SASI, Cassandra on the full text search ride

DuyHai DOAN – Apache Cassandra™ Evangelist

1	5 minutes introduction to Apache Cassandra™
2	SASI introduction
3	SASI cluster-wide
4	SASI local read/write path
5	Query planner
6	Some benchmarks
7	Take away

Trademark Policy

From now on ...

Cassandra == Apache Cassandra™

5 minutes introduction to Apache Cassandra™

APACHE: BIG_DATA EUROPE

@doanduyhai

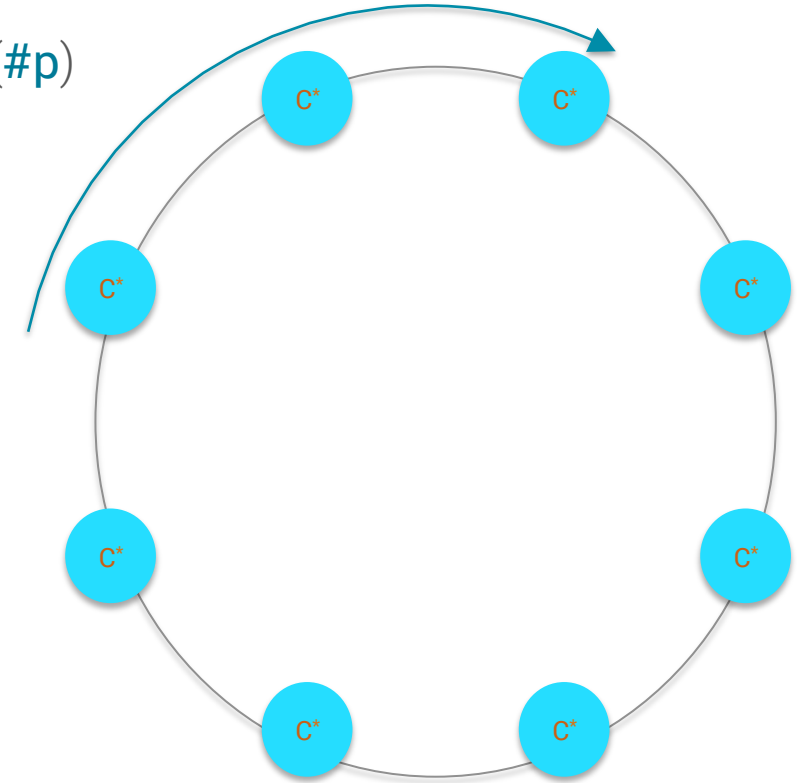
The tokens

Random hash of #partition → token = hash(#p)

Hash:] −x, x]

hash range: 2^{64} values

x = $2^{64}/2$



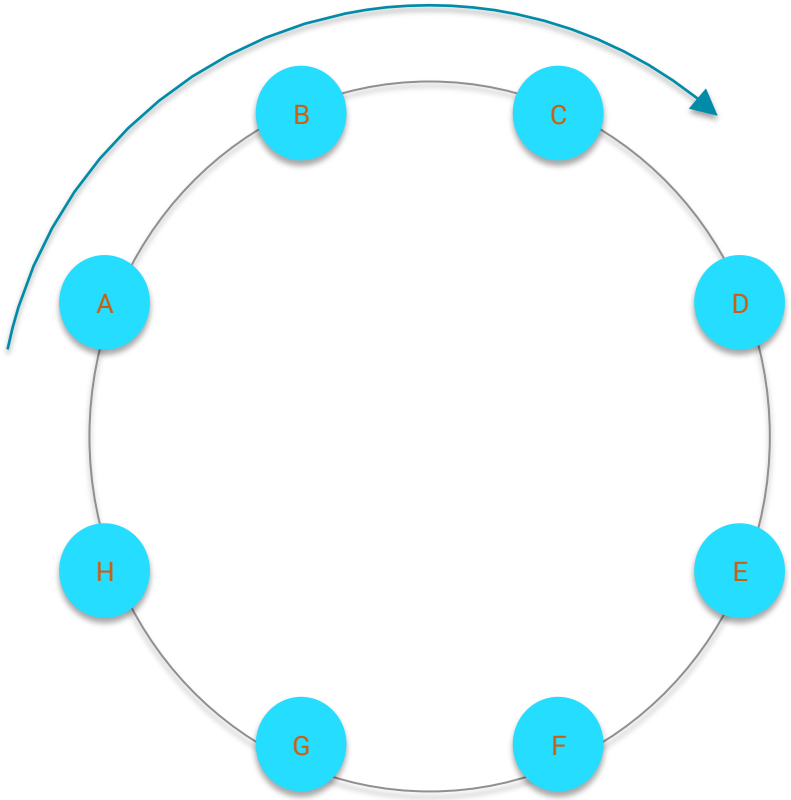
Token ranges

$$A:\left]-x,-\frac{3x}{4}\right] \qquad E:\left]0,\frac{x}{4}\right]$$

$$B:\left]-\frac{3x}{4},-\frac{2x}{4}\right] \qquad F:\left]\frac{x}{4},\frac{2x}{4}\right]$$

$$C:\left]-\frac{2x}{4},-\frac{x}{4}\right] \qquad G:\left]\frac{2x}{4},\frac{3x}{4}\right]$$

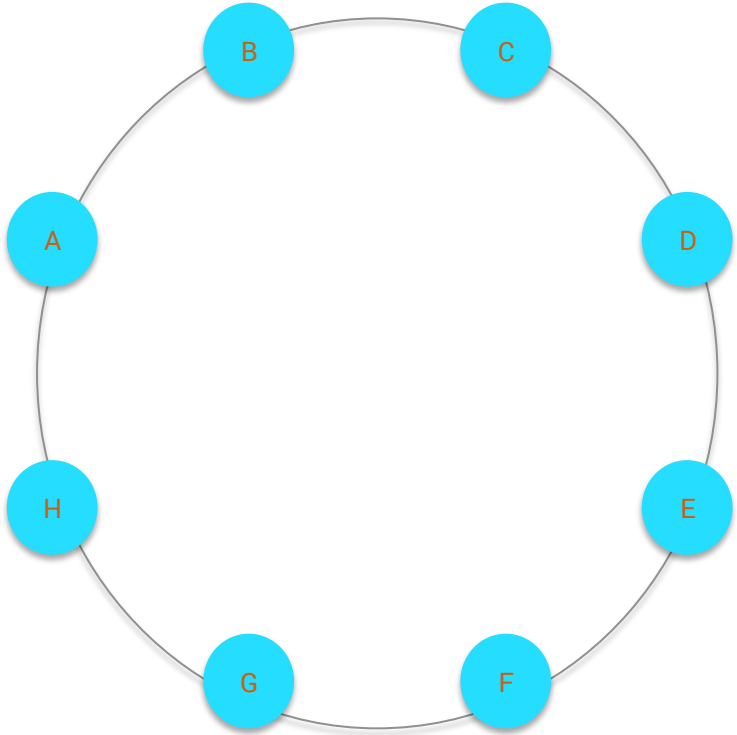
$$D:\left]-\frac{x}{4},0\right] \qquad H:\left]\frac{3x}{4},x\right]$$



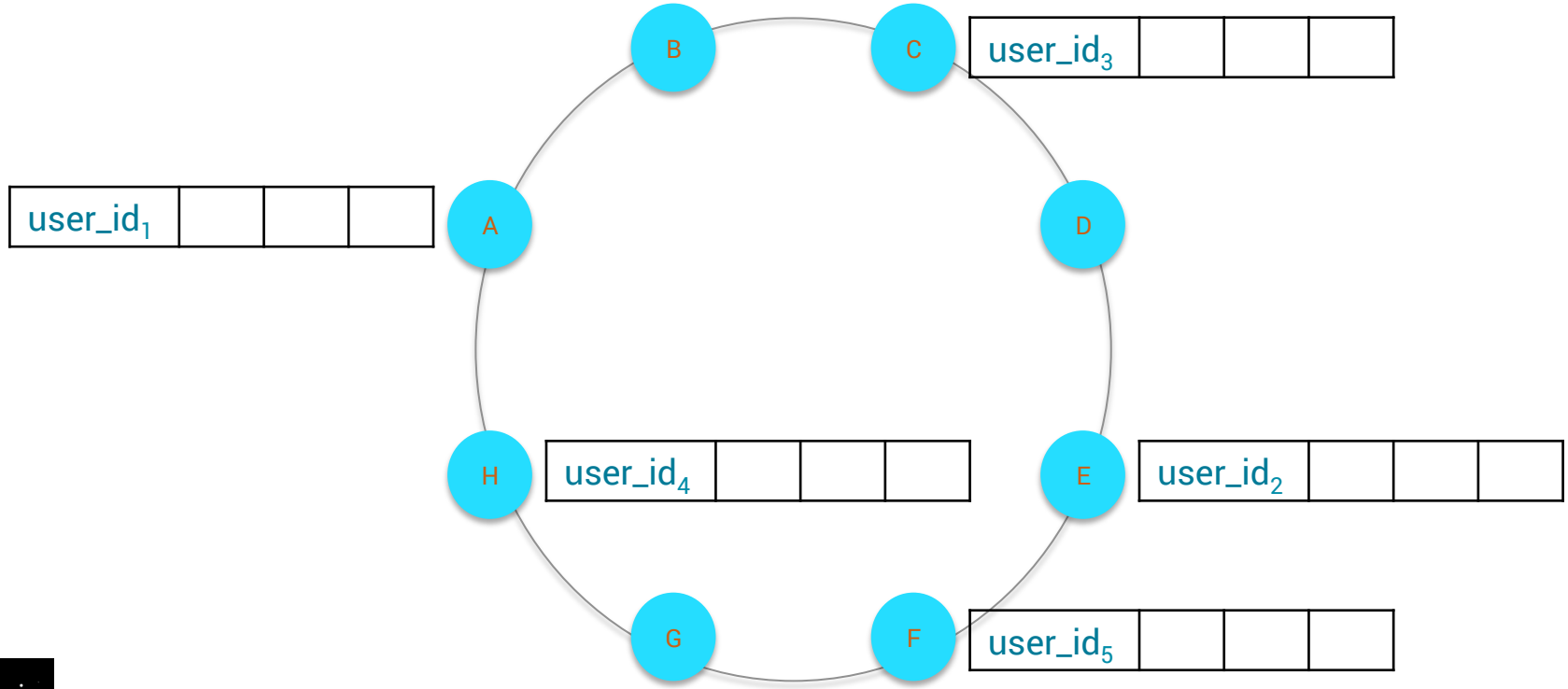
Distributed tables

```
CREATE TABLE users(
   user_id int,
   ...,
   PRIMARY KEY(user_id)
),
```

$user\_id_1$			
$user\_id_2$			
$user\_id_3$			
$user\_id_4$			
$user\_id_5$			



Distributed tables

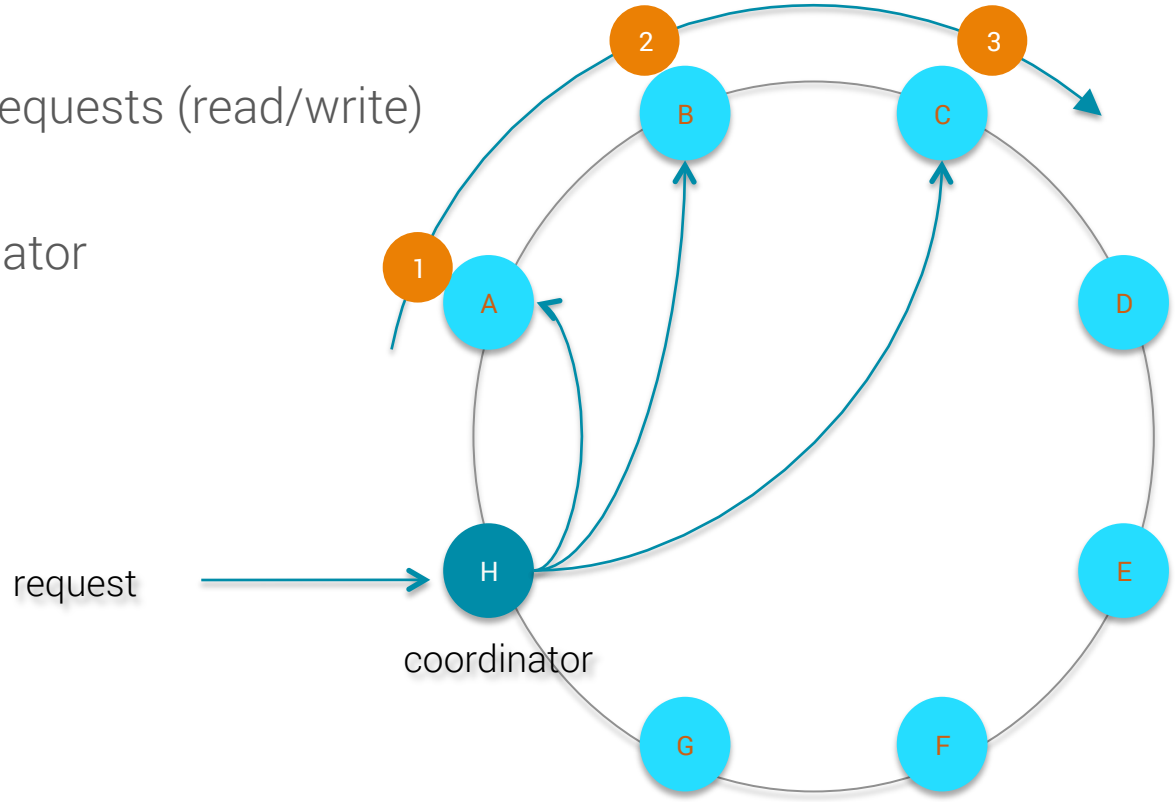


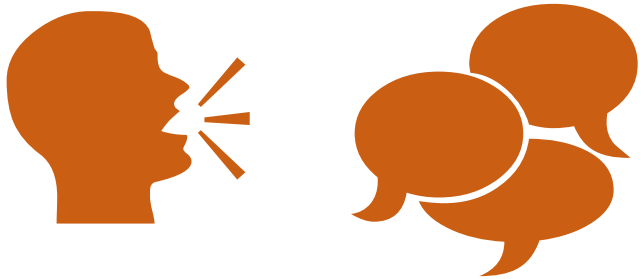
Coordinator node

Responsible for handling requests (read/write)

Every node can be coordinator

- masterless
- no SPOF
- proxy role





Q & A

SASI introduction

@doanduyhai

What is SASI ?

- **S**STable-**A**ttached **S**econdary **I**ndex → new 2nd index impl that follows SSTable life-cycle

- Objective: provide more performant & capable 2nd index

Who created it ?

Open-source contribution by an engineers team

Pavel Yaskevich
xedin

- Apple Inc.
- San Francisco, CA
- xedin@apache.org
- http://tinkerpop.com
- Joined on Aug 20, 2008

Jordan West
jrwest

- Apple
- San Francisco, CA
- jordanrw@gmail.com
- http://blog.jordanwest.me
- Joined on Dec 15, 2009

Jason Brown
jasobrown

- Apple
- Silicon Valley, CA
- http://www.apple.com
- Joined on Feb 1, 2012

Mikhail Stepura
Mishail

- Apple
- mikhail.stepura@outlook.com
- Joined on May 17, 2010

mkjellman

- Joined on May 11, 2012

Why is it better than native 2nd index ?

- follow SSTable life-cycle (flush, compaction, rebuild …) → more optimized
- new data-strutures
- range query (<, ≤, >, ≥) possible
- full text search options

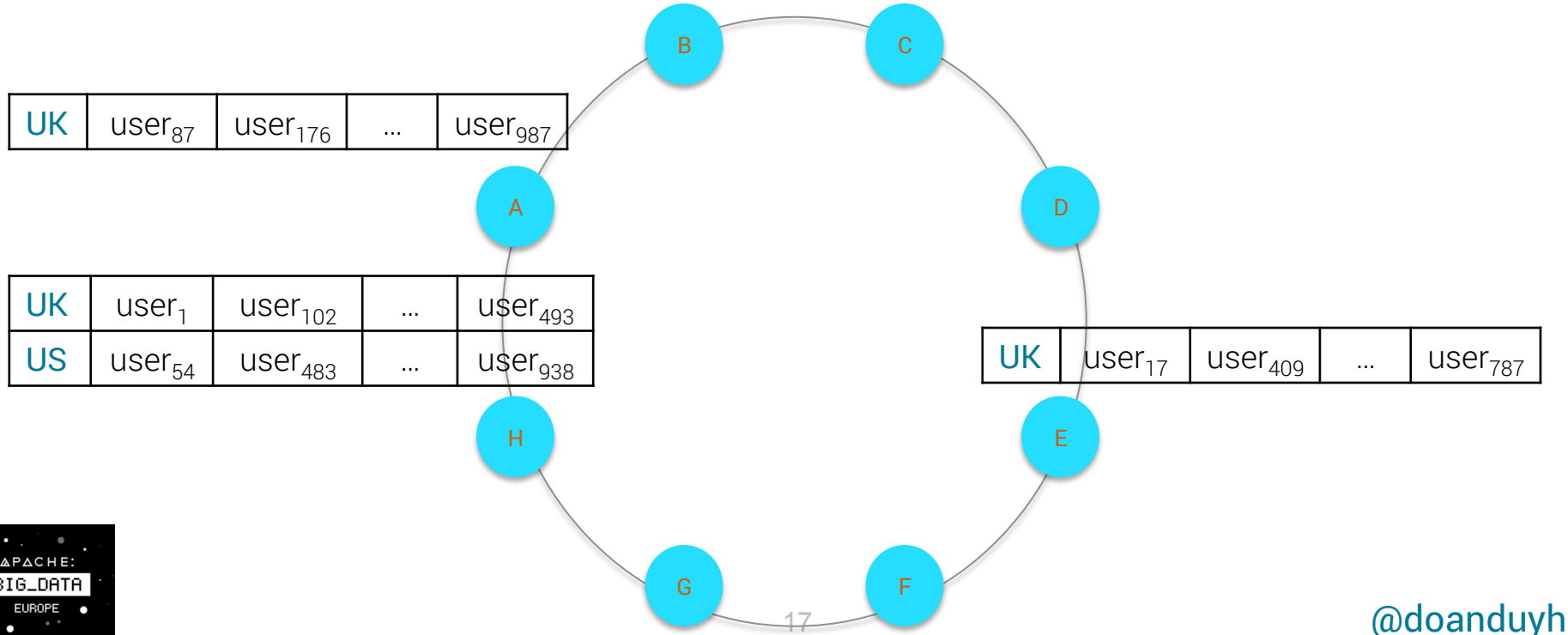
Demo

SASI cluster-wide

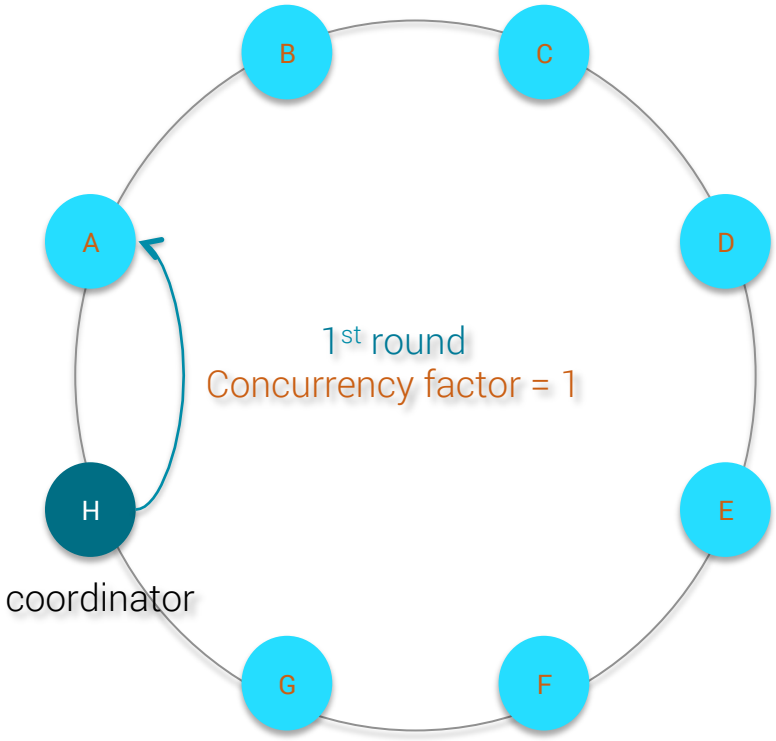
@doanduyhai

Distributed index

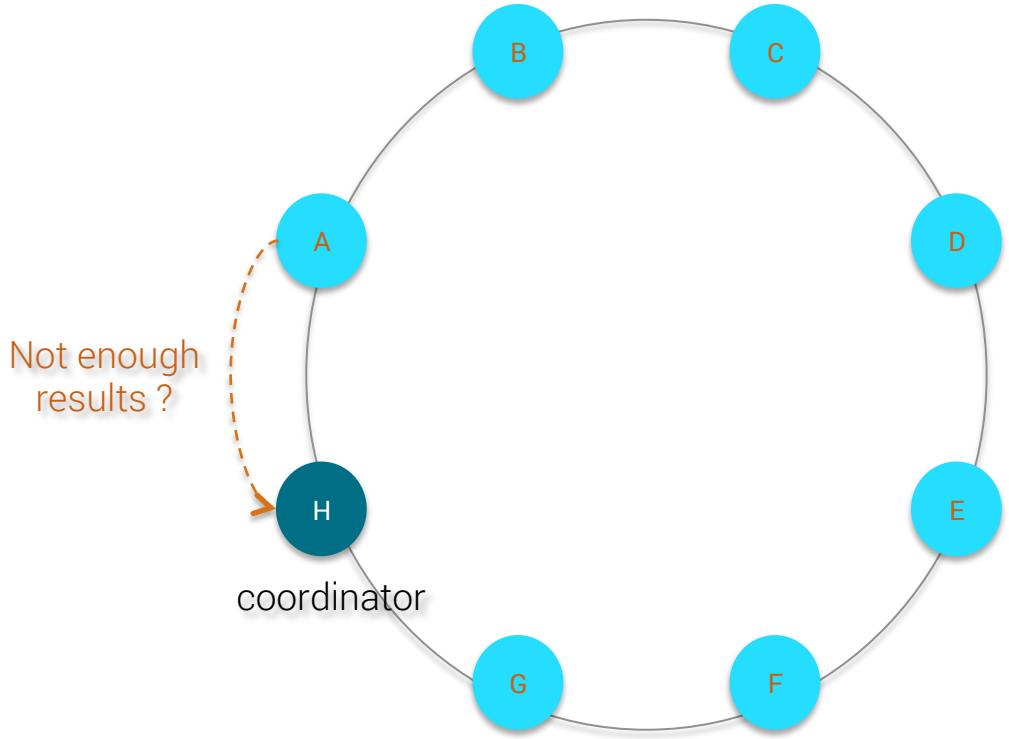
On cluster level, SASI works exactly like native 2nd index



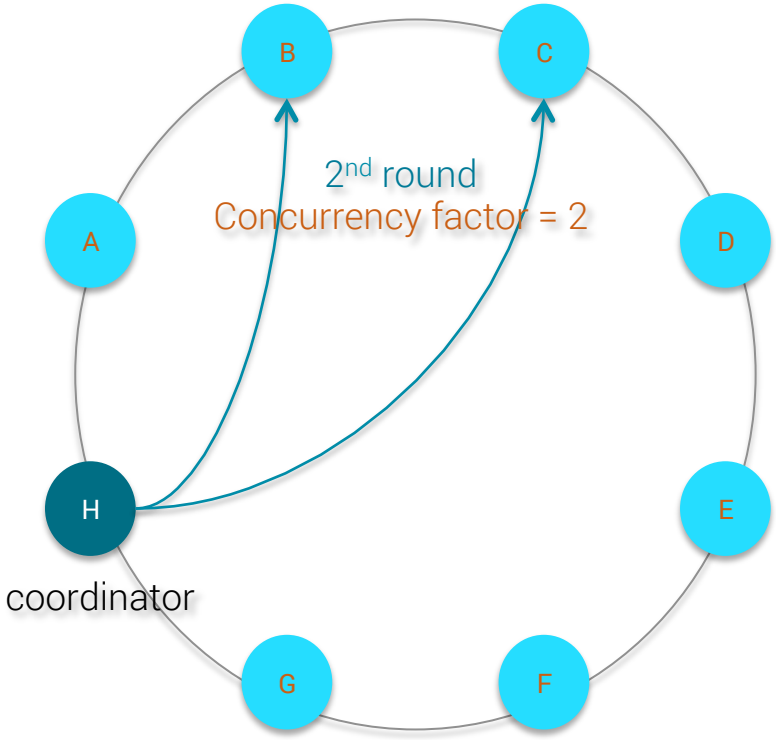
Distributed search algorithm



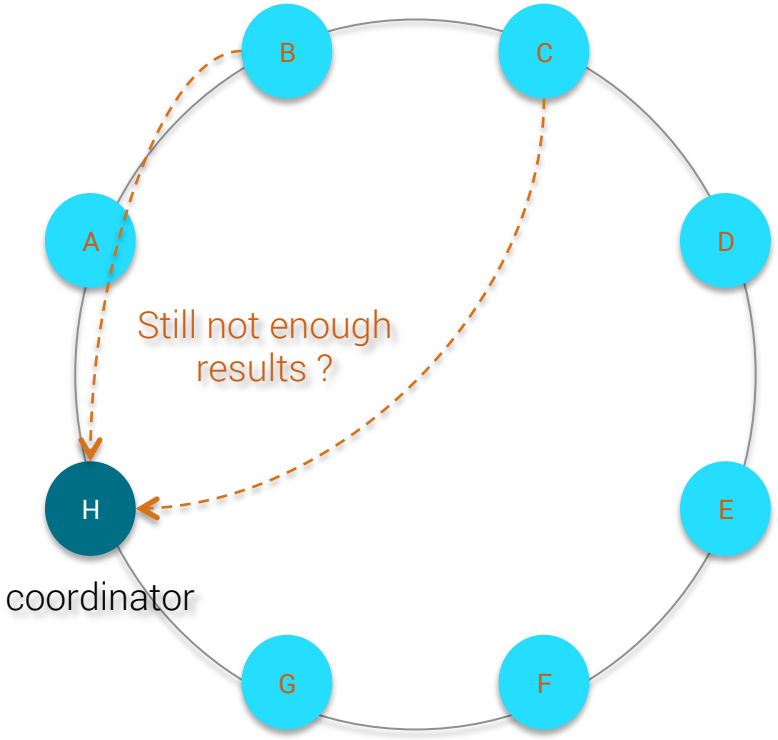
Distributed search algorithm



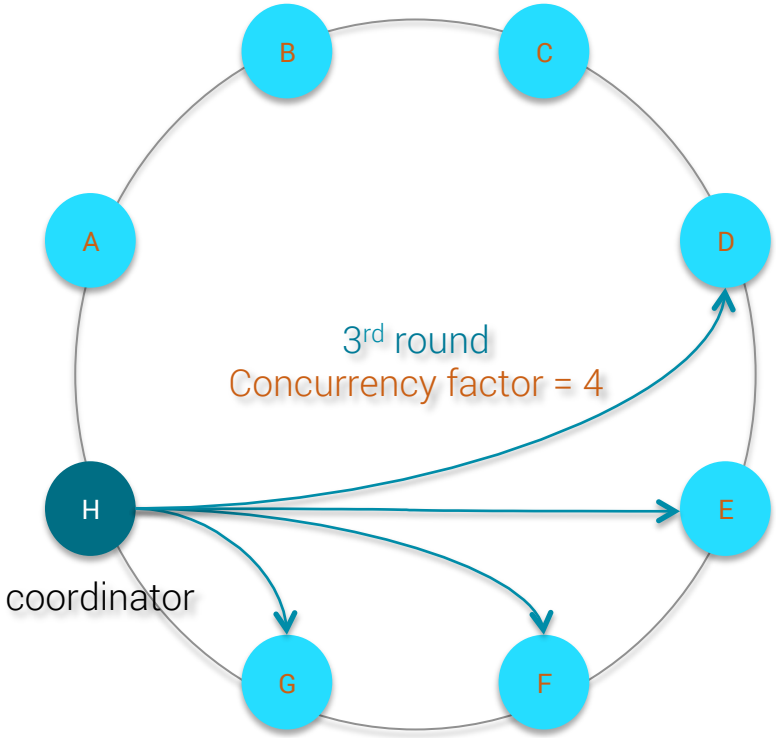
Distributed search algorithm



Distributed search algorithm



Distributed search algorithm



Concurrency factor formula

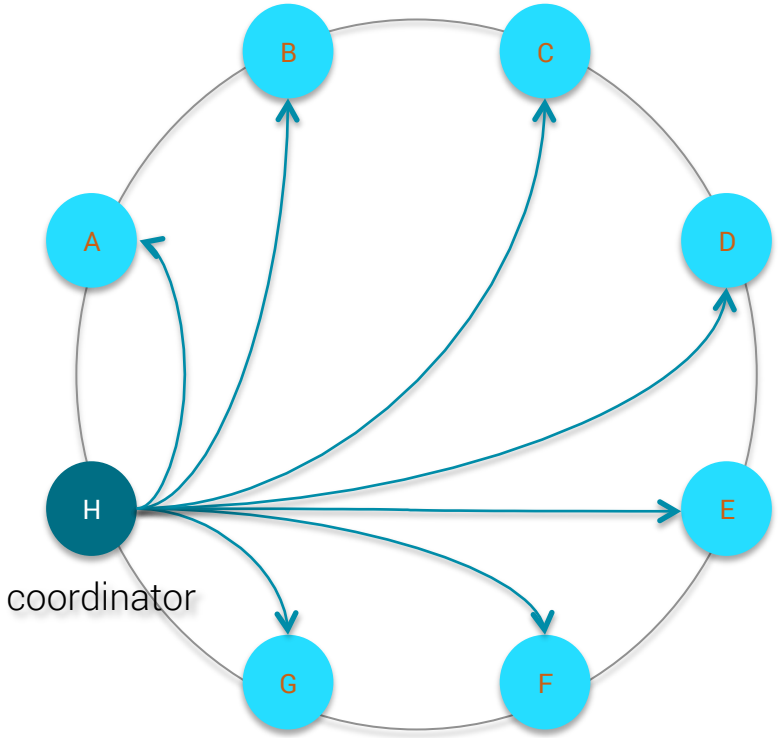
$$\text{CONCURRENCY\_FACTOR} = \max\left(1, \min\left(token\_range\_count, \left\lceil \frac{requested\_LIMIT}{estimate\_rows\_by\_token\_range} \right\rceil\right)\right)$$

$$estimate\_rows\_by\_token\_range = \frac{estimate\_rows}{token\_range\_count \times replication\_factor}$$

- more details at: http://www.doanduyhai.com/blog/?p=13191

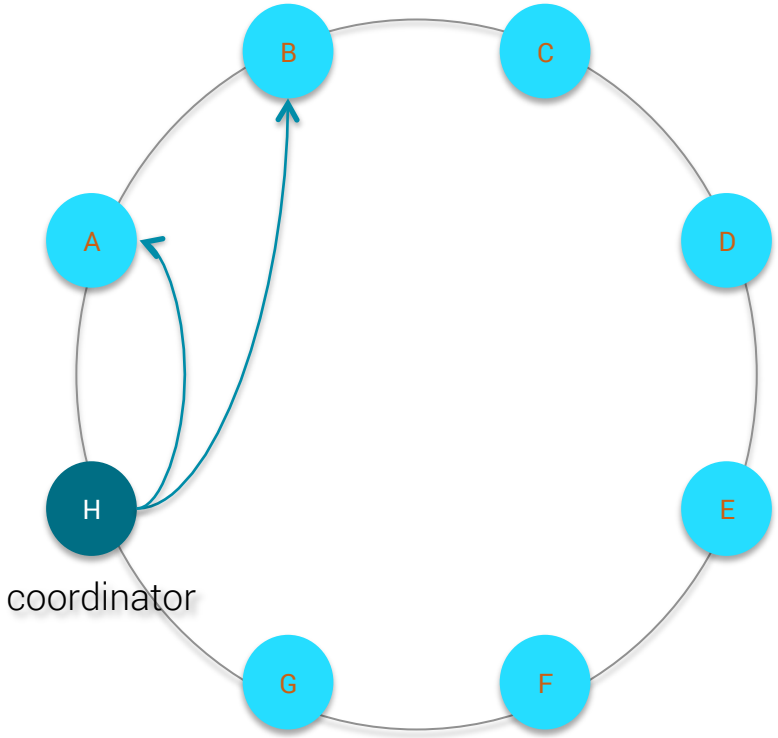
Caveat 1: non restrictive filters

Hit all nodes eventually ☹

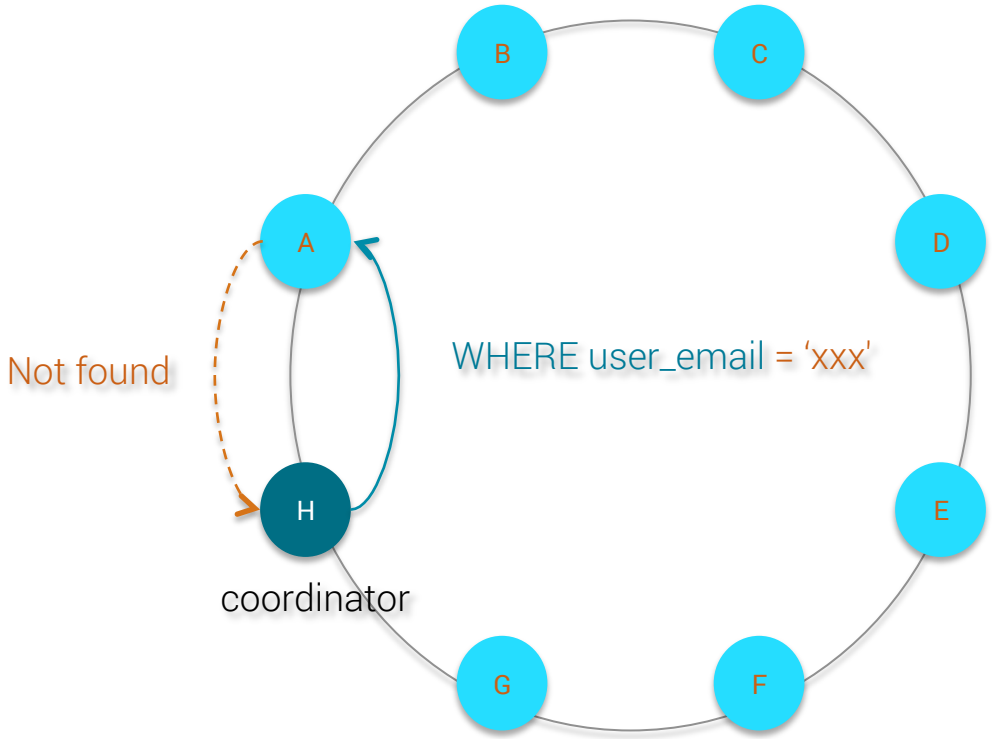


Caveat 1 solution : always use LIMIT

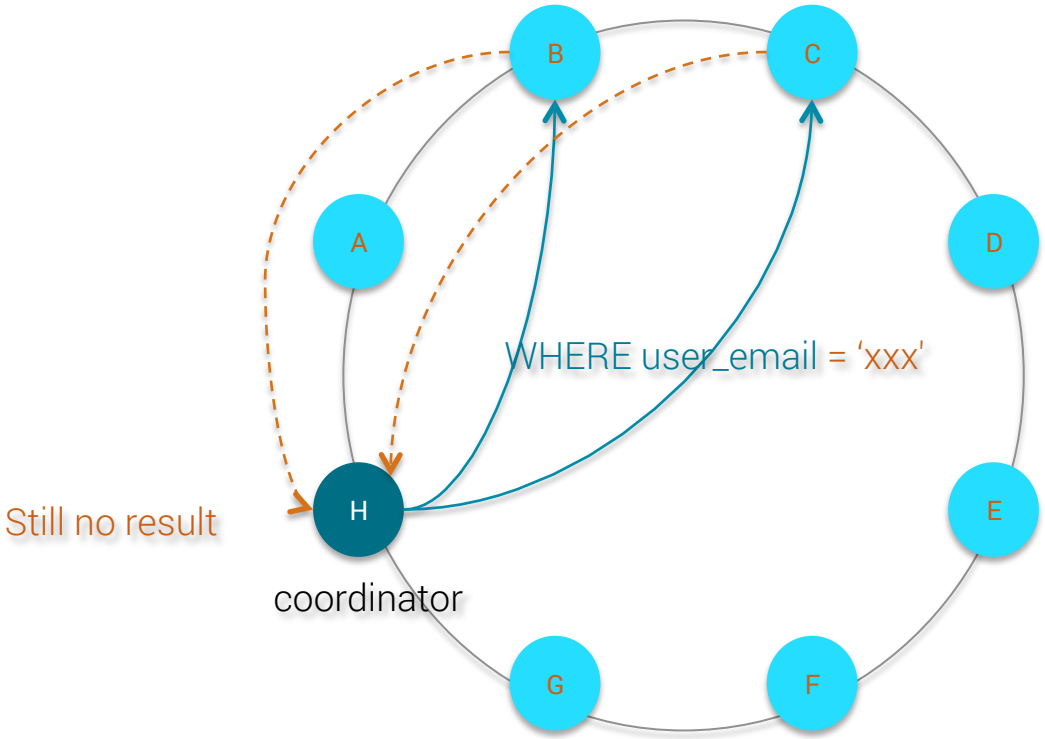
SELECT *
FROM ...
WHERE ...
LIMIT 1000



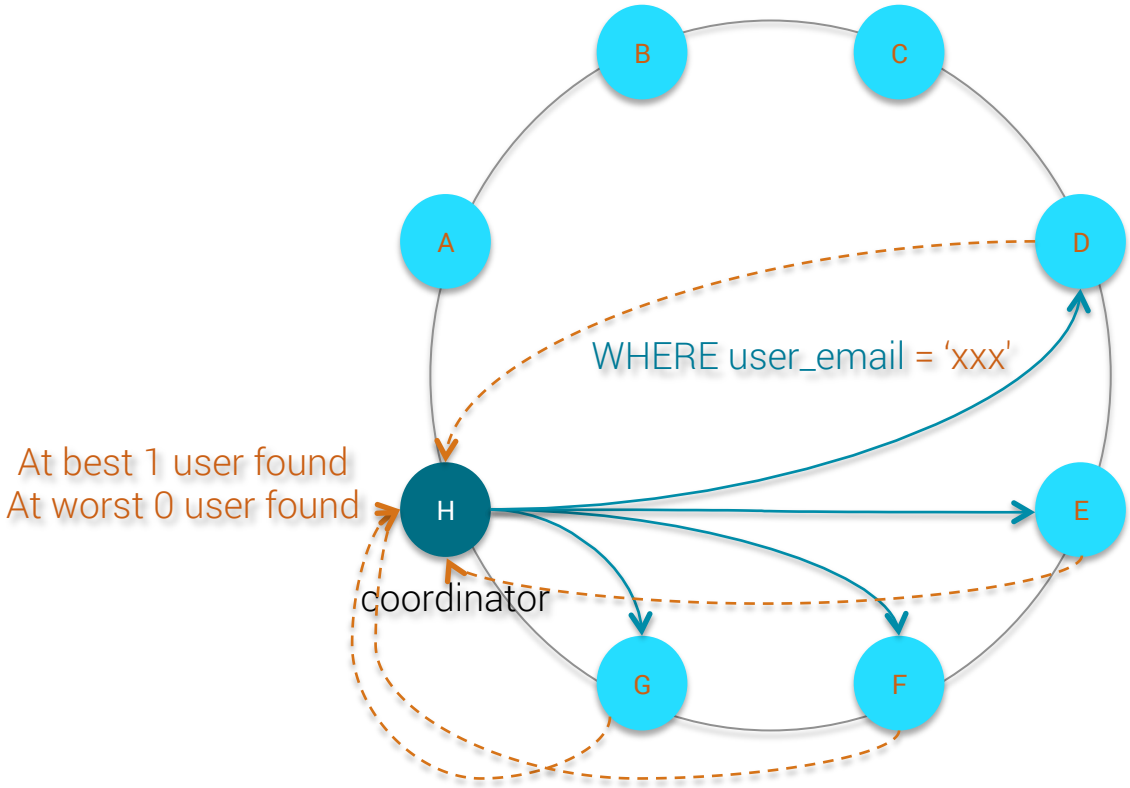
Caveat 2: 1-to-1 index (*user_email*)



Caveat 2: 1-to-1 index (*user_email*)



Caveat 2: 1-to-1 index (*user_email*)



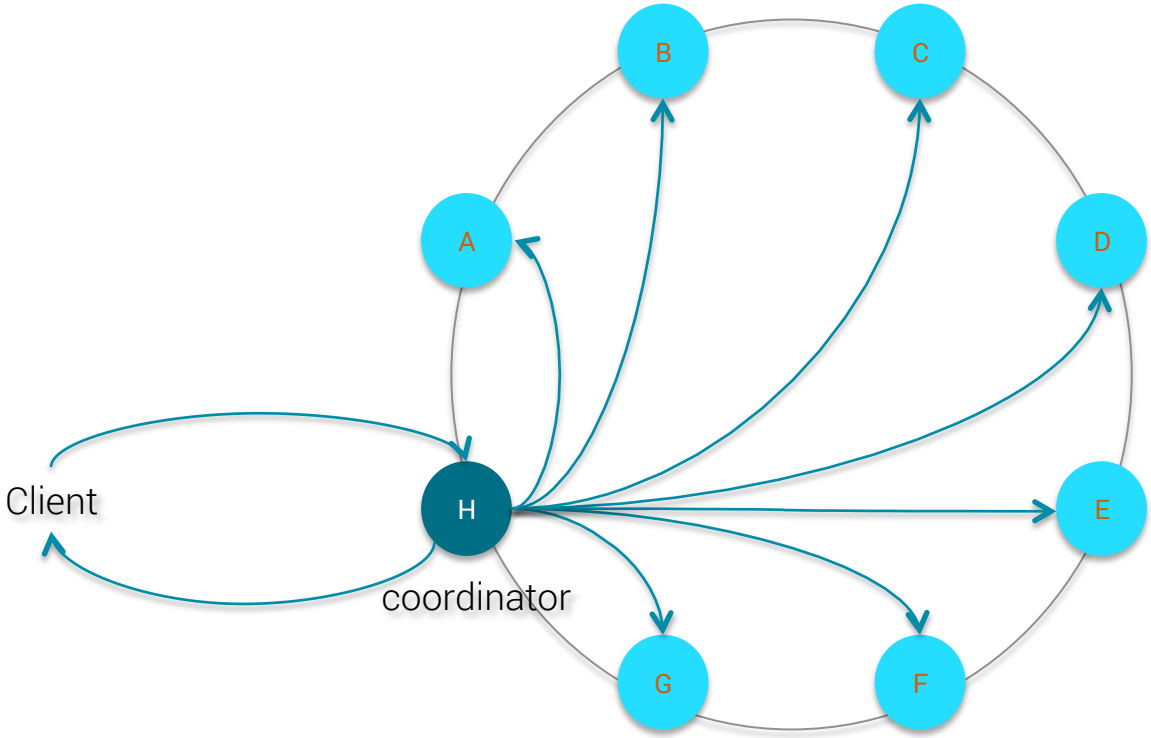
Caveat 2 solution: materialized views

For 1-to-1 index/relationship, use materialized views instead

```
CREATE MATERIALIZED VIEW user_by_email AS
SELECT * FROM users
WHERE user_id IS NOT NULL and user_email IS NOT NULL
PRIMARY KEY (user_email, user_id)
```

But range queries (<, >, ≤, ≥) not possible …

Caveat 3: fetch all rows for analytics use-case



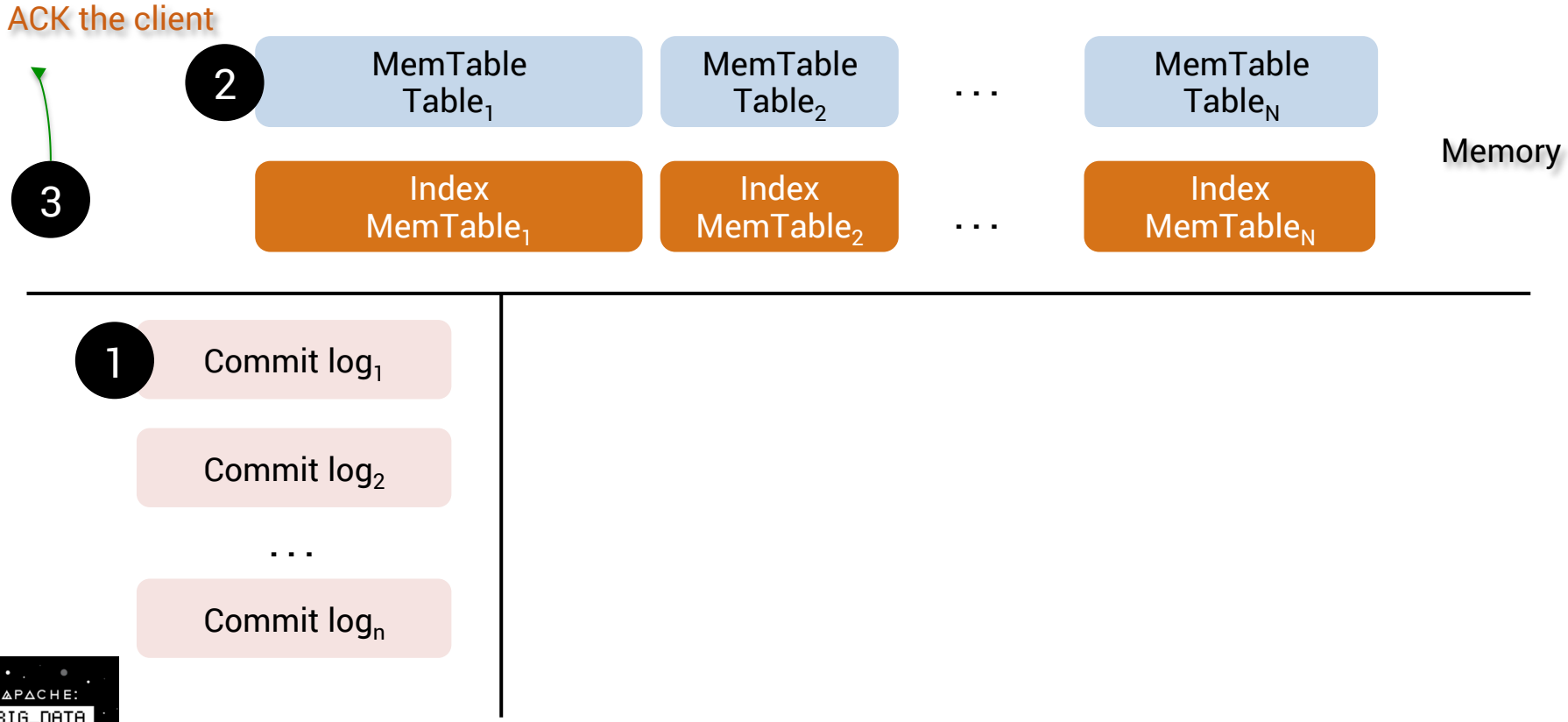
Caveat 3 solution: use co-located Spark

Local index query

SASI local read/write path

@doanduyhai

SASI Life-cycle: in-memory

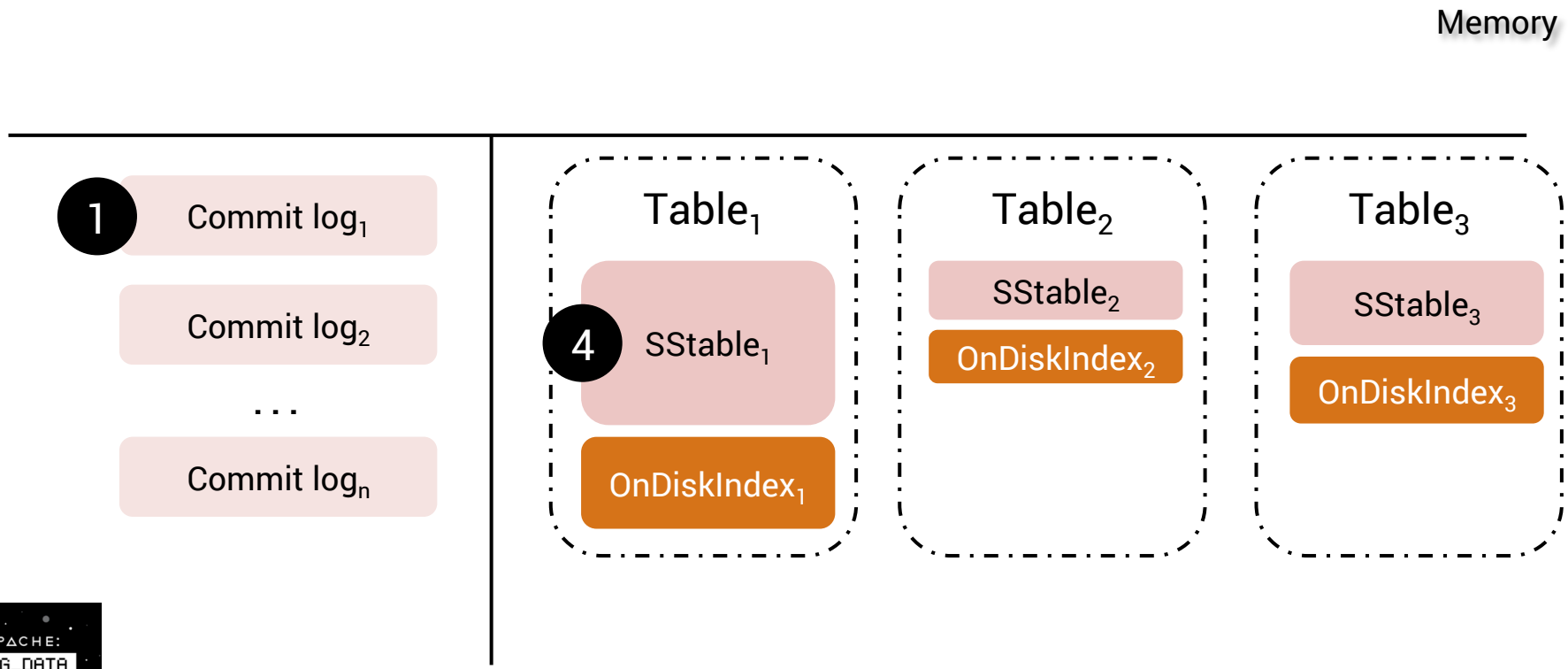


Local write path data structures

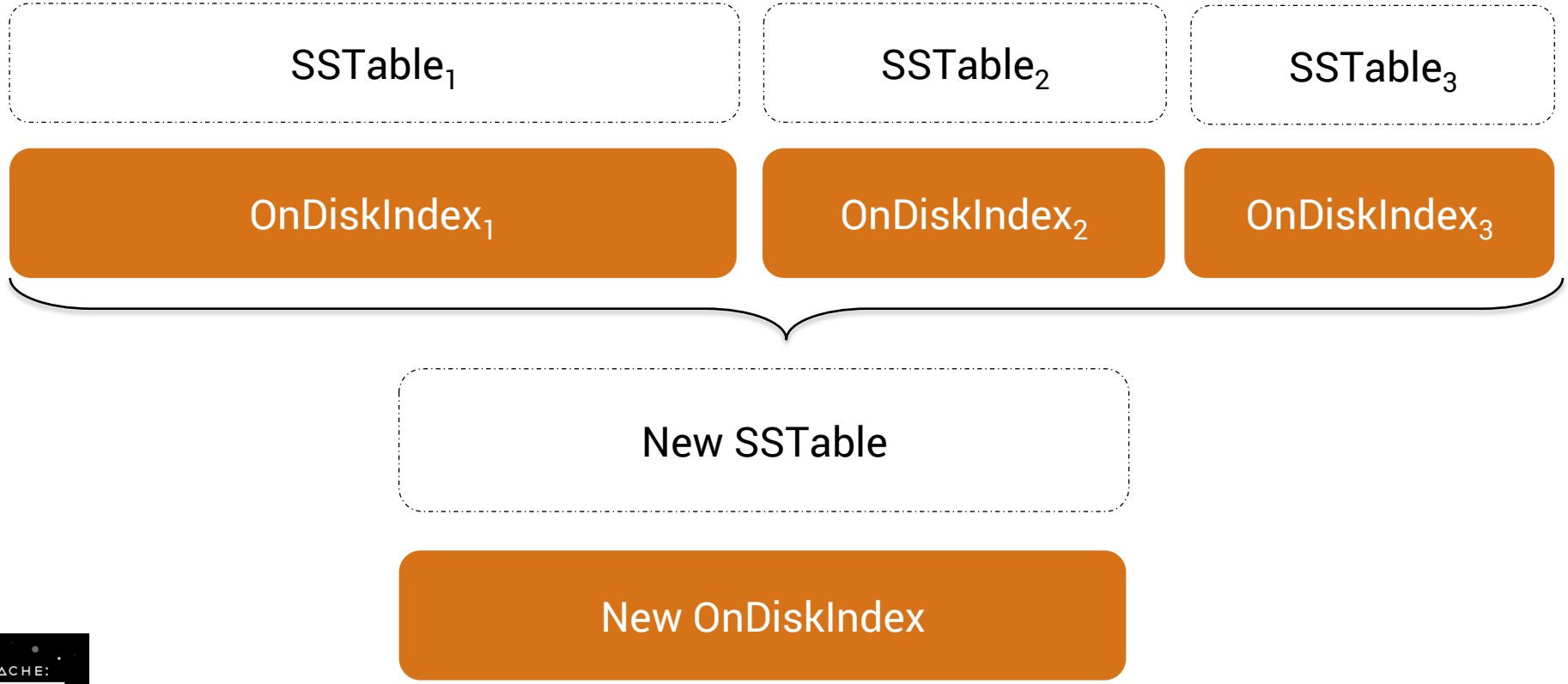
Index mode, data type	Data structure	Usage
PREFIX, text	Guava *ConcurrentRadixTree*	name LIKE 'John%'
CONTAINS, text	Guava *ConcurrentSuffixTree*	name LIKE '%John%' name LIKE '%ny'
PREFIX, other	JDK *ConcurrentSkipListSet*	age = 20 age >= 20 AND age <= 30
SPARSE, other	JDK *ConcurrentSkipListSet*	age = 20 age >= 20 AND age <= 30

suitable for 1-to-N index with N ≤ 5

SASI Life-cycle: flush to SSTable



SASI Life-cycle: compaction



Local write path summary

Index files are built

- on memtable flush
- on compaction flush

To avoid OOM, index files are split into chunk of

- 1Gb for memtable flush
- *max_compaction_flush_memory_in_mb* for compaction flush

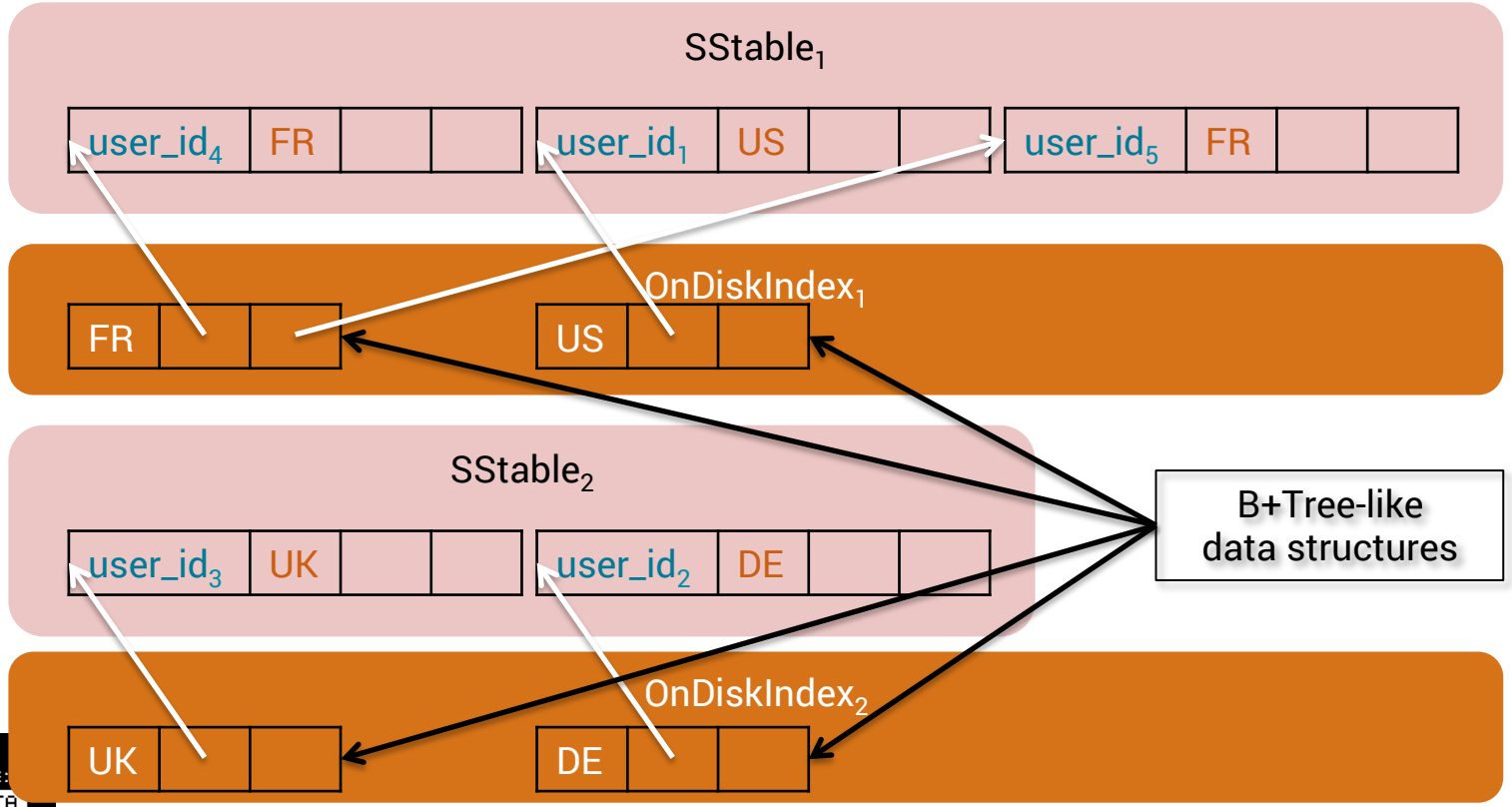
→ consequences: SASI has impact on write bandwidth (CPU & disk I/O)

Local read path

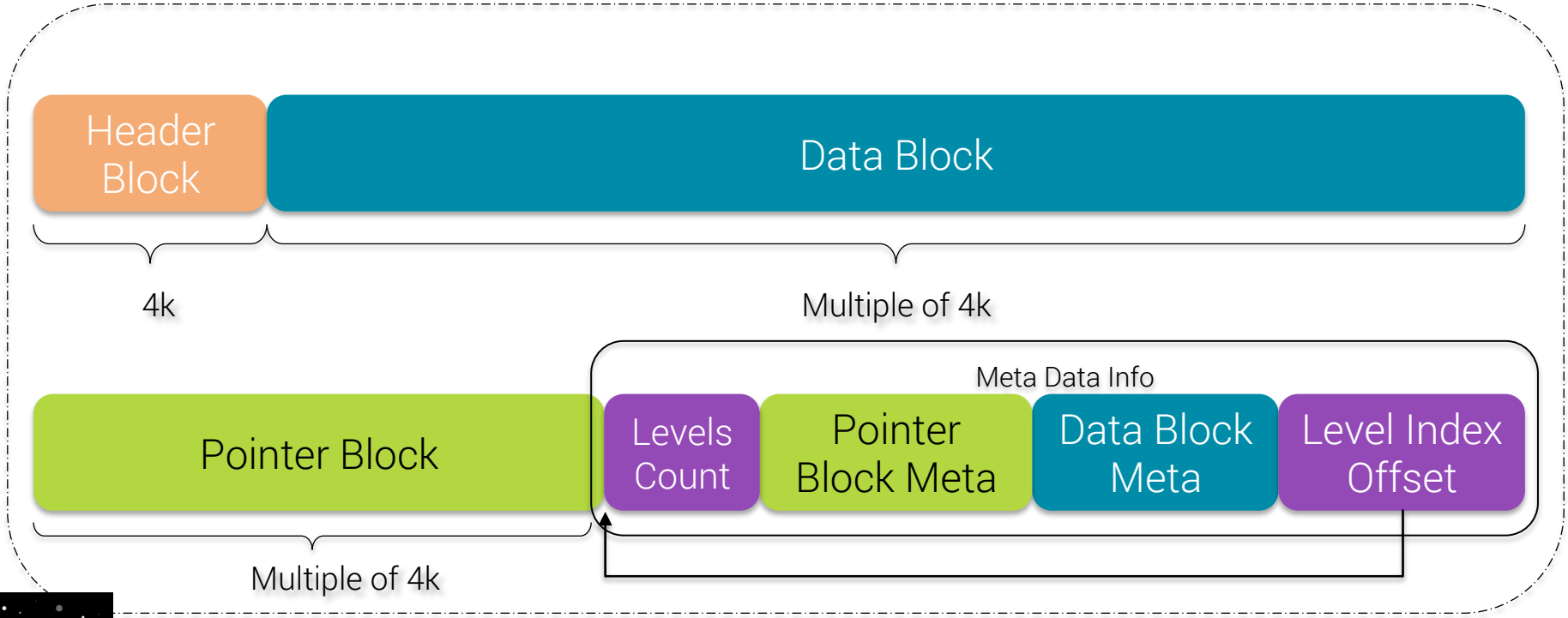
- first, optimize query using Query Planer (see later)
- then load chunks (4k) of index files from disk into memory
- perform binary search to find the indexed value(s)
- retrieve the corresponding **partition keys** and push them into the **Partition Key Cache**

→ Yes, **currently** SASI only keep partition key(s) so on wide partition it's not very optimized ...

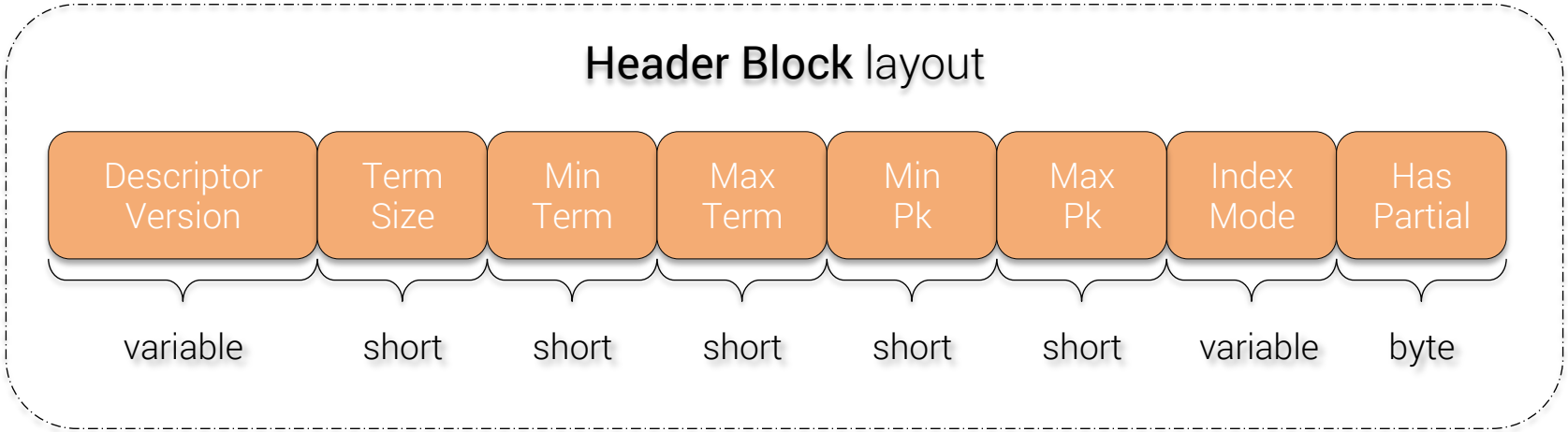
OnDiskIndex files



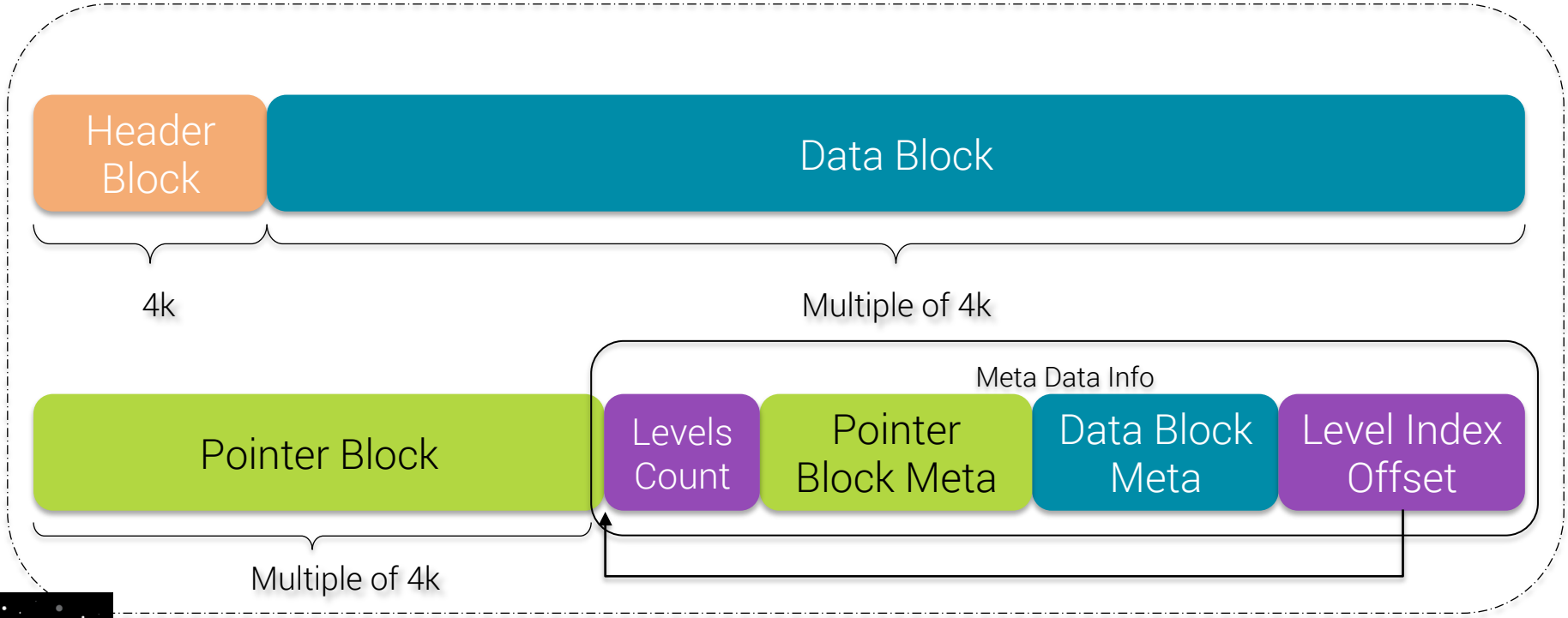
OnDiskIndex Layout



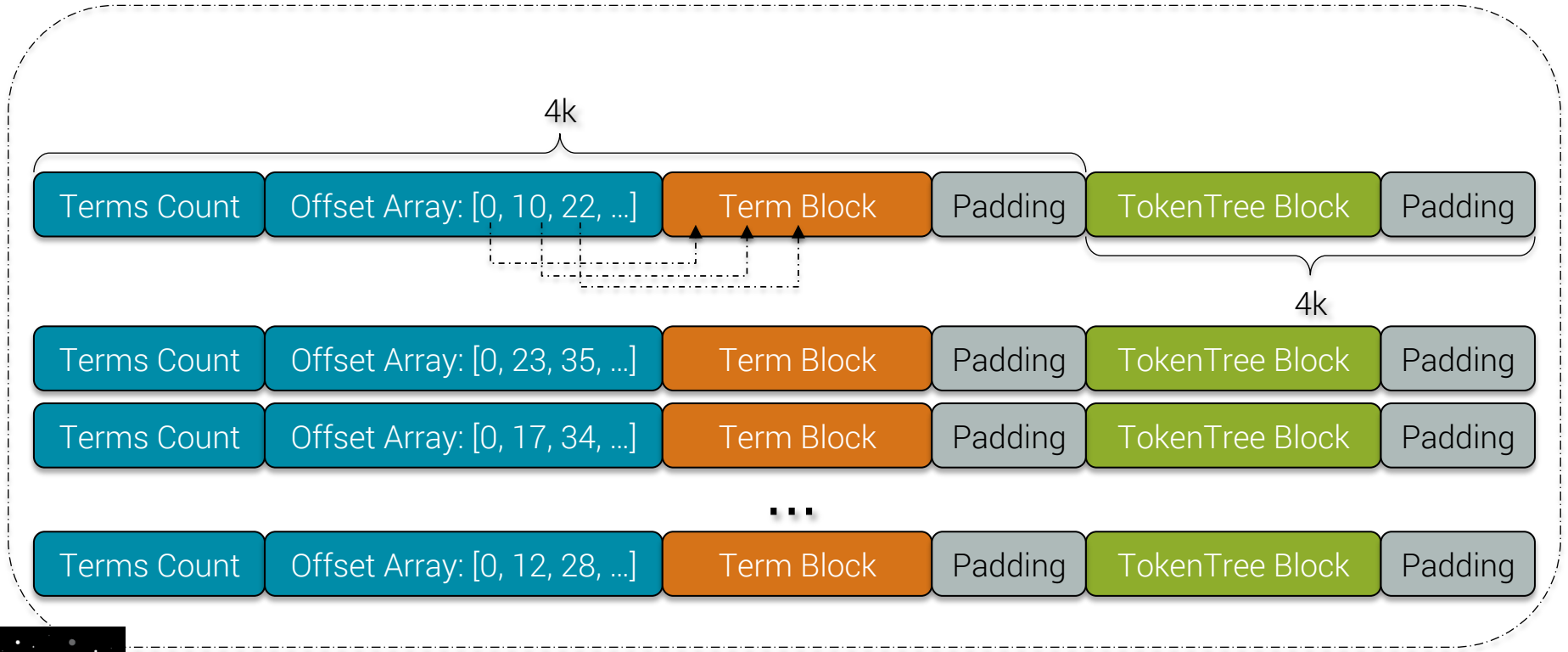
Header Block Layout



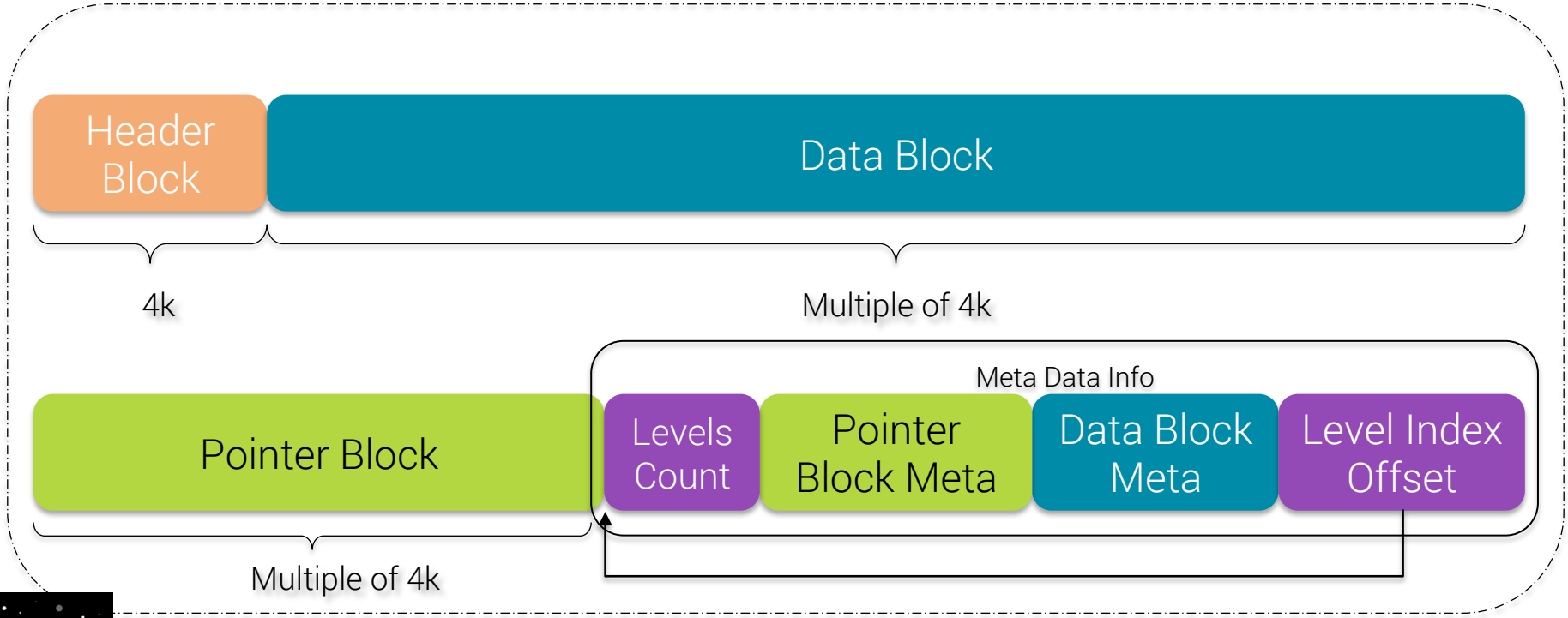
OnDiskIndex Layout



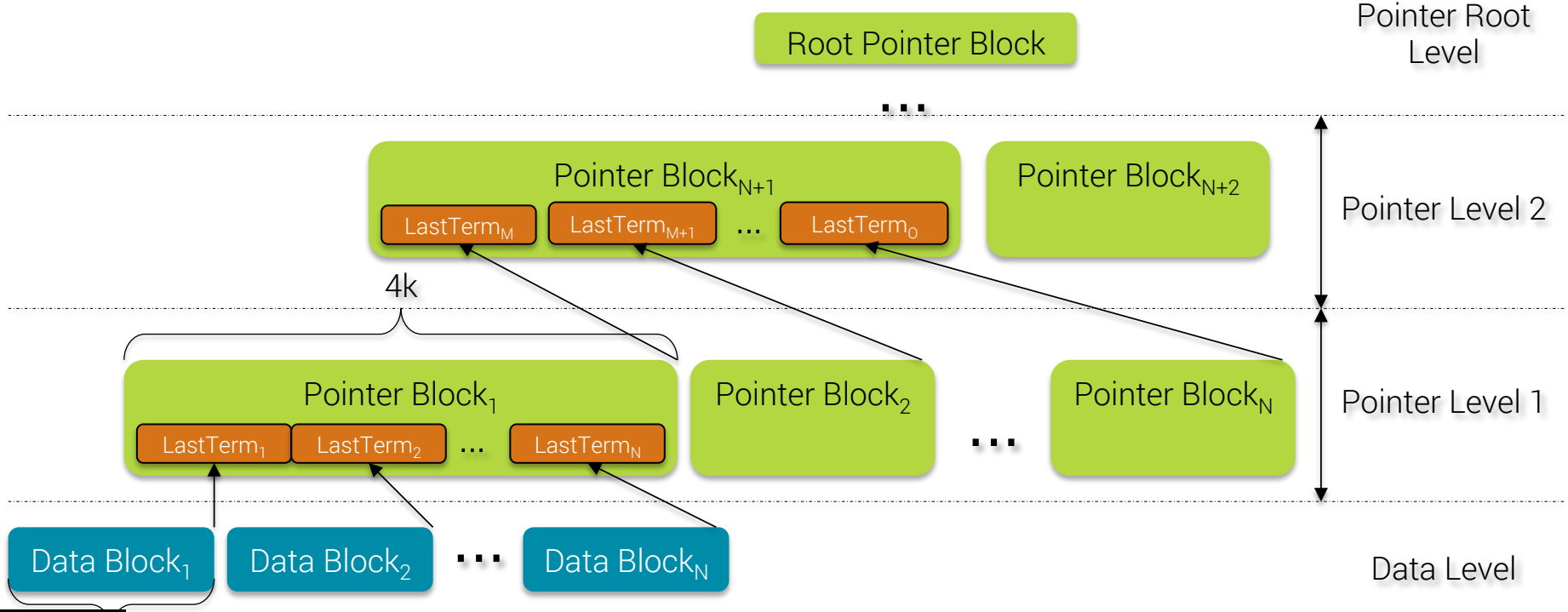
Data Block layout



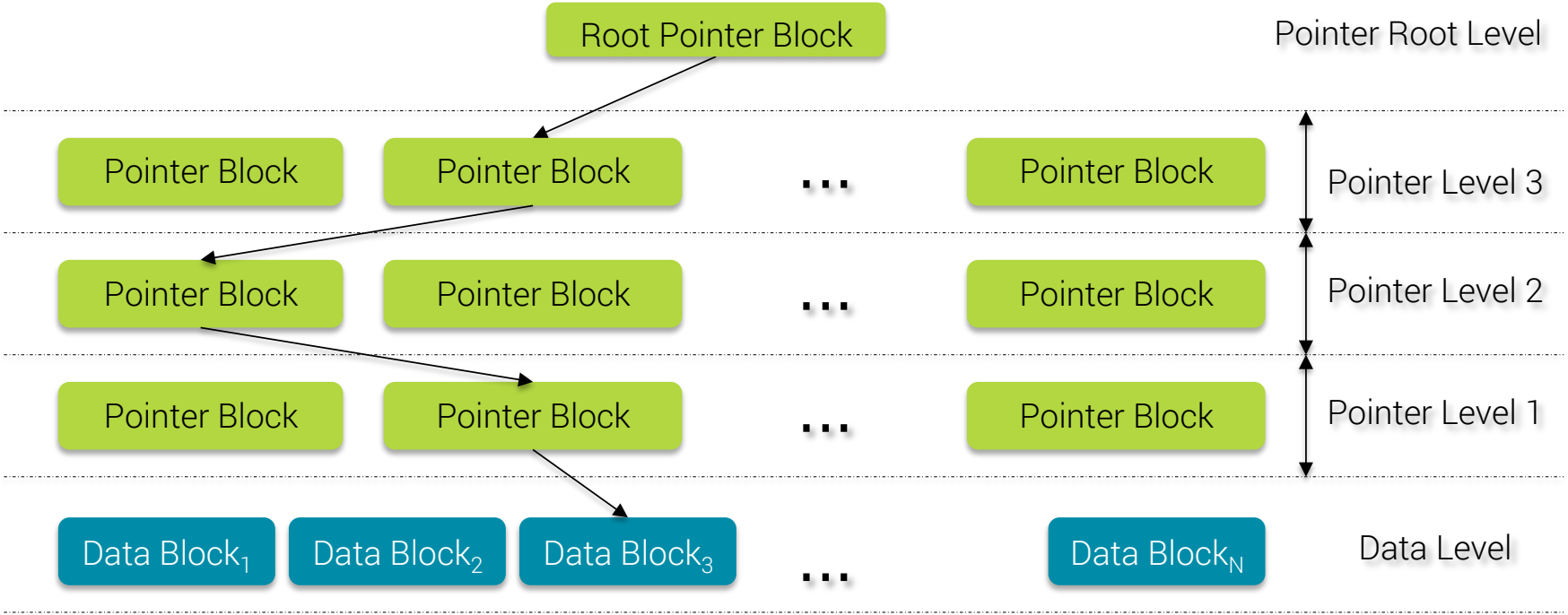
OnDiskIndex Layout



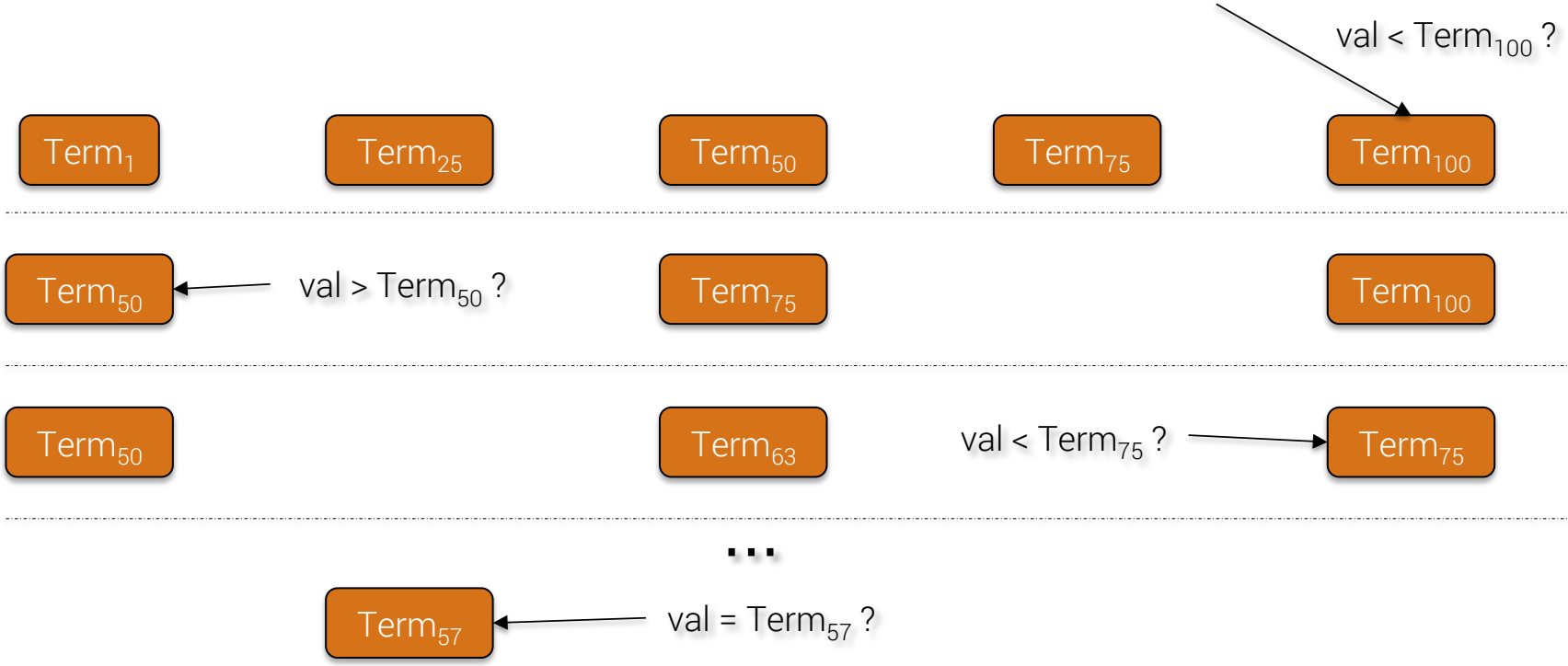
Pointer Block building



Binary search using OnDiskIndex files



Term Block Binary Search



Query Planner

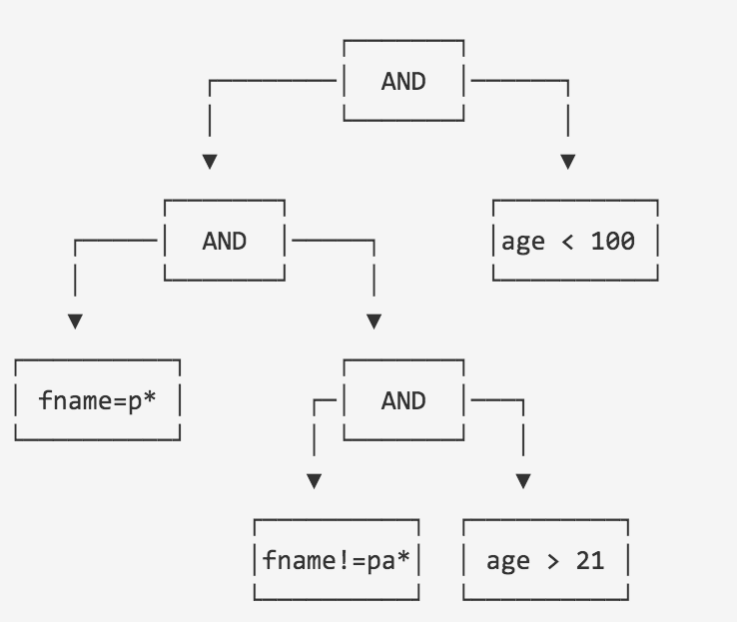
@doanduyhai

Query planner

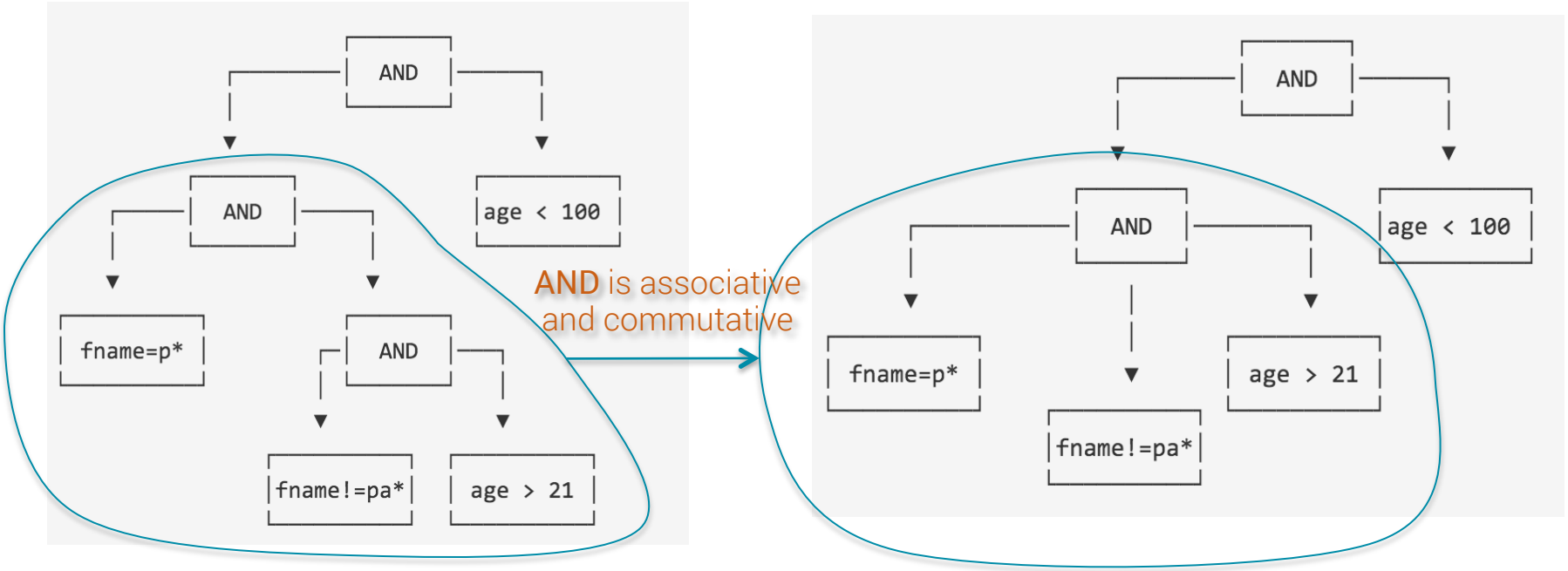
- build predicates tree
- predicates push-down & re-ordering
- predicate fusions for != operator

Query optimization example

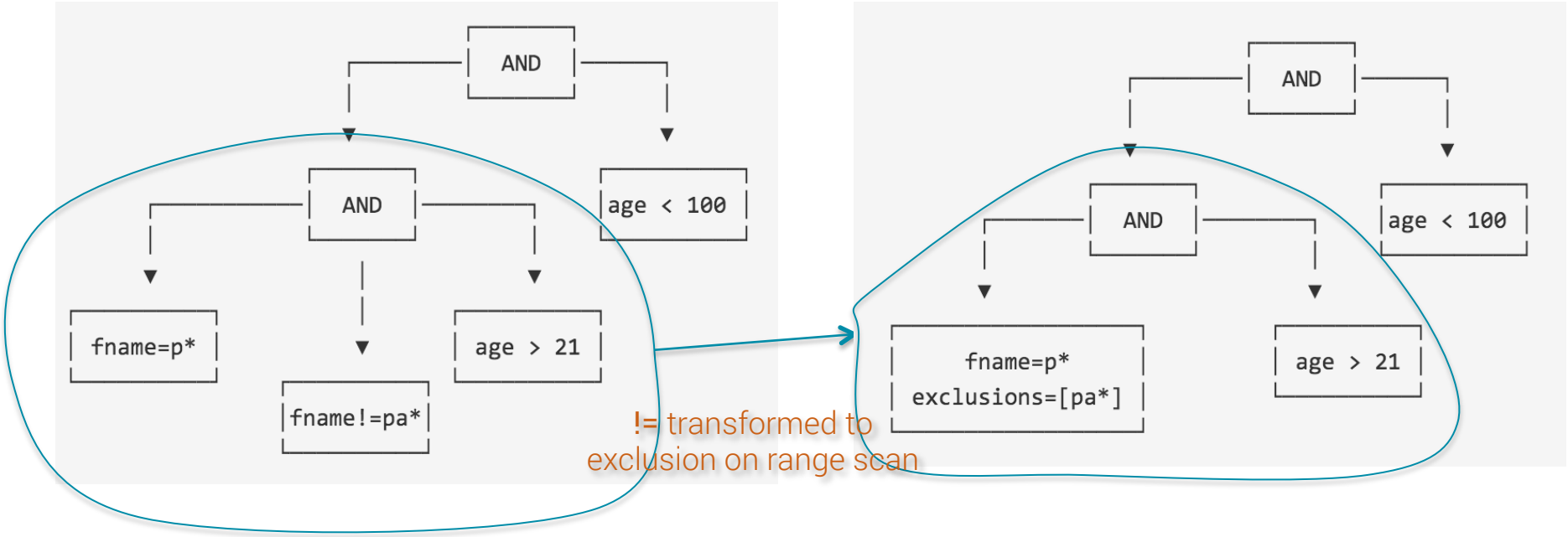
WHERE age < 100 AND fname LIKE 'p%' AND fname != 'pa%' AND age > 21



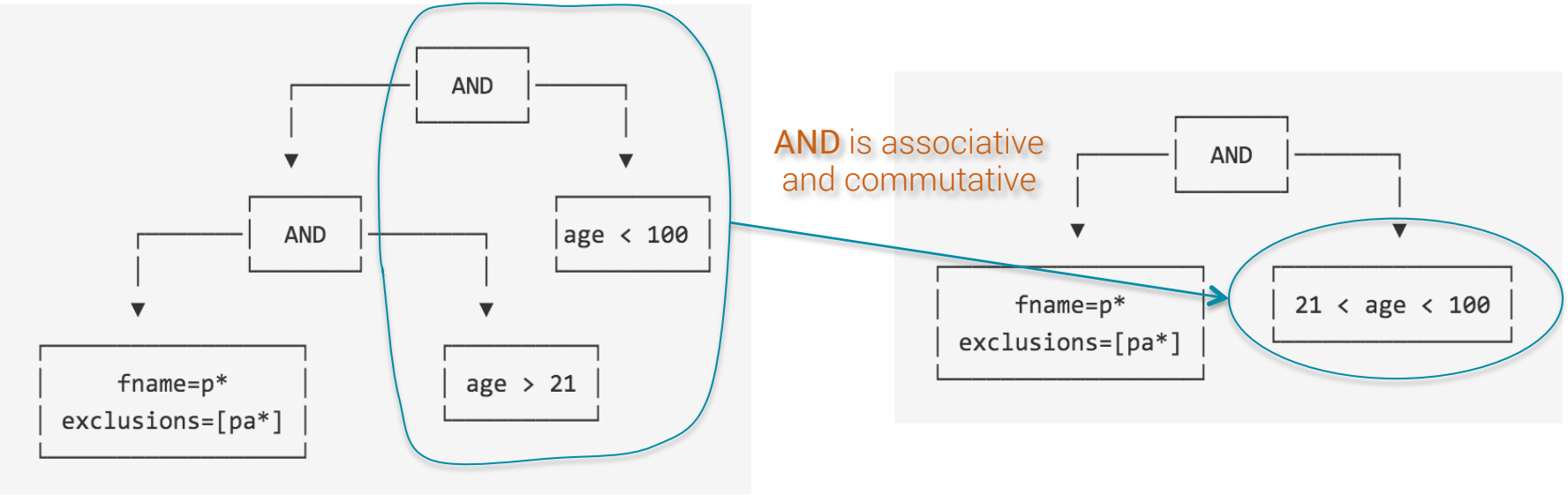
Query optimization example



Query optimization example



Query optimization example



Some benchmarks

Hardware specs

13 bare-metal machines

- 6 CPU HT (12 vcores)
- 64Gb RAM
- 4 SSDs in RAID0 for a total of 1.5Tb

Data set

- 13 billions of rows
- 1 numerical index with 36 distinct values
- 2 text index with 7 distinct values
- 1 text index with 3 distinct values

Benchmark results

Full table scan using co-located Spark (no LIMIT)

Predicate count	Fetched rows	Query time in sec
1	36 109 986	609
2	2 781 492	330
3	1 044 547	372
4	360 334	116

Benchmark results

Full table scan using co-located Spark (no LIMIT)

Predicate count	Fetched rows	Query time in sec
1	36 109 986	609
2	2 781 492	330
3	1 044 547	372
4	360 334	116

Benchmark results

Beware of disk space usage for full text search !!!

Table *albums* with ≈ 110 000 records, 6.8Mb data size

Index Name	Index Mode	Analyzer	Index Size	Index Size/SSTable Size Ratio
albums_country_idx	**PREFIX**	**`NonTokenizingAnalyzer`**	2Mb	0.29
albums_year_idx	**PREFIX**	N/A	2.3Mb	0.34
albums_artist_idx	**CONTAINS**	**`NonTokenizingAnalyzer`**	**30Mb**	**4.41**
albums_title_idx	**CONTAINS**	**`StandardAnalyzer`**	**41Mb**	**6.03**

Take Away

SASI vs search engines

SASI vs Solr/ElasticSearch ?

- Cassandra is not a search engine !!! (database = durability)
- always slower because 2 passes (SASI index read + original Cassandra data)
- no scoring
- no ordering (~~ORDER BY~~)
- no grouping (~~GROUP BY~~) → Apache Spark for analytics

If you don't need the above features, SASI is for you!

SASI sweet spots

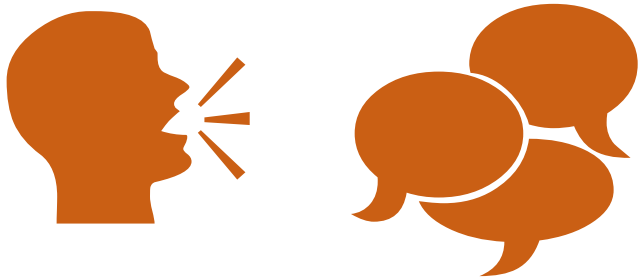
SASI is a relevant choice if

- you need **multi criteria search** and you don't need ordering/grouping/scoring
- you mostly need **100 to 10000** of rows for your search queries
- you **always know the partition keys** of the rows to be searched for (this one applies to native secondary index too)
- you want to index **static columns** (**SASI** has no penalty since it indexes the whole partition)

SASI blind spots

SASI is a poor choice if

- you have strong SLA on search latency, for example few millisecs requirement
- ordering of the search results is important for you



Q & A

Thank You

@doanduyhai

duy_hai.doan@datastax.com