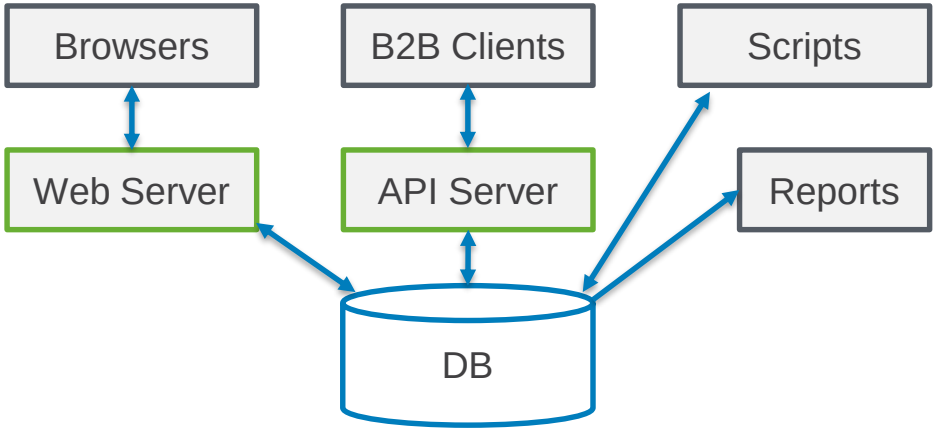
Understanding PostgreSQL IO

Jan Wieck, Principle Database Engineer

PGConf 2018

System Overview (simplified version)

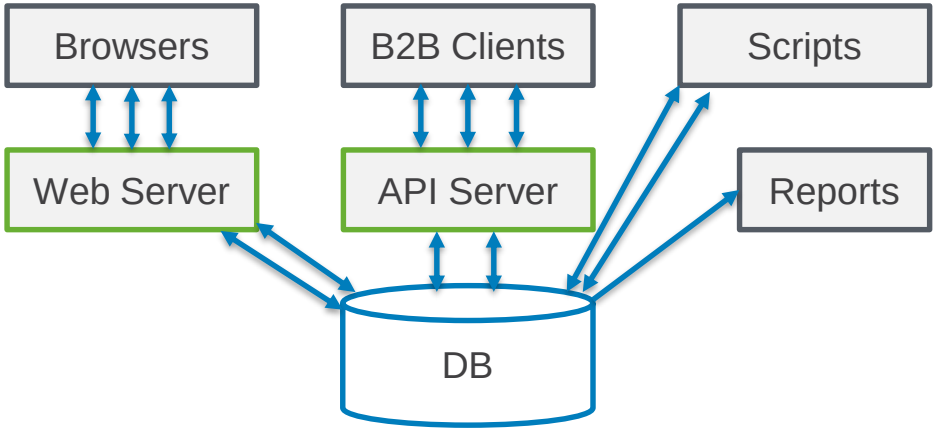


One Database usually serves multiple types of clients with different access patterns and IO requirements.

To satisfy all these requirements we need to understand the nature and priorities of the different clients.

aws

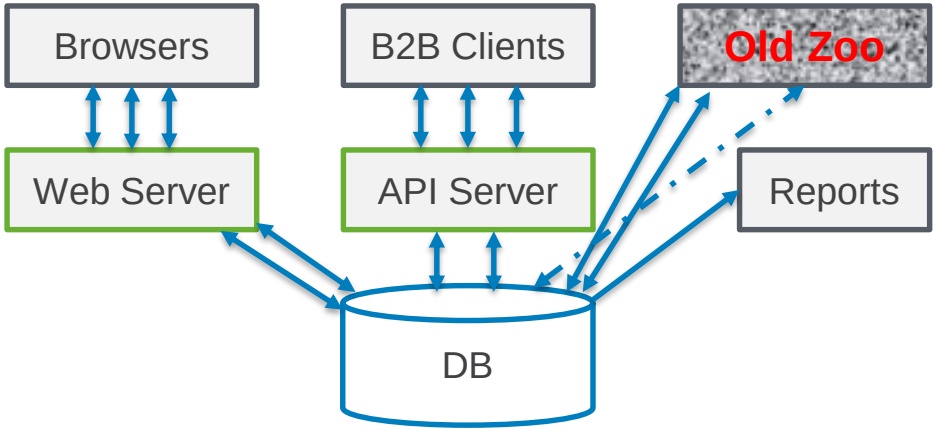
System Overview (less simplified)



Most of the clients have multiple threads or processes.

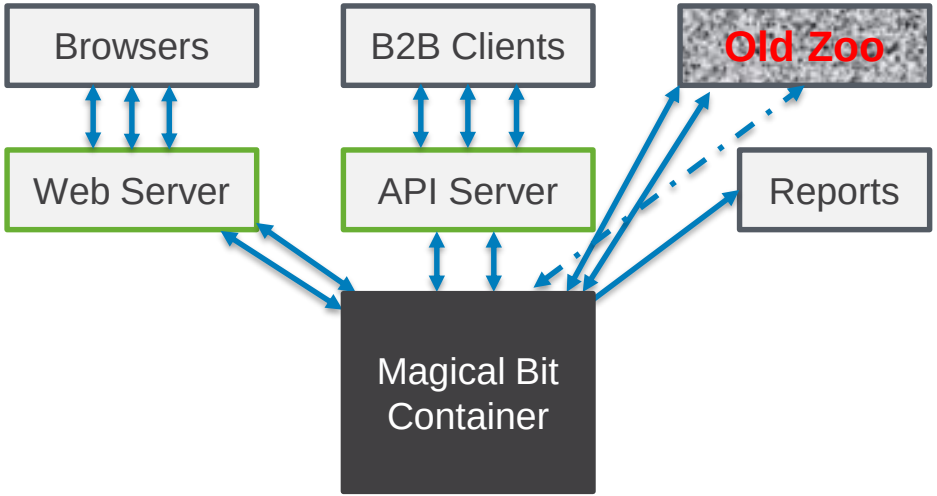
This can lead to dozens if not hundreds of long living database connections.

System Overview (closer to the truth)



Always expect to find some “legacy code”.

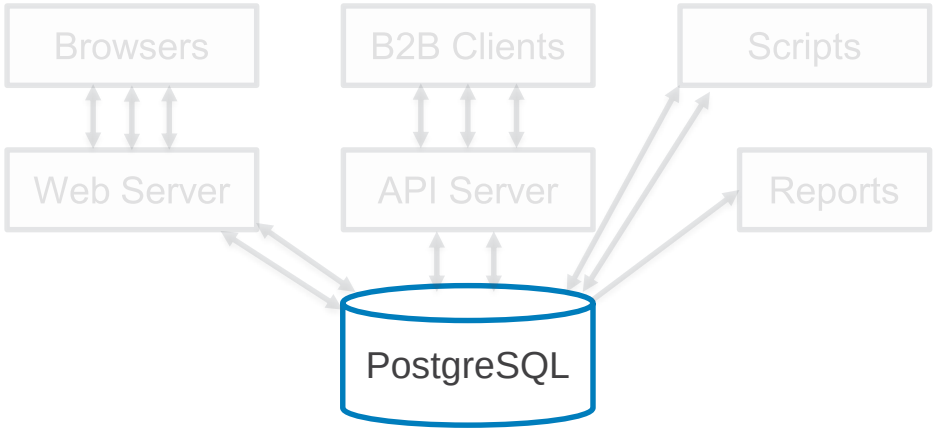
System Overview (even closer to the truth)



Many of those, who draw these types of system diagrams, don't really know what a database is or does.

aws

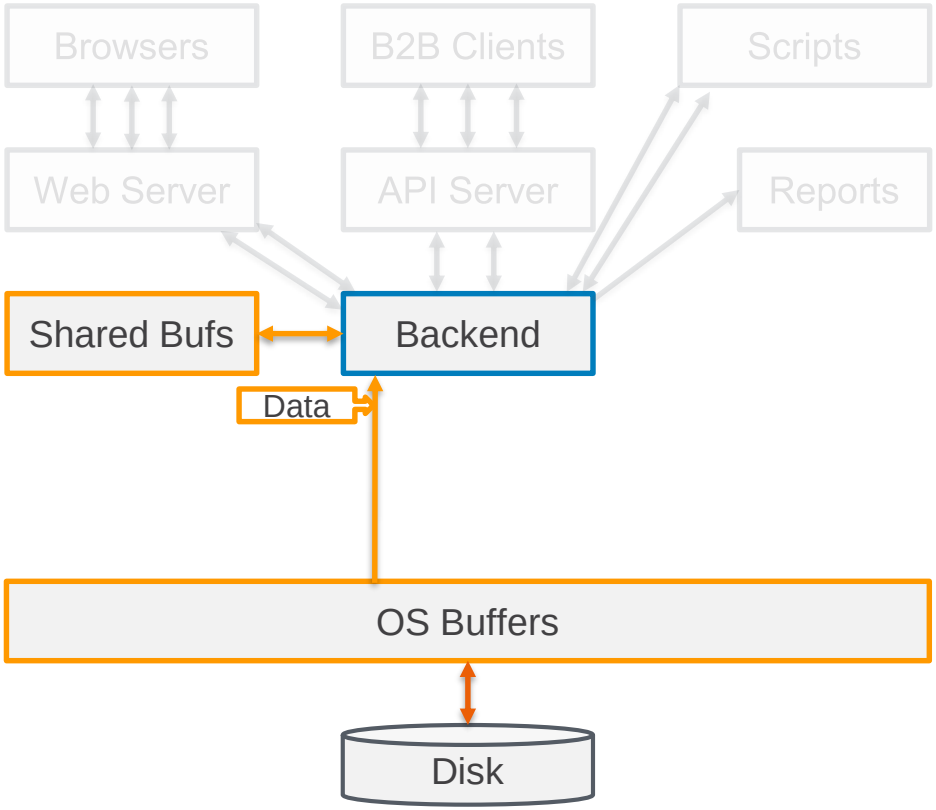
System Overview



Today we focus on the IO of PostgreSQL

aws

Buffering

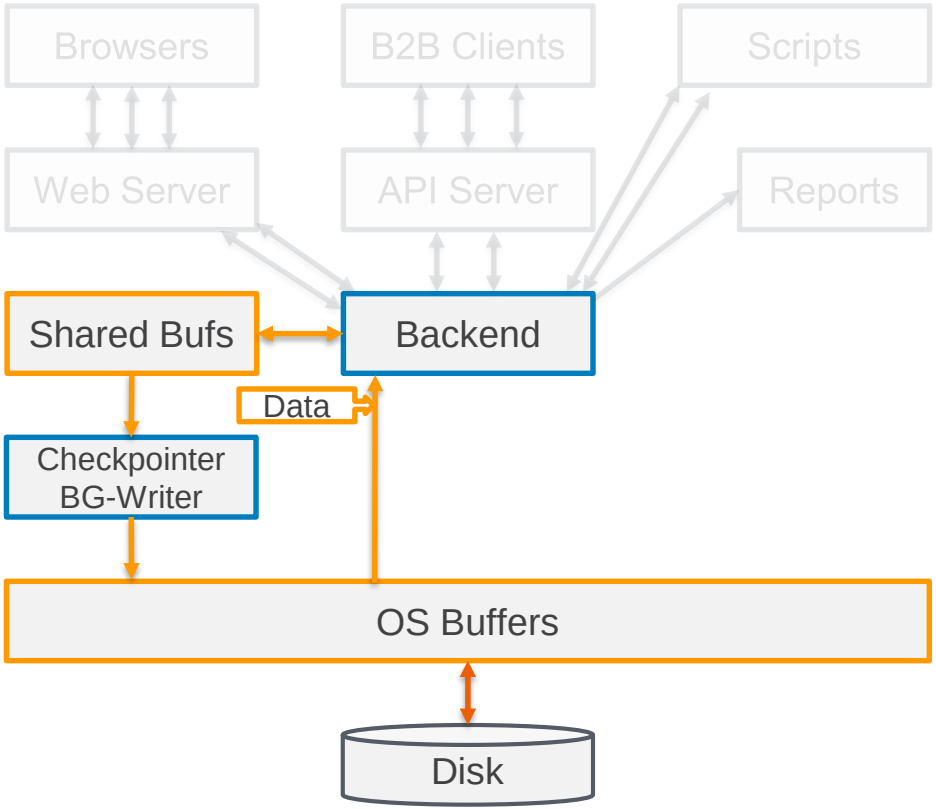


PostgreSQL does not operate on raw devices. It relies on a regular file system and the OS buffer cache.

Because of that tuning must not be limited to changing parameters in the file postgresql.conf. The OS kernel parameters are not “off limits”.

aws

Normal IO for Data



Under normal circumstances the backend should seldom to never write data (heap and index blocks).

postgresql.conf:

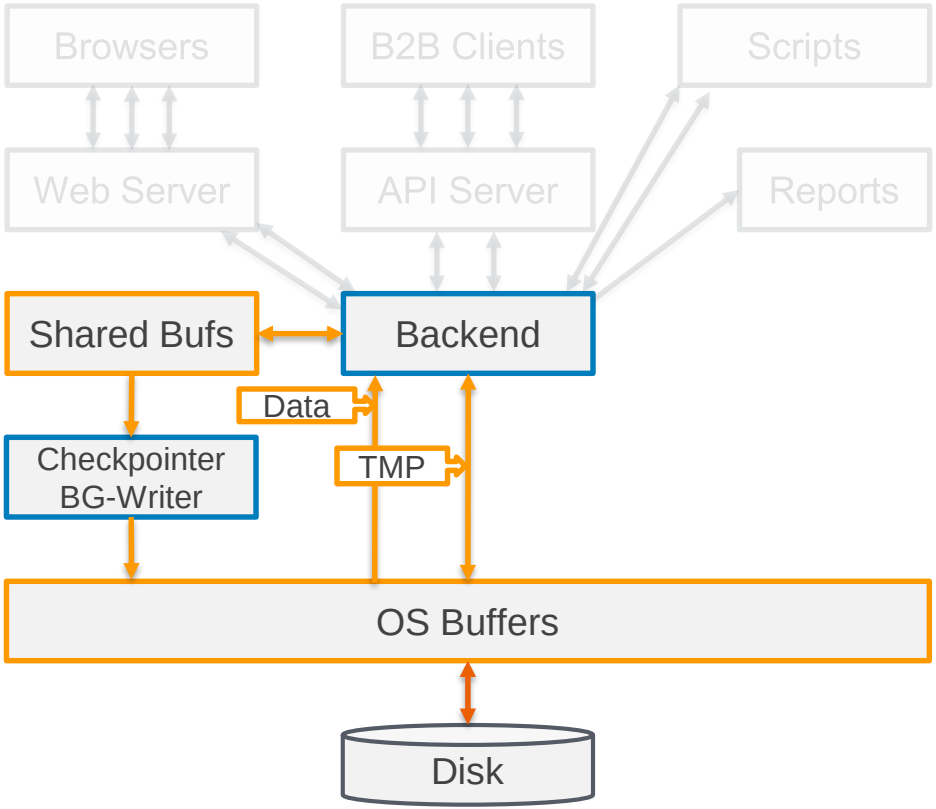
- `shared_buffers`
- `checkpoint_timeout`
- `checkpoint_completion_target`

Kernel parameters:

- `dirty_background_bytes`
- `dirty_ratio`
- `deadline IO scheduler`

aws

Temporary Files



PostgreSQL can spill large sort operations to disk.

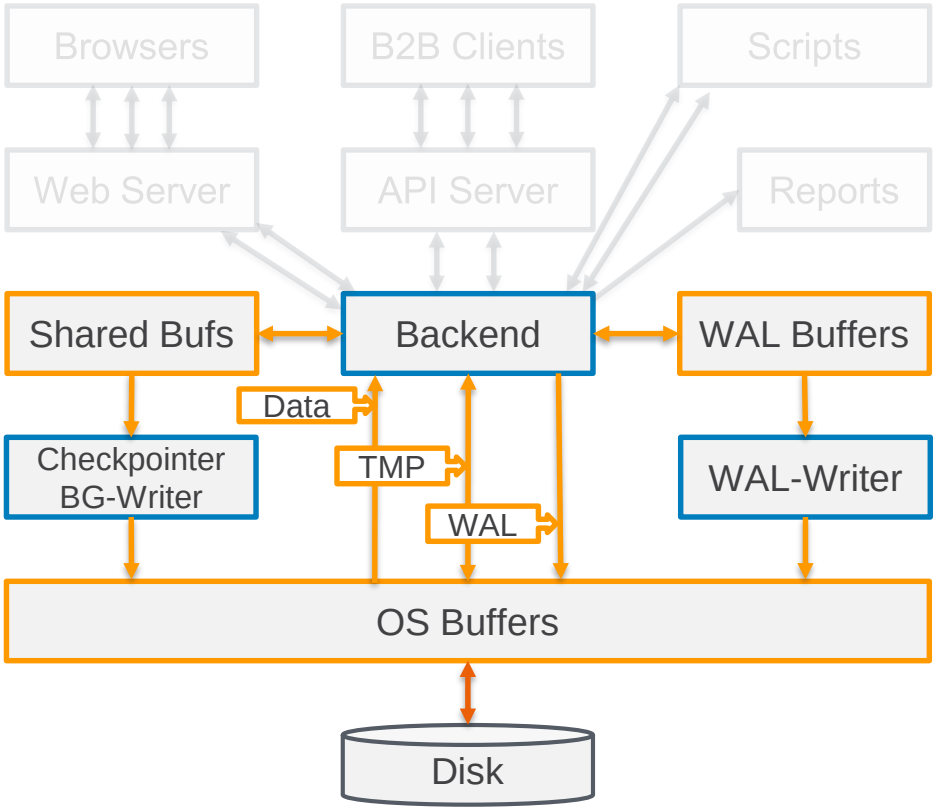
This is done on purpose because glibc's malloc()/free() clings to memory.

postgresql.conf:

- `work_mem`
- `maintenance_work_mem`

aws

Add WAL to the mix



WAL (redo log) has a very different IO pattern compared to data.

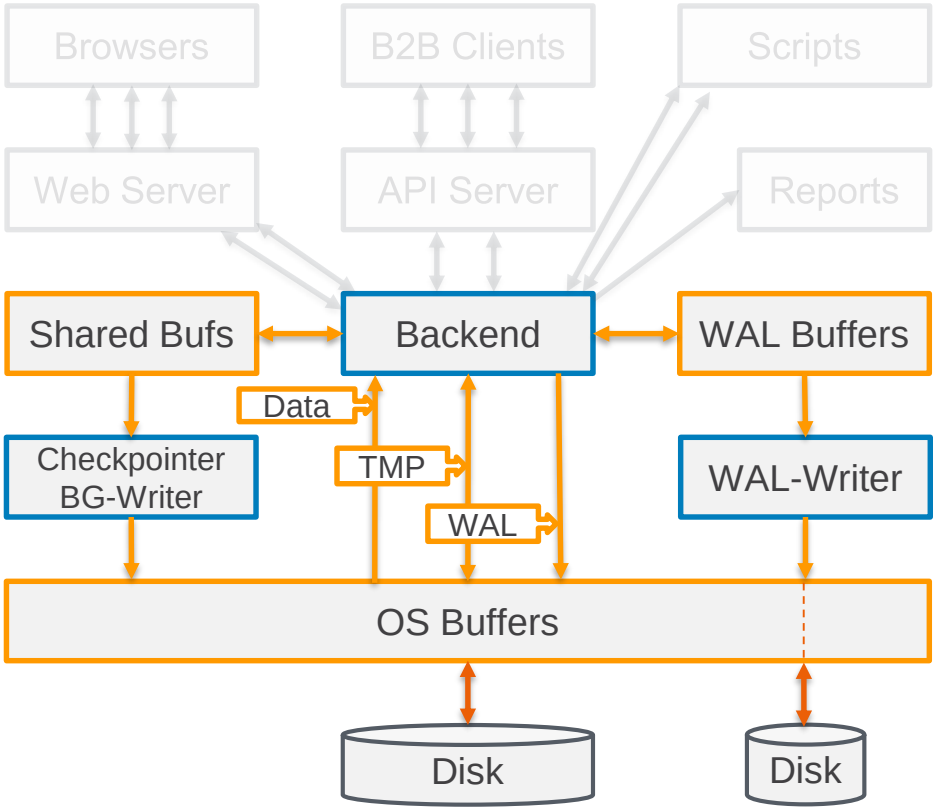
WAL is written sequentially (more or less) and has a high fdatasync() rate for OLTP.

postgresql.conf

- `wal_buffers`
- `max_wal_size`
- `synchronous_commit`

aws

Separate the WAL

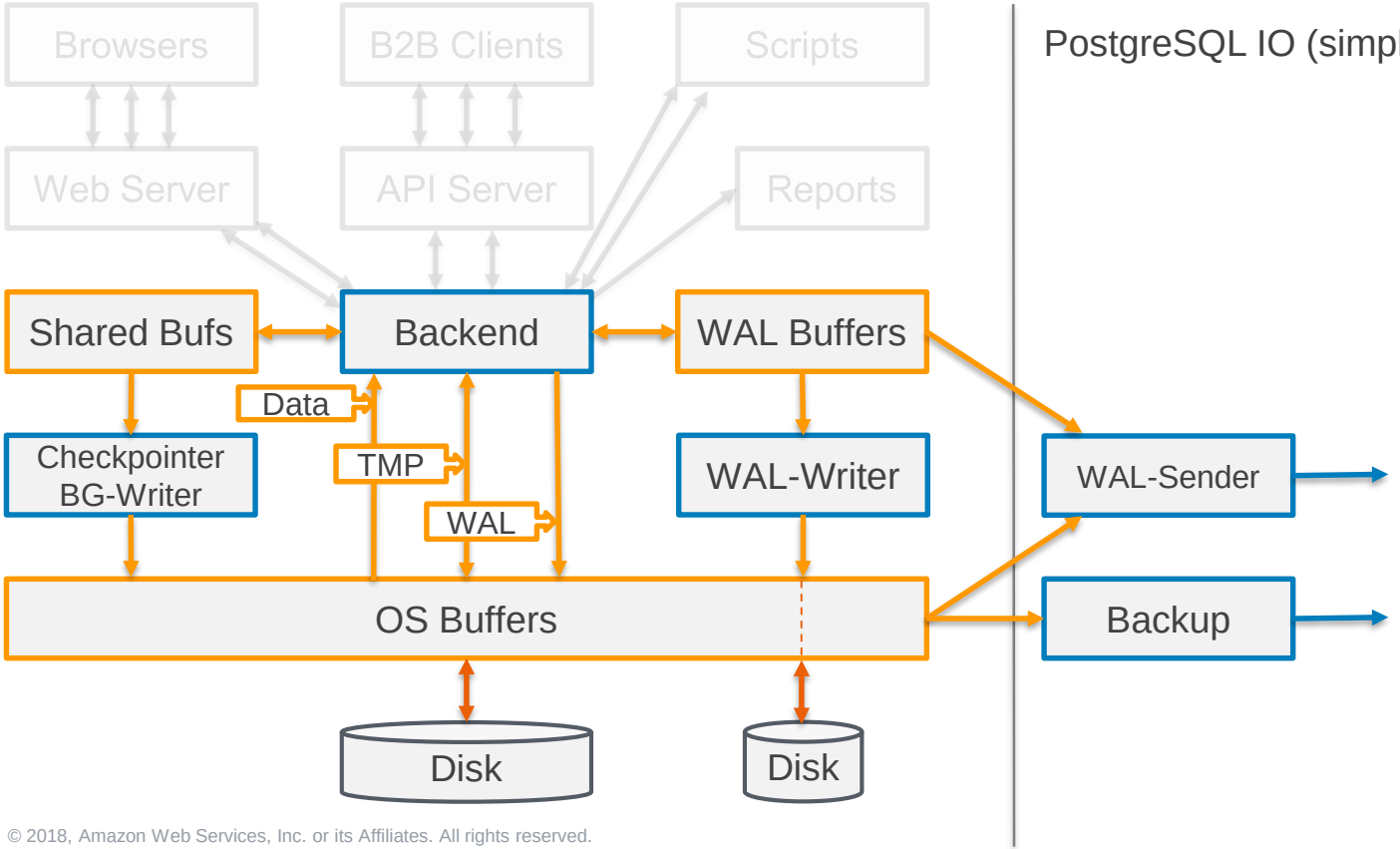


Due to its high fdatasync() rate (O_DIRECT on Linux) it is best to put WAL on a separate IO subsystem.

Even if that is not possible, having it on its own filesystem still has advantages (like that the database will start up when $PGDATA throws “no space left on device”).

aws

And all the rest …



aws