Setting up a Large PostgreSQL Server: A Case Study

Vivek Khera, Ph.D.

O'Reilly Open Source Convention

July 2004

The Project

- Replace existing dedicated DB server

- Use old DB server as live spare with replication software

Why?

- Improve interactive response
- Increase service availability
- Upgrade from PG 7.2 to 7.4

What Needs Fixing?

- You can’t fix something if you don’t know what isn’t working right
- Measure everything

Measurement Tools

- systat -vmstat
  - disk: MB/s, tps, KB/t, % busy
  - memory & CPU usage
- “Feel” of application
- Log files
- DB Statistics

The Existing Server

- Dual Pentium III
- 4 Disk SCSI hardware RAID10 for data
- 2 GB RAM
- Postgres 7.2
- FreeBSD 4.x

Database

- Customers
- Customers’ List Members
- Customers’ Messages
- Tracking Information

```
    1 users    Load  0.42  0.36  0.34                  May 25 10:04

Mem:KB    REAL            VIRTUAL                     VN PAGER  SWAP PAGER
        Tot   Share      Tot    Share    Free         in  out     in  out
Act  369588    2648   426060     3604   78940 count
All 2055180    3048  3320020     5908         pages
                                                                 Interrupts
Proc:r  p  d  s  w    Csw  Trp  Sys  Int  Sof  Flt     74 cow     508 total
          55  8        43  593  642  507   18  596 256684 wire        ata0 irq14
                                                   209888 act     260 aac0 irq2
 2.6%Sys   0.2%Intr  6.6%User  0.0%Nice 90.6%Idl  1264784 inact    10 fxp0 irq7
|    |    |    |    |    |    |    |    |    |      74288 cache       fdc0 irq6
=>>>>                                                4652 free        atkbd0 irq
                                                          daefr    10 sio0 irq4
Namei         Name-cache    Dir-cache                 388 prcfr       sio1 irq3
    Calls     hits   %     hits   %                       react   100 clk irq0
      457      430   94                                   pdwak   128 rtc irq8
                                      282 zfod            pdpgs
Disks aacd0 aacd1  acd0   fd0         213 ofod            intrn
KB/t  26.41 10.20  0.00  0.00          75 %slo-z   204096 buf
tps     257     2     0     0         482 tfree         8 dirtybuf
MB/s   6.64  0.02  0.00  0.00                      130807 desiredvnodes
% busy  100     4     0     0                       32701 numvnodes
                                                    31197 freevnodes
    1 users    Load  0.42  0.36  0.34                  May 25 10:04

Mem:KB    REAL            VIRTUAL                     VN PAGER  SWAP PAGER
        Tot   Share      Tot    Share    Free         in  out     in  out
Act  369588    2648   426060     3604   78940 count
All 2055180    3048  3320020     5908         pages
                                                                 Interrupts
Proc:r  p  d  s  w    Csw  Trp  Sys  Int  Sof  Flt     74 cow     508 total
          55  8        43  593  642  507   18  596 256684 wire        ata0 irq14
                                                   209888 act     260 aac0 irq2
 2.6%Sys   0.2%Intr  6.6%User  0.0%Nice 90.6%Idl  1264784 inact    10 fxp0 irq7
|    |    |    |    |    |    |    |    |    |      74288 cache       fdc0 irq6
=>>>>                                                4652 free        atkbd0 irq
                                                          daefr    10 sio0 irq4
Namei         Name-cache    Dir-cache                 388 prcfr       sio1 irq3
    Calls     hits   %     hits   %                       react   100 clk irq0
      457      430   94                                   pdwak   128 rtc irq8
                                      282 zfod            pdpgs
Disks aacd0 aacd1  acd0   fd0         213 ofod            intrn
KB/t  26.41 10.20  0.00  0.00          75 %slo-z   204096 buf
tps     257     2     0     0         482 tfree         8 dirtybuf
MB/s   6.64  0.02  0.00  0.00                      130807 desiredvnodes
% busy  100     4     0     0                       32701 numvnodes
                                                    31197 freevnodes
```

DB Statistics

- Enable in postgresql.conf
- Query from psql:
  - `SELECT * FROM pg_stat_activity;`
  - `SELECT relname,relkind,relpages FROM pg_class WHERE relname NOT LIKE 'pg_%';`

Results

- Disk capacity saturated
- RAM insufficient for both queries and disk cache
- Index bloat partly to blame

New Server Goals

- Improve disk subsystem speed
    - increase number of spindles
    - separate pg_xlog from main data disks
- Increase RAM
    - 4GB is max on 32-bit CPU without trickery

New Configuration

- Dual Xeon
- 14 disk external SCSI array for data
- 2 disk internal array for system + log
- 4 GB RAM
- Postgres 7.4
- FreeBSD 4.x

Disk Array Options

- Internal 2-disk array only one choice: RAID1 mirror
- External 14-disk:
  - RAID5
  - RAID10
  - RAID50

Evaluating Arrays

- DB restore of live database snapshot
- Sample queries
- bonnie++ and iozone

And the Winner Is...

RAID5

PostgreSQL Tuning

Shared Buffers

- 30,000 buffers
  - Wisdom from mailing list
  - Personal experience with old server
  - Wire SHM pages to RAM: `kern.ipc.shm_use_phys=1`

Other Settings

- Sort Memory
- Vacuum Memory
- Free Shared Map
- Checkpoint Segments
- Commit Delay

Logging

- Log via syslog
- Log long running queries

Query Profiling

- Examine logs to identify slow queries
- EXPLAIN
- Analyze program holistically

Problems Persist

- Indexes still growing
- Log disk highly utilized

The Causes

- Long running transactions were idle
- Open transactions prevented data and index rows from being reused
- Tracking data recorded as it came in

Lather, Rinse, Repeat

- Never stop monitoring
- Keep looking for optimizations
- Take good notes

The End

PostgreSQL