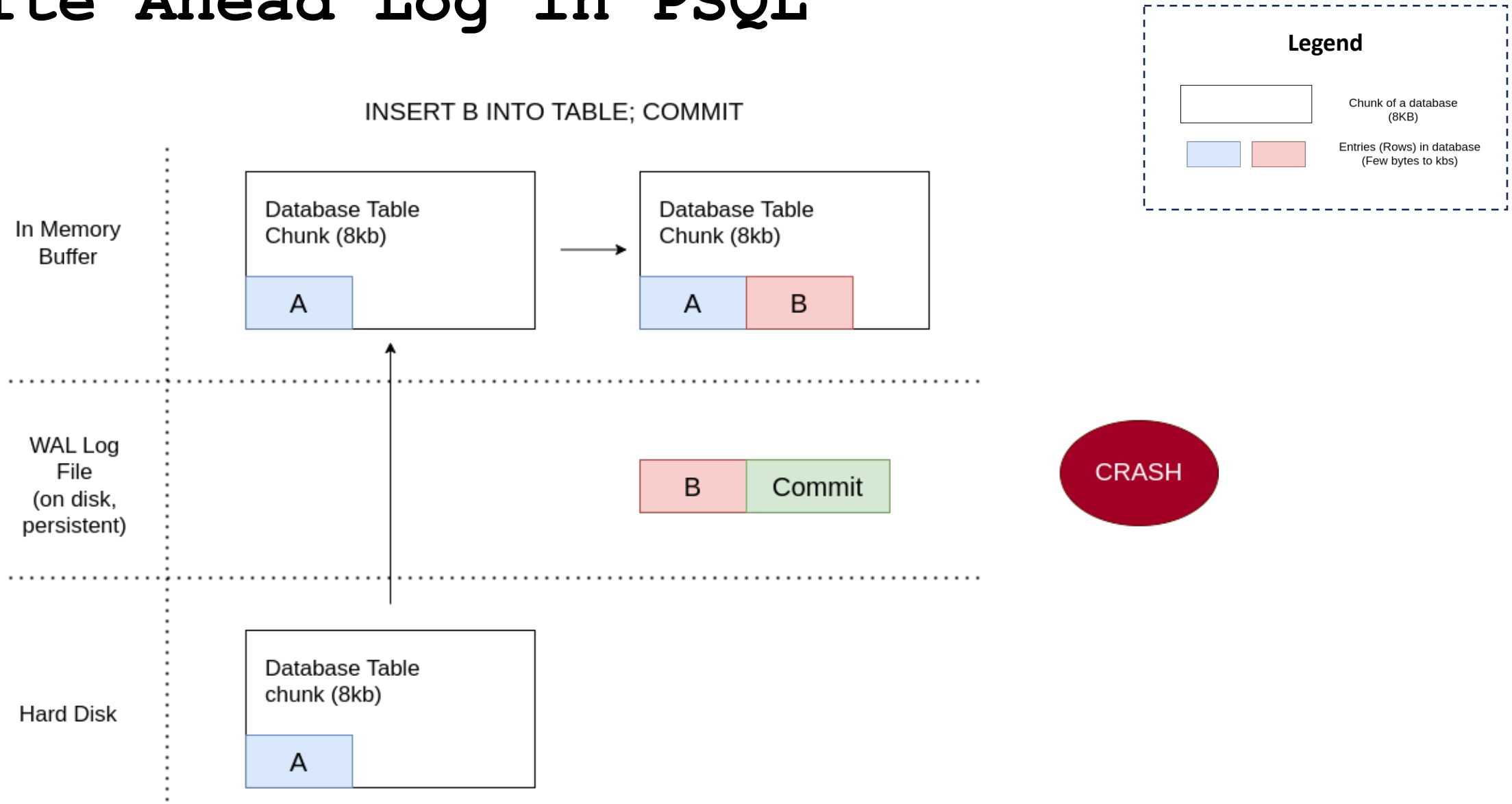
Atomic Writes & Modern Databases

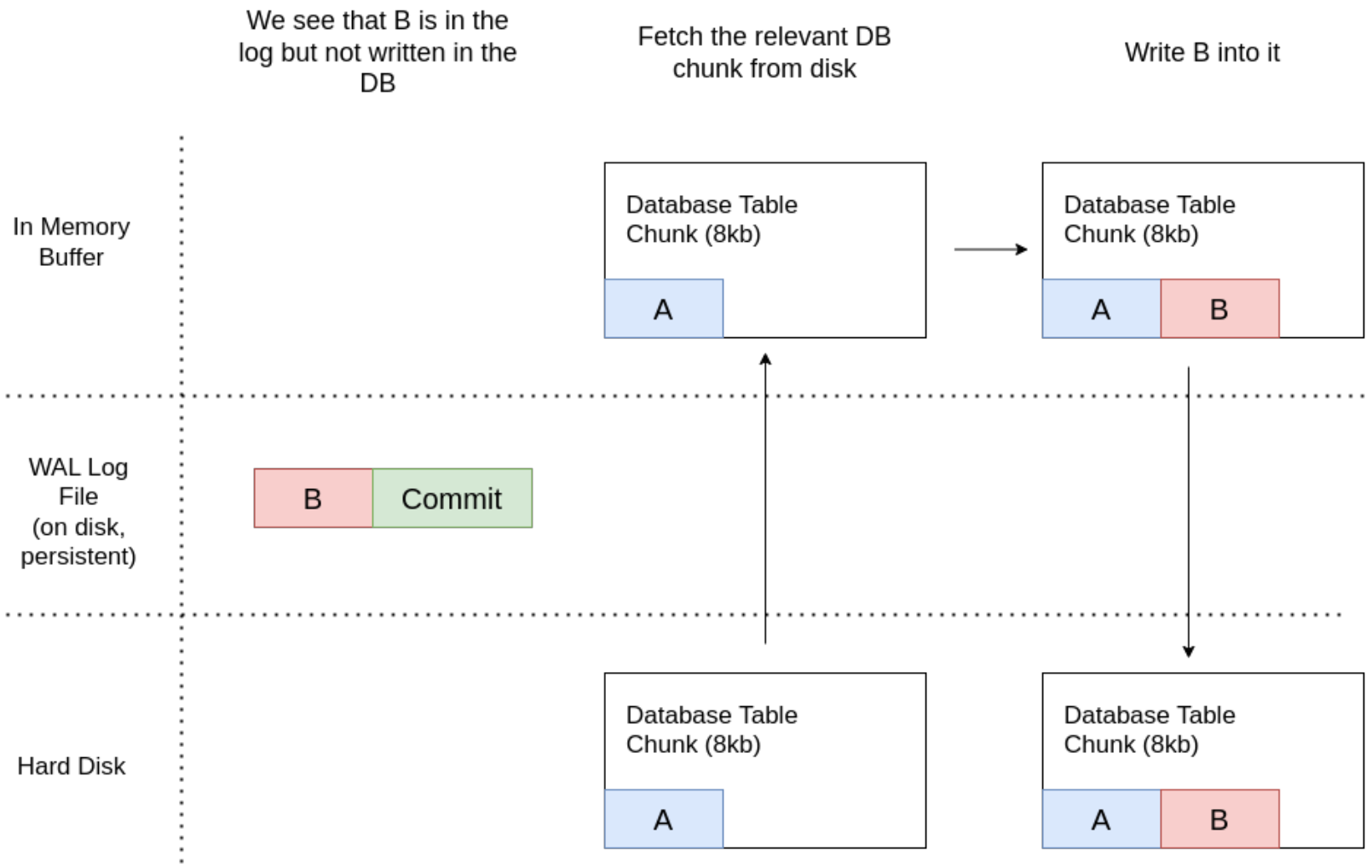
Ojaswin Mujoo

Linux Technology Center, IBM

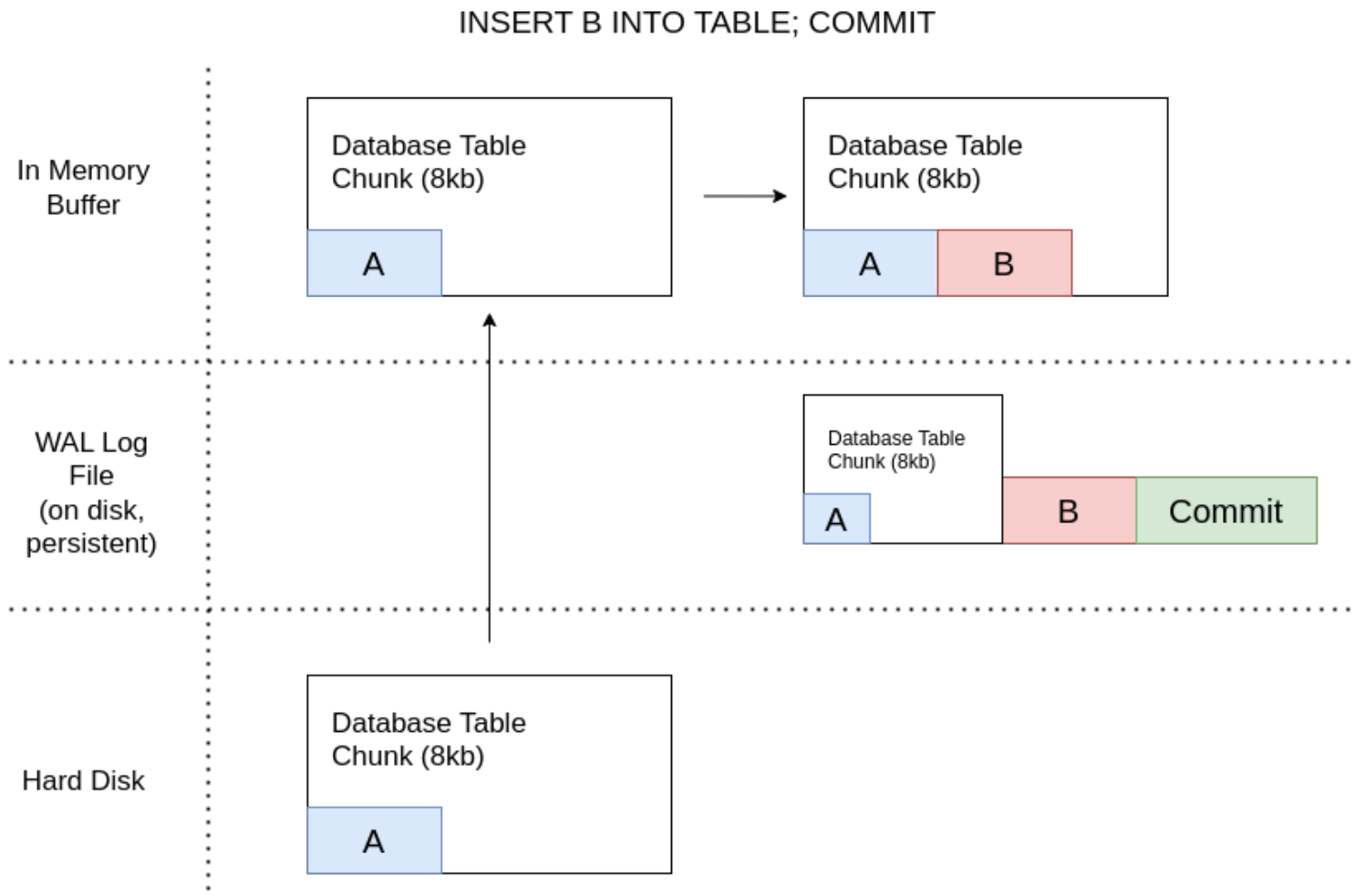
Write Ahead Log in PSQL



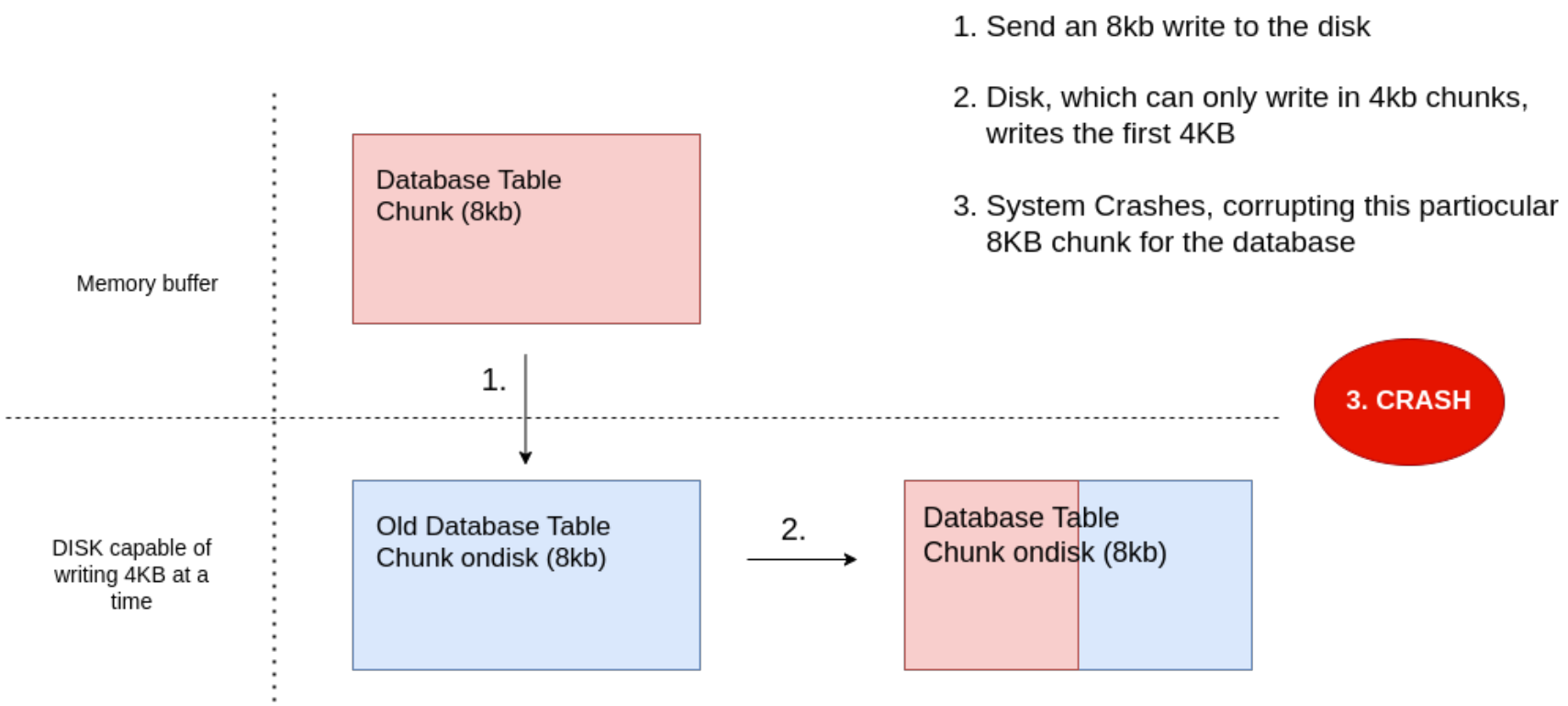
Log Recovery in PSQL



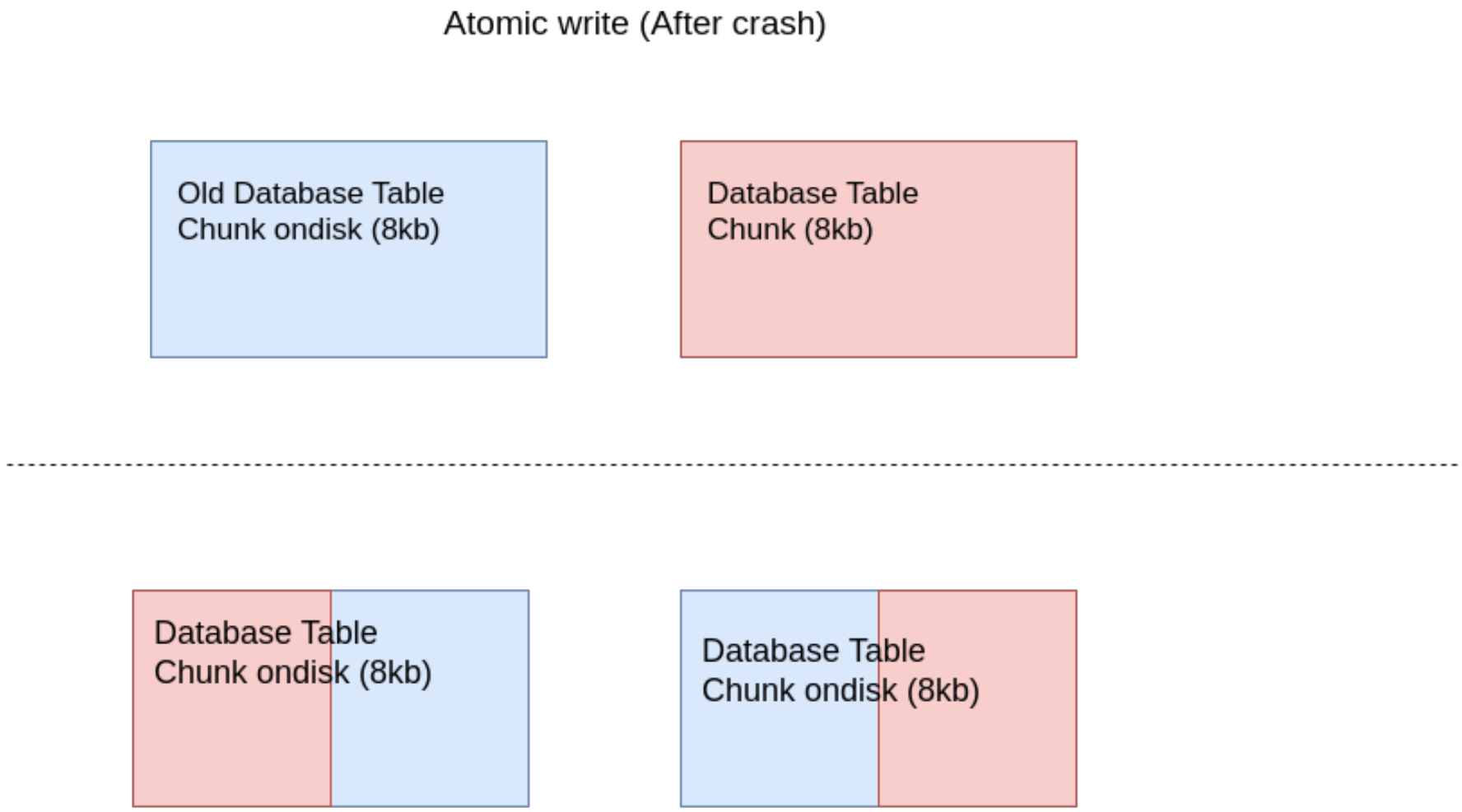
WAL with full page writes



When would the page get corrupt?



Atomic vs Non Atomic Behavior



Challenges in Atomic Write

Userspace

- Userspace applications need to be modified to utilize the atomic capabilities

Kernel Filesystem Layer

- File system allocations must respect atomic alignments
- File systems must not split atomic writes before submitting them to block layer

Kernel Block Layer

- Block Layer should have support to dynamically detect the atomic capabilities of block device.
- Block Layer must also make sure to not split IO before submitting to block device layer.

Storage Device

- Storage Device needs atomic IO capabilities
- Storage device needs to advertise its atomic capabilities to the kernel

Current Work

- We have started seeing storage devices that provide upto 64KB atomic writes
- There's ongoing work to support atomic writes for direct IO.
  - Userspace must pass an **RWF_ATOMIC flag** during write via the pwritev2() syscall.
  - Ongoing work in XFS and EXT4 **ensures allocations will respect atomic constraints**
  - The block layer will **not split the IO**, and only proceed if atomic write is possible by the device, else fail.

- **Future Work**: Stabilize the design for direct IO and continue working on bufferred IO

Legal Statement

- This work represents the view of the author and does not necessarily represent the view of IBM.
- IBM and IBM(logo) are trademarks or registered trademarks of International Business Machines Corporation in the United States and/or other countries.
- Linux is a registered trademark of Linus Torvalds
- Other company, product, and service names may be trademarks or service marks of others.

Thank You