Kam kráčíš, PostgreSQL?

Tomáš Vondra <tomas.vondra@enterprisedb.com>
<tv@fuzzy.cz>

Agenda

- trochu o historii
- komunita a ekosystém
- hlavní projekty
  - dlouhodobé, long-shot
  - interní (komunita)
  - externí (firmy okolo)

EDB

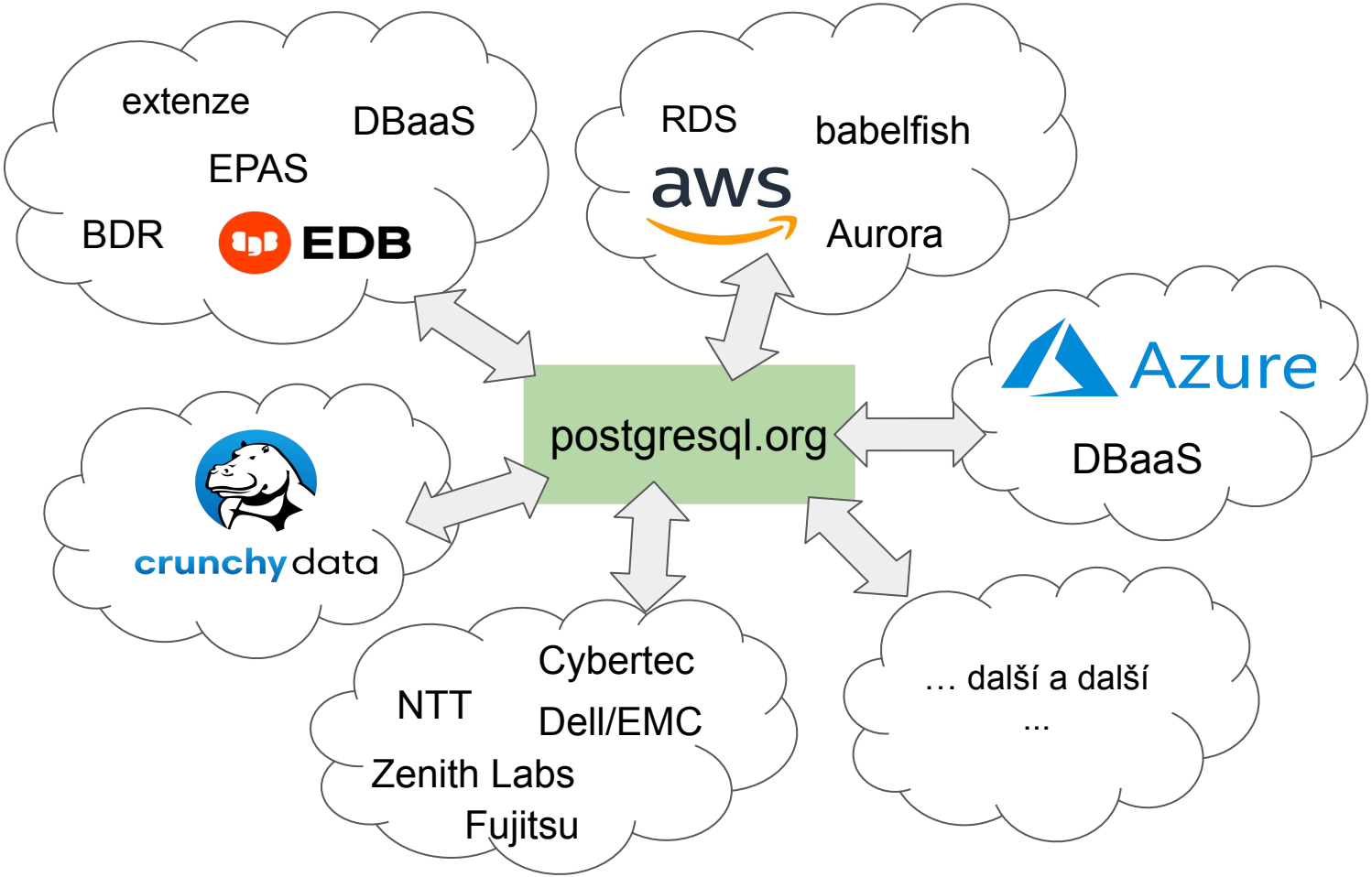
Vše co tu prezentuji jsou moje názory,

odhady a (často) spekulace.

Historie

- relační databáze
- původně výzkumný projekt na UC Berkeley
  - LSD, BSD, PostgreSQL, …
- od roku 1995 komunitní vývoj
  - žádný "vlastník" zdrojáků
  - různé (proprietární / specializované) forky
  - spousta firem spolupracuje na vývoji
  - důraz na kontinuitu, zpětnou kompatibilitu, flexibilitu

EDB


extenze
DBaaS
EPAS
BDR
EDB
RDS
babelfish
aws
Aurora
Azure
DBaaS
postgresql.org
crunchy data
Cybertec
NTT
Dell/EMC
Zenith Labs
Fujitsu
… další a další
...

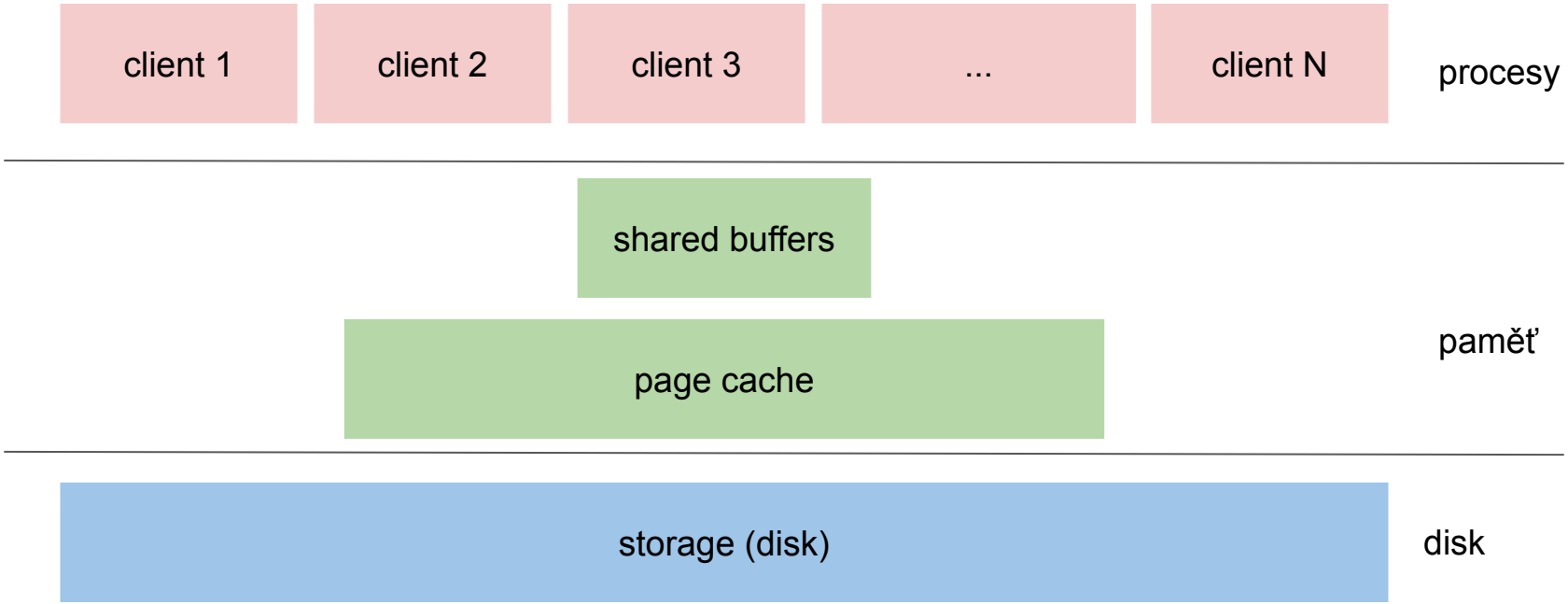
EDB

Architektura

- víceméně tradiční architektura
  - general purpose
  - nedistribuovaný systém
  - procesy (1:1 s klienty)
  - OLTP, postupně i OLAP
- ale i mnoho "revolučních" vlastností
  - rozšiřitelnost (UDF, datové typy, indexy, callbacky, …)
  - snaha o flexibilní API + externí řešení
  - ...

EDB

Architektura



EDB

Forky

Nikdo nechce udržovat invazivní forky!

On-premise forky

- víceméně každá firma má nějaký svůj fork
- způsob jak zákazníkům dodat vylepšení rychleji
  - Postgres má 1 rok release cyklus
- část vylepšení je nepřijatelná pro komunitu
  - podpora specifického replikačního řešení
  - vylepšení pro kompatibilitu s jinými DB
- balíček s dalšími extenzemi / produkty
  - HA, management, …

EDB

RDBA/DBaaS Forky

- víceméně "čistý" PostgreSQL
  - akorát ho "spravuje" někdo další
- nikdo nemá zájem udržovat invazivní forky
- patche zlepšující monitoring, …
- vylepšení HA, replikace, …
- 99% vylepšení se vrací komunitě

EDB

Cloud-native forky

- výrazně modifikovaný PostgreSQL
  - vesměs silně "opinionated"
- poměrně invazivní změny
  - alternativní storage model (Zenith)
  - multi-master replikace (BDR), sharding, …
- specifické změny zůstanou ve forcích
  - competitive advantage + nepřijatelnost pro komunitu
- infrastrukturní patche / API se vrací komunitě
  - logická replikace, FDW, partitioning, paralelismus, ...

EDB

Komunitní patche

https://commitfest.postgresql.org/

EDB

Commitfests

The following commitfests exist in the system. Current review work is done in commitfest 2021-11. New patches should be submitted to commitfest 2022-01.

- 2022-03 (Future - 2022-03-01 - 2022-03-31)
- 2022-01 (Open - 2022-01-01 - 2022-01-31)
- 2021-11 (In Progress - 2021-11-01 - 2021-11-30)
- 2021-09 (Closed - 2021-09-01 - 2021-09-30)
- 2021-07 (Closed - 2021-07-01 - 2021-07-31)
- 2021-03 (Closed - 2021-03-01 - 2021-03-31)
- 2021-01 (Closed - 2021-01-01 - 2021-01-31)
- 2020-11 (Closed - 2020-11-01 - 2020-11-30)
- 2020-09 (Closed - 2020-09-01 - 2020-09-30)
- 2020-07 (Closed - 2020-07-01 - 2020-07-31)
- 2020-03 (Closed - 2020-03-01 - 2020-03-31)
- 2020-01 (Closed - 2020-01-01 - 2020-01-31)
- 2019-11 (Closed - 2019-11-01 - 2019-11-30)
- 2019-09 (Closed - 2019-09-01 - 2019-09-30)
- 2019-07 (Closed - 2019-07-01 - 2019-07-31)
- 2019-03 (Closed - 2019-03-01 - 2019-03-31)
- 2019-01 (Closed - 2019-01-01 - 2019-01-31)
- 2018-11 (Closed - 2018-11-01 - 2018-11-30)
- 2018-09 (Closed - 2018-09-01 - 2018-09-30)
- 2018-07 (Closed - 2018-07-01 - 2018-07-31)
- 2018-03 (Closed - 2018-03-01 - 2018-03-31)
- 2018-01 (Closed - 2018-01-01 - 2018-01-31)

EDB

Commitfest 2021-11

Search/filter | Shortcuts

Status summary: Needs review: 165. Waiting on Author: 54. Ready for Committer: 16. Committed: 35. Moved to next CF: 1. Returned with Feedback: 8. Rejected: 1. Withdrawn: 7. Total: 287.

Active patches

Patch	Status	Ver	Author	Reviewers	Committer	Num cfs	Latest activity	Latest mail
Bug Fixes								
standby recovery fails when re-replaying due to missing directory which was removed in previous replay.	Needs review	stable	Kyotaro Horiguchi (horiguti), Paul Guo (paulguo)			13	2021-10-01 06:45	2021-11 12:34
pg_upgrade fails with non-standard ACL	Waiting on Author		Anastasia Lubennikova (lubennikovaav), Artur Zakirov (a.zakirov)	Grigory Smolkin (g.smolkin.postgrespro.ru)	nmisch	12	2021-10-01 17:57	2021-03 07:25
Corruption during WAL replay	Needs review	stable	Teja Mupparti (tejam)			8	2021-10-04 20:14	2021-09 08:30

EDB

Budoucí patche (bez záruky)

- storage
- connection pooling
- failover / HA
- sharding
- logická replikace

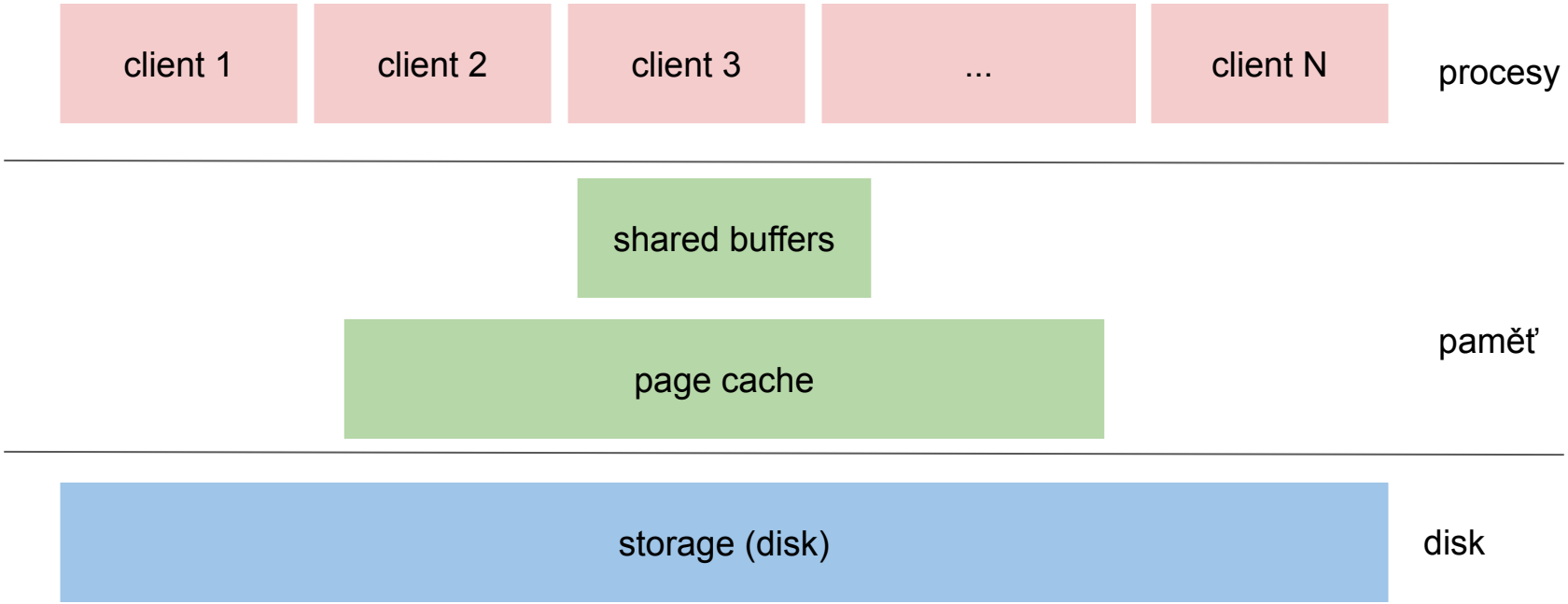
EDB

Asynchronní / direct I/O

- https://commitfest.postgresql.org/35/3316/
- dosud synchronní buffered I/O
  - malá kontrola, spoléhá se na kernel OS
  - jednoduchost, přenositelnost, slušný výkon
  - good enough
- asynchronní a direct I/O
  - detailnější kontrola dle potřeb DB (co cachovat, co ne, …)
  - zřejmě jen vybrané platformy, ...
- velmi aktivní práce / postup

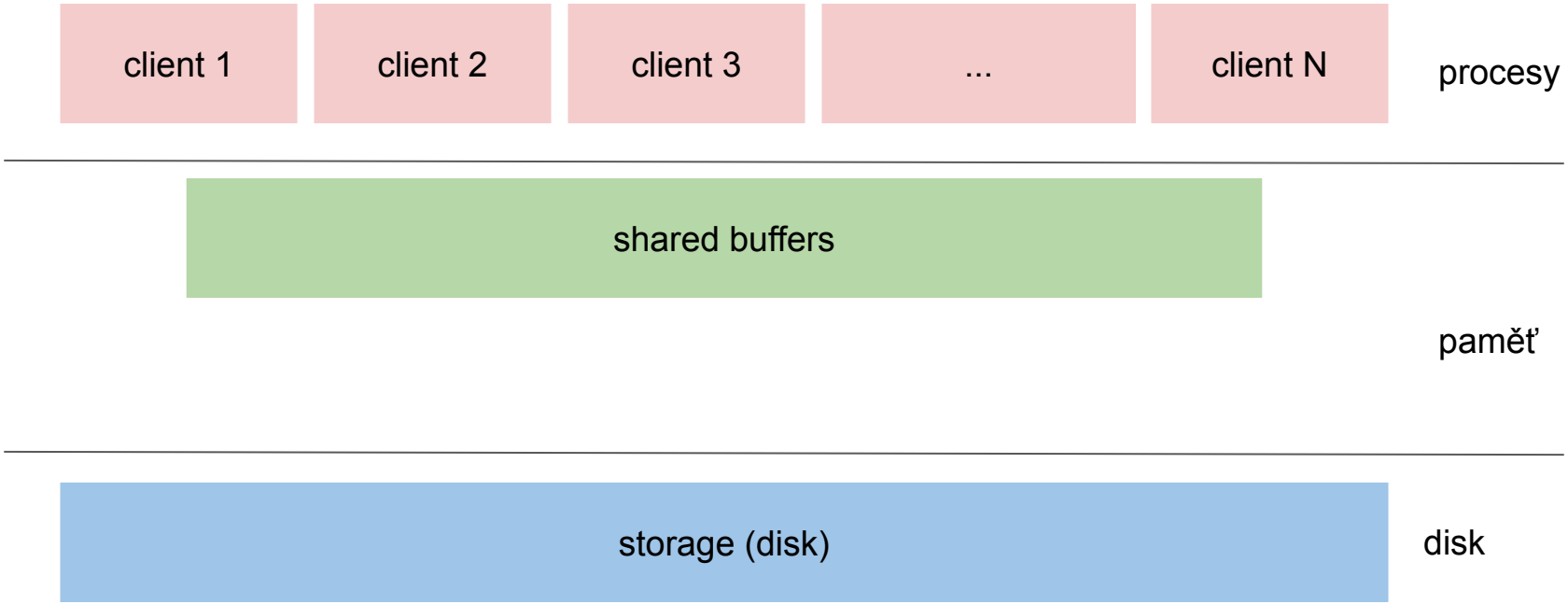
EDB

Buffered I/O



EDB

Direct + asynchronní I/O



EDB

Alternativní "storage engines"

- tradičně "row store"
  - slušné pro OLTP, horší pro analytické dotazy
- snaha umožnit alternativní formáty
  - interní API
  - zheap - jiný "row store" formát, řeší některé problémy
  - zedstore - formát pro "column store" (komprese, …)
- vývoj se "zasekl"
  - máme (čistší) interní separaci / API
  - vývoj zheap/zedstore aktuálně příliš nepostupuje

EDB

Transparent Data Encryption

- žádné built-in transparentní šifrování
- data-at-rest typicky přes dm-crypt
  - jednoduché, osvědčené řešení
  - ne vždy možné použít (restricted environments)
  - mohu číst soubory => nešifrované
- patch přidává šifrování přímo v DB (při zápisu)
  - složitější než se zdá (různé typy dat, replikace, ...)
  - potenciálně "use cases" které dm-crypt neumí
- poměrně aktivní vývoj
  - dlouhé diskuse o šifrovací metodě (XTS, …)

EDB

Sharding

- oficiální "built-in sharding" neexistuje
  - z minulosti existují různá "zbastlená" řešení
- používá se partitioning + FDW (a to se vylepšuje)
  - Citus, Timescale
- setrvalý inkrementální vývoj
  - push-down co nejvíce operací
  - asynchronní exekuce
  - batching operací
  - …

EDB

Logická replikace

- dekódování "logických změn" z WAL
- aktuální patche
  - column/row filtering
  - dekódování sekvencí
  - dekódování na (fyzické) standby
- privátní projekty (BDR)
  - async multi-master
  - conflict resolution

EDB

Connection pooling

- klasicky problém s velkým počtem spojení
  - max_connections = 10000 :-(
- každé spojení má dedikovaný "vlastní" proces
  - částečně historické důvody, částečně jednodušší
- procesy mají vyšší režii oproti threadům (zavádějící)
  - sdílení méně dat vs. více zamykání, jasnější oddělení
- overhead kvůli počtu "sessions" (snapshotů)
  - je jedno jestli to je proces nebo thread
- časté připojování / odpojování => naprd je obojí

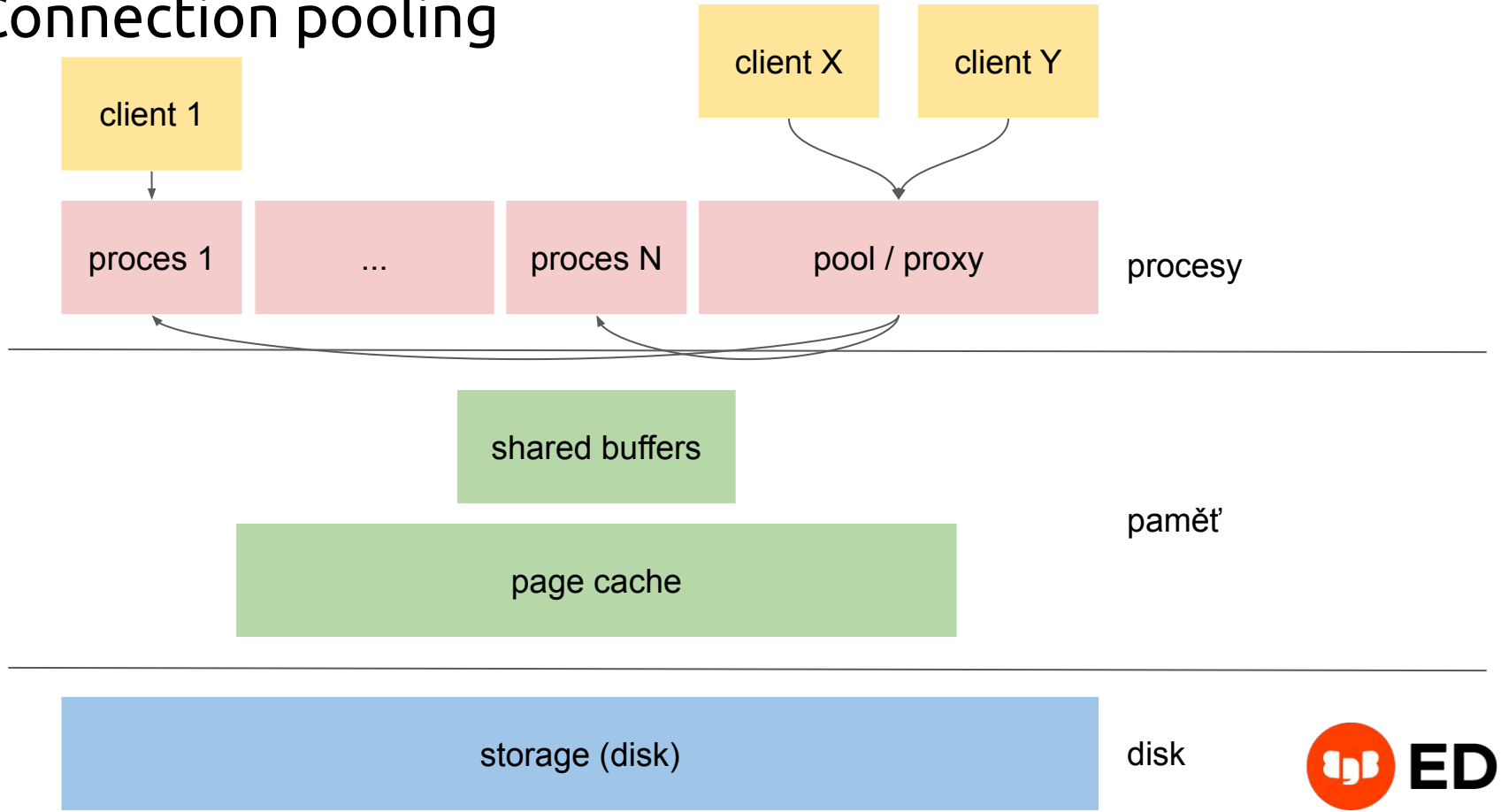
EDB

Connection pooling

- dlouhodová optimalizace "per session" overheadu
  - stále platí že stále máte jenom X jader procesoru
  - … ale když už musíte mít hromadu "idle" spojení
- patch přidává zabudovaný connection pool
  - funguje podobně jako externí pool, ale managed
  - obdobná omezení (např. prepared statements)
  - vývoj trochu zamrzl, snad pokročí pro PG 15

EDB

Connection pooling



EDB

Q&A