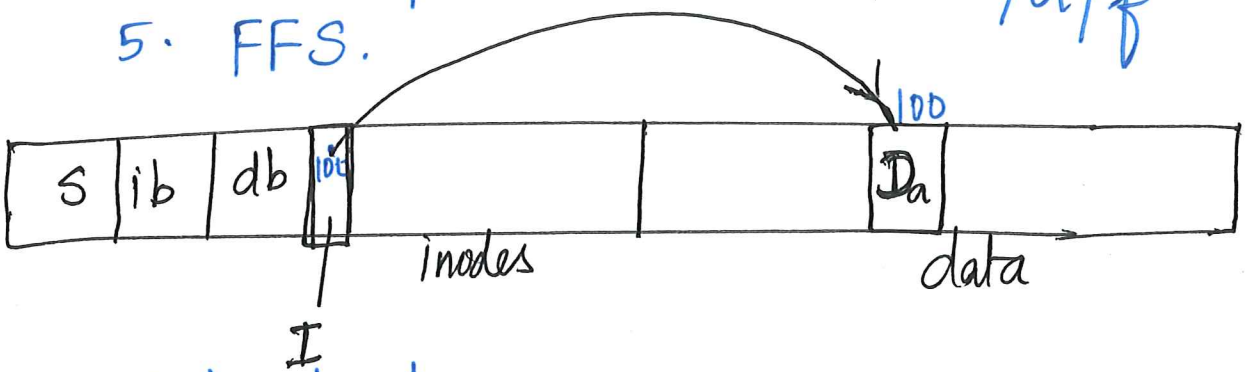
Crash Consistency

Review:

1. Disks
2. RAID
3. FS – files & dirs
4. FS Impl. (VSFS)
5. FFS.

/d/f



1. inode of d
2. data of d
3. ib
4. db

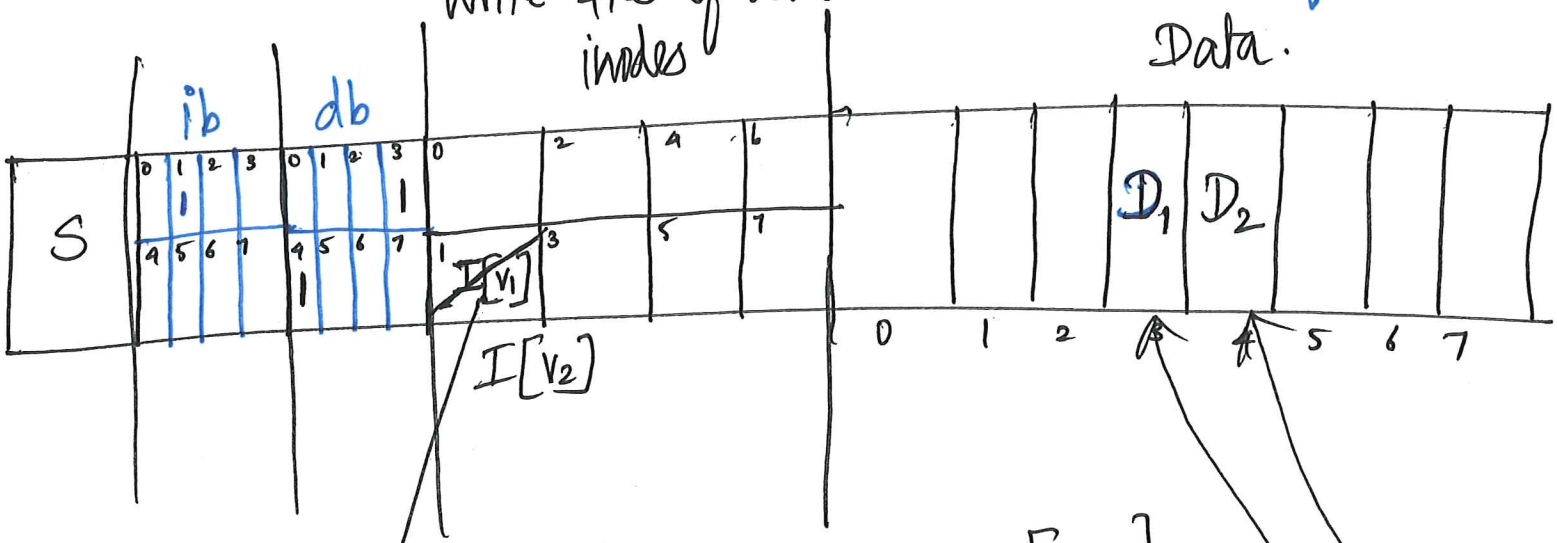
5. inode of f.
6. write Data block. (f)

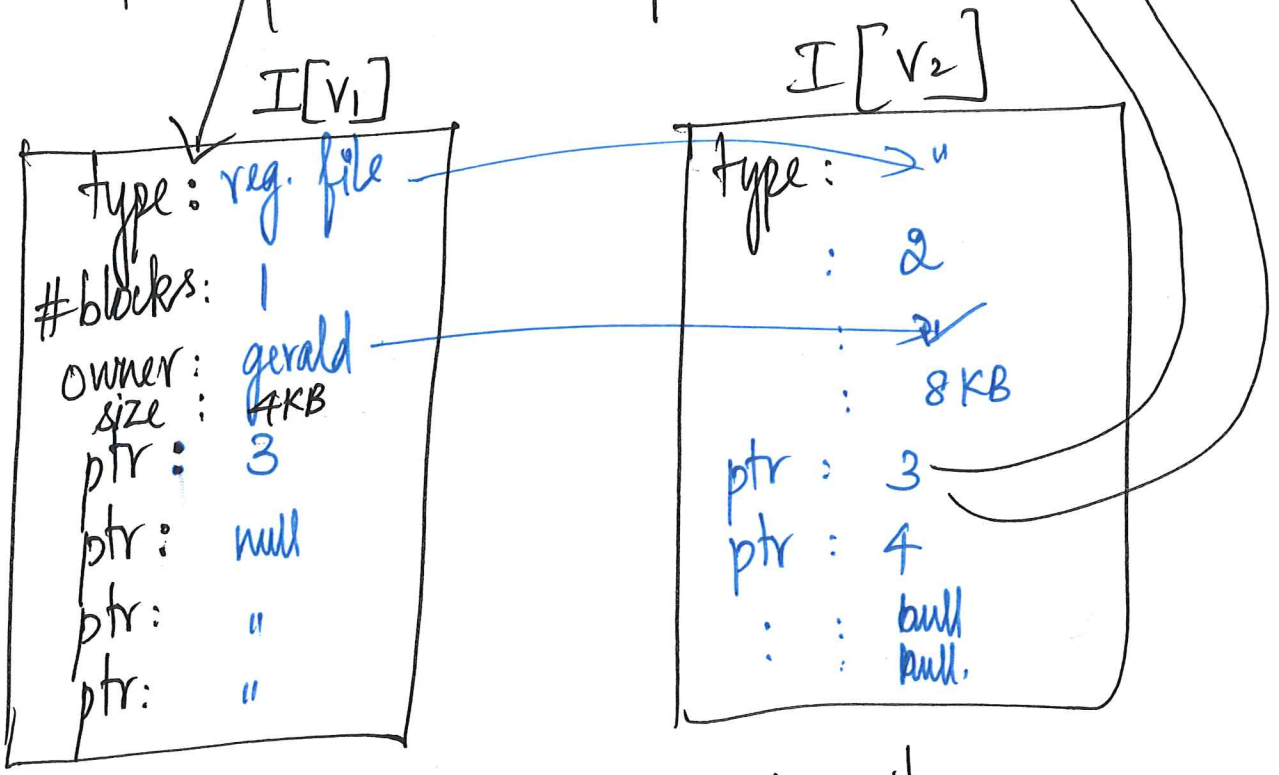
CRASH

Workload: Append to a file.

↓currently has 1 data block (full).

write 4KB of data





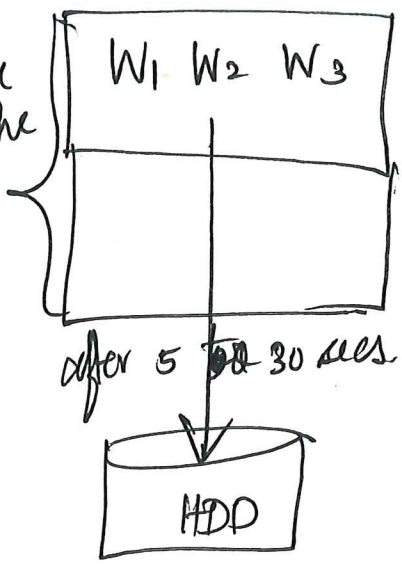
append()

1. open()
2. lseek() to the end.
3. write()
4. close()

writes

1. db[V1]: 0001 0000
   db[V2]: 0001 1000
2. I[V1] → I[V2]
3. D2.

CRASH

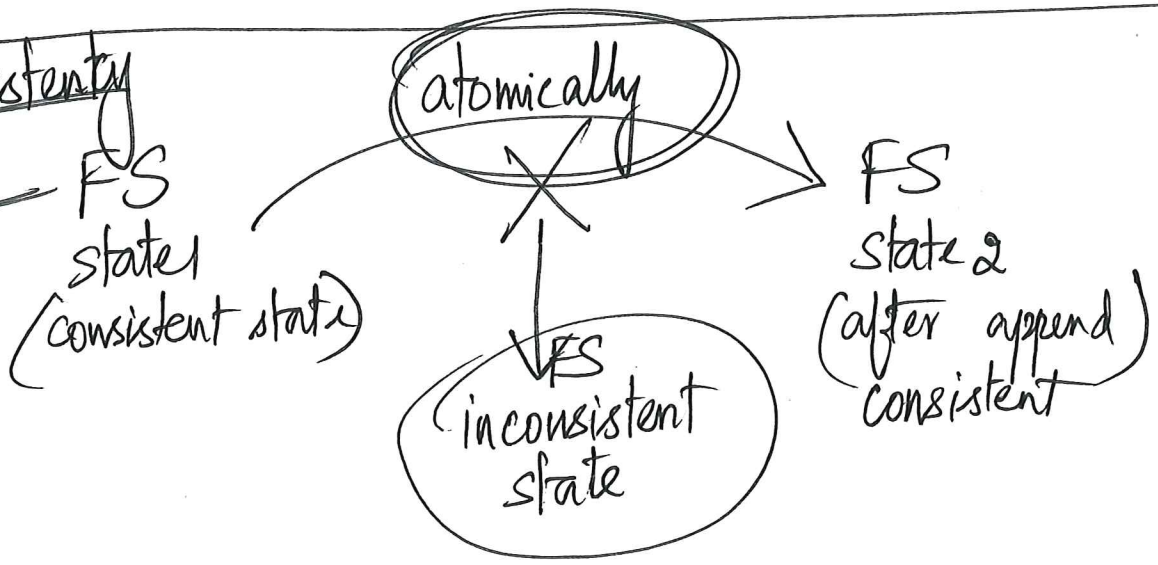


1. Write D_2 : – consistent from OS perspective
   ∵ Inode & db – OK
2. only Inode is written.
   NOT consistent
   ∵ inode – db ⇒ don't match.
3. only db is written.
   – inconsistent ≡ space leak
   ∵ inode – db – mismatch.

---

1. only $I[V_2]$ & D_2 are written.
   – inconsistent
2. " $I[V_2]$ & $db[V_2]$ are written.
   – consistent.
3. " D_2 & $db[V_2]$ are written.
   – inconsistent.

Crash consistency Problem



1. FSCK - File System Checker

→ no other FS operations should run.

→ fsck should run before mount the FS.

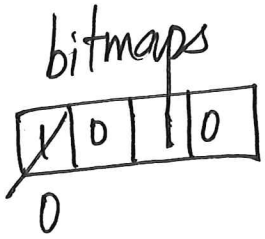
How fsck works?

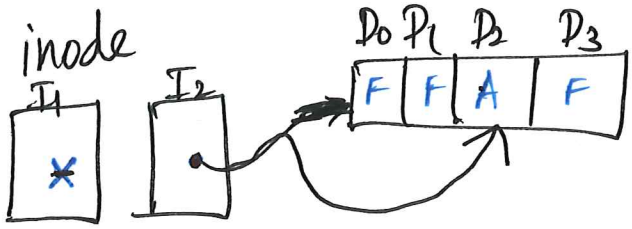
1. Detect a problem.
2. Recovery

---

1. **Superblock**
   - sanity check e.g. size of FS > # of blocks allocated

2. **Freeblocks**





3. **Inode state**

type: d / f / s / (K)

- remove inode.

4. **Inode Links**

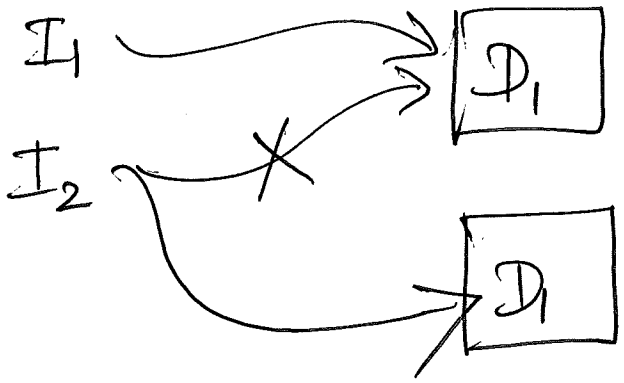
Links : 3

f 1004

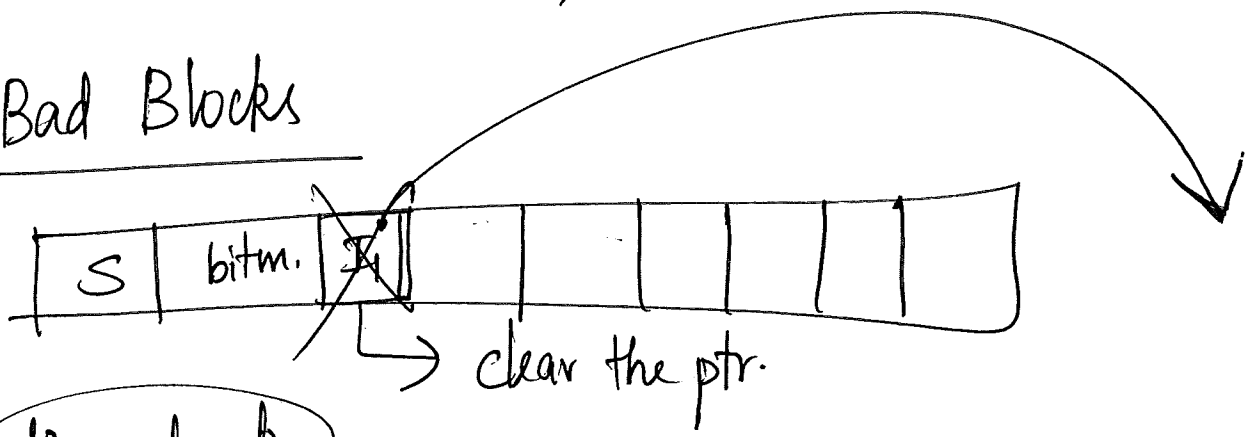
/d/g 1004

/h 1004

5. Duplicates



6. Bad Blocks



7. dir check

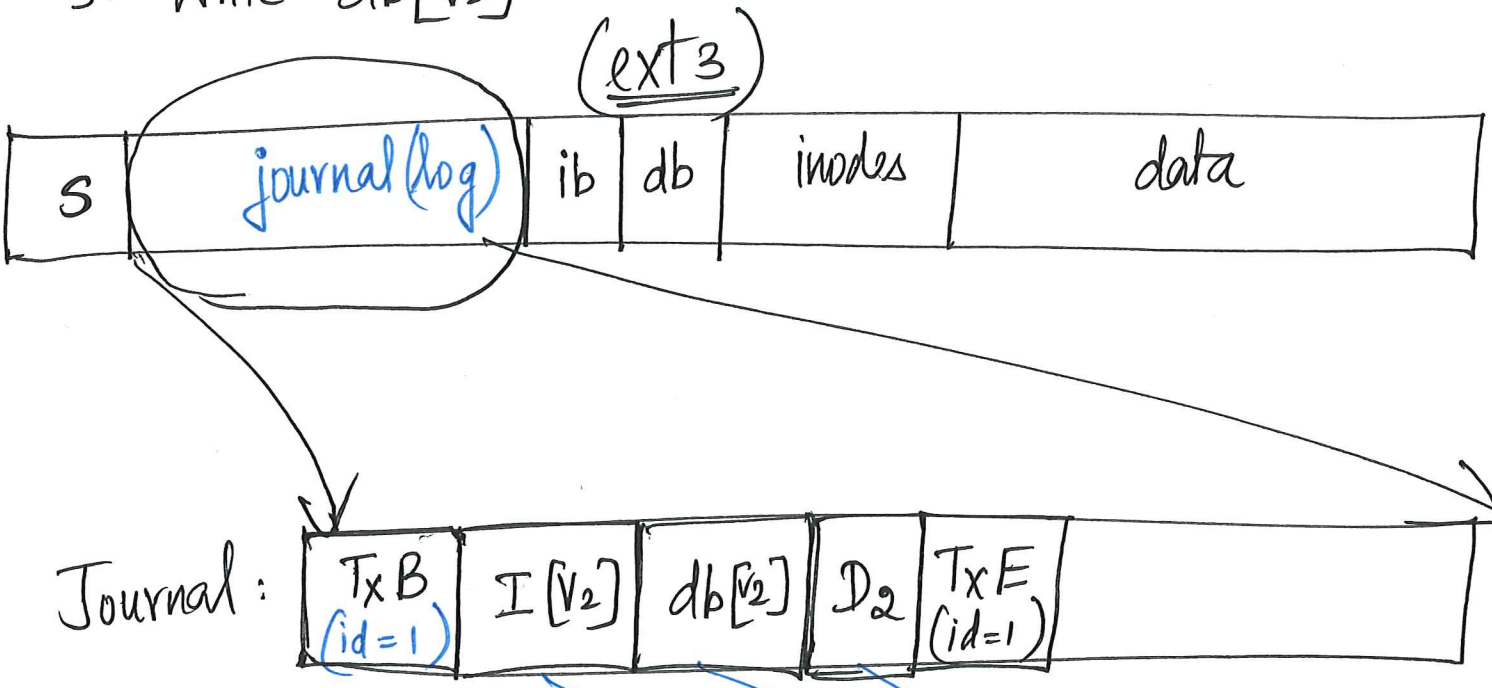
Human.	low-level name
.	2
..	4
f	10
g	16.

Problems?

Too Slow!

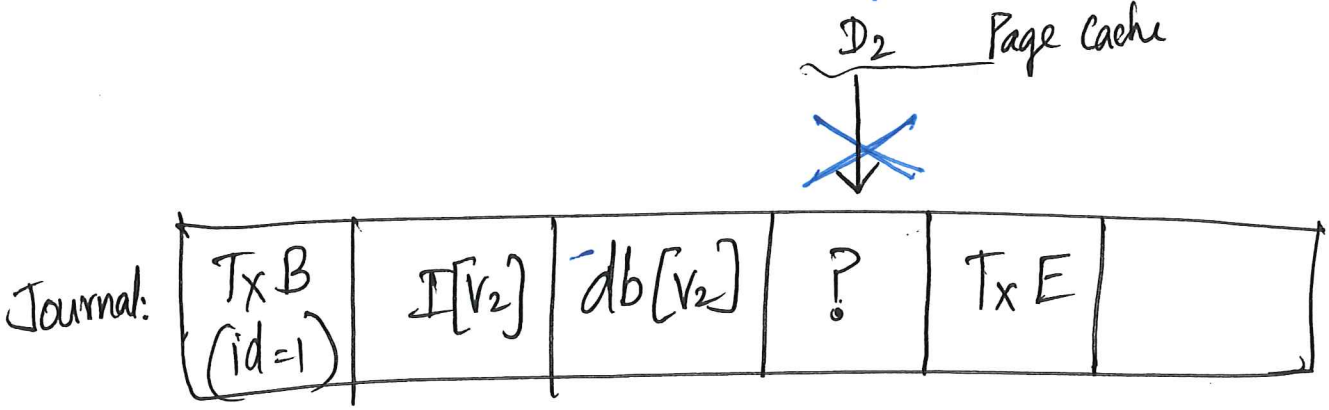
Journaling (Write-ahead Logging)

1. Write D_2
2. Write $I[v_2]$
3. Write $db[v_2]$



Protocol:

1. Journal Write (Write TxB, contents, TxE)
2. Checkpoint (Write $I[v_2]$, $db[v_2]$, D_2 to the final addr on the FS).



Data Journaling

Protocol

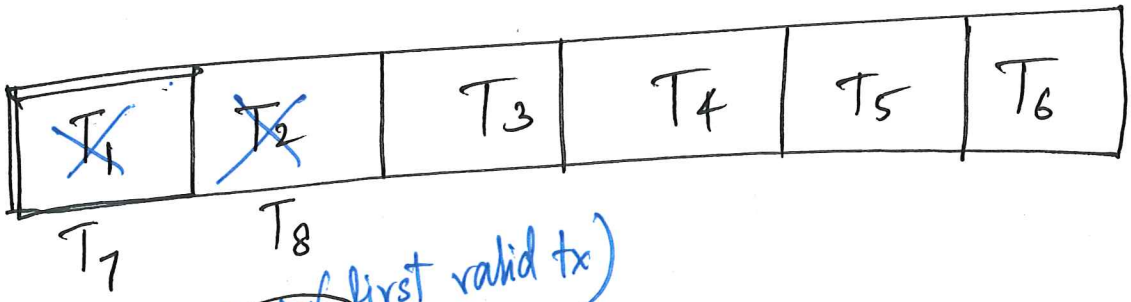
1. Journal Write (Tx B, contents). Wait! ~~this this~~
2. Journal Commit (Write Tx E). Wait.
3. Checkpoint. Wait.
4. Free the tx in the journal

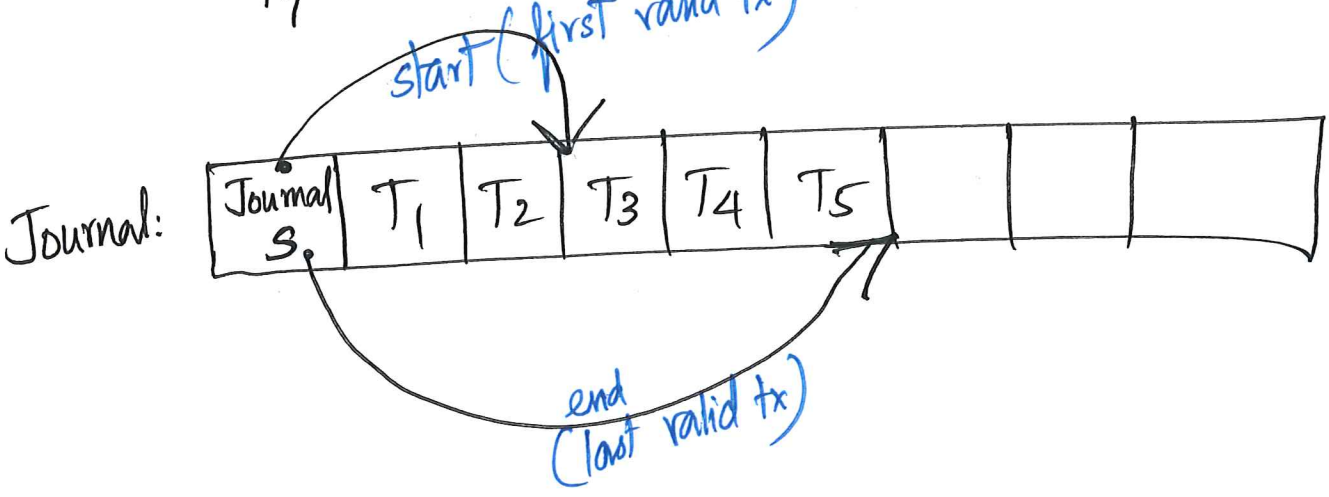
/d/f1 → wwwww

/d/f2 → wwwww

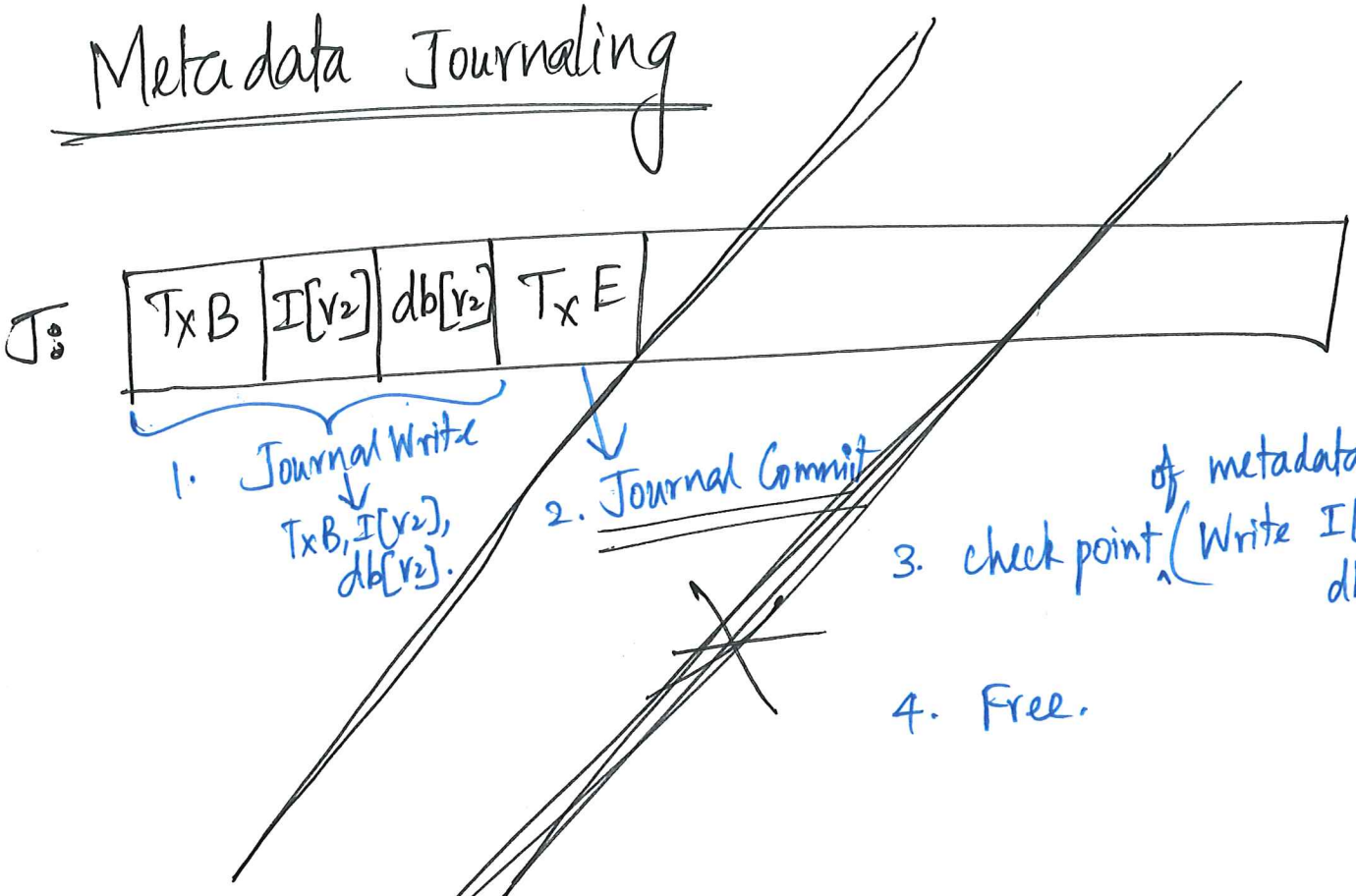
Page Cache - helps us.

circular log!





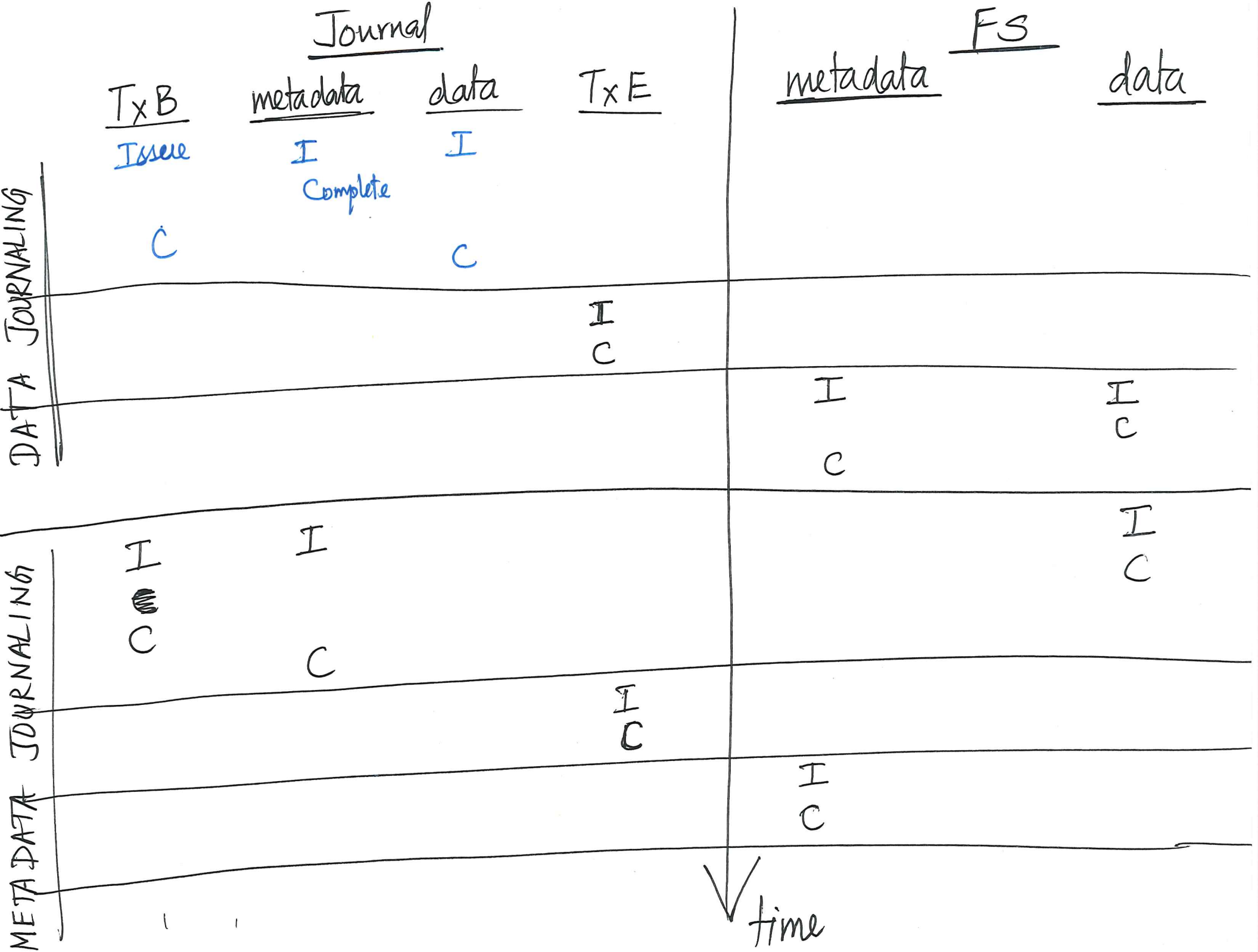
Metadata Journaling



Data - should be written before Journal commit.

Protocol

1. Write Data
2. Journal metadata Write — WAIT
3. Journal ~~checkpoint~~ Commit — WAIT
4. ~~Free~~ Checkpoint metadata. — WAIT
5. Free.


Journal
FS
TxB
metadata
data
TxE
metadata
data
Issue
Complete
I
C
DATA JOURNALING
METADATA JOURNALING
time