклатуры SEAGATE ST973402SS.

Для случая использования программного RAID диски экспортировались контроллером Zware как одиночные диски (Single Drive). Соответственно сравнение на первый взгляд не совсем честное, т. к. программный RAID использует возможности аппаратного контроллера, например, дисковый кэш. Однако если окунуться в статью [Benchmarking hardware RAID vs. Linux kernel software RAID](#) редакторы linux.com находятся в аналогичной ситуации тестирования и в параграфе "Single disk performance" приходят к выводу: «Because the Adaptec card does not offer a clear advantage when accessing a single disk, using the Adaptec controller to expose the six disks for software RAID should not provide an unfair advantage relative to other software RAID setups». В своем тестировании мы также будем полагаться на это допущение.

Каждое тестирование сопровождалось созданием нового хранилища RAID, поэтому влияние какого-либо кэширования исключено.

PostgreSQL использовался последней на текущий момент версии 8.4.1 в следующей конфигурации

```
~$ cat /var/lib/pgdata/data/postgresql.conf
auto_explain.log_min_duration = '3s'
autovacuum_analyze_scale_factor = 0.03
autovacuum_analyze_threshold = 250
autovacuum_naptime = 10min
autovacuum_vacuum_scale_factor = 0.05
autovacuum_vacuum_threshold = 500
bgwriter_lru_maxpages = 700
checkpoint_completion_target = 0.8
checkpoint_segments = 50
checkpoint_timeout = 10min
constraint_exclusion = on
custom_variable_classes = 'auto_explain,pg_stat_statements'
datestyle = 'iso, dmy'
default_text_search_config = 'pg_catalog.english'
effective_cache_size = 10GB
escape_string_warning = off
lc_messages = 'en_US.UTF-8'
lc_monetary = 'en_US.UTF-8'
lc_numeric = 'en_US.UTF-8'
lc_time = 'en_US.UTF-8'
listen_addresses = '*'
log_autovacuum_min_duration = 0
log_checkpoints = on
log_destination = 'syslog'
log_min_duration_statement = 2000ms
log_min_error_statement = warning
log_statement = 'none'
maintenance_work_mem = 256MB
max_connections = 300
pg_stat_statements.max = 10000
pg_stat_statements.track = all
shared_buffers = 2GB
shared_preload_libraries = 'auto_explain,pg_stat_statements'
silent_mode = on
statement_timeout = 35000
synchronous_commit = off
timezone = 'Europe/Moscow'
track_activity_query_size = 16384
wal_buffers = 8MB
work_mem = 64MB
```

Заполнение данными производилось командой
~\$ pgbench -i -s 5000 -U postgres test

Тестирование

~\$ pgbench -c 10 -T 600 -M prepared -U postgres test

В соседнем окне работал dstat. Из всех результатов вывели на диаграмму tps (transactions per second) с установкой соединения (including connections establishing).

Используемые сокращения:

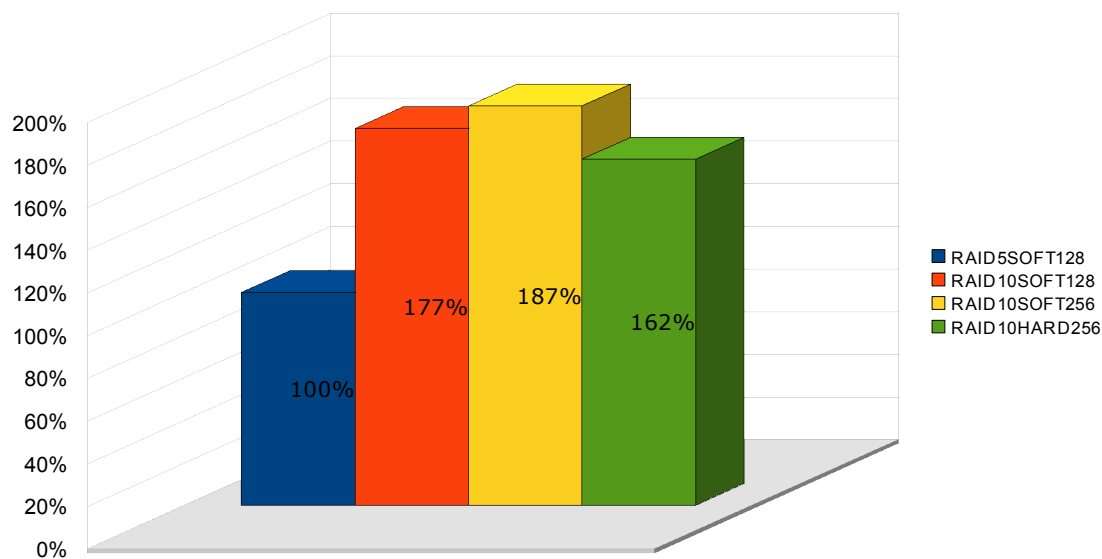
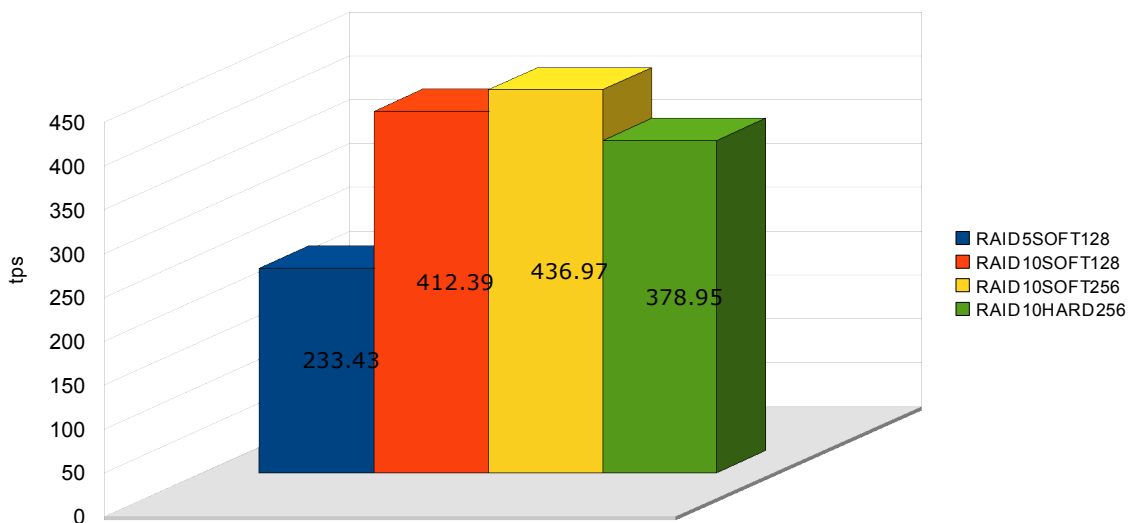
HARD - PGDATA размещается на аппаратном массиве;

SOFT - PGDATA размещается на программном массиве;

128/256 - размер Stripe Size в Кбайтах.

Например, RAID10SOFT128 означает, что тестирование проходило на программном массиве RAID10 с размером Stripe равным 128 Кбайт.

Итоговые диаграммы



Рассмотрим полученные нами результаты в сравнение с заключениями, приведенными на страницах журнала [Storage Advisors](#) от Adaptec.

RAID5 проиграл ровно вдвое своему конкуренту RAID10. Объясняется это тем, что каждая операция записи распадается на операцию записи и операцию чтения. Из-за этого производительность записи падает ровно вдвое. Более детальное пояснение [Yet another RAID-10 vs RAID-5 question](#) .

Выбор размера Stripe описывается в заметке [RAID Stripe width](#) . Нас интересует полученный вывод: «There really aren't many practical reasons to go with a stripe width much larger, unless you have a hugely performance dependent application like -for real-time data capture - holding temporary/transitory data meaning, you don't plan on keeping the data there very long, at least not without backing it up or having a copy somewhere else». По результатам наших замеров разница при выборе размера страйпа размером 128 Кбайт и 256Кбайт составила в районе 5%.

И самый главный вывод, в котором мы полностью согласны с сотрудниками Adaptec и статьей [RAID 5 and database](#) «Simply for most database applications you should consider using RAID 10. RAID 10 does not do parity, but simply writes the same data in two separate disks within the array. Consequently for a variety of technical reasons RAID 10 has faster random writes than RAID 5».

Материалы:

1. [Обсуждение статьи \(комментарии\)](#);
2. [Hardware 3ware RAID10 vs Linux Software RAID10](#) (рус.);
3. [Benchmarking hardware RAID vs. Linux kernel software RAID](#) (eng.);
4. [Storage Advisors](#) от Adaptec (eng.);

Необработанные данные:

- [RAID5SOFT128-dstat.log](#)
- [RAID5SOFT128-pgbench.log](#)
- [RAID10HARD256-dstat.log](#)
- [RAID10HARD256-pgbench.log](#)
- [RAID10SOFT128-dstat.log](#)
- [RAID10SOFT128-pgbench.log](#)
- [RAID10SOFT256-dstat.log](#)
- [RAID10SOFT256-pgbench.log](#)

При использовании материалов ссылка на сайт «[Чип и Дейл спешат на помощь. Обитель UNIX](#)» **обязательна**.