Rapid Upgrades With Pg_Migrator

BRUCE MOMJIAN,
ENTERPRISEDB

May, 2009

EnterpriseDB™

Abstract

Pg_Migrator allows migration between major releases of Postgres without a data dump/reload. This presentation explains how pg_migrator operates.

http://momjian.us/presentations

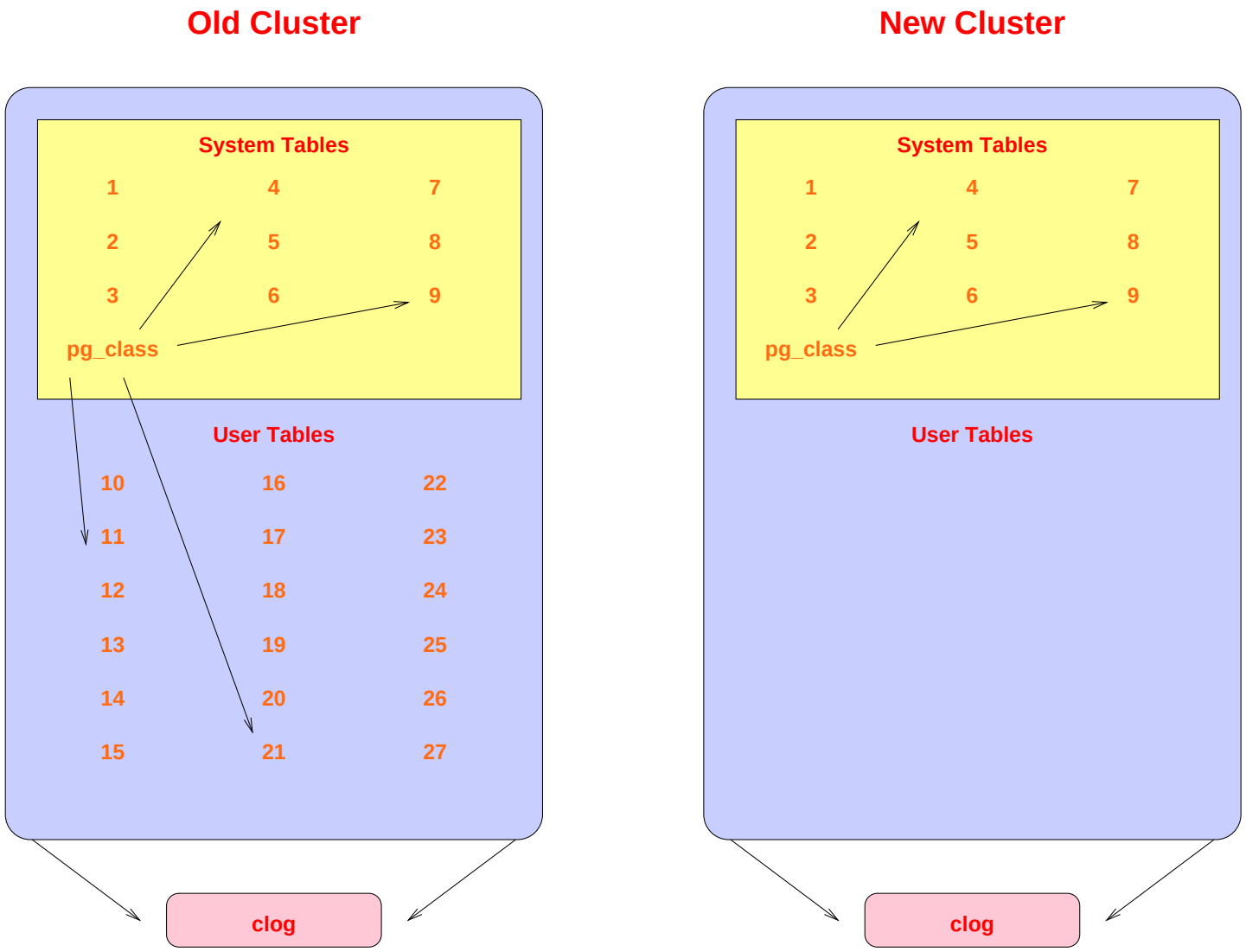
Why Pg_Migrator

- Very fast upgrades
- Optionally no additional disk space

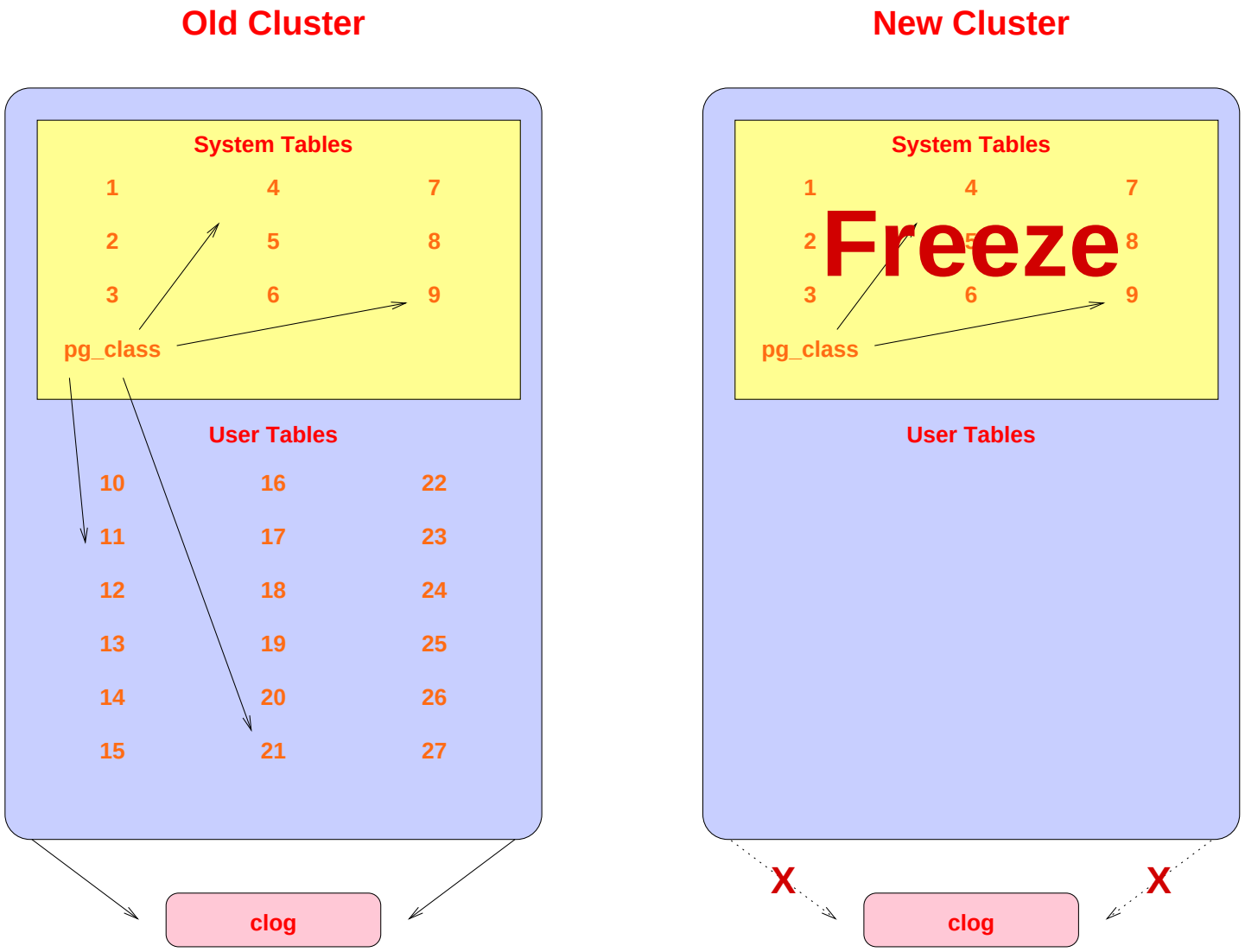
Other Upgrade Options

- dump/restore
- Slony

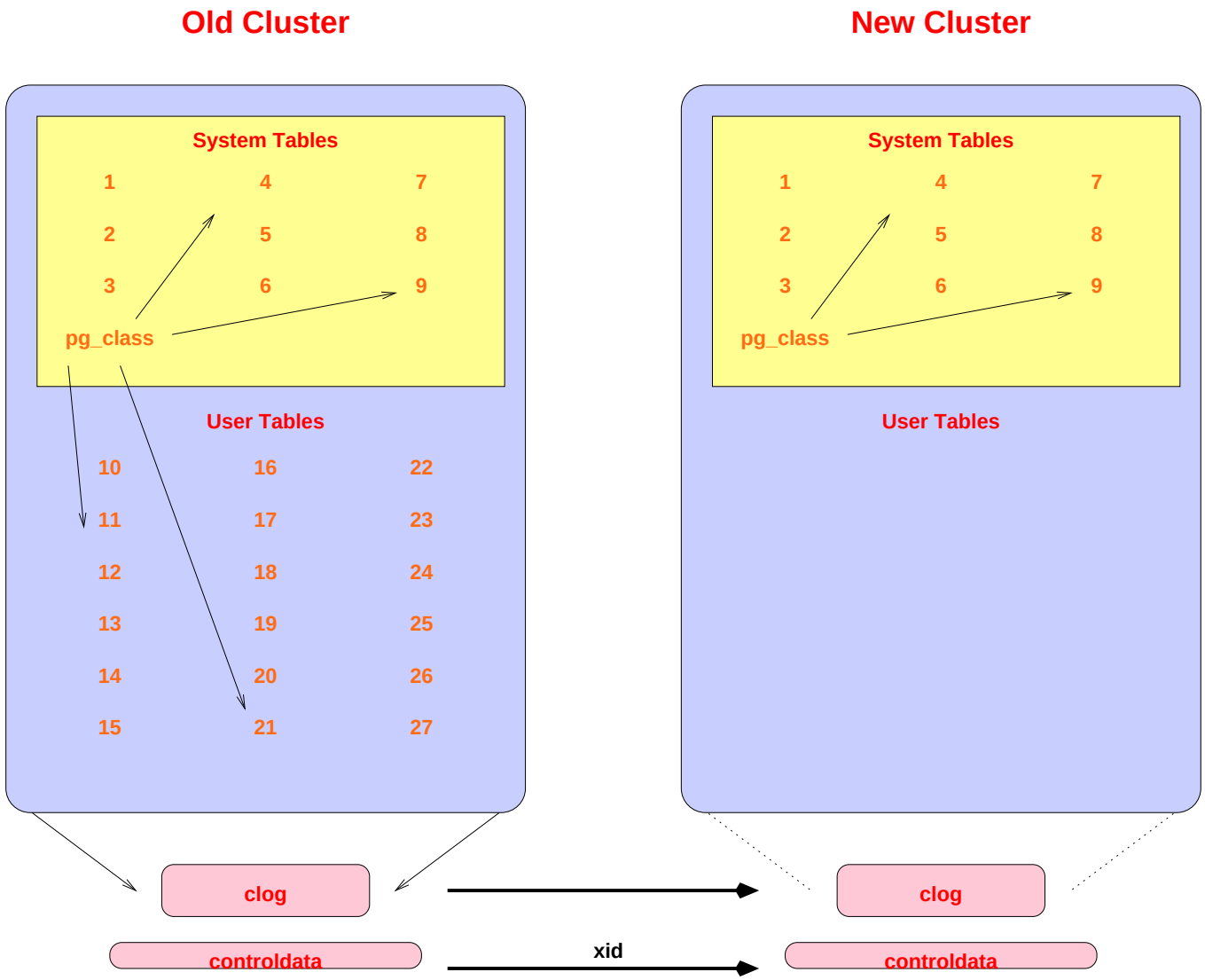
How It Works: Initial Setup



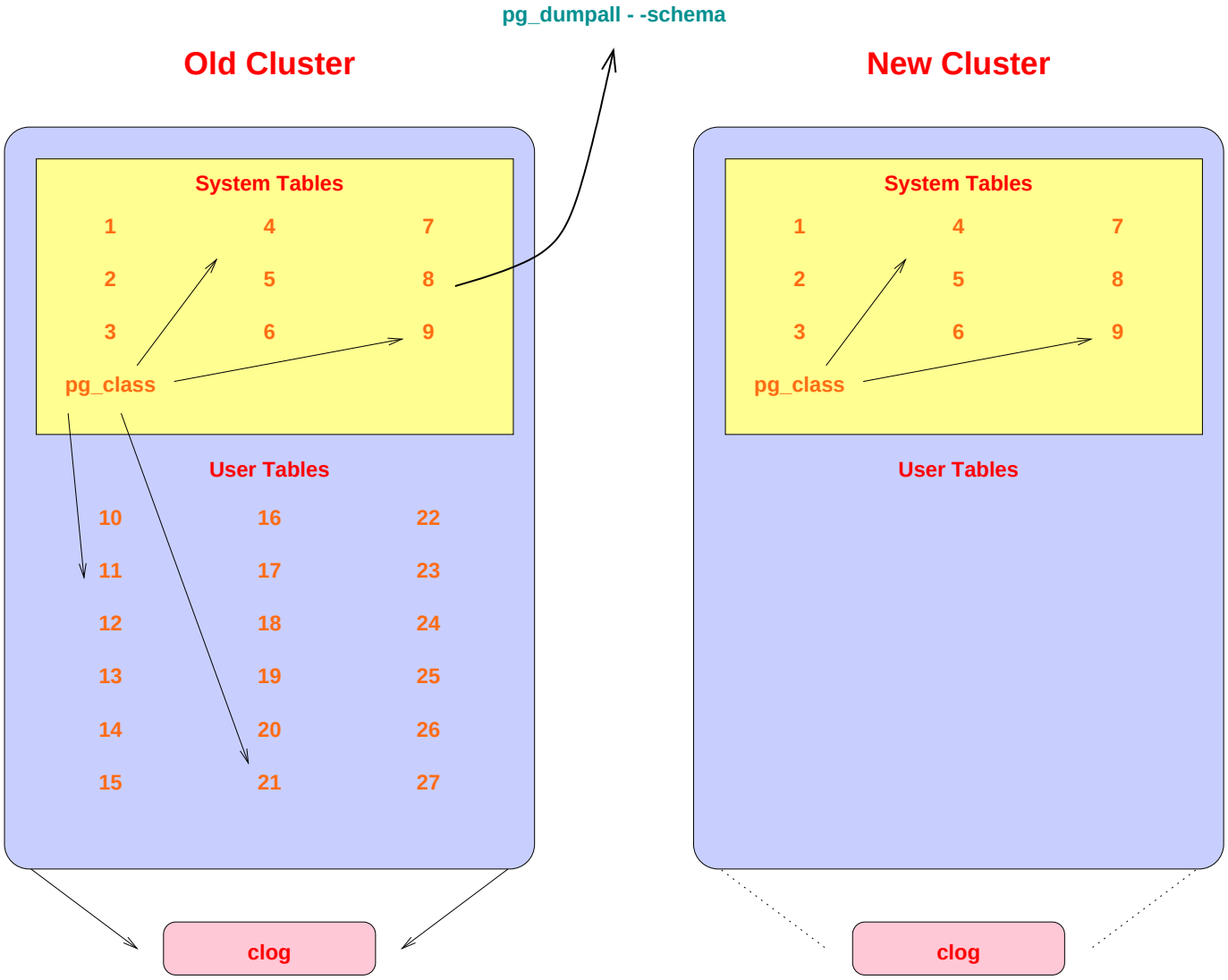
Decouple New Clog Via Freezing



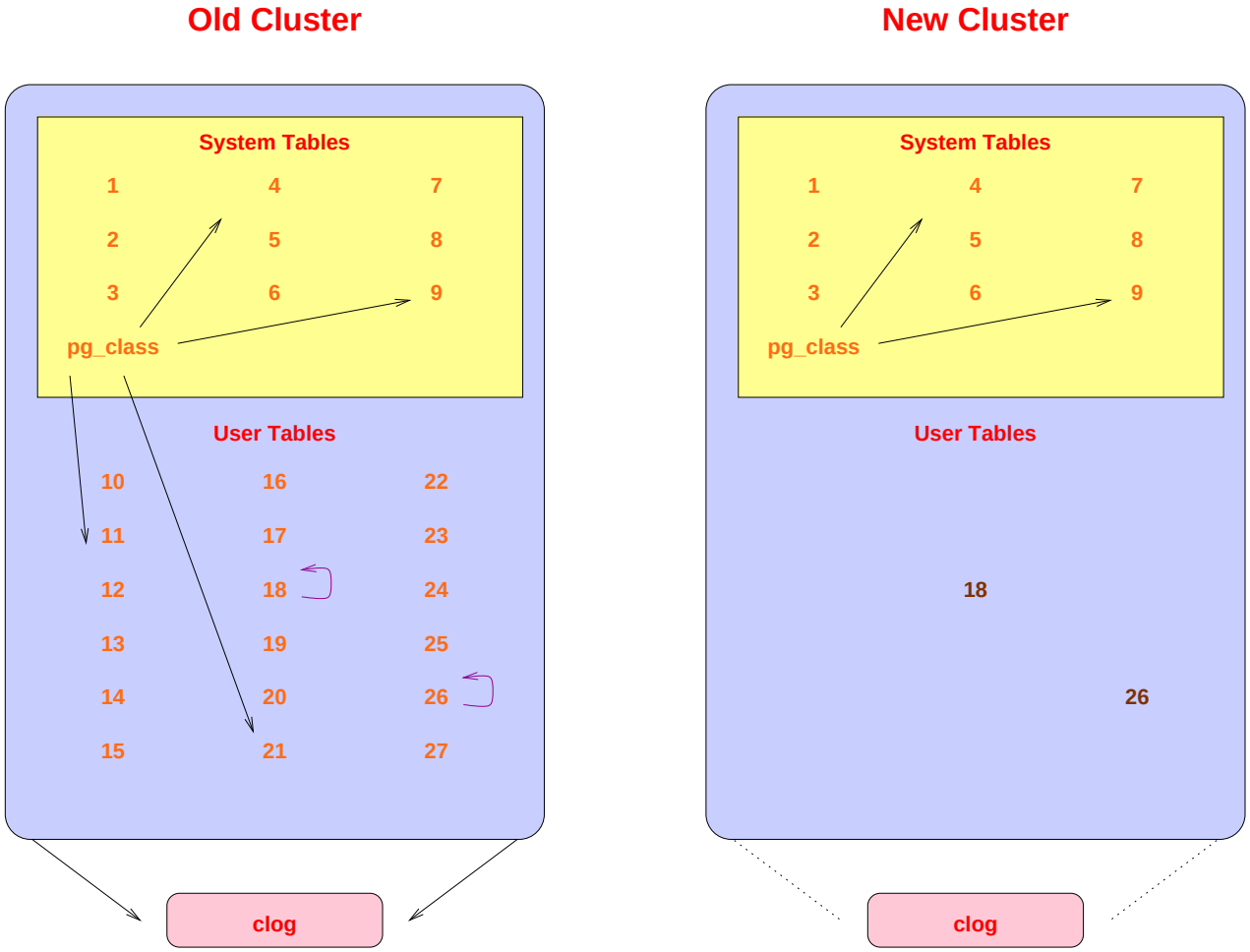
Transfer Clog and XID



Get Schema Dump

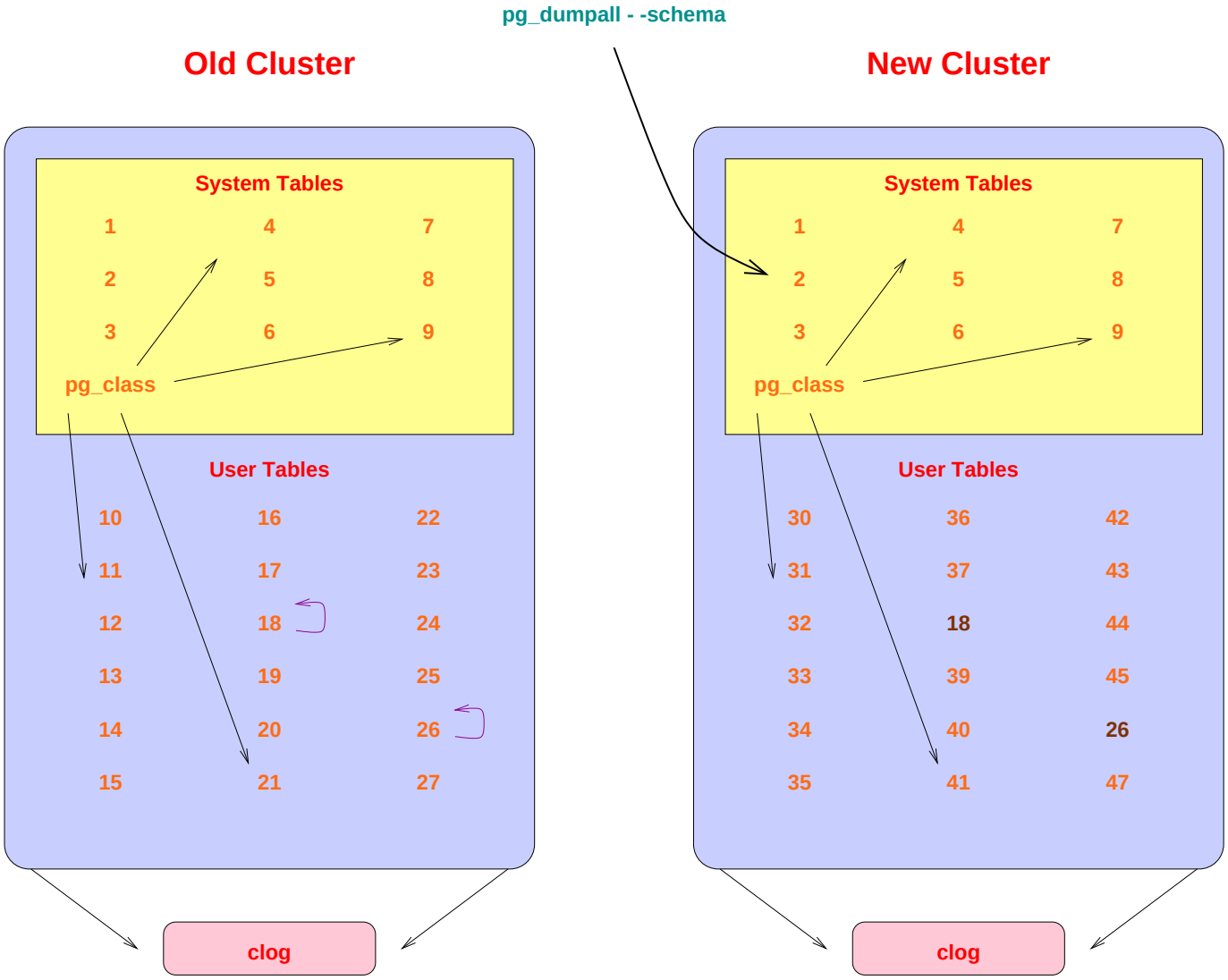


Reserve TOAST OIDs Using Relfilenodes

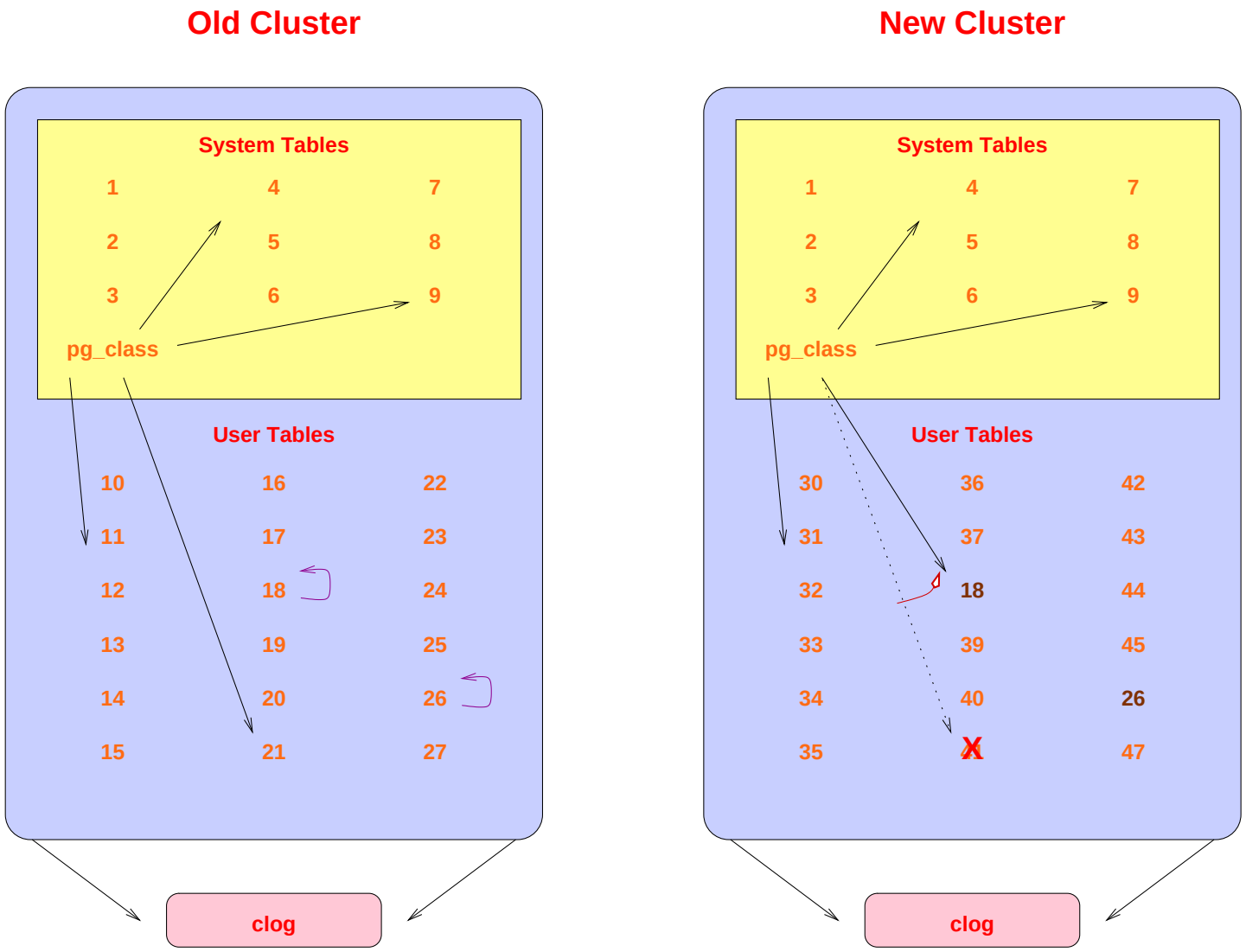


This is necessary because heap references to TOAST tables contain the TOAST oids for easy lookup.

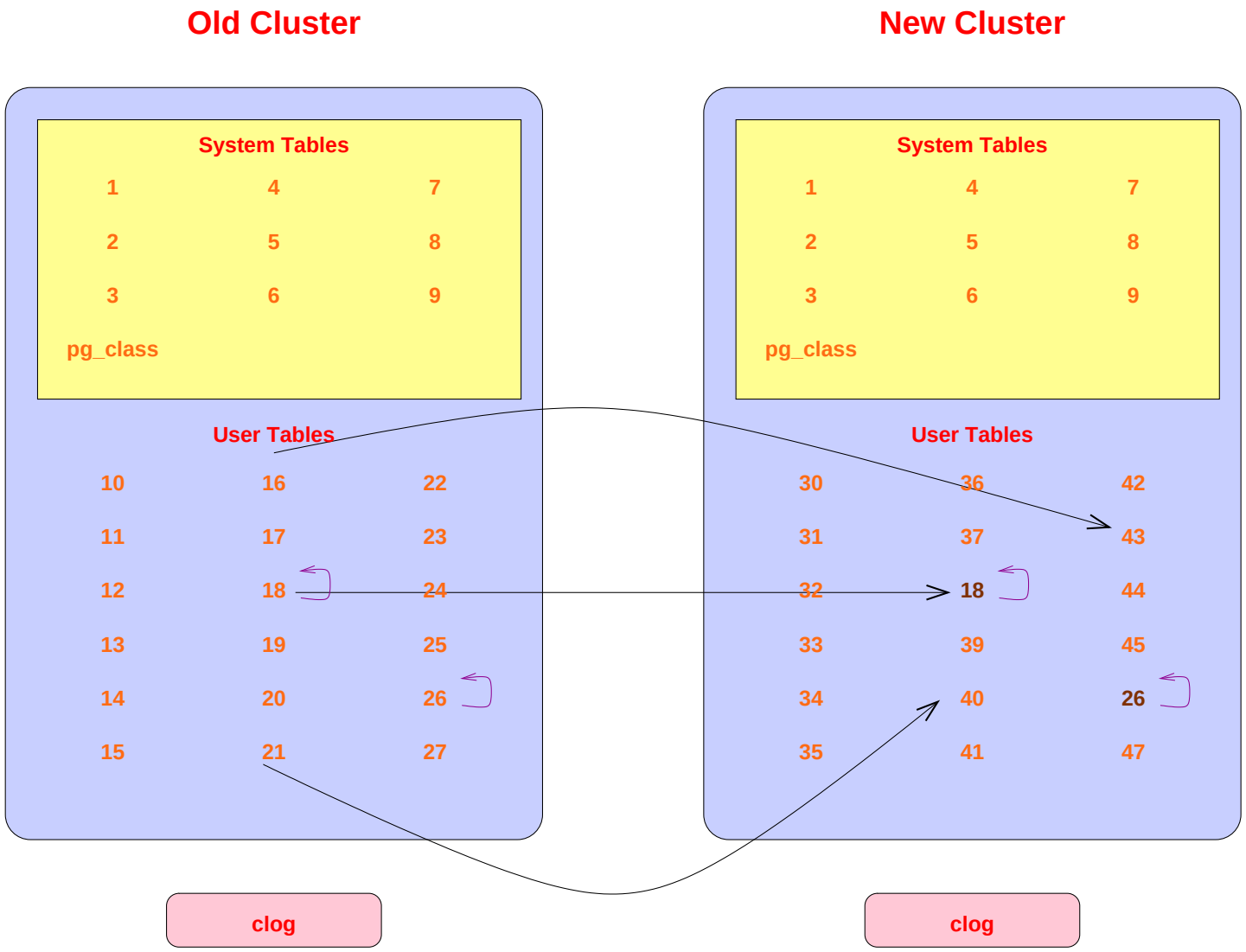
Restore Schema In New Cluster



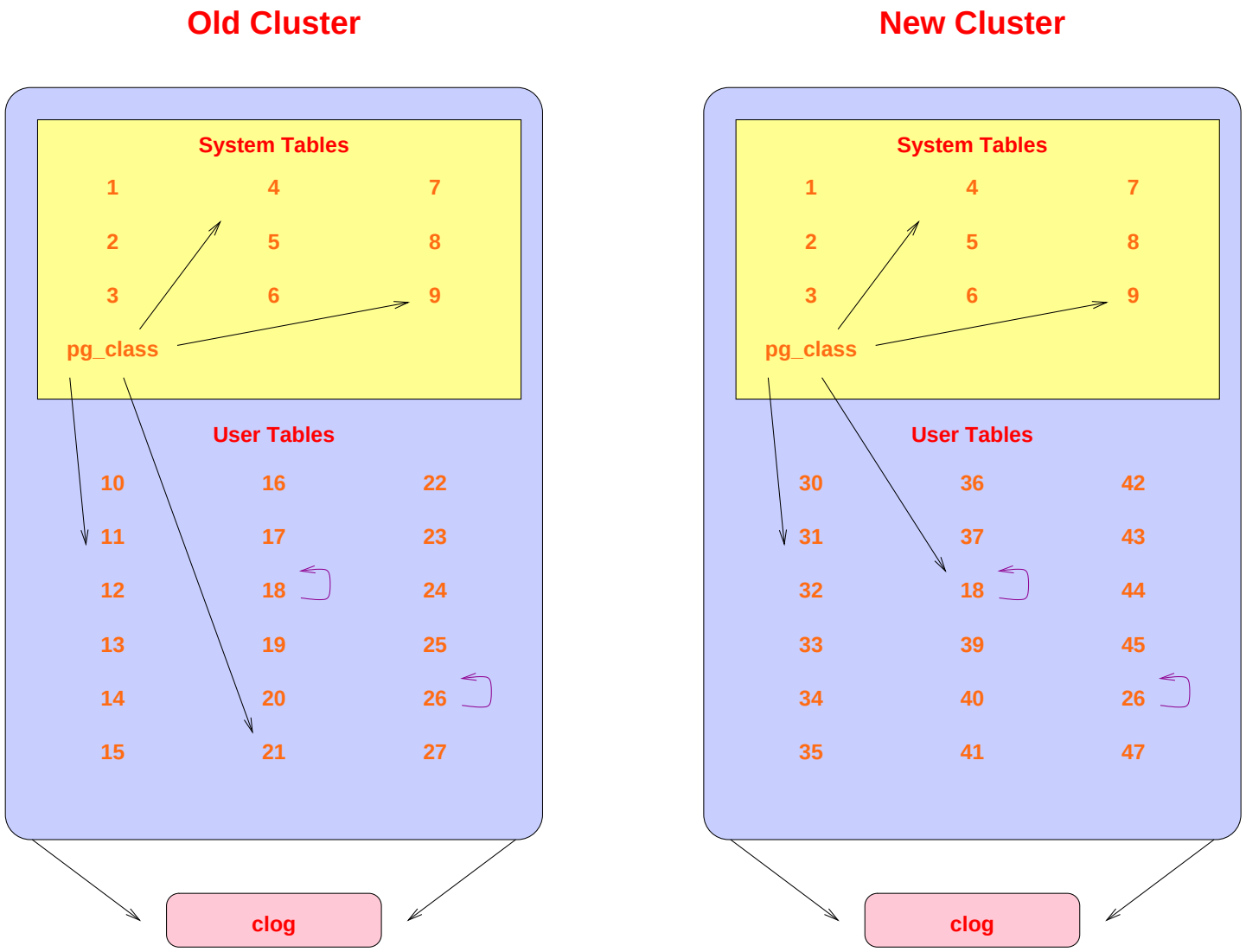
Connect TOAST Placeholders To the Proper Relations



Copy User Heap/Index Files



Complete


Old Cluster
System Tables
1
4
7
2
5
8
3
6
9
pg_class
User Tables
10
16
22
11
17
23
12
18
24
13
19
25
14
20
26
15
21
27
clog
New Cluster
System Tables
1
4
7
2
5
8
3
6
9
pg_class
User Tables
30
36
42
31
37
43
32
18
44
33
39
45
34
40
26
35
41
47
clog

How It Works: In Detail

- Check for cluster compatability
  - locale
  - encoding
  - integer datetimes (default changed from 8.3 -> 8.4)
- Use pg_dumpall to dump old cluster schema (no data)
- Freeze all new cluster rows (remove reference to clog entries)
- Rename tablespaces to *_old
- New cluster uses old xid counter value (see freeze above)
  - Set system table frozen xids to match the current xid
- Create new users/databases

- Collect cluster information
- Install support functions that call internal backend functions
- Create placeholder files to reserve relfilenode file names
- Create schema in new cluster
- Adjust new cluster to use reserved relfilenode names
  - Delete placeholder toast relfilenode files
  - Remove new cluster toast tables
  - Create new cluster toast table using reserved relfilenode
  - Assign new toast tables with proper relfilenodes to relations
- Copy or link files from old cluster to new cluster
  - Toast tables have the same relfilenodes as in the old cluster
- Warn about any remaining issues, like REINDEX requirements

Sample Run: Validation 1

```
Performing consistency checks
-----------------------------
Checking old data directory /u/pgsql.old/data
   checking base                                             ok
   checking global                                           ok
   checking pg_clog                                          ok
   checking pg_multixact                                     ok
   checking pg_subtrans                                      ok
   checking pg_tblspc                                        ok
   checking pg_twophase                                      ok
   checking pg_xlog                                          ok
Checking new data directory /u/pgsql/data
   checking base                                             ok
   checking global                                           ok
   checking pg_clog                                          ok
   checking pg_multixact                                     ok
   checking pg_subtrans                                      ok
   checking pg_tblspc                                        ok
   checking pg_twophase                                      ok
   checking pg_xlog                                          ok
Checking binaries in old cluster (/u/pgsql.old/bin)
   checking postgres                                         ok
   checking pg_ctl                                           ok
   checking pg_dumpall                                       ok
```

Sample Run: Validation 2

```
Checking binaries in new cluster (/u/pgsql/bin)
   checking postgres                                        ok
   checking pg_ctl                                          ok
   checking pg_dumpall                                      ok
   checking psql                                            ok
Starting postmaster to service old cluster
   waiting for postmaster to start                          ok
Creating catalog dump                                       ok
Splitting old dump file                                     ok
Checking for invalid 'name' user columns                    ok
Stopping postmaster servicing old cluster                   ok
Starting postmaster to service new cluster
   waiting for postmaster to start                          ok
Stopping postmaster servicing new cluster                   ok

*Checks complete*


| If pg_migrator fails after this point, you must
| re-initdb the new cluster before continuing.
| You will also need to rename your old tablespace
| directories to remove the ".old" suffix before
| continuing.
```

Sample Run: Migration

```
Performing migration
--------------------
Starting postmaster to service new cluster
   waiting for postmaster to start                          ok
Analyzing all rows on the new cluster                       ok
Freezing all rows on the new cluster                        ok
Stopping postmaster servicing new cluster                   ok
Renaming tablespaces to *.old                               ok
Deleting old commit clogs                                   ok
Copying commit clogs                                        ok
Setting next transaction id for new cluster                 ok
Resetting WAL archives                                      ok
Starting postmaster to service new cluster
   waiting for postmaster to start                          ok
Setting frozenxid counters in new cluster                   ok
Creating databases in new cluster                           ok
Adding support functions to new cluster                     ok
Creating placeholder relfiles for toast relations           ok
Restoring database schema                                   ok
Restoring relations to use fixed toast file names           ok
Restoring user relations                                    ok
Stopping postmaster servicing new cluster                   ok
Setting next oid for new cluster                            ok
*Upgrade complete*
| Optimizer statistics and free space information is not transfered by
| pg_migrator, so consider running 'vacuumdb --all --analyze'
| on the newly-upgraded database.
```

Possible Post-8.4 Data Format Changes

- clog
- heap page format
- page header, include bitmask
- tuple header, including bitmask
- data value format
- index page format

Migration Timings

Migration Method	Minutes
dump/restore	300.0
dump with parallel restore	180.0
pg_migrator in copy mode	44.0
pg_migrator in link mode	0.7

Database size: 150GB, 850 tables

The last duration is 44 *seconds.*

Timings courtesy of
Stefan Kaltenbrunner
(mastermind on IRC)

Conclusion

PG 8.3

PG 8.4