

MongoDB Storage Internals

Mathias Stearn
@mathias_mongo
#mongosv

10gen 

\$ ls -lh /data/db

mongod.lock 6 bytes

test.ns 16 MB

test.0 64 MB

test.1 128 MB

test.2 256 MB

...
test.30 2048 MB

\$ ls -lh /data/db

mongodb.lock

6 bytes

test.ns 16 MB

test.0 64 MB

test.1 128 MB

test.2 256 MB

...
test.30 2048 MB

other.ns 16 MB

other.0 64 MB

other.1 128 MB

208 MB

MMAP

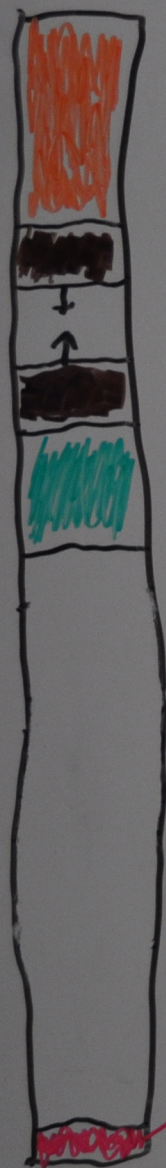
Maps a file on disk
into memory

MMAP

Maps a file on disk
into memory

Virtual Memory NOT Physical RAM

0xFFFFFFF



Kernel Space

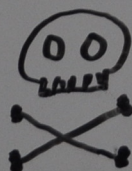
(can't touch this)

Stack

Heap

Mongod

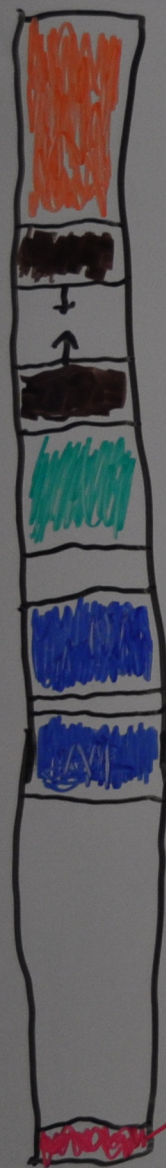
NULL



0x00000000

Virtual
Address
Space

0xFFFFFFF



Kernel Space

(can't touch this)

Stack

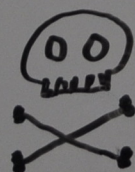
Heap

Mongod

test.ns

test.0

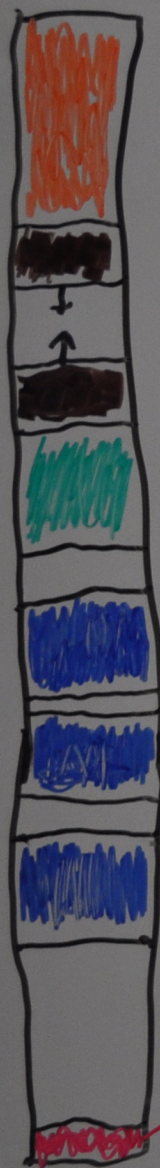
NULL



0x00000000

Virtual
Address
Space

0xFFFFFFF



Kernel Space

(can't touch this)

Stack

Heap

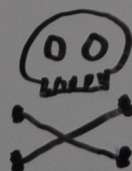
Mongod

test.ns

test.0

test.5

NULL



0x00000000

Virtual
Address
Space

32-Bit

$$2^{32} = 4 \text{ GB?}$$

Virtual
Address
Space

32-Bit

4GB

1GB

Kernel

Stack Mongod

Misc

~0.5GB

Virtual
Address
Space

32-Bit

4GB

1GB

— ~0.5GB

Kernel

Stack Mongod
Misc

~2.5GB

Room for
Data

Virtual
Address
Space

32-Bit

4 GB

1 GB

— ~0.5 GB

Kernel

Stack Mongod
Misc

~ 2.5 GB

Room for
Data

BAD

Virtual
Address
Space

64-bit

$$2^{64} = 1.7 \times 10^{10} \text{ GB?}$$

Virtual
Address
Space

~~64~~-Bit
48

0xF0 ... 0xFF **kernel**
0x00 ... 0x7F **user**

$$2^{47} = 128 \text{ TB}$$

Virtual
Address
Space

~~64~~-Bit
48

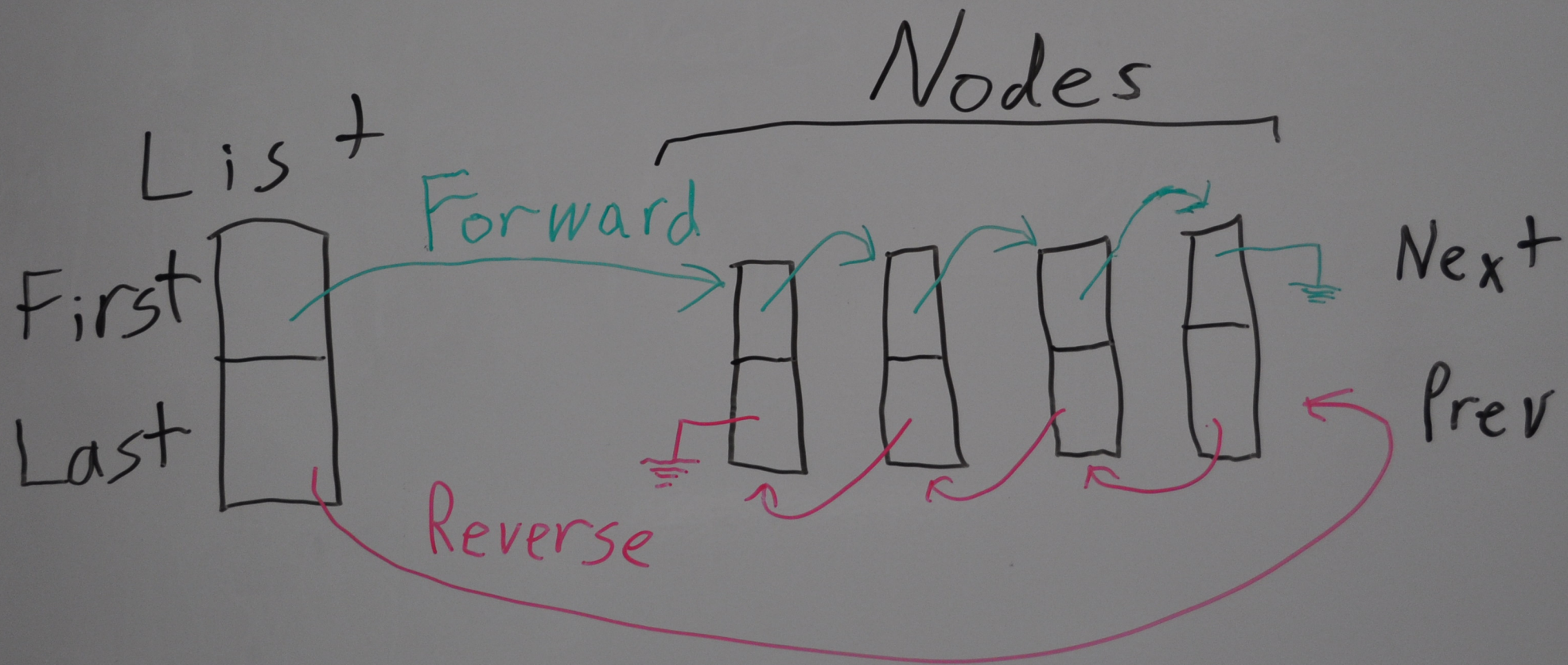
0xF0 ... 0xFF **kernel**
0x00 ... 0x7F **user**

$$2^{47} = 128 \text{ TB}$$

GOOD

Virtual
Address
Space

Doubly-Linked List



\$ ls -lh /data/db

mongod.lock

6 bytes

test.ns 16 MB

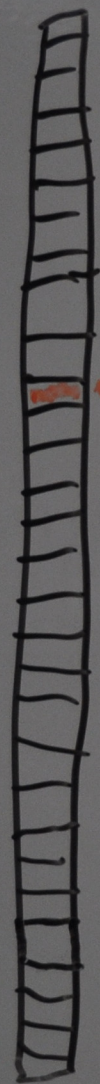
test.0 64 MB

test.1 128 MB

test.2 256 MB

test.30 2048 MB

test.ns



~2kB

Name: test.foo
Stats: size, count,
etc.

First Extent

Last Extent

Deleted List [19]

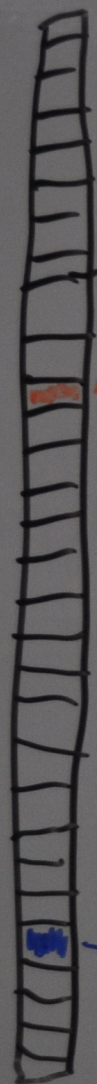
Index Data

Namespace
Details

Disk Loc 8B

Offset 4B

test.ns



~2kB

Name: test.foo
Stats: size, count,
etc.

First Extent
Last Extent
Deleted List [19]
Index Data

\$freelist
for extents

Namespace
Details

Disk Loc 8B

Offset 4B

Disk Loc:

A pointer to a
location on disk


```
struct DiskLoc {  
    int fileNum; // the '0' in test.0  
    int offset; // position in file  
}; // 64 Bits
```



```
struct DiskLoc {  
    int fileNum; // the '0' in test.0  
    int offset; // position in file  
}; // 64 Bits
```

Max offset is $2^{32}-1$

Same as max file size

\$ ls -lh /data/db

mongod.lock 6 bytes

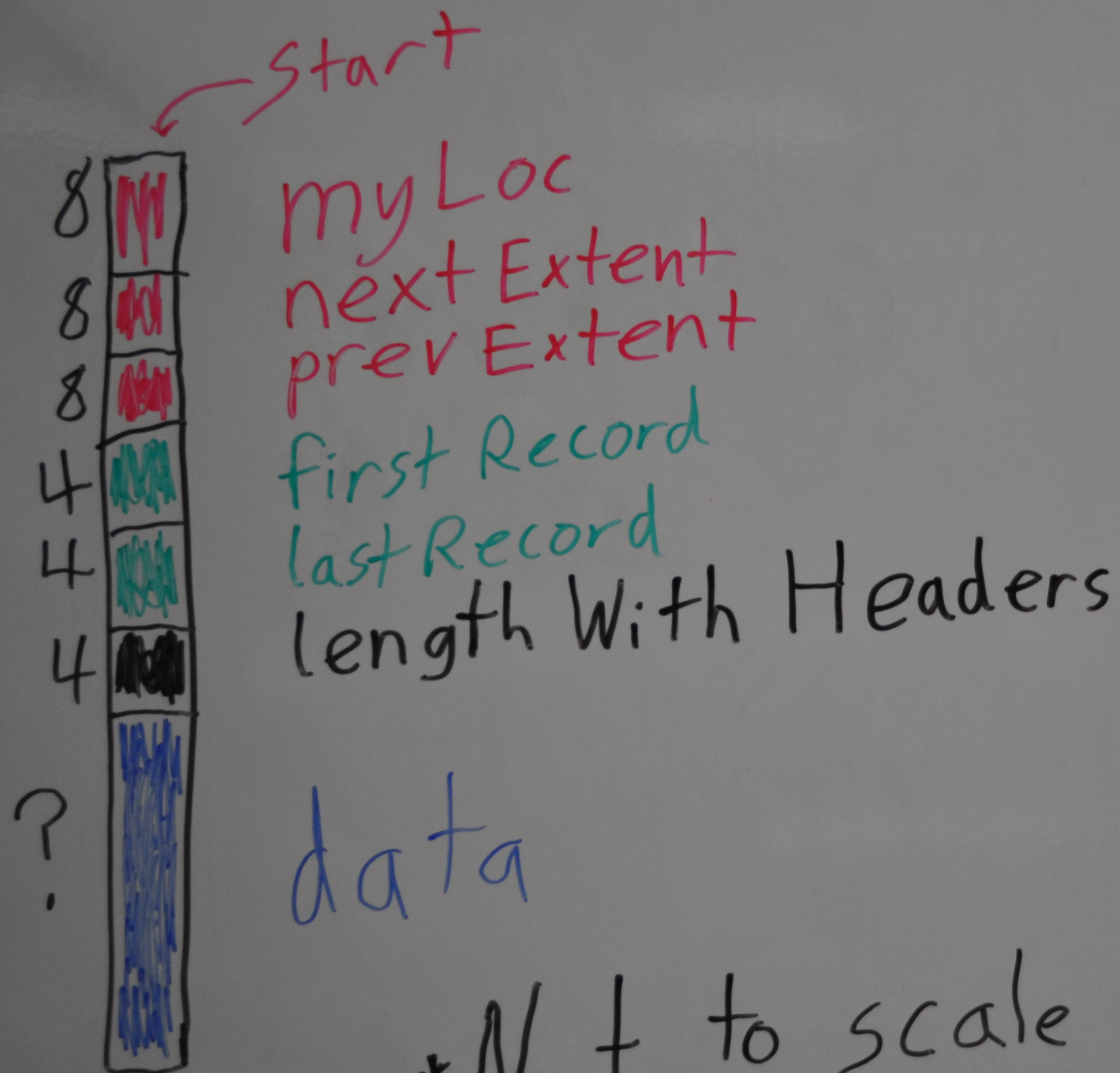
test.ns 16 MB

test.0 64 MB

test.1 128 MB

test.2 256 MB

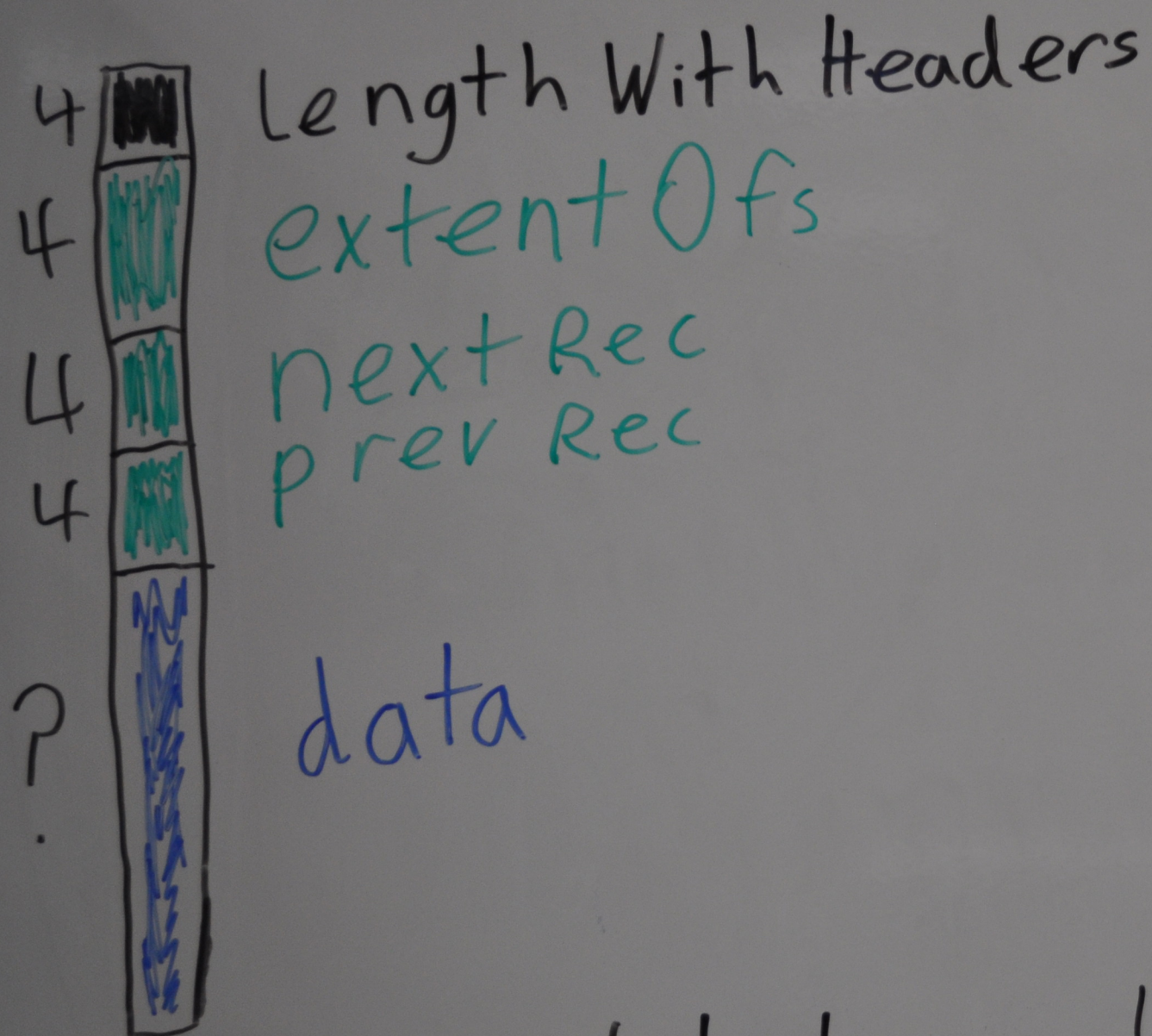
...
test.30 2048 MB



Extent

Disk Loc 8B
Offset 4B

*Not to scale

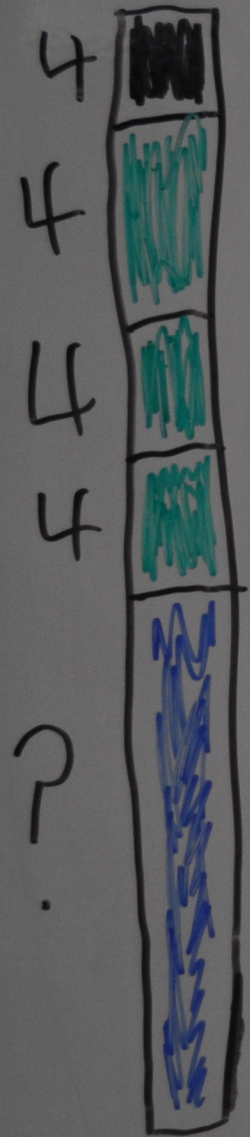


Record

*Not to scale

Disk Loc 8B

Offset 4B



Length With Headers

extent Of's — from self

next Rec
prev Rec

Record

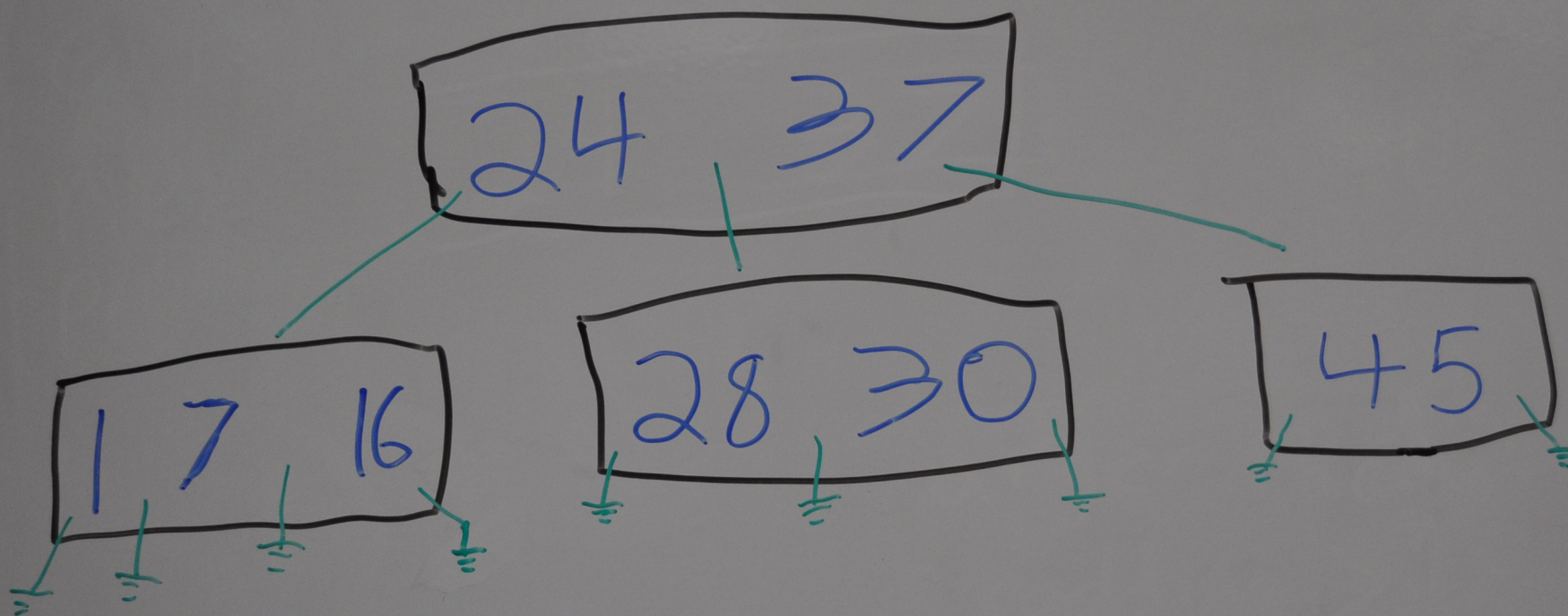
data — may be larger than
object if padding > 1

Disk Loc 8B

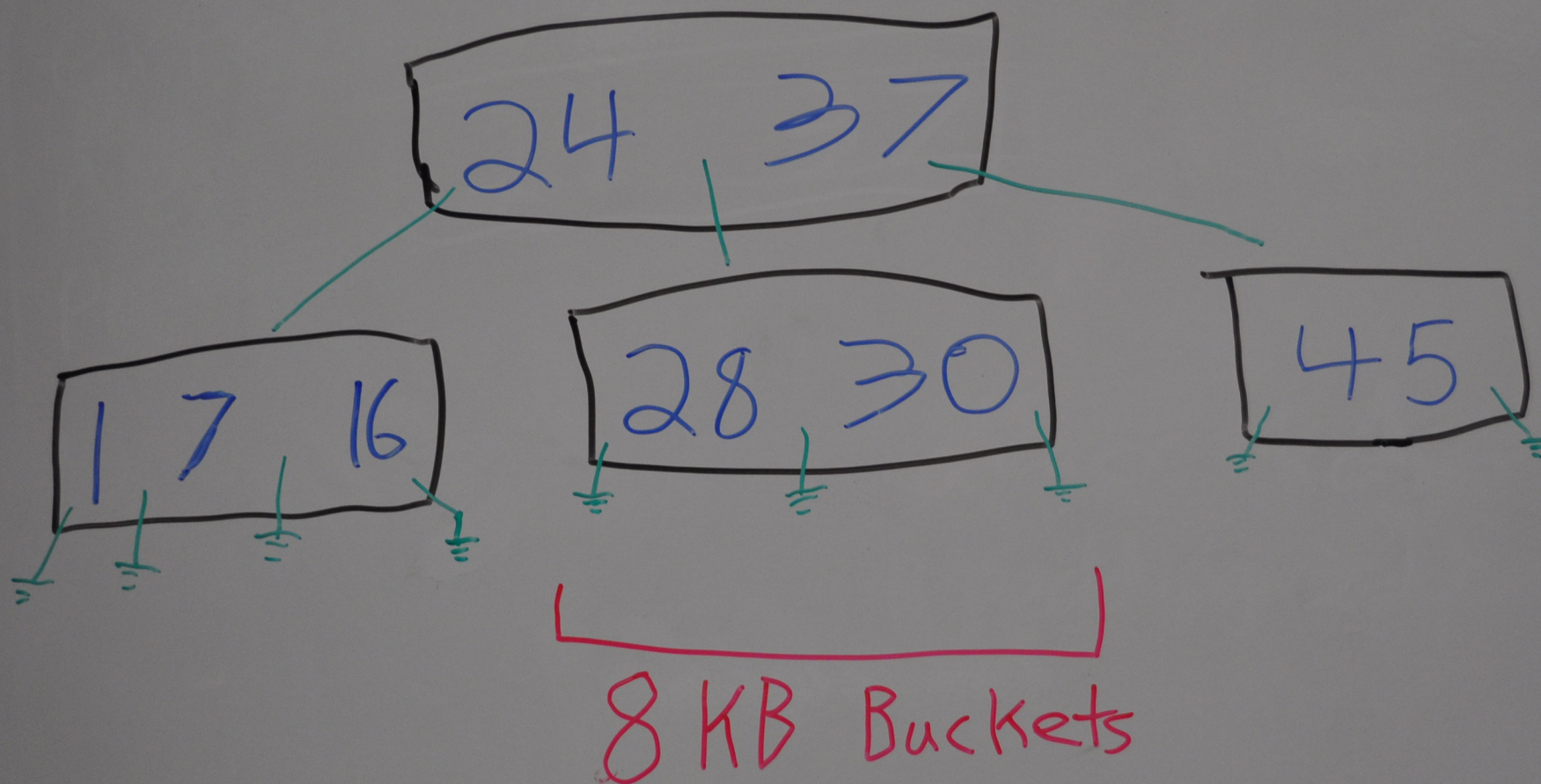
Offset 4B

*Not to scale

B-Tree



B-Tree



B-Tree

1 2 3 4

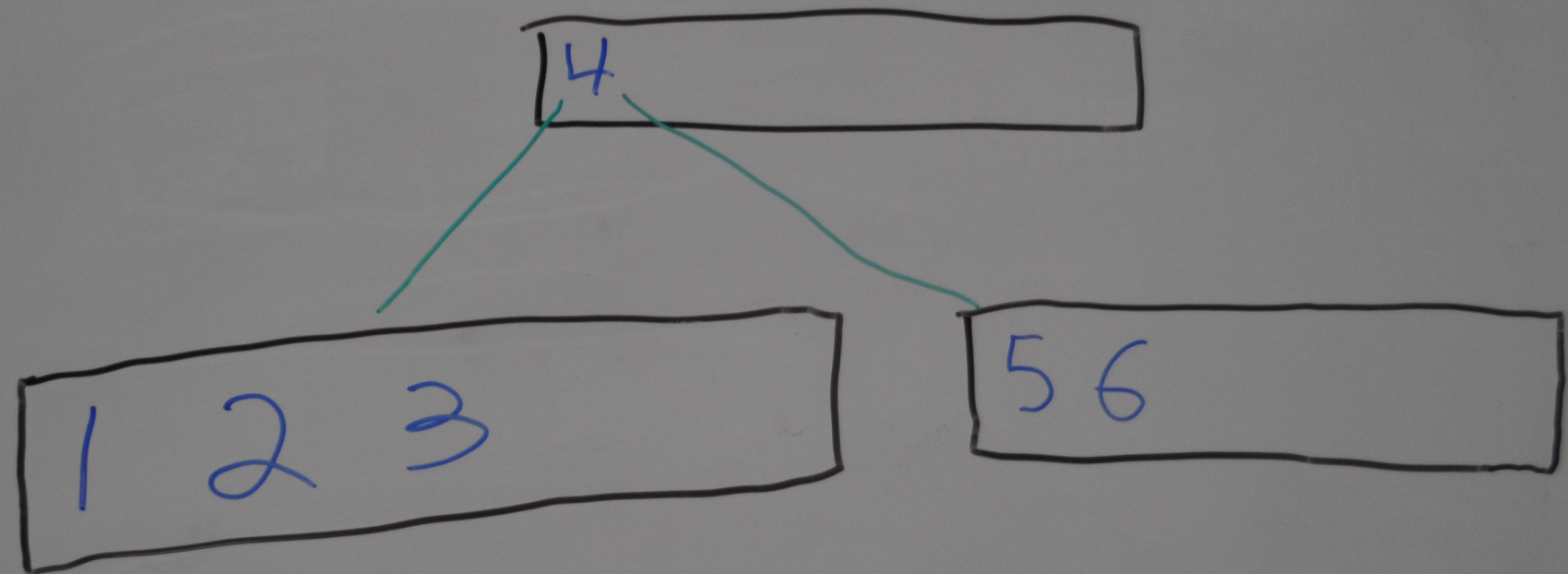
B-Tree

1 2 3 4 5

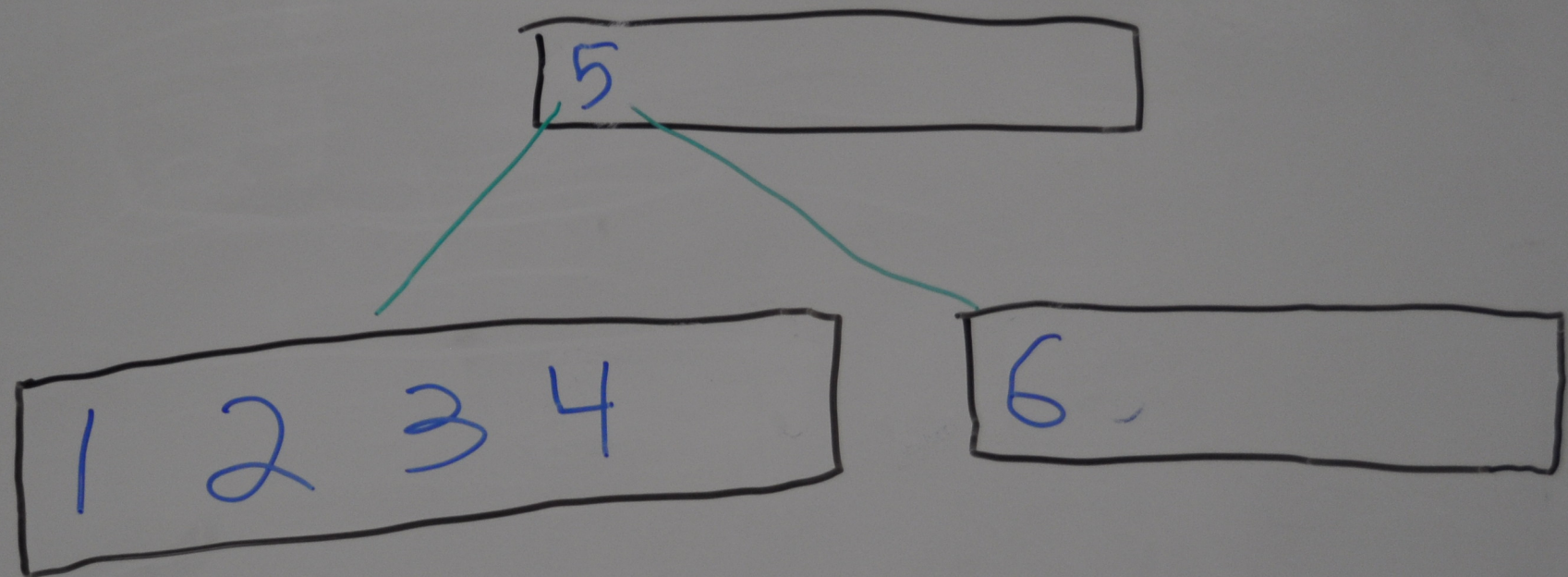
B-Tree

1 2 3 4 5 6

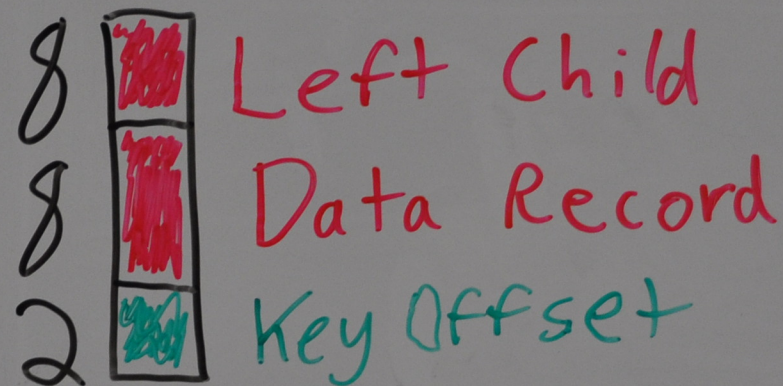
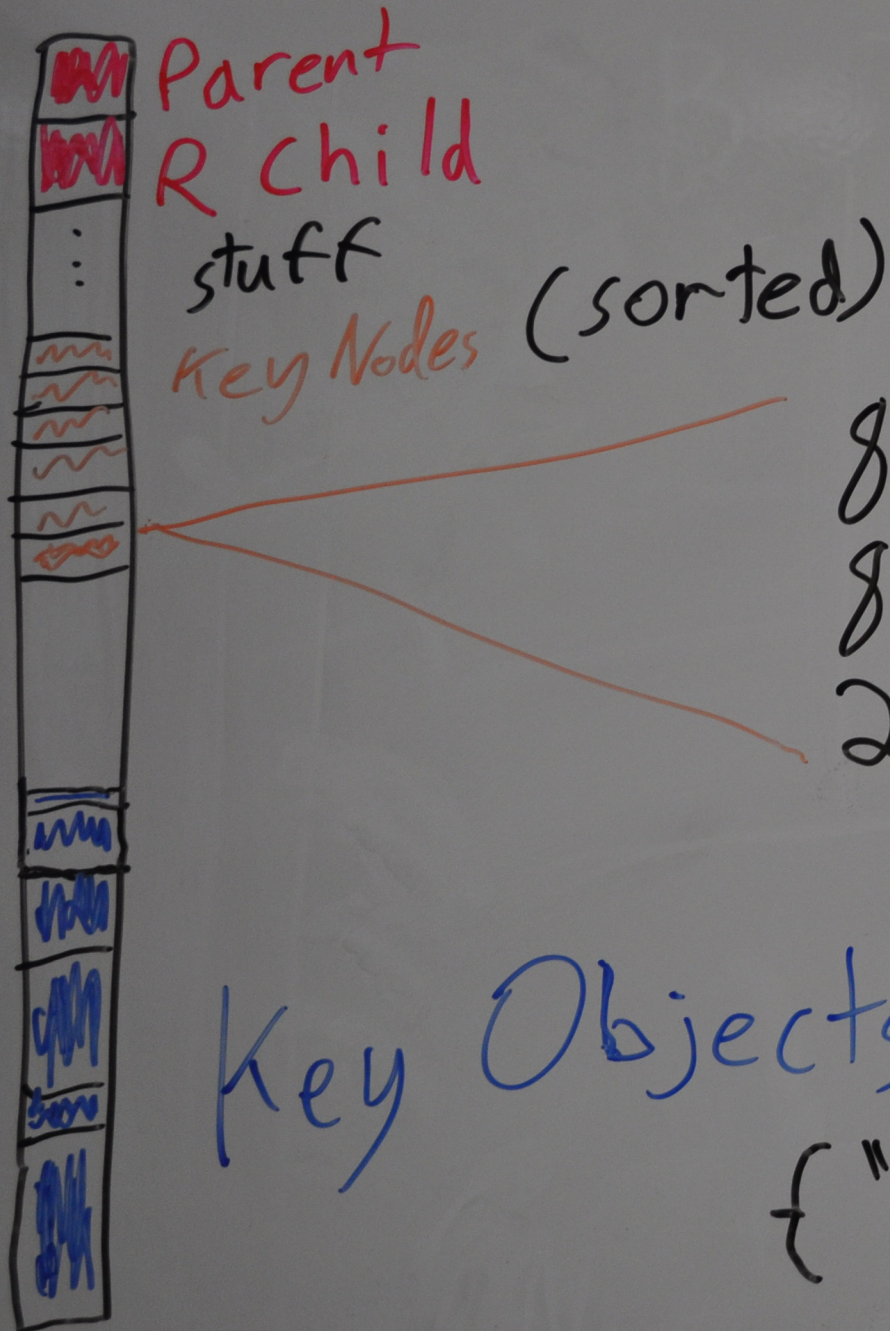
B-Tree



B-Tree



Bucket



Key Objects (unsorted)

{ "": 1, "": "Hello" }

Zero-length Keys

MongoDB

Storage Internals

Mathias Stearn
@mathias_mongo
#mongosv

10gen 