Kam data chodí spát

Zdeněk Kotala

Agenda

- MVCC
  - ACID
  - Izolační úrovně
- Vnitřní architektura
- Datové úložiště
  - Data
  - Commit Log
  - Write Ahead Log

MVCC

- MultiVersion Concurrency Control je metoda jak zajistit paralelní přístup k databázi více uživatelům, při zajištění ACID kritérií při co největší propustnosti
- PostgreSQL používá multigenerační architekturu (MGA) převzatou z InterBase.
- Záznamy se nepřepisují, vždy se vytváří nová kopie.

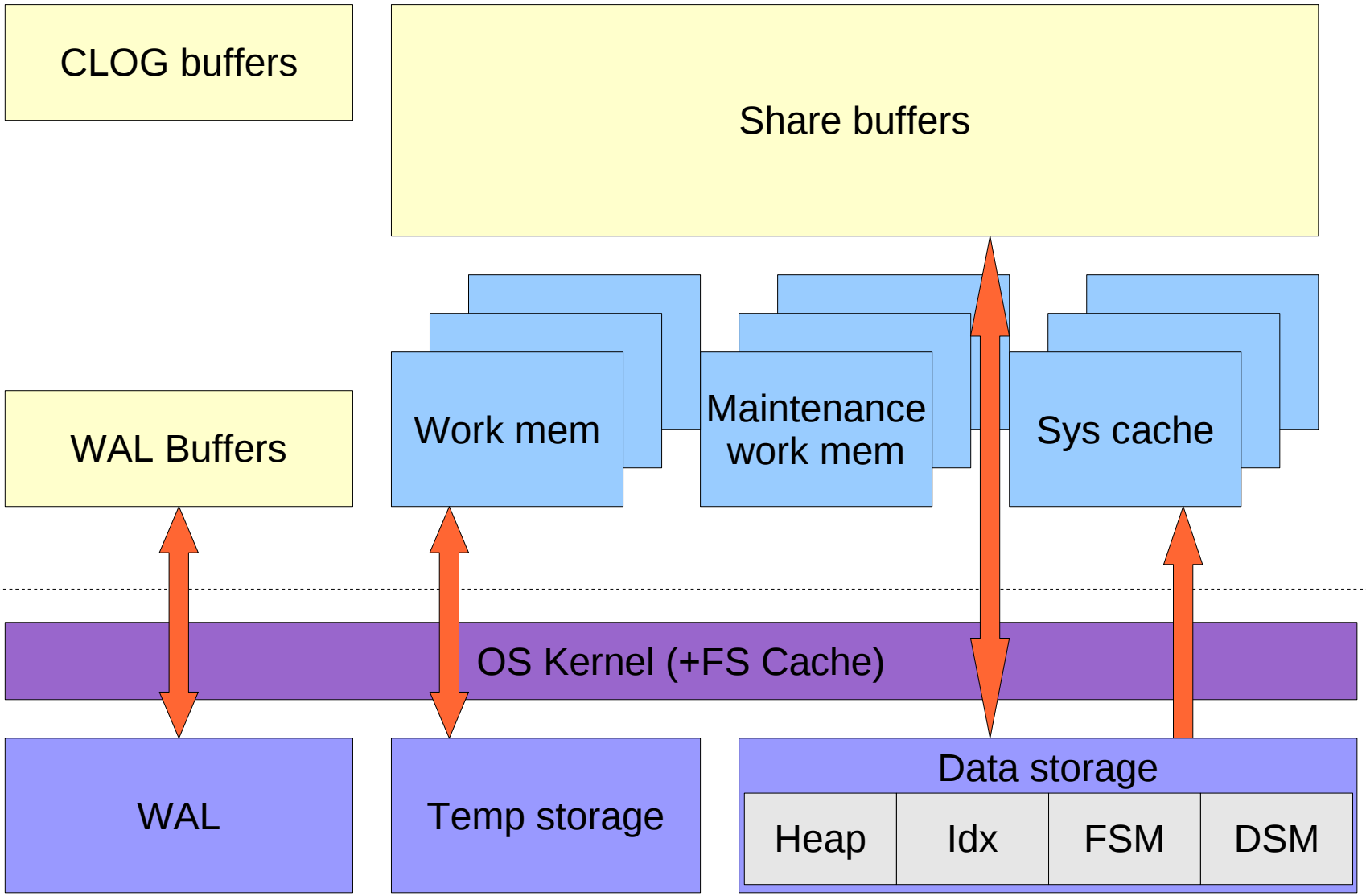
ACID

- **Atomicity** – Vše a nebo nic
- **Consistency** – Databáze je vždy konzistentní
- **Isolation** – Jedna transakce nevidí co dělá jiná
- **Durability** – Potvrzená transakce je neměnná

Izolační úrovně

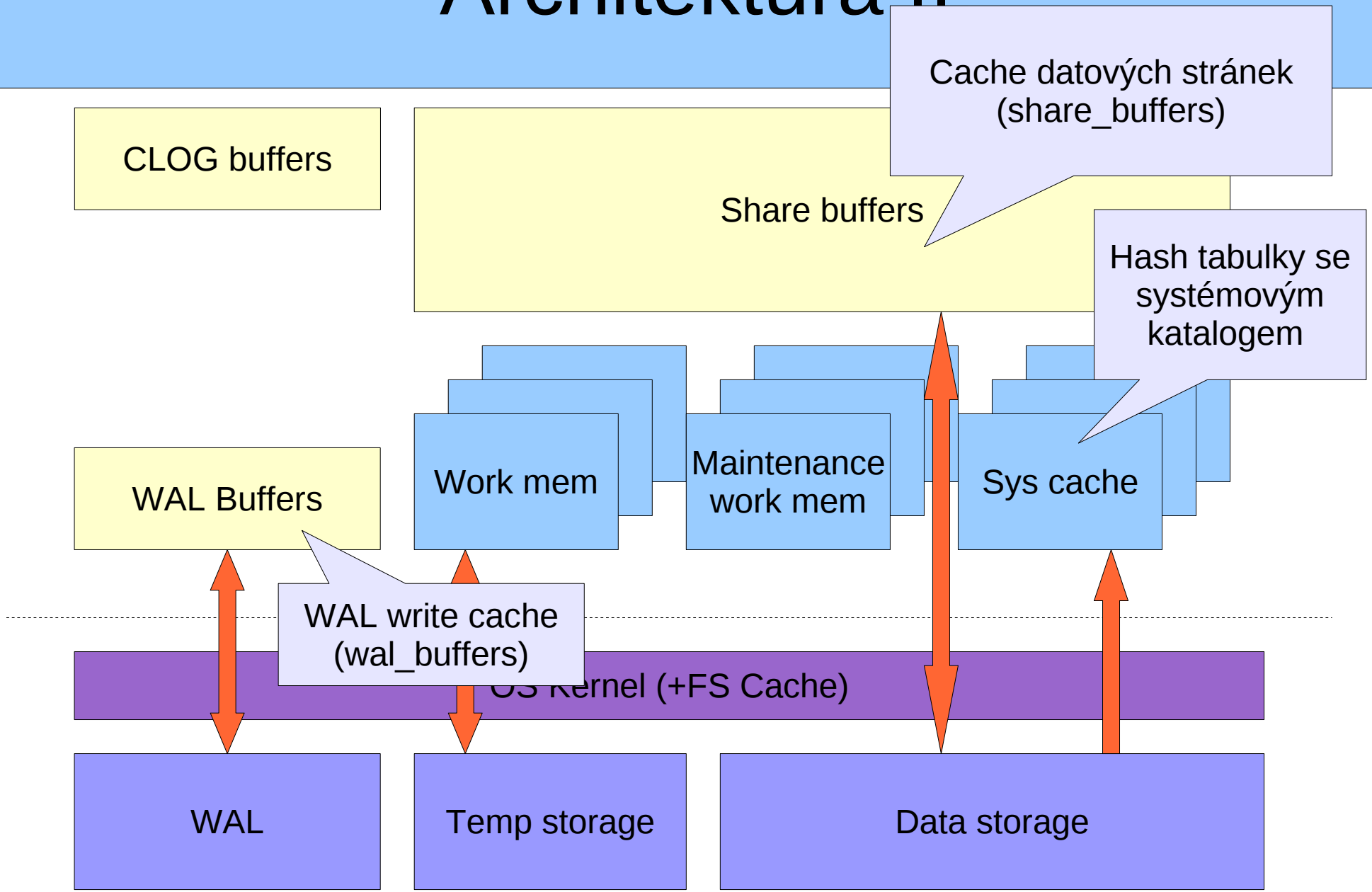
- Read Uncommitted – čte i nepotvrzená data
- Read Committed – opakovaný dotaz v transakci může vrátit jiný výsledek
- Repeatable Read – opakované čtení vrací stejný výsledek (s vyjímkou “phantom reads”)
- Serializable – emuluje sériové provádění transakcí

Architektura I.

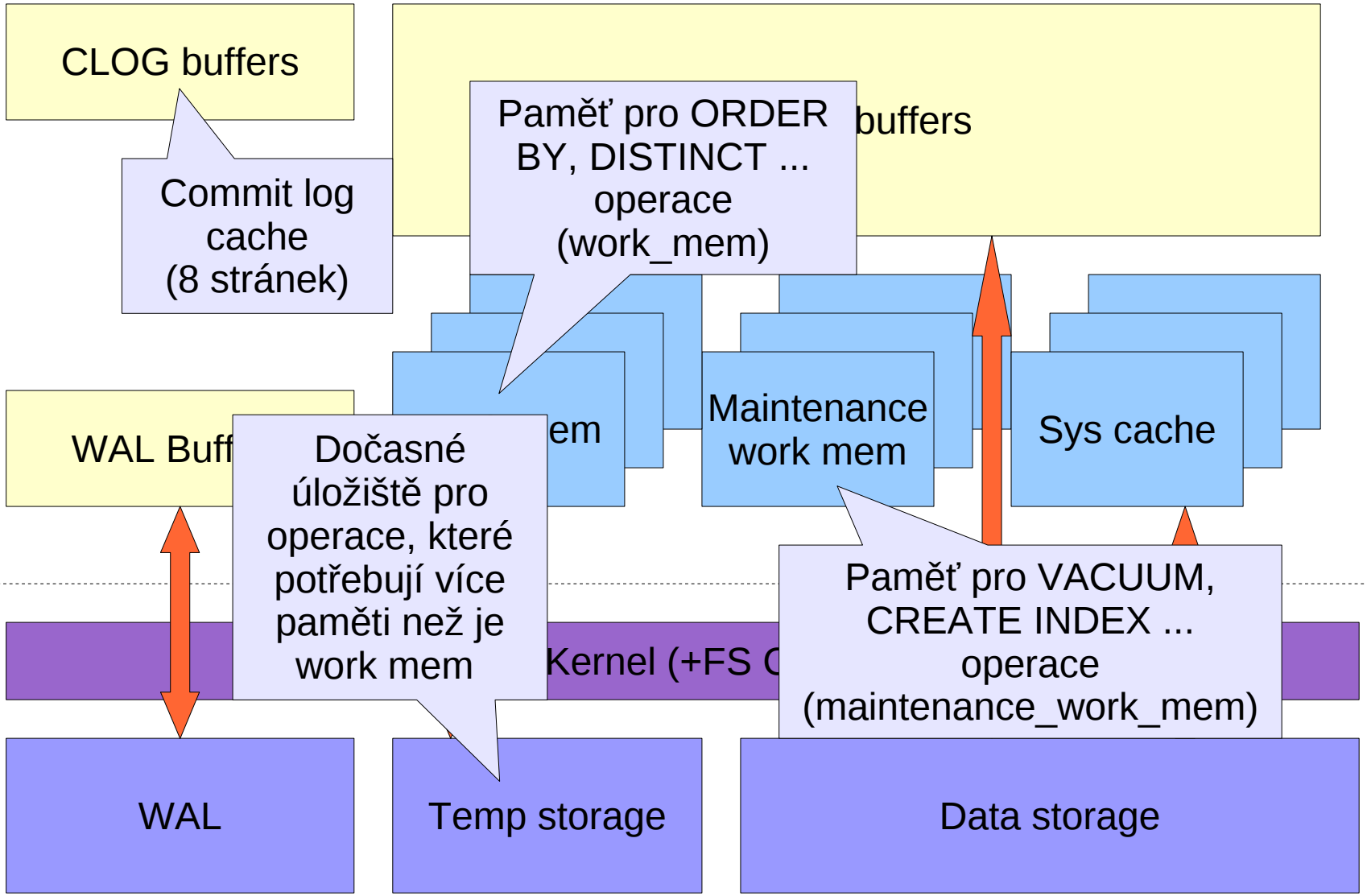


Poznámka: Clog, FSM, Dead space map zde nejsou zmíněny.

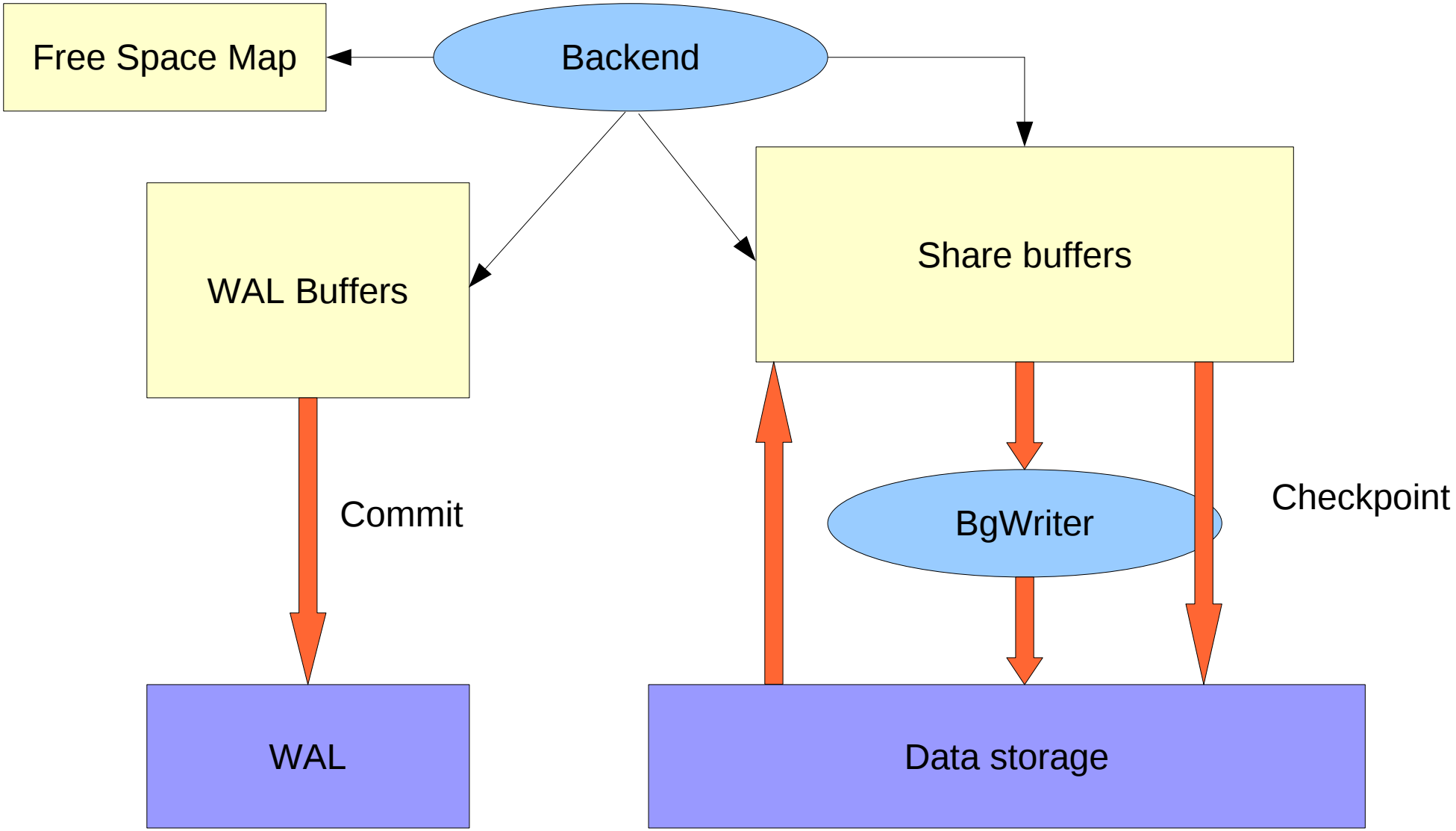
Architektura II



Architektura III.



Jak pracuje INSERT



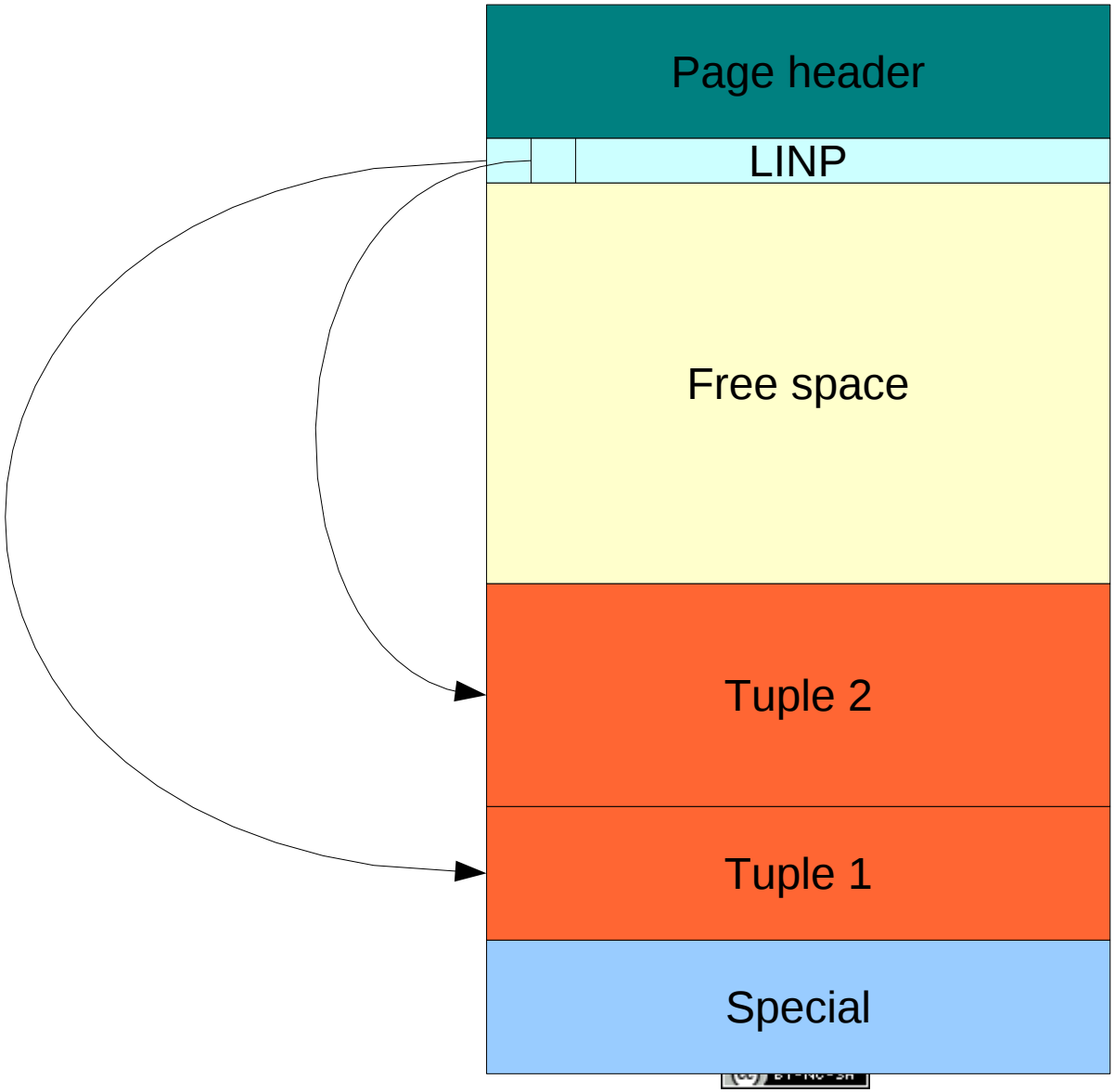
Datové úložiště

- $DATA/global
- $DATA/base/<OID DB>/<OID REL>
- $DATA/pg_clog
- $DATA/pg_xlog
- $DATA/pg_tblspc/<OID TBL SPC>

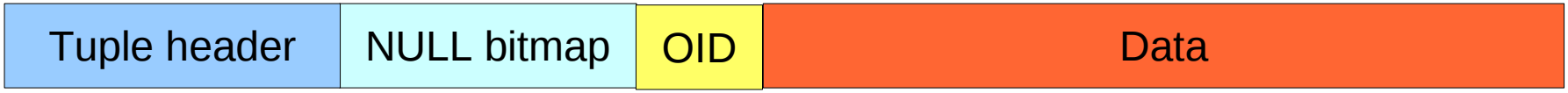
Relace

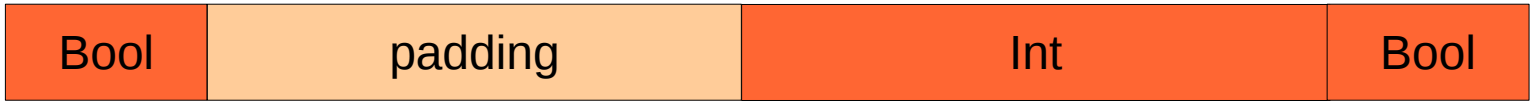
- Jméno pg_class.relfilenode
- Maximální velikost 32TB
- Rozdělené na 1GB segmenty, každý segment je samostatný soubor s číselnou příponou
- Velikost stránky 8kB
- Free Space Map a Dead Space Map jsou implementovány jako “forks” (přípony _fsm, _vm)

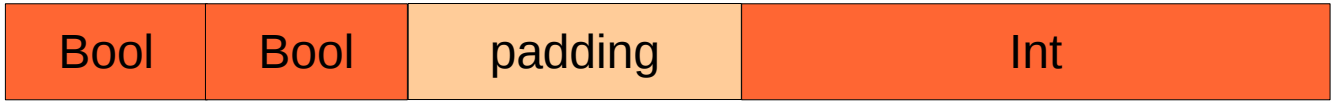
Datová stránka



Tuple/Datum







TOAST

- Atributy, které přesáhnou velikost ~2kB jsou uloženy v TOAST tabulce
- V hlavní relaci je uložen jen TOAST pointer
- Data jsou rozděleny na “chunky” po ~2kB, které se skládají z ID, chunk ID a vlastních data
- TOAST tabulka má vlastní index pro rychlé vyhledávání data
- TOAST tabulka obsahuje různé atributy
- TOAST data mohou být komprimovány

CLOG

- Commit log slouží k uložení informací o transakcích, nutných pro určení viditelnosti záznamů
- Každá transakce je vyjádřena dvoubitovou informací
- Rozdělen do segmentů po 256KB (32 stránek)
- VACUUM odklízí nepotřebné segmenty

WAL

- Write Ahead Log slouží pro ukládání veškerých změn v datech
- Zajišťuje konzistenci dat v případě výpadku
- Segment 16M je předalokován, aby vždy byl dostatek místa na disku
- Zapisuje se incrementálně – není nutná cache systému

Dotazy

?