新手机 新应用 新娱乐

PostgreSQL Introduction

Digoal.Zhou

7/20/2011

Catalog

- PostgreSQL Origin
- Layout
- Features
- Enterprise Class Attribute
- Case

Origin

Extract From Wiki

INGRES 1973

DARPA

POSTGRES 1985

Postgres95 1995

Michael Stonebraker

Informix

VERTICA

C-Store

SciDB

VoltDB

H-Store

MARIPOSA

Federated database system

PeopleSoft

ORACLE

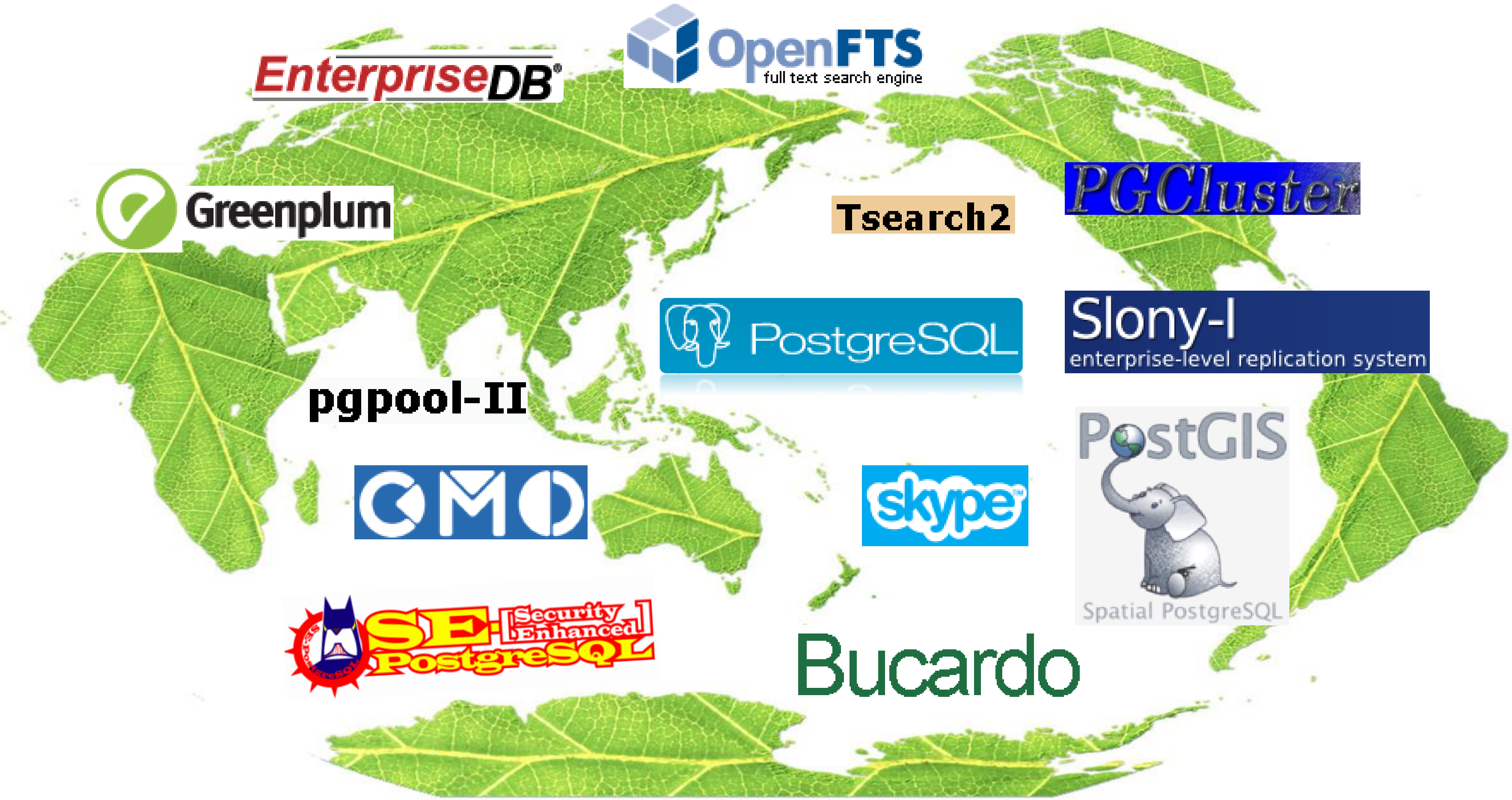
PostgreSQL 1996

EnterpriseDB OLTP

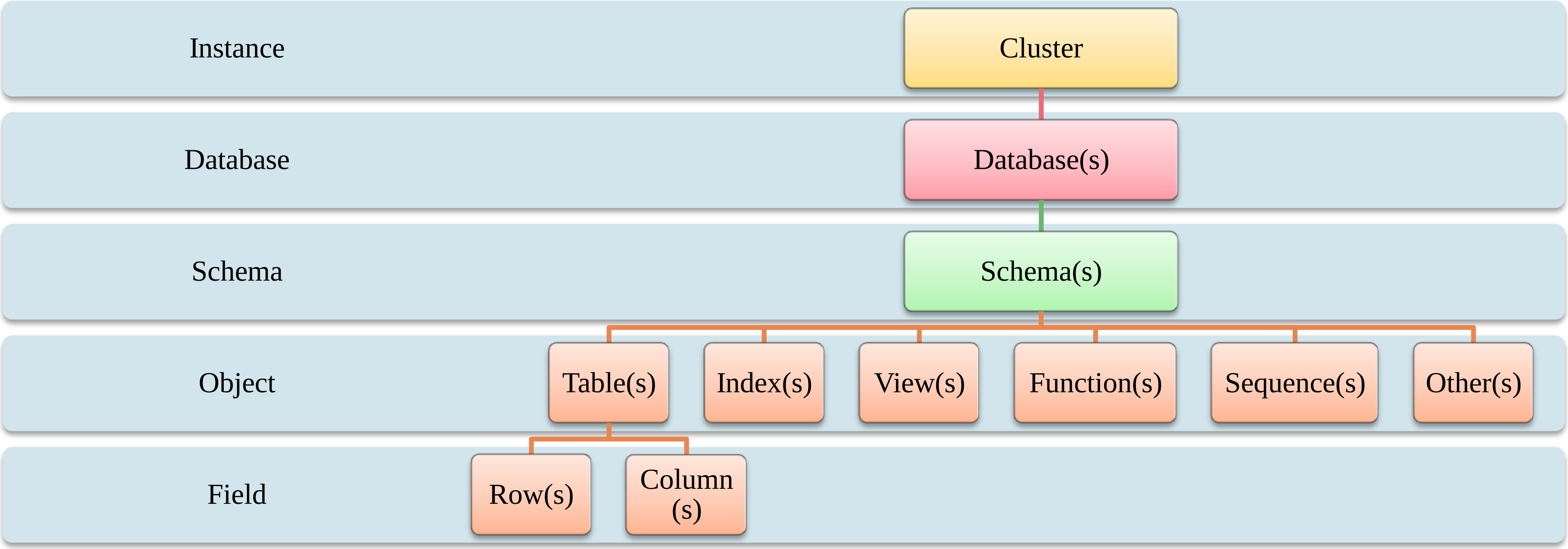
Greenplum DW

aster data DW
big data. fast insights.

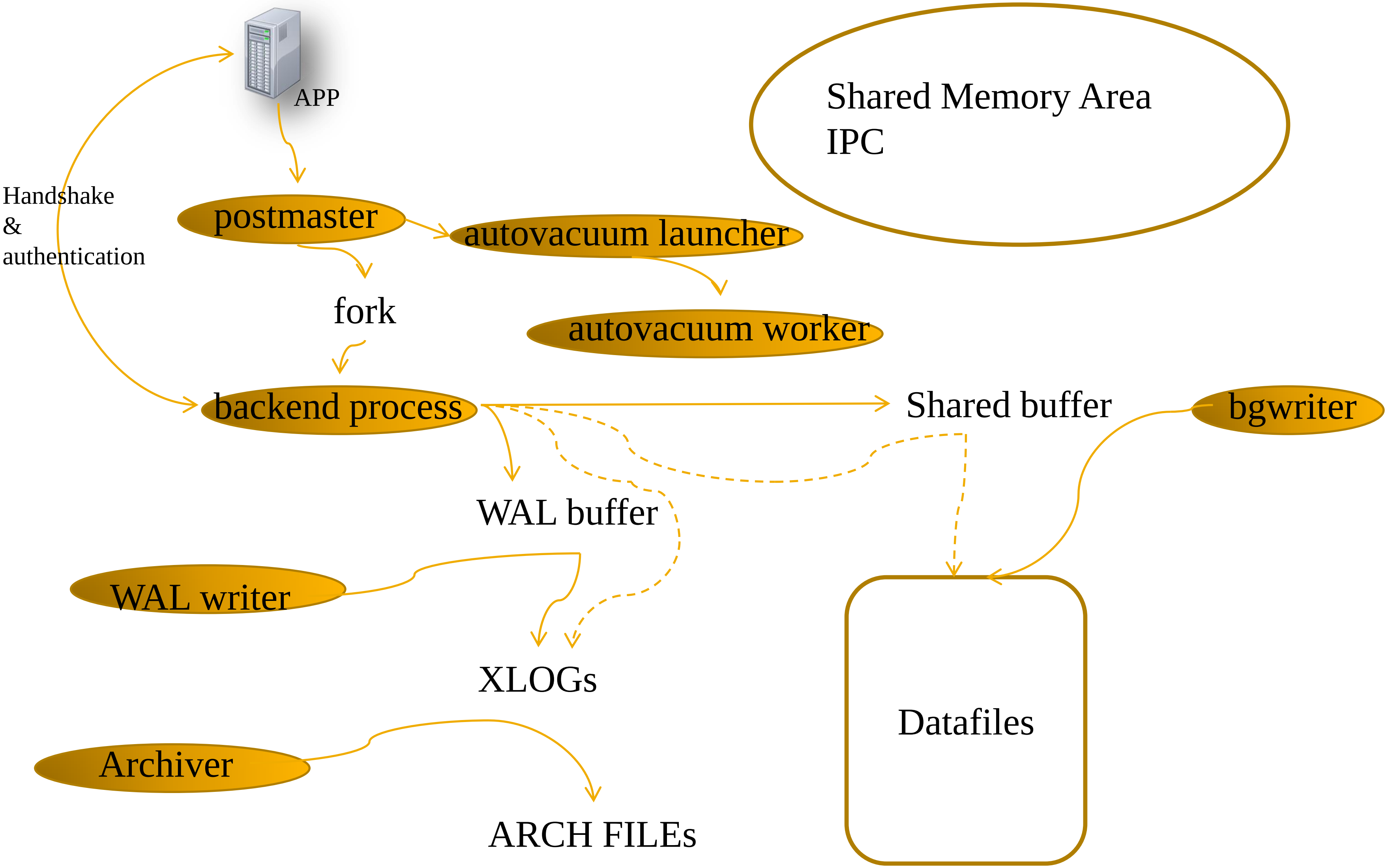
Portion Contributers


EnterpriseDB
OpenFTS full text search engine
Greenplum
Tsearch2
PGCluster
PostgreSQL
Slony-I enterprise-level replication system
pgpool-II
CMD
skype
PostGIS Spatial PostgreSQL
SE-PostgreSQL [Security Enhanced]
Bucardo

Logical Layout



Process Introduction



APP

Handshake
&
authentication

postmaster

fork

backend process

autovacuum launcher

autovacuum worker

Shared Memory Area
IPC

Shared buffer

bgwriter

WAL buffer

WAL writer

XLOGs

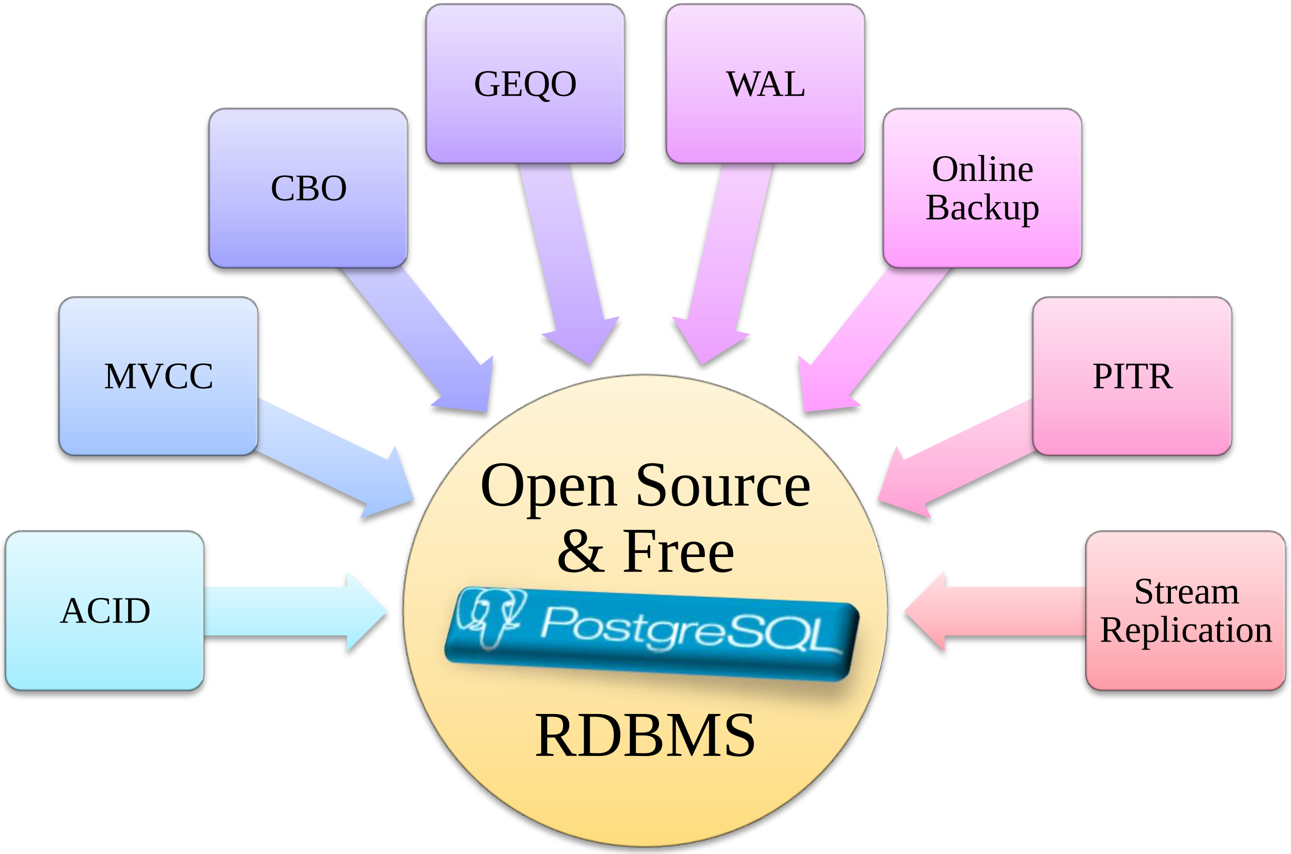
Archiver

ARCH FILEs

Datafiles

SKYmobi 新手机 新应用 新娱乐

Potion Features



Powerful Localization Support

- Supported Character Sets
  - http://www.postgresql.org/docs/9.1/static/multibyte.html
- Support Database and Column level COLLATE
  - Example : CREATE TABLE test1 (a text COLLATE "de_DE", b text COLLATE "es_ES", ...);

Powerful Platform Support

- X86
- X86_64
- IA64
- PowerPC
- PowerPC 64
- S/390
- S/390x
- Sparc
- Sparc 64
- Alpha
- ARM
- MIPS
- MIPSEL
- M68K
- PA-RISC

PostgreSQL

- Linux
- Windows
- FreeBSD
- OpenBSD
- NetBSD
- Mac OS X
- AIX
- HP/UX
- IRIX
- Solaris
- Tru64 Unix
- UnixWare

Rich Extensions

- adminpack
- auto_explain
- btree_gin
- btree_gist
- chkpass
- citext
- cube
- dblink
- dict_int
- dict_xsyn
- earthdistance
- fuzzystrmatch
- hstore
- intagg
- intarray
- isn
- lo
- ltree
- oid2name
- pageinspect
- passwordcheck
- pg_buffercache
- pg_freespacemap
- pg_standby
- pg_stat_statements
- pg_test_fsync
- pg_trgm
- pg_upgrade
- pgbench
- pgcrypto
- pgrowlocks
- pgstattuple
- seg
- sepgsql
- spi
- sslinfo
- start-scripts
- tablefunc
- test_parser
- tsearch2
- unaccent
- uuid-ossp
- vacuumlo
- xml2

Potion Compare

ORACLE

1. Language

SQL/Plsql

2. Index

Global / Partition

3. DDL Rollback

Cann’t rollback but can recovery from Backup or Flash Recovery Area.

4. Compress

Table Level

5. Trigger

6. Data Type

……

PostgreSQL

1. Language

SQL/Plpgsql/Pltcl/Plperl/Plpython…

2. Index

Global(non-partition TABLE)

Partition

Partial Index

3. DDL Rollback

Can rollback every ddl sql.

4. Compress

Column Level(Limited)

5. Trigger / Rule

6. Data Type extention

IP / MAC / XML / UUID / …

……

Limit

Limit	Value
Maximum Database Size	Unlimited
Maximum Table Size	32 TB
Maximum Row Size	1.6 TB
Maximum Field Size	1 GB
Maximum Rows per Table	Unlimited
Maximum Columns per Table	250 - 1600 depending on column types
Maximum Indexes per Table	Unlimited

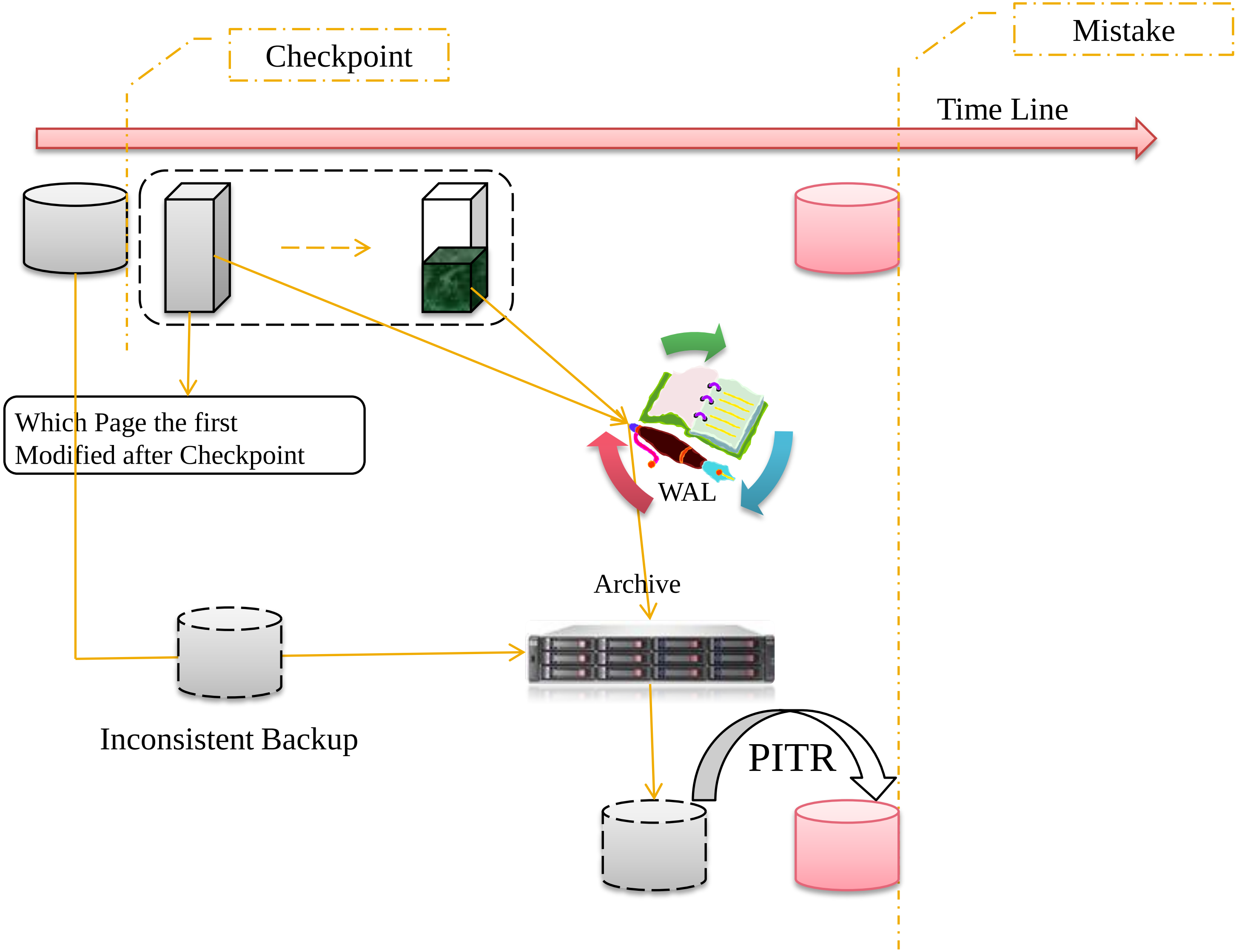
Reliability

- ACID
  - Atomicity
    - All Success or All Fail
  - Consistency
    - Only valid data will be written to the database
    - Example： check (age>=0)
  - Isolation
    - SERIALIZABLE | REPEATABLE READ | READ COMMITTED | READ UNCOMMITTED
  - Durability
    - The ability of the DBMS to recover the committed transaction updates against any kind of system failure (hardware or software).

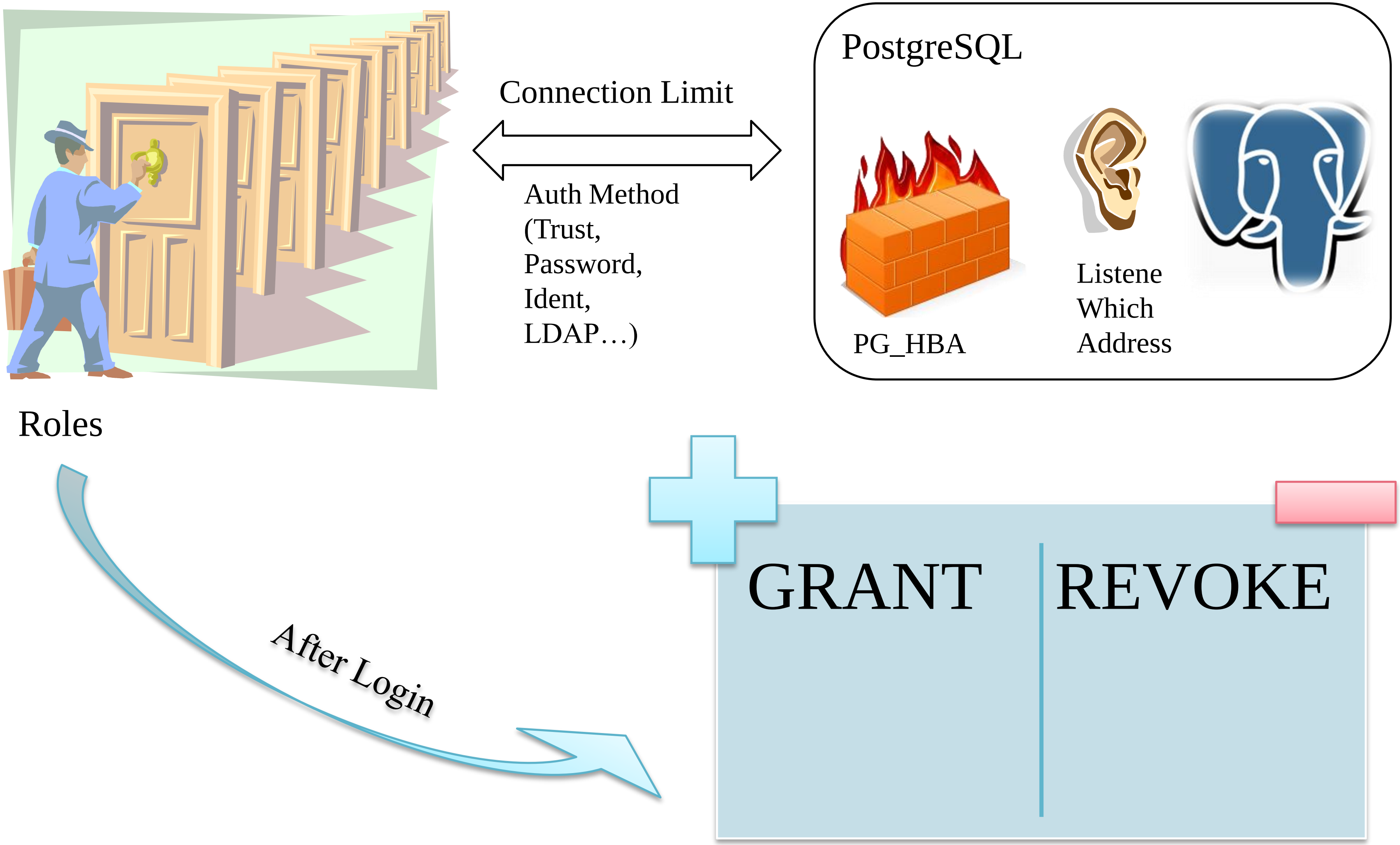
Recoverability

- Requirement
  - Baseline Backup
  - Parameter
    - Open fsync,full_page_writes
    - Optional open synchronous_commit
  - Open WAL Backup

Recoverability



Security



Roles

Connection Limit

Auth Method
(Trust,
Password,
Ident,
LDAP…)

PostgreSQL

PG_HBA

Listene
Which
Address

After Login

GRANT

REVOKE

Scalability

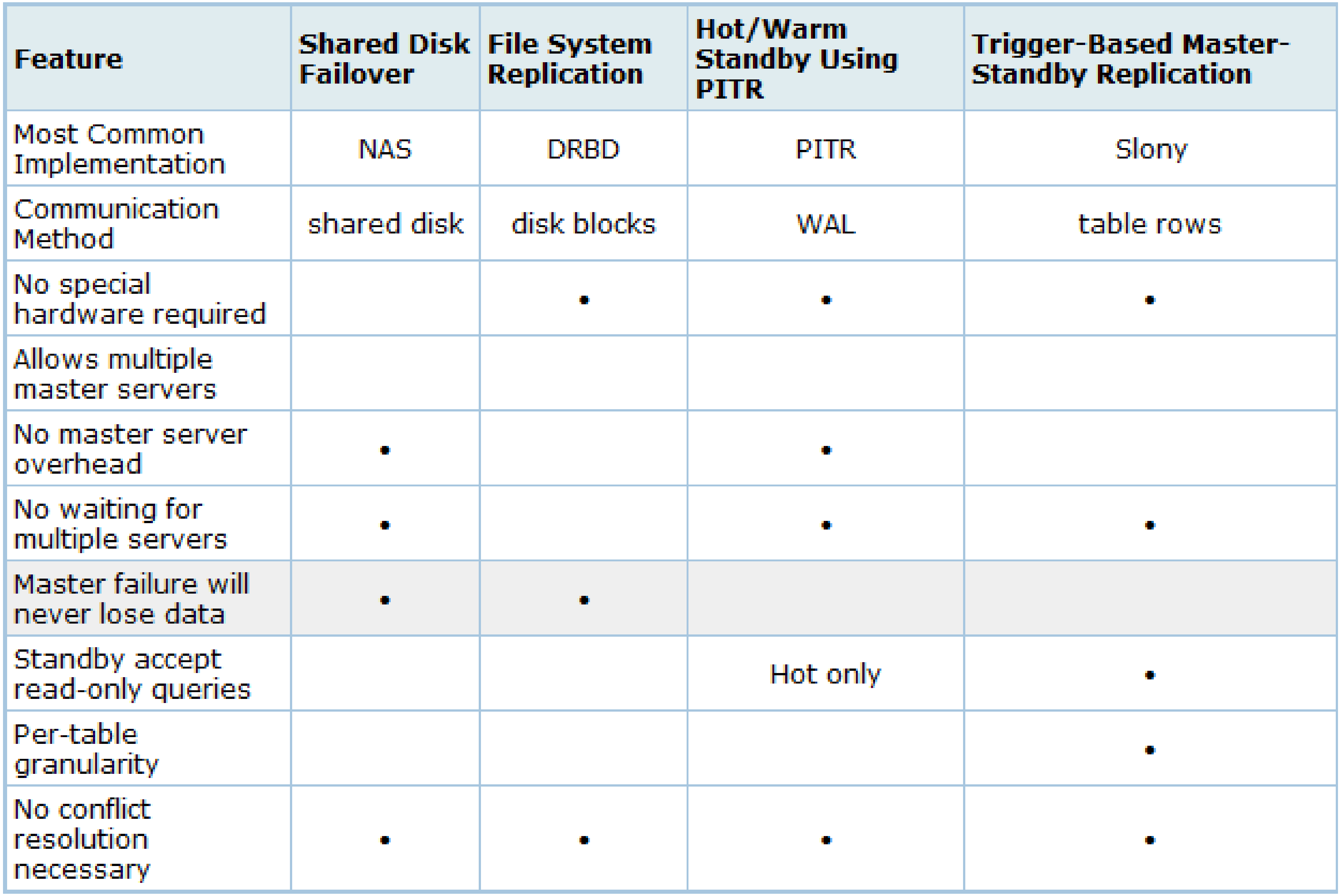
- Hardware
- Software

Project	Type	Method	Storage
Plproxy	OLTP	Distributed	Can Shared-nothing
GridSQL	DW	Distributed	Can Shared-nothing
GreenPlum	DW	Distributed	Shared-nothing
Aster Data	DW	Distributed	Shared-nothing
Postgres-XC	OLTP	Distributed	Can Shared-nothing
Pgpool-II	DW	Distributed	Can Shared-nothing
Sequoia/Continuent	OLTP	Distributed	Can Shared-nothing
PGMemcache	OLTP	Distributed	Cache

Performance

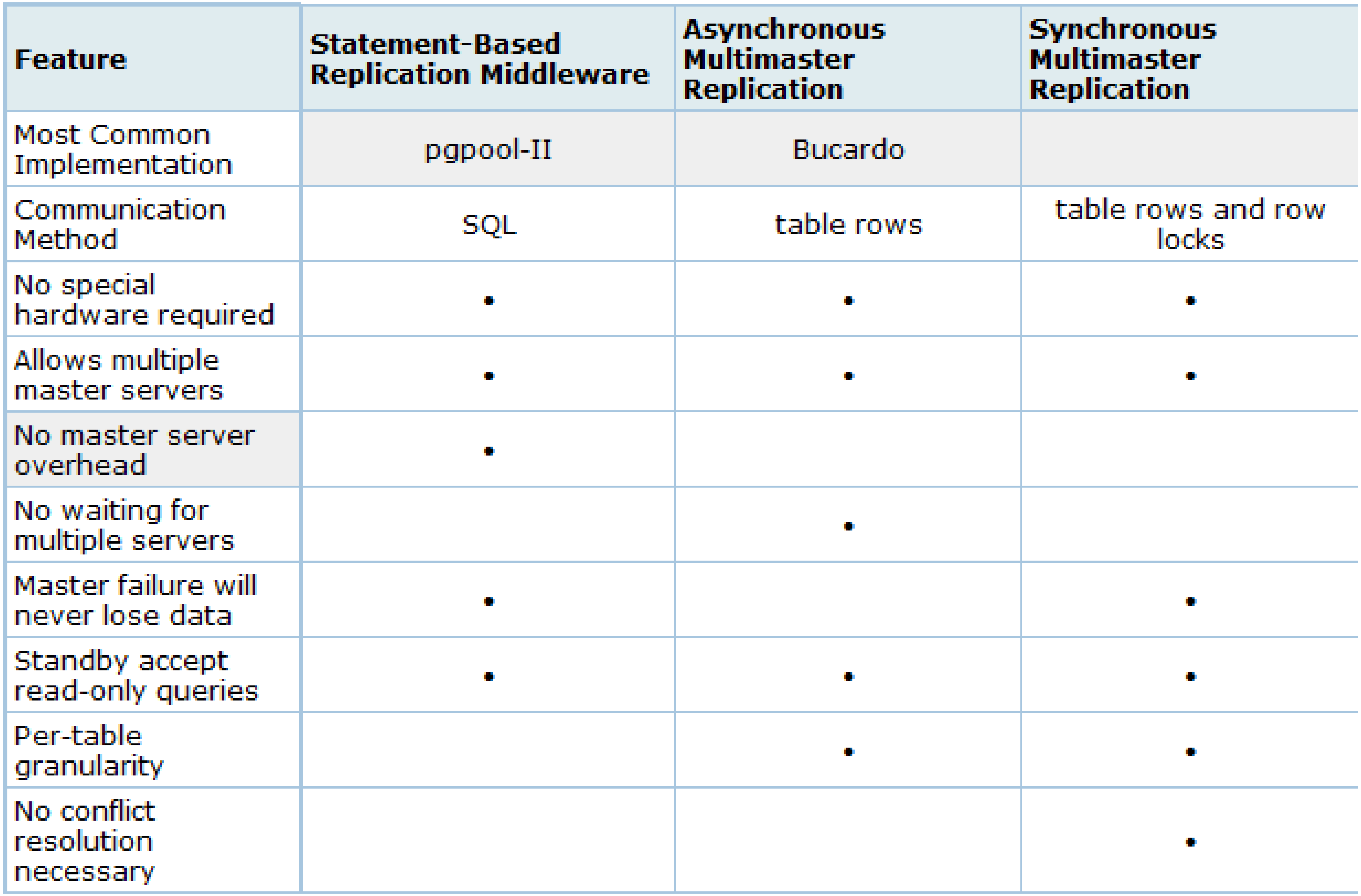
- SAIO Optimizer
    - wulczer.org
- Virtual Index
- Prefetch
- Cache State Persistent
- Tablespace Based IO Cost Value
- Async IO
- Partial Index
- Parallel restore

High-Availability



Feature	Shared Disk Failover	File System Replication	Hot/Warm Standby Using PITR	Trigger-Based Master-Standby Replication
Most Common Implementation	NAS	DRBD	PITR	Slony
Communication Method	shared disk	disk blocks	WAL	table rows
No special hardware required		•	•	•
Allows multiple master servers				
No master server overhead	•		•	
No waiting for multiple servers	•		•	•
Master failure will never lose data	•	•		
Standby accept read-only queries			Hot only	•
Per-table granularity				•
No conflict resolution necessary	•	•	•	•

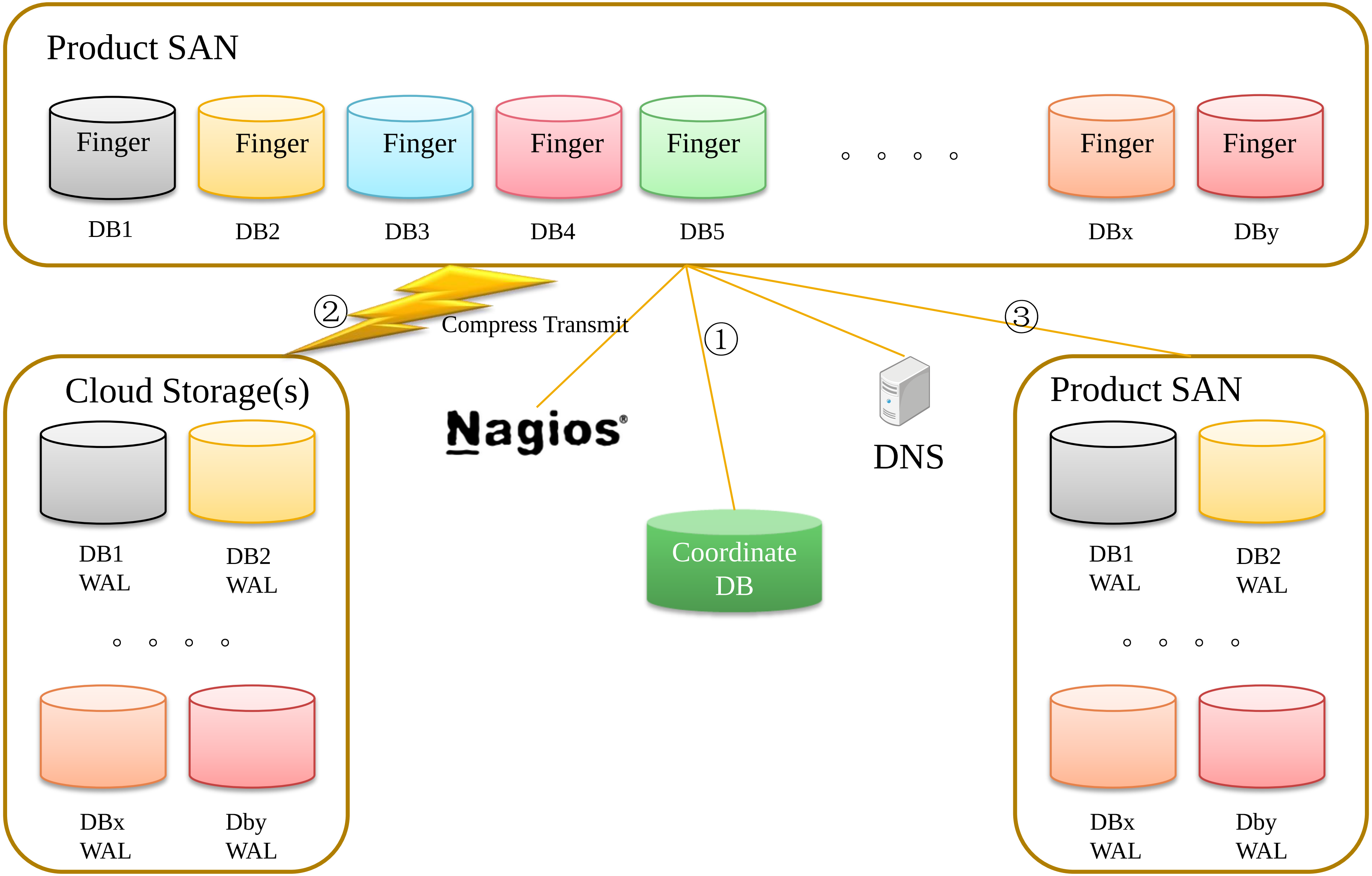
High-Availability



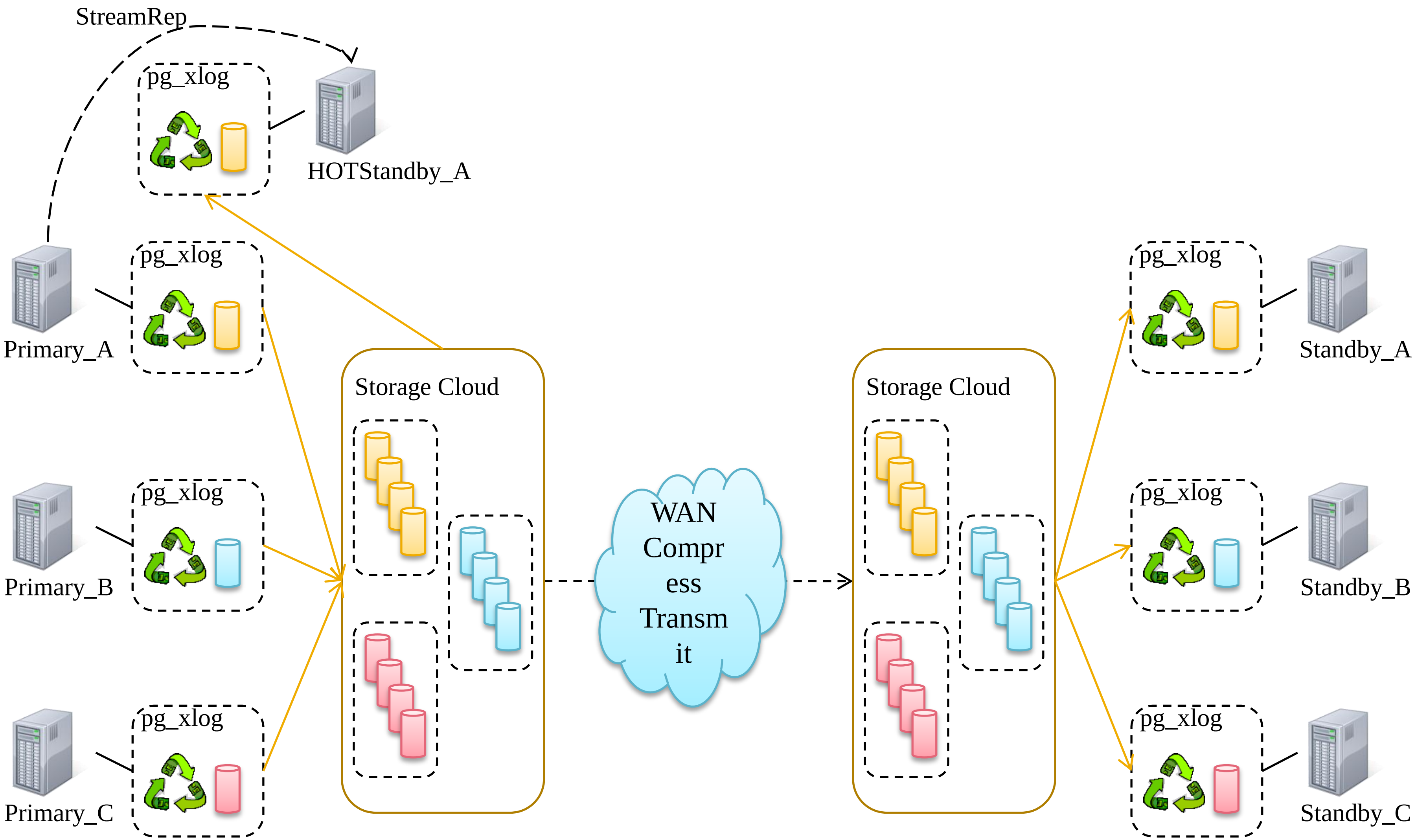
Feature	Statement-Based Replication Middleware	Asynchronous Multimaster Replication	Synchronous Multimaster Replication
Most Common Implementation	pgpool-II	Bucardo	
Communication Method	SQL	table rows	table rows and row locks
No special hardware required	•	•	•
Allows multiple master servers	•	•	•
No master server overhead	•		
No waiting for multiple servers		•	
Master failure will never lose data	•		•
Standby accept read-only queries	•	•	•
Per-table granularity		•	•
No conflict resolution necessary			•

新手机 新应用 新娱乐

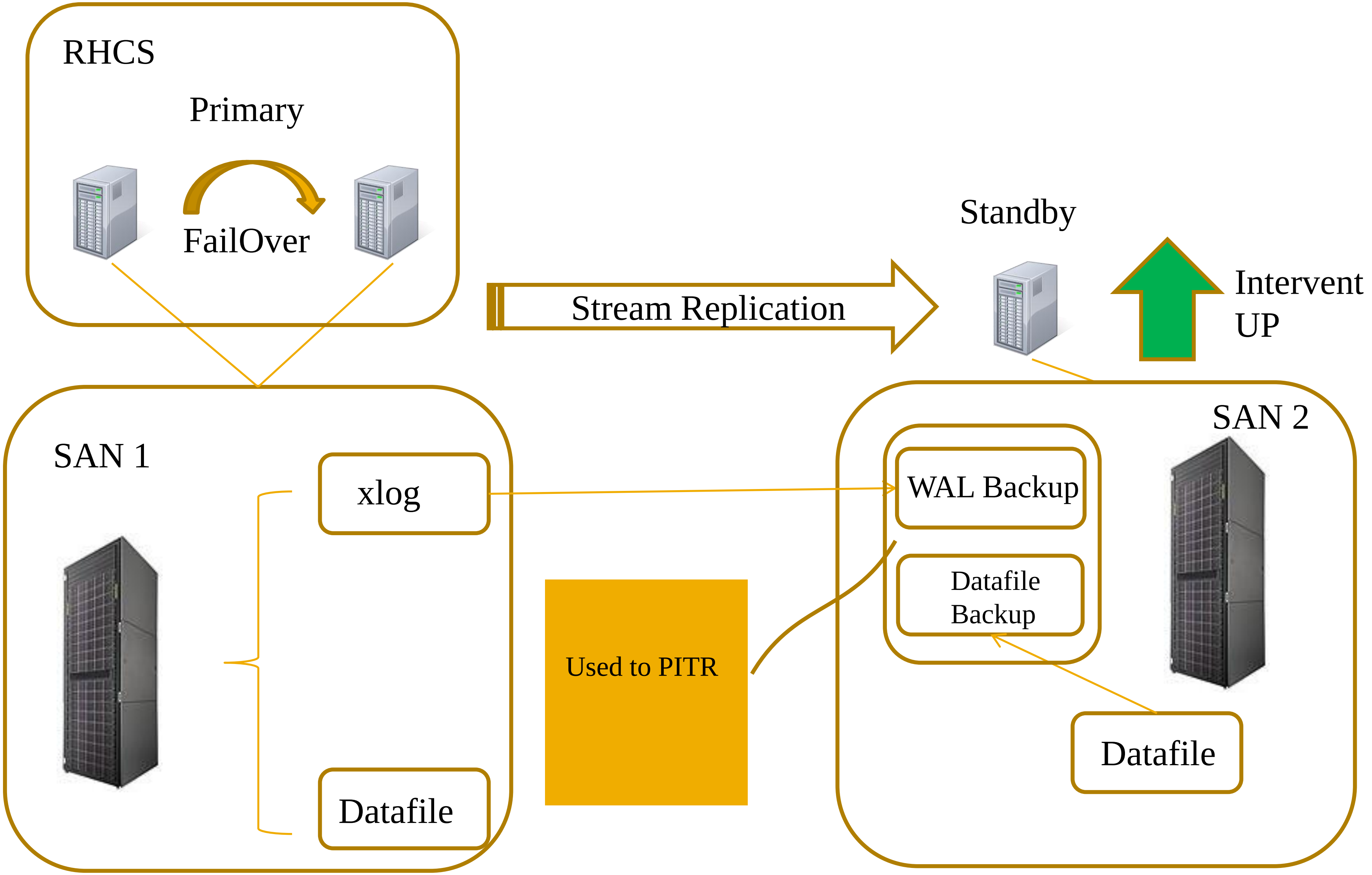
Archive Case



HA & DR Case


StreamRep
pg_xlog
HOTStandby_A
pg_xlog
Primary_A
pg_xlog
Primary_B
pg_xlog
Primary_C
Storage Cloud
WAN
Compr
ess
Transm
it
Storage Cloud
pg_xlog
Standby_A
pg_xlog
Standby_B
pg_xlog
Standby_C

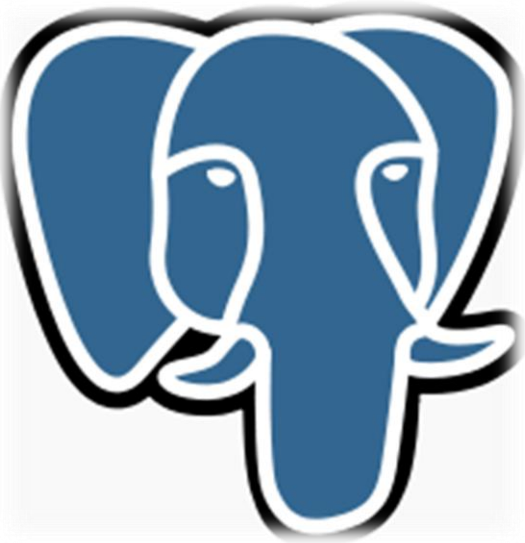
Shard-everything HA Case



SKYmobi 新手机 新应用 新娱乐

Thanks

- Thanks all people contribute to PostgreSQL.



- Digoal.Zhou
- Blog
- http://blog.163.com/digoal@126