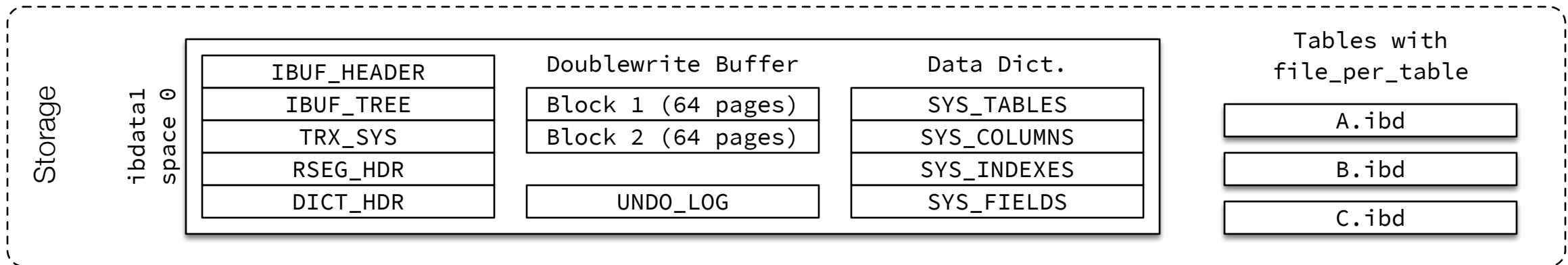
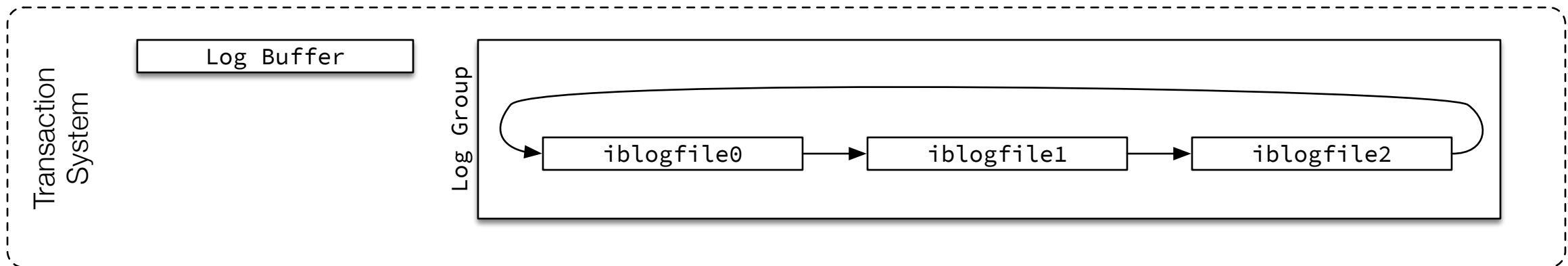
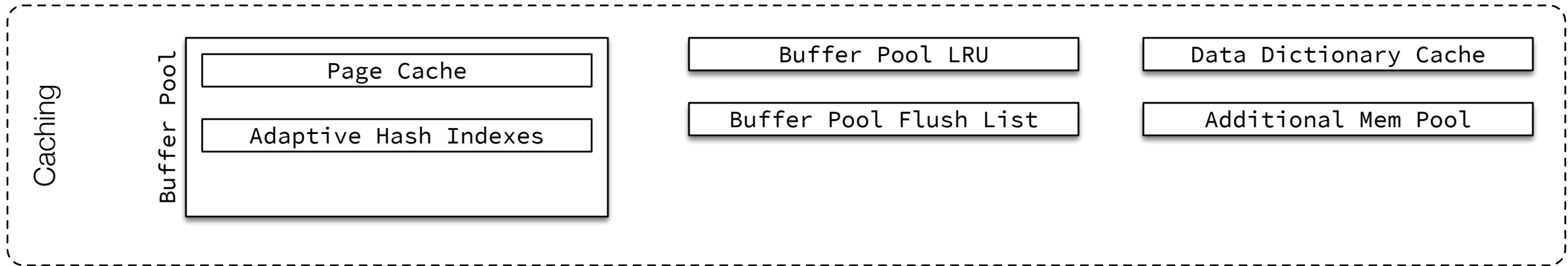
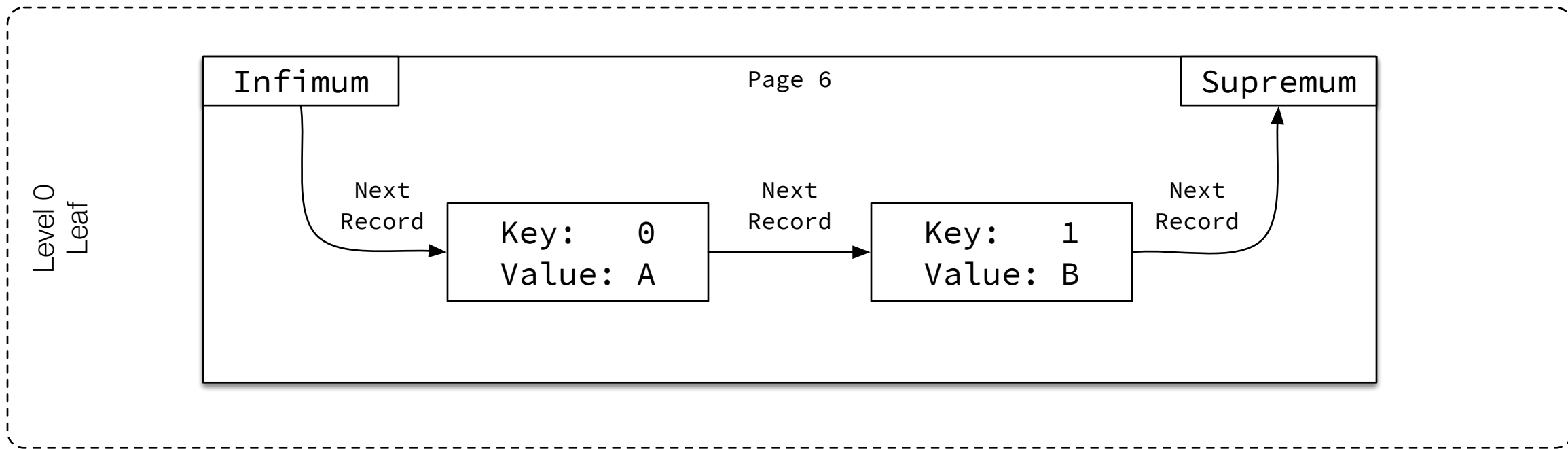


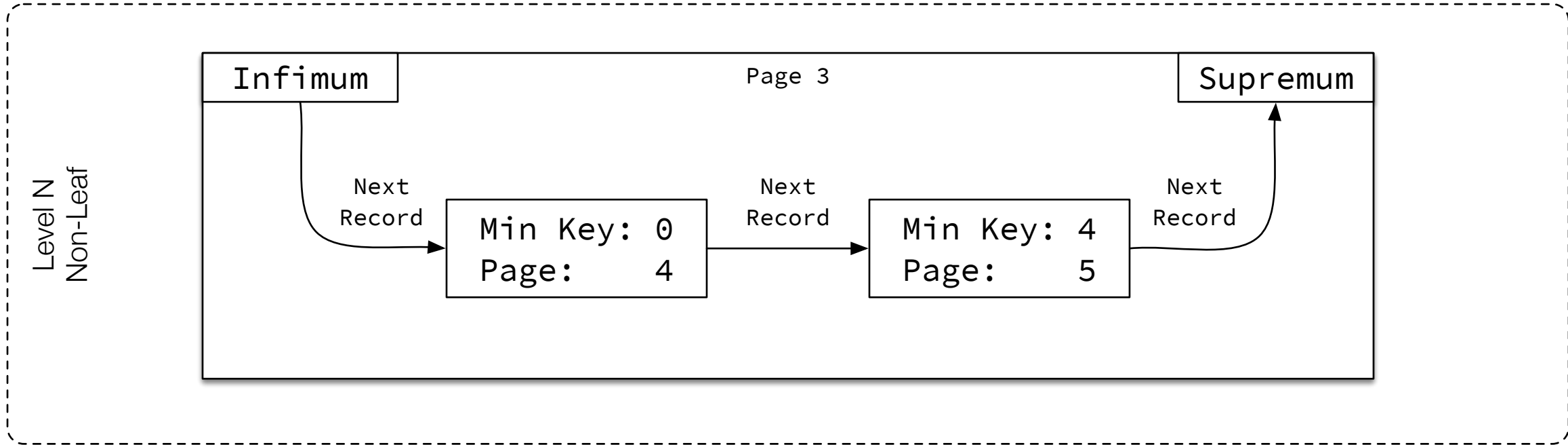
# High-level Overview



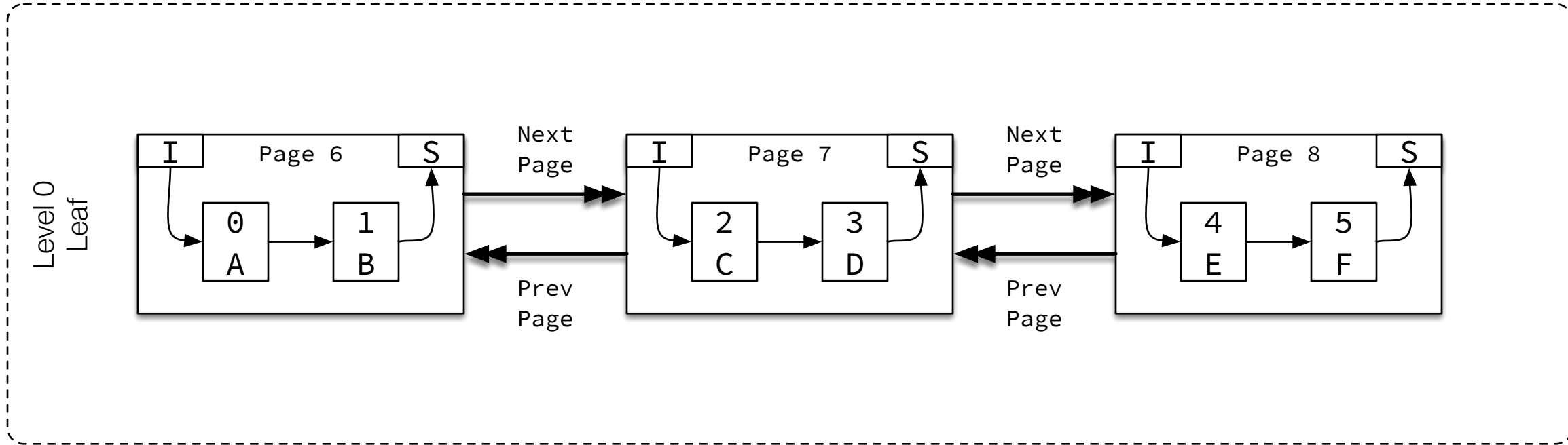
# B+Tree Simplified Leaf Page



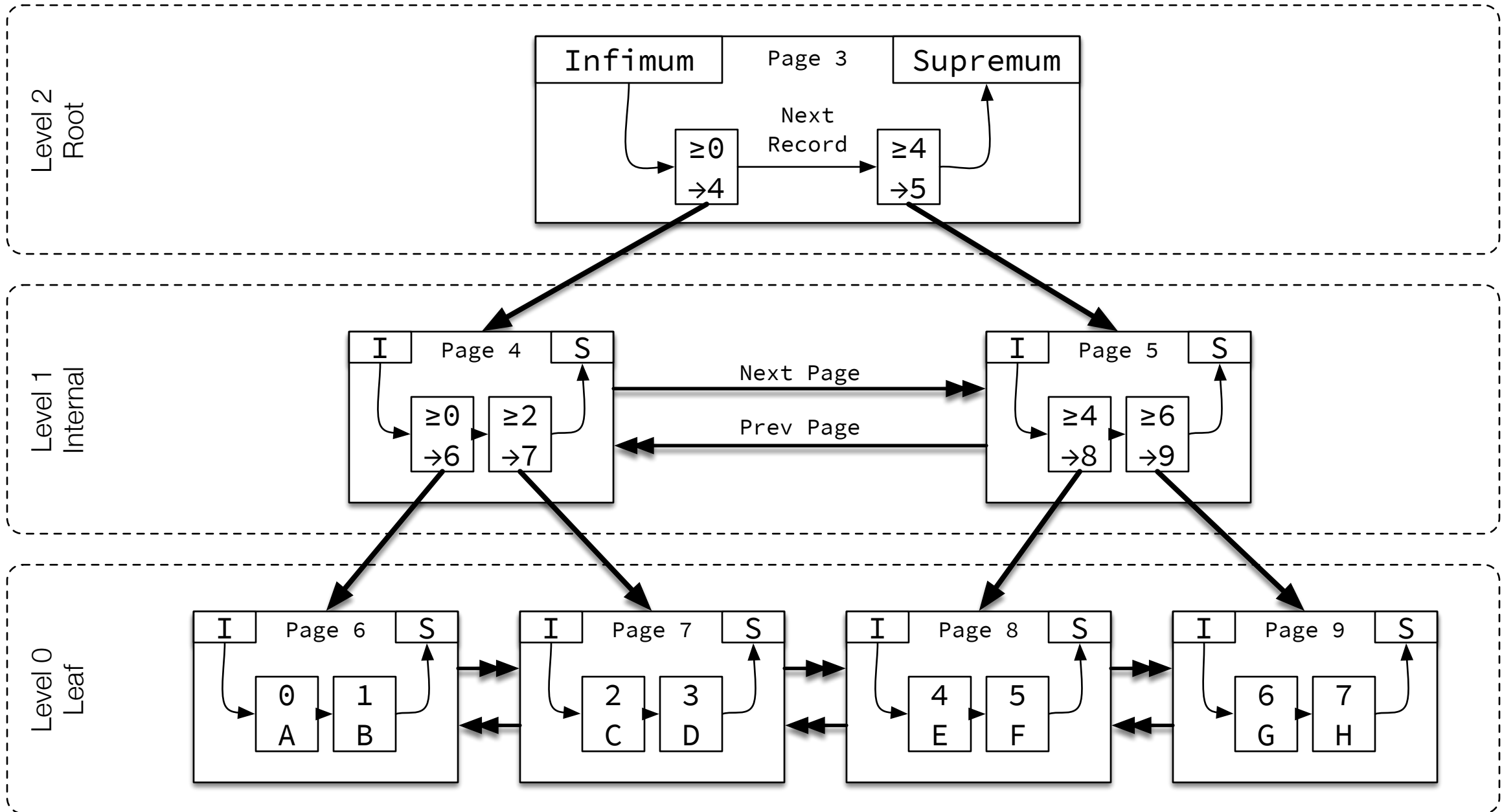
# B+Tree Simplified Non-Leaf Page



# B+Tree Simplified Level



# B+Tree Structure



Levels are numbered starting from 0 at the leaf pages, incrementing up the tree.

Pages on each level are doubly-linked with previous and next pointers in ascending order by key.

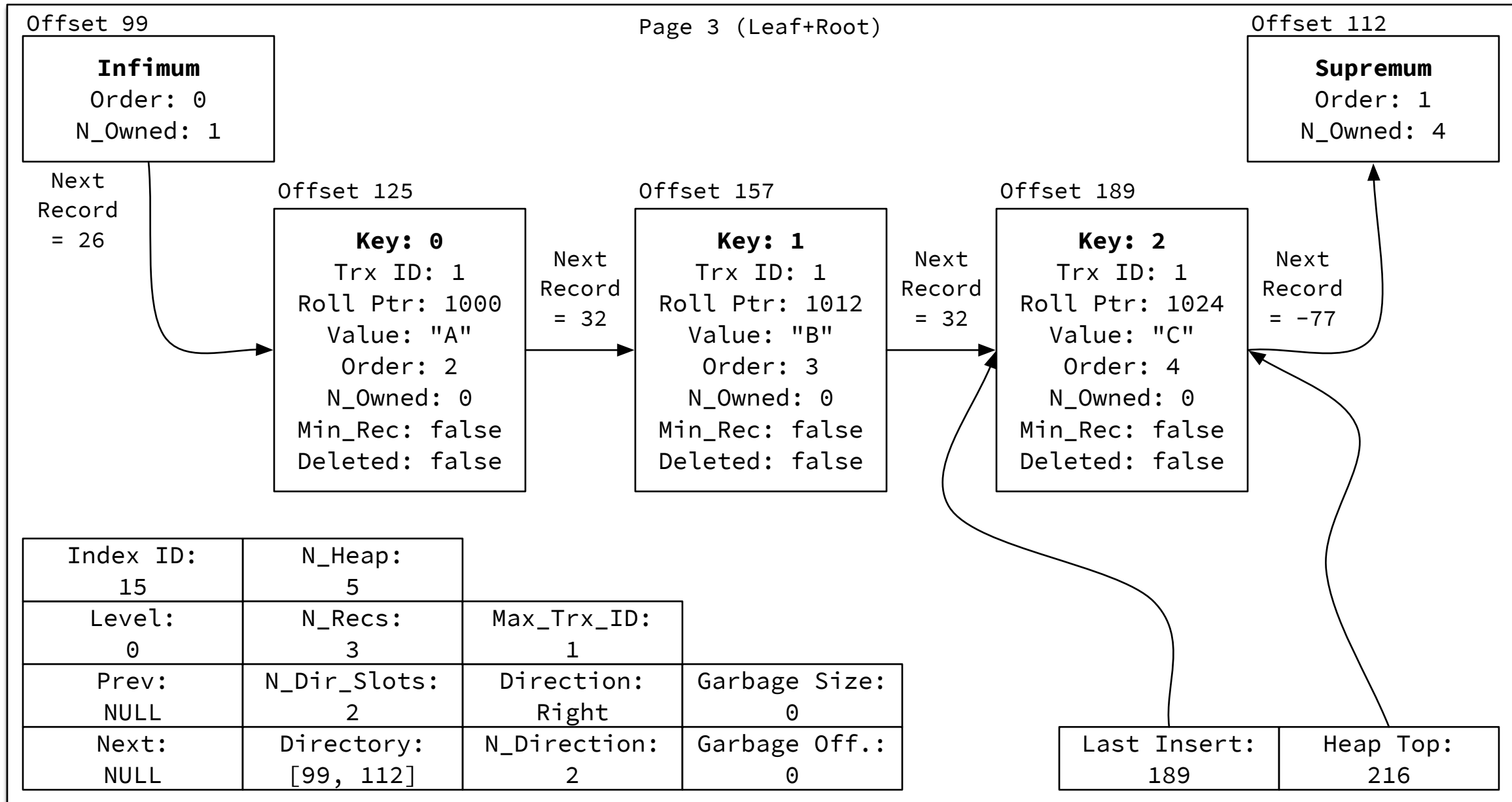
Records within a page are singly-linked with a next pointer in ascending order by key.

Infimum represents a value lower than any key on the page, and is always the first record in the singly-linked list of records.

Supremum represents a value higher than any key on the page, and is always the last record in the singly-linked list of records.

Non-leaf pages contain the minimum key of the child page and the child page number, called a "node pointer".

# B+Tree Detailed Page Structure



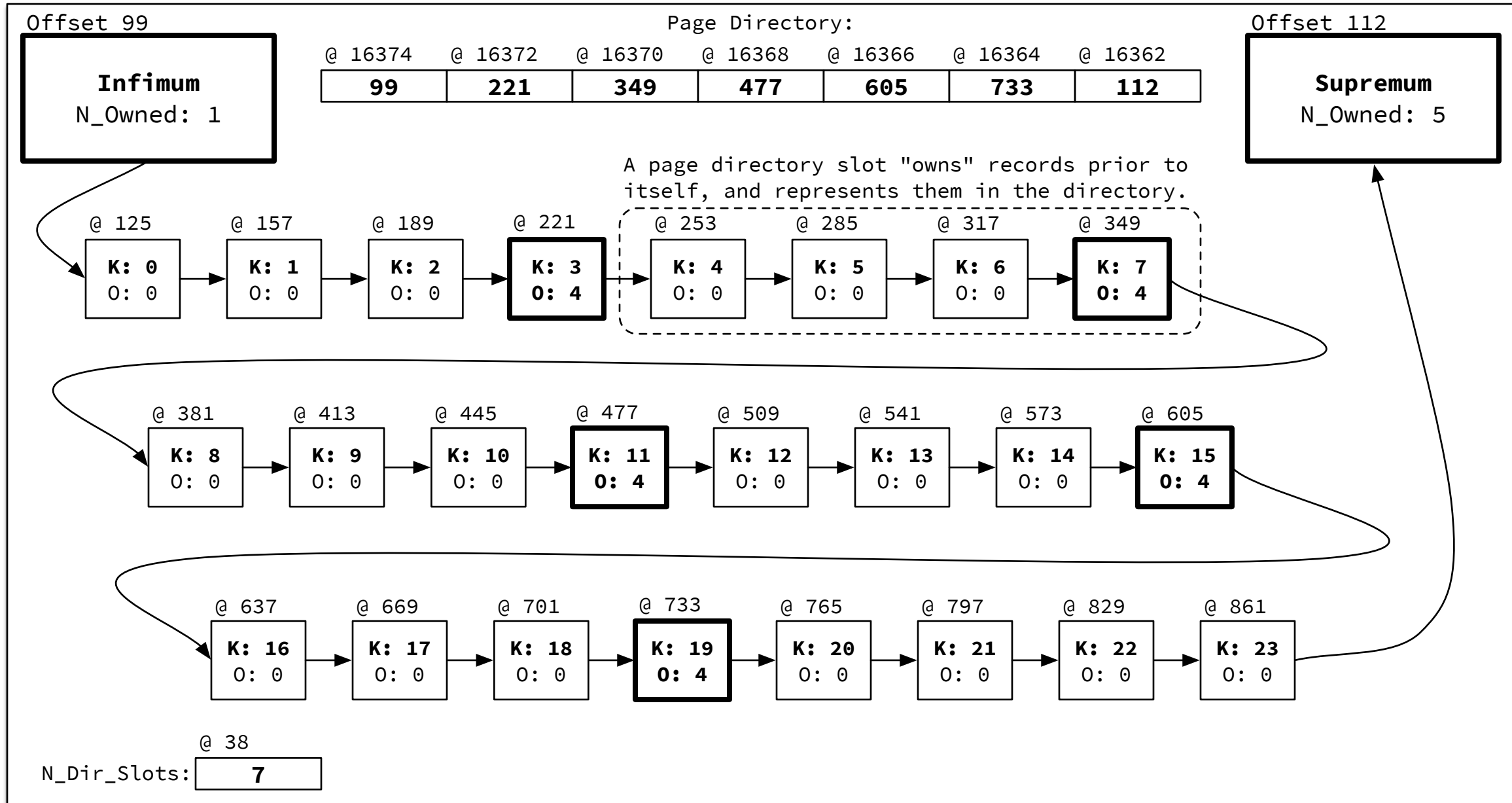
InnoDB table format is Barracuda with "compact" record structure, non-compressed.

Table created with: `CREATE TABLE t (i INT NOT NULL, s CHAR(10) NOT NULL, PRIMARY KEY(i)) ENGINE=InnoDB;`

Table populated with: `INSERT INTO t (i, s) VALUES (0, "A"), (1, "B"), (2, "C");`

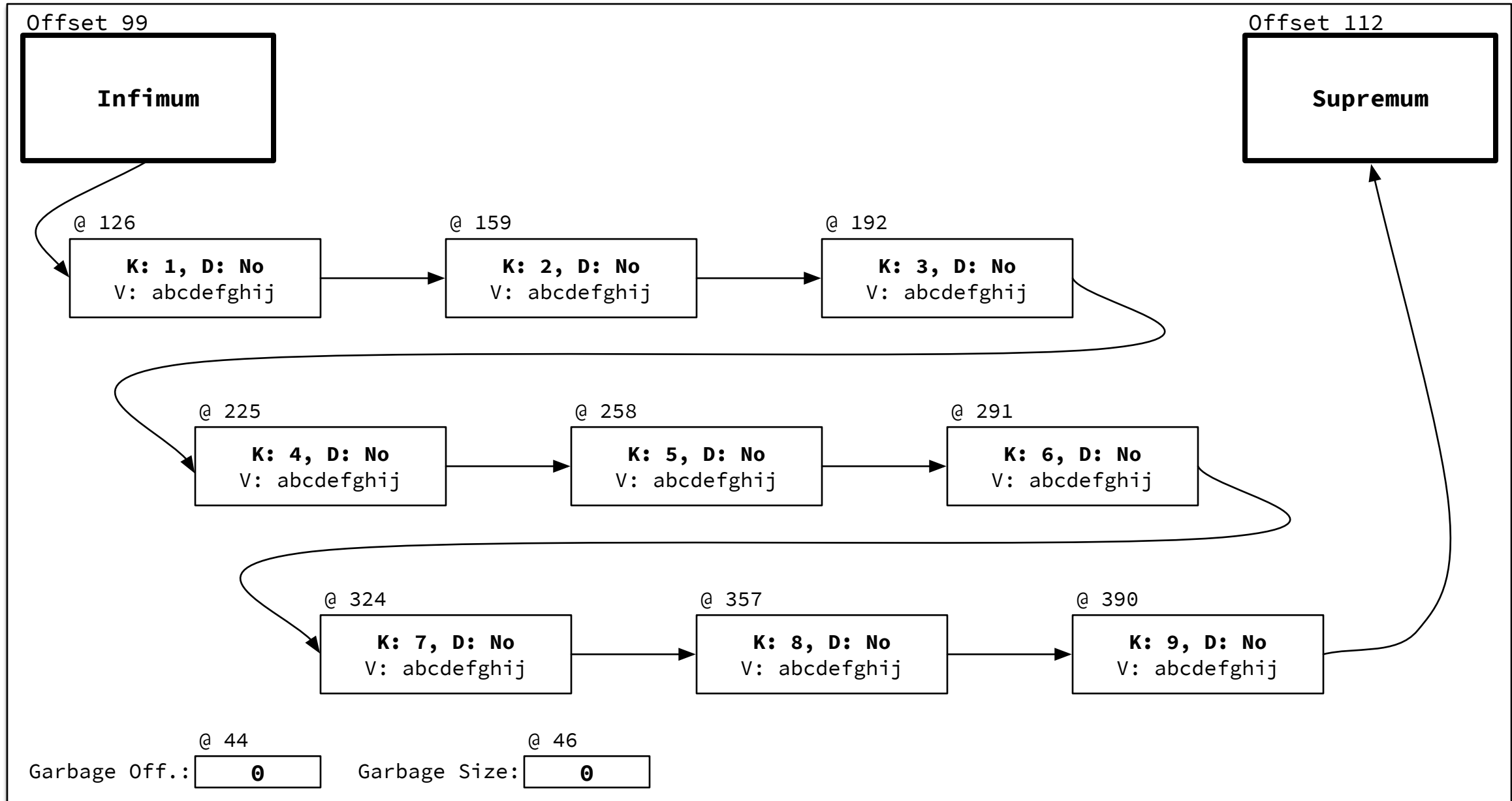
Record size: 5 (header) + 4 (PK) + 6 (TRX\_ID) + 7 (ROLL\_PTR) + 10 (non-key fields) = 32 bytes

# B+Tree Page Directory Structure



Infimum always owns only itself, so will always have a slot in the page directory with N\_Owned = 1.  
 Supremum always owns the last few records in the page, and is allowed to own less than 4 records (if the page has fewer).  
 All directory slots will own a minimum of 4 and maximum of 8 records, except supremum, which may own fewer.  
 The page directory grows "downwards" from offset 16376, the beginning of the FIL trailer; the first directory entry starts at 16374.

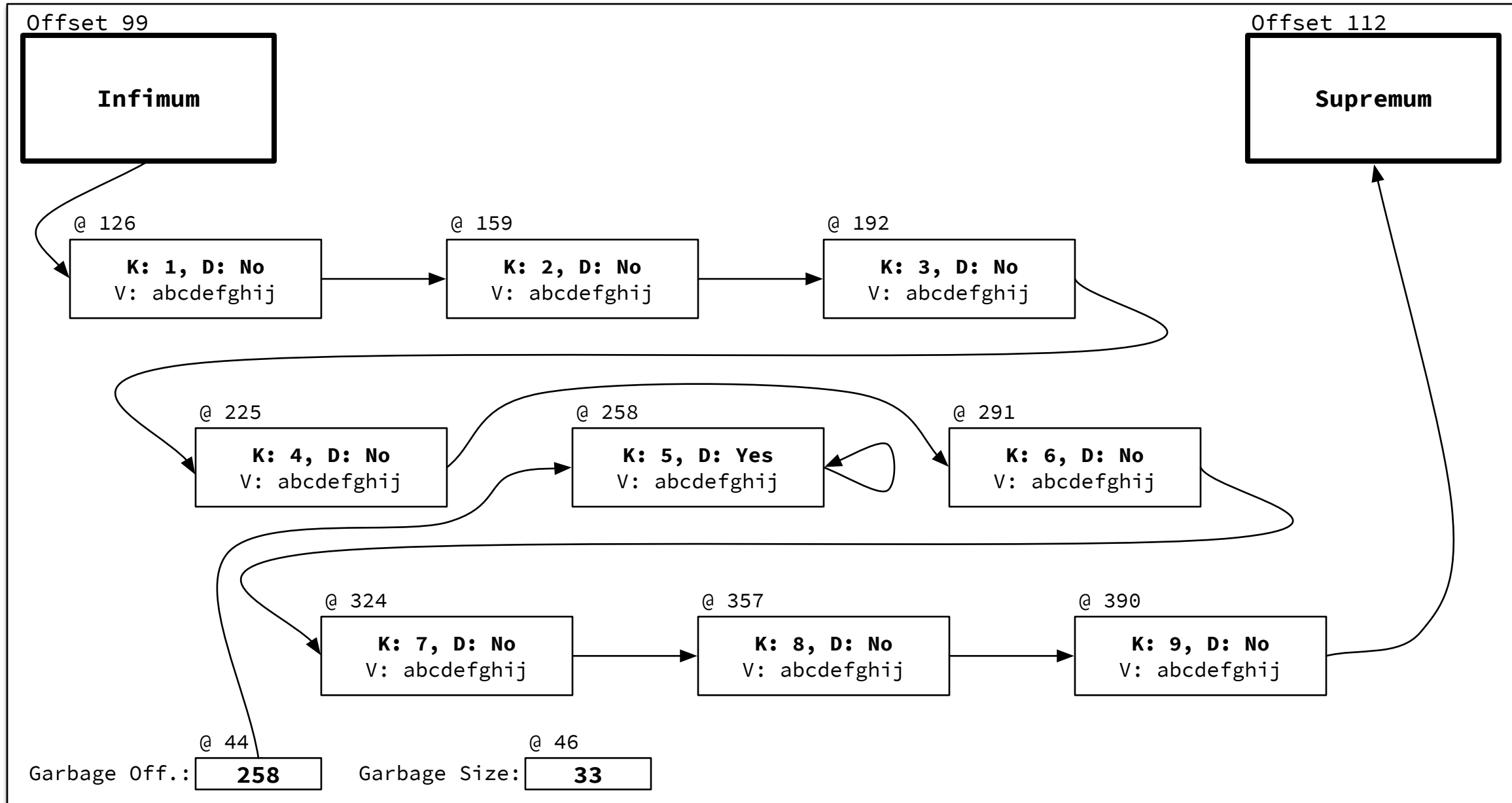
# B+Tree Record Initial State



SQL: create table t (i int not null, s varchar(100) not null, primary key(i)) engine=innodb;  
SQL: insert into t (i, s) values (1, "abcdefghij"); for i in 1..9



# B+Tree Record Delete 1



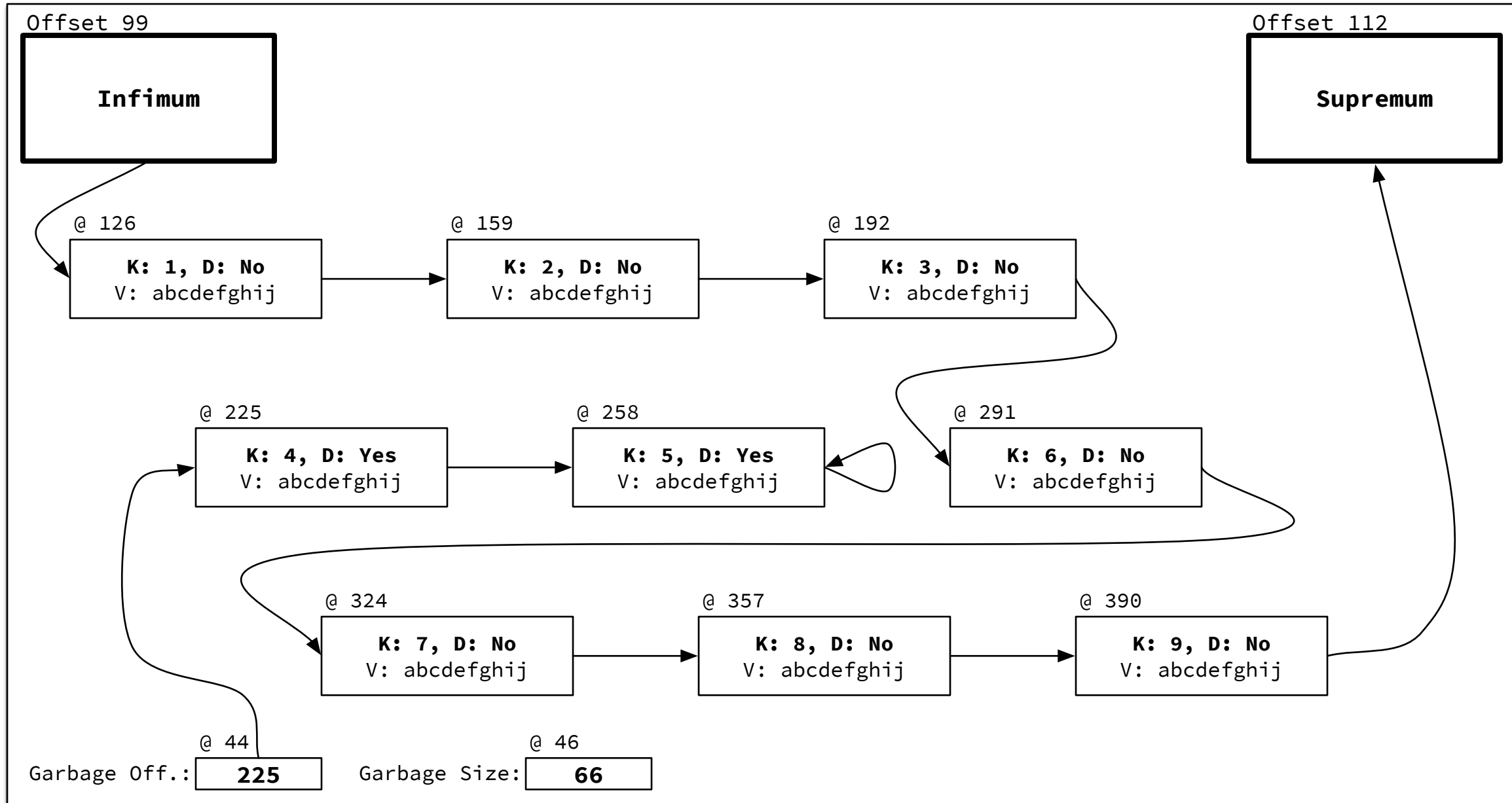
SQL: delete from t where i = 5;

Row is marked as deleted.

Garbage size is incremented by total row size.

Garbage offset is pointed to row, and row next pointer is pointed back to self.

# B+Tree Record Delete 2



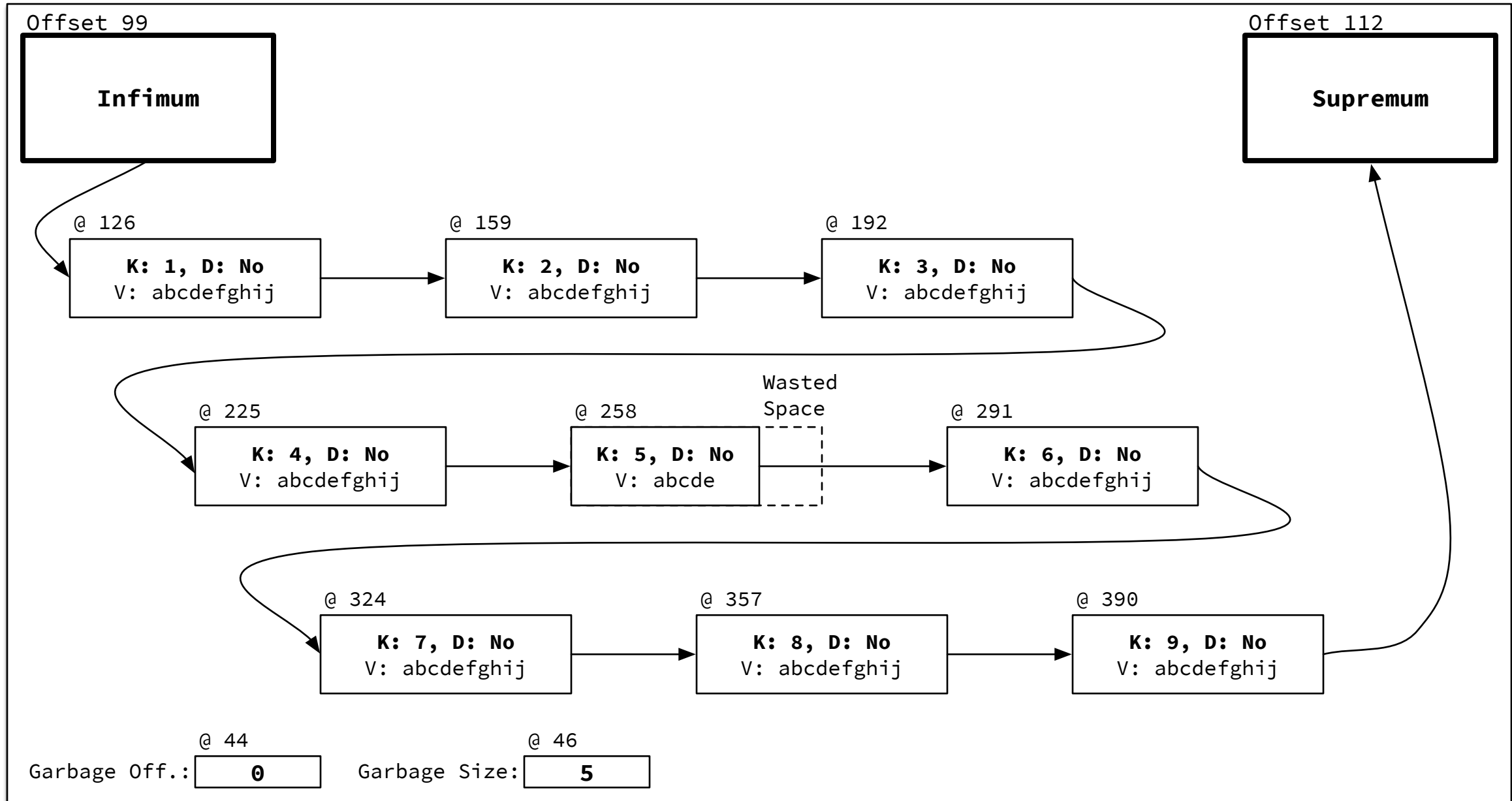
SQL: delete from t where i = 5; delete from t where i = 4;

Garbage size is incremented by total row size for each delete.

Garbage offset is pointed to row @ 258 initially, and row next pointer is pointed back to self.

Garbage offset is updated to row @ 225, and row next pointer is pointed to previous garbage offset (garbage is added to head of list).

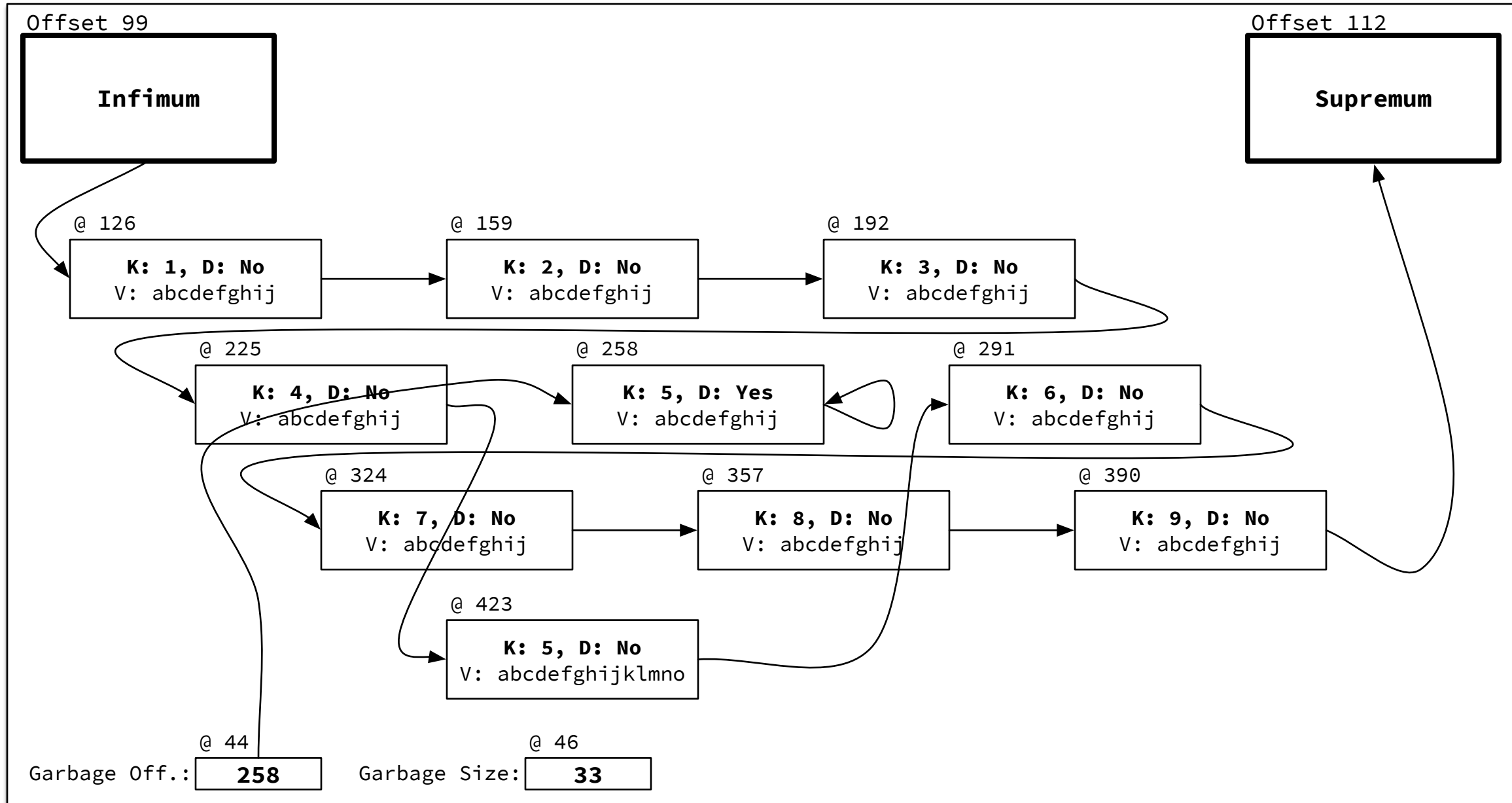
# B+Tree Record Update - Smaller



SQL: update t set s="abcde" where i = 5;

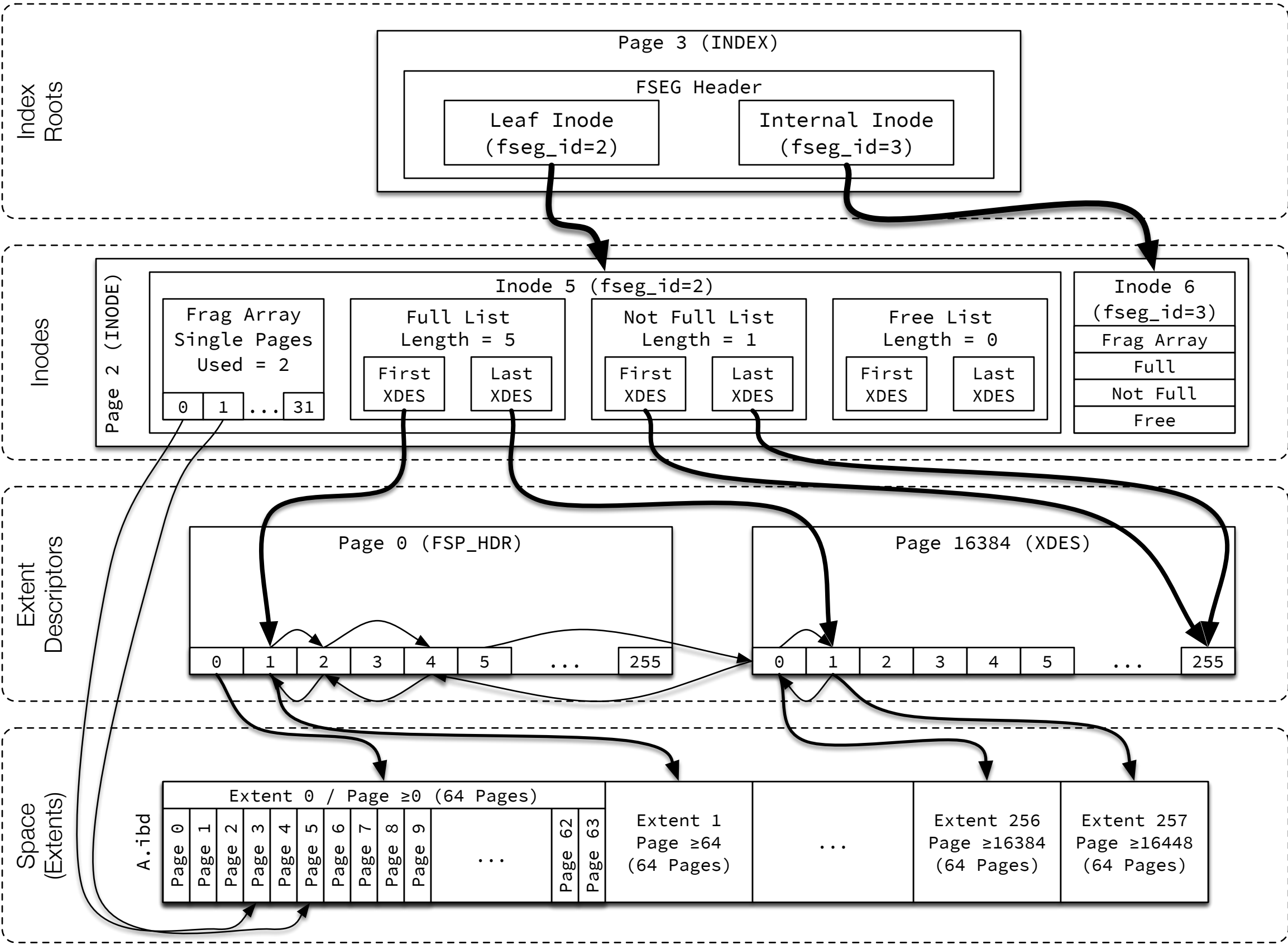
Garbage size is incremented by size of row shrinkage, but wasted space is not tracked in garbage list.

# B+Tree Record Update - Larger



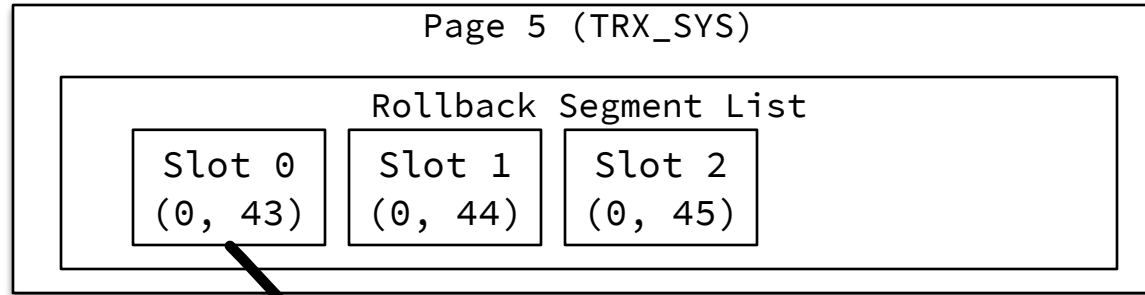
SQL: update t set s="abcdefghijklmno" where i = 5;  
 Row is deleted, and a new row is inserted into the heap.

# Index File Segment Structure

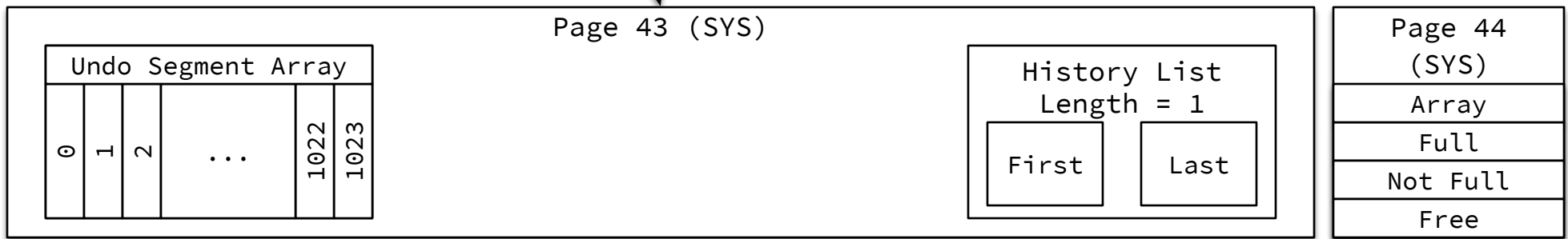


# History Structure

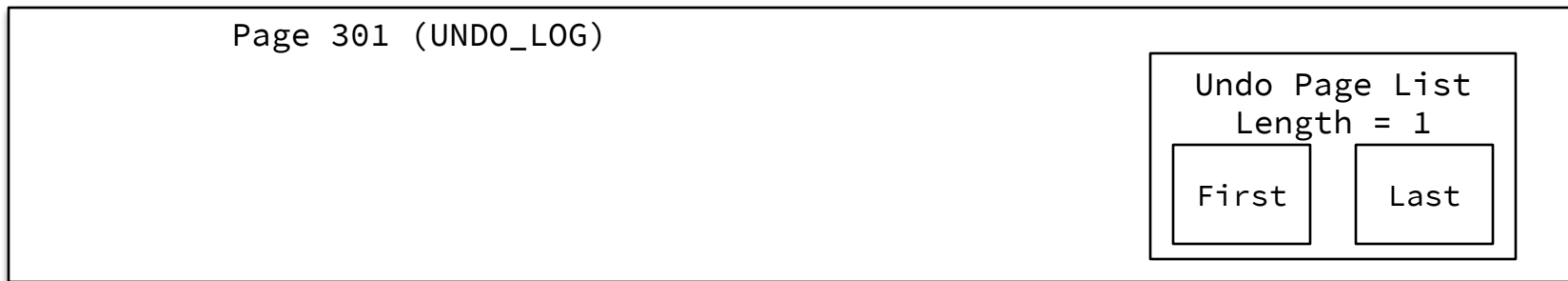
Transaction System



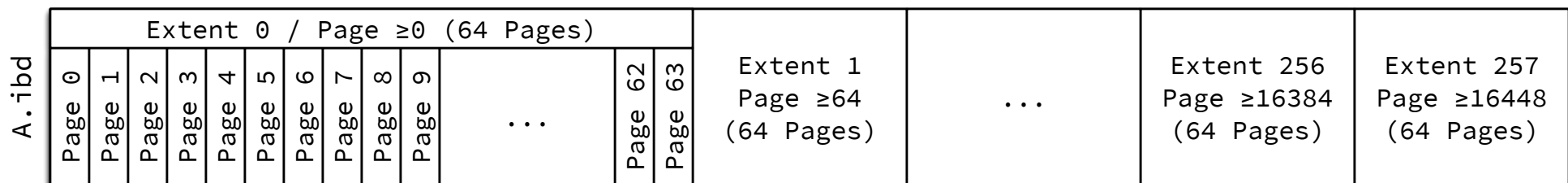
Rollback Segments



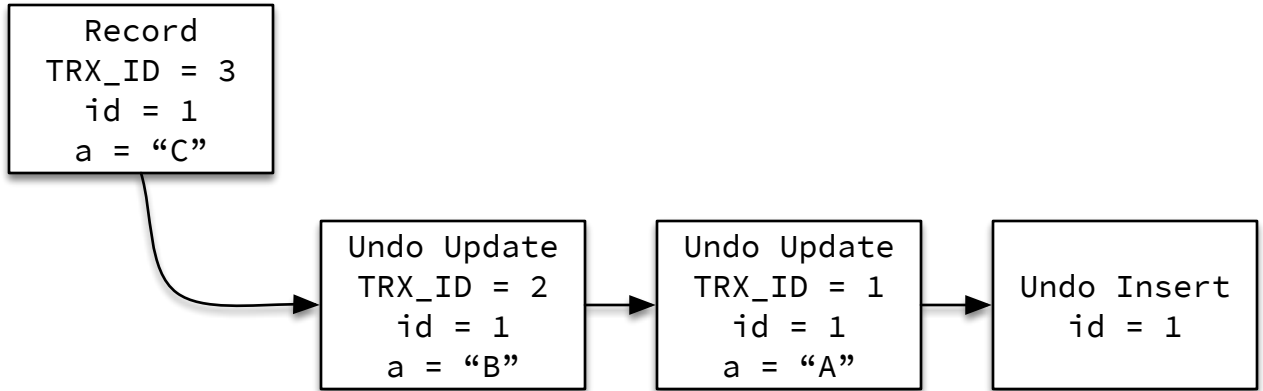
Undo Segments



Undo Logs



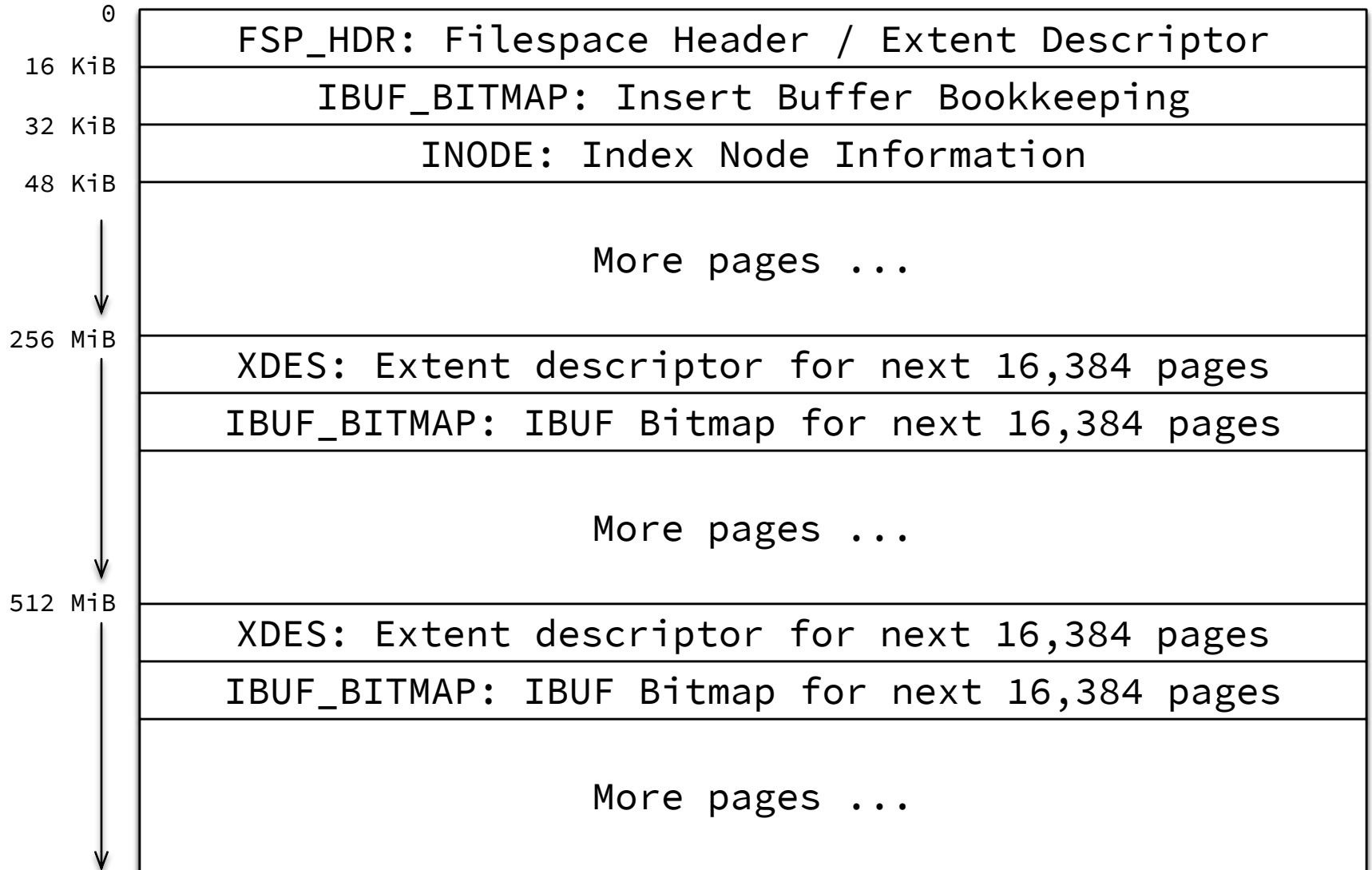
# Record History



History above represents the following SQL statements:

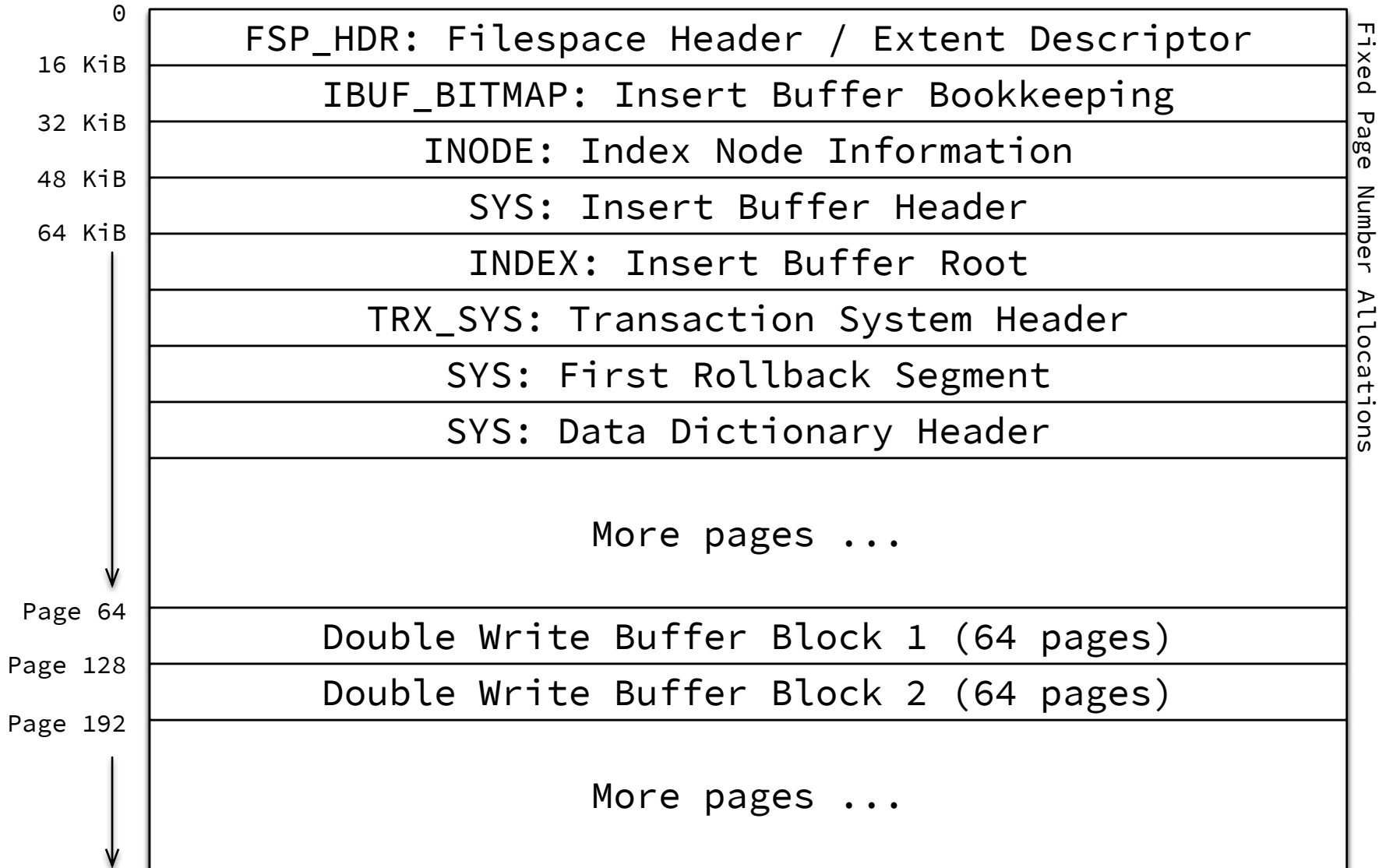
```
INSERT INTO t (id, a) VALUES (1, "A");  
UPDATE t SET a="B" WHERE id = 1;  
UPDATE t SET a="C" WHERE id = 1;
```

# Space File Overview

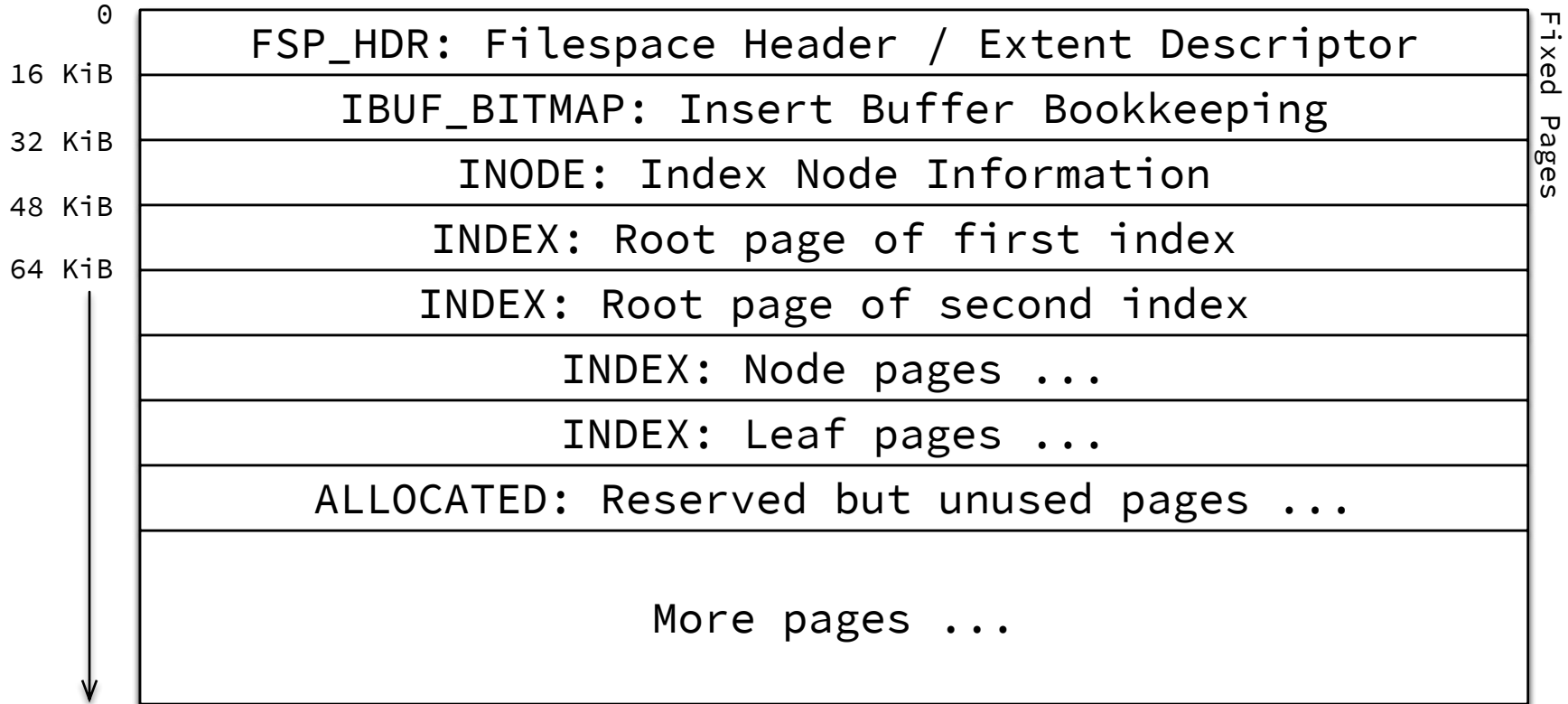




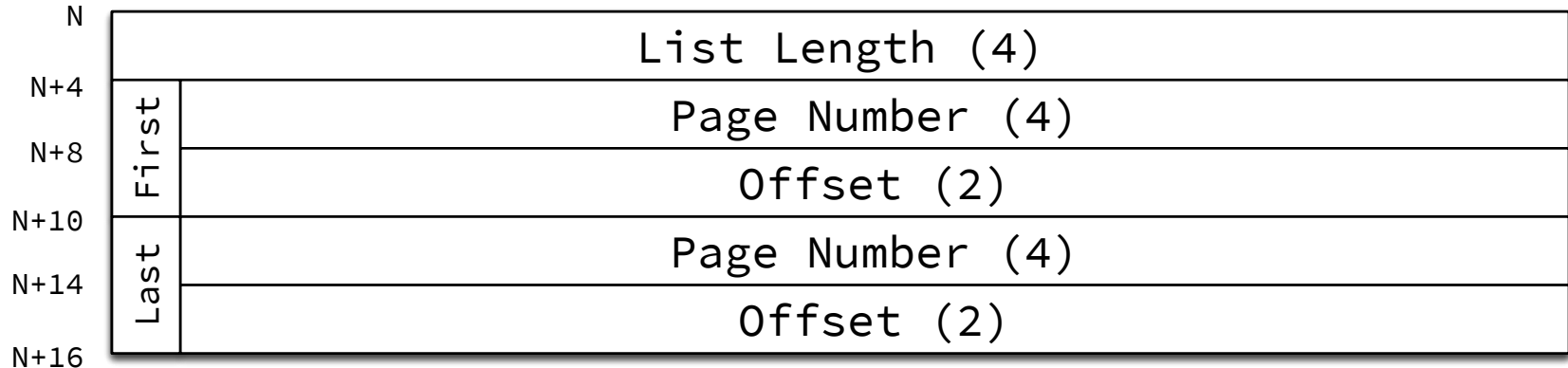
# ibdata1 File Overview



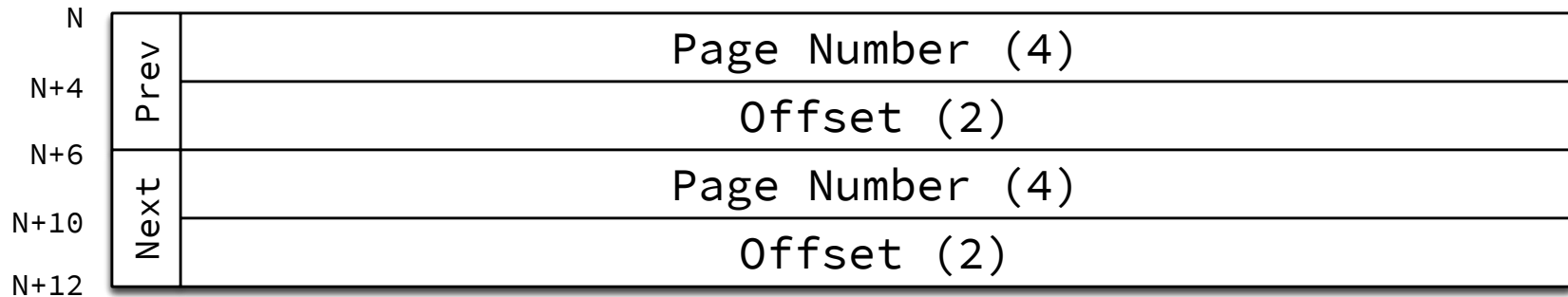
# IBD File Overview



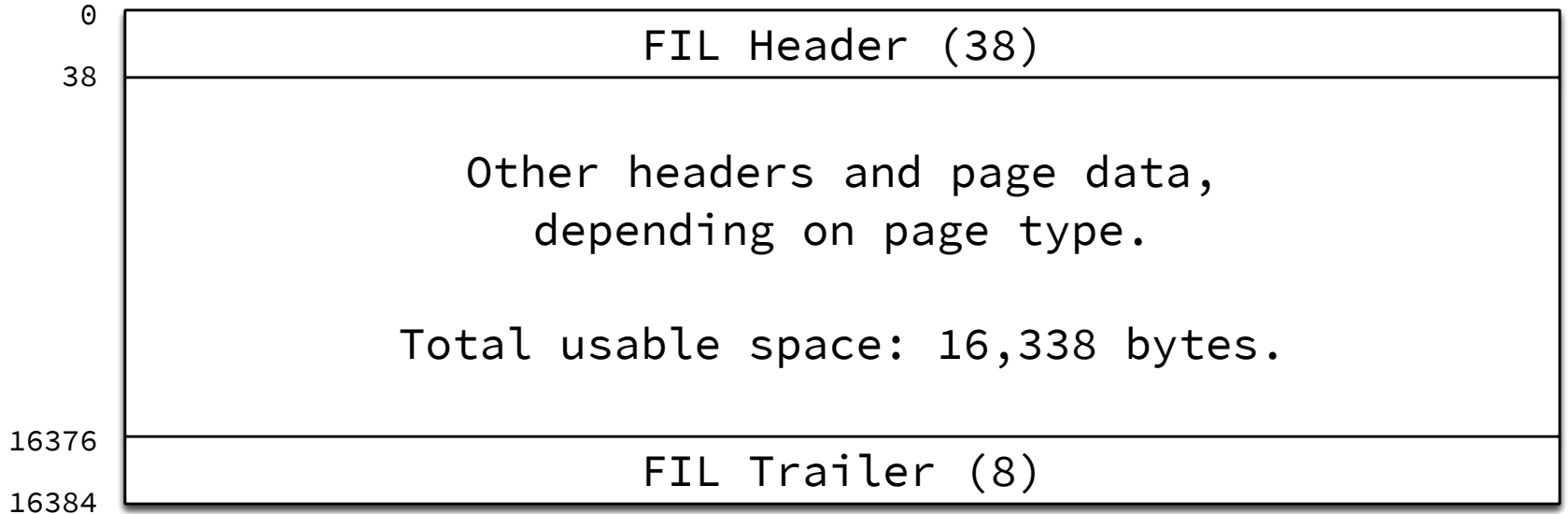
# List Base Node



# List Node



# Basic Page Overview



# FIL Header/Trailer

0	Checksum (4)
4	Offset (Page Number) (4)
8	Previous Page (4)
12	Next Page (4)
16	LSN for last page modification (8)
24	Page Type (2)
26	Flush LSN (0 except space 0 page 0) (8)
34	Space ID (4)
38	...
16376	Old-style Checksum (4)
16380	Low 32 bits of LSN (4)
16384	

# FSP\_HDR/XDES Overview

0	FIL Header (38)			
38	FSP Header (zero-filled for XDES pages) (112)			
150	XDES Entry	0 (pages	0- 63)	(40)
190	XDES Entry	1 (pages	64- 127)	(40)
230	XDES Entry	2 (pages	128- 191)	(40)
270	XDES Entry	3 (pages	192- 255)	(40)
310	...			
10310	XDES Entry	254 (pages	16256-16319)	(40)
10350	XDES Entry	255 (pages	16320-16383)	(40)
10390	(Empty Space: 5,986 bytes)			
16376	FIL Trailer (8)			
16384				

# FSP Header

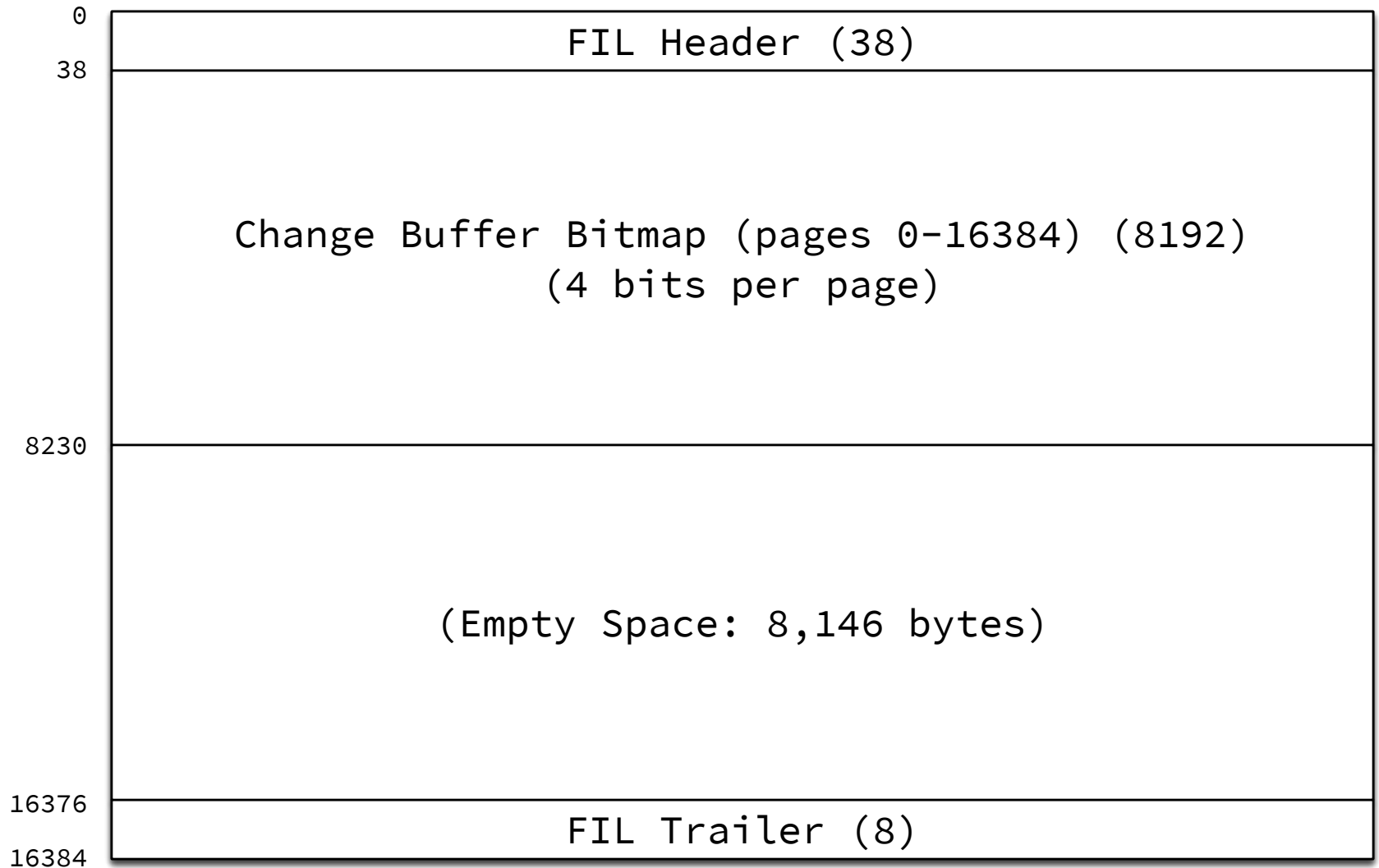
38	Space ID (4)
42	(Unused) (4)
46	Highest page number in file (size) (4)
50	Highest page number initialized (free limit) (4)
54	Flags (4)
58	Number of pages used in "FREE_FRAG" list (4)
62	List base node for "FREE" list (16)
78	List base node for "FREE_FRAG" list (16)
94	List base node for "FULL_FRAG" list (16)
110	Next Unused Segment ID (8)
118	List base node for "FULL_INODES" list (16)
134	List base node for "FREE_INODES" list (16)
150	



# XDES Entry

N	File Segment ID (8)
N+8	List node for XDES list (12)
N+20	State (4)
N+24	Page State Bitmap (16) 2 bits per page, 1=free, 2=clean
N+40	

# IBUF\_BITMAP Overview



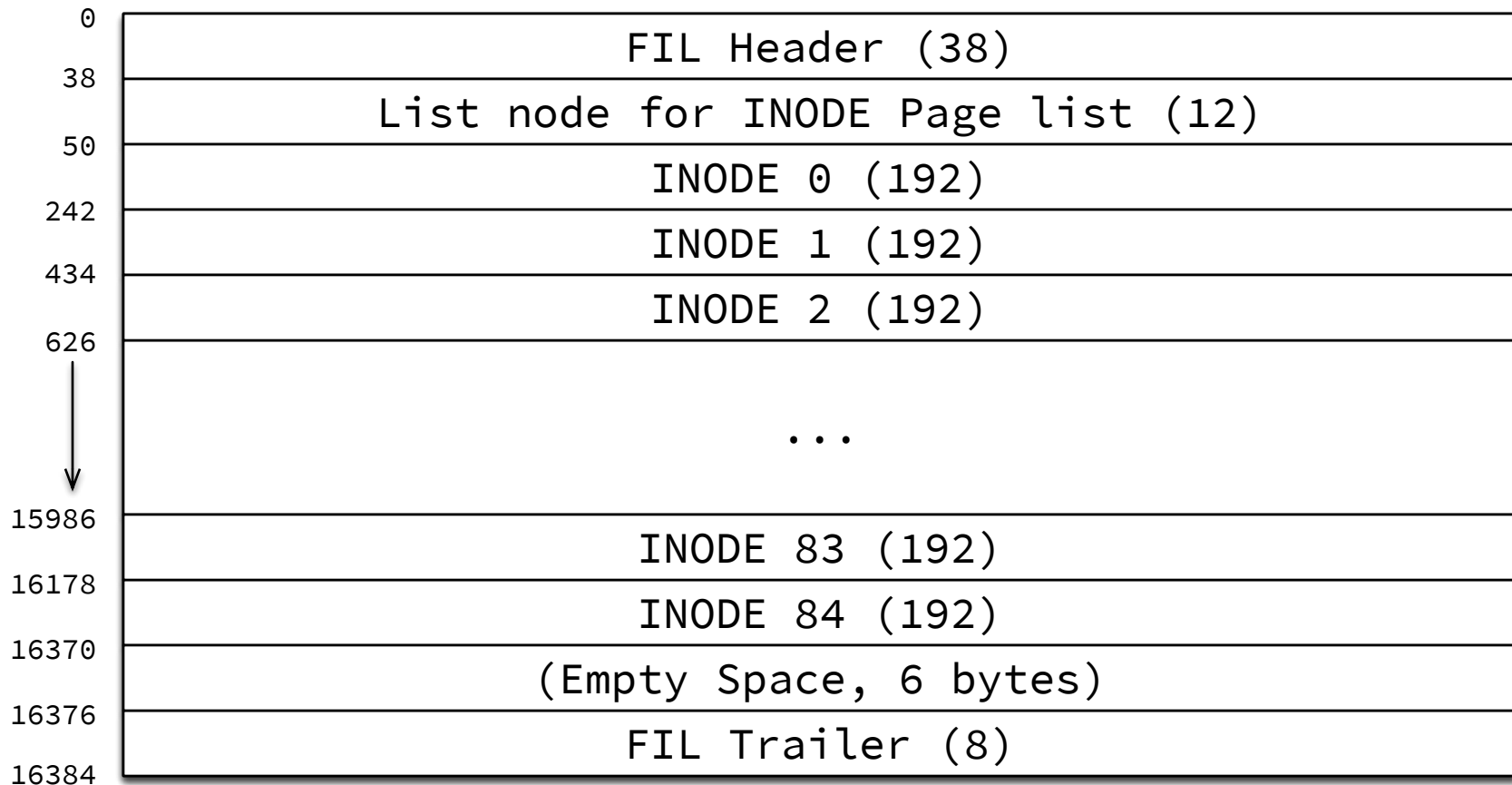
# IBUF\_BITMAP Page Entry

Free Space (2 bits)

Buffered Flag (1 bit)

Change Buffer Flag (1 bit)

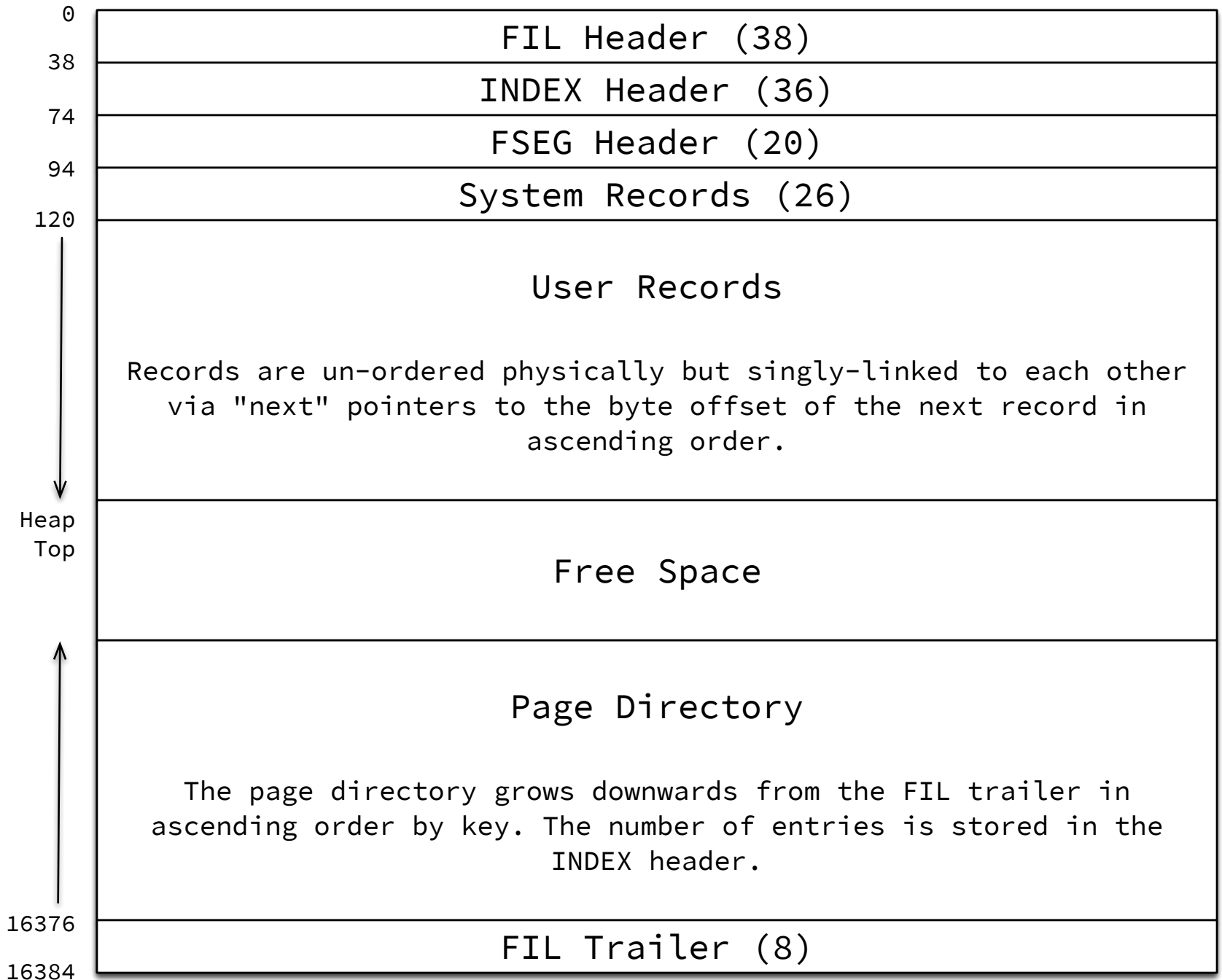
# INODE Overview



# INODE Entry

N	FSEG ID (8)
N+8	Number of used pages in "NOT_FULL" list (4)
N+12	List base node for "FREE" list (16)
N+28	List base node for "NOT_FULL" list (16)
N+44	List base node for "FULL" list (16)
N+60	Magic Number = 97937874 (4)
N+64	Fragment Array Entry 0 (4)
N+68	...
N+188	Fragment Array Entry 31 (4)
N+192	

# INDEX Overview



# INDEX Header

38	Number of Directory Slots (2)
40	Heap Top Position (2)
42	Number of Heap Records / Format Flag (2)
44	First Garbage Record Offset (2)
46	Garbage Space (2)
48	Last Insert Position (2)
50	Page Direction (2)
52	Number of Inserts in Page Direction (2)
54	Number of Records (2)
56	Maximum Transaction ID (8)
64	Page Level (2)
66	Index ID (8)
74	

# FSEG Header

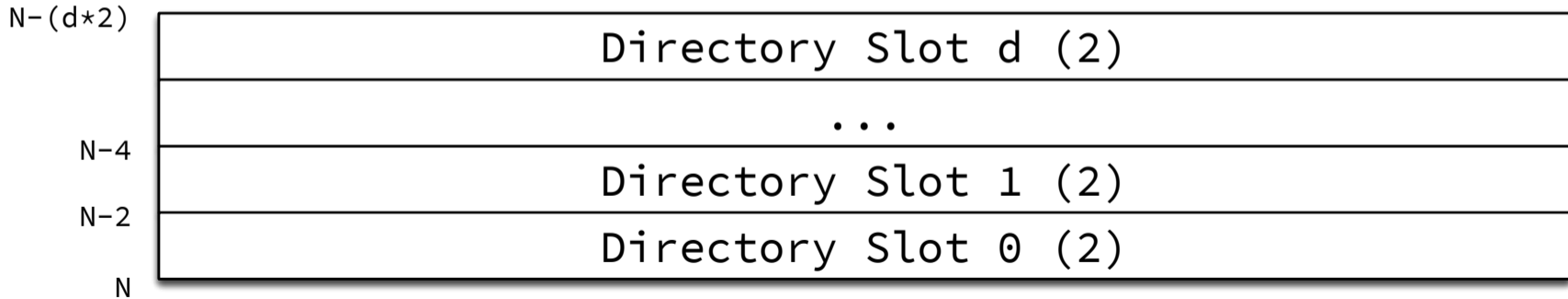
74	Leaf Pages Inode Space ID (4)
78	Leaf Pages Inode Page Number (4)
82	Leaf Pages Inode Offset (2)
84	Internal (non-leaf) Inode Space ID (4)
88	Internal (non-leaf) Inode Page Number (4)
92	Internal (non-leaf) Inode Offset (2)
94	



# INDEX System Records

94	Info Flags (4 bits)
	Number of Records Owned (4 bits)
95	Order (13 bits)
	Record Type (3 bits)
97	Next Record Offset (2)
99	"infimum\0" (8)
107	Info Flags (4 bits)
	Number of Records Owned (4 bits)
108	Order (13 bits)
	Record Type (3 bits)
110	Next Record Offset (2)
112	"supremum" (8)
120	

# INDEX Page Directory



# TRX\_SYS Overview

0	FIL Header (38)
38	Transaction ID (8)
46	TRX_SYS FSEG Entry (10)
56	Rollback Segment 0: Space (4)
60	Rollback Segment 0: Page (4)
64	...
	Rollback Segment 127: Space (4)
	Rollback Segment 127: Page (4)
1080	(Empty Space: 13304 bytes)
14384	Master Log Info (112)
14496	(Empty Space: 888 bytes)
15384	Binary Log Info (112)
15396	(Empty Space: 980 bytes)
16184	Doublewrite Buffer Info (38)
16222	(Empty Space: 154 bytes)
16376	FIL Trailer (8)
16384	

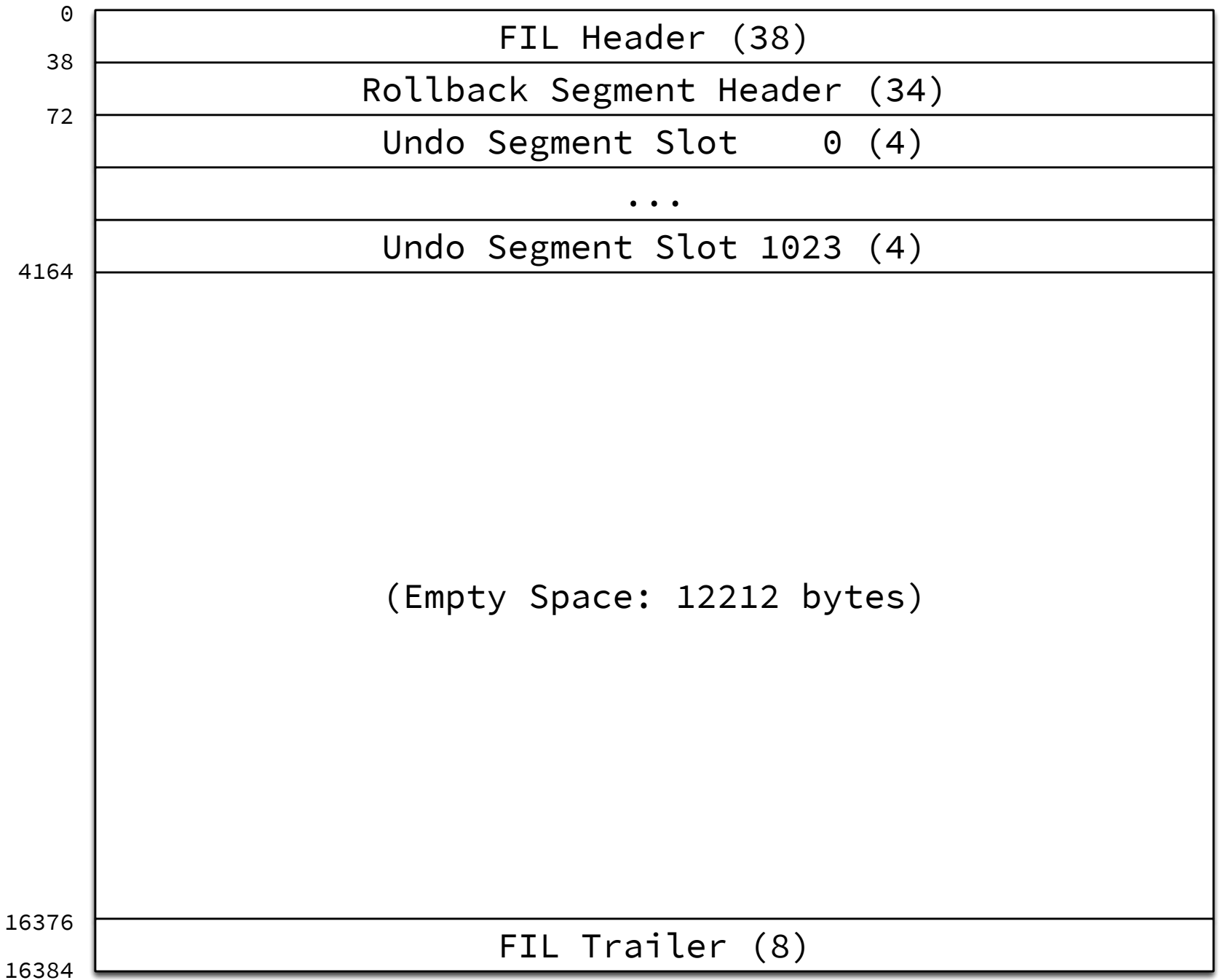
# TRX\_SYS MySQL Log Info

N	Magic Number (4)
N+4	Log Offset (8)
N+12	Log Name (100)
N+112	

# TRX\_SYS Doublewrite Buffer Info

16184	Doublewrite Buffer FSEG Entry (10)
16194	Doublewrite Magic Number (4)
16198	Block 1 Start Page (4)
16202	Block 2 Start Page (4)
16206	Doublewrite Magic Number (4)
16210	Block 1 Start Page (4)
16214	Block 2 Start Page (4)
16218	Space ID Stored Magic Number (4)
16222	

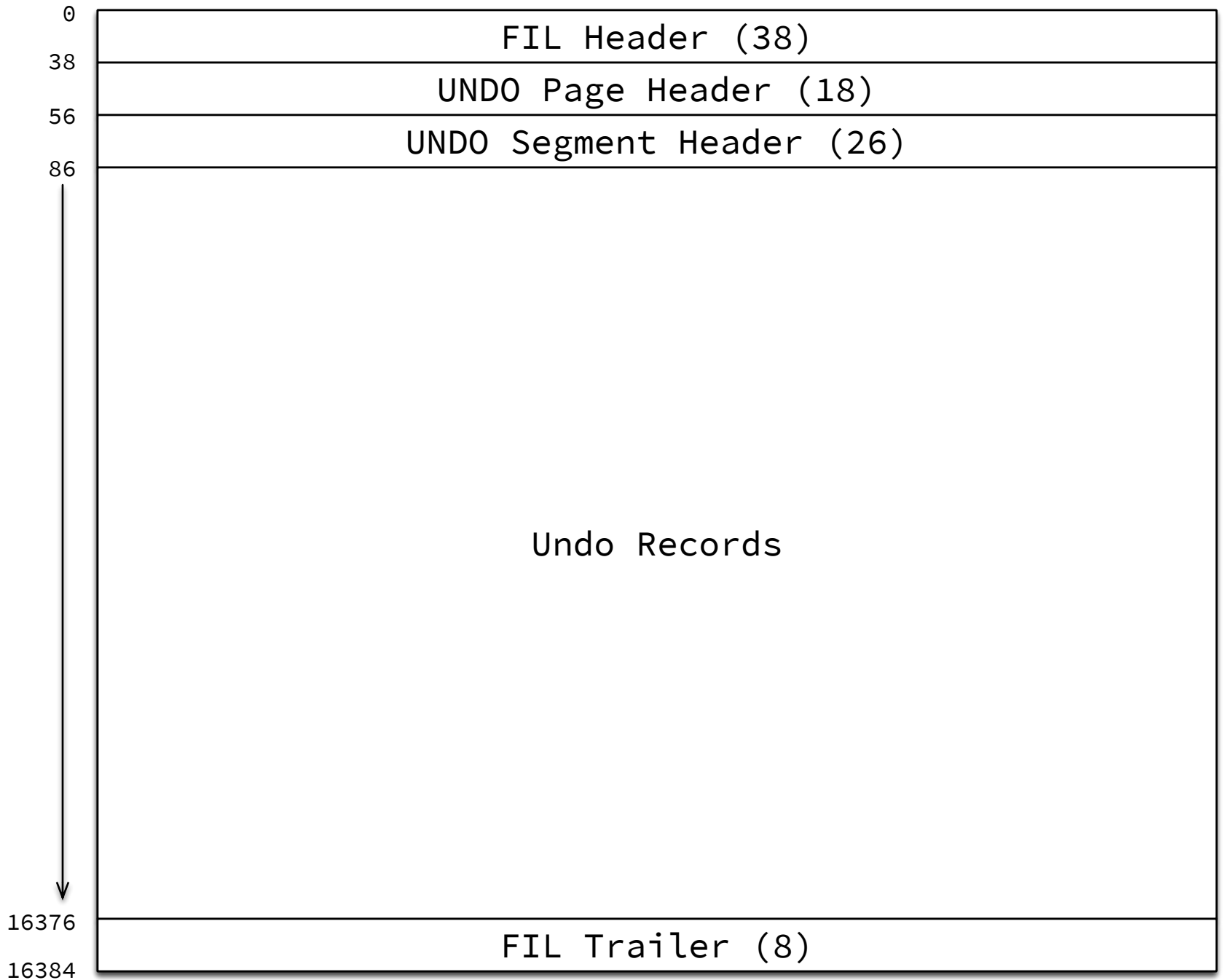
# SYS\_RSEG\_HEADER Overview



# Rollback Segment Header

38	Max Size (4)
42	History Size (4)
46	History List Base Node (16)
62	Rollback Segment FSEG Entry (10)
72	

# UNDO\_LOG Overview





# Undo Page Header

38	Undo Page Type (2)
40	Latest Log Record Offset (2)
42	Free Space Offset (2)
44	Undo Page List Node (12)
56	

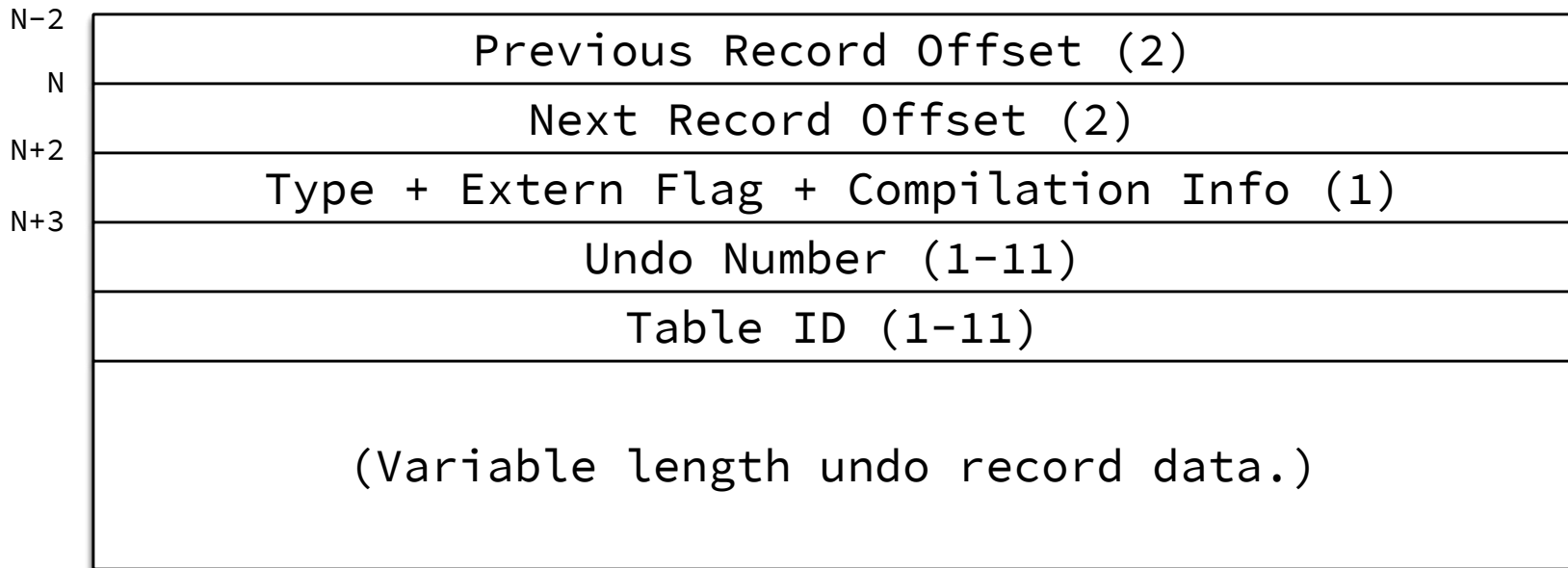
# Undo Segment Header

56	State (2)
58	Last Log Offset (2)
60	Undo Segment FSEG Entry (10)
70	Undo Segment Page List Base Node (16)
86	

# Undo Log Header

0		Transaction ID (8)
8		Transaction Number (8)
16		Delete Marks Flag (2)
18		Log Start Offset (2)
20		XID Flag (1)
21		DDL Transaction Flag (1)
22		Table ID if DDL Transaction (8)
30		Next Undo Log Offset (2)
32		Prev Undo Log Offset (2)
34		History List Node (12)
46	If XID Flag	XID Format (4)
50		TRID Length (4)
54		BQUAL Length (4)
58		XID Data (128)
186		

# Undo Record



# Undo Record for Update

N-2	Previous Record Offset (2)	
N	Next Record Offset (2)	
N+2	Type + Extern Flag + Compilation Info (1)	
N+3	Undo Number (1-11)	
	Table ID (1-11)	
Key Fields	Key Field 1 Length (1-5)	
	Key Field 1 Content (n)	
	...	
	Key Field N Length (1-5)	
	Key Field N Content (n)	
	Updated Field Count	
Updated Fields	Updated Field 1 Field Number (1-5)	
	Updated Field 1 Length (1-5)	
	Updated Field 1 Content (n)	
	...	
	Updated Field N Field Number (1-5)	
	Updated Field N Length (1-5)	
	Updated Field N Content (n)	

Rec. Offset ->

(Variable Length Record Header)

(Variable Length Record Data)

# Record Format - Overview

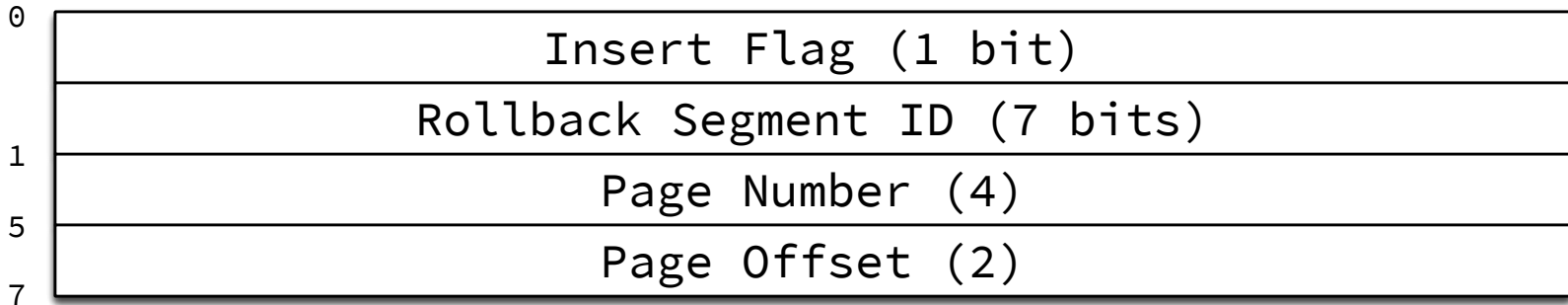
	Variable field lengths (1-2 bytes per var. field)	
	Nullable field bitmap (1 bit per nullable field)	
N-5	Info Flags (4 bits)	"Header"
	Number of Records Owned (4 bits)	
N-4	Order (13 bits)	
	Record Type (3 bits)	
N-2	Next Record Offset (2)	
N	(Record Data)	

# Record Format - Header

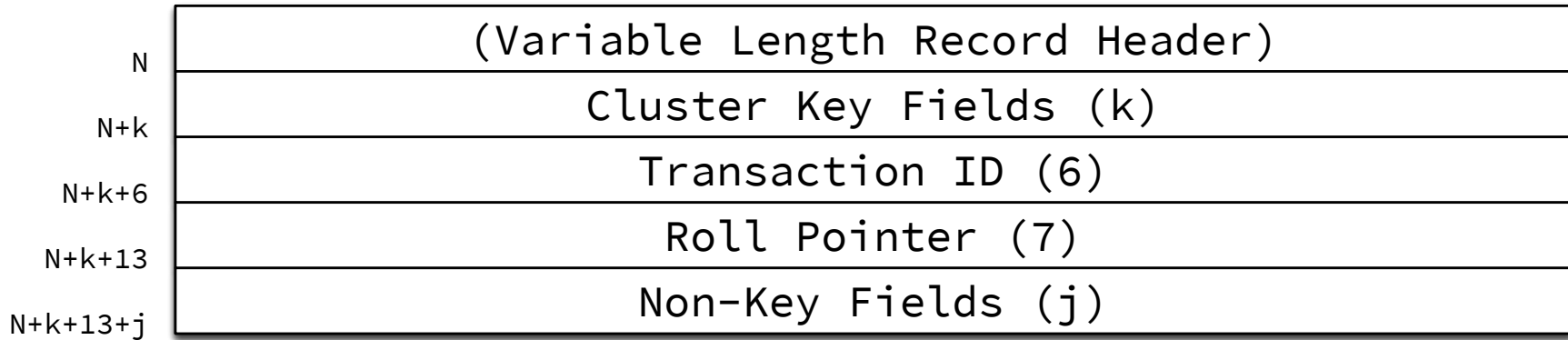
	Variable field lengths (1-2 bytes per var. field)
	Nullable field bitmap (1 bit per nullable field)
N-5	Info Flags (4 bits)
	Number of Records Owned (4 bits)
N-4	Order (13 bits)
	Record Type (3 bits)
N-2	Next Record Offset (2)
N	



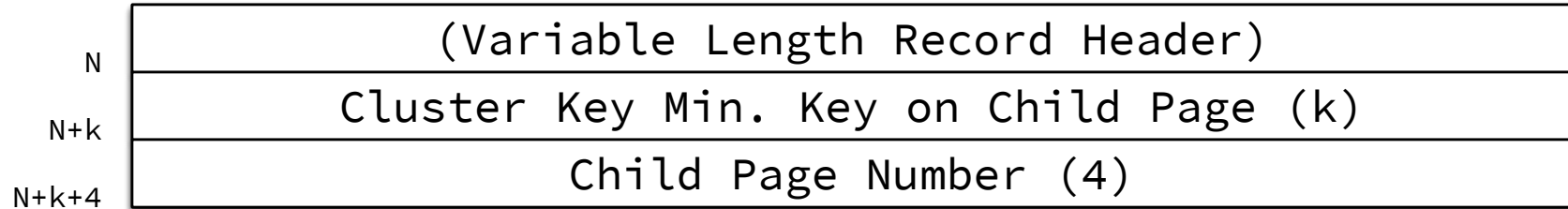
# Rollback Pointer (ROLL\_PTR)



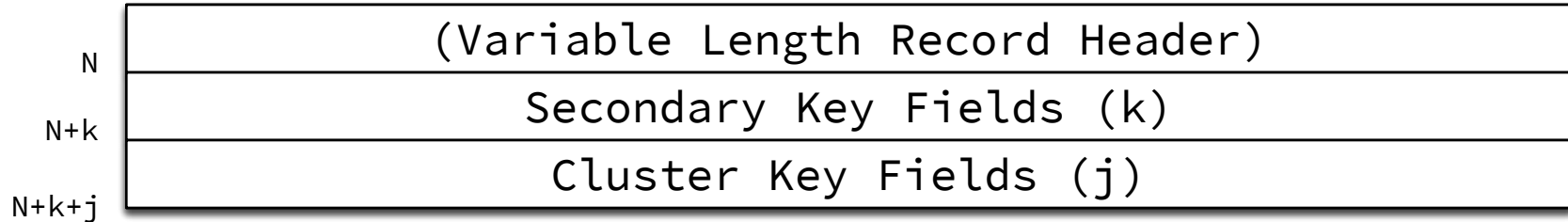
# Record Format - Clustered Key - Leaf Pages



# Record Format - Clustered Key - Non-Leaf Pages



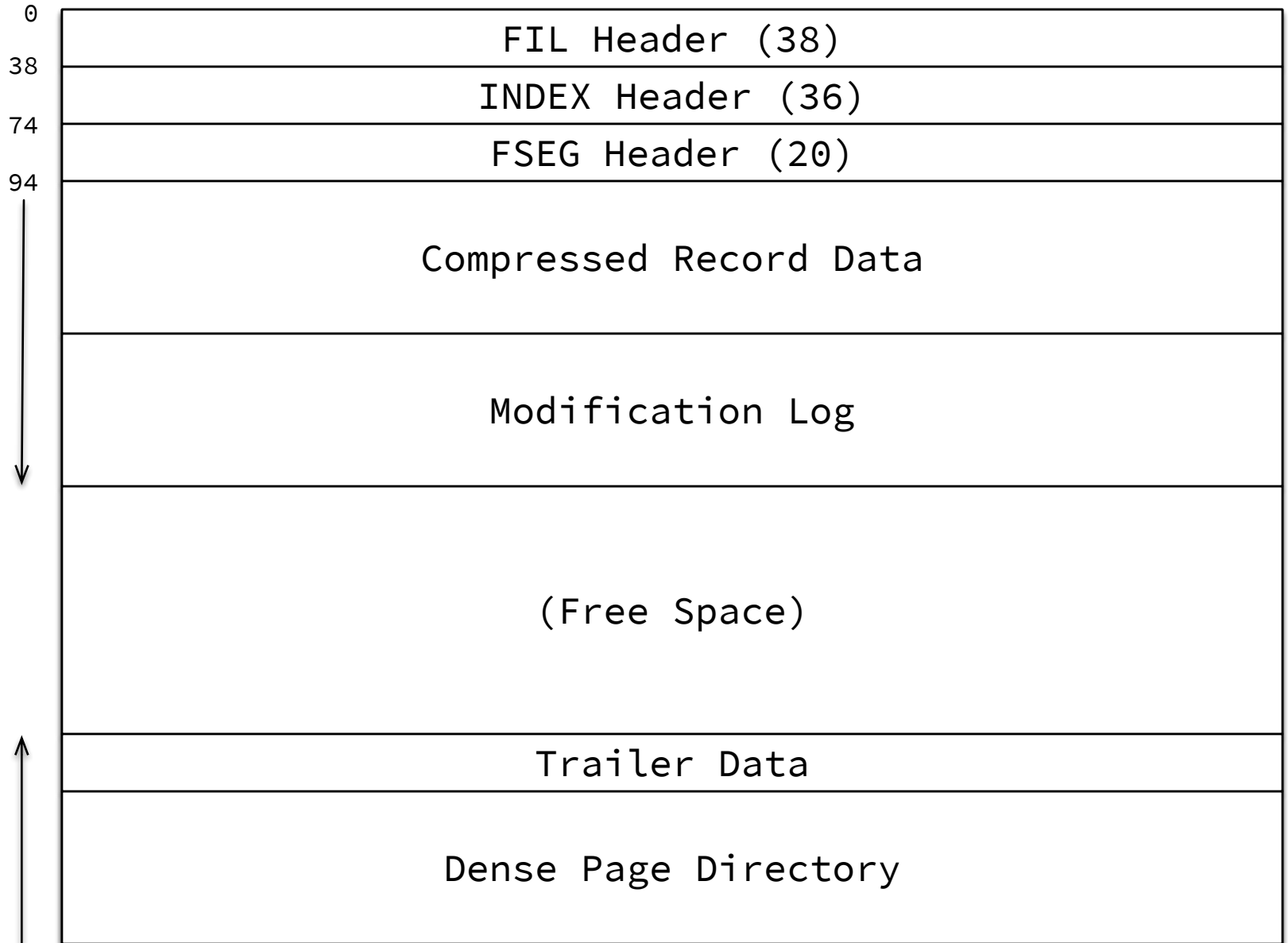
# Record Format - Secondary Key - Leaf Pages



# Record Format - Secondary Key - Non-Leaf Pages

N	(Variable Length Record Header)
N+k	Secondary Key Min. Key on Child Page (k)
N+k+j	Cluster Key Min. Key on Child Page (j)
N+k+j+4	Child Page Number (4)

# INDEX Overview



# Compressed Index Data

## Uncompressed Index Records

Record 1: "A"
Key fields
System Fields
Non-key fields

Record 2: "C"
Key fields
System Fields
Non-key fields

Record 3: "B"
Key fields
System Fields
Non-key fields

Record 4: "D"
Key fields
System Fields
Non-key fields

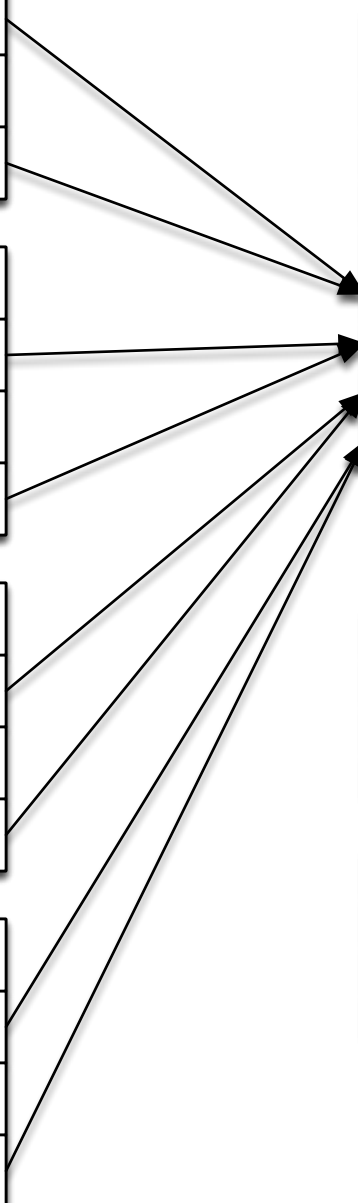
## Compressed Data

ZLIB Header
Description 1
Description 2
Description 3
Description 4
Record Data 1
Record Data 2
Record Data 3
Record Data 4

## Uncompressed Data

TRX_ID/ROLL_PTR 4
TRX_ID/ROLL_PTR 3
TRX_ID/ROLL_PTR 2
TRX_ID/ROLL_PTR 1

Slot 3
Slot 2
Slot 1
Slot 0



# Compressed Index Data

## Uncompressed Index Records

Record 1: "A"
Key fields
System Fields
Non-key fields

Record 2: "C"
Key fields
System Fields
Non-key fields

Record 3: "B"
Key fields
System Fields
Non-key fields

Record 4: "D"
Key fields
System Fields
Non-key fields

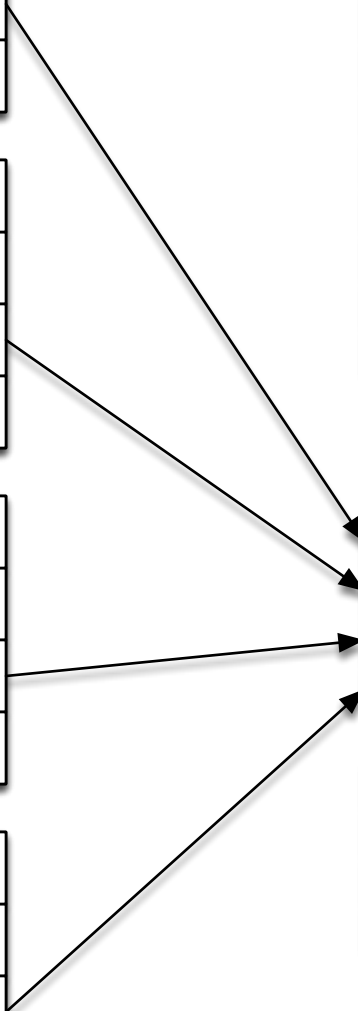
## Compressed Data

ZLIB Header
Description 1
Description 2
Description 3
Description 4
Record Data 1
Record Data 2
Record Data 3
Record Data 4

## Uncompressed Data

TRX_ID/ROLL_PTR 4
TRX_ID/ROLL_PTR 3
TRX_ID/ROLL_PTR 2
TRX_ID/ROLL_PTR 1

Slot 3
Slot 2
Slot 1
Slot 0





# Compressed Index Data

## Uncompressed Index Records

Record 1: "A"
Key fields
System Fields
Non-key fields

Record 2: "C"
Key fields
System Fields
Non-key fields

Record 3: "B"
Key fields
System Fields
Non-key fields

Record 4: "D"
Key fields
System Fields
Non-key fields

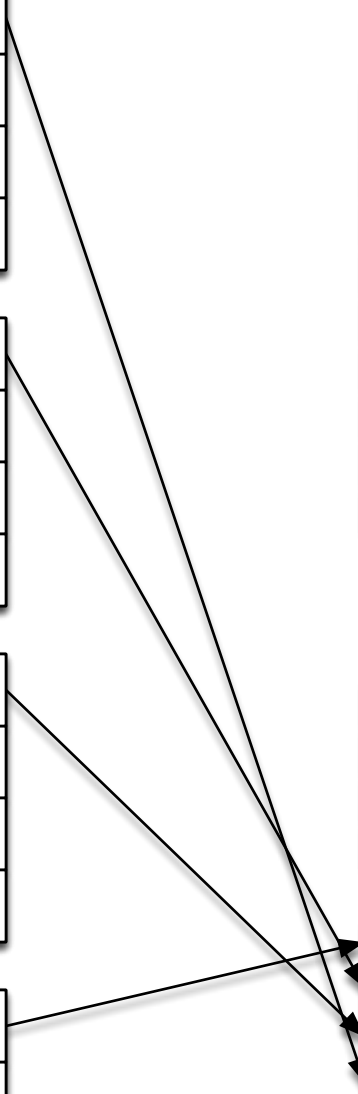
## Compressed Data

ZLIB Header
Description 1
Description 2
Description 3
Description 4
Record Data 1
Record Data 2
Record Data 3
Record Data 4

## Uncompressed Data

TRX_ID/ROLL_PTR 4
TRX_ID/ROLL_PTR 3
TRX_ID/ROLL_PTR 2
TRX_ID/ROLL_PTR 1

Slot 3
Slot 2
Slot 1
Slot 0



# Modification Log

Entry 1	Heap Number (1-2)
	Record Data
Entry 2	Heap Number (1-2)
	Record Data
...	
Entry N	Heap Number (1-2)
	Record Data
End Marker (1) = 0	

# Trailer Data

Entry N	TRX_ID (6)
	Roll Pointer (5)
...	
Entry 2	TRX_ID (6)
	Roll Pointer (5)
Entry 1	TRX_ID (6)
	Roll Pointer (5)

# Dense Page Directory

Slot N	Deleted Flag (1 bit)
	Owned Flag (1 bit)
	Record Offset (14 bits)
...	
Slot 1	Deleted Flag (1 bit)
	Owned Flag (1 bit)
	Record Offset (14 bits)
Slot 0	Deleted Flag (1 bit)
	Owned Flag (1 bit)
	Record Offset (14 bits)

The dense page directory contains one entry per records, in the key's collation order. All directory slots will own a minimum of 4 and maximum of 8 records. The page directory grows "downwards" from the end of the page.

# Record Format - Change Buffer - Leaf Pages

Space ID (4)	
Field Marker (1)	
Page Number (4)	
Metadata	Operation Counter (2)
	Operation Type (2)
	Flags (1)
Type Info. 1	Data Type (1)
	“Precise” Data Type (1)
	Length (2)
	Collation Code (2)
...	
Type Information N	
Secondary Index Fields (j)	