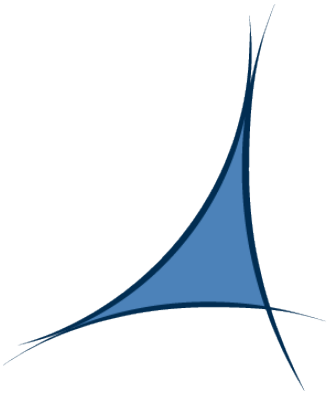
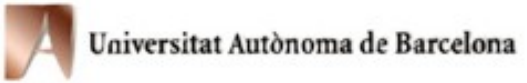


PIC
port d'informació
científica

PostgreSQL tips for dCache

Gerard Bernabeu
Francisco Martinez



PIC
port d'informació
científica

PostgreSQL tips for dCache

- Common customization (autovacuum, backups, basic performance)
- PNFS server (more performance enhacements)

PIC
port d'informació
científica

Common customization - autovacuum

- Enabling Autovacuum
  (/var/lib/pgsql/data/postgresql.conf)
  - autovacuum=on (by default on 8.3.5)
  - *ps faux* | *grep autovacuum* should find something like this (on 8.3.5)

```
postgres  3612  0.0  0.0 228980 1604 ?       Ss    2008   0:00  \_ postgres: autovacuum launcher process
```

PIC
port d'informació
científica

Common customization - backups

- DB backups

/usr/bin/pg_dumpall -U postgres | /bin/gzip > backupFile.sql.gz

cat /usr/local/sbin/pgsql_backup.sh

```
#!/bin/bash
#Backups the gziped output of $DUMPALL to $LPATH

LPATH=/DBbackups
LFILE=`date +%y%m%d_%H%M`_backup_postgres_${HOSTNAME}.sql.gz
LOG=/DBbackups/pgsql_backup.log
DUMPALL="/usr/bin/pg_dumpall -U postgres"
GMETRIC="/usr/bin/gmetric --type int16 --units Seconds --dmax 100000"
DATE=/bin/date
GZIP=/bin/gzip

# Backup en cuestion
mkdir -p $LPATH
DATE1=`date +%s`
$DATE --iso-8601=seconds  >> $LOG 2>&1
$DUMPALL | $GZIP > $LPATH/$LFILE

DATE2=`date +%s`; DELTAT=`expr $DATE2 - $DATE1`
echo "pg_dumpall and gzip finished in $DELTAT secs"  >> $LOG
$GMETRIC --name Backup_Dump_T --value $DELTAT

# Remove older than 24h backups
find $LPATH/ -mtime +1 -exec rm -f {} \;
```

PIC
port d'informació
científica

Common customization - performance

- How big is your DB? (this is PIC's SRM)

```
dcache=> select sum(relpages) from pg_class;
  sum
--------
 761234 (pages)
```

Each page is tipcally 8Kb so 761234*8/1024=5947MB. To crosscheck:

```
du -sh /var/lib/pgsql/data/base/
5.9G/var/lib/pgsql/data/base/
```

- Increase shared buffers/work_mem to decrease HDD ussage.
- We should have in mind that OS and other apps (JVM) need memory too.

PIC
port d'informació
científica

Common customization - performance

- How to increase shared buffers/work_mem?
  (a must on SRM, PNFS and, if you use it for monitoring, billingDB)
  - PostgreSQL (8.3.5 /var/lib/pgsql/data/postgresql.conf)
    - **shared_buffers**
    - **work_mem**
    - **maintenance_work_mem**
    - **checkpoint_segments**
    - **effective_cache_size**
    - **max_fsm_pages**
  - System level (linux)
    - **/proc/sys/kernel/shmmax** >= PostgreSQL shared_buffers
    - **/prox/sys/kernel/shmall** >=shmmax/PAGE_SIZE.
      getconf PAGE_SIZE (usually 4096)

PIC
port d'informació
científica

Common customization - performance

- Which are the “magic numbers”?

  According to

  www.postgresql.org/docs/8.3
  www.powerpostgresql.com/PerfList/

PIC
port d'informació
científica

Common customization - performance

- Which are the magic numbers?
  - shared_buffers

    This is NOT the total memory PostgreSQL has to work with.
    shared_buffers is the block of dedicated **memory used for active opreations** and should be **at least 16k*max_connections**. However settings significantly higher than the minimum are usually needed for good performance. **512MB** seems to be a good number for PNFS/SRM, other servers shouldn't have problems with the default 32MB, but if posible better 128MB or more.

  - work_mem

    Specifies the amount of memory to be used by **internal sort operations and hash tables before switching to temporary disk files**. The right number is a ceiling on the amount of RAM any single operation can grab before being forced to disk. **The more concurrent connections you have doing complex queries, the smaller it needs to be**. Since we're using 100 connections and we don't want it to use more than 2GB we set it to **20MB** (default is 1MB)

PIC
port d'informació
científica

Common customization - performance

- Which are the magic numbers?
  - maintenance_work_mem

    Amount of memory used in **maintenance operations such as VACUUM**, CREATE INDEX, etc. 50-75% of the on-disk size of largest table/index is a good rule according to powerpostgresql.com. In PIC's PNFS CMS has almost 12GB so we can not follow the rule. We'll set it to **512MB** (default is 16MB).

  - checkpoint_segments

    If you have messages like the one below you should increase

    ```
    # cat /var/lib/pgsql/data/pg_log/postgresql-2009-01-01_000000.log
      LOG:  checkpoints are occurring too frequently (22 seconds apart)
      HINT:  Consider increasing the configuration parameter "checkpoint_segments".
    ```

    If it's a bussy, non read-only DB you probably need to increase this value. On the SRM we've it at the default 3 and an increase is needed (pg_xlog dir is using 129MB). On the PNFS we've it at 20 and it's OK now (pg_xlog using 673MB). pg_xlog size=(chekcpoint_segments*2+1)*16. It's a good idea to have pg_xlog and in different HDDs.

PIC
port d'informació
científica

Common customization - performance

- Which are the magic numbers?
  - effective_cache_size

    Tells the planner the largest possible database object that could be expected to be cached, should be about 70% RAM if on a dedicated server. For instance our 8GB PNFS could use 2GB for JVM + 2GB for other pgsql stuff&OS so we could set 3-4GB.

  - max_fsm_pages

    Sizes the register which tracks partially empty data pages for population with new data; if set right, **makes VACUUM faster and removes the need for VACUUM FULL or REINDEX**. **Require very little memory so it's better to be generous here**. To determine the number you can run “VACUUM VERBOSE ANALYZE” and grep for *page slots are required to track all free space* or use *on-disk db size/16KB* (default is choosen by initdb depending on the amount of available memory and can range from 20k to 200k pages)

PIC
port d'informació
científica

PNFS server

- Performance enhancing
  - Place different DBs in different HDDs.
    http://www.postgresql.org/docs/8.3/static/manage-ag-tablespaces.html
    CREATE TABLESPACE new_tablespace LOCATION '/data/fastHDD';
    ALTER TABLE name SET TABLESPACE new_tablespace;
    - That will not migrate indexes; move them by hand