PGC POSTGRESCONF CN 2020 | PGConf.Asia 11.17-11.20

Brin index on AO

陈金豹

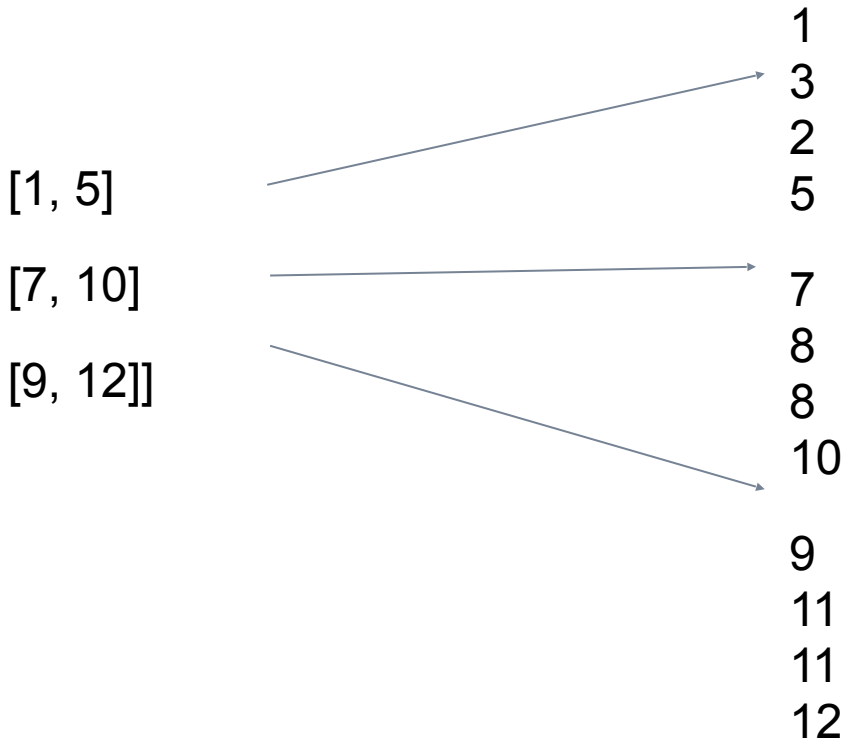
https://cn.greenplum.org
https://2020.postgresconf.cn

PART 01

Brin Index Introduction

Block Range Index

Min Max of a bock range



When use brin

The table is extremely large.

We don't want to pay too much for the index.

Data has some distribution characteristics.

Selection rate of brin

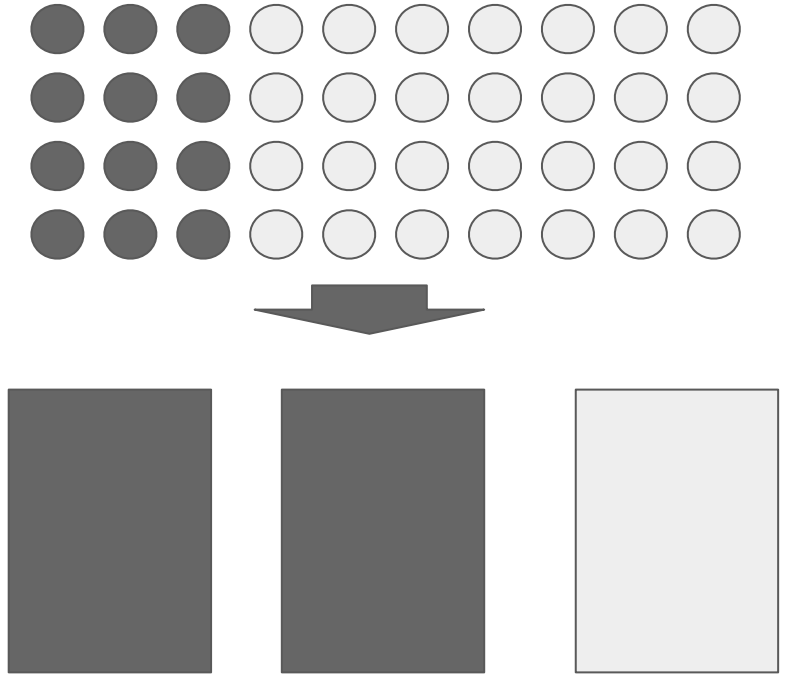
BlockNum: B = 1000

TupleNum: N = 1000000

TuplePerblock: M = 1000

Selection: a = 1%

Brin Selection: 1-((B-1)/B)^(N*a) =
1- 0.000045



Brin scan

select * from t where a > 1 and a < 8;

1, 3, 2, 5	7, 8, 8, 10	9, 11, 11, 12	10, 19, 11, 100

bit map

1 1 0 0

Brin build insert update delele

Generate a record for each block

Record maximum and minimum

Extend the maximum or minimum when the inserted data is out of range

Do nothing when the data is deleted

Brin vacuum

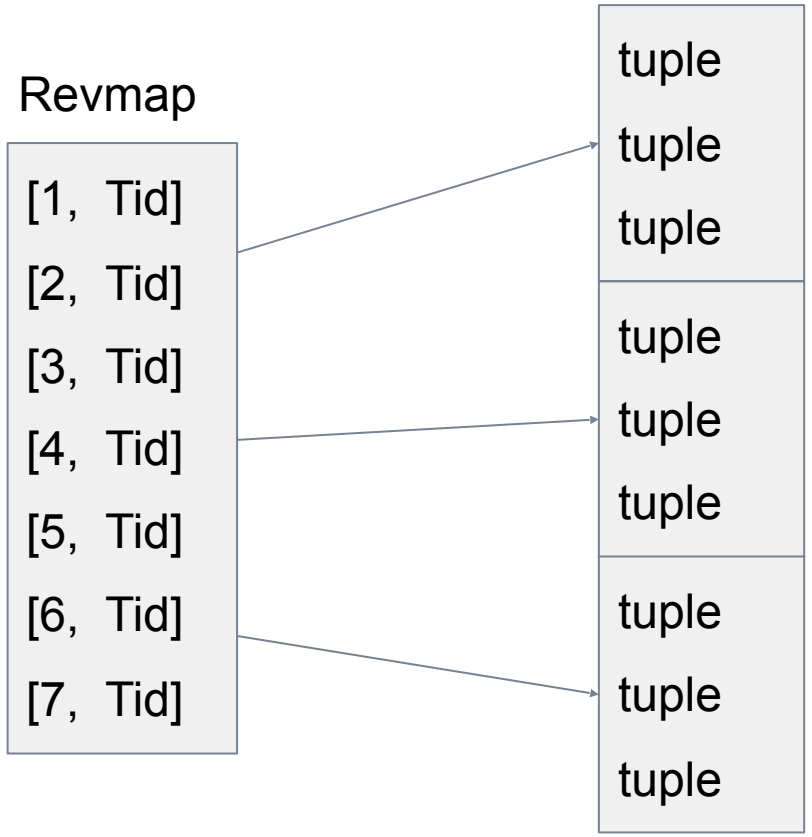
Do nothing on normal vacuum

Rebuild the index after full vacuum

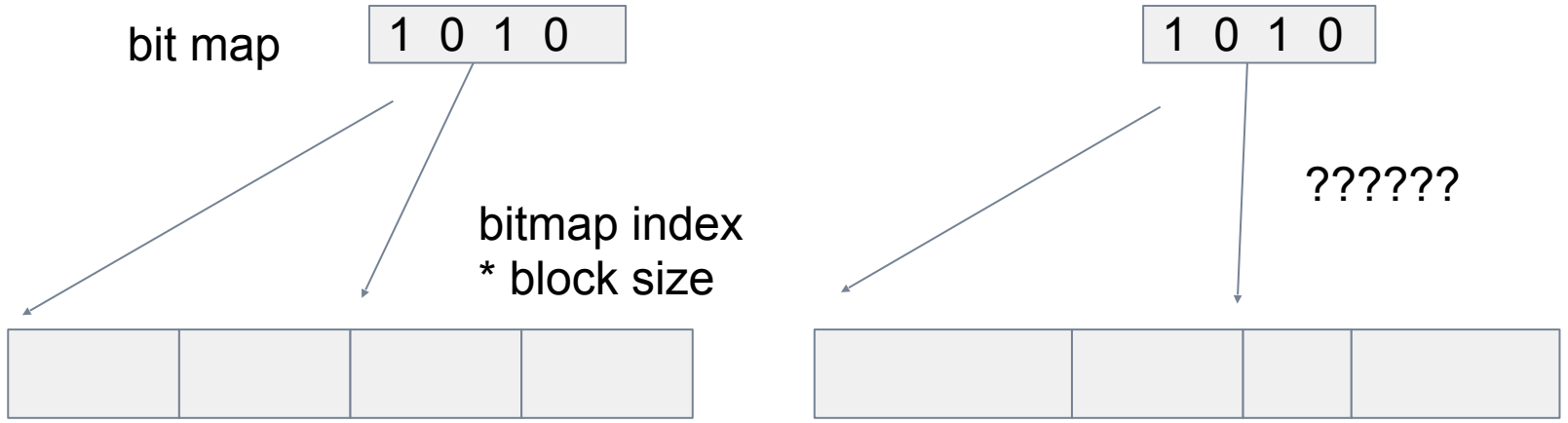
Brin storage

insert a new brin tuple when the inserted data is out of range.

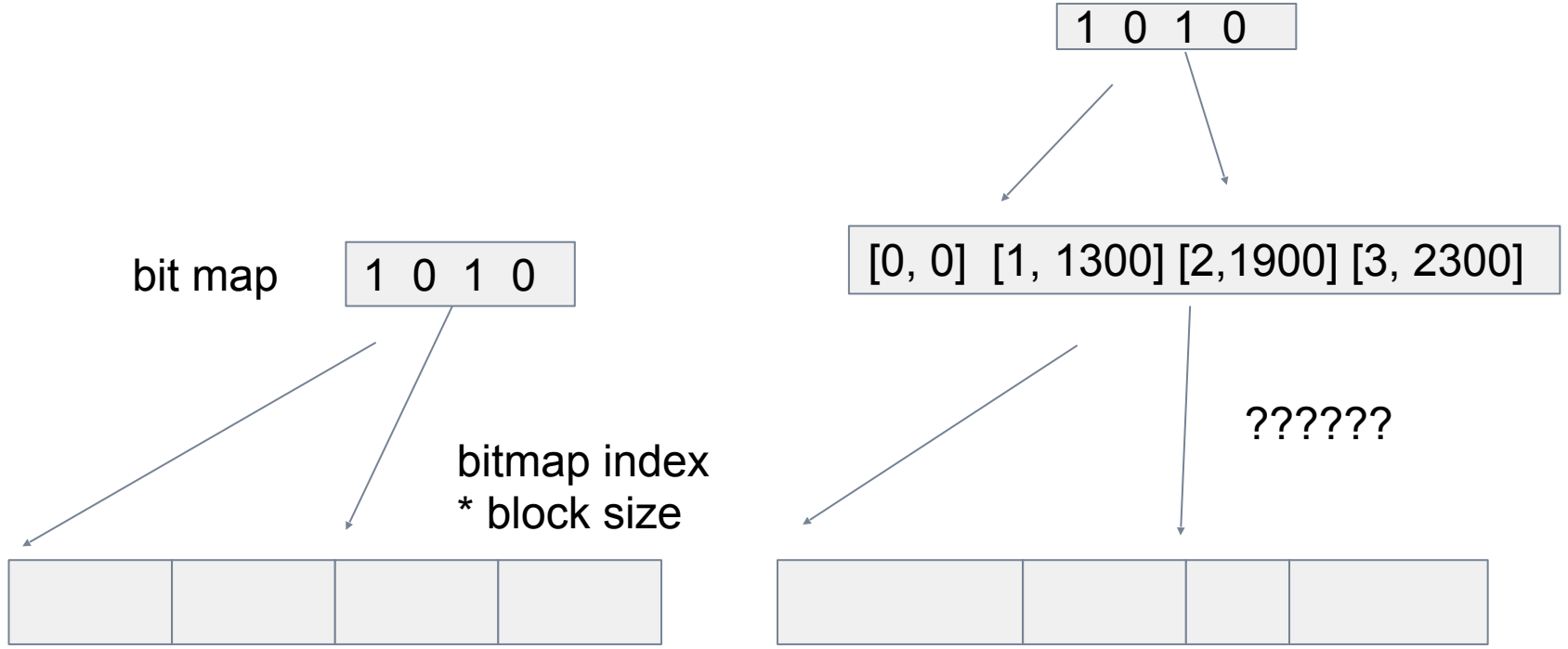
update tid in Revmap record and point to the new brin tuple



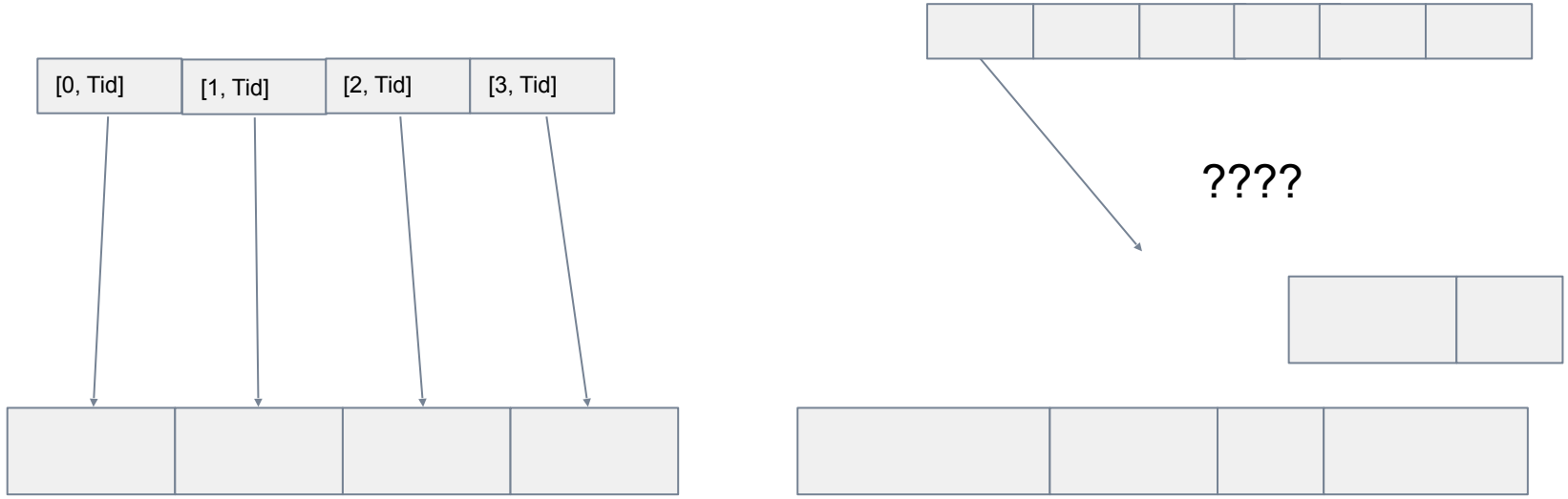
brin scan for heap and ao



brin scan for heap and ao



Revmap for heap and ao



delete and update for ao

Do nothing

Advantages of using brin on Ao

Bigger table

No need to update existing brin tuple

Disadvantages of using brin on Ao

Muti ao seg

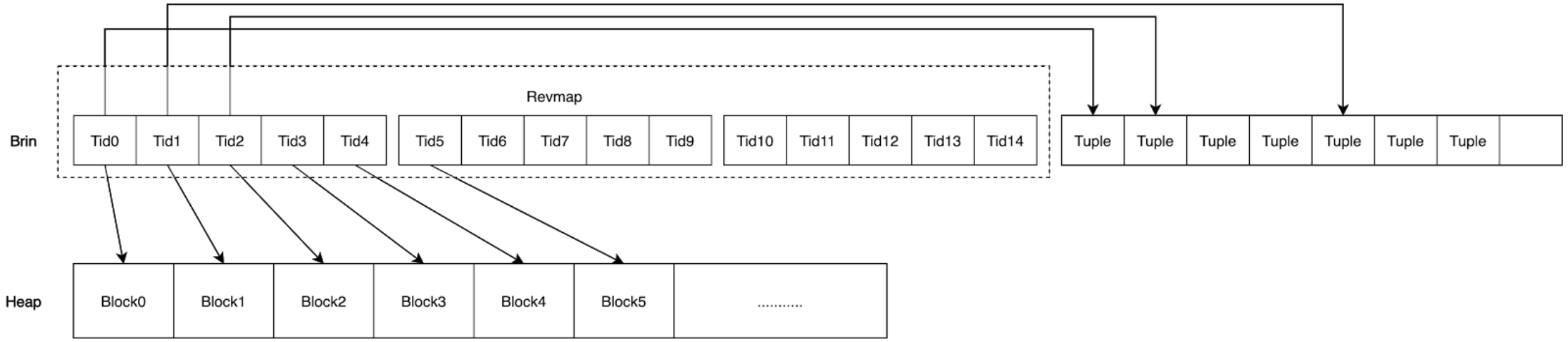
Different block size

Blockdir table cost

PART 02

Brin Index On Append Only Table

Brin on heap


Revmap
Brin
Tid0
Tid1
Tid2
Tid3
Tid4
Tid5
Tid6
Tid7
Tid8
Tid9
Tid10
Tid11
Tid12
Tid13
Tid14
Tuple
Tuple
Tuple
Tuple
Tuple
Tuple
Tuple
Heap
Block0
Block1
Block2
Block3
Block4
Block5
.....

Ao table

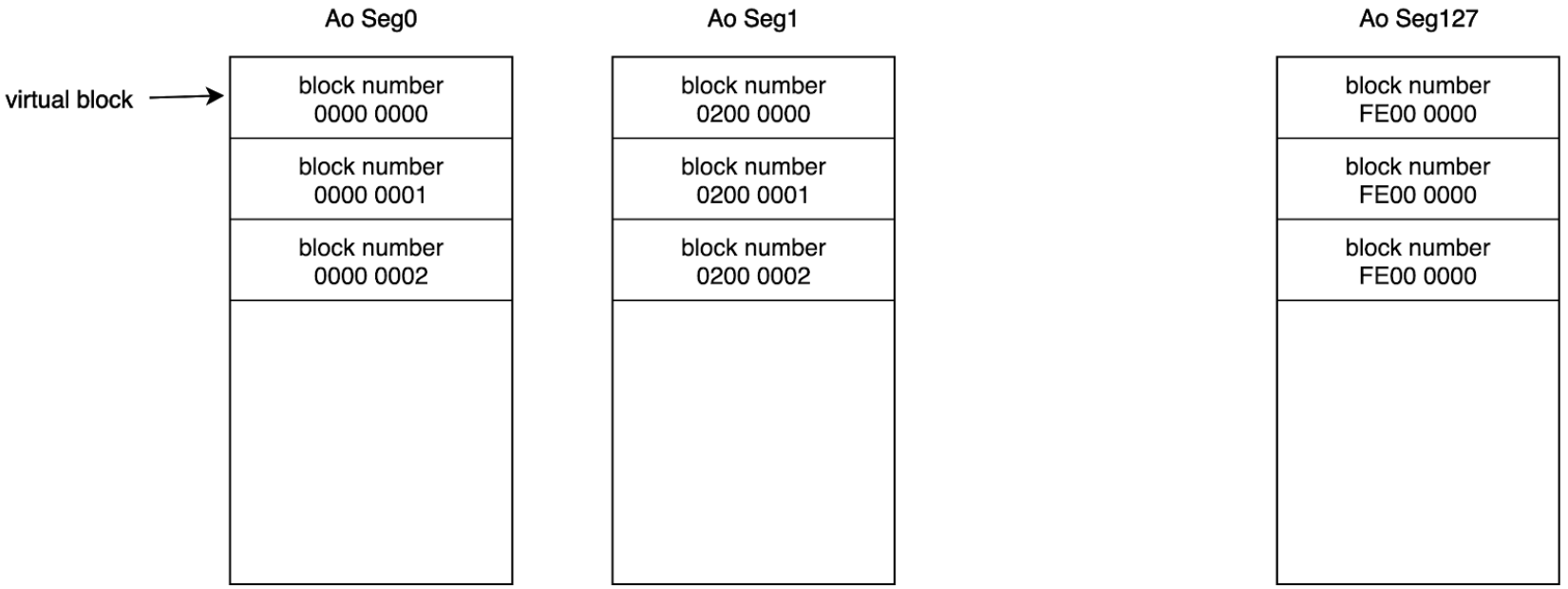
The Ao table is logically composed of 128 aosegs to support concurrent inserts

Each tuple in the Ao table corresponds to a virtual tid

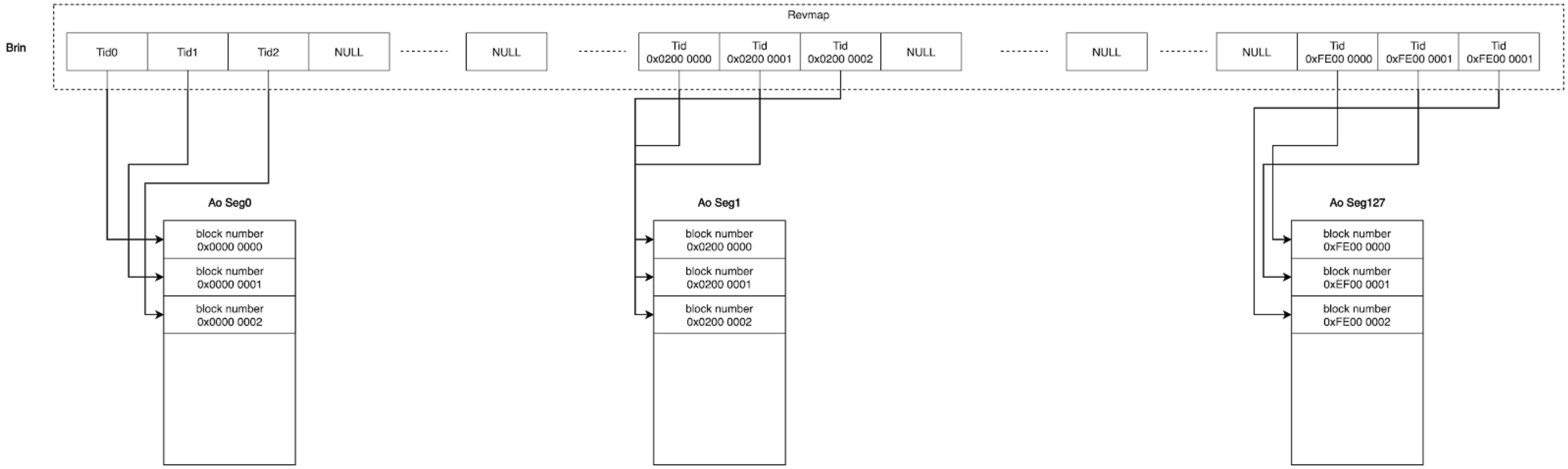
The virtual tid of the first tuple of each Aoseg is equal to (248/128)*segnum

The first virtual block number of each Aoseg is equal to (232/128) * segnum

Ao table


Ao Seg0
Ao Seg1
Ao Seg127
virtual block
block number 0000 0000
block number 0000 0001
block number 0000 0002
block number 0200 0000
block number 0200 0001
block number 0200 0002
block number FE00 0000
block number FE00 0000
block number FE00 0000

Revmap with aoseg



Extend an upper level

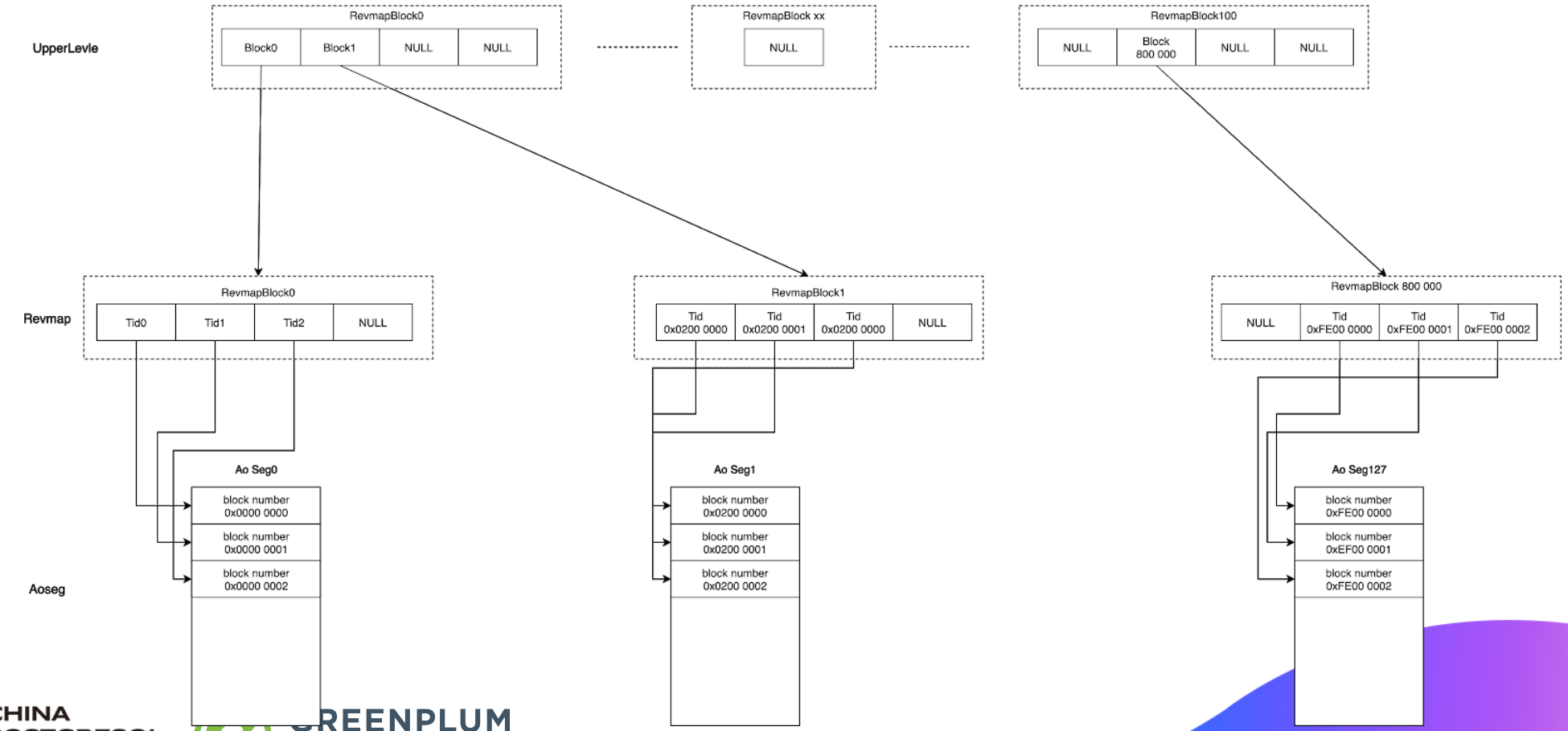
Added an extra upper level on top of the revmap

Record Number on Upper level = 232/TidNumPerPage = 800,000

upper_index = blocknum/TidNumPerPage

revmap_offset = blocknum%TidNumPerPage

Extend an upper level



PART 03

Performance test

Performance test

```
create table aocs(a int, b int) with(appendonly=true, ORIENTATION = column);
insert into aocs select i,i from generate_series(1,10000000)i;
create index abidx on ao using brin(b) with(pages_per_range=1);
create index atidx on ao using btree(b);
```

Test1

select gp_segment_id,* from ao where b=999999;

seqscan: 3957.205 ms

brin-bitmapscan: 36.456 ms

btree-bitmapscan: 18.111 ms

Test2

select gp_segment_id,* from ao where b > 1000000 and b < 1010000;

seqscan: 5757.855 ms

brin-bitmapscan: 57.838 ms

btree-bitmapscan: 42.250 ms

Test3

select gp_segment_id,* from ao where b > 1000000 and b < 2000000;

seqscan: 6413.329 ms

brin-bitmapscan: 2241.363 ms

btree-bitmapscan: 2141.896 ms

Size

ao: 180,198,032

atidx-btree: 222,920,704

abidx-brin: 6,553,600

GREENPLUM DATABASE®

微信技术讨论群
添加入群小助手：gp_assistant

微信公众号
技术干货、行业热点、活动预告

欢迎访问Greenplum中文社区：cn.greenplum.org

GREENPLUM DATABASE®

CONTACT

THANKS

CONTACT INFORMATION

CHINA POSTGRESQL ASSOCIATION | PGC POSTGRESCONF CN 2020 | PGConf.Asia