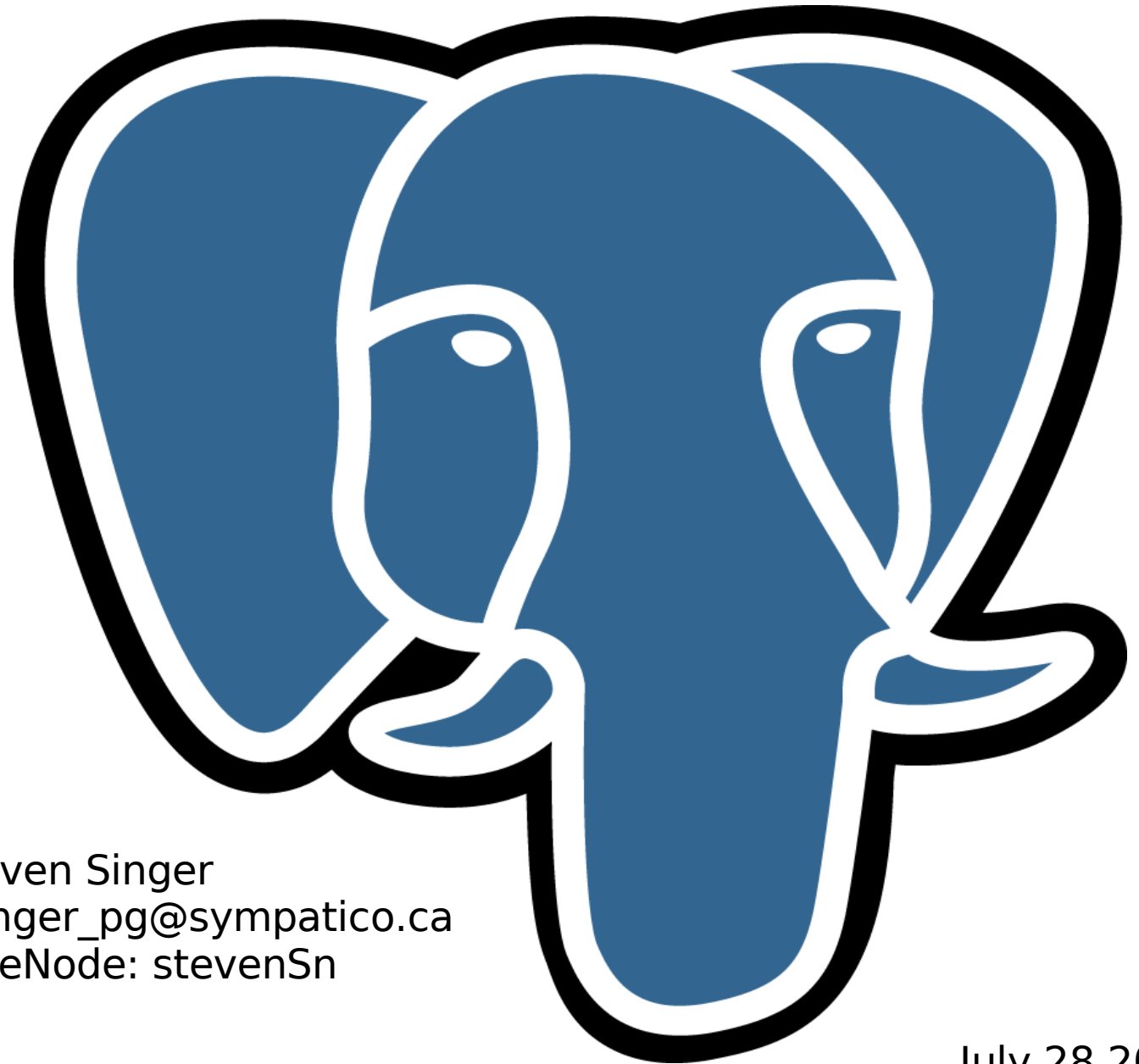


PostgreSQL Replication

F
O
R
O
P
O
S
I
T
I
O
N
S



Steven Singer
ssinger_pg@sympatico.ca
FreeNode: stevenSn

July 28 2008

Who am I

```
SELECT * FROM pg_user where name='Steve'
```

- PostgreSQL user since 6.5
- Author of contrib/dbmirror included with 7.x
- Software developer for a local consulting company
- Occasional contributor to side projects

The Golden Rule

- **You can't have everything**
- What you want \neq what you need
- Determine *need* from requirements
- Be prepared to accept trade-offs
- If you've already been sold on Oracle RAC, go and buy RAC

The Choices

pgcluster
pgcluster-II
slony
dbmirror
longdist
pgpool
pgpool-II
Sequoia
PITR
postgres-r
pyReplica
pg_dump/pg_restore
EDB Replicator
pl/proxy
Mammoth
RepDB
Bucardo
Cybercluster
erserve
rserv
NTT WAL shipping

Why Replication?

- Because it is fun?
- So you can buy twice the number of licenses?
- To distribute your data?
- So you can be buzzword compliant?

Fail-Over

- Goal: To have additional database servers standing by in case of:
 - Hardware Failure
 - DBA/Application mistakes
 - Maintenance
- Turning on a standby means failing the master

Automatic or manual

- Who makes the decision to promote a slave to a master?
- Computer or human?
- Look at your possible failures, how can you detect them?
- When will the wrong decision be made?

Load Balancing

- Goal: Improve performance by adding more servers
- Master slave or multi-master?
- You can only do queries on slaves
- Multi-Master is really hard

Load Balancing : slaves

- Send queries to slaves
- Run query parts on machines and combine results
- Partition some data across servers (sharding)
- Topic on its own

Question: How stale will you allow your slaves to be?

Trigger based Replication

- Slony, Longdiste, Bucardo
- Uses *ON INSERT, ON UPDATE, ON DELETE* triggers
- Slaves jump between consistent snapshots of master
- Asynchronous
- Performance impact on master
- No triggers will fires on DDL

Slony

- www.slony.info
- Developed by Afillias
- Async trigger based, master/slave
- Most widely deployed PostgreSQL replication solution
- Very flexible
- Allows for complicated setups
- Some rough edges

Longdiste

- <https://developer.skype.com/SkypeGarage/DbProjects/SkyTools>
- Async trigger based, master slave
- Part of SkyTools from Skype
- Based on experience with Slony but simpler; less features
- Can't cascade slaves
- No switchover support (Today)

Bucardo

- <http://bucardo.org>
- Async trigger based
- Developed by backcountry.com
- Allows multi-master writes with user specified conflict resolution

Middleware based statement replication

- pgpool-II, pgcluster, Continuent, GridSQL
- Sits in-between your application and PostgreSQL
- Redirects SQL to one or more of your databases
- How do you know all databases contain the same data?

Middleware - Issues

- INSERT INTO x VALUES (rand())
- Functions/Stored Procedures, or timestamps are a bad idea
- How do you ensure things happen in the same order on all nodes?
- What if a COMMIT on the second node fails?
- Limitations fine for some applications

Sequoia

- <http://sequoia.continuent.org/Ho>
- Middleware
- Supports multiple RDBMS
- Has its own journaling
- **Ensure all SQL writers understand how it works**

PITR

- Point-In Time Recovery
- WAL segments are sent to the slaves
- Aysnc (today)
- Can't query slaves (today) unless you bring them up
- Rolling slaves with ZFS?
- Only good for failover (today)

Multi-Master

- If the same row is changed on two servers *around* the same time?
- Solving this in the general sense is really hard.
- **Don't go here unless you need to**

Works in progress

- Not Production Ready
- Serious technical problems to be worked out
- If your up for a challenge
- Good place to spend your RAC budget

postgres-r

- Multi Master based on a group communication system
- All nodes process all statements in the same order (total order)
- Depends on a GCS like Spread
- Recently open-sourced

pgcluster-II

- Multi Master
- Shared disc
- Presented at pgconn 07 not released
- Status unknown

Rules of Thumb

- PITR if it will meet your needs
- Slony or Longdiste if you need to query your slaves
- Find a way around multi-master
- Avoid statement based solutions if consistency is important

Questions?

Steven Singer
ssinger_pg@sympatico.ca
FreeNode: stevenSn