Frangipani: A Scalable Distributed File System
Thekkath, Mann, Lee
SOSP 1997

Why not primary copy?
  Actually they do use primary copy inside Petal
  But many Petal primary/backup pairs in one FS

Work out our own design...
  How to divide data among the network disks.

What does Petal do / guarantee?

What happens if a client e.g. creates a file?
  What steps does the server go through?
  acquire lock, append to *local* log, update local meta-data,
    release lock locally, reply to client.

What if a client on a different server reads that file?
  S1 gets the REVOKE
  writes log to Petal, writes meta-data to Petal, RELEASEs lock

Why must it write the log entry to Petal before writing the meta-data?
Why must it write the meta-data to Petal before releasing the lock?

What if two clients try to create the same file at the same time?

The locks are doing two things:
  Atomic multi-write transactions.
  Serializing updates to meta-data (cache consistency).

What if a server dies and it is not holding any locks?
  Can the other servers totally ignore the failure?

What if a server dies while holding locks?
  Can we just ignore it until it comes back up and recovers itself?
  Can we just revoke its locks and continue?
  What does Frangipani do to recover?

What's in a log record?

S1 creates f2, crashes while holding lock
  how does replay work?
  if S1 crashed before any flush of anything?
  mid-way through flushing log?
  mid-way through flushing data?
  just after all flushing, before releasing lock?
  just after releasing the lock?

What effect will the logging have on ordinary performance?

Suppose S1 deletes f1, flushes its block+log, releases lock.
  Then S2 acquires lock and creates a new f1.
  Then S1 crashes.
  Will recovery re-play the delete?
  Details depend on whether S2 has written the block yet.

Does the recovery manager have to acquire locks before playing records?
  What if some other server currently holds the lock?
  Might the other server have stale data cached? From before replay?

What if two servers crash at about the same time?
  And they both modified the same file, then released lock.
  How do we know what order to replay their logs in?
  I.e. can we replay one, then the other?
  Or must we interleave in the original order?

What if power failure affects all servers?

Suppose S1 creates f1, creates f2, then crashes.
  What combinations of f1 and f2 are allowed after recovery?

What if a server runs out of log space?
  What if it hasn't yet flushed corresponding blocks to Petal?

What happens if the network partitions?
  Could a partitioned file server perform updates?
  Serve stale data out of its cache?

What if the partition heals just before the lease expires?
  Could file server and lock server disagree about who holds the lock?

Why isn't the lock service a performance bottleneck?

What if a lock server crashes?

Why does their lock service use Paxos?

Why does Frangipani have a disk-like interface to Petal?
  Frangipani was never intended to use a disk, so no compatibility reason
  might some other interface work better?

Table 2: why are creates relatively slow, but deletes fast?

Why is figure 5 flat?
  Why not more load -> longer run times?