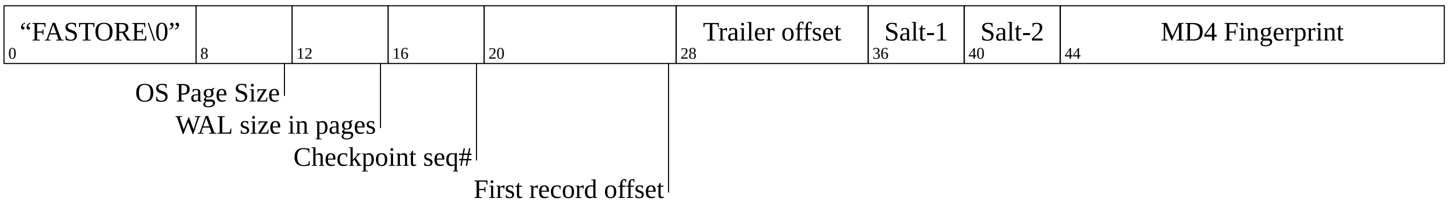
WAL Data Structures

Each WAL file is organized as succession of revision records in a ring. The status of the ring is maintained in two header records, located on pages 0 and 1, updated alternately in classic “ping-pong” fashion after each write to the file.

Recovery consists of reading both pages, determining which is more recent, seeking to the first record, and reading records sequentially until the EOF marker.

Both header records and all data records contain an MD4 checksum, giving the reader great protection against file corruption.

WAL Header Record



The first 8 bytes contain the **magic** string *FASTORE* with a trailing NULL. These serve to identify the file as a Fastore log file, and to denote the version of the file. For consistency with future versions the first four bytes should be considered constant (to identify the file type) and the last four bytes a version number.

The **OS Page Size** serves tells the reader where to look for the other copy of the header record. It is also a factor in computing the physical file size, allowing **WA: size in pages** to describe a file up to 4 “gigapages”, or 1 TB using a page size of 1 KB.

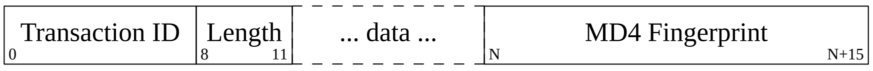
The **Checkpoint sequence number** begins with zero R and is incremented at every checkpoint. It wraps back to zero after 0xFFFF. The **First record offset** will likely be an `off_t`; it can be used with **lseek(2)** or, when the file is mapped into memory, with pointer arithmetic.

The **Trailer offset** gives the logical EOF for the file. Note that the **trailer offset** will be less than the **First record offset** if the file has wrapped around.

Salt-1 is a random integer incremented with each checkpoint, whereas **Salt-2** is a random integer regenerated with each checkpoint.

The **MD4 fingerprint** is computed over the whole header each time it is written. The random numbers help to prevent any mixing of a header with the wrong file.

WAL Page Header



(The WAL record header is provisional until the mechanism for checkpointing and obsoleting records is clarified.)

Each WAL record begins with a **Transaction ID** and a **Length** in bytes. The **MD4 Fingerprint** is computed over the bytes in the interval [0-N).