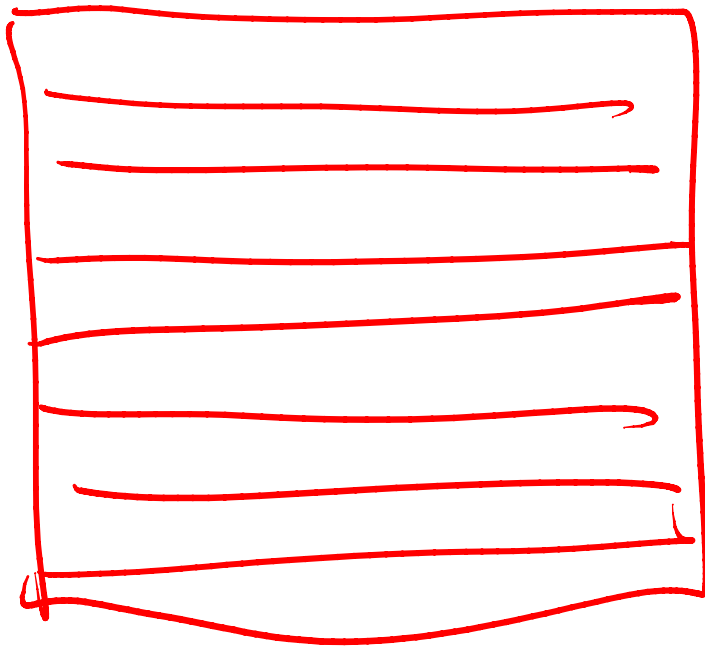
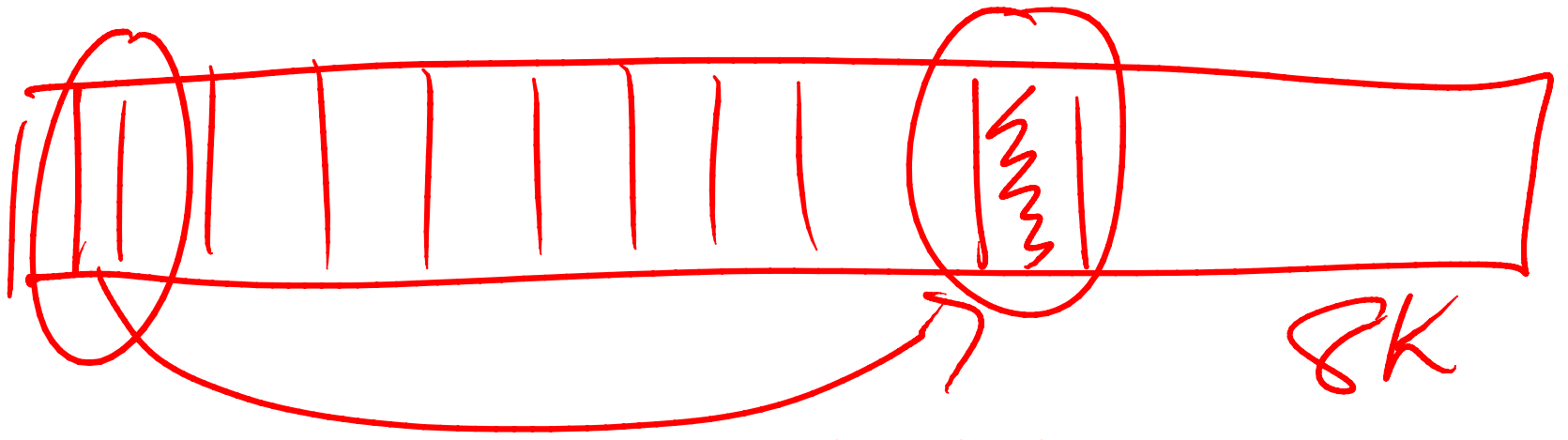




CACHE LINES

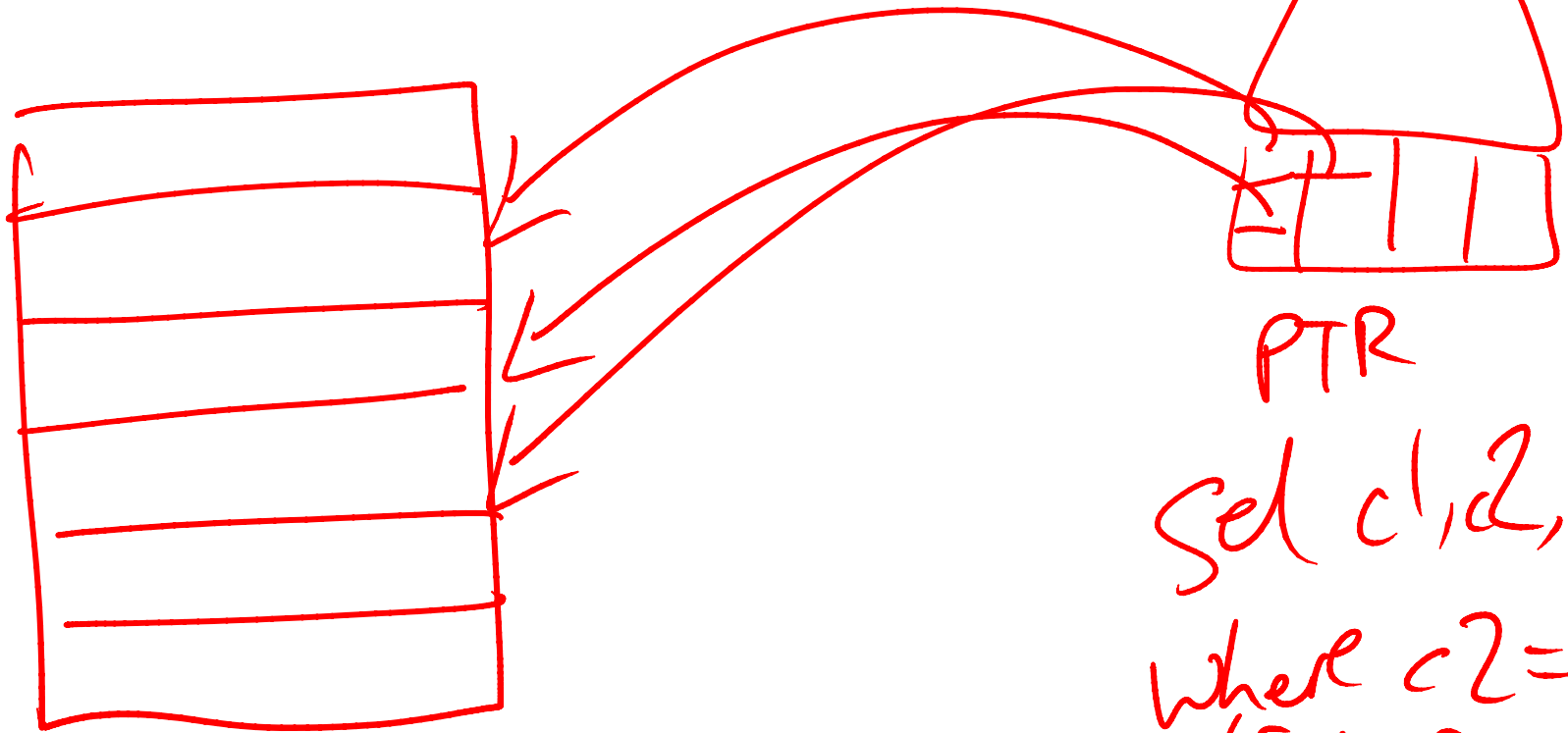
64b/128b

INVALIDATIONS

M1S06 Discussing why the NULL
bitmap is a perf optimization by
avoiding having to read record
parts into the CPU's L1 cache

$c_1, c_2, \dots, c_{10}$

NCI c_2

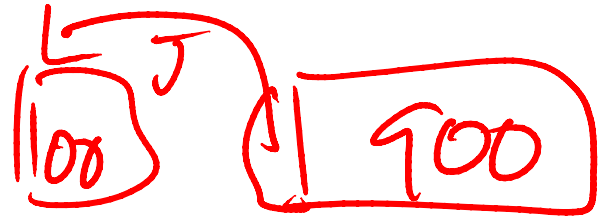
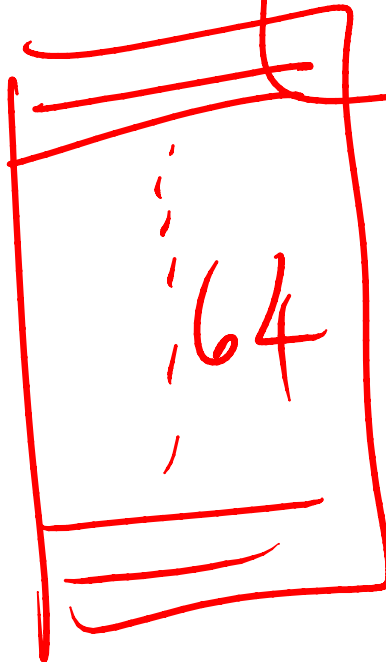
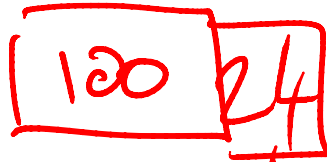
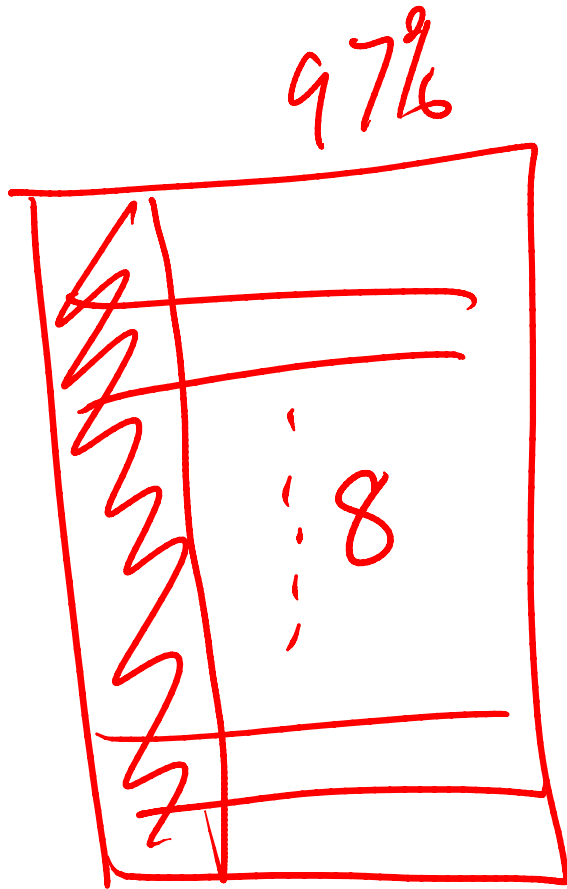


sel c_1, c_2, c_3

where $c_2 = X$

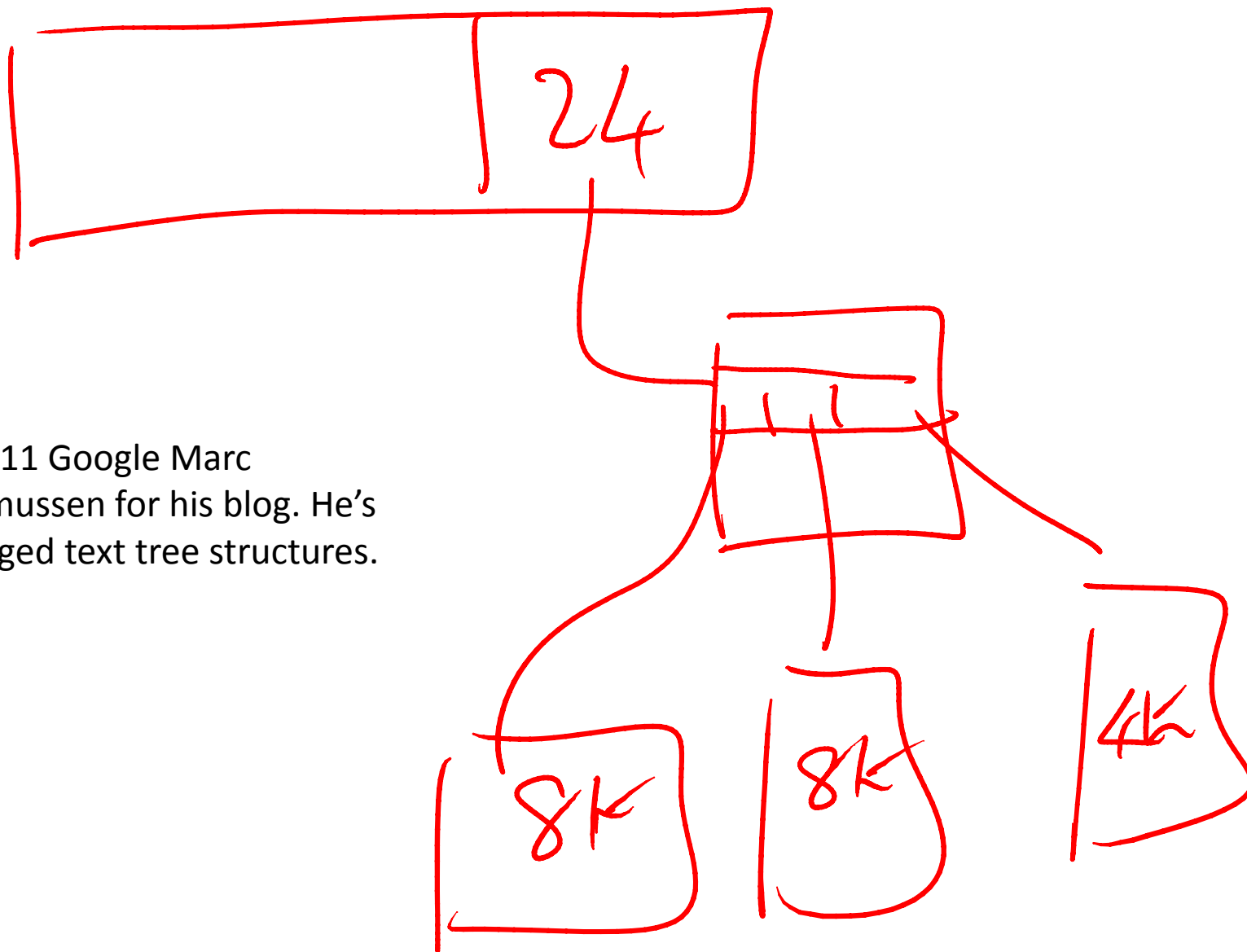
HEAP PTR = PHYSICAL RID (File, Page, Slot)
CL PTR = LOGICAL RID (cluster keys)

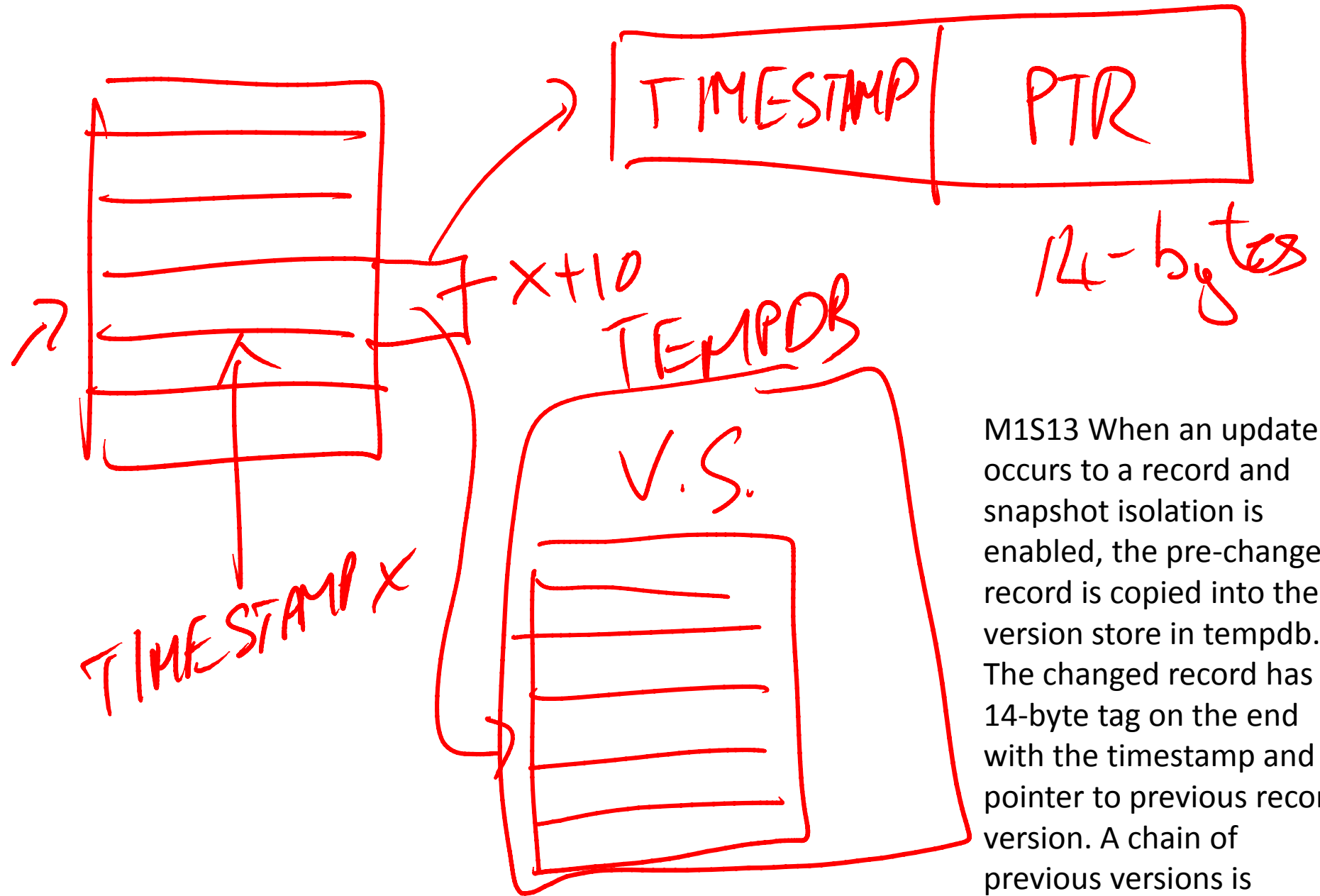
M1S09 Explaining NC index back-links to the data records as part of describing why forwarding/forwarded records exist. Without them, moving a heap row would entail changing all NC index records with a back-link to that row.



M1S12 Having a large LOB value in row that's not used much makes for large rows and low data density. Choose how to store LOB data wisely, either in-row, off-row, or in another table with a join.

M1S11 Google Marc
Rasmussen for his blog. He's
blogged text tree structures.

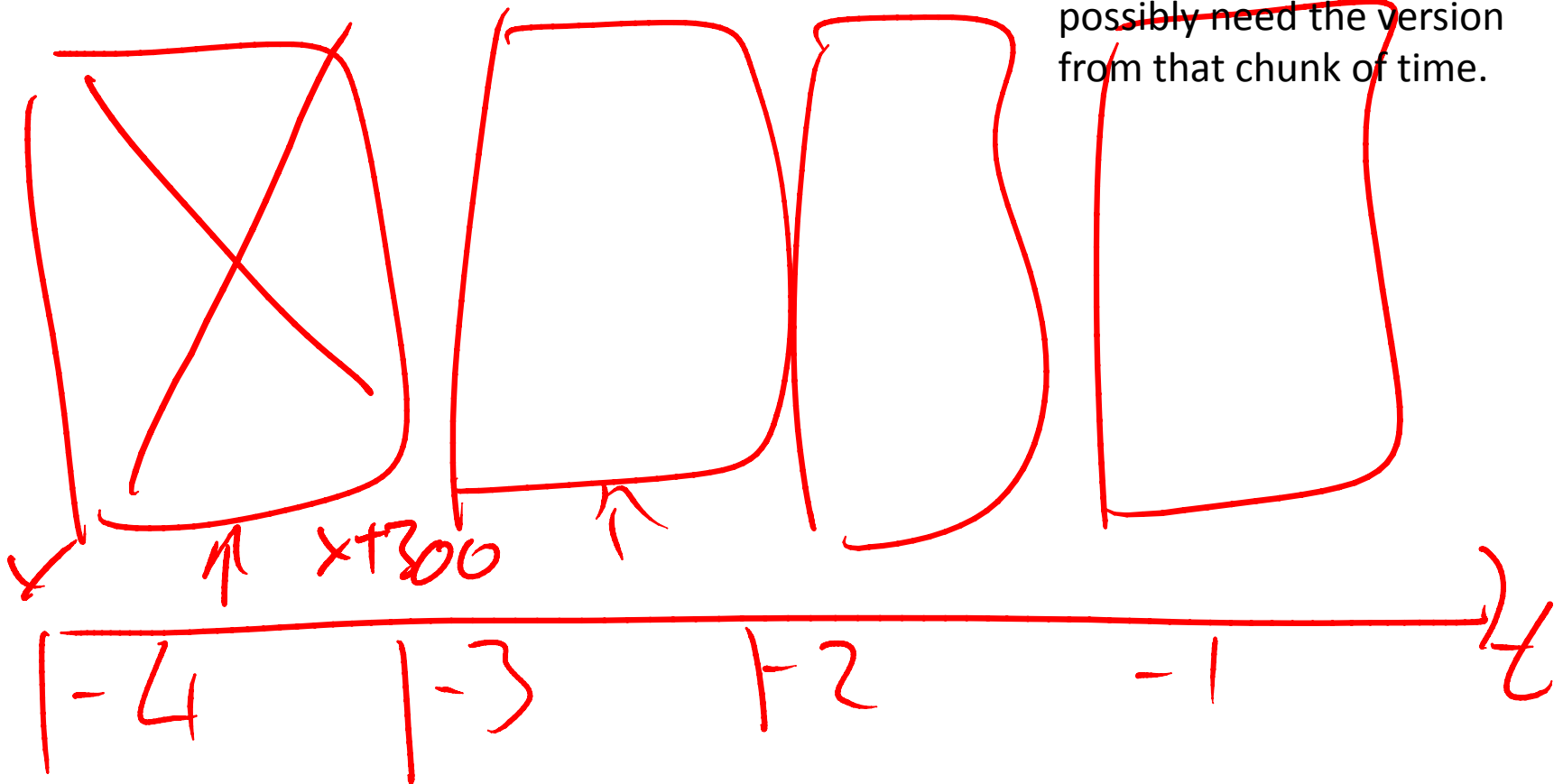


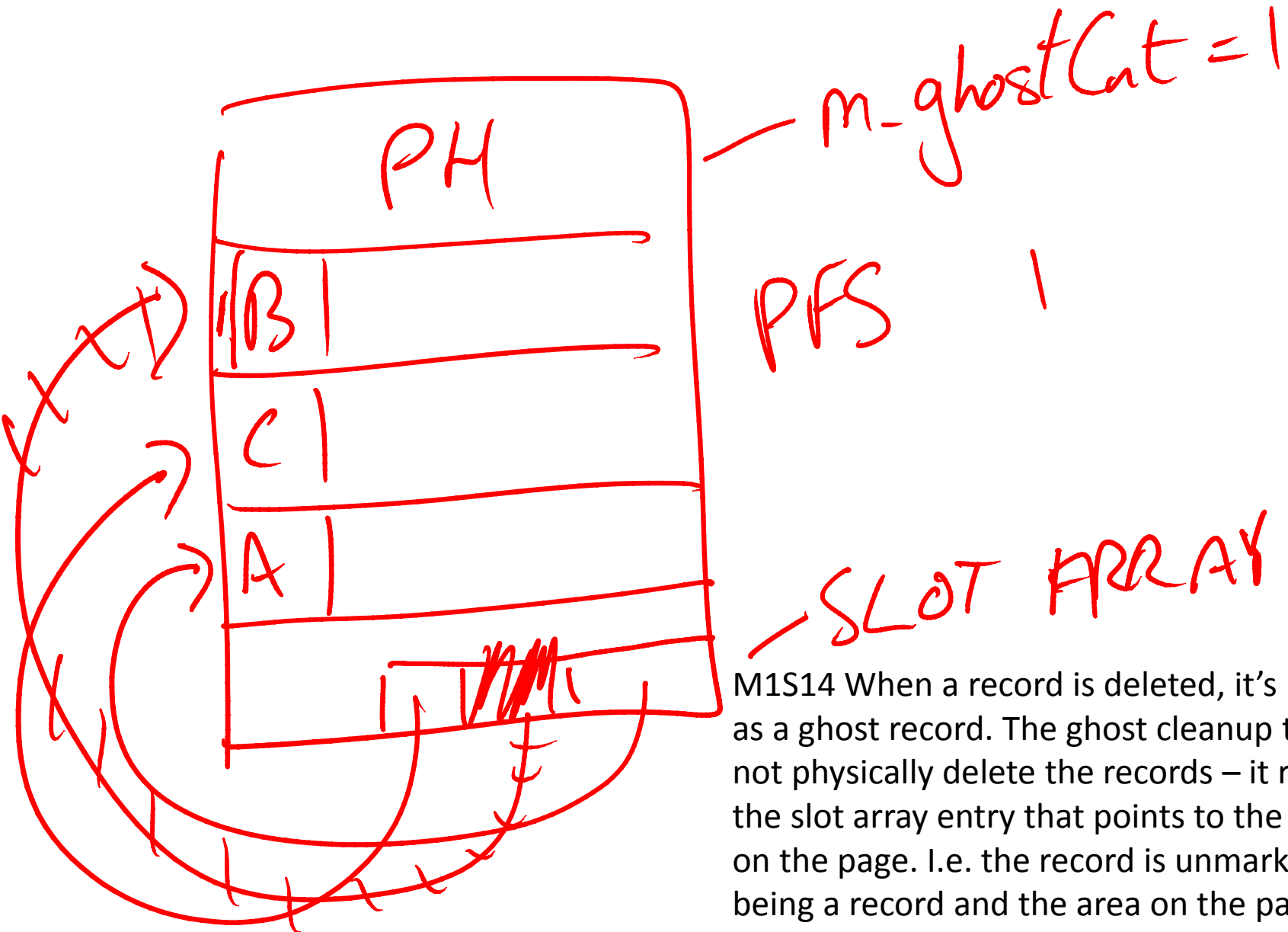


M1S13 When an update occurs to a record and snapshot isolation is enabled, the pre-change record is copied into the version store in tempdb. The changed record has a 14-byte tag on the end with the timestamp and pointer to previous record version. A chain of previous versions is possible.

APPEND-ONLY ALLOCATION UNIT

M1S13 & M2S54 Explaining how there's a new portion of the version store created every minute and these portions can only be removed by the background cleanup task when no running queries could possibly need the version from that chunk of time.



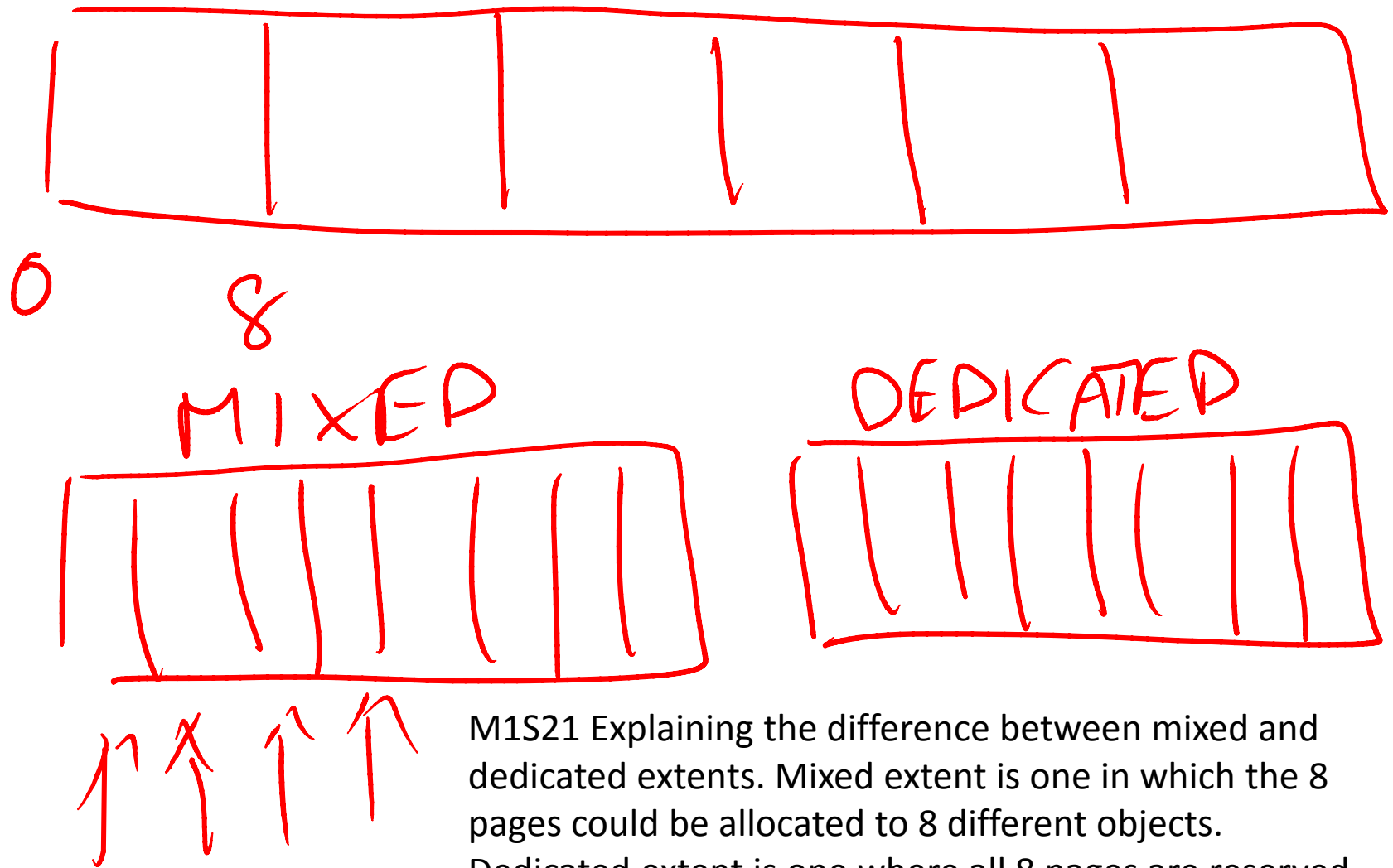


M1S14 When a record is deleted, it's marked as a ghost record. The ghost cleanup task does not physically delete the records – it removes the slot array entry that points to the record on the page. I.e. the record is unmarked as being a record and the area on the page is now free space – that just happens to have bits and bytes that used to be a valid record.

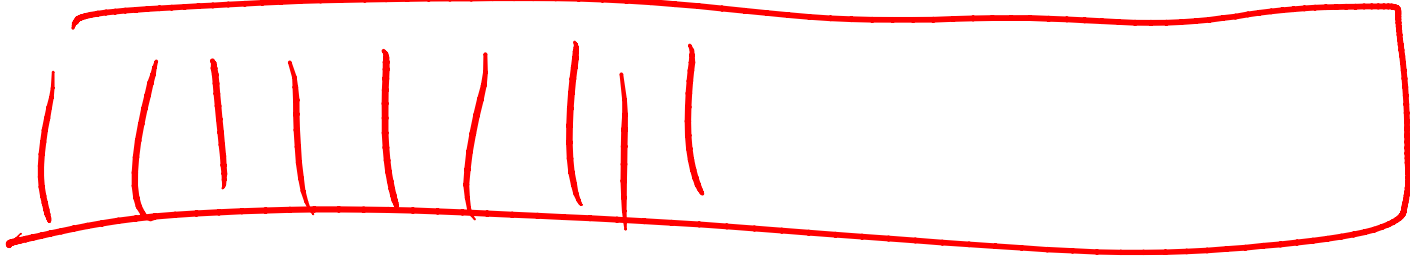
DBCC TRACEON
(662, -1)

LOP_EXPUNCE_GHOST

M1S14 How to enable the trace flag globally to watch the ghost cleanup task. You may also need to enable 3605 as well. That one allows debug traceprints to go to the error log.

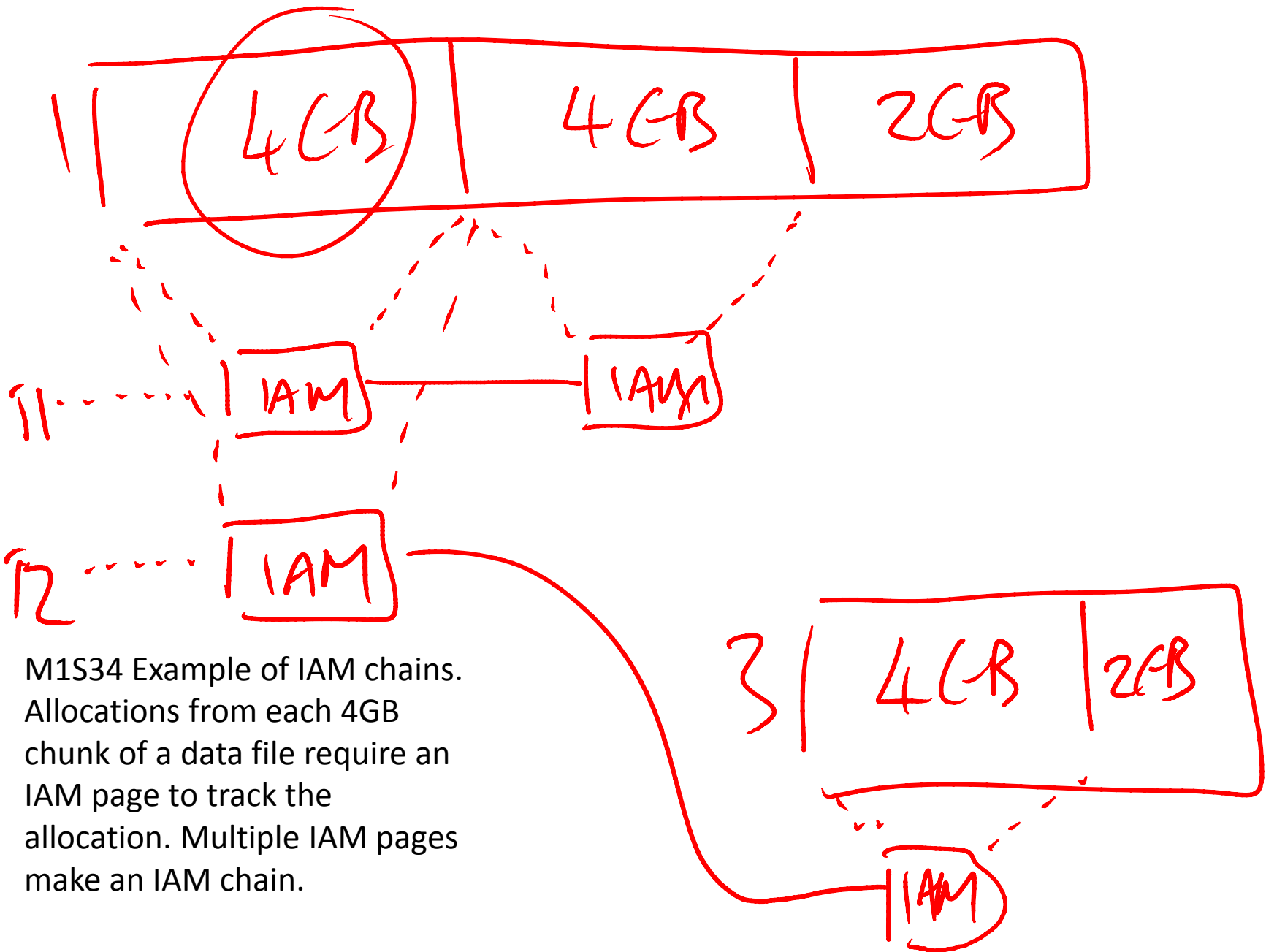


M1S21 Explaining the difference between mixed and dedicated extents. Mixed extent is one in which the 8 pages could be allocated to 8 different objects. Dedicated extent is one where all 8 pages are reserved for exclusive use of a single object. Later in the deck I refine 'object' to really be 'allocation unit'.



0 - FH	4	} - UNUSED
1 - PFS	5	
2 - CAM	6	- DIFF
3 - SCAM	7	- MZ

M1 The page types in the first extent in a data file



M1S34 Example of IAM chains.
Allocations from each 4GB
chunk of a data file require an
IAM page to track the
allocation. Multiple IAM pages
make an IAM chain.

1TB

SALES

2011

2015

M2S6 Splitting a large database up into multiple filegroups can enable you to do small, targeted restores or even a partial restore from scratch – in the event that the database is damaged or destroyed.



WA 3

CR 4

PARIS 1

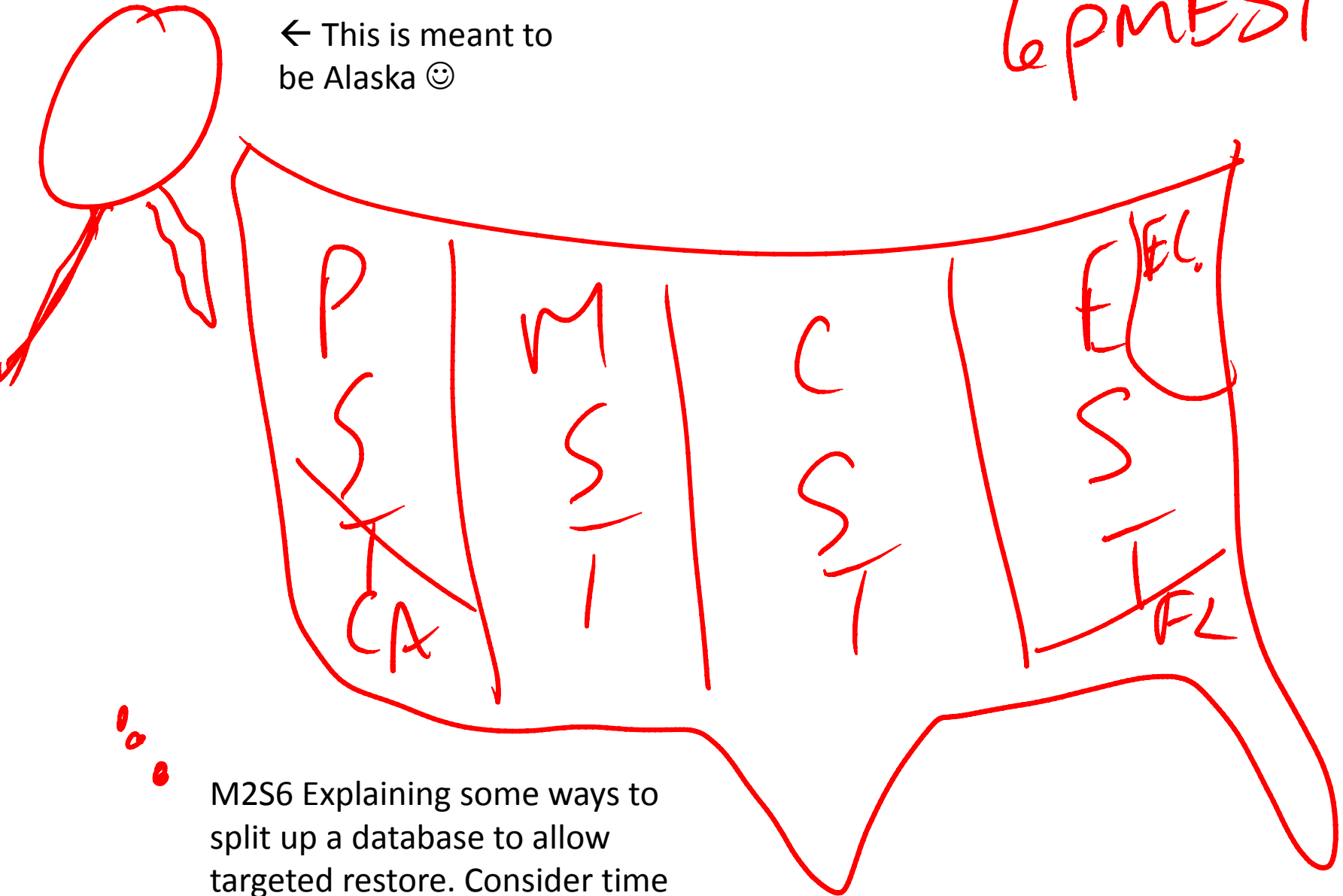
CA 2

ID 5

M2S6 Explaining some ways to split up a database to allow targeted restore. Consider time zone, population density, or other demographics.

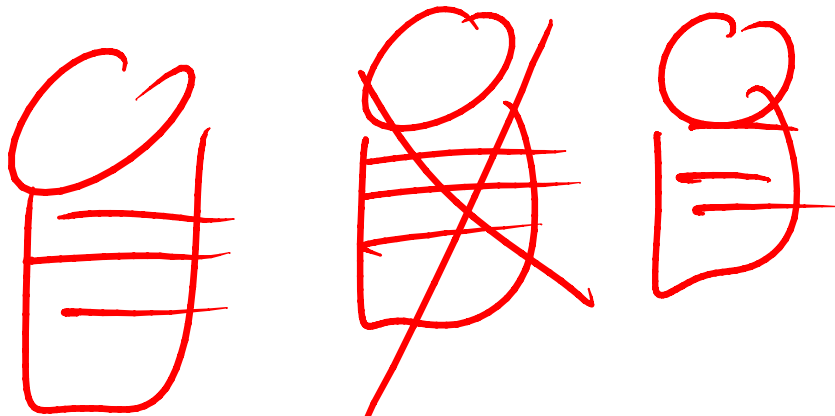
6 PM EST

← This is meant to
be Alaska ☺



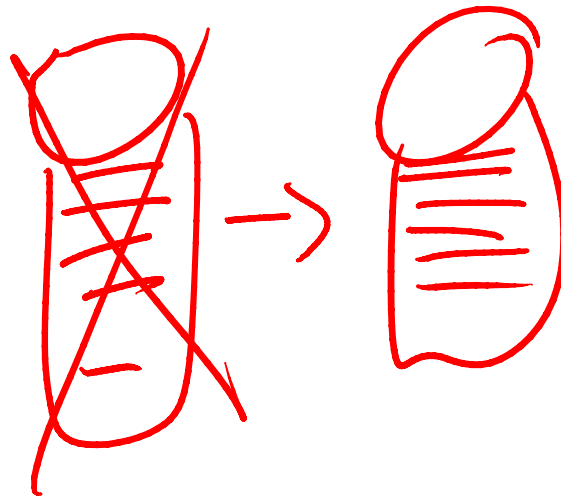
M2S6 Explaining some ways to
split up a database to allow
targeted restore. Consider time
zone, population density, or other
demographics.

RAID 0 STRIPING



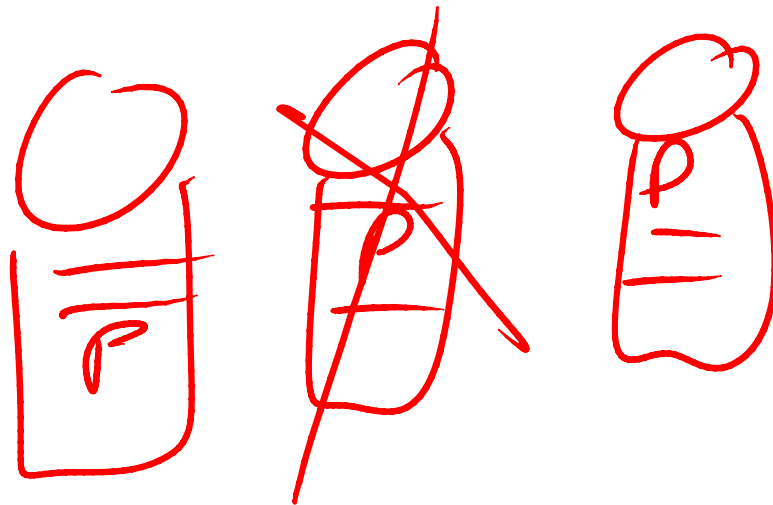
http://en.wikipedia.org/wiki/Standard_RAID_levels

R-1
MIRRORING



http://en.wikipedia.org/wiki/Standard_RAID_levels

R-5 ROTATING PARITY



http://en.wikipedia.org/wiki/Standard_RAID_levels

R-10

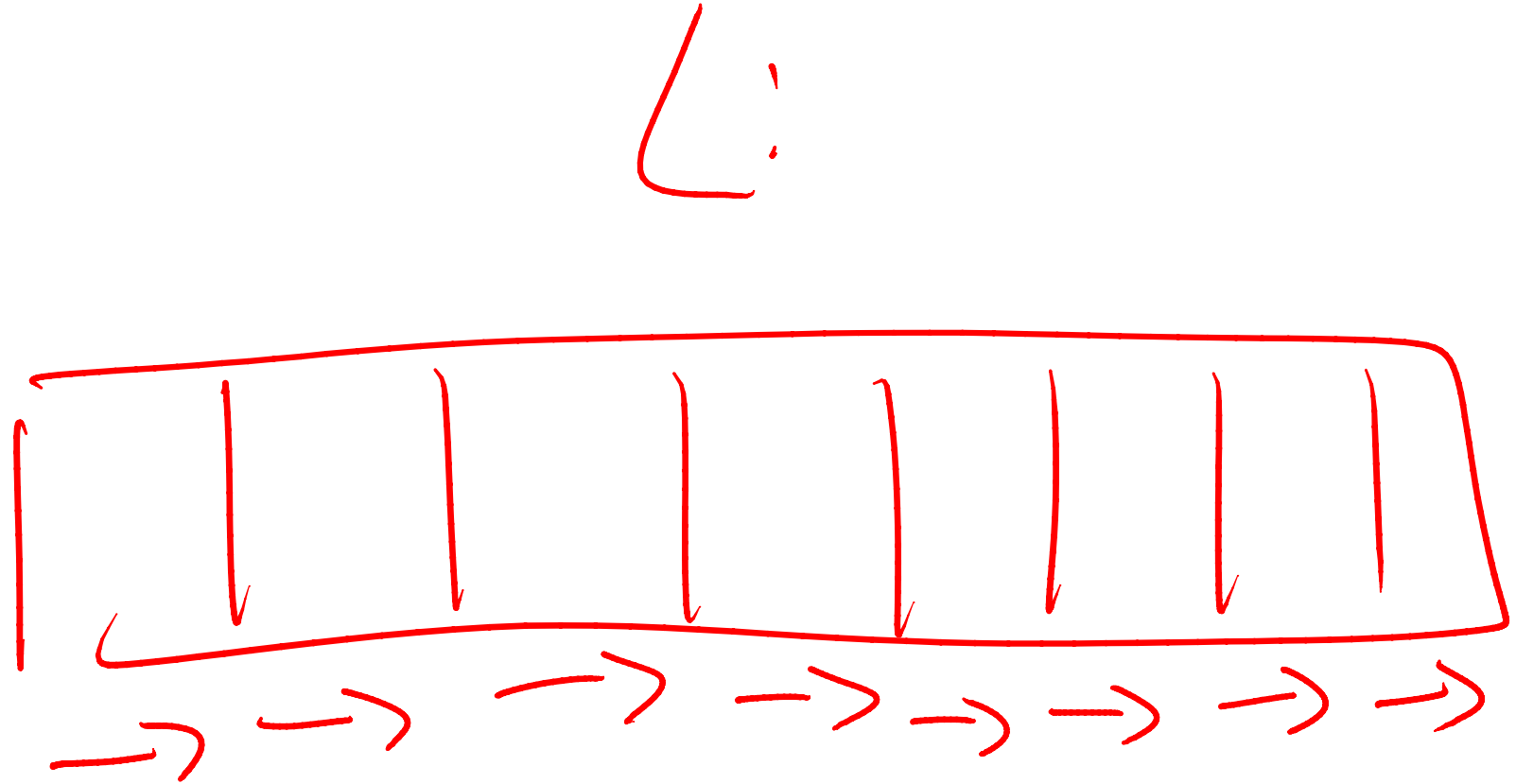
MIRRORING + STRIPING



R1

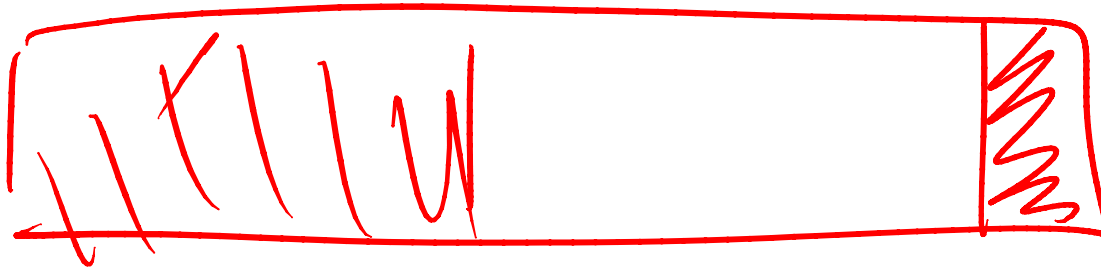
R1

http://en.wikipedia.org/wiki/Standard_RAID_levels

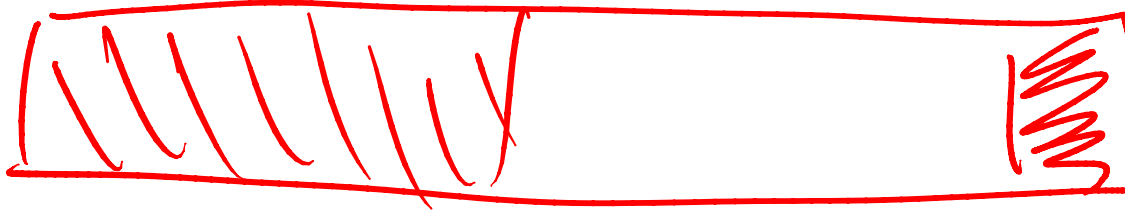


M2S12 If you put all your log files on one volume, its I/O characteristics become random. This may or may not become a performance problem, depending on the capabilities of the I/O subsystem.

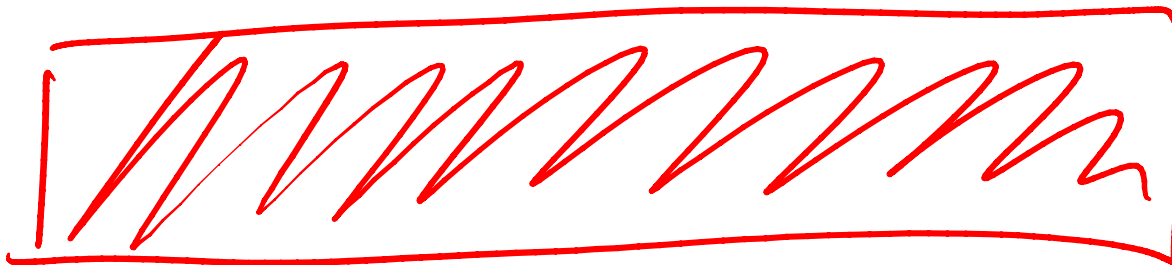
WEIGHTING



~~8164~~ ~~8163~~ 8162
16 ALLOCS



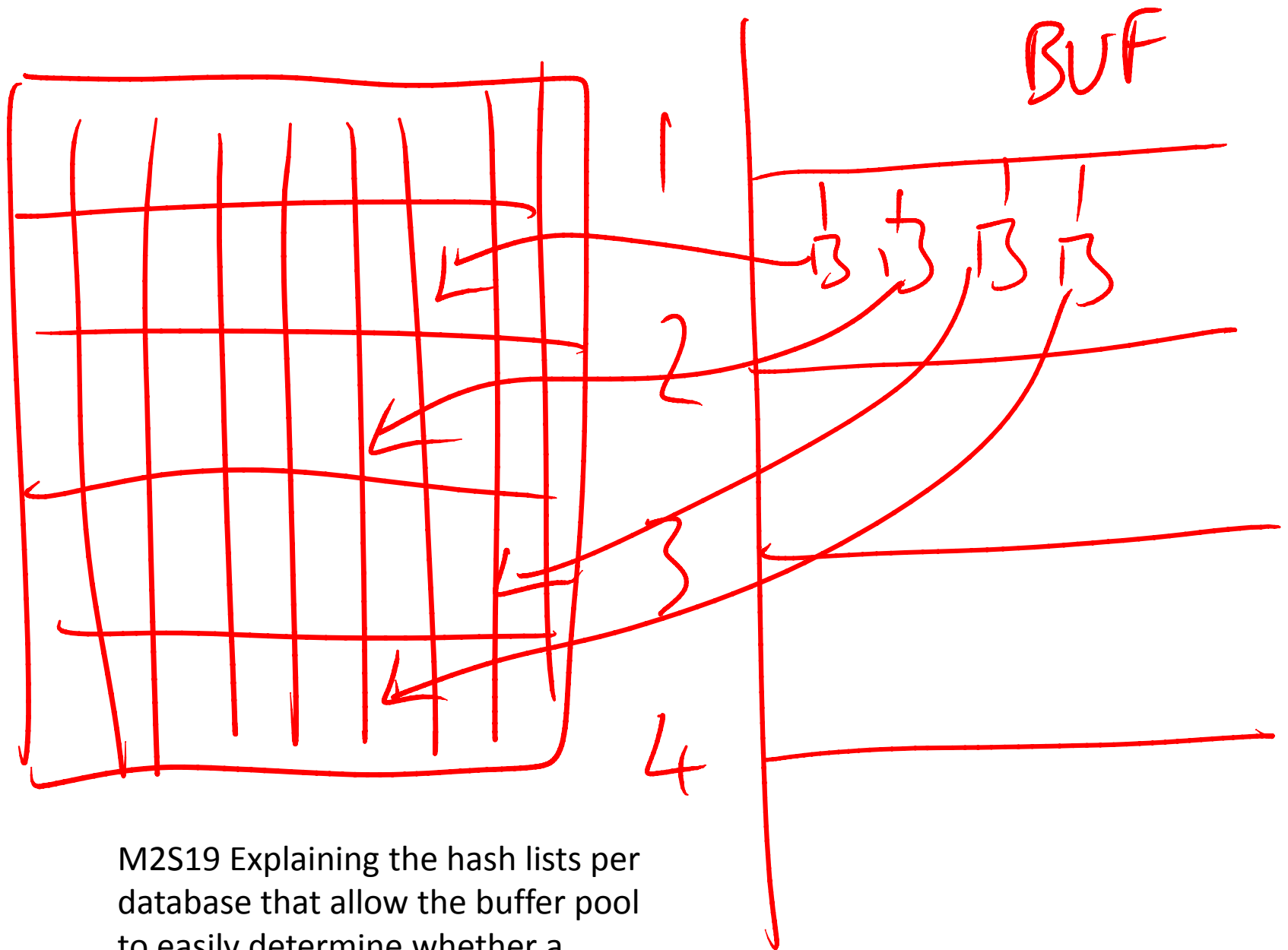
~~8164~~ ~~8163~~ 8162
16 ALLOCS



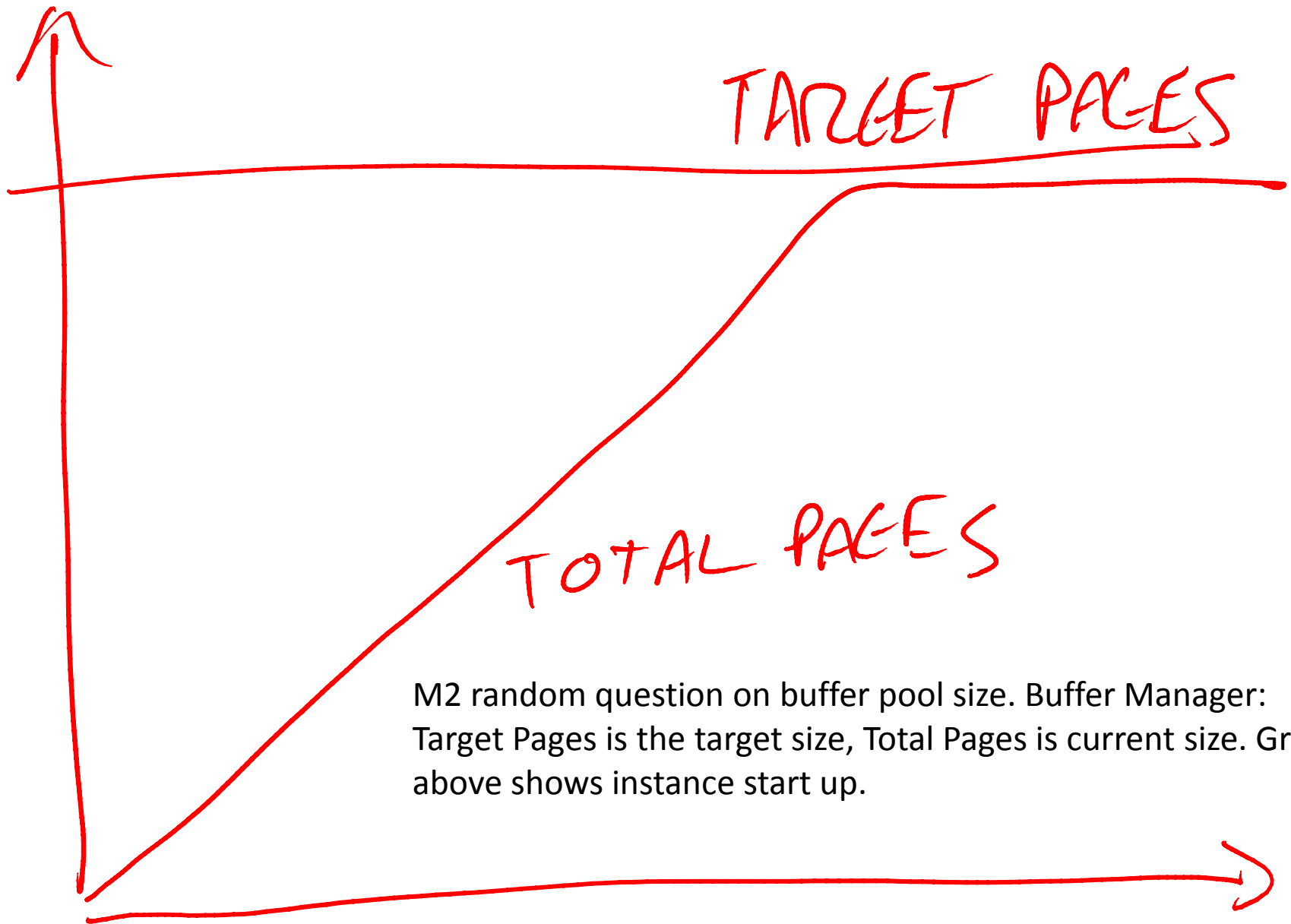
~~8164~~ 1 ✓ ✓

8164 ALLOCS

M2S18 Explaining round-robin allocation and proportional fill. Each file has a proportional fill weighting. Allocations happen proportionally more frequently from files with more free space.



M2S19 Explaining the hash lists per database that allow the buffer pool to easily determine whether a page is already in memory or not

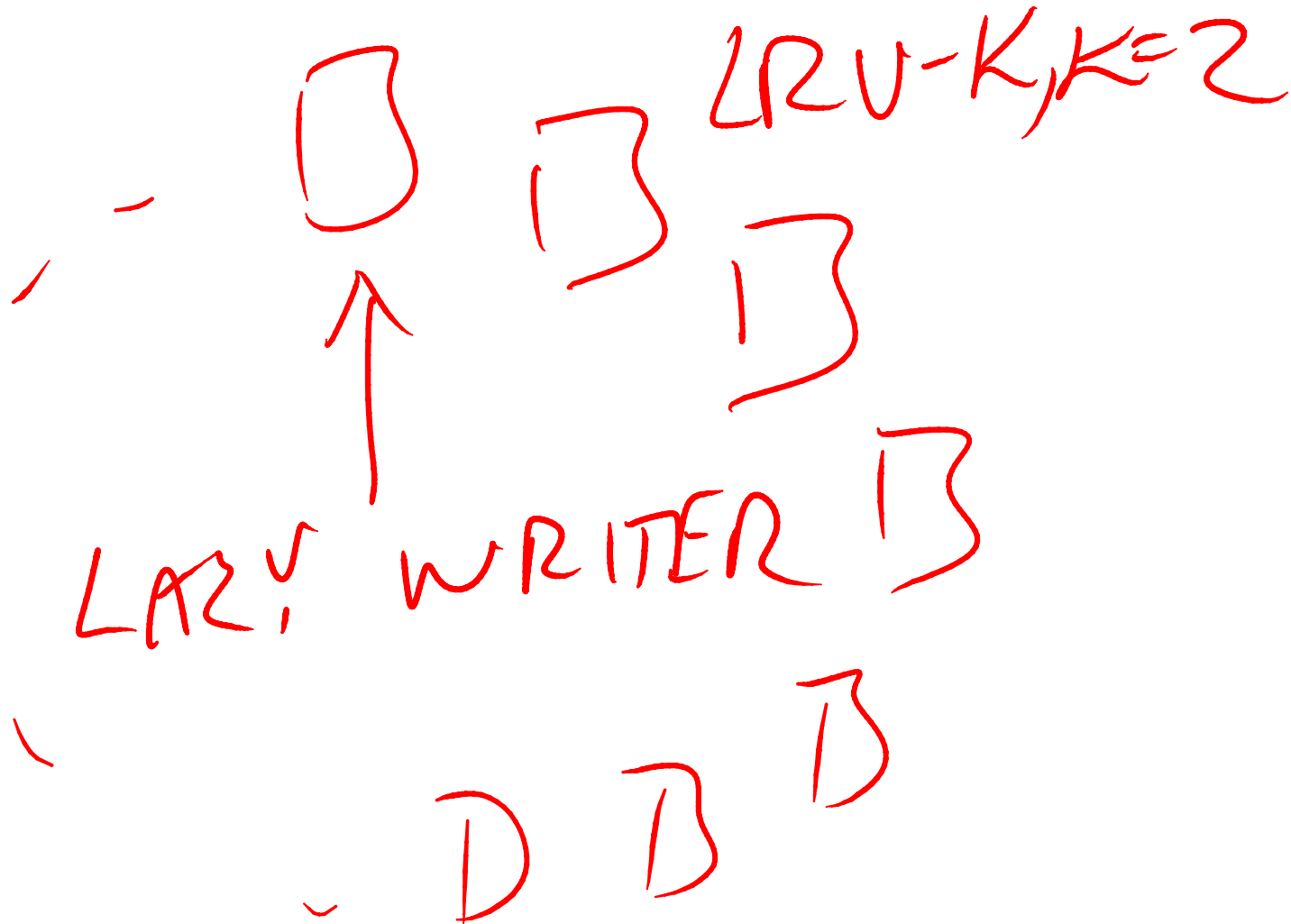


M2 random question on buffer pool size. Buffer Manager:
Target Pages is the target size, Total Pages is current size. Graph
above shows instance start up.

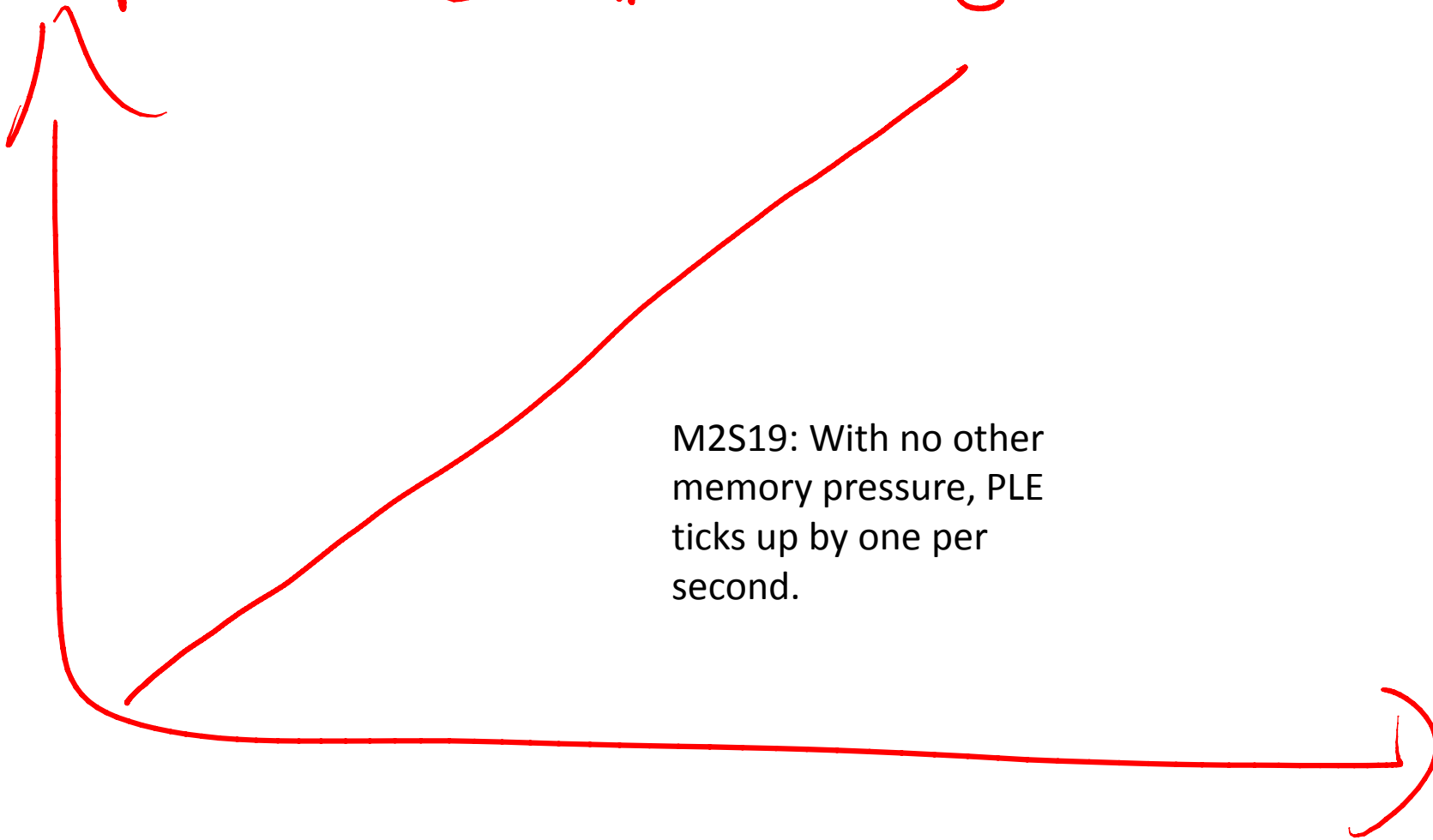
P.L.E. = 300

M2S19: PLE (page life expectancy) of 300 is an old, old threshold. PLE is a measure of how long a page will live in the buffer pool, in seconds.

The buffer pool implements a Least Recently Used algorithm, based on last two times the page was accessed. Done by the Lazy Writer background task.

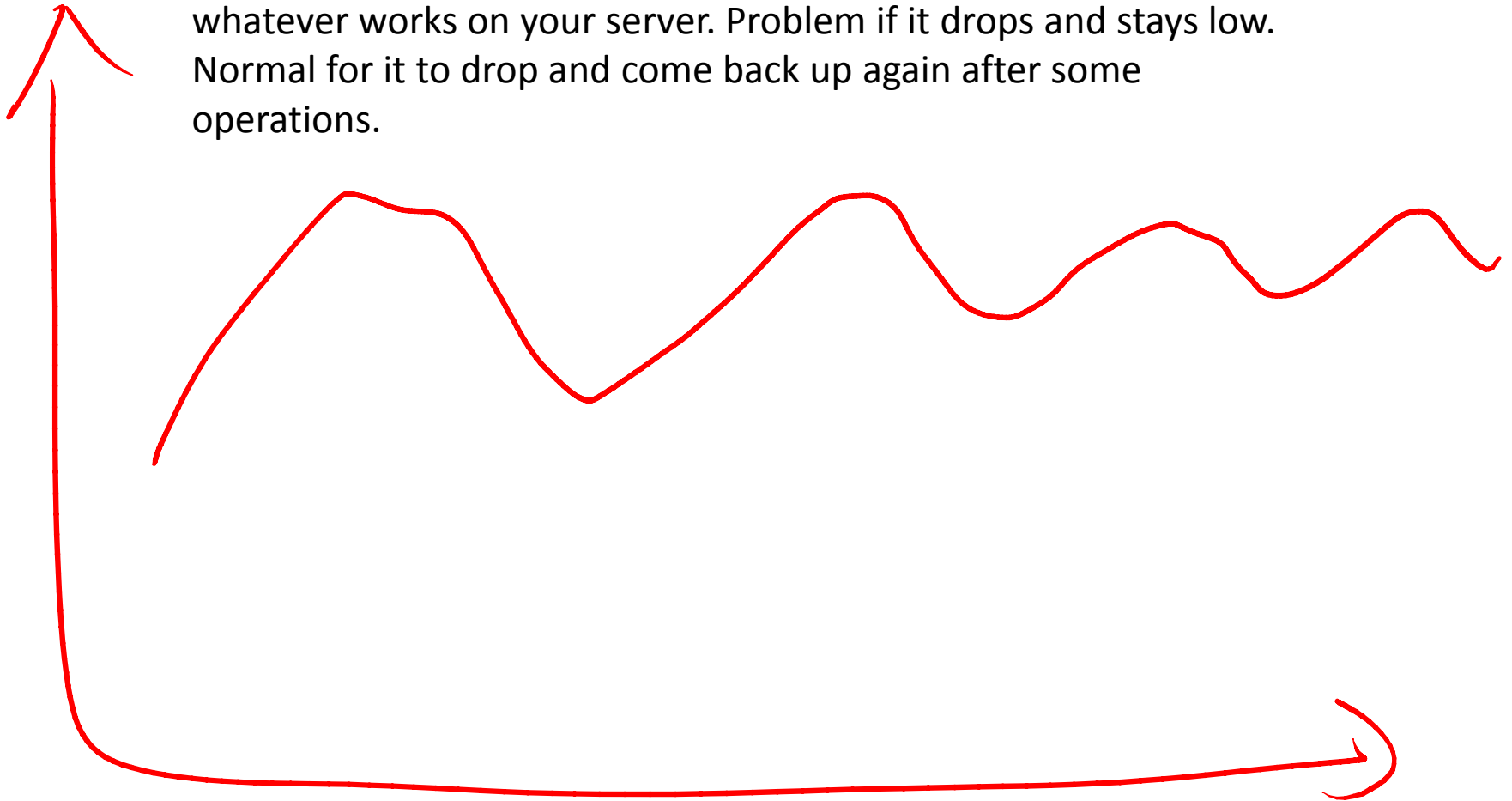


PLE (Buffer Manager PO)

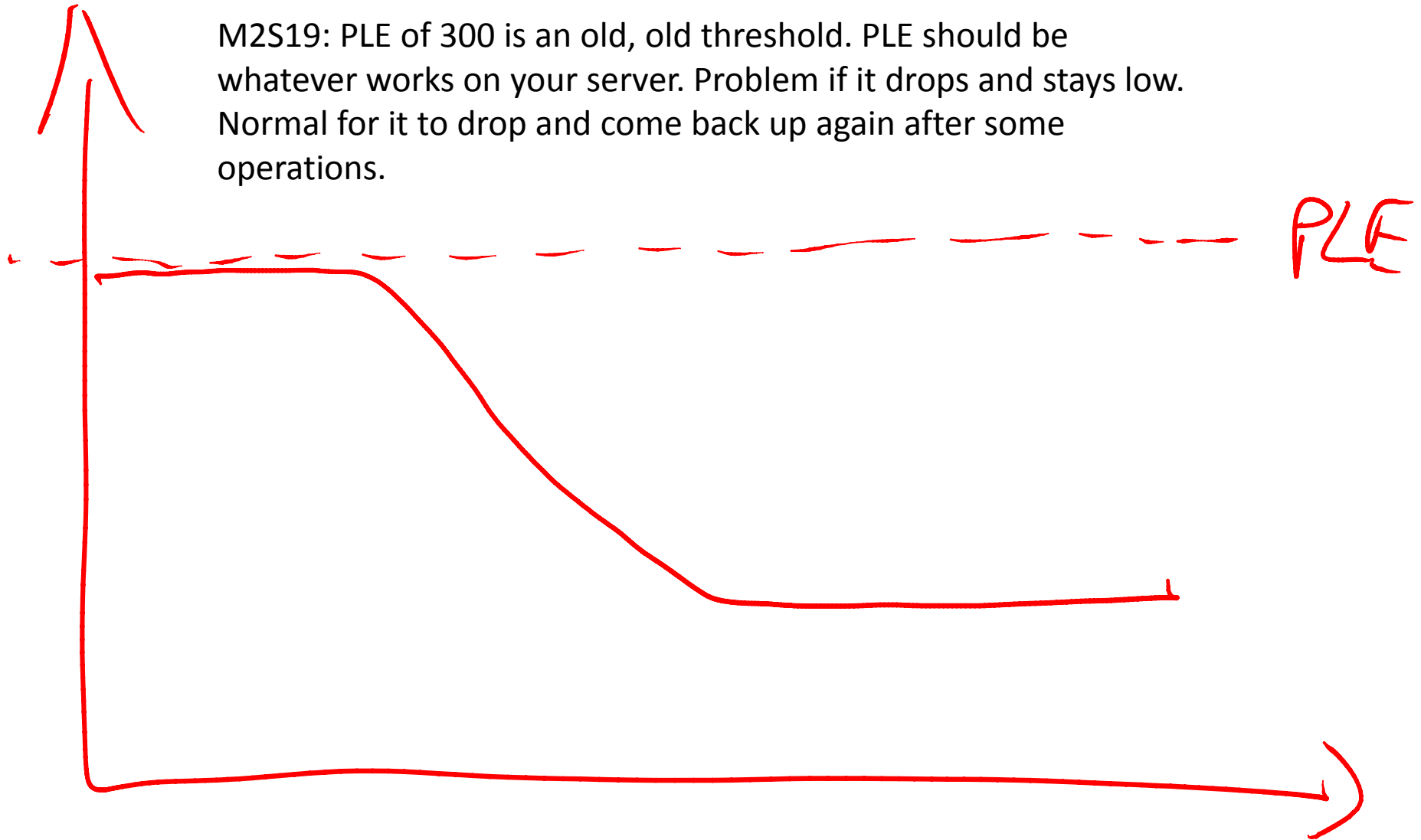


M2S19: With no other memory pressure, PLE ticks up by one per second.

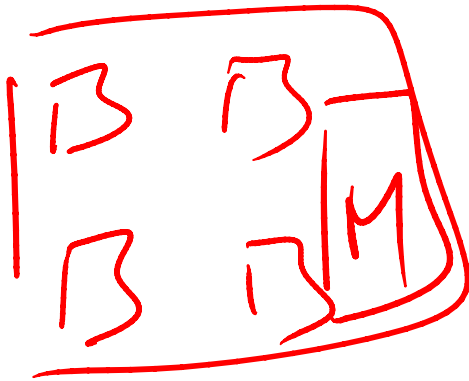
M2S19: PLE of 300 is an old, old threshold. PLE should be whatever works on your server. Problem if it drops and stays low. Normal for it to drop and come back up again after some operations.



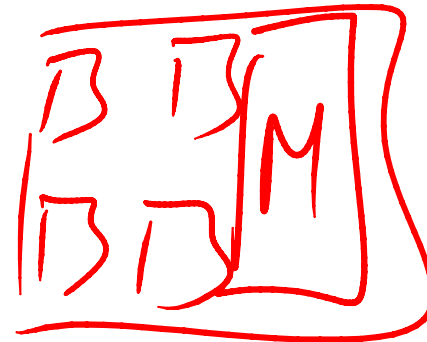
M2S19: PLE of 300 is an old, old threshold. PLE should be whatever works on your server. Problem if it drops and stays low. Normal for it to drop and come back up again after some operations.



NUMA



M2S19: NUMA – Non Uniform Memory Access. Each processor (or group) has local memory, containing a partition of the buffer pool. 4 NUMA nodes = 4 equal-size partitions of the buffer pool.



See <http://www.sqlskills.com/blogs/paul/page-life-expectancy-isnt-what-you-think/>

Buffer Manager P.L.E.

= harmonic mean

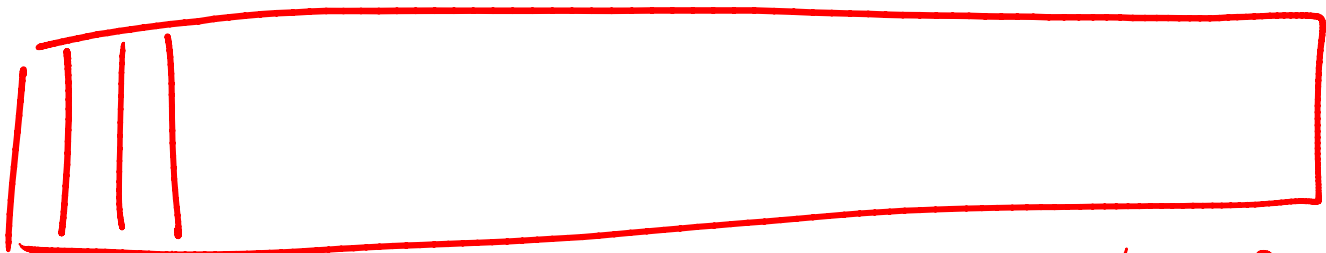
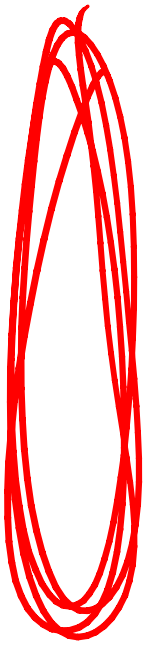
Buffer Node 1 PLE

1 2 PLE

11 3 PLE

11 4 PLE

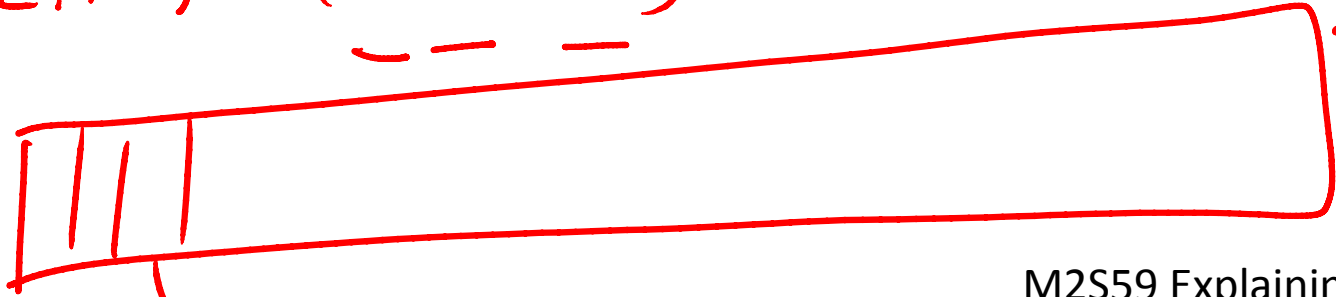
M2S19: If you have NUMA, you need to monitor the PLE in each Buffer Node perfmon object, not the Buffer Manager PLE.



PFS
(2:1:1)

SGAM
(2:1:3)

PAGE LATCH UP
_EX



PFS

SGAM

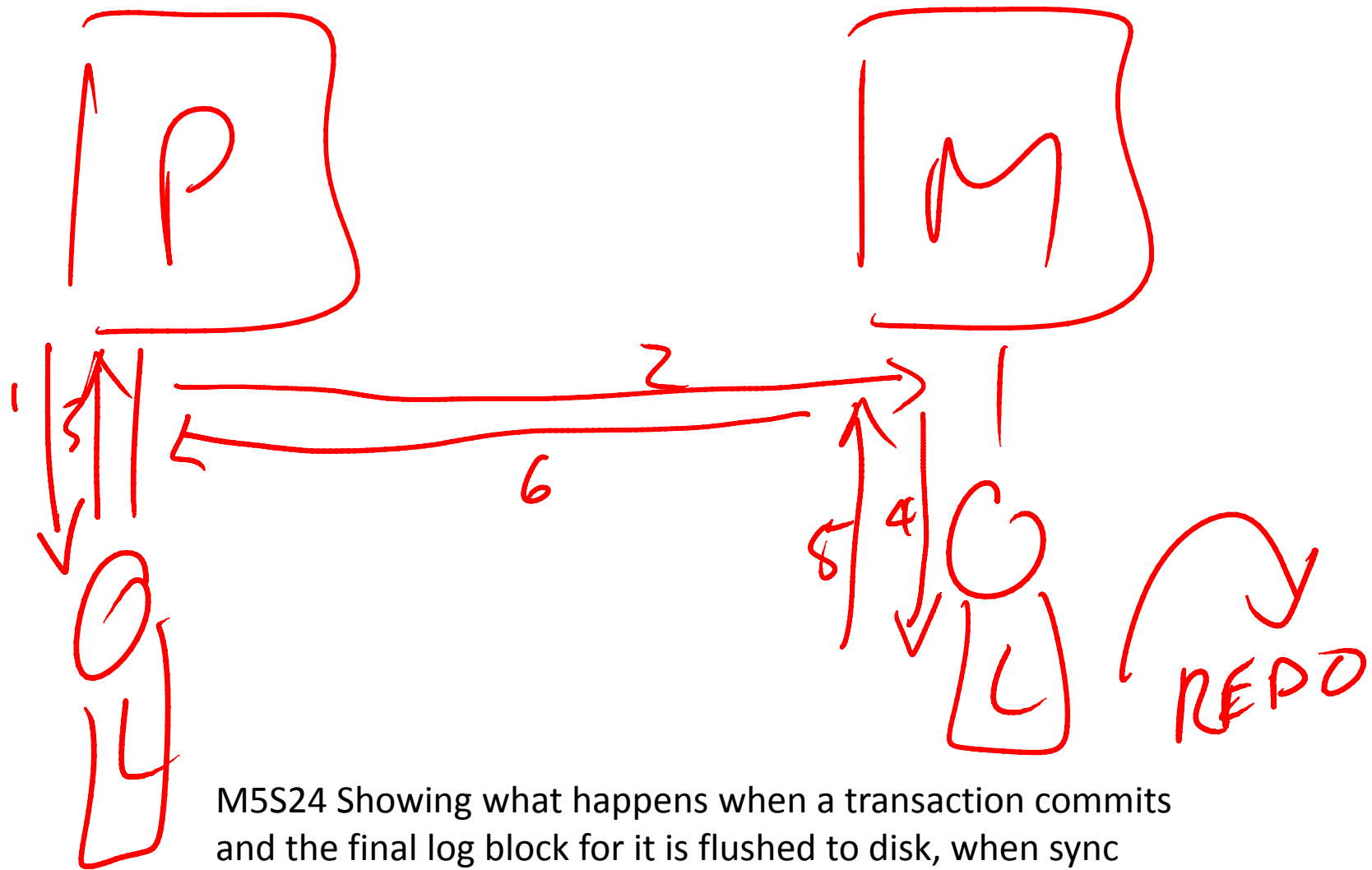
M2S59 Explaining PFS and SGAM contention in tempdb and how adding multiple files spreads the contention over multiple files, thus reducing it.

1	FIXED	1	v1	v2	v3
---	-------	---	----	----	----

MODIFY ROW

MODIFY COLUMNS

M5S19 Explaining how the Access Methods determines how to log UPDATES efficiently by splitting the logging by portion of the record. Each variable length column is a portion, as is the fixed-width portion (up to 16 bytes in each log record)

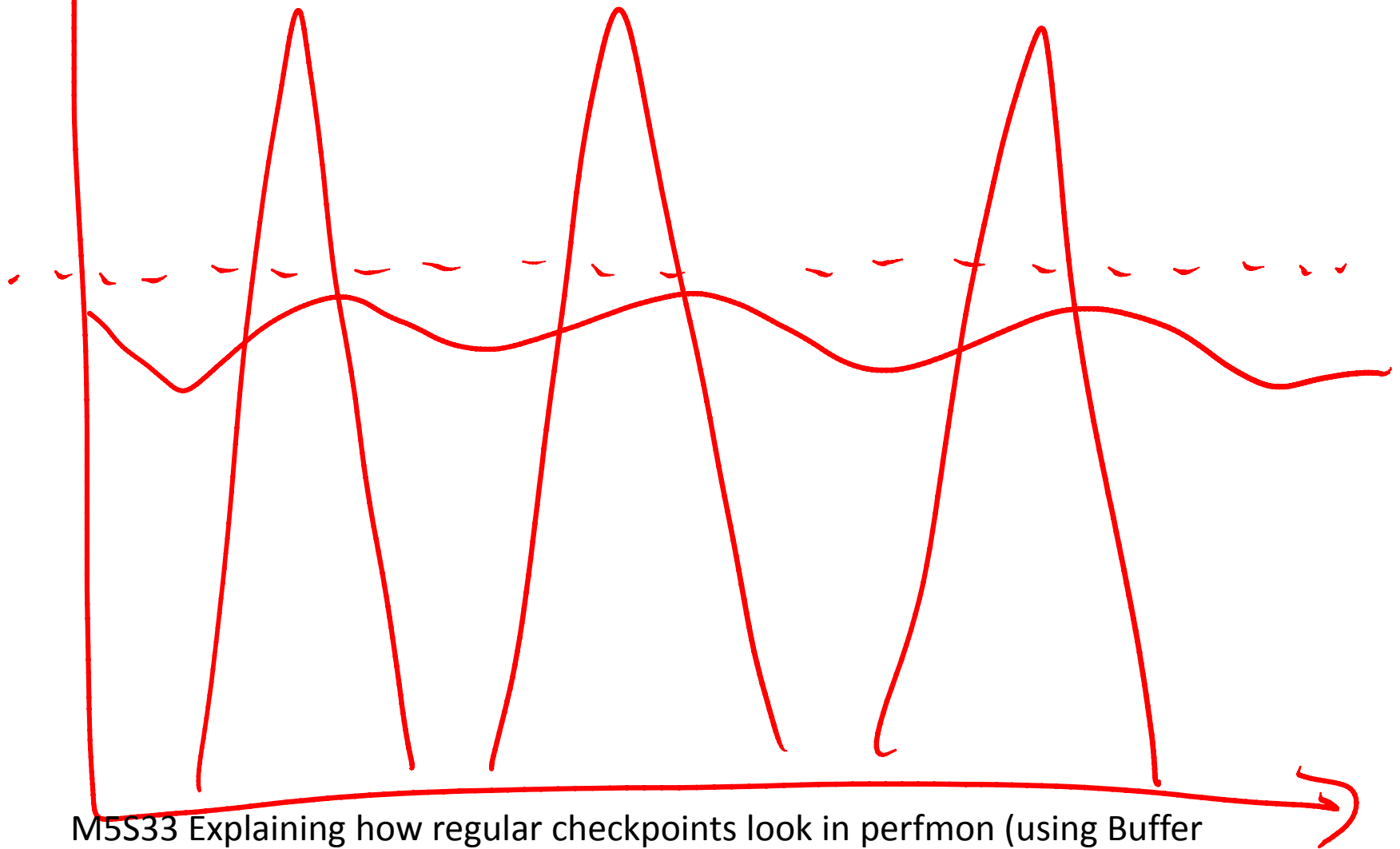


M5S24 Showing what happens when a transaction commits and the final log block for it is flushed to disk, when sync DBM or AG is involved. The local I/O will complete first (3) but the transaction commit cannot complete until the mirror acknowledges the remote I/O (6).

sys.dm_io_pending_io-
requests

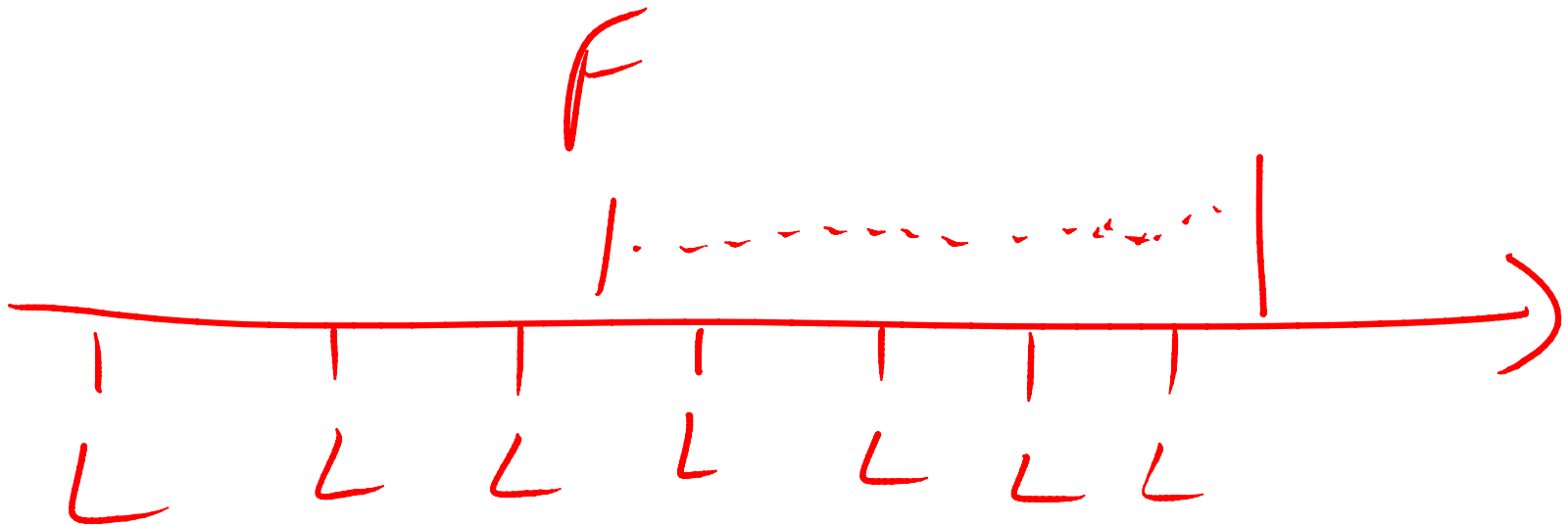
Use this DMV to find in-flight writes.

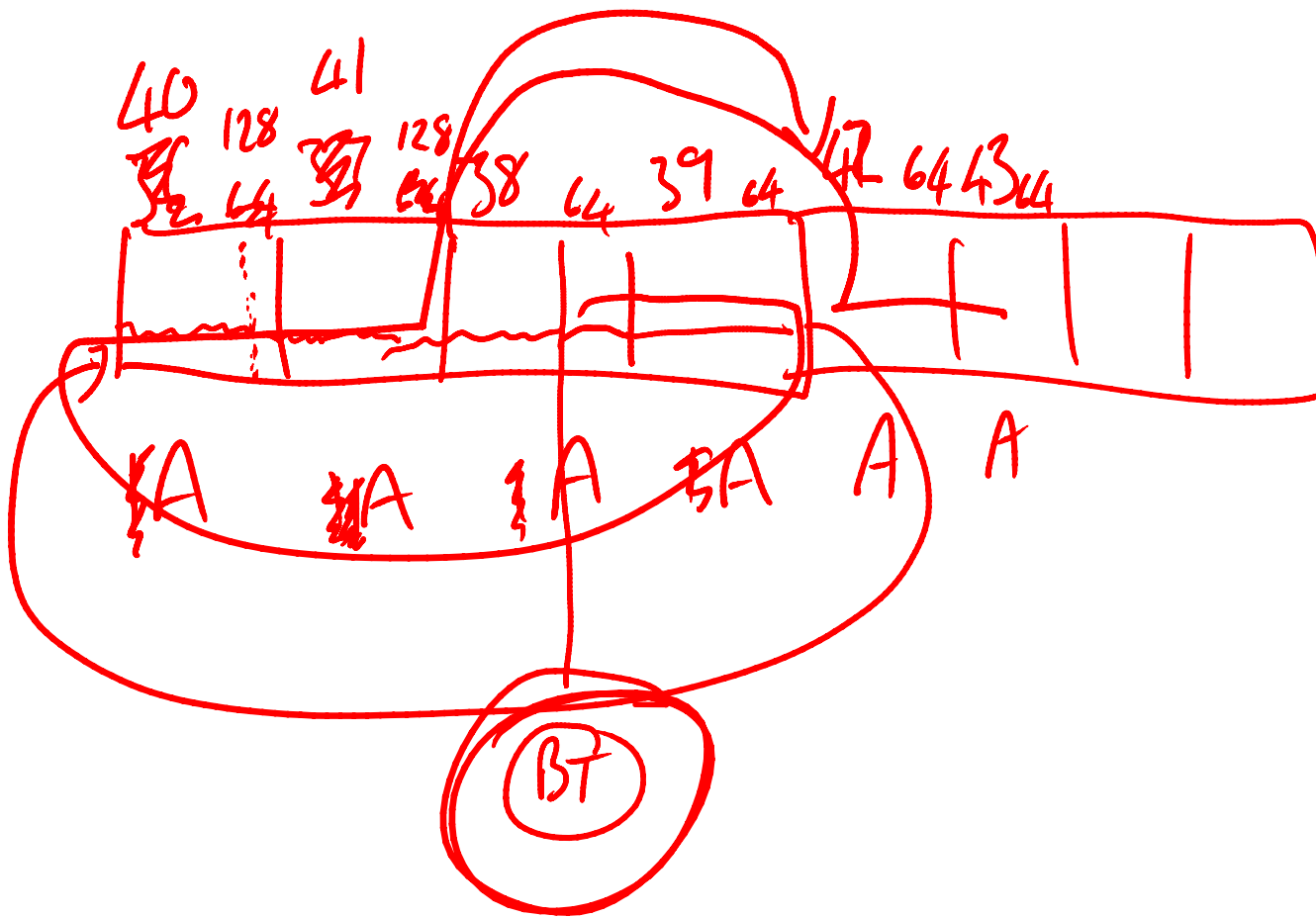
BUFFER MGR: CKPT Pages/sec



M5S33 Explaining how regular checkpoints look in perfmon (using Buffer Manager/Checkpoint pages/sec counter) and how they may exceed I/O capacity (the dotted line). Doing more frequent manual checkpoints, or using the indirect checkpoints in SQL 2012+, can result in more constant, lower IO load.

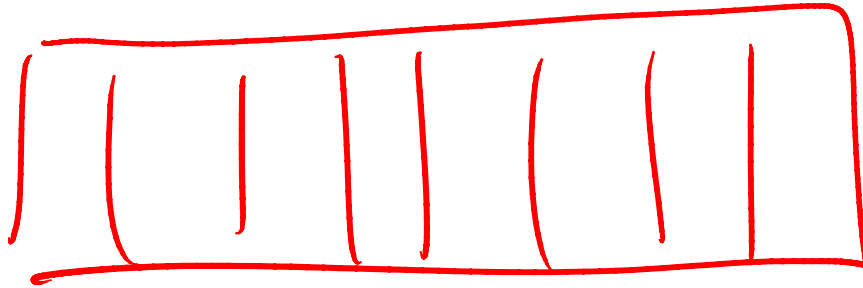
M5S45 Explaining the misconception that a full backup clears the log. It's deferred log clearing from concurrent log backups.





M5S51 Explaining the circular log demo. We discussed log space reservation that happens to ensure the active transactions can roll back without having to grow the log – this accounts for extra log growths. Space reserved in the log does not make a VLF become active – as there are no log records written out, just empty space reserved.

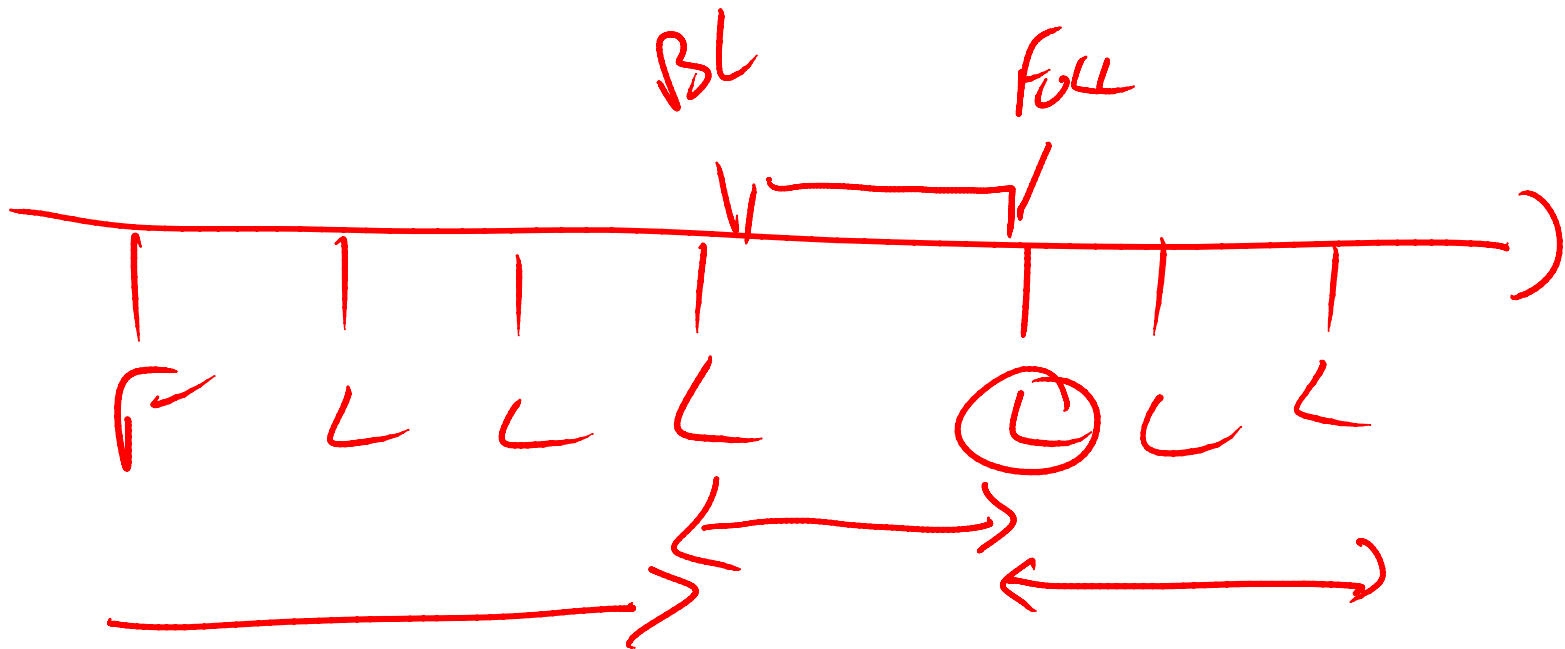
EXTENT X



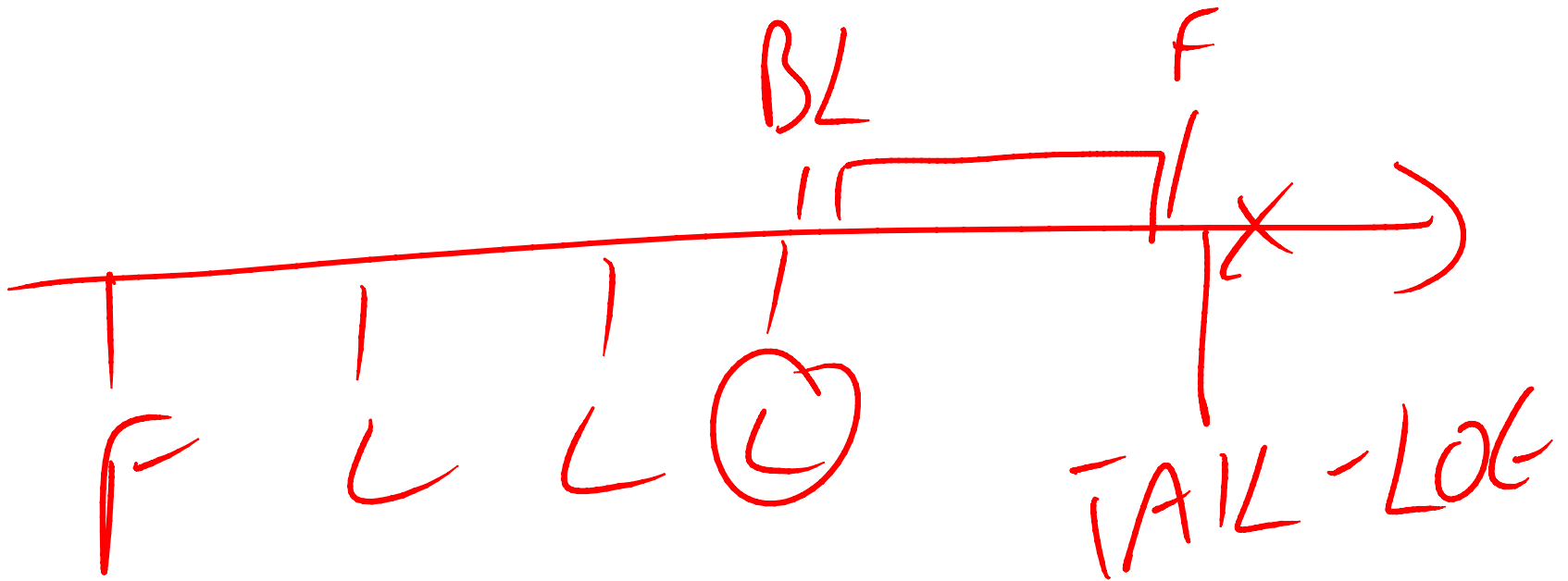
PROBE X

M5S65 Explaining part of the deferred-drop functionality. Before an extent can be deallocated, an X lock is acquired on it and all 8 page locks are probed (instant acquire X lock and release) to make sure nothing else has them locked.

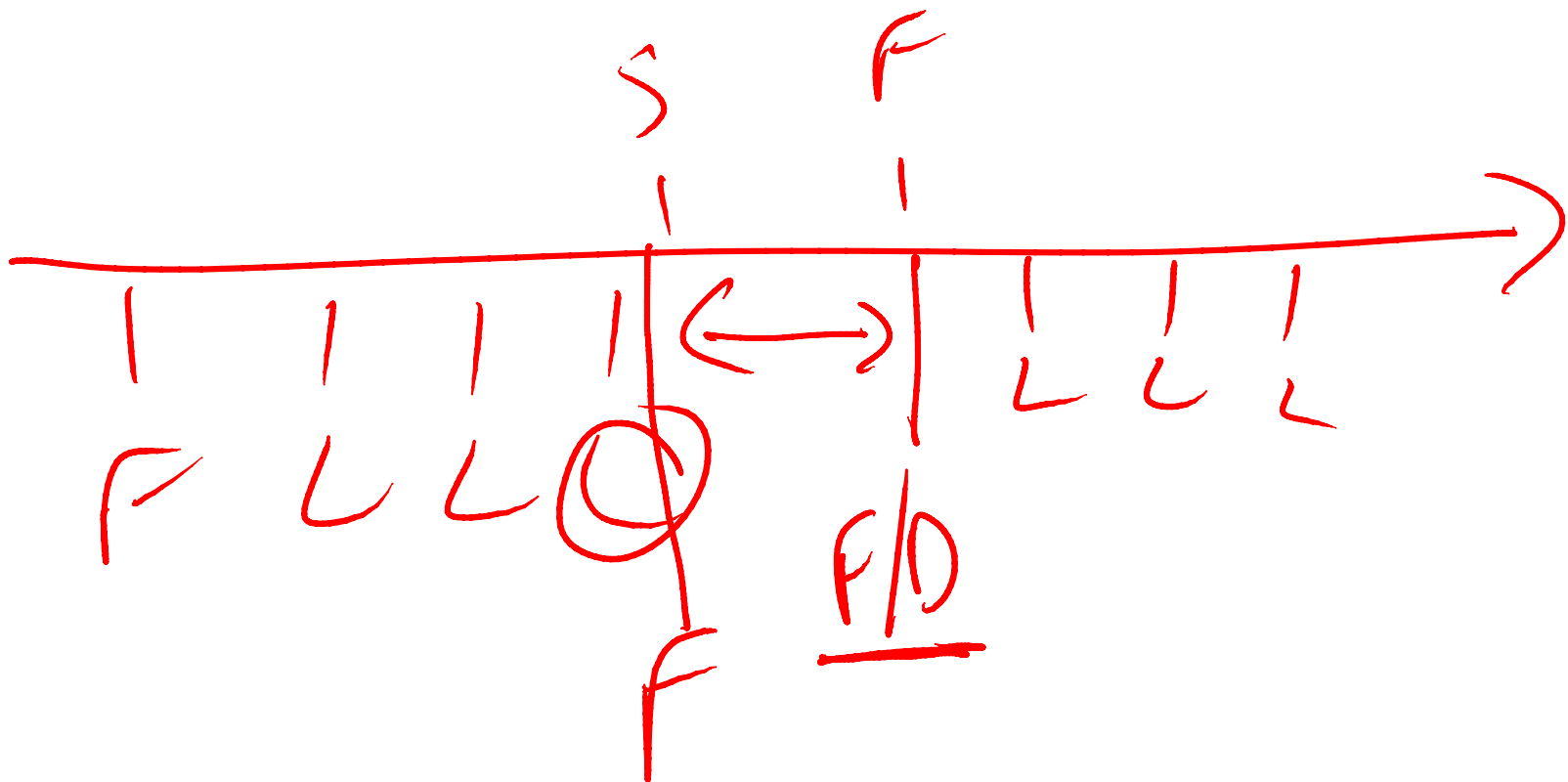
M5S61 Explaining that if a minimally-logged operation is present in a log backup, you cannot do a point-in-time restore to any time covered by that backup (except the start or end of that time period).

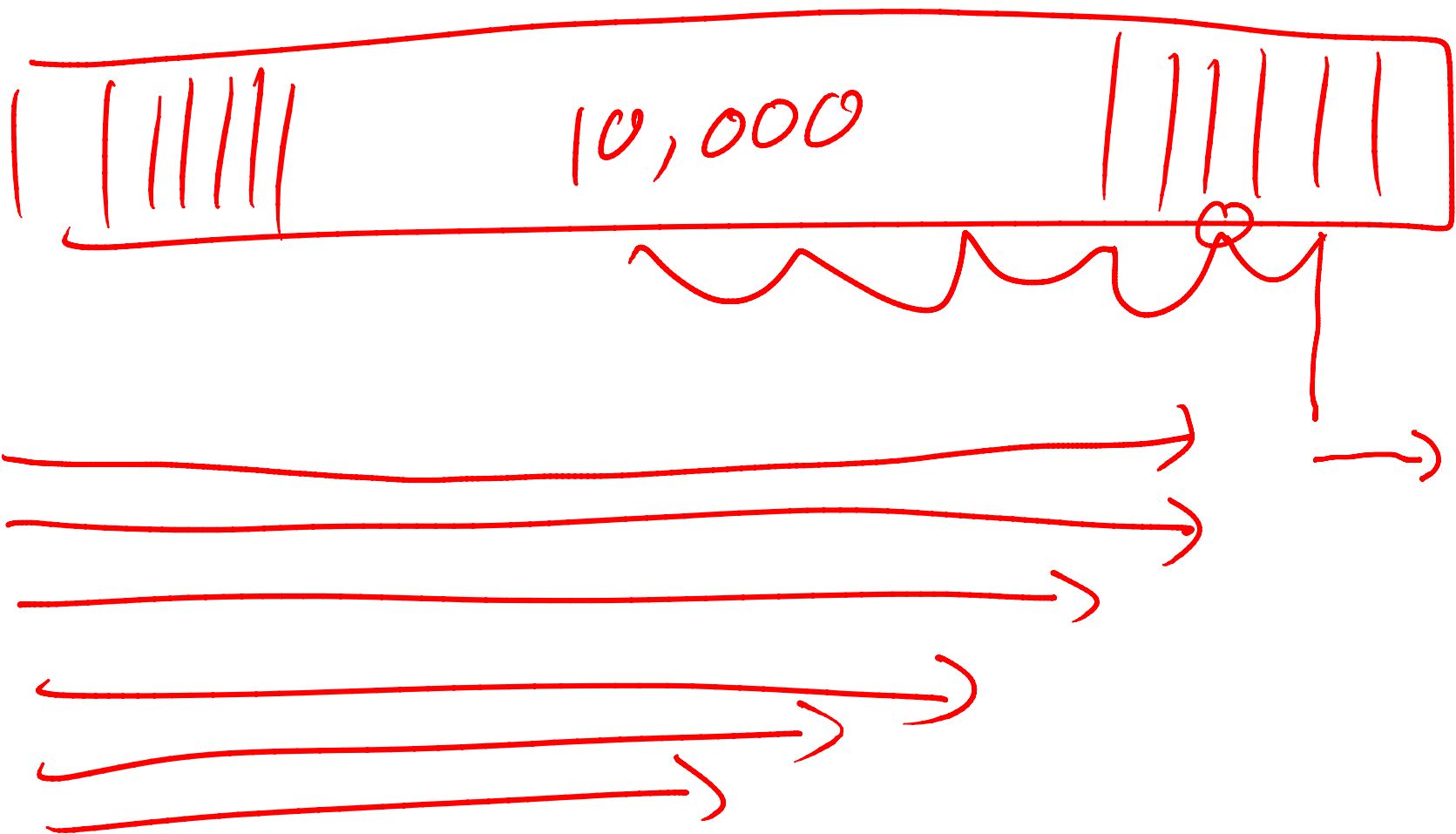


M5S61 You cannot do a tail-of-the-log backup after a crash that destroys the data files if there has been a minimally-logged operation since the last log backup. Well in 2008 R2/2012, you can, but the log backup is invalid.



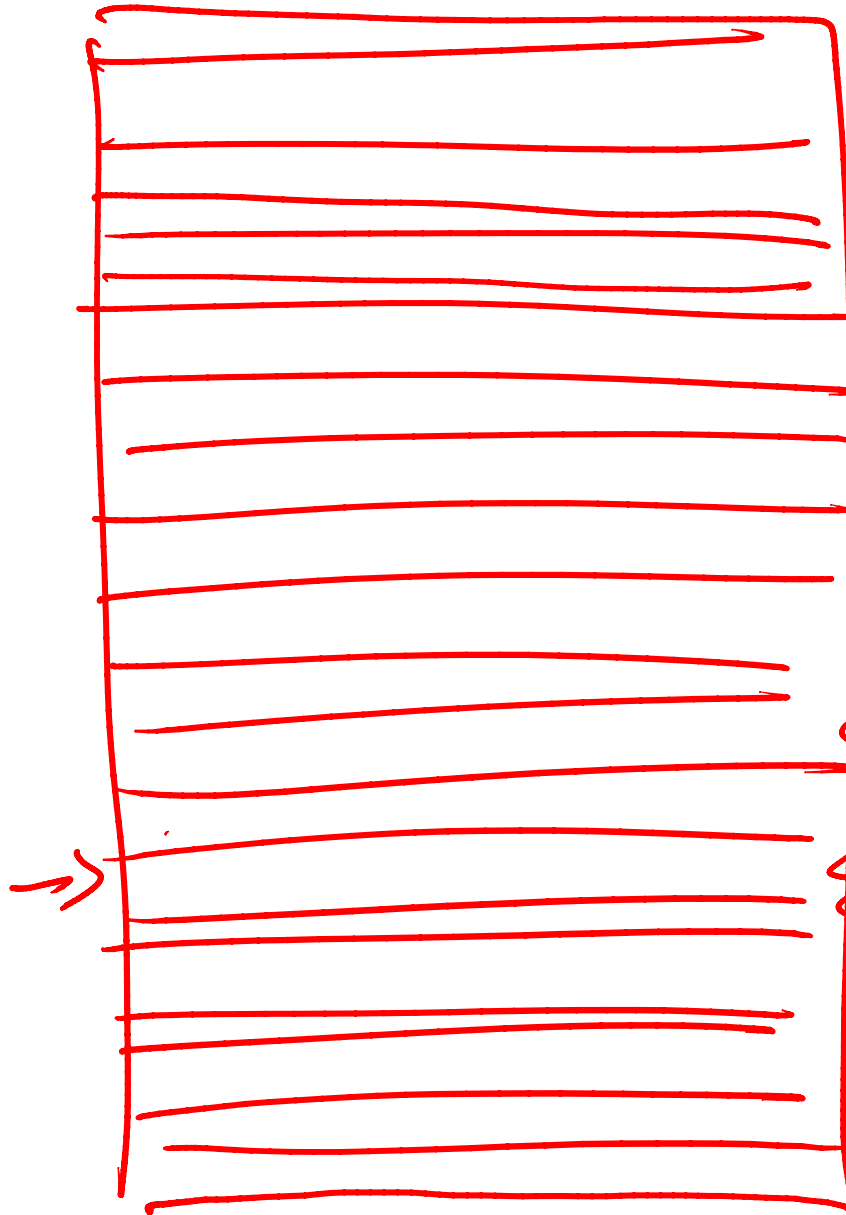
M5S66 Explaining how to restart the log backup chain after switching to SIMPLE. It requires a full or differential backup. Try not to switch to SIMPLE though, as it means there's only a single data backup you can restore from (the full/differential after the switch). If it's damaged, you can't restore using log backups to cross the gap between S and F.





M5S66 VLF fragmentation is bad because of having to search through the VLFs for a particular VLF sequence number. See the KB article link for more info.

14



BINARY SEARCH



M7S5 Explaining how a record is found on a page using binary search rather than brute force search.

MEMBER ID

PAUL	ALLAN	MICHAEL
------	-------	---------

KIMBERLY

M7S19 The MySpace story. The 'K' is the tiny portion of comments on Kimberly's page because she wasn't very popular back then...



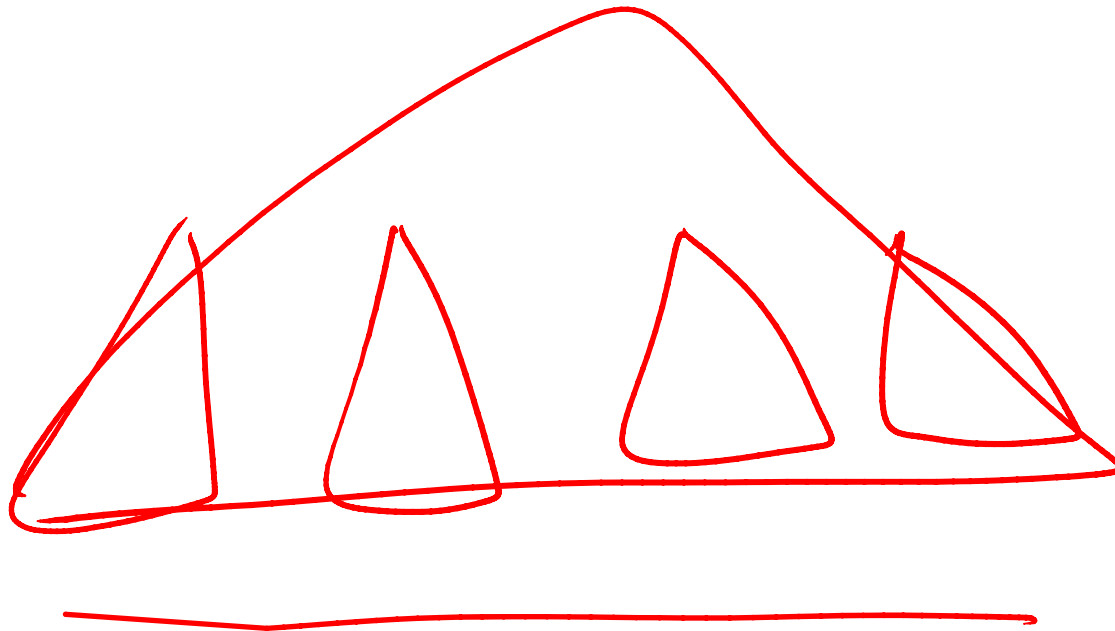
M7S28 The numbers from the page split logging cost demo.

SIZE	NO SPLIT	SPLIT	x
1000	1256	7580	6
100	356	6772	18
10	268	1172	45

LOP_DELETE_SPLIT

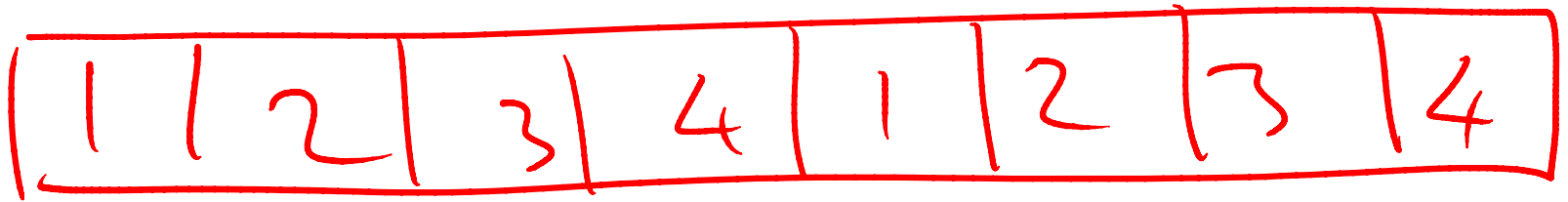
M8S18 Append-only insert pattern fills pages on right-hand side of index. New page allocations in this case are called page splits too – making all methods of tracking page splits largely useless (until 2012 with Xevents). You can look for LOP_DELETE_SPLIT log records to see splits occurring.

$$DOP = 4$$



$$DOP = 1$$

M7S38 Discussing how DOP affects contiguity. With DOP 4, 4 threads are building portions of the new index in parallel. They will all be allocating extents, leading to extent fragmentation (and hence logical fragmentation).

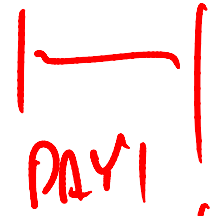


M7S38 Discussing how DOP affects contiguity. With DOP 4, 4 threads are building portions of the new index in parallel. They will all be allocating extents, leading to extent fragmentation (and hence logical fragmentation).

100GB



.....



STAGGERED
INDEX
MAINT

M7S44 Explaining staggered index maintenance with database mirroring. See <http://bit.ly/IS04W5> for more explanation