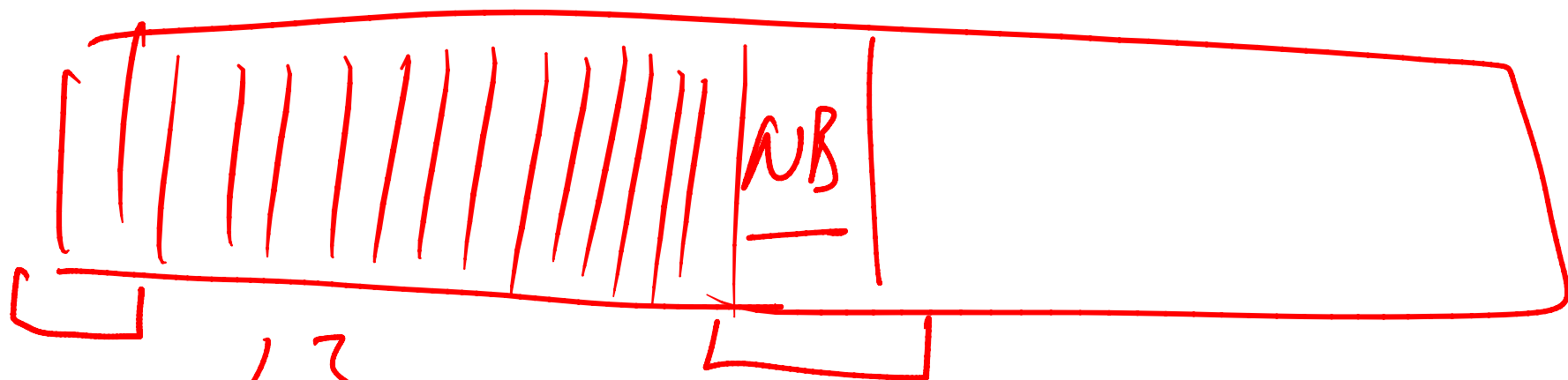




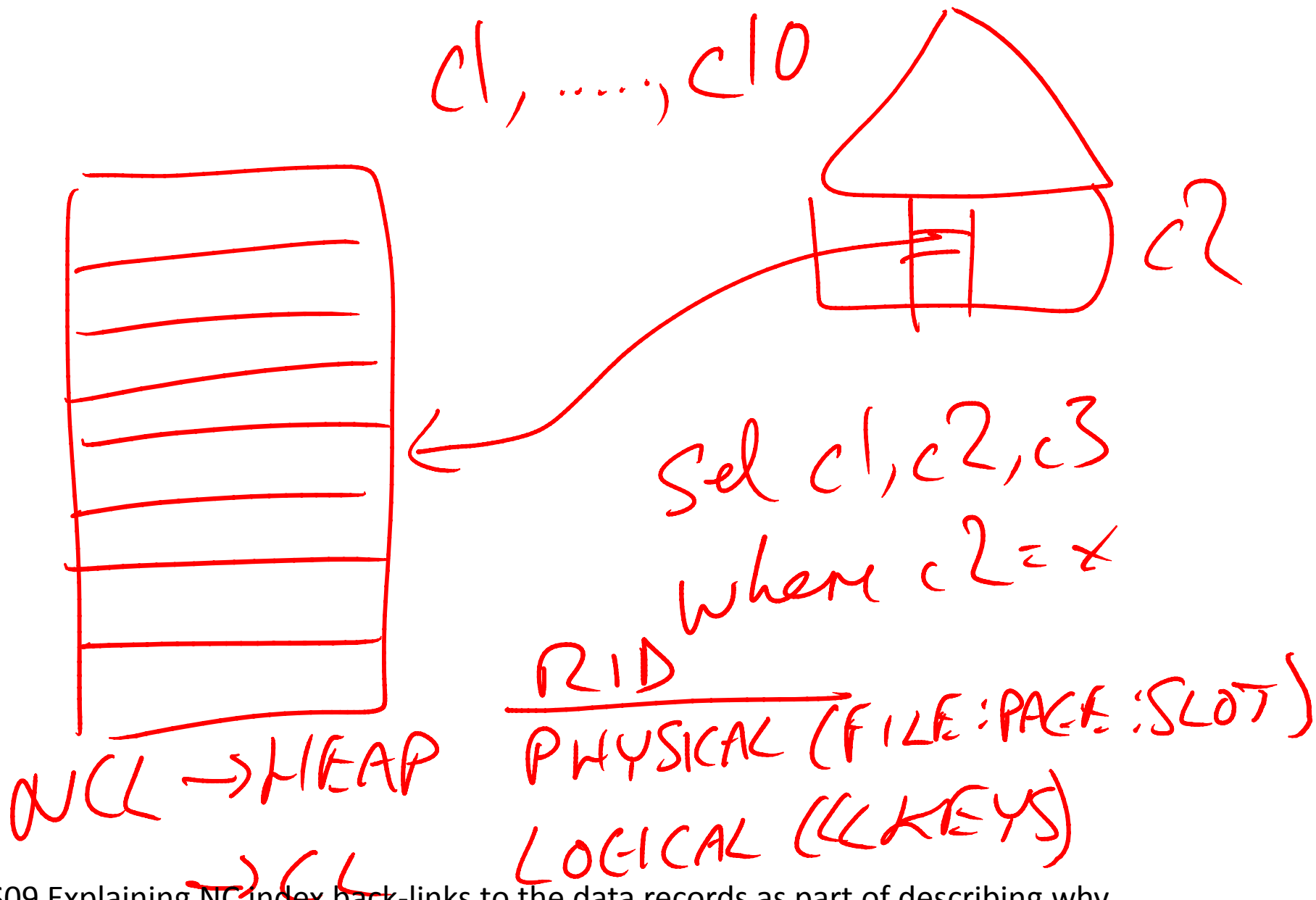
L3

CACHE LINES

64b
128b

INVALIDATION

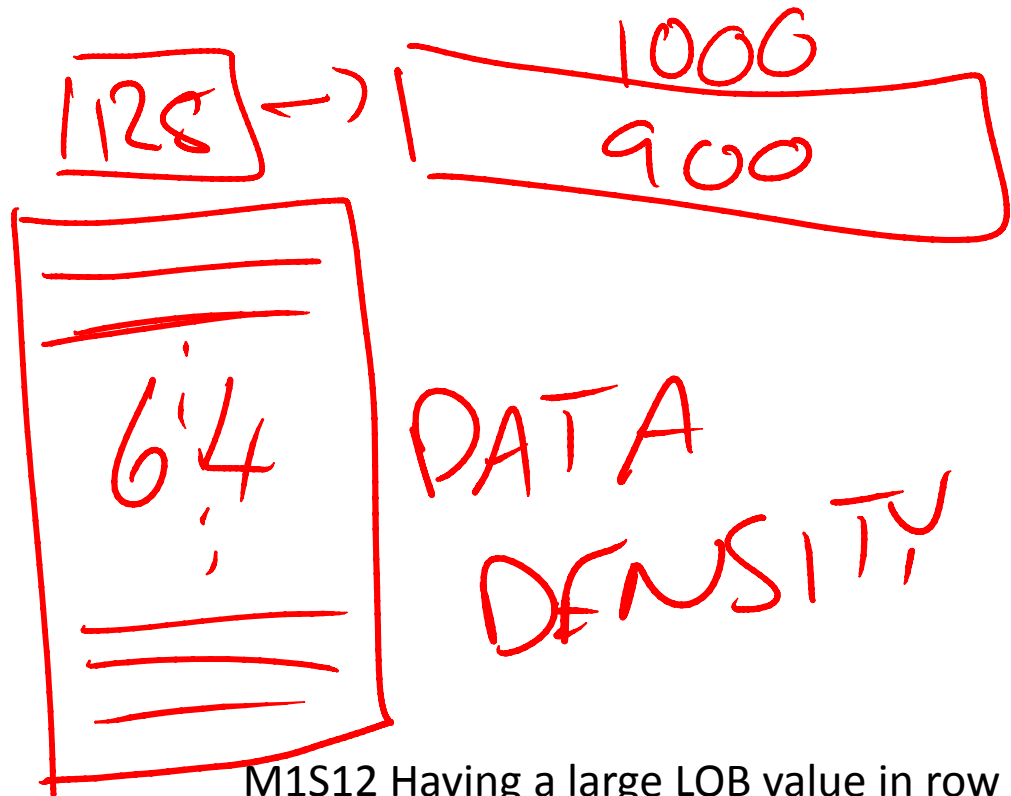
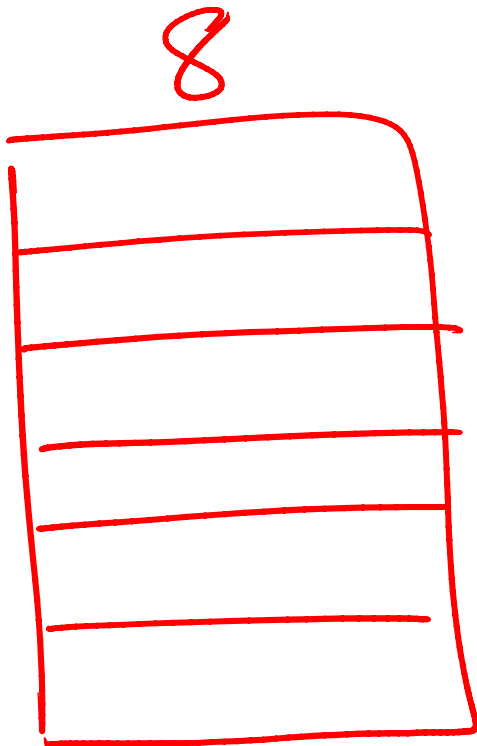
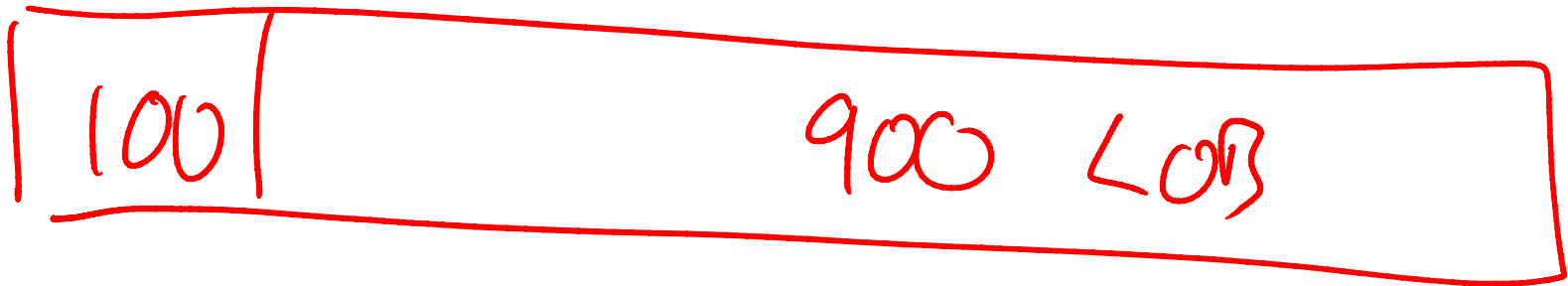
M1S06 Discussing
why the NULL
bitmap is a perf
optimization by
avoiding having to
read record parts
into the CPU's L3
cache



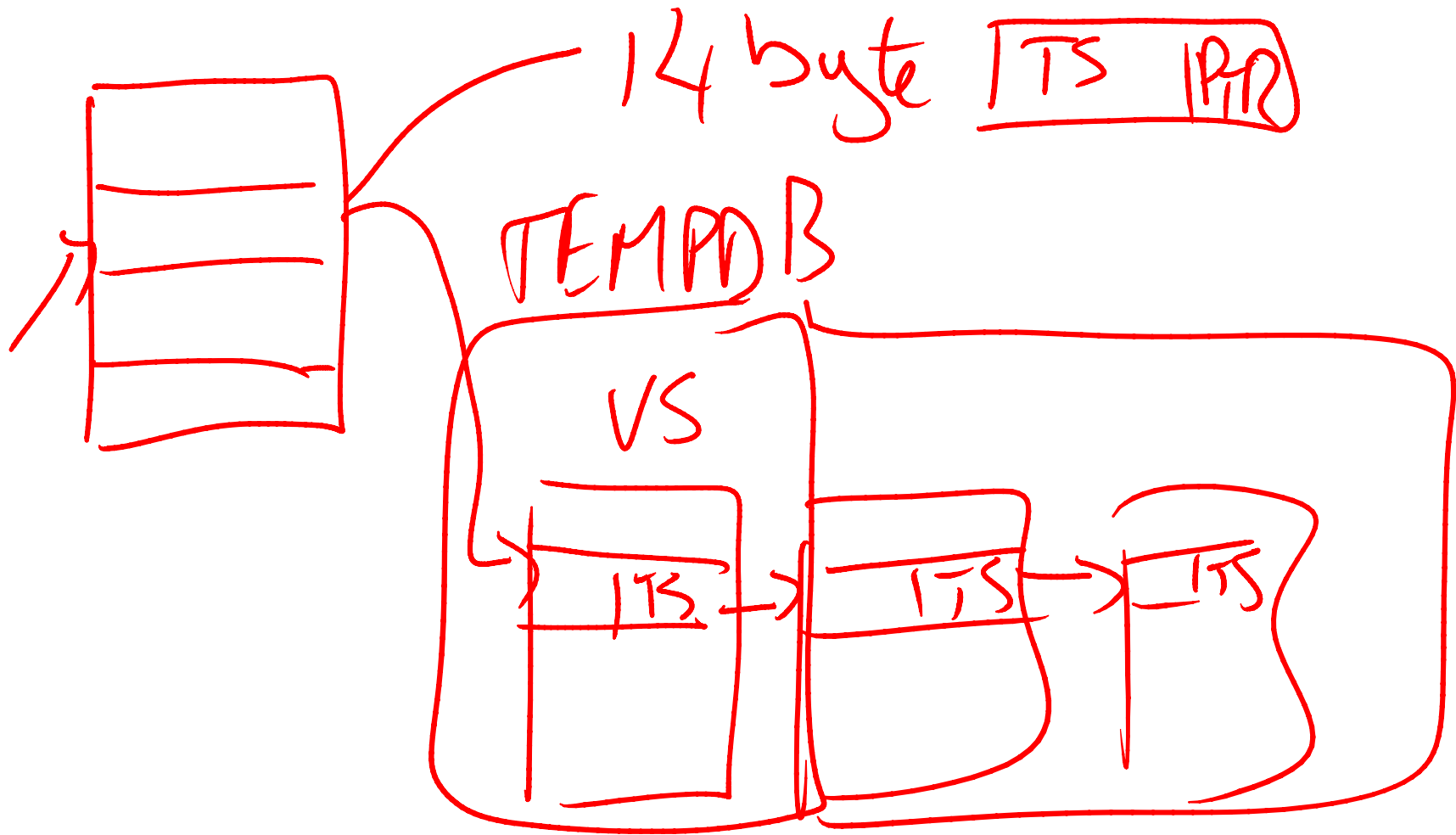
M1S09 Explaining NC index back-links to the data records as part of describing why forwarding/forwarded records exist. Without them, moving a heap row would entail changing all NC index records with a back-link to that row.

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



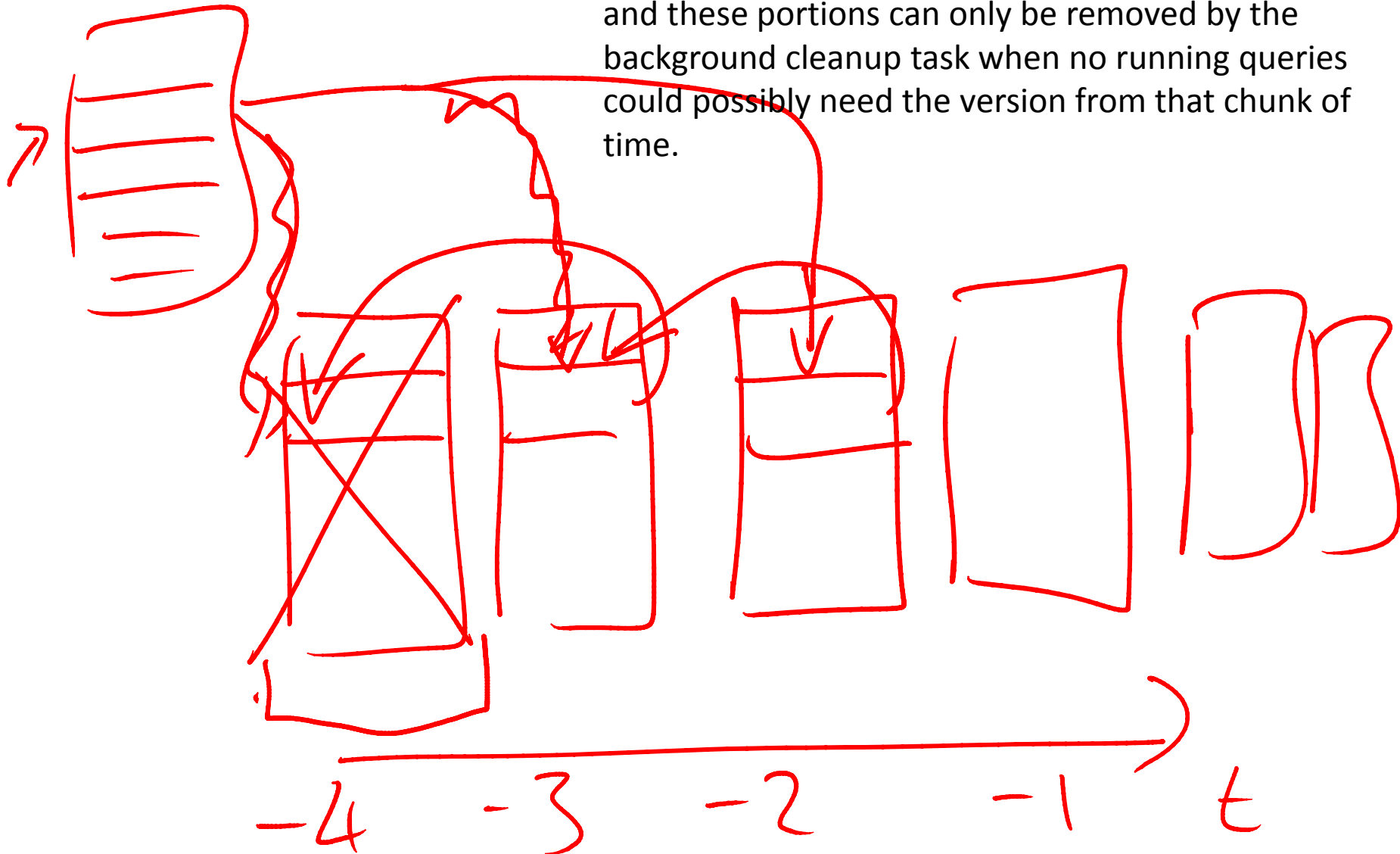


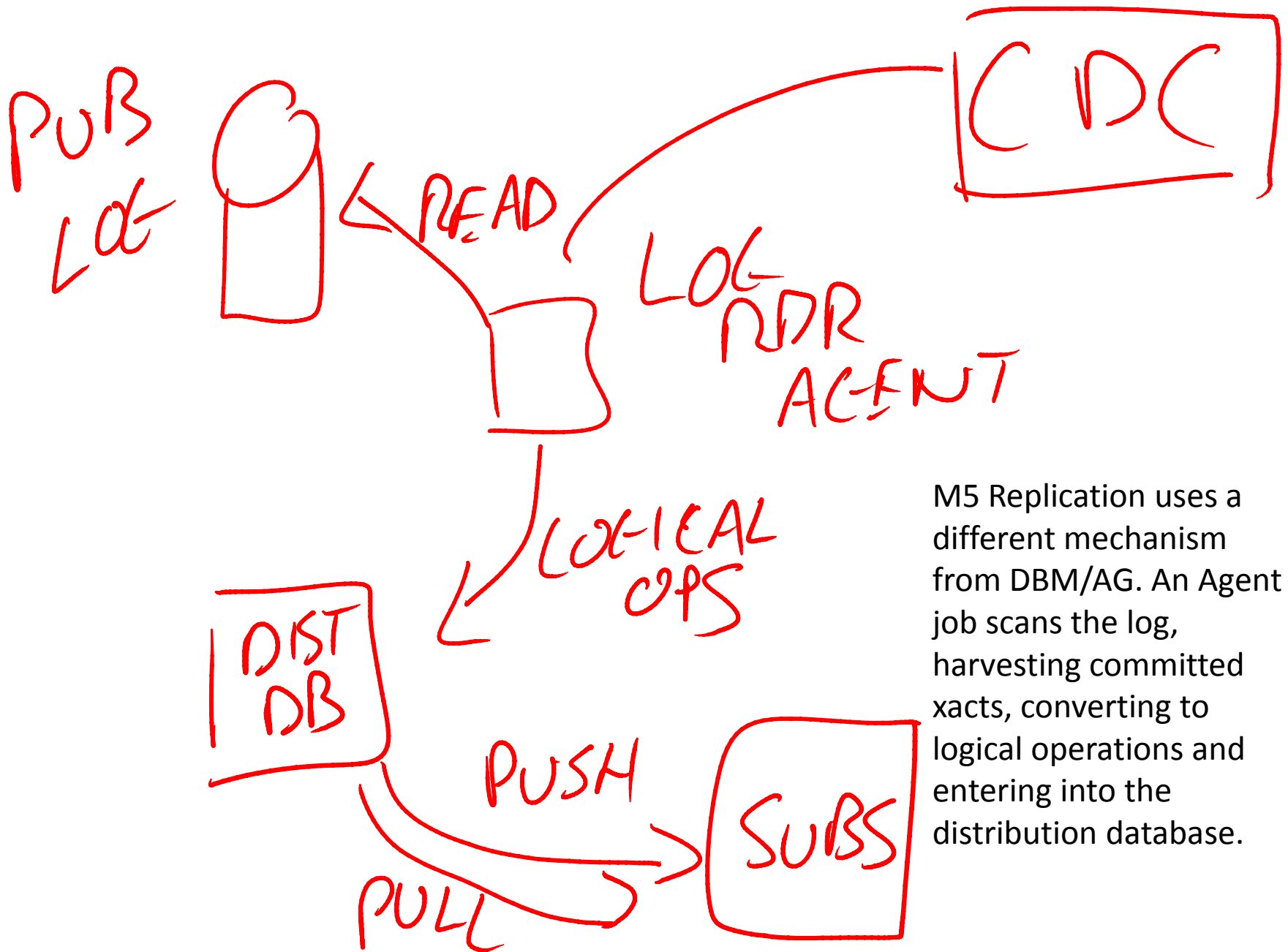
M1S12 Having a large LOB value in row that's not used much makes for large rows and low data density. Choose how to store LOB data wisely.



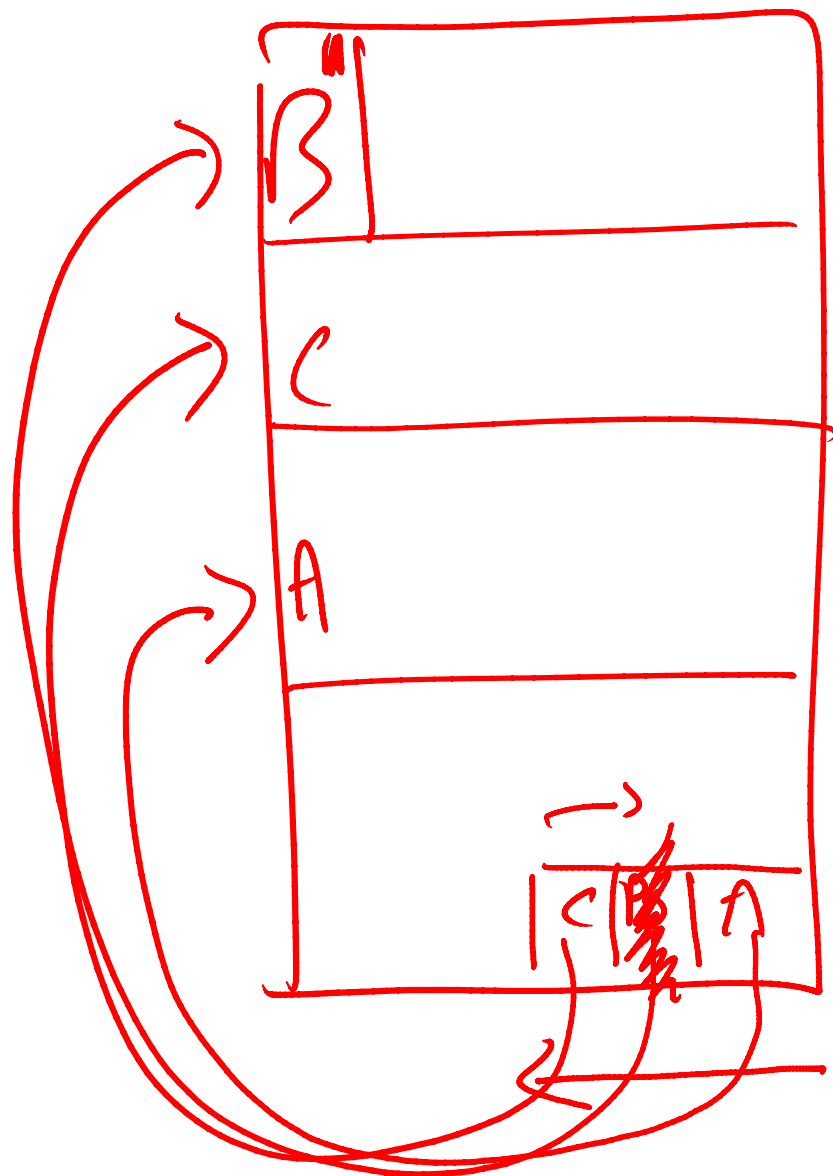
M1S14 When an update occurs to a record and snapshot isolation is enabled, the pre-change record is copied into the version store in tempdb. The changed record has a 14-byte tag on the end with the timestamp and pointer to previous record version. A chain of previous versions is possible.

M1S14 & M2S56 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.





M5 Replication uses a different mechanism from DBM/AG. An Agent job scans the log, harvesting committed xacts, converting to logical operations and entering into the distribution database.



LOP EXPUNCE -
GHOST

SLOT ARRAY

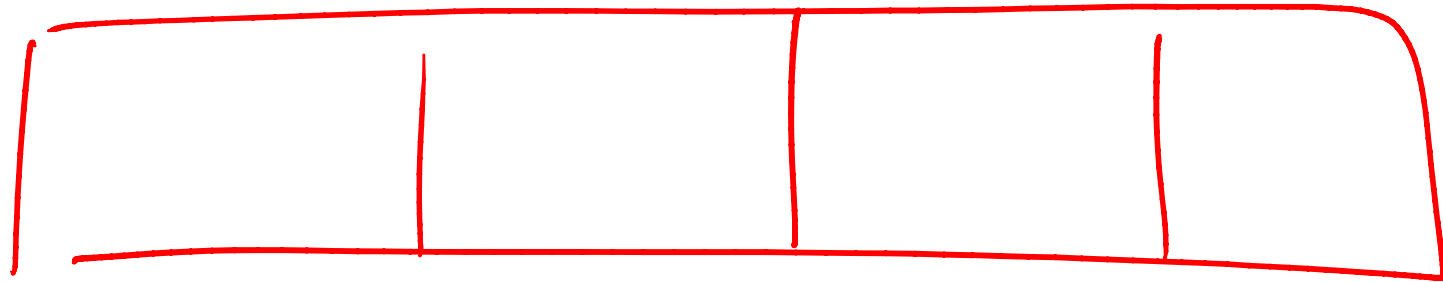
M1S15 When a record is deleted, it's marked as a ghost record. The ghost cleanup task does not physically delete the records – it removes the slot array entry that points to the record on the page. I.e. the record is unmarked as being a record and the area on the page is now free space – that just happens to have bits and bytes that used to be a valid record.

DBCC

TRACEON(662, -1)

3605

M1S15 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.



0 — 7 8 — 15 16 — 23

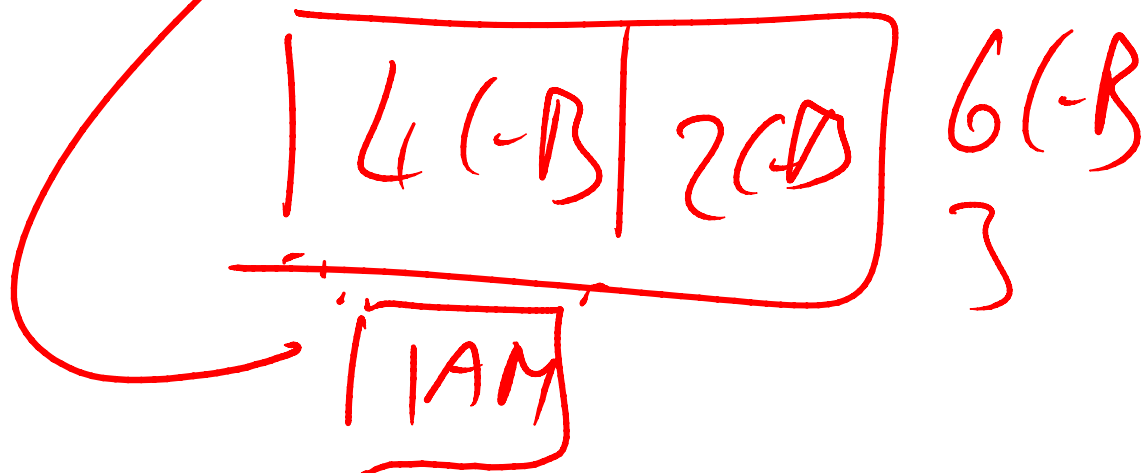


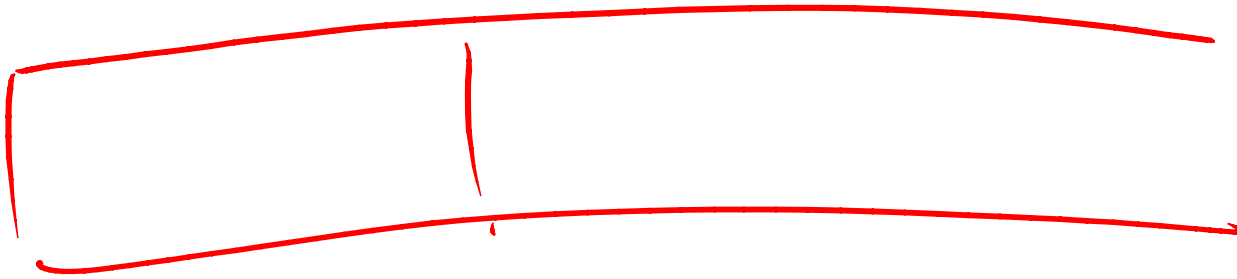
M1S22 Explaining the difference between mixed and dedicated extents. Mixed extent is one in which the 8 pages could be allocated to 8 different objects. Dedicated extent is one where all 8 pages are reserved for exclusive use of a single object.

Later in the deck I refine 'object' to really be 'allocation unit'.



M1S35 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.





0 - FH

1 - PFS

2 - GAM

3 - SGAM

4 - UNUSED

5 - UNUSED

6 - DIFF

7 - ML

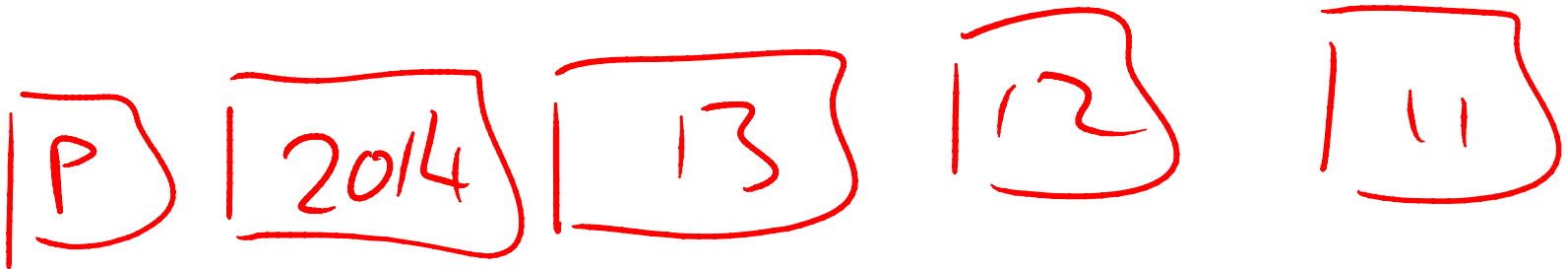
M1 The page types in the first extent in a data file

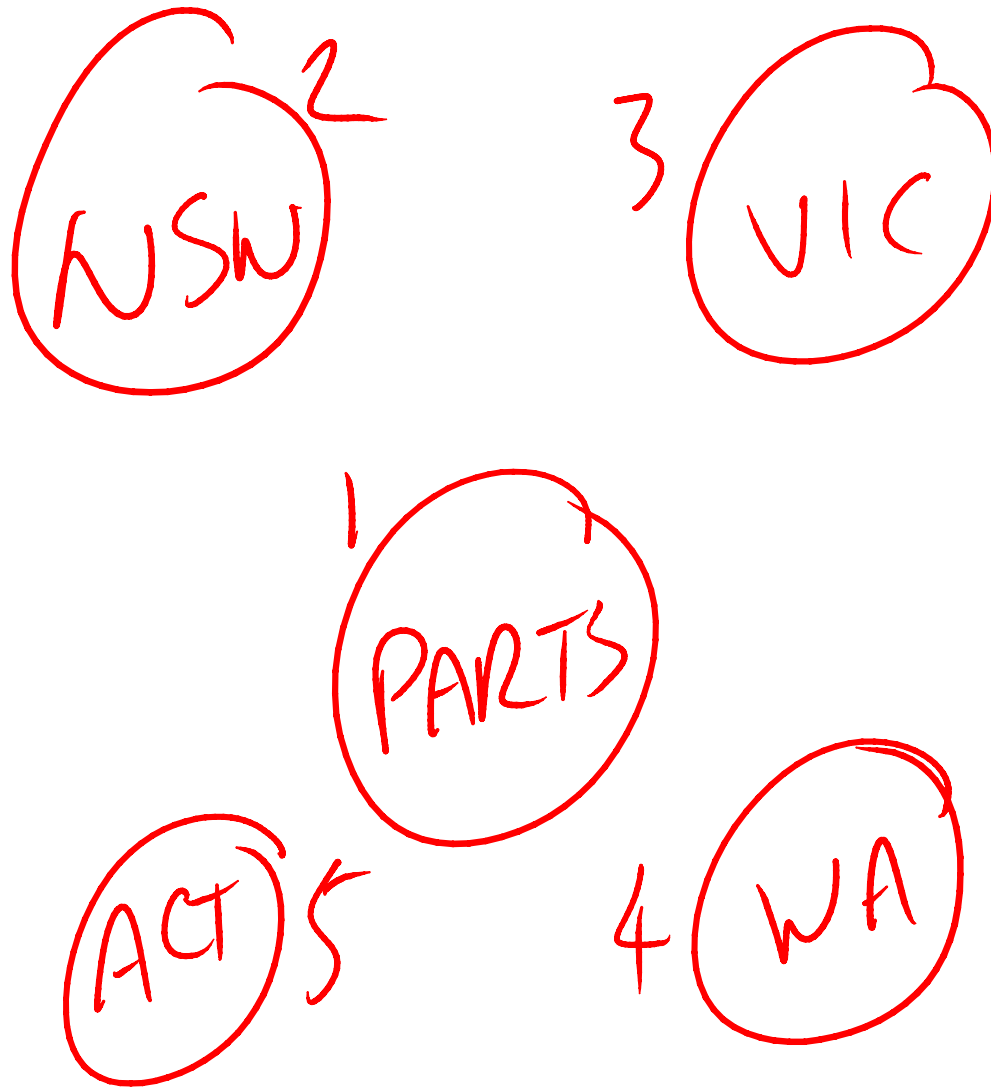
1 TB SALES

2011

2014

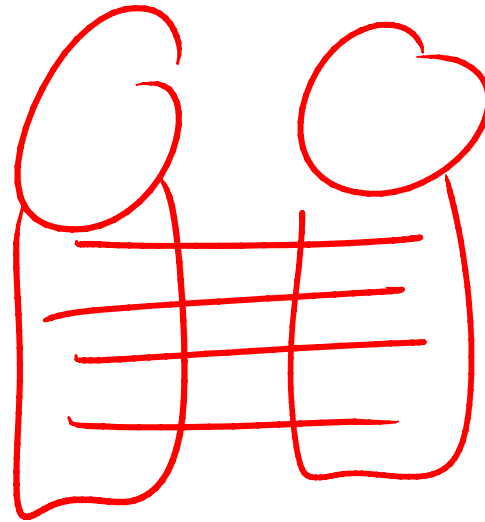
M2S7 Splitting a large database up into multiple filegroups can enable you to do small, targeted restores or even a partial restore from scratch – in the event that the database is damaged or destroyed.





M2S7 Explaining some ways to split up a database to allow targeted restore. On the left, auto parts sales system – makes sense to bring parts and NSW online first.

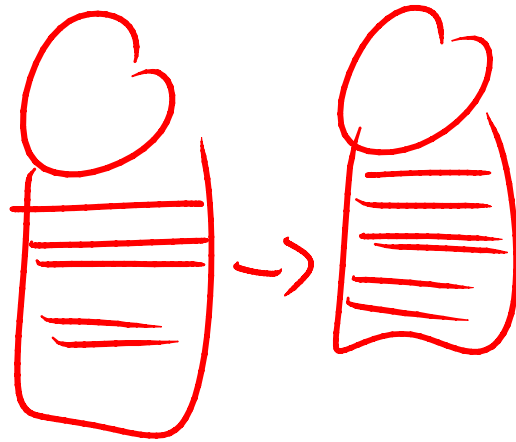
STRIPING



RAID 0

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

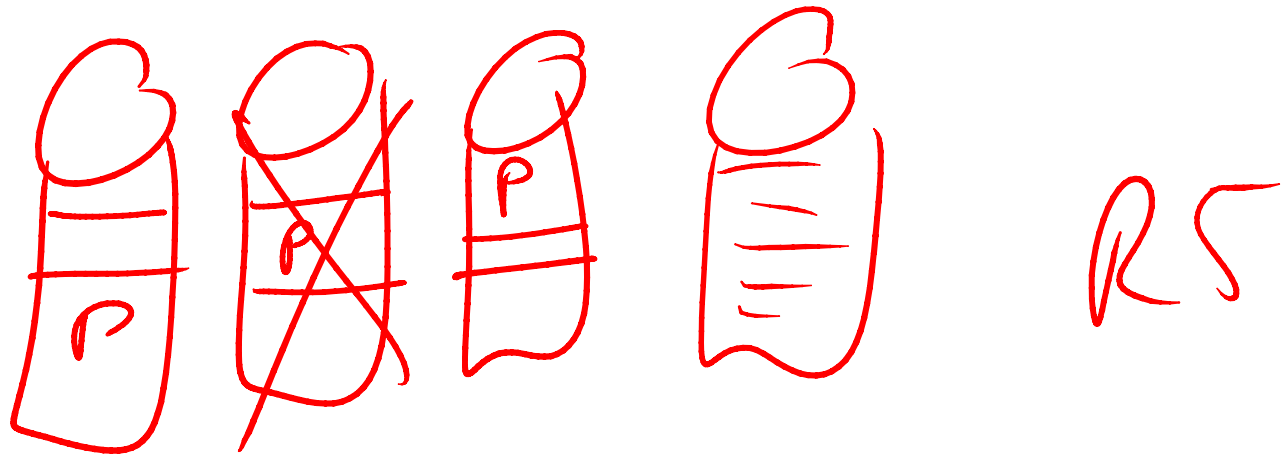
MIRRORING



RA1

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

NOTATING PARITY

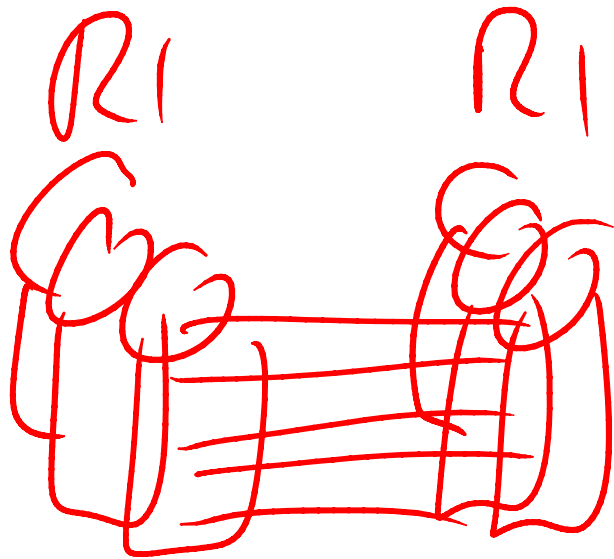


http://en.wikipedia.org/wiki/Standard_RAID_levels

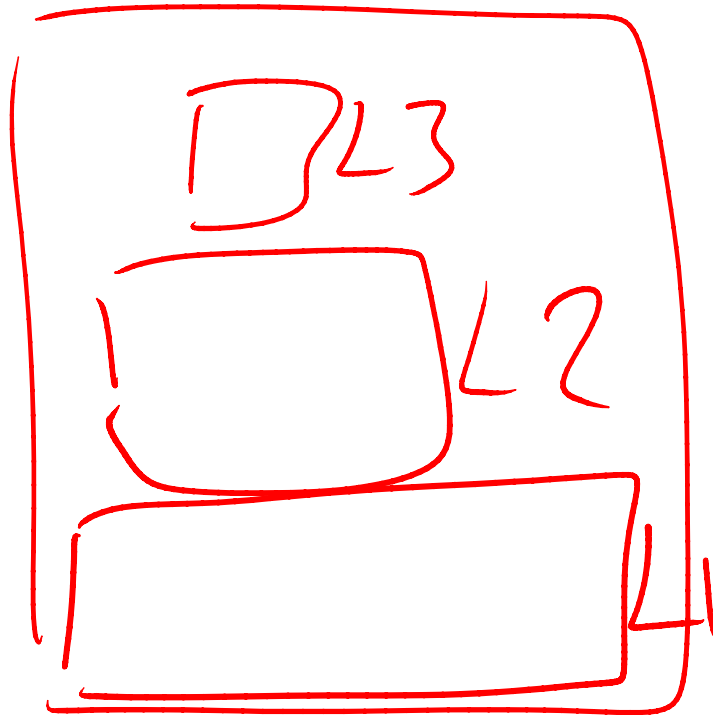
$$R10 = R1 + G$$

= MIRRORING

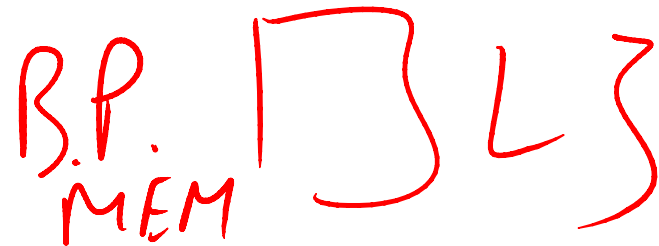
+ STRIPING



CPU

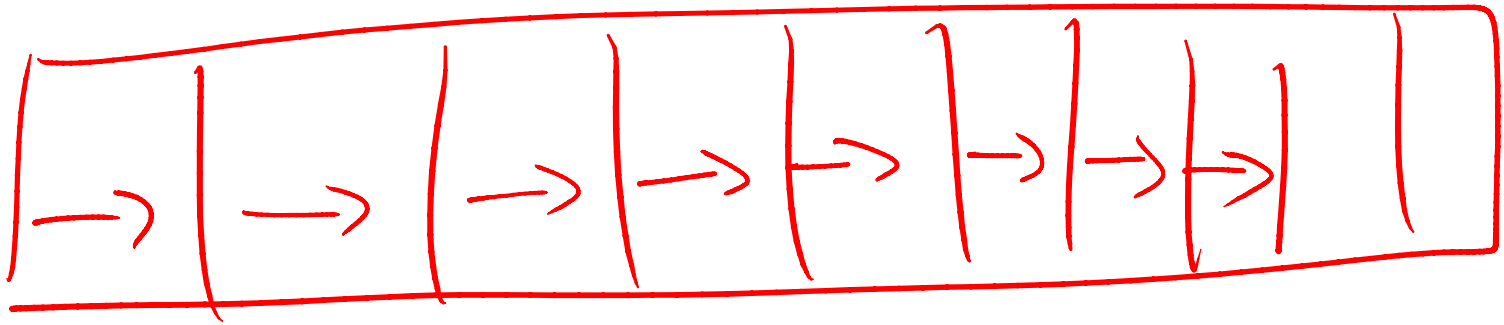


I/O LATENCY



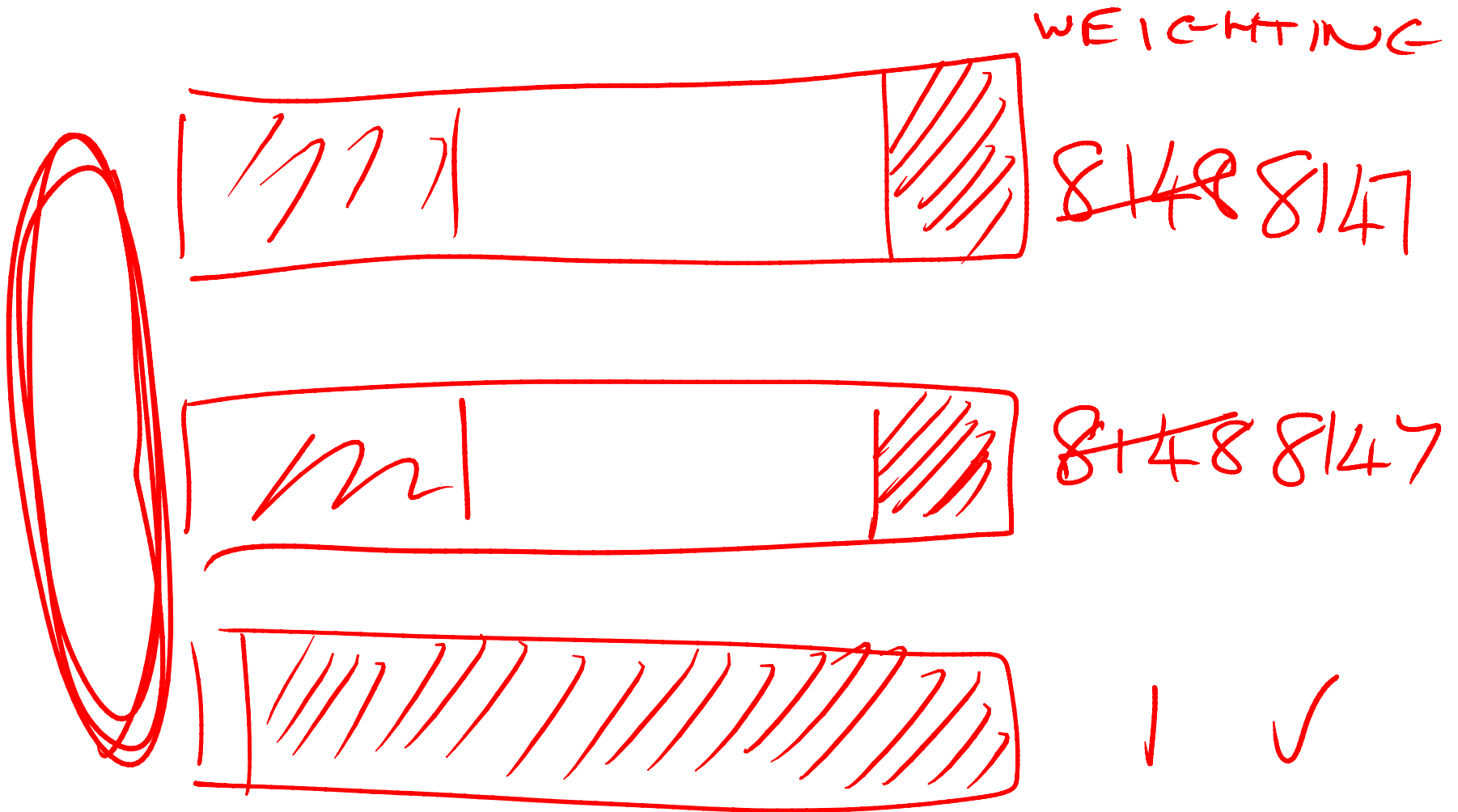
I/O SUBSYSTEM

M2S11 Explaining 2014 Buffer Pool Extension. It's like have L1-L3 caches for your data in memory. BPE sits in between memory and the I/O subsystem.



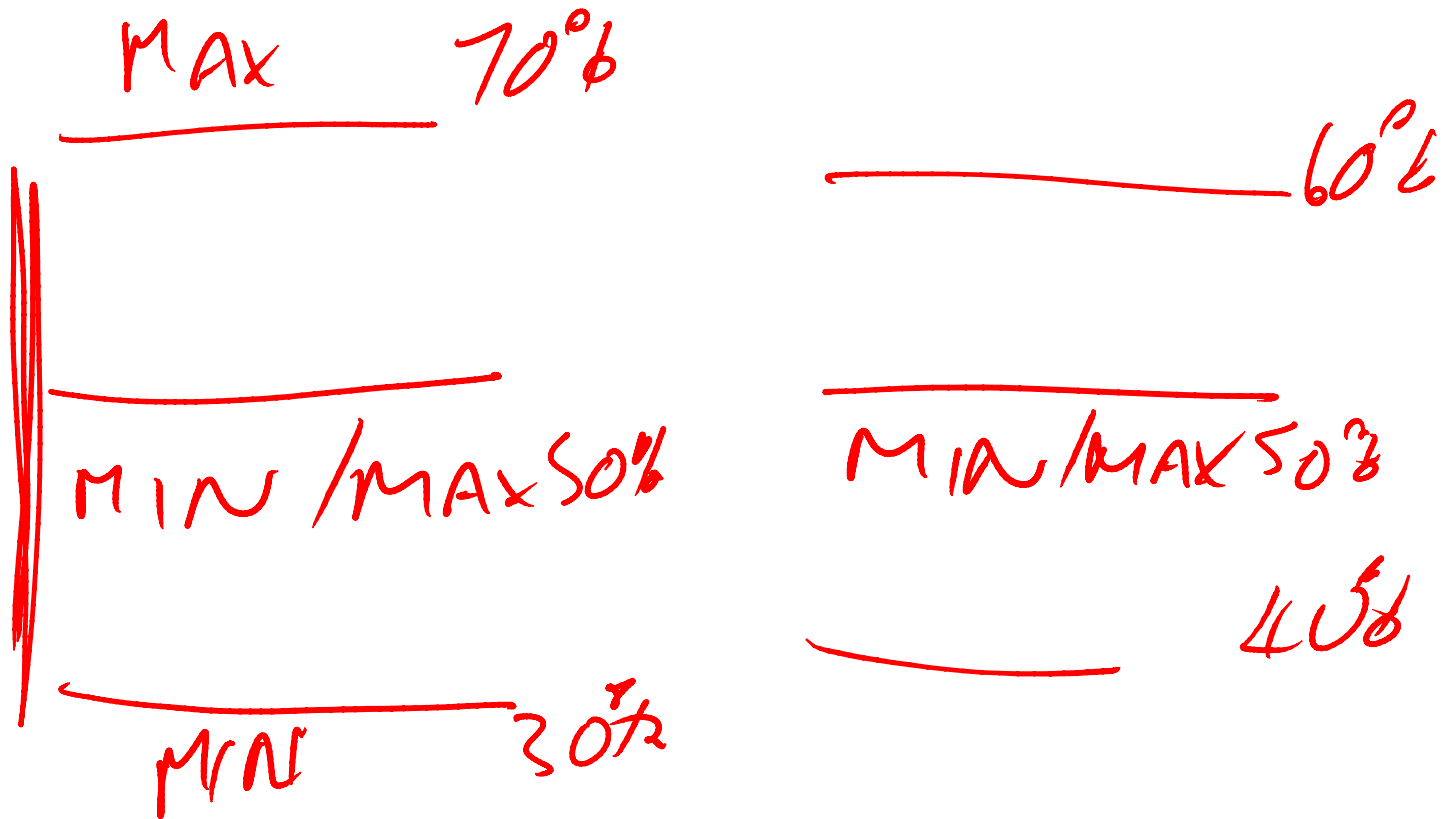
∠:

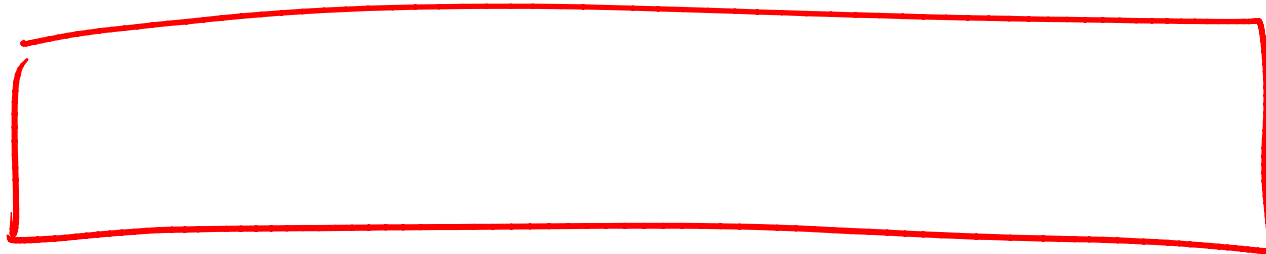
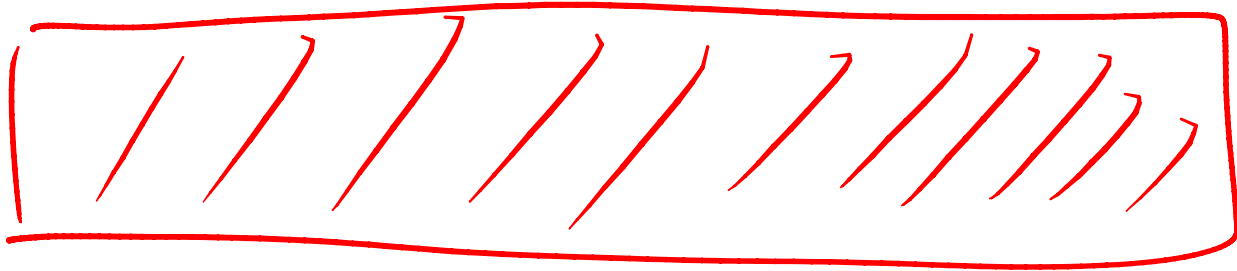
M2S14 If you put all your log files on one volume, its I/O characteristics become random



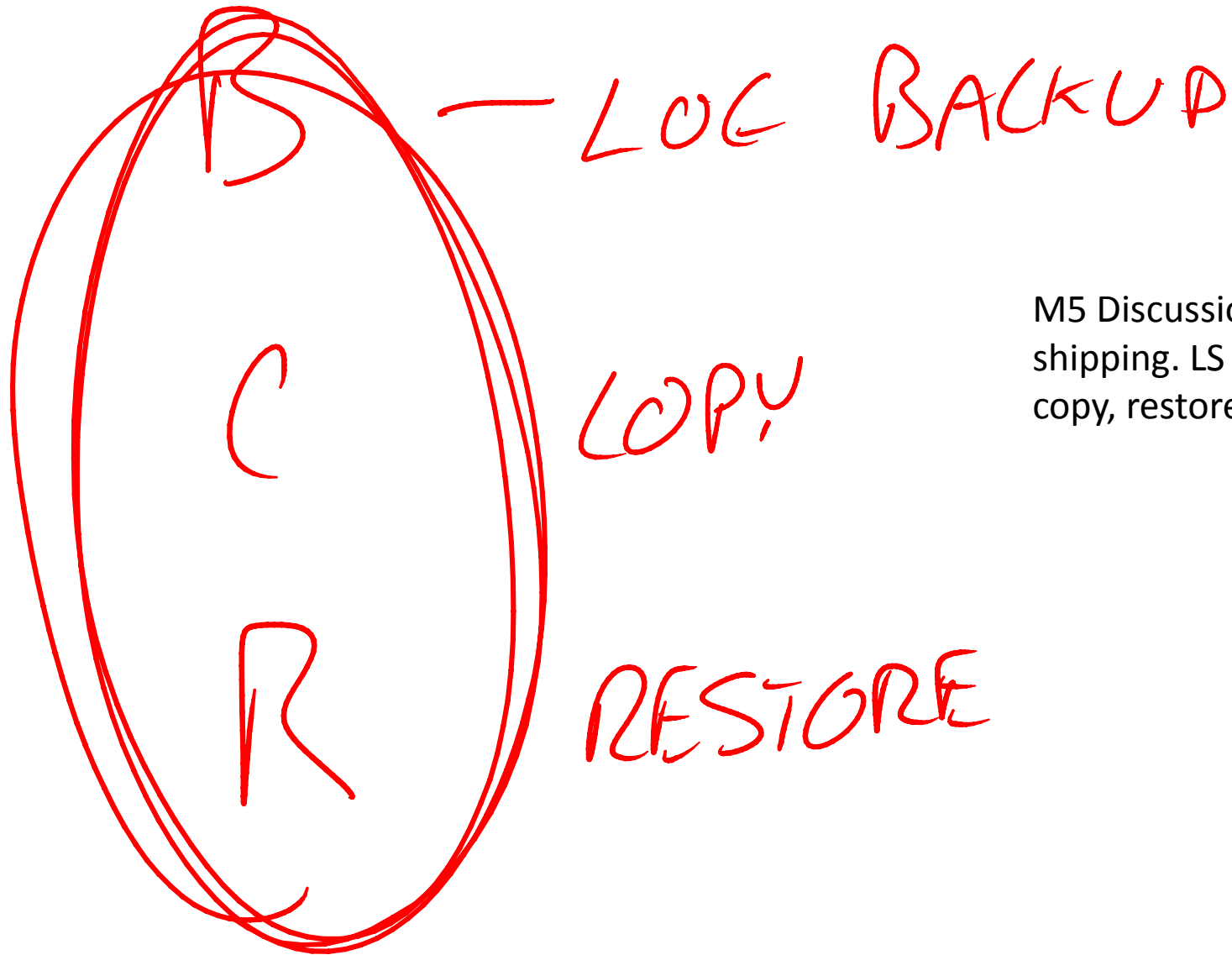
M2S20 Explaining round-robin allocation and proportional fill. Each file has a proportional fill weighting. Allocations happen proportionally more frequently from files with more free space.

Question about how to set min/max memory with two instances on a server. You could set both to have min=max=50%, but then you're wasting memory. So maybe min=30, max=70 for one, min=40, max=60 for another so there's some give/take. There is no generalized correct answer.





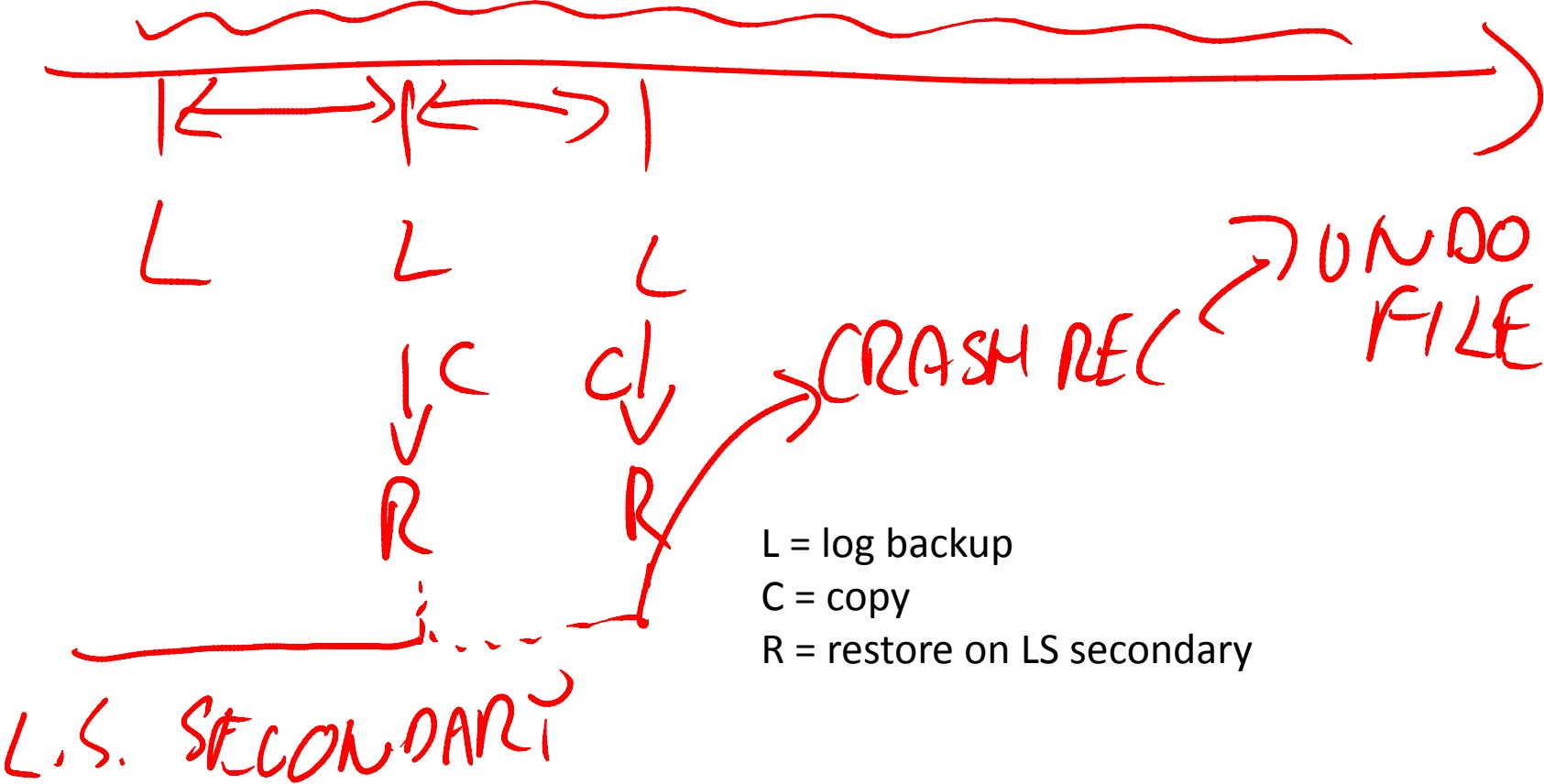
M2S63 You might add a second tempdb data file to alleviate contention and see no gain in perf. That could be because the first file is very full, and the new one very empty, so proportional fill makes all the allocations come from the new file -> no drop in allocation contention.



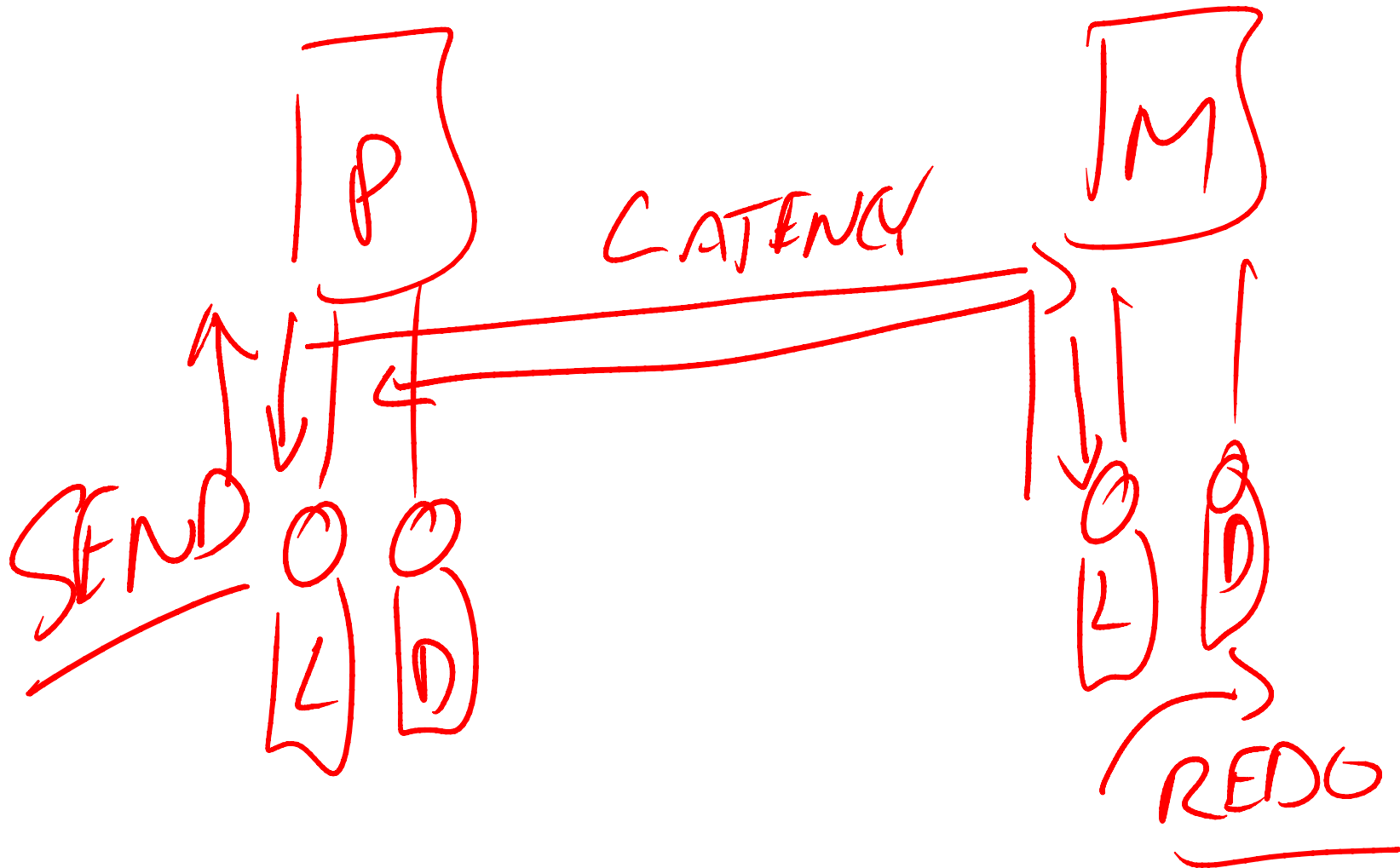
M5 Discussion on log shipping. LS = backup, copy, restore, repeat.

More on LS. When you use WITH STANDBY on the restore, it runs UNDO recovery, putting the log records into the undo file and then lets you in to the database read-only. Subsequent restore undoes the undo, throws away the undo file, and continues restoring more log.

NORECOVERY STANDBY



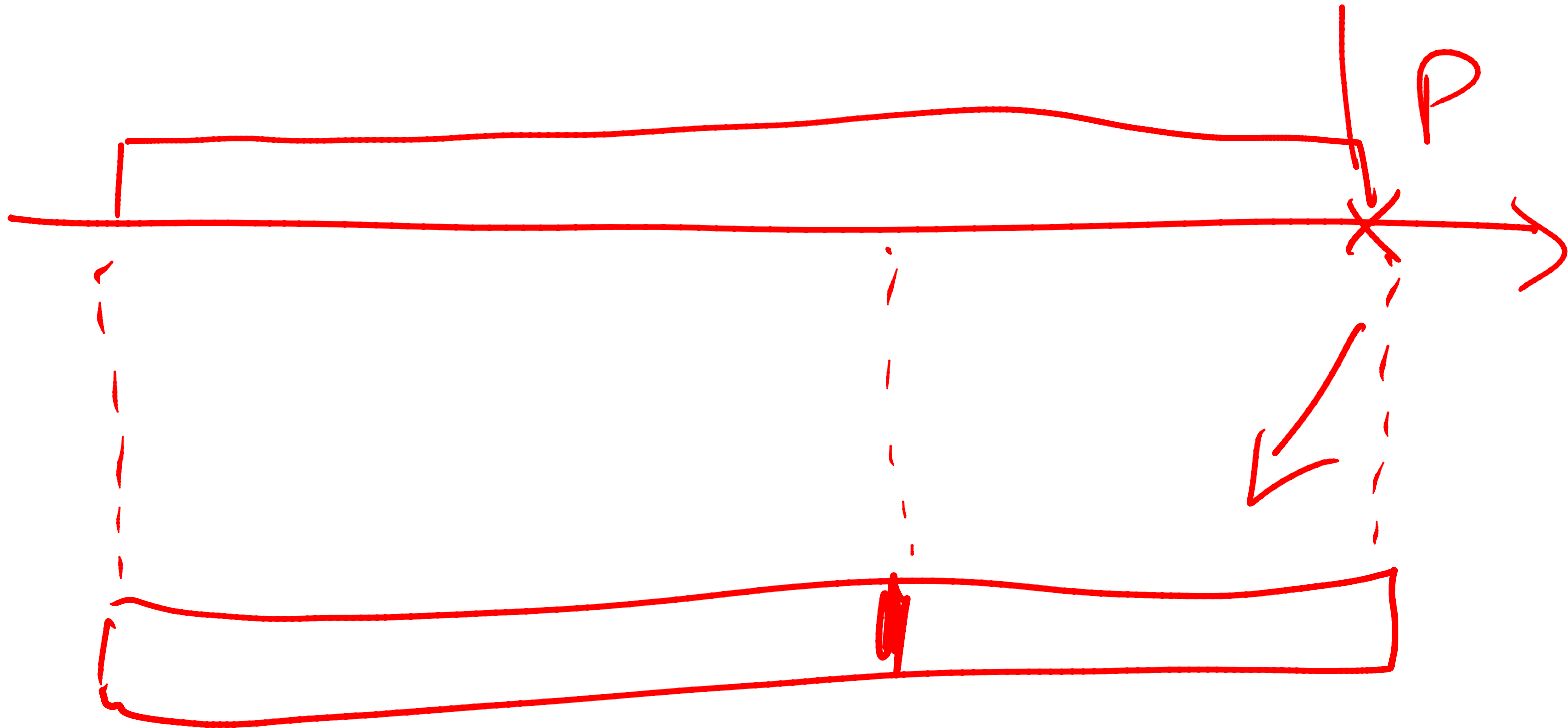
M5 discussing mirroring. Local log flush happens and completes before remote log copy. With sync mirroring/AG, transaction cannot commit until the remote log copy has completed. Remote system is constantly running redo of log records.



PASSIV

829

LSN

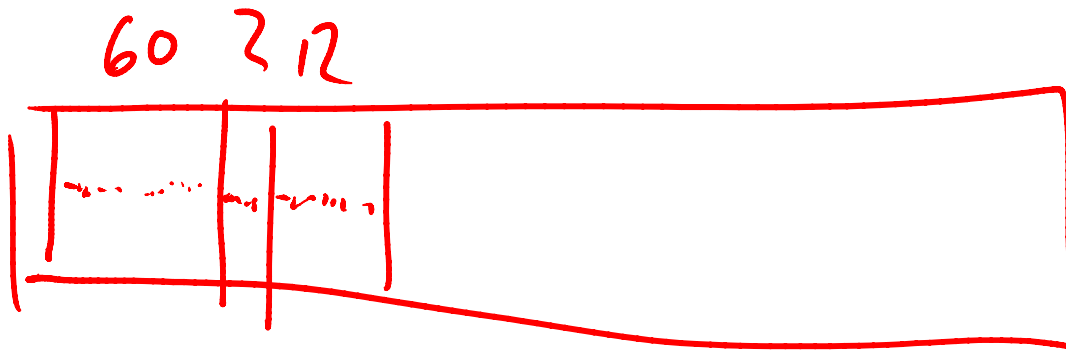
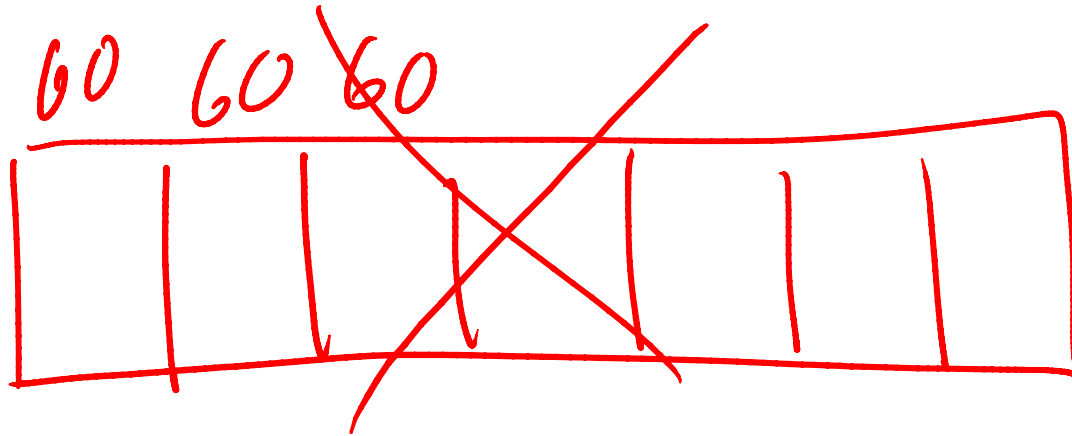


M5 discussing automatic page repair. When the principal asks the mirror for a page, the mirror has to continue running redo until it reaches the LSN at which the principal asked for the page. This makes sure it gives back the correct version of the page.



M5 The two rows with X, inserted in the xact, are X locked until the end of the transaction. Other records on the page can be read and updated, without being blocked by these two X locks. A select *, however, would block.

M5S9 (recap on Weds morning) The log blocks in a VLF are not pre-sized to 60KB. They start at 512bytes and grow as needed.



Name of a spinlock to
do with the buffer pool.

BUF_FREE-
LIST

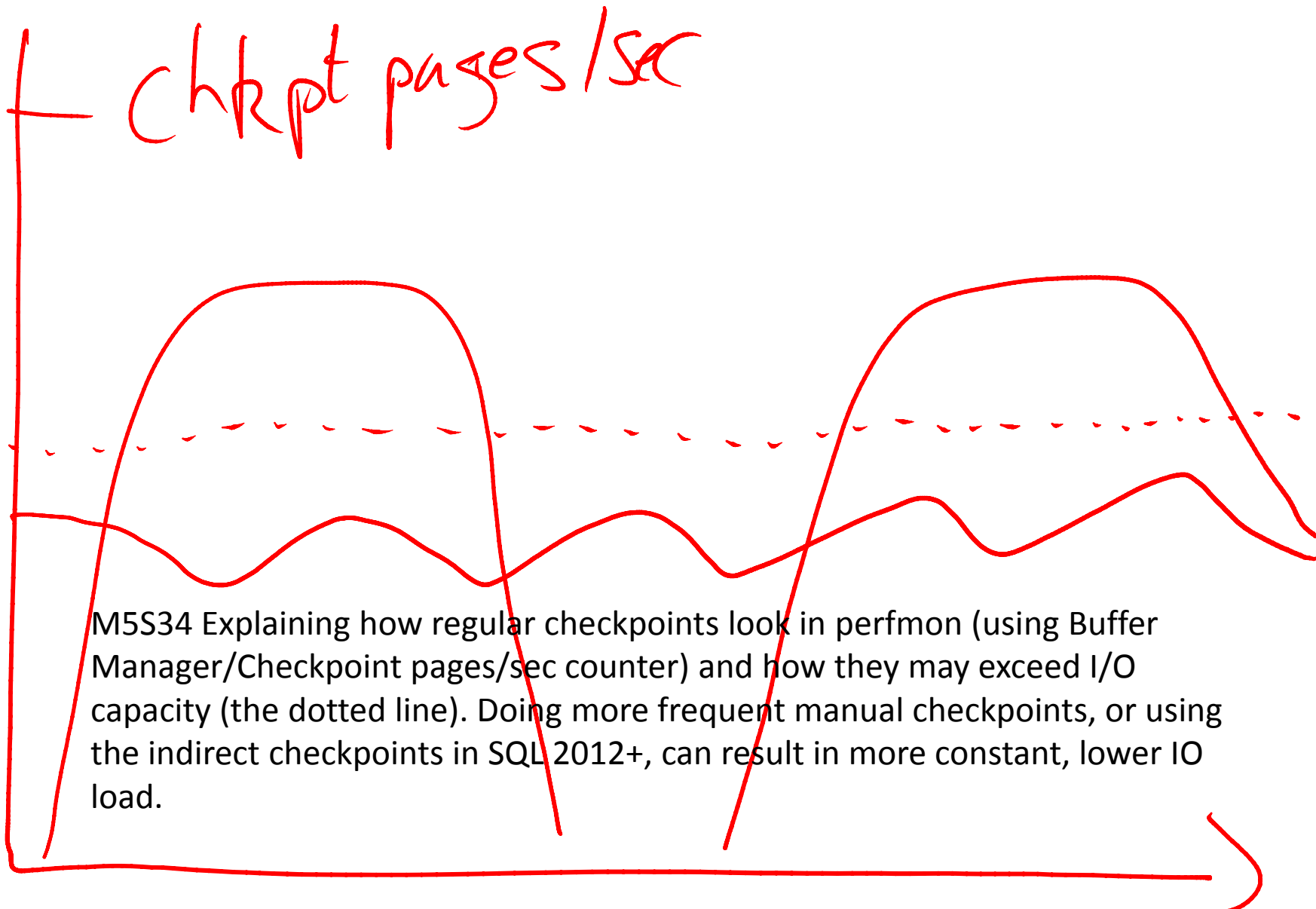
LATCH
BUF

EX
OP

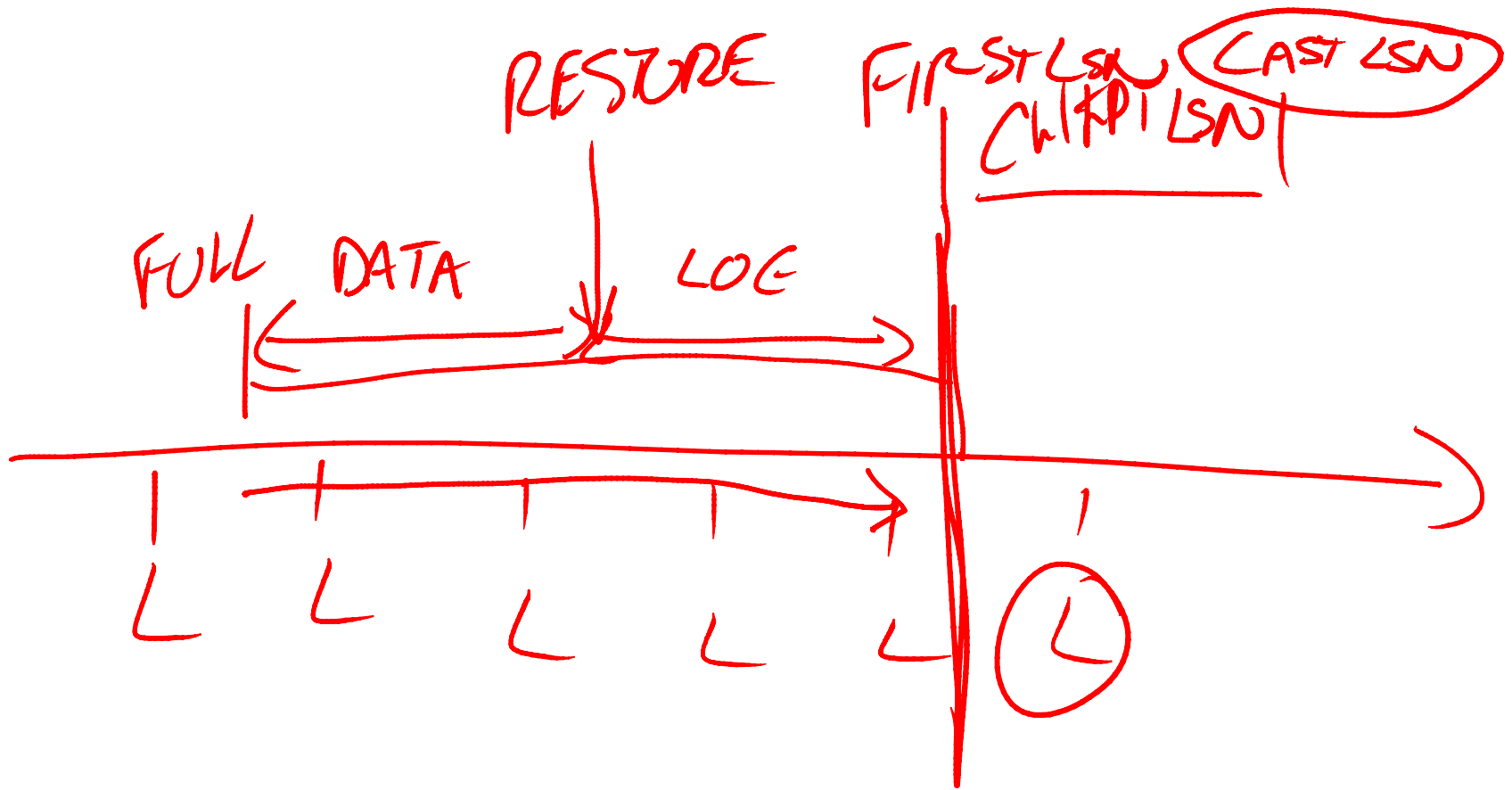
M5 Discussing latch protection
on a page being written by a
checkpoint. It's SH latched to
prevent changes while it's
being written out.



chkpt pages/sec

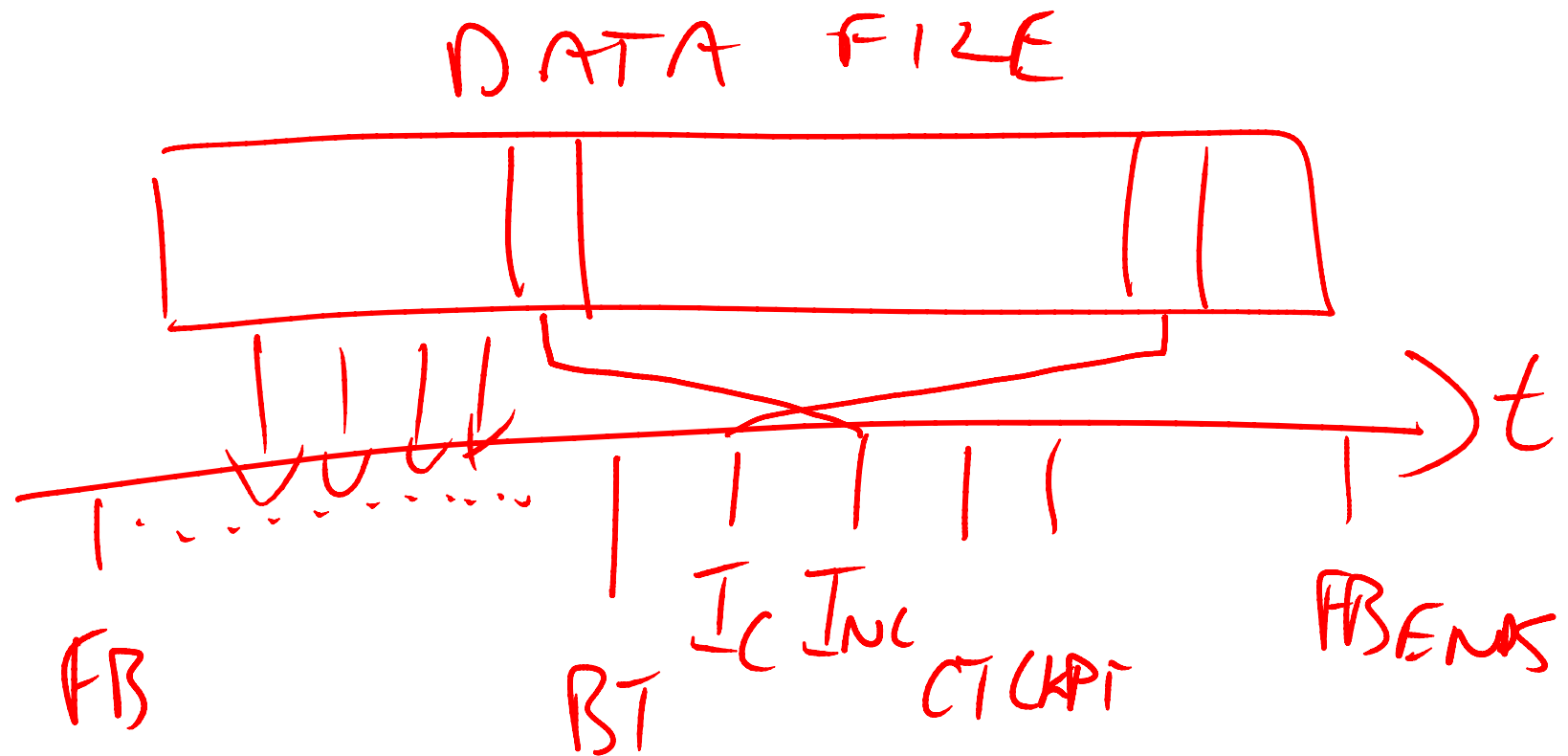


M5S34 Explaining how regular checkpoints look in perfmon (using Buffer Manager/Checkpoint pages/sec counter) and how they may exceed I/O capacity (the dotted line). Doing more frequent manual checkpoints, or using the indirect checkpoints in SQL 2012+, can result in more constant, lower IO load.



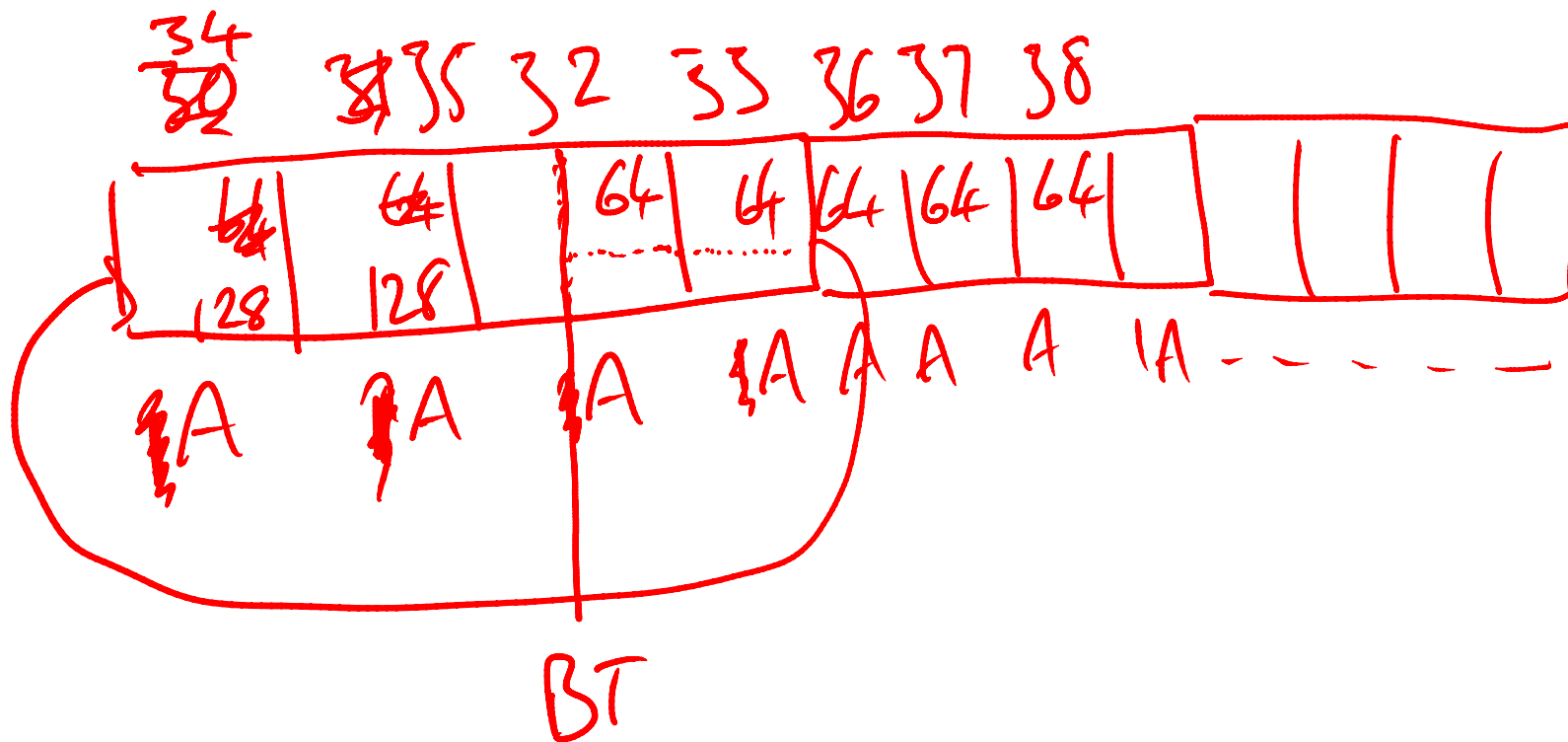
M5S46 It's possible for concurrent log backups to occur with a data backup (full or differential). Log clearing cannot occur though, as the data backup will back up the log. When the data backup finishes, deferred log clearing takes place.

The point in time you get when you restore a data backup (and don't restore any more) is the point when the data reading portion of the backup ended (marked RESTORE in the diagram).

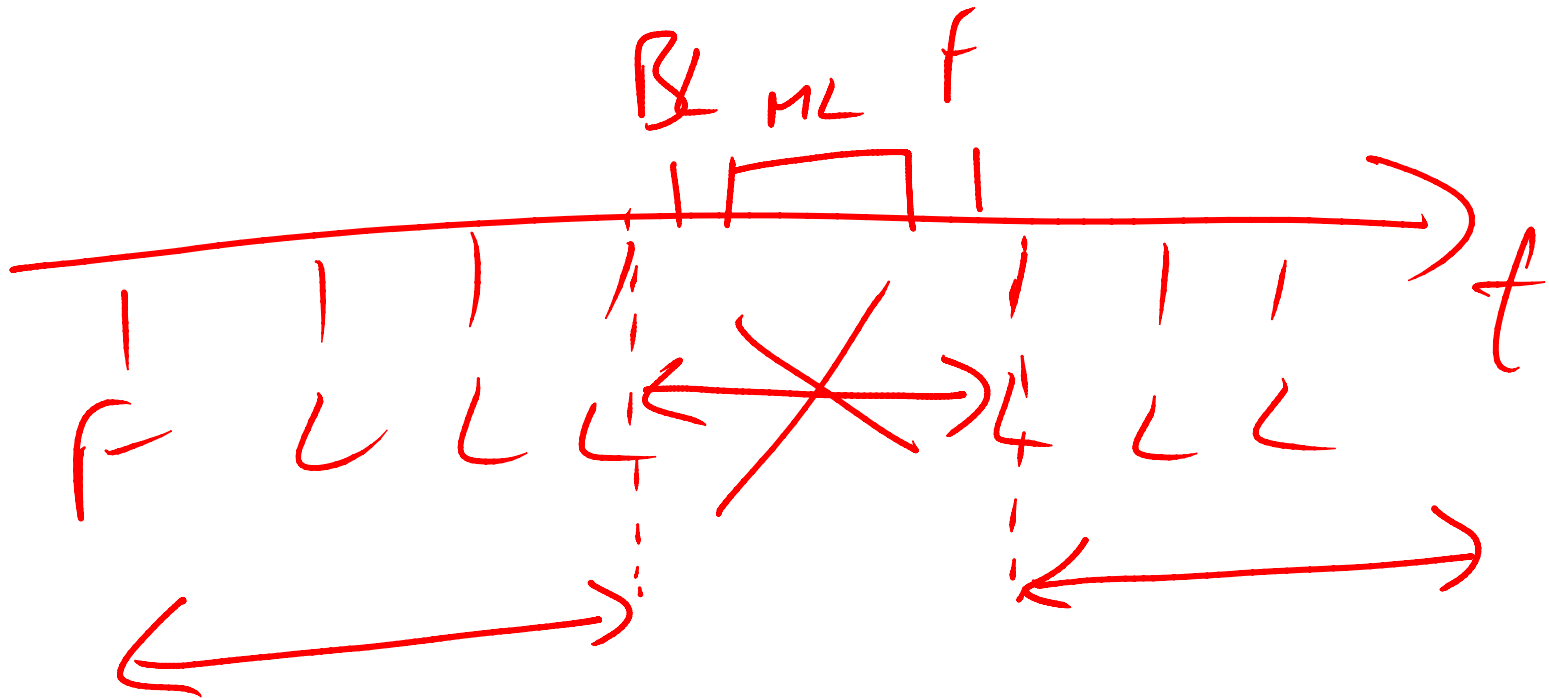


M5S46 Why does a full/diff backup contain log? It's to make sure that the restored database is transactionally consistent.

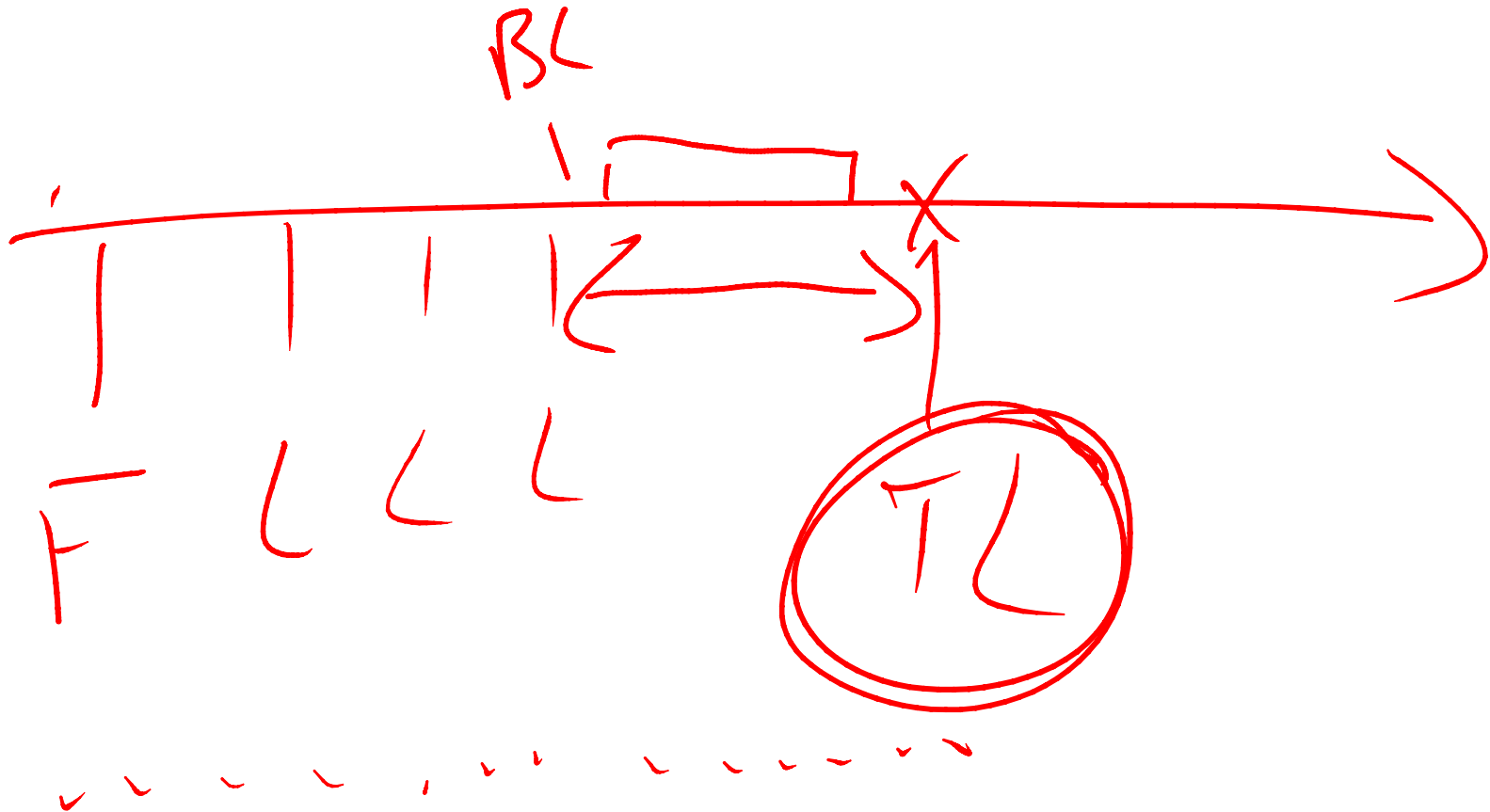
Scenario: full backup starts and reads nonclustered index page before insert. Transaction starts and inserts a row into clustered index page (Ic) and nonclustered index page (Inc). Checkpoint flushes out both pages. Backup then reads clustered index page with insert. If no log in the backup, restored database would be corrupt – missing nonclustered index record.



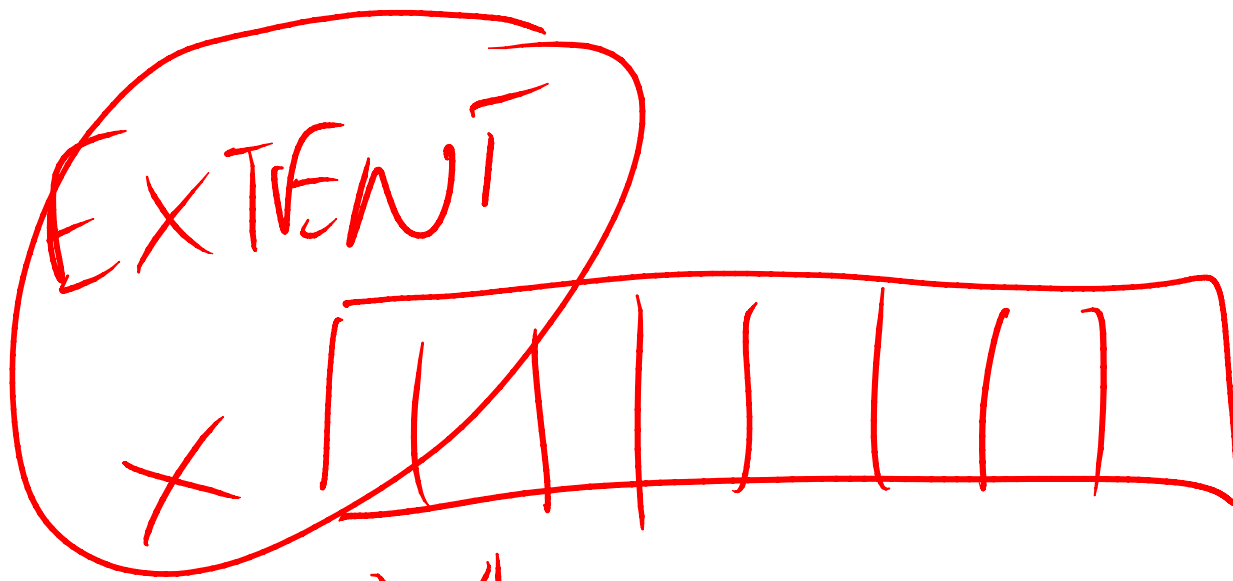
M5S52 Explaining the circular log demo. We discussed log space reservation that happens to ensure the active transactions can roll back without having to grow the log – this accounts for extra log growths. Space reserved in the log does not make a VLF become active – as there are no log records written out, just empty space reserved.



M5S64 If a log backup contains a minimally-logged operation, it cannot be used to do a point-in-time restore using STOPAT for any point in time between the beginning and end of the backup. It can be used as part of a restore sequence as normal though.



M5S64 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). You also 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+, you can, but the log backup is invalid.

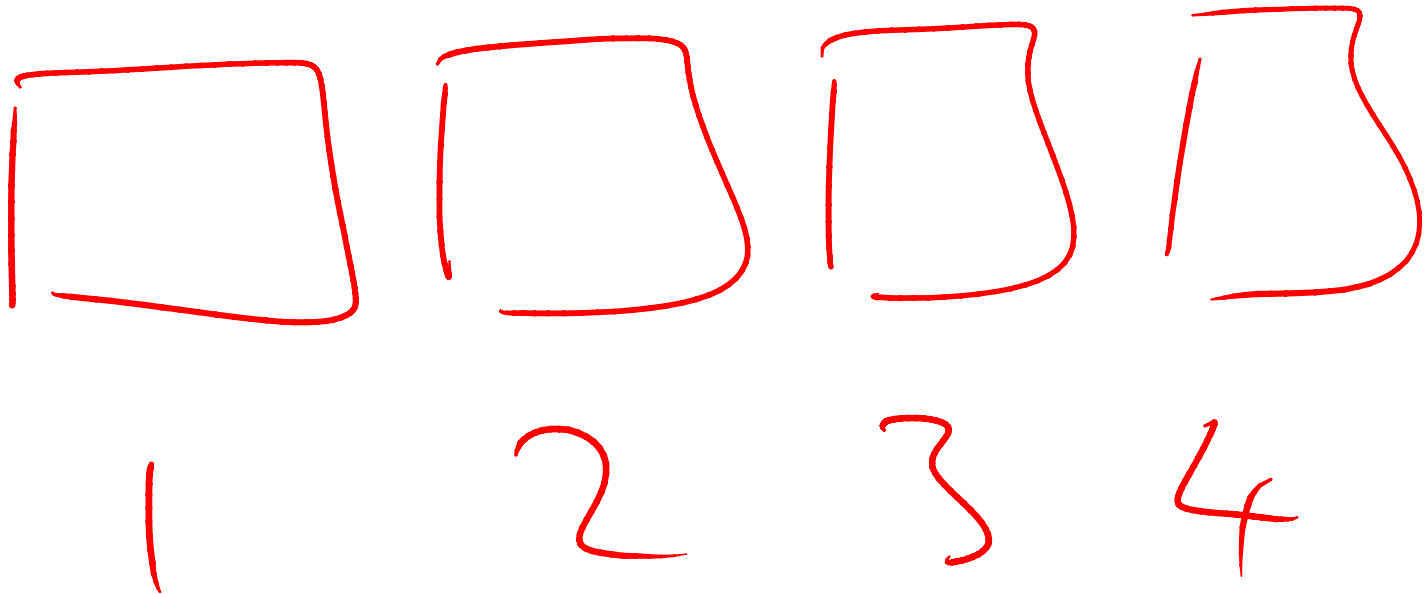


PROBE

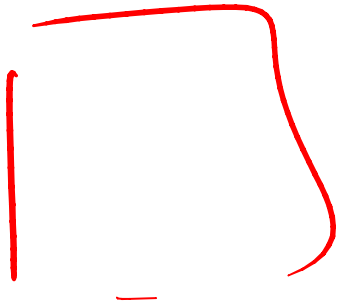
M5S66 Explaining part of the deferred-drop functionality. Before an extent can be deallocated, an X lock is acquired on it and all 8 page locks are probed (instant acquire X lock and release) to make sure nothing else has them locked.



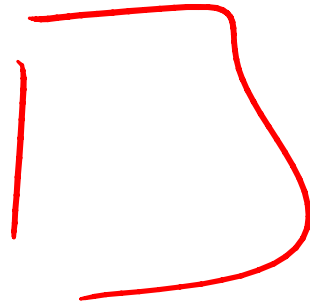
MSS67 VLF fragmentation is bad because of having to search through the VLFs for a particular VLF sequence number. See the KB article link for more info.



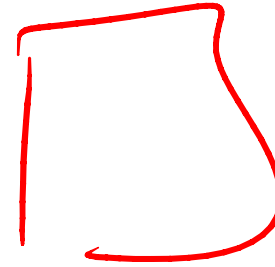
M7S11 These pages are in the same logical order as physical order



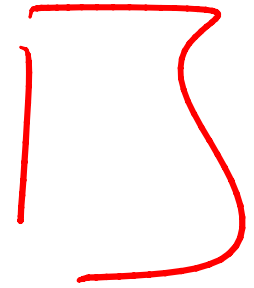
1



46



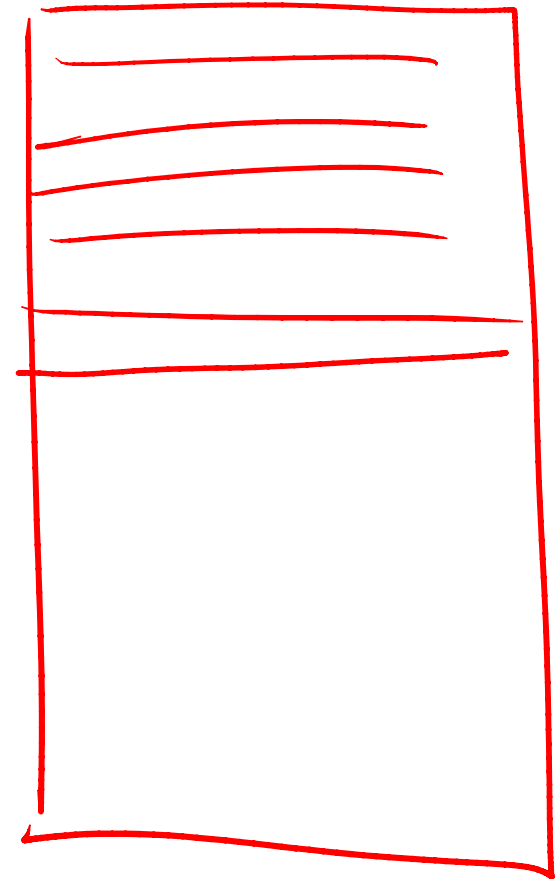
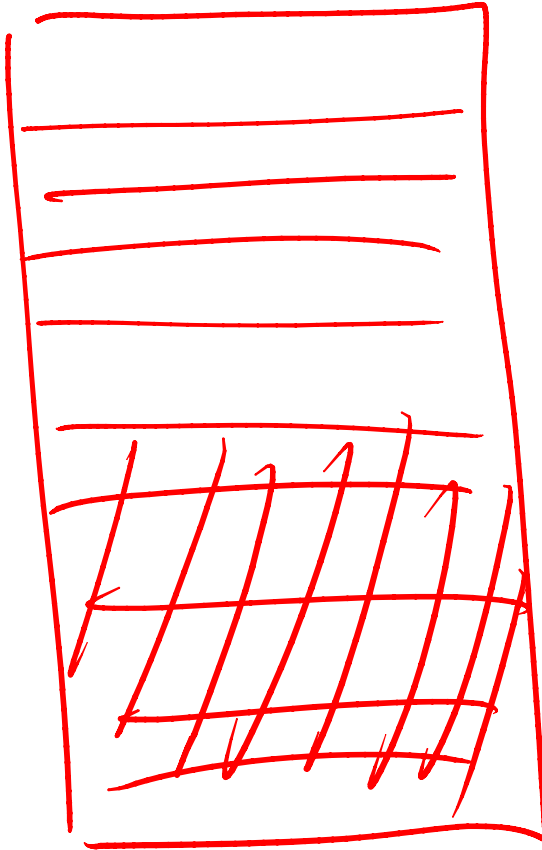
2



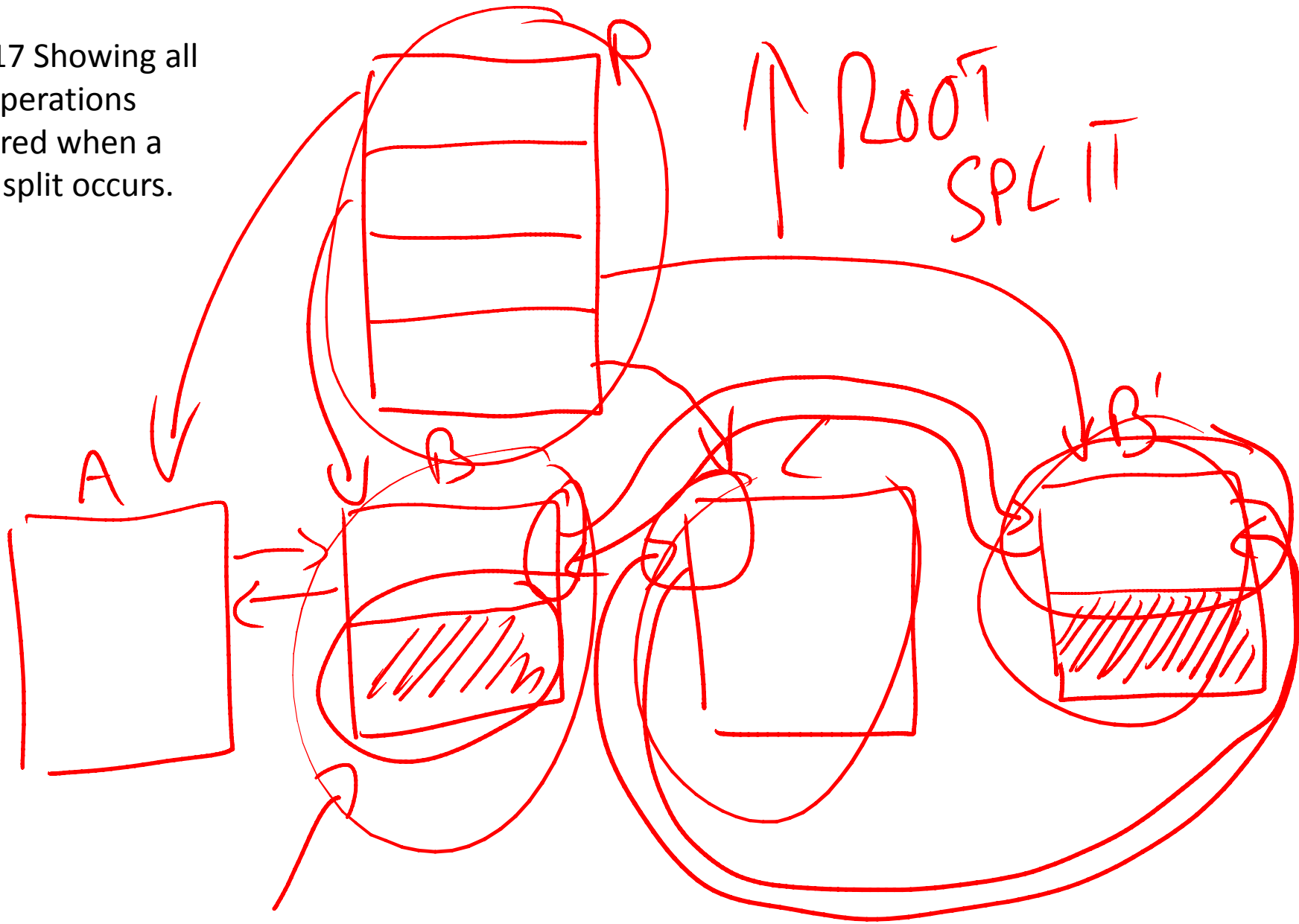
374

M7S11 These pages are not in the same logical order as physical order

M7S16 When a page split occurs, it's usually 50/50 and creates 50% free space on each page.



M7S17 Showing all the operations required when a page split occurs.



BT

MS

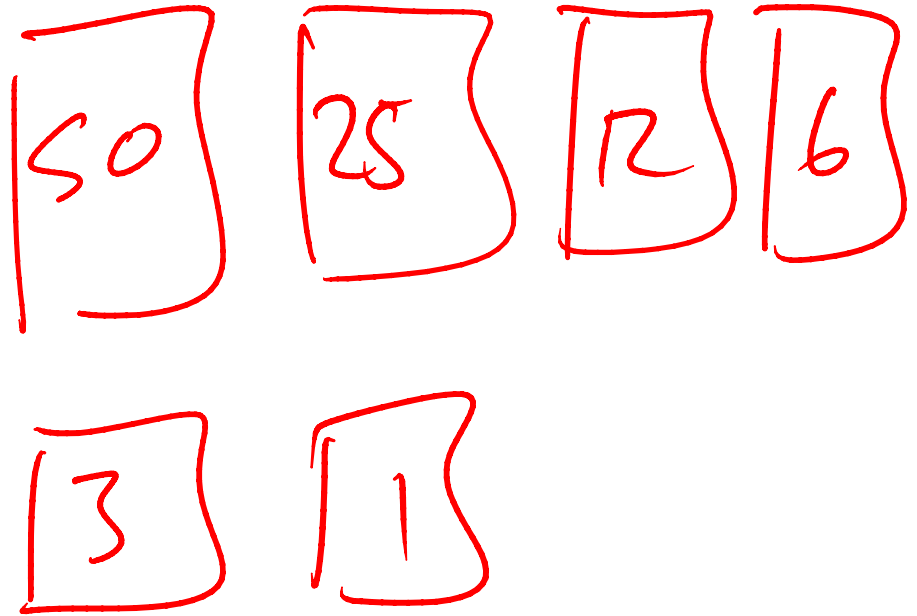
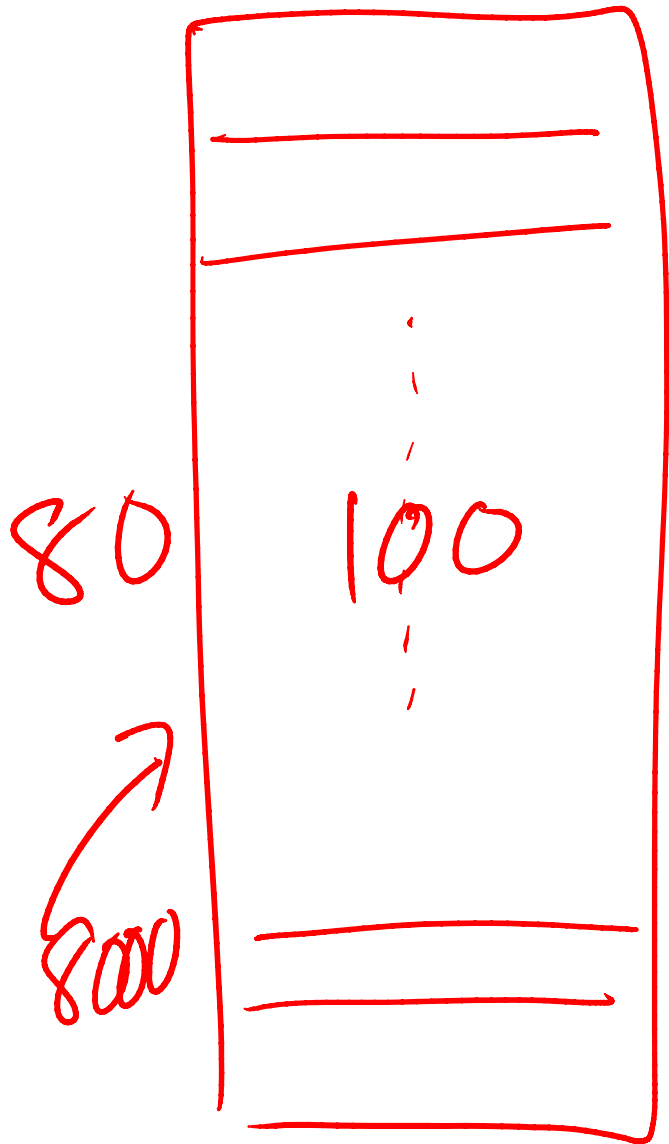
BT

~

CT

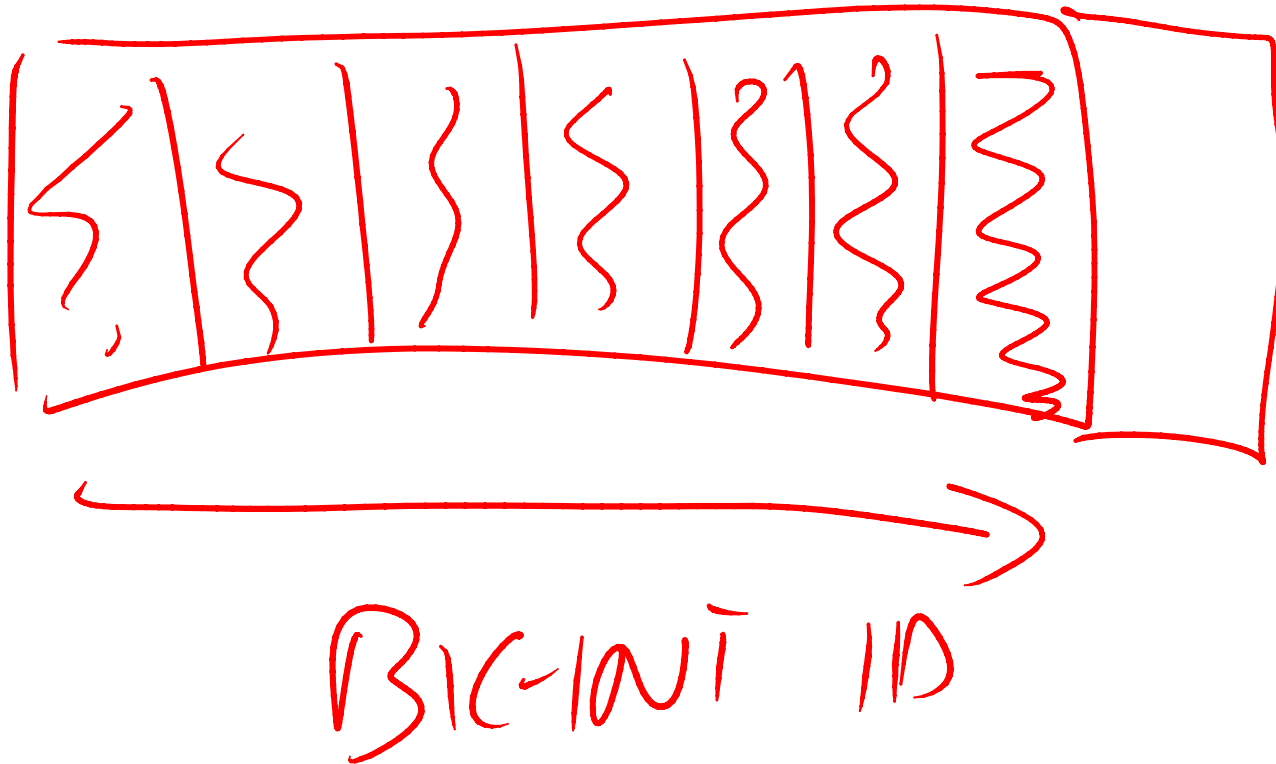
SYSTEM XACT

M7S17 Explaining how the page split is done under the covers as a system transaction, owned by your transaction. Once committed, the system transaction will not roll back if your transaction does – so the page split is permanent.



M7S18 Explaining how the Storage Engine is not smart sometimes. Imagine a page full of 80-byte rows. Inserting an 8000-byte row there will lead to 6 splits before enough space is created.

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.

	NO SPLIT	SPLIT	
1000	1256	7580	x5
100	356	6772	x18
10	268	11172	x45

M7S23 The numbers from the page split logging cost demo.

COMMENTS

PAUL	JON	ERIN	ELENA
MEMBER_ID			K

M7S24 The MySpace story. The 'K' is the tiny portion of comments on Kimberly's page because she wasn't very popular back then...



ITB



M7S34 Explaining staggered index maintenance with database mirroring. See <http://bit.ly/IS04W5> for more explanation