# PostgreSQL互联网应用

Building Economy,Scalable,Security Database System

Digoal

2010-06-19

# Contents

- PostgreSQL Introduction
- Database Life Cycle Introduction
- Specialized
- Building Block

# PostgreSQL Introduction

- Origin
- Standard
- Platform
- Globalization Support
- Features
- Limit
- Contribute

# PostgreSQL Introduction

- Origin

INGRES 1973

DARPA

POSTGRES 1985

Postgres95 1995

PostgreSQL 1996

Michael Stonebraker

Informix

VERTICA

C-Store

EnterpriseDB OLTP

Greenplum OLAP

aster data OLAP
big data. fast insights.

SciDB

VoltDB

H-Store

MARIPOSA

Federated database system

PeopleSoft

ORACLE

# PostgreSQL Introduction

- Standard


SQL-92

SQL:1999

SQL:2003

SQL:2008

# PostgreSQL Introduction

- Platform

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


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

# PostgreSQL Introduction

- Globalization Support

ENCODING

LC_COLLATE

LC_CTYPE

TIMEZONE

# PostgreSQL Introduction

- Features

1.Functions
Returning rows
Returning void
PL/pgSQL
PL/lua
PL/LOLCODE
PL/Perl
plPHP
PL/Python
PL/Ruby
PL/sh
PL/Tcl
PL/Scheme
C
C++
PL/Java
PL/R

2.Indexes
Btree
HASH
GiST
GiN
Express Index
Partial index
Bitmap Index
3.Trigger event
DML
Truncate
Before
After
Row
Statement
4.MVCC
Ensure ACID

5.Rules
Rewrite QTree
After Parsing
Before Qplan
DO INSTEAD
DO ALSO
DO NOTHING
6.Data Types
Up to 1G Field
Geometric
IPv4 / IPv6
CIDR / MAC
XML
UUID
User Type

7.User-Def Obj
Casts
Conversions
Data types
Domains
Functions
Indexes
Operators
Procedural LG
8.Inheritance
Parent-TAB
Child-TAB
Partition-TAB
9.SSL Conn
10.Tablespace
11.Savepoints
12.PITR

13.TOAST
14.Regular Exp
15.Embed SQL
16.Transaction
DDL

# PostgreSQL Introduction

- 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 |

# PostgreSQL Introduction

- Contribute















# PostgreSQL Introduction

- Commencial,Where we are?

SYBASE

400

ORACLE

1000

TERADATA

900

DB2

500

Microsoft SQL Server

2000

# PostgreSQL Introduction

- Open Source
  - Where we are?
    - No Bound,No Limit

# Database Life cycle Introduction

- Investigate
- Develop
- Testing
- Deploy
- Administrate

# Database Life cycle Introduction

- Investigate

RDBMS

DB2

PostgreSQL

ORACLE

MySQL

VB/LDB

GridSQL®

Greenplum

EDB-Icache

Postgres-XC

VoltDB

ORACLE TIMESTEN

Special

Tsearch2

mongoDB

apache CouchDB relax

PostGIS Spatial PostgreSQL

# Database Life cycle Introduction

- Develop
  - Develop IDE
    - PGAdmin
    - EMS
    - TOAD
    - 。 。 。
  - Develop Language
    - PLpgsql
    - PLJava
    - SQL
    - 。 。 。

# Database Life cycle Introduction

- Testing
  - Building Testing Model
  - Function Testing
  - Press Testing
- Deploy
  - Deploy Database
  - Apply Database Scripts
- Administrate
  - Monitor
  - Performance Tuning
  - Backup Maintenance
  - HA Deploy
  - 。 。 。

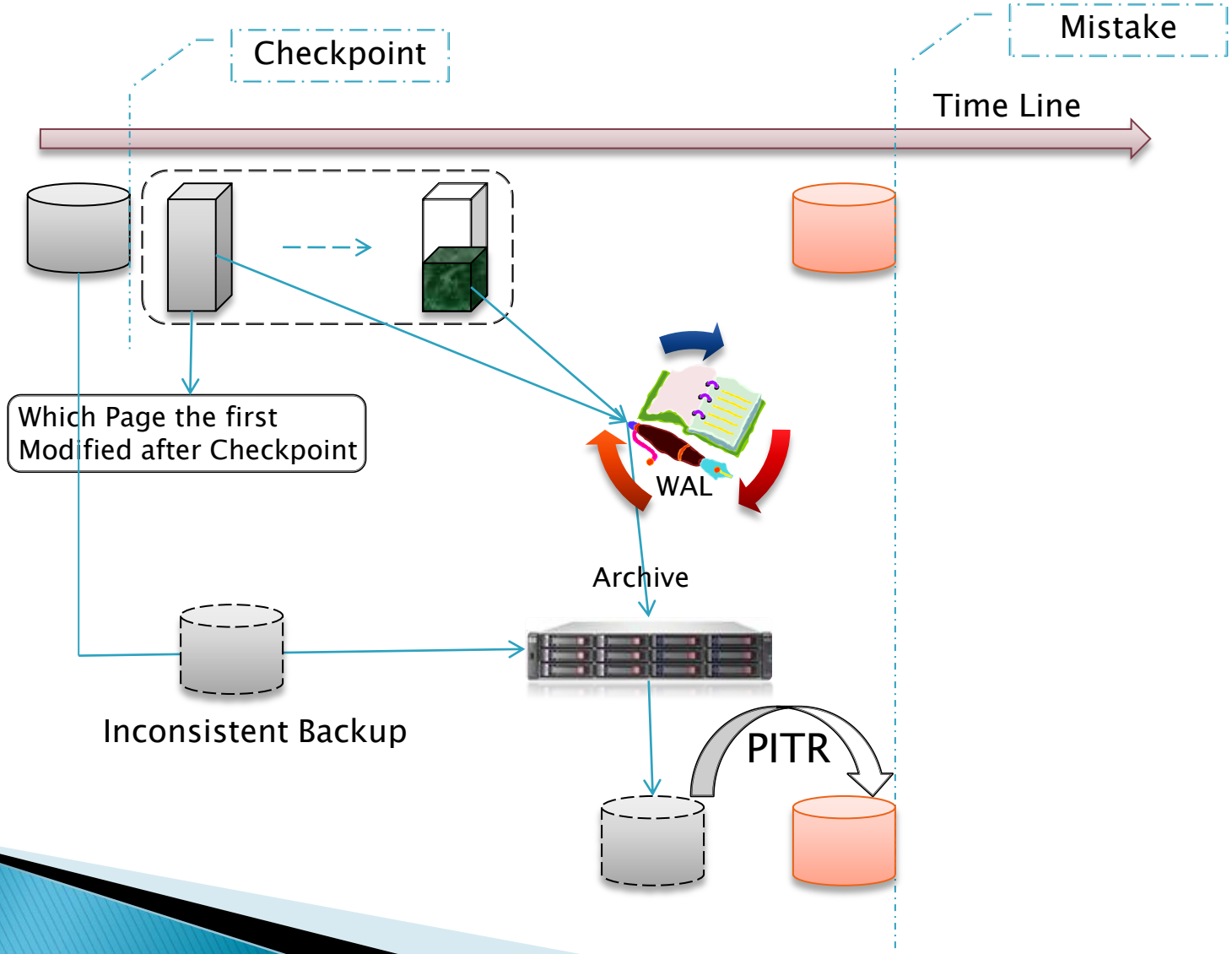
# Specialized

- Reliability
- Security
- Scalable
- Performance
- High-Availability
- Warehouse
- Monitor
- Administrate

# Reliability

- WAL
  - fsync,full_page_writes
- Checkpoints
- Archive
- PITR

# Reliability



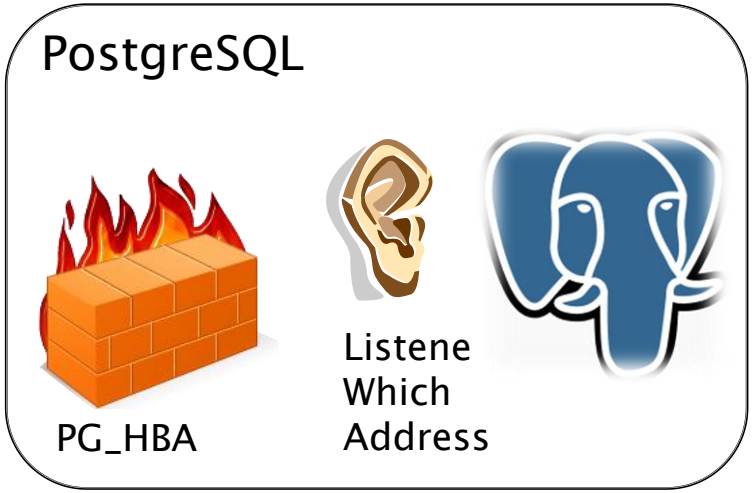
# Security

- Authenticate

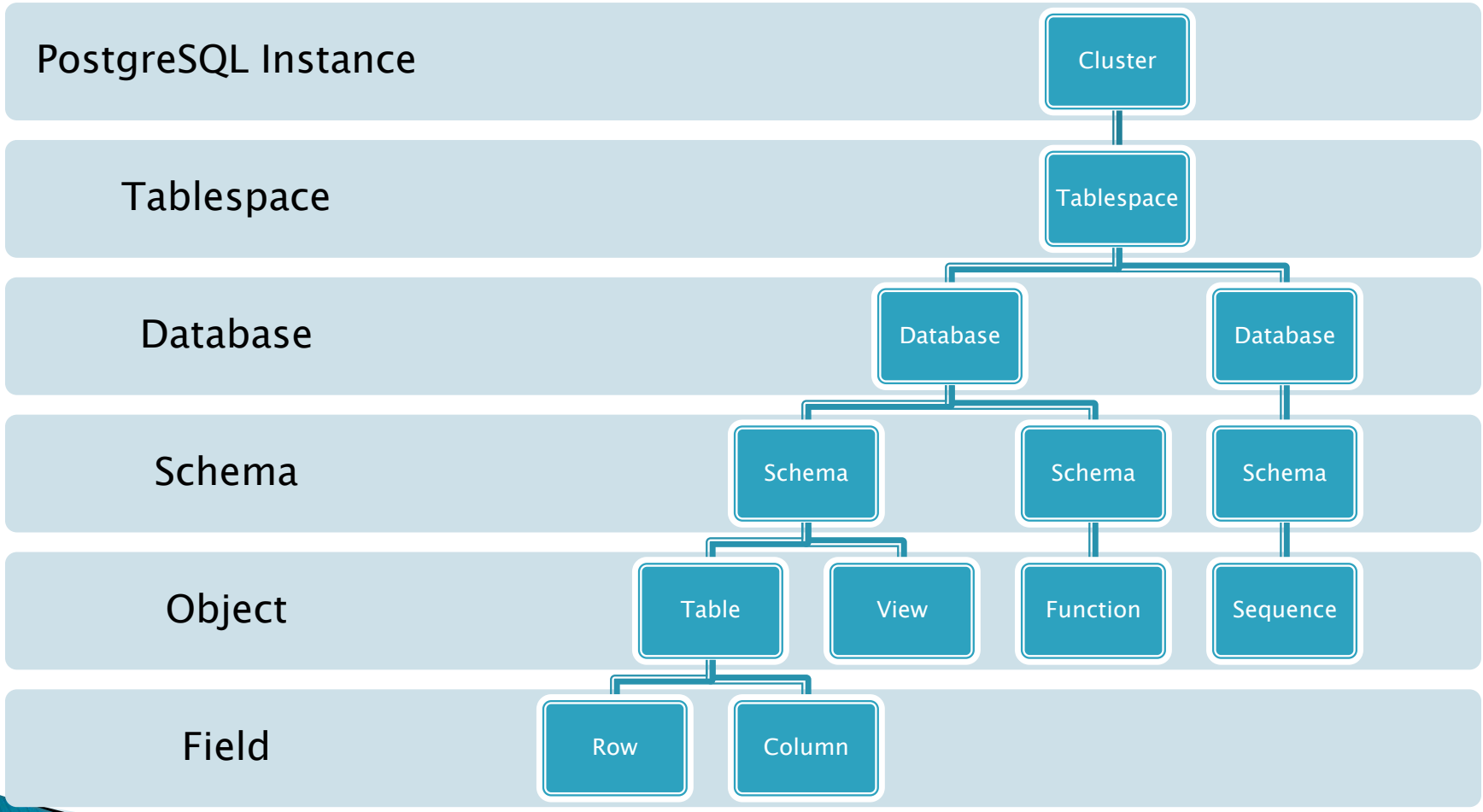
Roles

Connection Limit

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

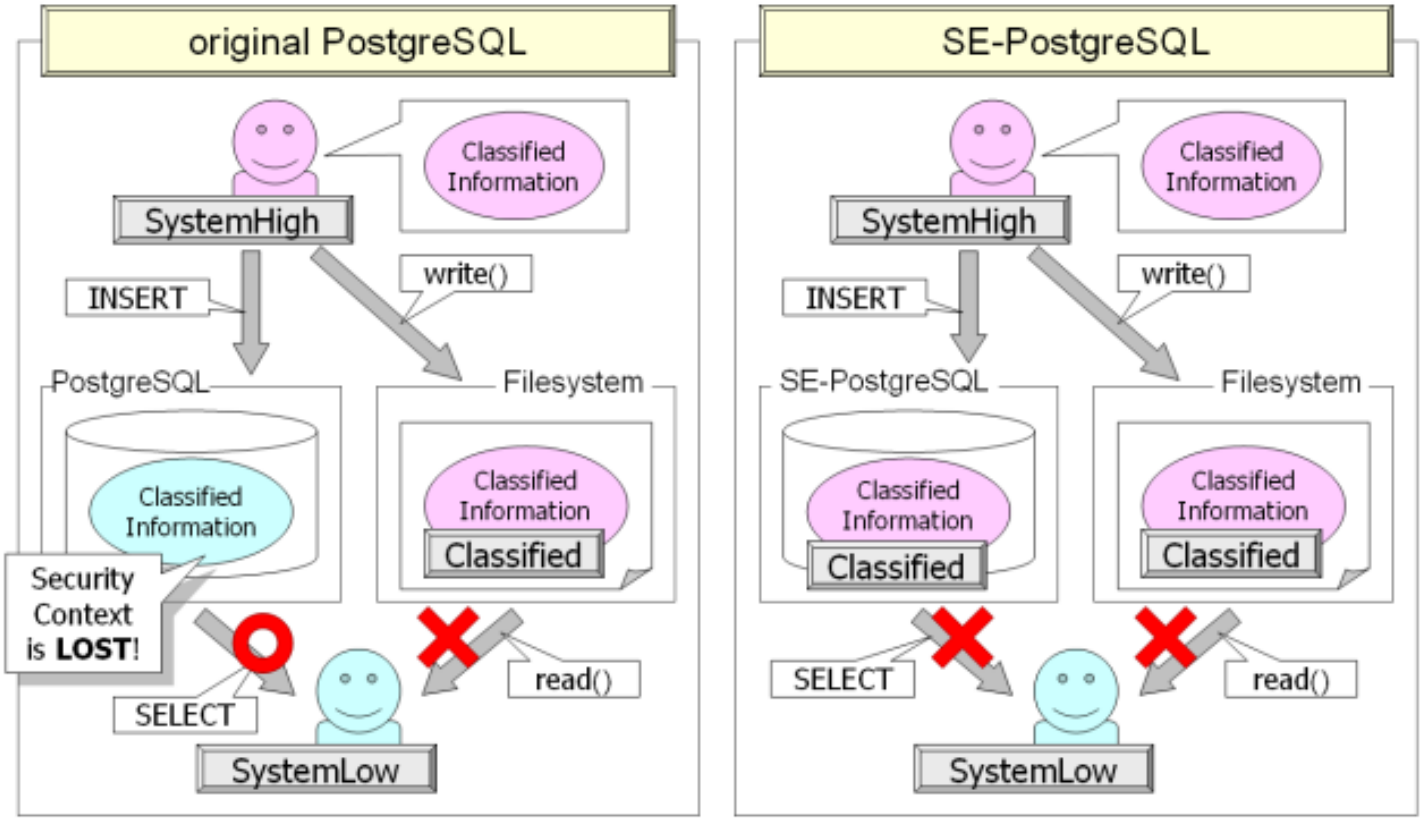


# Security


PostgreSQL Instance
Tablespace
Database
Schema
Object
Field
Cluster
Tablespace
Database
Database
Schema
Schema
Schema
Table
View
Function
Sequence
Row
Column

# Security

- SE-PostgreSQL



# Scalable

- Hardware
- Scalable

| 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 | 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-Slave 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 | • | • | | |
| Slaves 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 | • | | • |
| Slaves accept read-only queries | • | • | • |
| Per-table granularity | | • | • |
| No conflict resolution necessary | | | • |

# Warehouse

- Window Function
- Massively Parallel Processing
- Transparent Compress
- Table Partitioning

Greenplum

aster data
big data. fast insights.

# Monitor

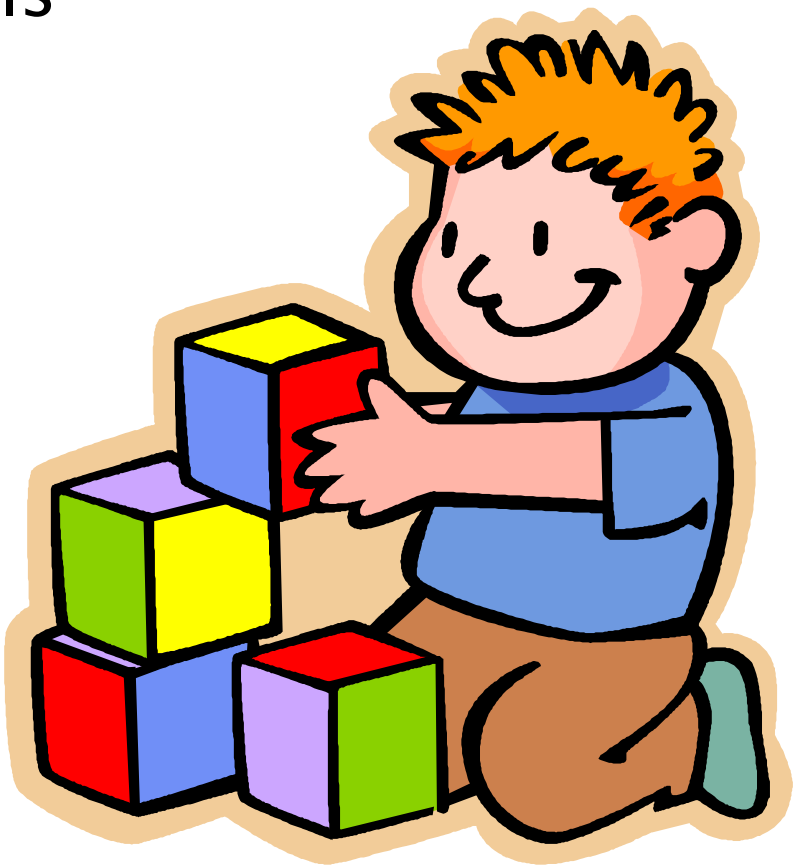
- Target
- Real time
- Contributed Tools
  - pg_statsinfo
  - HQ
  - pgsnap
  - pg_statspack
  - check_postgres
  - nagios,munin,cacti,mrtg,circonus,reconnoiter,traffic objects

# Administrate

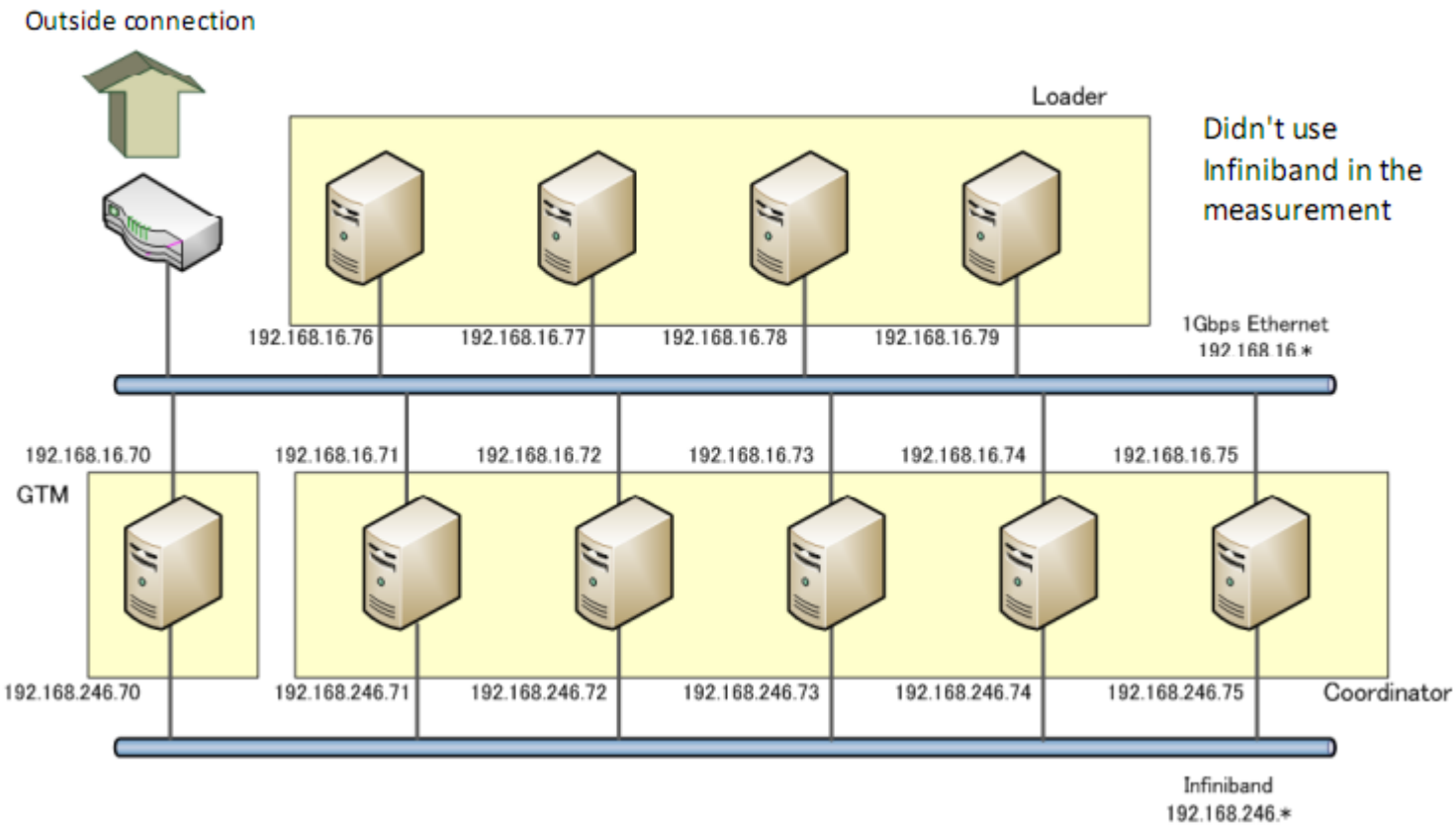
- Audit
- Backup
- Monitor
- Performace Tuning
- Maintenance

# Building Block

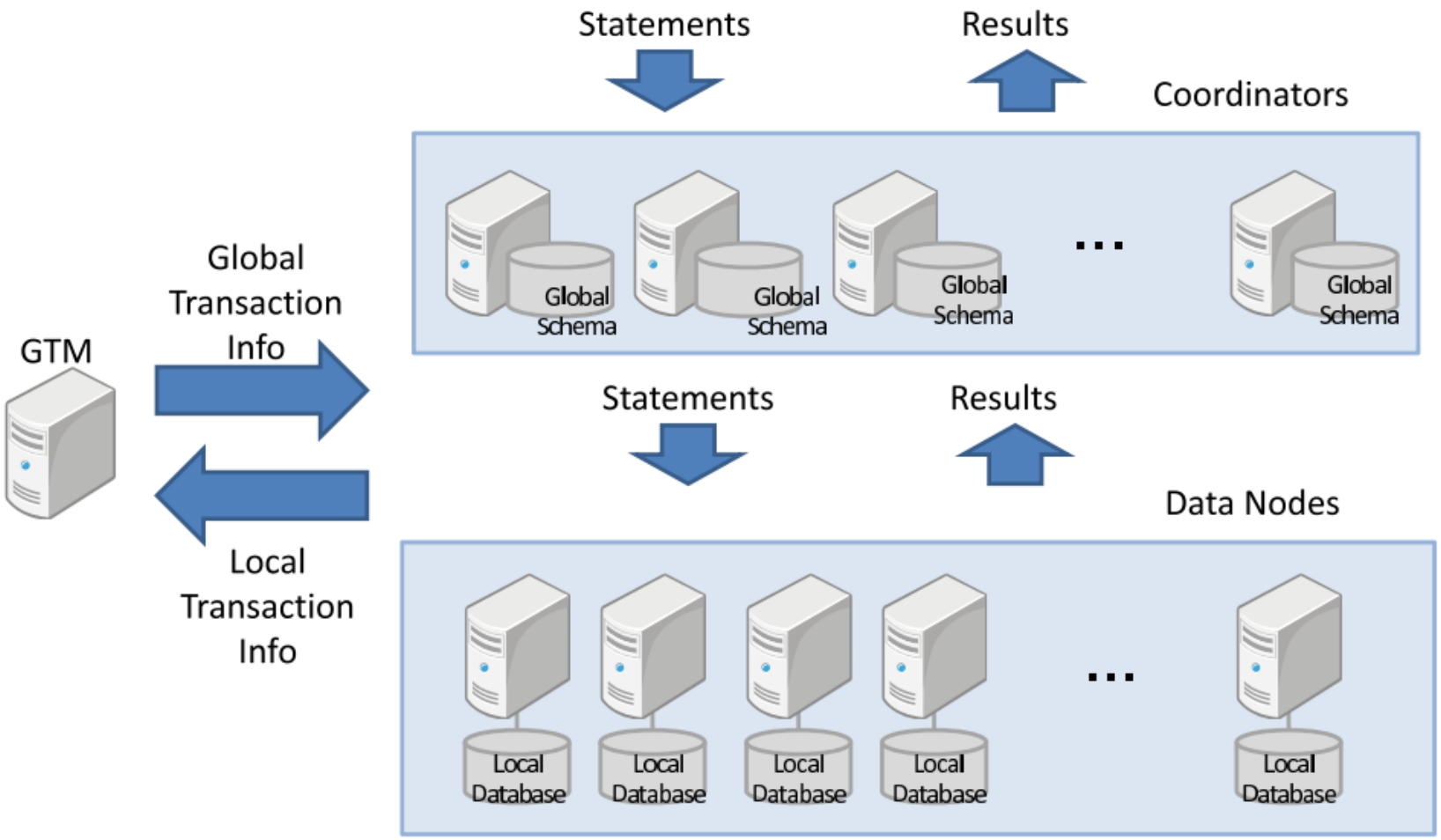
- Choose Blocks which fits your current application goals
- E.g.



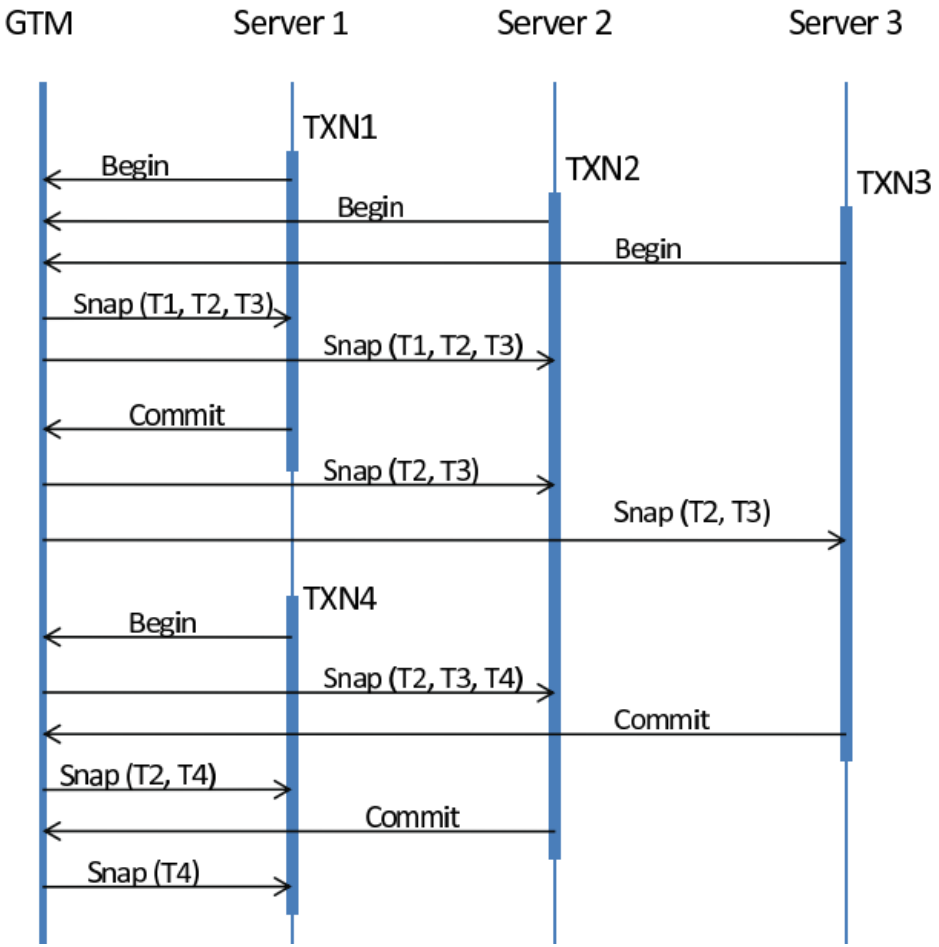
# Postgres-XC


Outside connection
Loader
Didn't use Infiniband in the measurement
192.168.16.76
192.168.16.77
192.168.16.78
192.168.16.79
1Gbps Ethernet
192.168.16.*
192.168.16.70
192.168.16.71
192.168.16.72
192.168.16.73
192.168.16.74
192.168.16.75
GTM
192.168.246.70
192.168.246.71
192.168.246.72
192.168.246.73
192.168.246.74
192.168.246.75
Coordinator
Infiniband
192.168.246.*

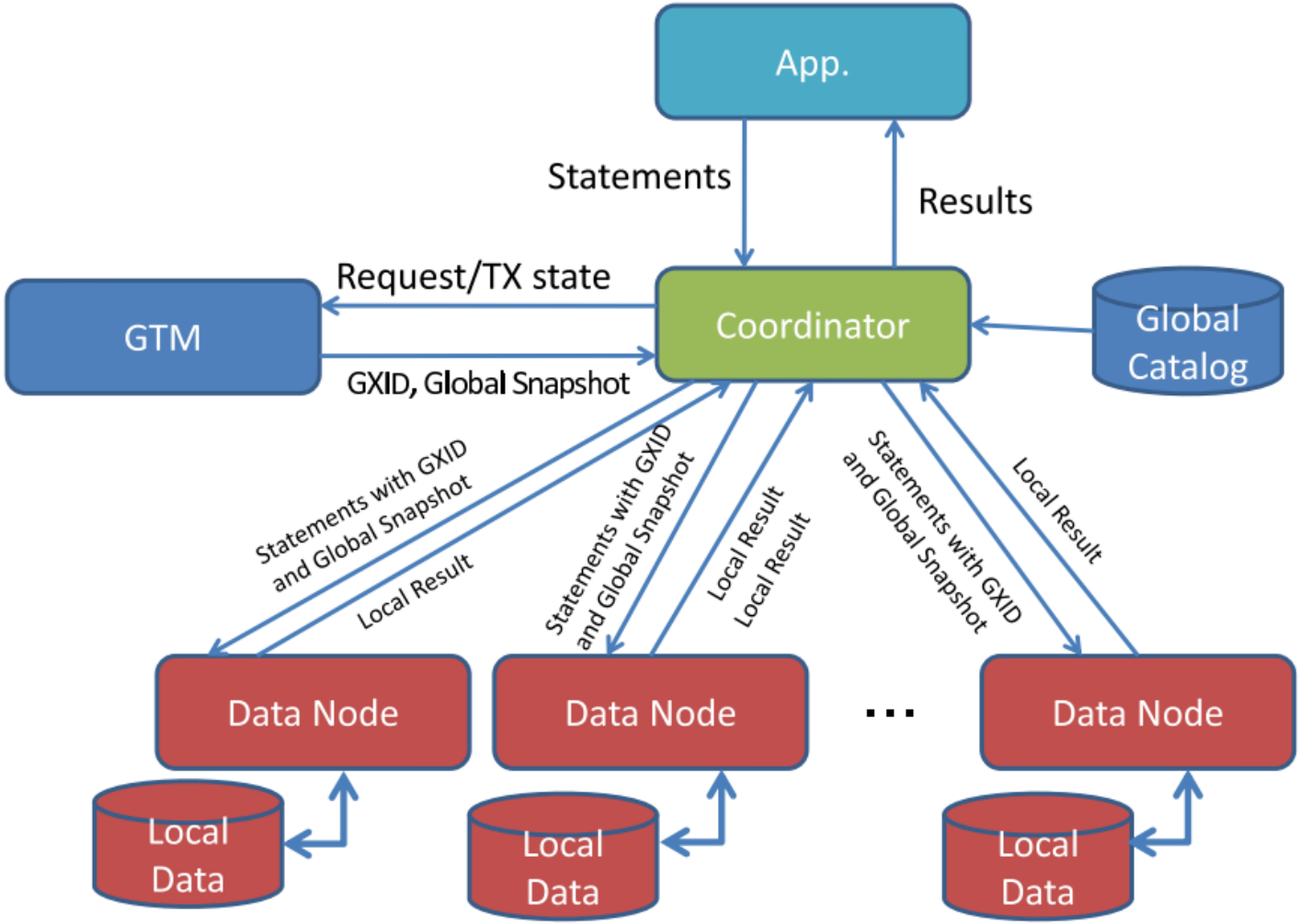
# Postgres-XC


Statements
Results
Coordinators
Global Schema
Global Schema
Global Schema
...
Global Schema
GTM
Global Transaction Info
Local Transaction Info
Statements
Results
Data Nodes
Local Database
Local Database
Local Database
Local Database
...
Local Database

# Postgres-XC


GTM
Server 1
Server 2
Server 3
TXN1
TXN2
TXN3
Begin
Begin
Begin
Snap (T1, T2, T3)
Snap (T1, T2, T3)
Commit
Snap (T2, T3)
Snap (T2, T3)
TXN4
Begin
Snap (T2, T3, T4)
Commit
Snap (T2, T4)
Commit
Snap (T4)

# Postgres-XC


App.
Statements
Results
Request/TX state
GTM
Coordinator
Global Catalog
GXID, Global Snapshot
Statements with GXID and Global Snapshot
Local Result
Statements with GXID and Global Snapshot
Local Result
Local Result
Statements with GXID and Global Snapshot
Local Result
Data Node
Data Node
...
Data Node
Local Data
Local Data
Local Data

# Postgres-XC

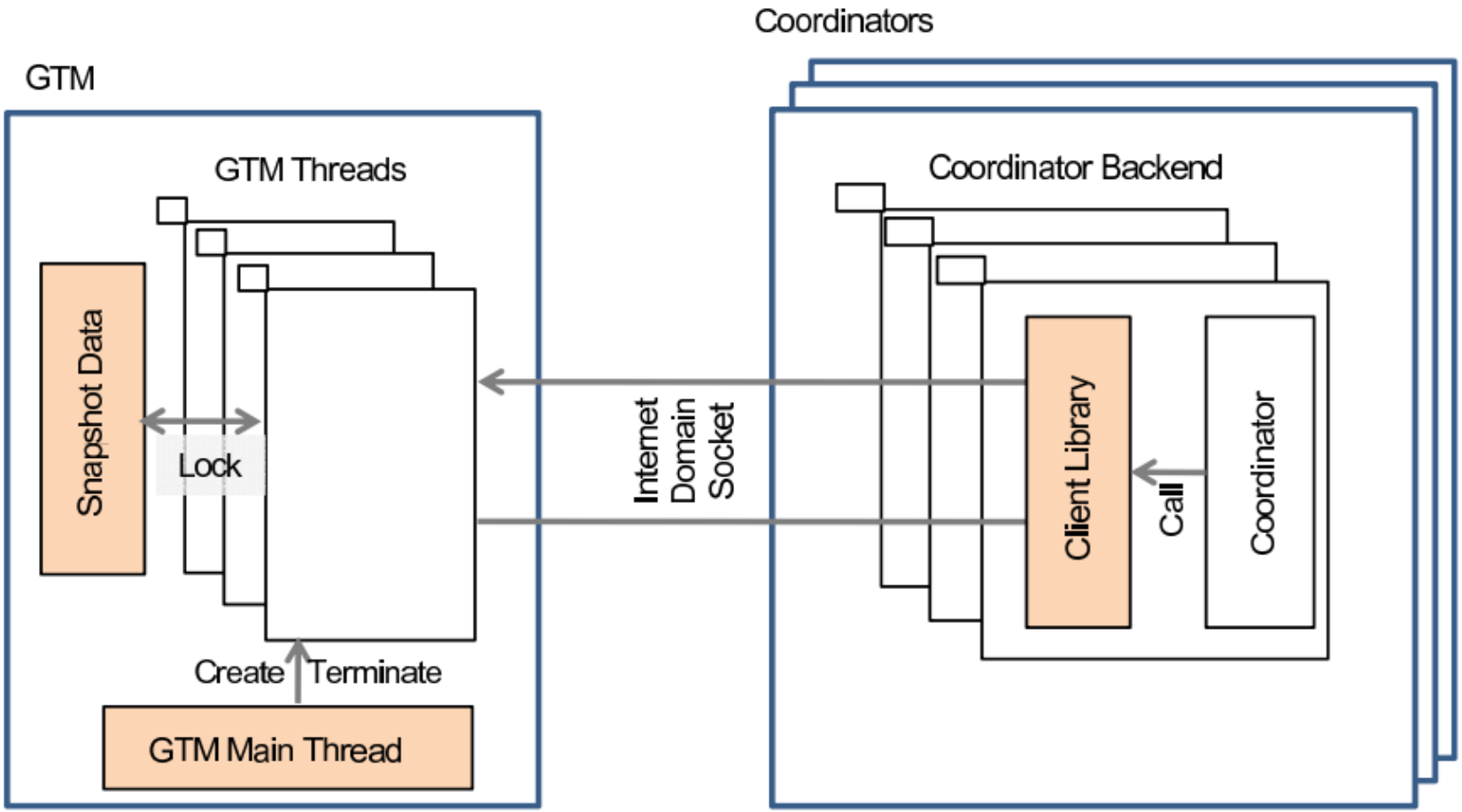



Figure 7: Primitive GTM structure

# Postgres-XC

| Database | Num.of Servers | Throughput (TPS) | Scale Factor |
|---|---|---|---|
| PostgreSQL | 1 | 2,500 | 1.0 |
| Postgres-XC | 1 | 1,900 | 0.72 |
| Postgres-XC | 2 | 3,630 | 1.45 |
| Postgres-XC | 3 | 5,568 | 2.3 |
| Postgres-XC | 5 | 8,500 | 3.4 |
| Postgres-XC | 10 | 16,000 | 6.4 |

# Postgres-XC

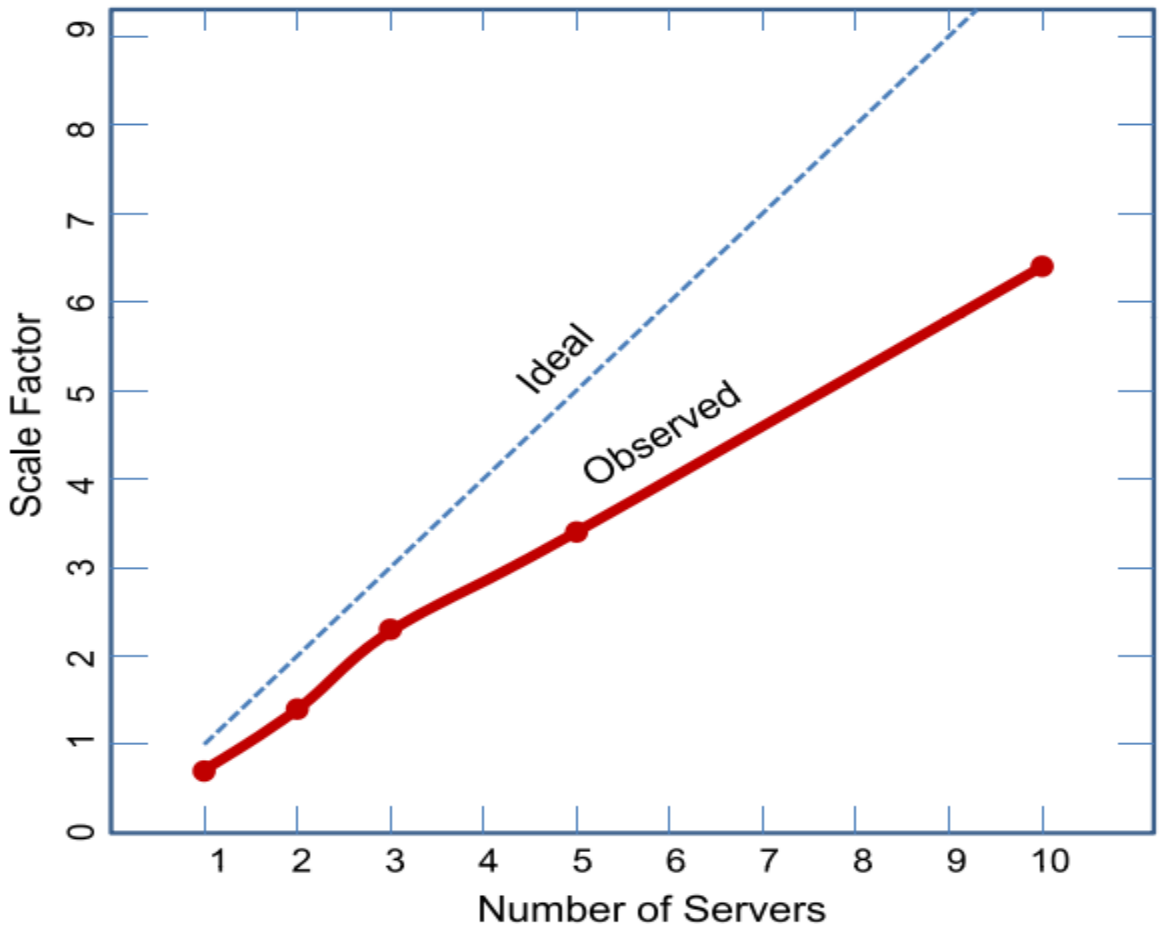
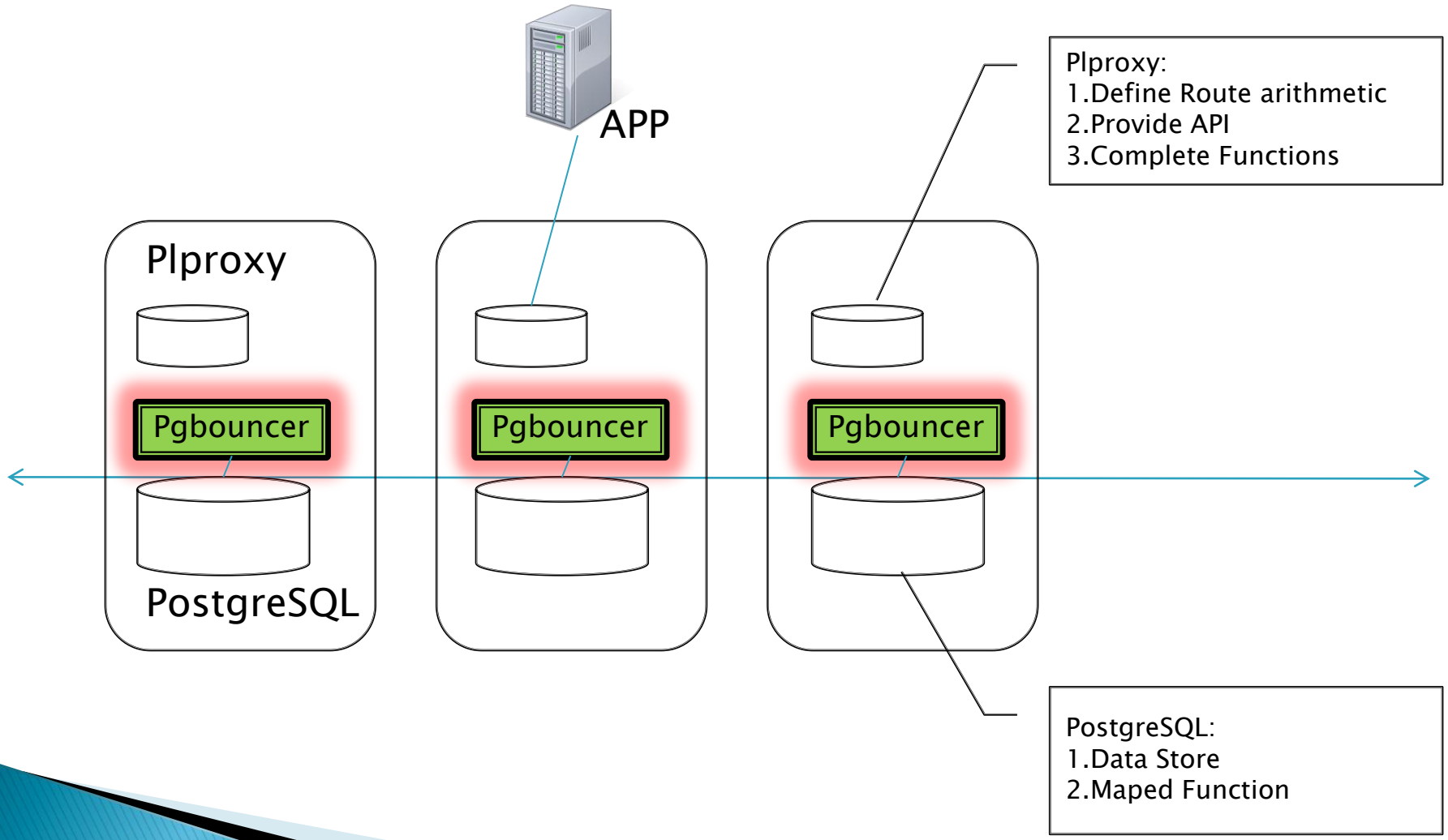


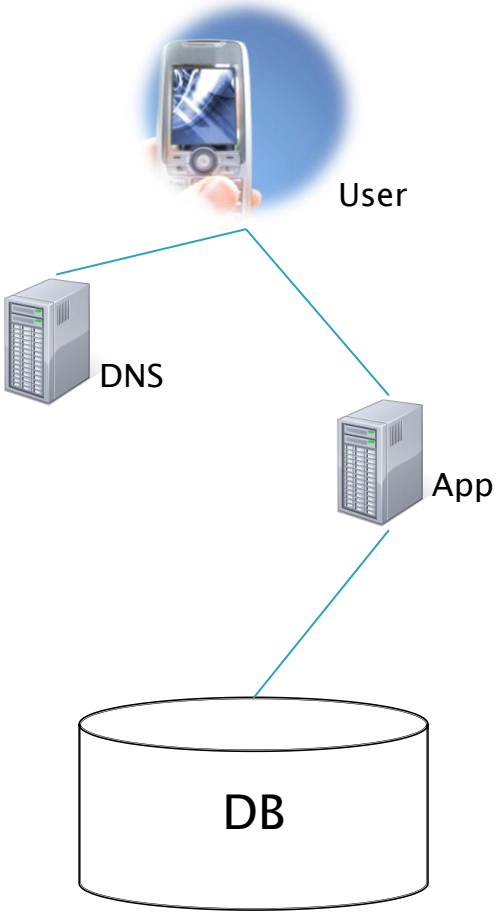

Figure 11: Postgres-XC Full Load Throughput

# Plproxy


APP
Plproxy
Pgbouncer
PostgreSQL
Pgbouncer
Pgbouncer
Plproxy:
1.Define Route arithmetic
2.Provide API
3.Complete Functions
PostgreSQL:
1.Data Store
2.Maped Function

# Plproxy Example

- Sample
  - Simulate login
- Object
  - Fastest
- Method
  - Shortest path



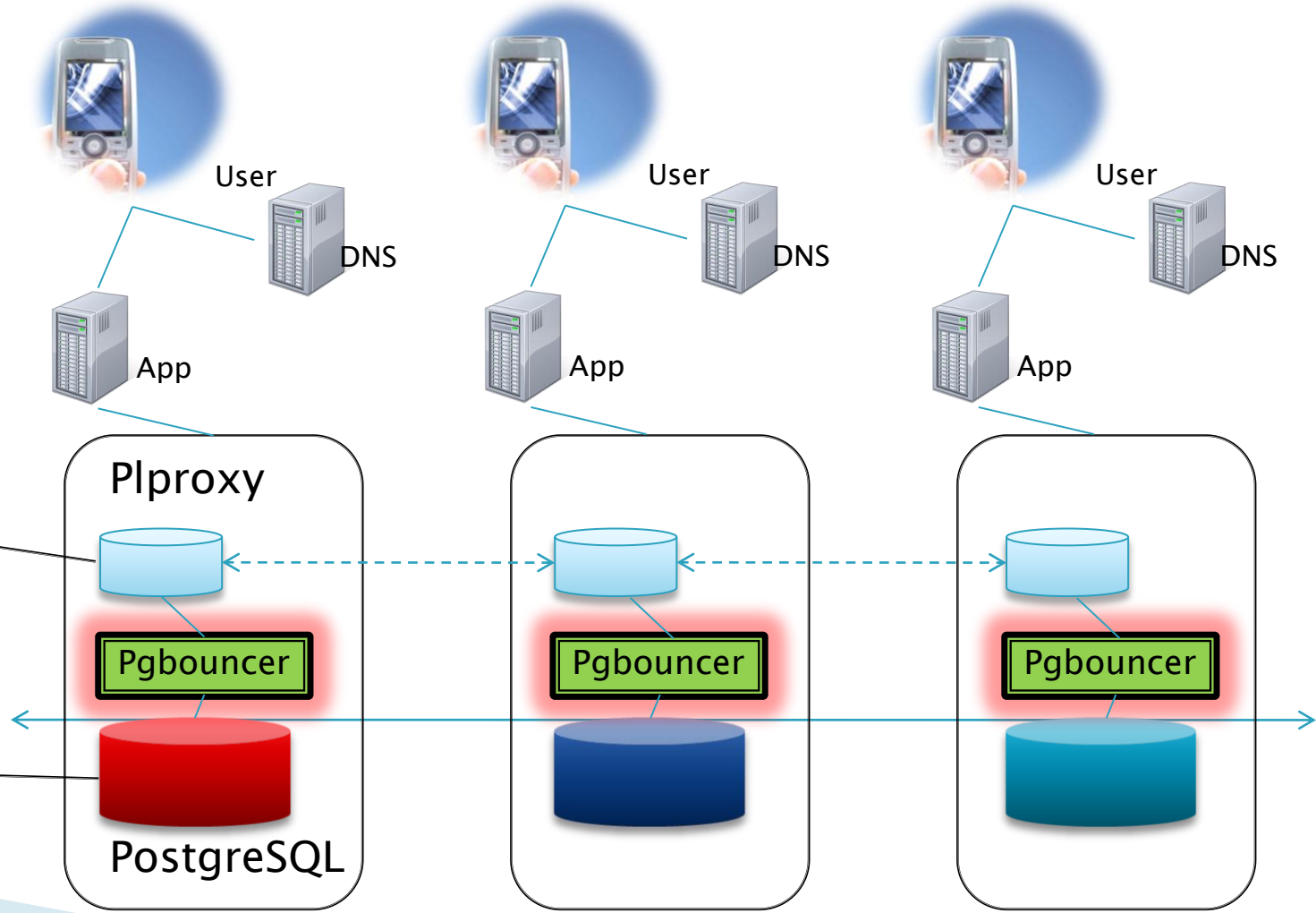
# Plproxy Example

User ID Aware:
Dns No
App Yes
DB Yes

Route Arithmetic:
1.Range
2.Hash
3.List ……

Plproxy:
1.Coordinator
2.Funs

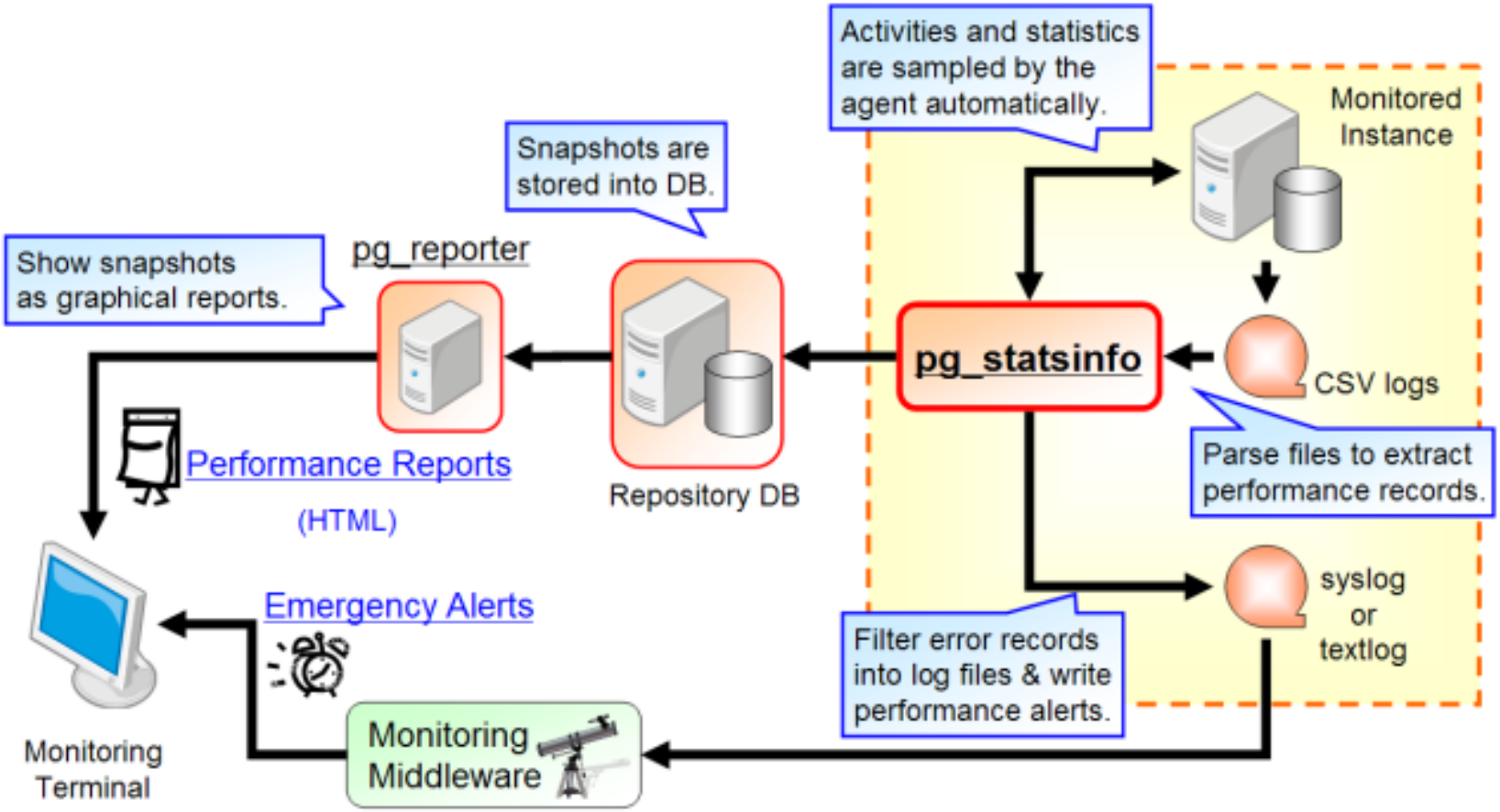
Data Node:
1.Mapped Funs
2.User Data



# PgMQ

- Embedded Message Client in PostgreSQL
- Support most major Message Systems
  - ActiveMQ
  - RabbitMQ
- Support event triger by time etc.
- Object
  - Final Consistent
- Support more than one Queues simultaneity
- Like per-row trigger on table

# Pg_statsinfo



# Pg_fincore

- Object
  - Keep good TPS when Server reboot
- PostgreSQL Buffer Cache
  - Shared Memory
  - Simple LRU List
  - Effective Cache Size
    - Mackert and lohman approximation
- OS Page Cache
  - Complex LRU List
    - Some piece of FIFO

# Pg_fincore

- Get stats per segment of table or index
- Restore the OS Page Cache state for a table or index
- Tools
  - mmap/mincore
  - posix_fadvise
- Impacts
  - more syscall
  - memory mapping

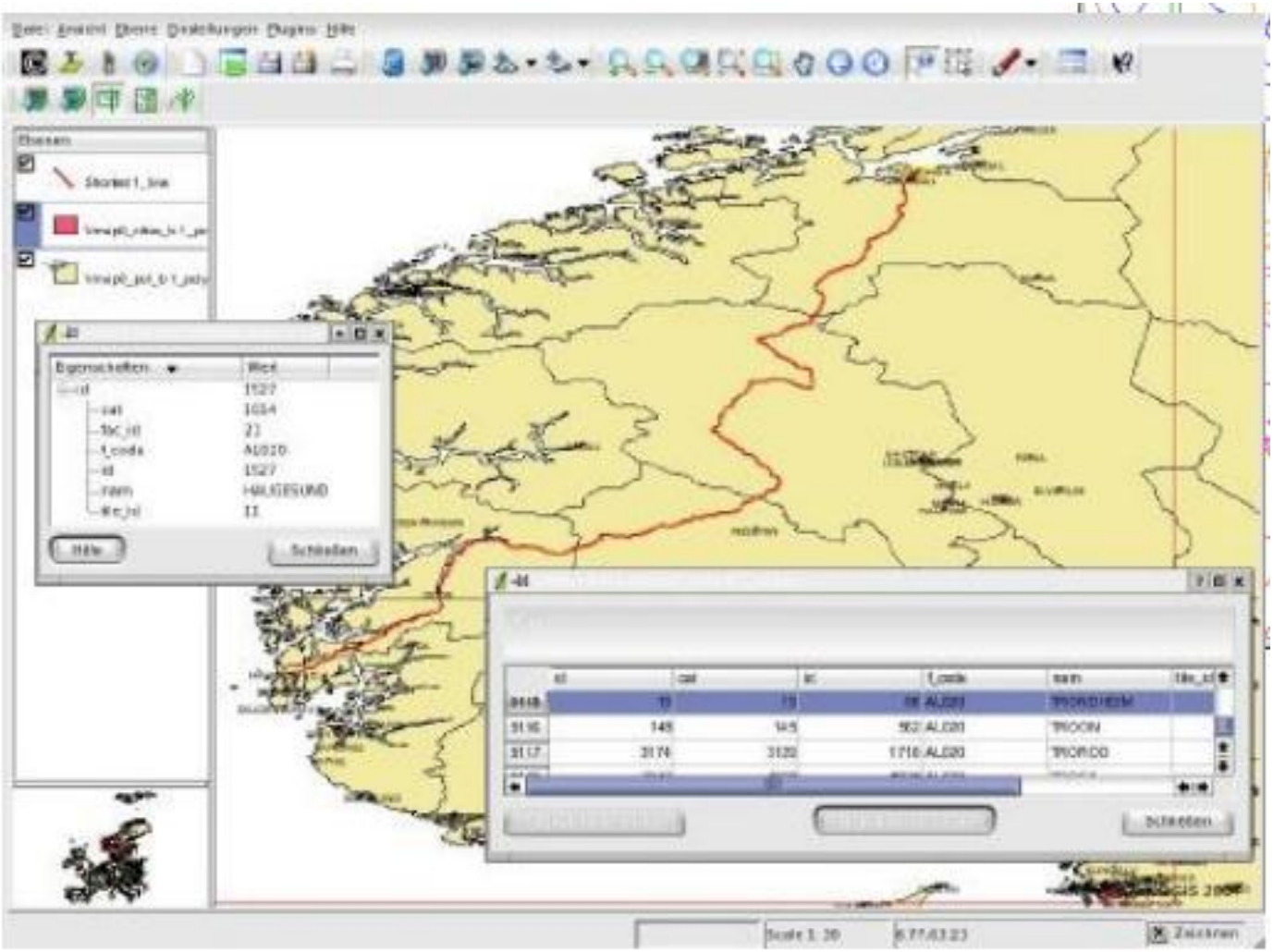
# Pg_fincore

- Functions for DBA
  - Debug
  - Pgsysconf()
  - Pgmincore(‘tablename’)
  - Pgfadv_WILLNEED(‘tablename’)
  - Pgfadv_DONTNEED(‘tablename’)
  - Pgmincore_snapshot(‘tablename’)
  - Pgfadv_willneed_snapshot(‘tablename’)
  - Pgfadv_nromal()
  - Pgfadv_sequential()
  - Pgfadv_random()

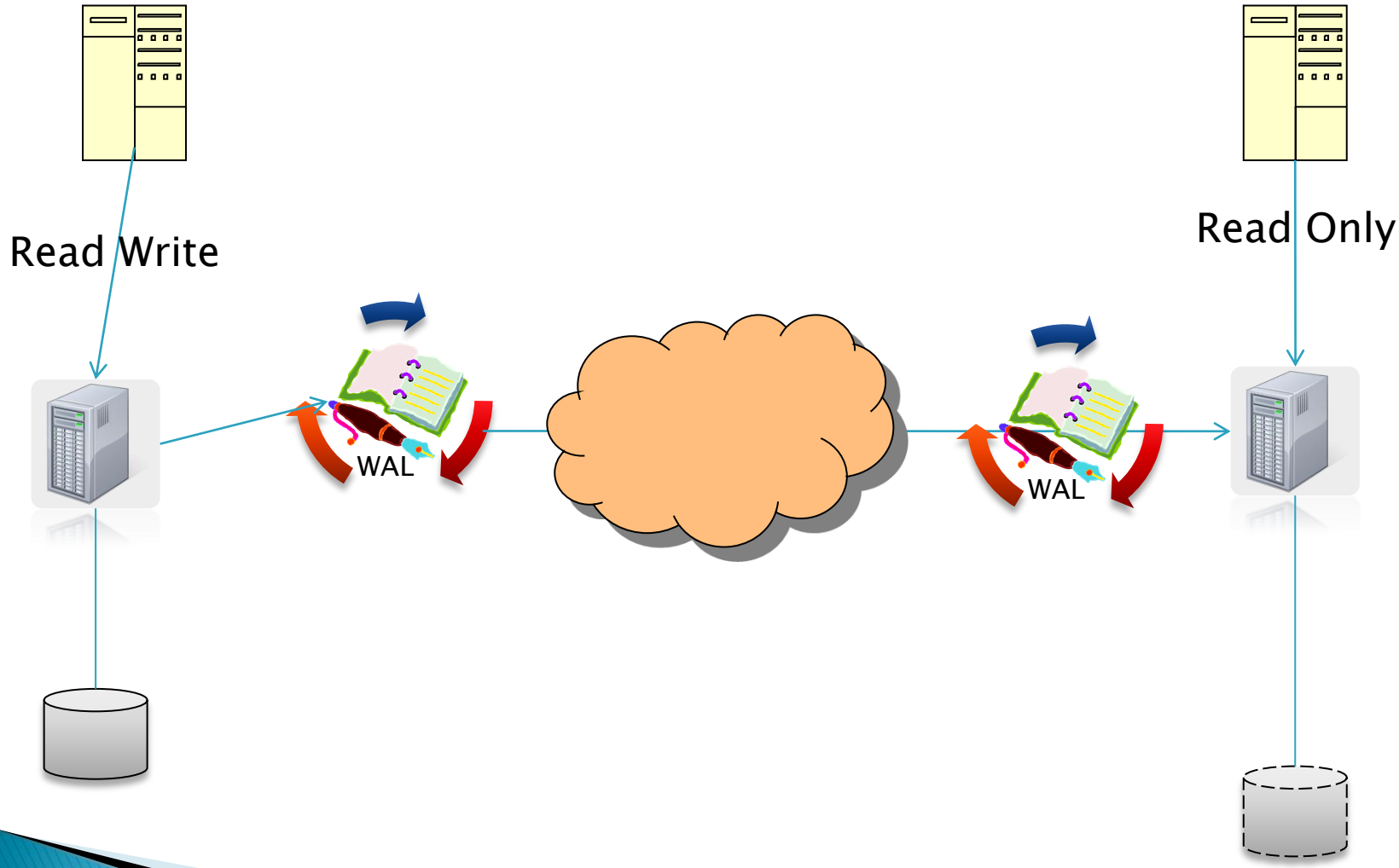
# Pg_fincore

- Usage case
  - Preload
  - Snapshot
  - Restore
  - Monitoring
  - Performance boost

# PostGIS


Eigenschaften
Wert
1527
1014
21
AL020
1527
HALGESUND
11
Hilfe
Schließen
TRONDHEIM
TRONDO
Schließen

# Active-Standby



# Not The End!