Bi-Direction Replication

P2D2 2015

Petr Jelínek, 2ndQuadrant
(petr@2ndQuadrant.com)

BDR

- Bi-Directional Replication
  - Je možné zapisovat na všech serverech
- Asynchronní
  - Nízká latence (zápisu)
  - Tolerance ke krátkodobým výpadkům
- DDL je replikováno transparentně
- Mohou nastat konflikty, celý cluster je eventuálně konzistentní

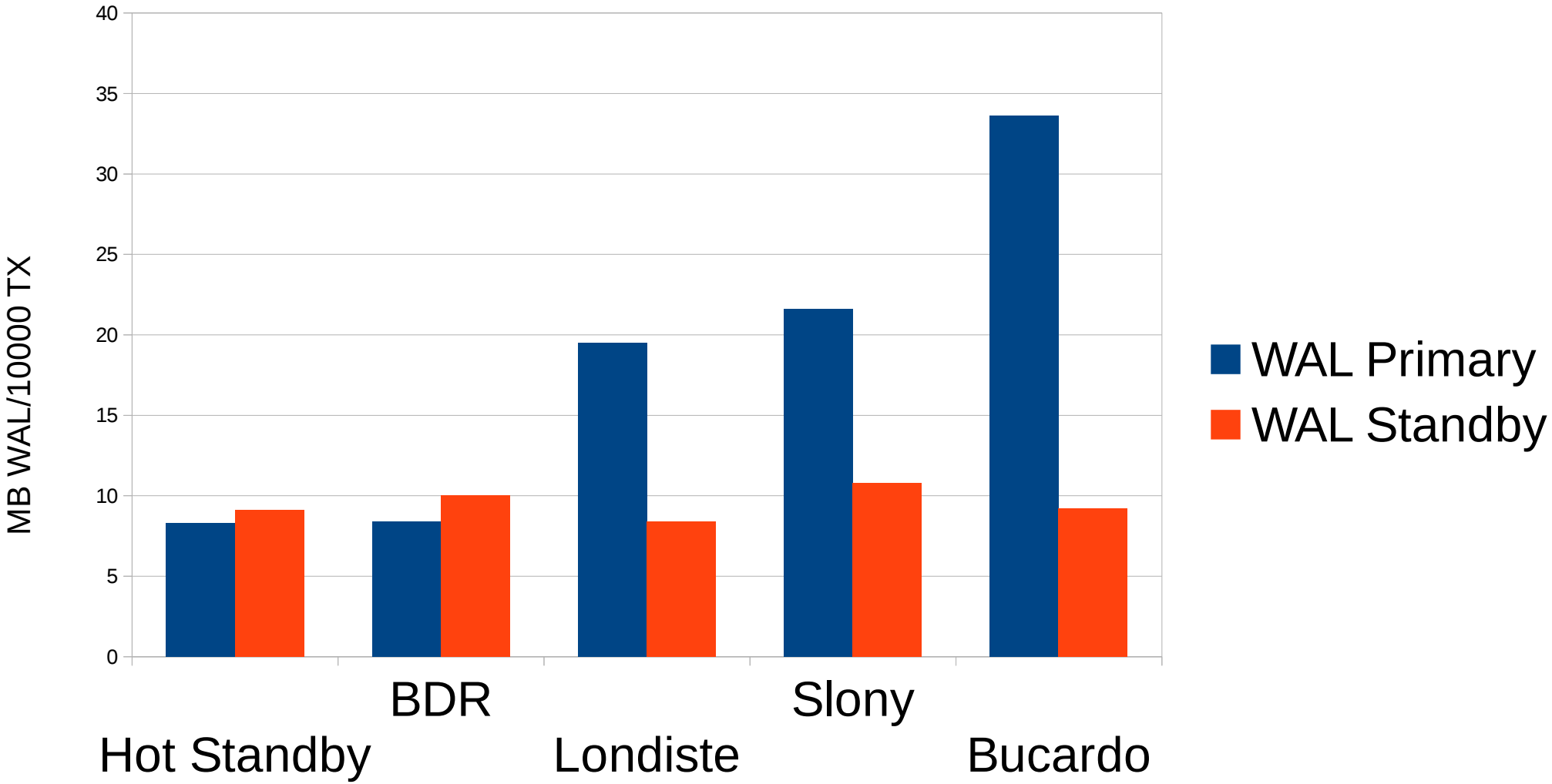
Základní stavební kameny BDR

- Relativně snadné nastavení
- Stabilita
- Malý overhead
- Snadný failover/failback – stačí se prostě připojit na jiný server
- Geograficky distribuované databáze
- Upgrade major verzí bez výpadku

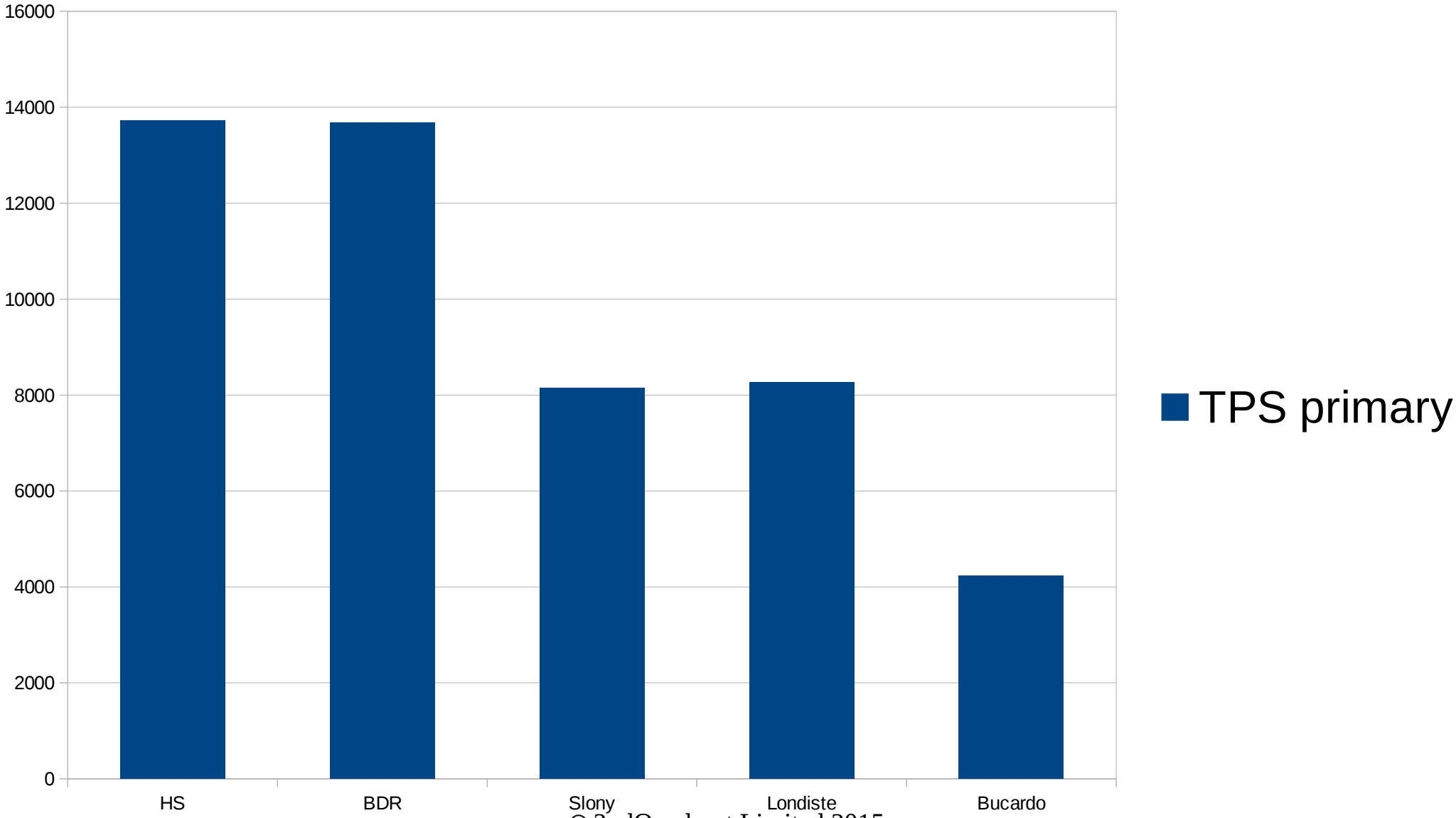
Výkon

- 2 x i2.4xlarge
  - Intel Xeon E5-2670 v2
  - 122GB RAM
  - 4x800GB SSDs Raid 0
- Minimální změny v konfiguraci PostgreSQL
- pgbench scale 50, -c 32 -j 32

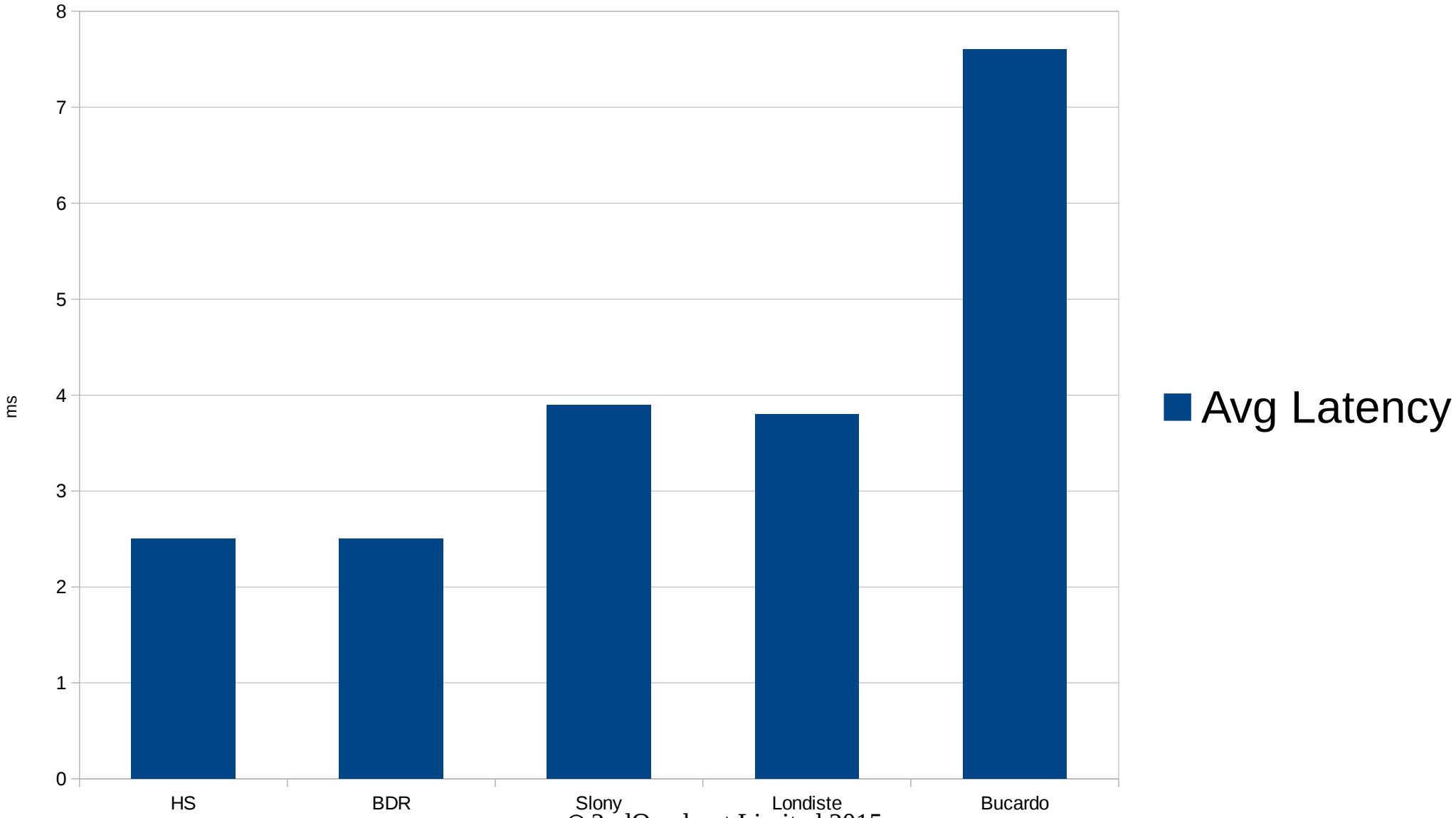
Objem WAL



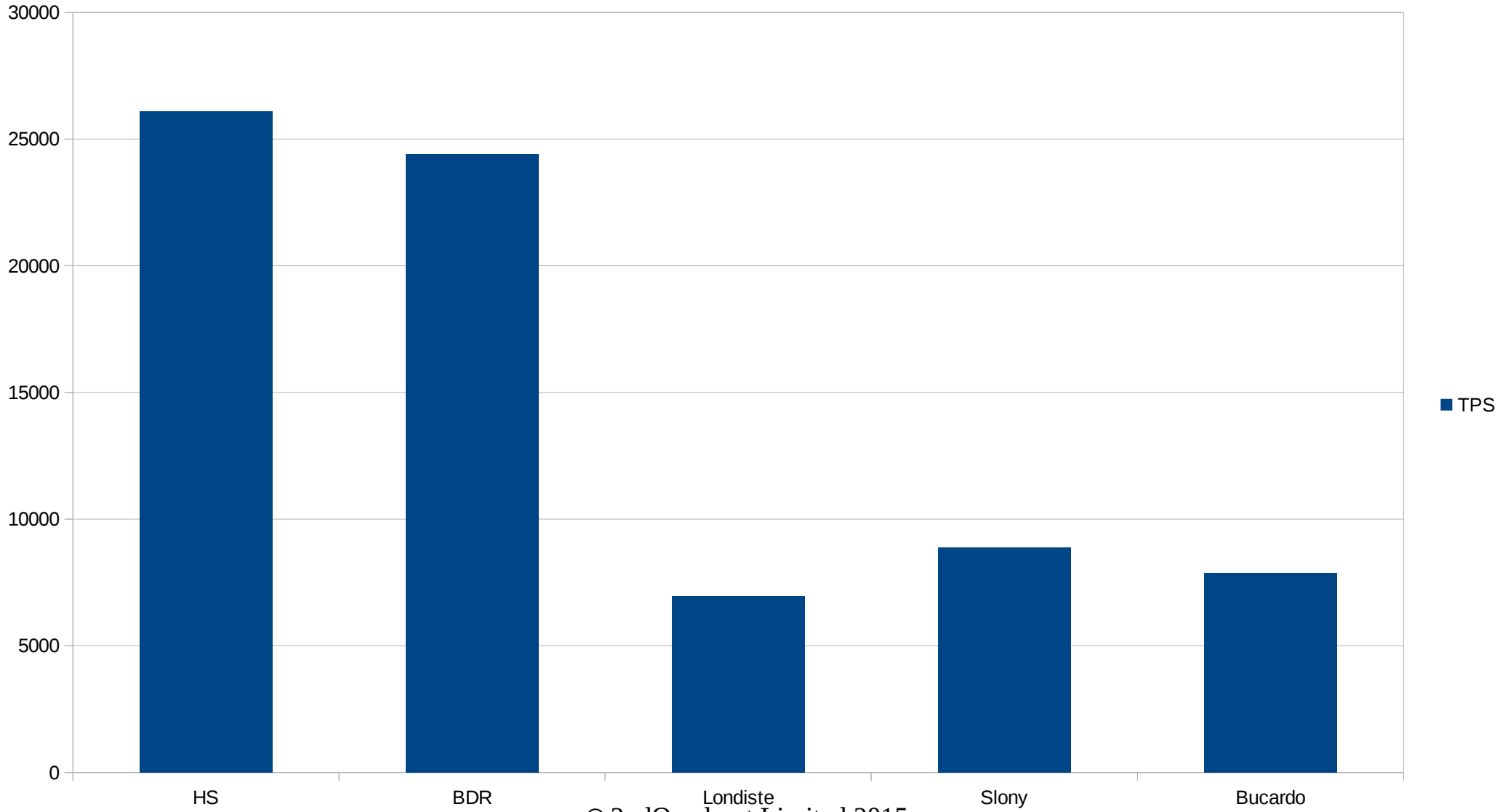
Propustnost transakcí



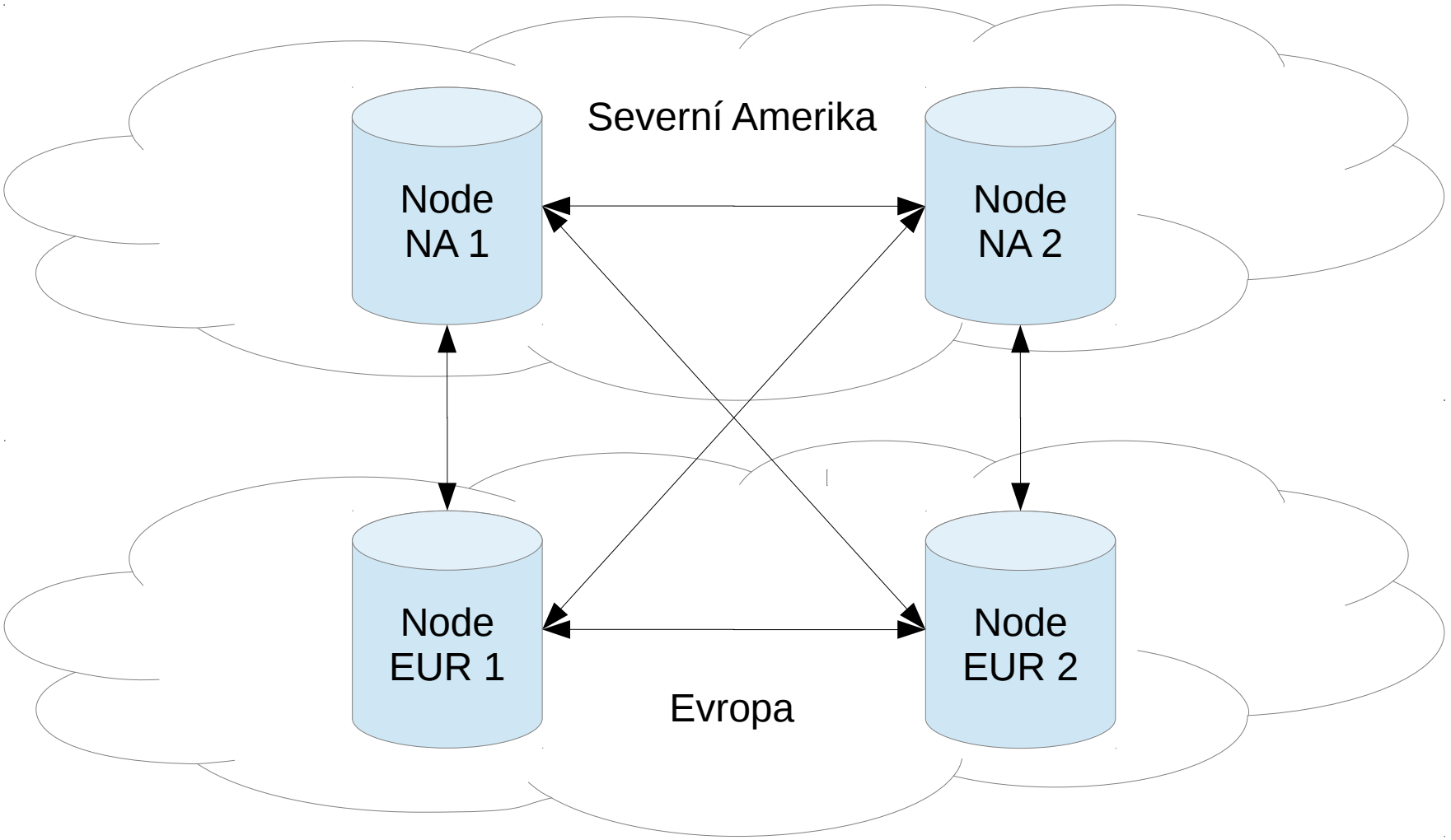
Latence transakcí



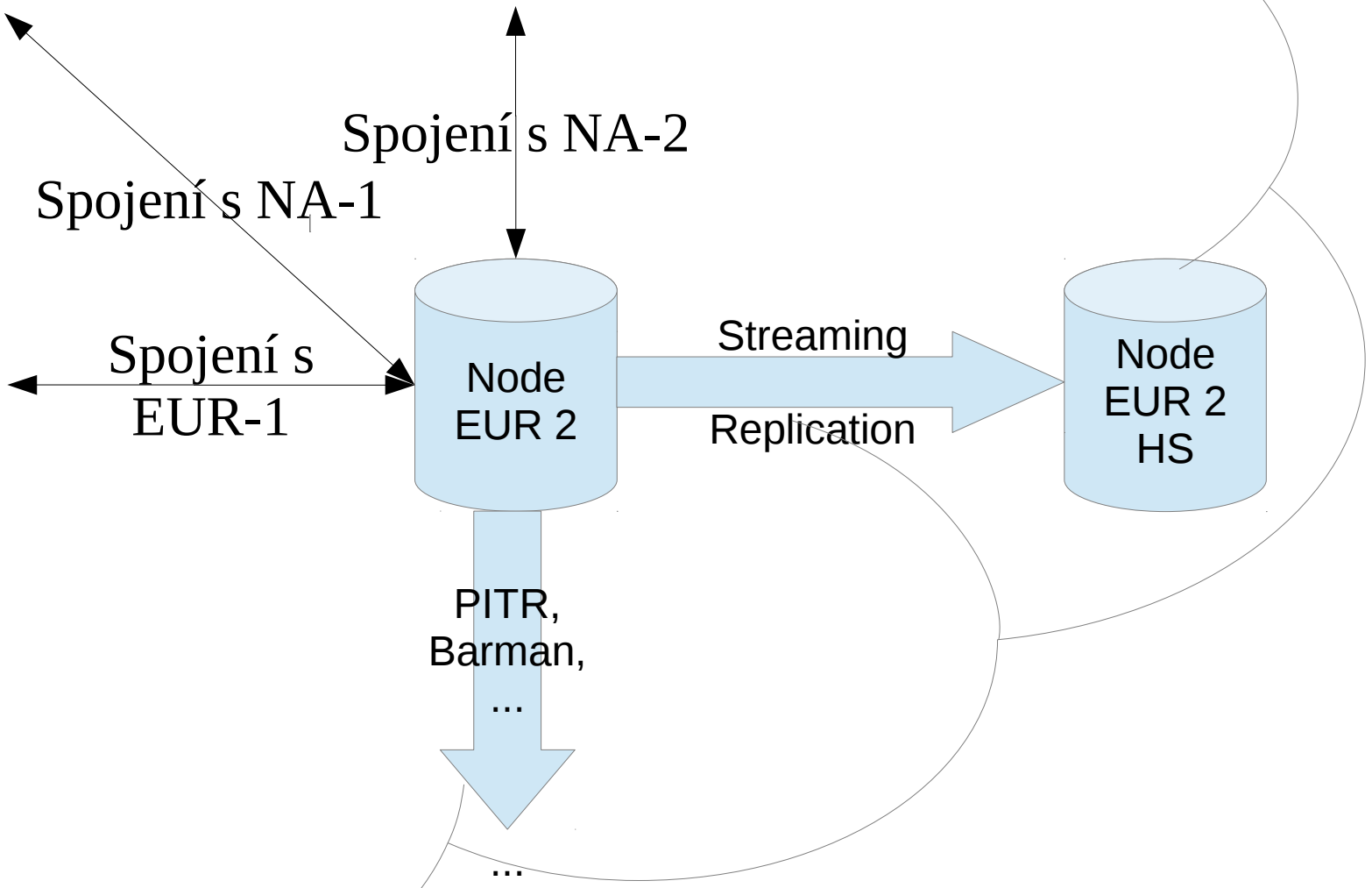
TPS se zpožděním replikace maximálně 2s



Příklad



Příklad - detail



Základní konfigurace

- Server1:
  - `CREATE EXTENSION bdr;`
  - `bdr.bdr_group_create('hostname=host1 dbname=db');`
- Server2..N:
  - `CREATE EXTENSION bdr;`
  - `bdr.bdr_group_join('hostname=host2 dbname=db', 'hostname=host1 dbname=db');`

Selektivní replikace

- Každá tabulka může být v jednom či více replikačních setech:
  - `SELECT bdr.table_set_replication_sets('tbl_ab', '{replicate-to-a,replicate-to-b}');`
  - `SELECT bdr.table_get_replication_sets('tbl_ab');`

```
-------------------------------

 {replicate-to-a,replicate-to-b,all}
```

- BDR obsahuje předdefinované sety 'all' a 'default'

Selektivní replikace

- Každý server může mít jiný seznam setů které ho zajímají:
  - `bdr.bdr_group_join('hostname=host2 dbname=db', 'hostname=host1 dbname=db', replication_sets := '{replicate-to-a,replicate-to-b}');`
- Výsledkem je UNION tabulek v daných setech

Prevence konfliktů

- Globální sekvence
  - Každý server dostane přidělenou množinu hodnot
- Zápis replikovaných dat používá UPSERT logiku

Možnosti řešení konfliktů

- Vestavěná funkce „poslední změna vyhrává“
- Uživatelsky definované funkce pro řešení konfliktů
  - Jakýkoli procedurální jazyk
  - Různé druhy konfliktů
  - Různé možnosti řešení: Skip, Keep Local, Keep Remote, Returned Tuple, Error
- Možnost zapisovat konflikty včetně dat do tabulky

DDL replikace

- Transparentní, stačí prostě spustit DDL
- DDL manipulující s globálními objekty serveru nejsou replikovány
  - `CREATE/ALTER/DROP DATABASE/USER/TABLESPACE`
- Některé příkazy nejsou povolené
  - `ALTER TABLE … ALTER COLUMN … USING();`
  - `ALTER COLUMN … ADD COLUMN … DEFAULT;`
- GRANT/REVOKE je aplikován pouze na lokální server (není replikován)

Další užitečné vlastnosti

- Synchronní replikace (stejný princip jako u HS)
- Servery mohou být inicializovány pomocí pg_basebackup
- Servery mohou být inicializovány pomocí paralelního pg_dump/restore
- Zpožděná replikace

Vhodné použití BDR

- Většina dat je modifikována na jednom serveru
- Aplikace je připravena na konflikty a jejich řešení
- INSERT-only databáze
- Geograficky distribuovaná databáze

Nevhodné použití BDR

- Nutnost okamžité a absolutní konzistence dat
- Aplikace, které nebyly napsané pro multi-master
- BDR není řešení pro neomezené škálování zápisu

Změny v PostgreSQL pro BDR

- Patche pro PostgreSQL 9.3, 9.4:
  - Background Workers (9.3, 9.4)
  - Event Triggers (9.3)
  - Logical Decoding (9.4)
- Vlastní patche nad PostgreSQL 9.4:
  - DDL deparse
  - Sequence AM
  - Replication Identifiers, Wal Messages, Commit Timestamps

Patche?!?

- Všechny vlastní patche jsou v procesu začlenění do Postgresu
- Všechny patche mají PostgreSQL licenci
- BDR PostgreSQL vychází společně s oficiálním PostgreSQL
- Dlouhodobě neplánujeme udržovat fork, ale chceme dostat všechny změny do PostgreSQL

Bez patchů - UDR

- Uni-Directional Replication
- Extenze pro 9.4, nevyžaduje žádné další patche
- Rychlejší náhrada za řešení jako londiste/Slony
- Replikace z více serverů na jeden server

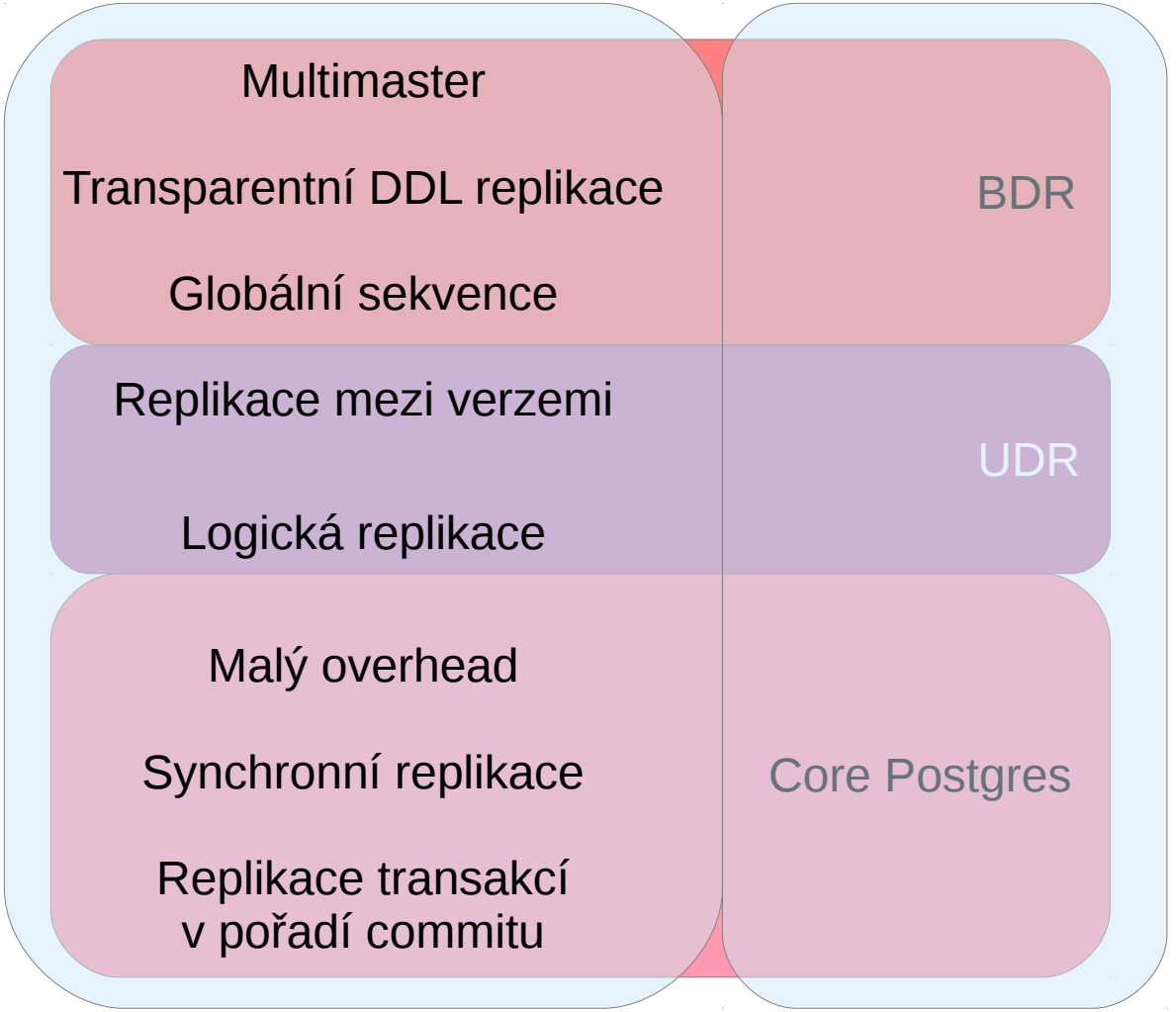
UDR

- Umí vše co BDR kromě:
  - Nepodporuje multi-master
  - Nepodporuje transparentní DDL replikaci
    - bdr.bdr_replicate_ddl_command
  - Nemá globální sekvence
  - Mírně větší overhead na slave serveru

UDR - konfigurace

- Master:
  - `CREATE EXTENSION bdr;`
- Standby:
  - `CREATE EXTENSION bdr;`
  - `bdr.bdr_subscribe('hostname=host1 dbname=db', 'hostname=host2 dbname=db');`

BDR vs. UDR



Budoucnost BDR

- Možnost filtrovat jednotlivé akce pro individuální tabulky (INSERT/UPDATE/DELETE)
- Plně synchronní replikace (bez konfliktů)
- Komplexní topologie (nejen mesh)

Užitečné odkazy

- http://2ndquadrant.com/bdr
- http://wiki.postgresql.org/wiki/BDR_User_Guide
- git://git.postgresql.org/git/2ndquadrant_bdr.git
  - bdr-plugin/stable – aktuální stable verze
  - bdr-plugin/next – aktuální vývojová verze
  - bdr-pg/REL9_4_STABLE – opatchovaný Postgres 9.4
- Otázky pro 2ndquadrant: bdr@2ndquadrant.com
- Otázky o použití: pgsql-general@postgresql.org