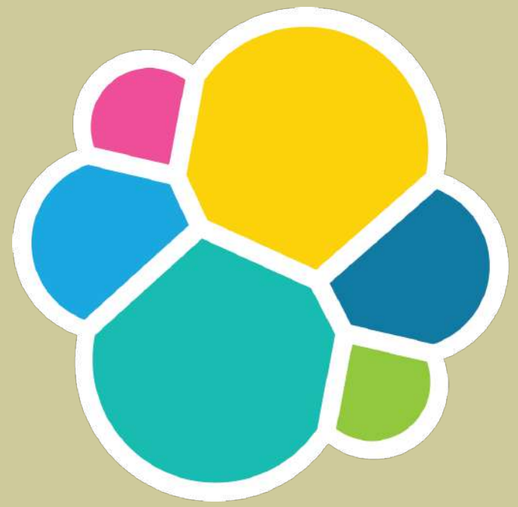
Elasticsearch

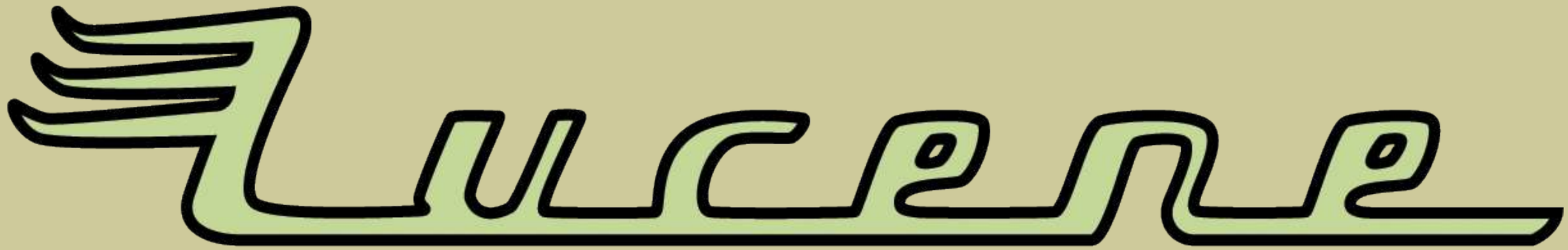
Under the Hood

Philipp Krenn @xeraa



elastic

Developer 🥑



elasticsearch

A Console Request

```
POST databases/_doc
{
  "name": "Elasticsearch",
  "author": "Shay Banon",
  "stable_version": "7.13.2"
}
```

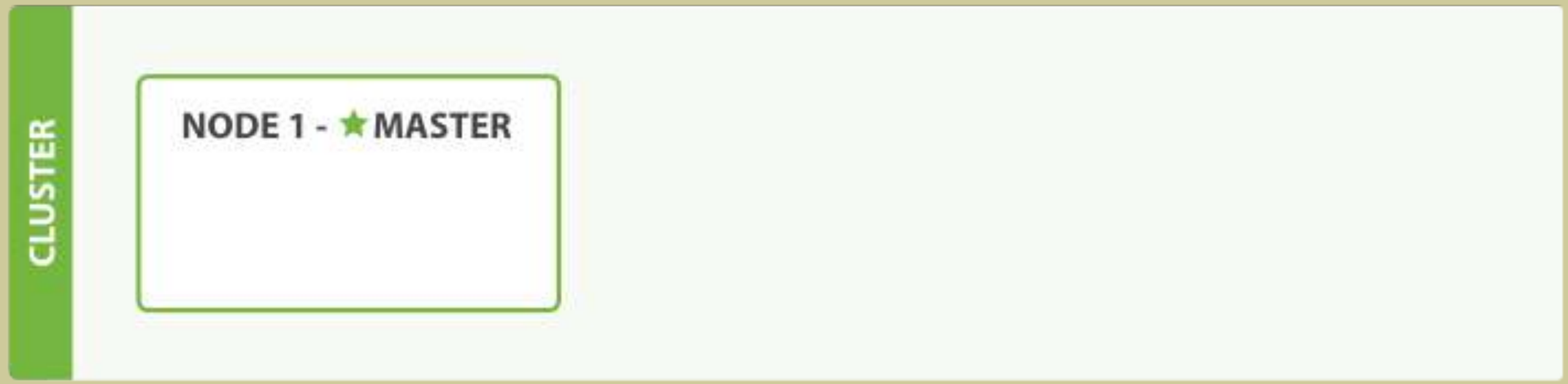
A Console Response


```
{
  "_index" : "databases",
  "_type" : "_doc",
  "_id" : "MmsCOHoBA2wTstDEjWFq",
  "_version" : 1,
  "result" : "created",
  "_shards" : {
    "total" : 2,
    "successful" : 2,
    "failed" : 0
  },
  "_seq_no" : 0,
  "_primary_term" : 1
}
```

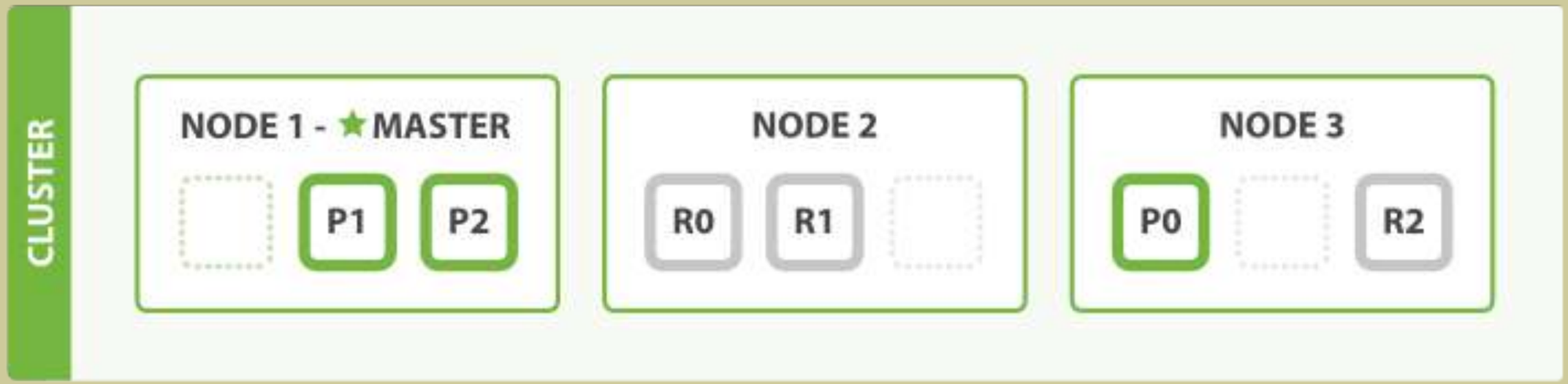

INITECH

Single node

```
$ cd elasticsearch-<version>
$ ./bin/elasticsearch
```


CLUSTER
NODE 1 - ★MASTER

Cluster


CLUSTER
NODE 1 - ★MASTER
P1
P2
NODE 2
R0
R1
NODE 3
P0
R2

Node Types

Master-Eligible Node

by default

Data Node

by default

Ingest / Transform Node

by default

Coordinating-Only Node

if no other role

Full Node List

c (cold), d (data), f (frozen), h (hot), i (ingest), l (machine learning), m (master-eligible), r (remote cluster client), s (content), t (transform), v (voting-only), w (warm), and - (coordinating-only)

Example

```
GET _cat/nodes?v

ip             heap.percent ram.percent cpu load_1m load_5m load_15m node.role master name
172.27.14.104            30          99  15    6.80    7.82     7.95 himrst    *      instance-0000000001
172.27.67.193            25          99  35   12.05   12.46    11.94 himrst    -      instance-0000000000
172.27.135.118           76          74   8    3.49    3.28     3.49 mv         -      tiebreaker-0000000002
```

Discovery

Cluster Coordination?

Cluster State?

Cluster Metadata

Cluster Settings

Index Metadata

Lots more

```
GET _cluster/state
```

Only move forward

Do not lose data

```
{
  "cluster_name" : "9febb36ac54b45059344df37b985d7e7",
  "cluster_uuid" : "WC_5Z-orS6qocUEWxMdReQ",
  "version" : 13007,
  "state_uuid" : "EorsIIRvT7uRM_N-QSzI_A",
  "master_node" : "_9mvVnPIQlSWbbXk6GMCdg",
  "blocks" : { },
  "nodes" : {
    "0ZvrOgwrS9Gsv34tZckwmQ" : {
      "name" : "tiebreaker-0000000002",
      "ephemeral_id" : "iomBboAxQ8uBnzgOvq5CCg",
      "transport_address" : "172.27.135.118:19885",
      "attributes" : {
        "logical_availability_zone" : "tiebreaker",
        "server_name" : "tiebreaker-0000000002.9febb36ac54b45059344df37b985d7e7",
        "availability_zone" : "eu-west-1c",
        ...
```

Main Components

Discovery

Master Election

Cluster State Publication

Zen

Zen to Zen2

Not pluggable


elastic
@xeraa

Why

https://www.elastic.co/guide/en/elasticsearch/resiliency/current/index.html

Repeated network partitions can cause cluster state updates to be lost (STATUS: DONE, v7.0.0)

And more

How

https://github.com/elastic/elasticsearch-formal-models

TLA+ specification

TLC model checking

https://github.com/elastic/elasticsearch-formal-models/blob/master/cluster/isabelle/Preliminaries.thy

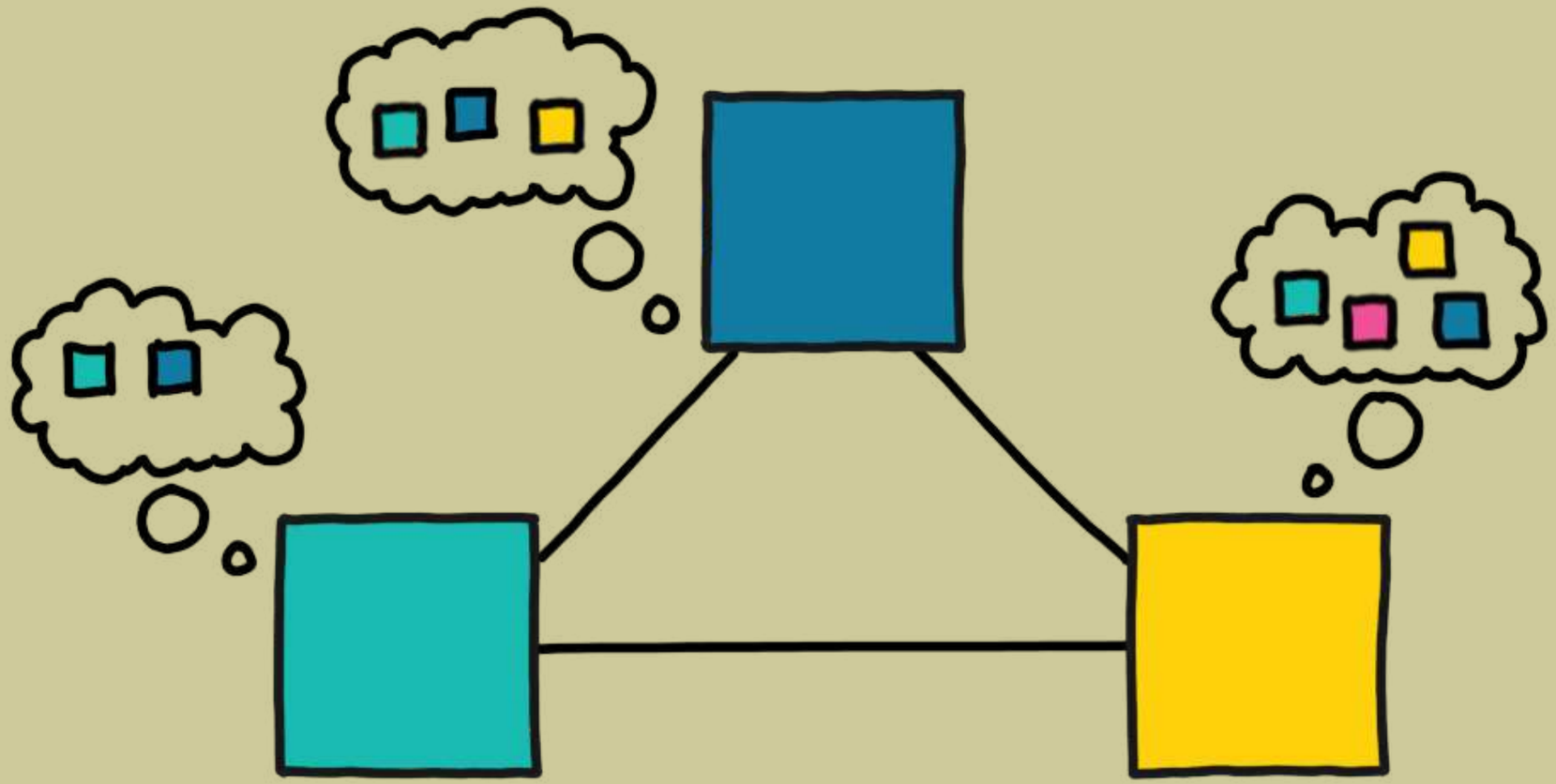
```
text \<open>It works correctly on finite and nonempty sets as follows:\<close>

theorem
  fixes S :: "Term set"
  assumes finite: "finite S"
  shows maxTerm_mem: "S \<noteq> {} \<Longrightarrow> maxTerm S \<in> S"
    and maxTerm_max: "\<And> t'. t' \<in> S \<Longrightarrow> t' \<le> maxTerm S"
proof -
  presume "S \<noteq> {}"
  with assms
  obtain t where t: "t \<in> S" "\<And> t'. t' \<in> S \<Longrightarrow> t' \<le> t"
  proof (induct arbitrary: thesis)
    case empty
    then show ?case by simp
    ...
```

Discovery

Where are master-eligible nodes?

Is there a master already?



Settings

`discovery.zen.ping.unicast.hosts` → `discovery.seed_hosts` **static**

`discovery.zen.hosts_provider` → `discovery.seed_providers` **dynamic (file, EC2, GCE,...)**

Master Election

Agree which node should be master

Form a cluster


FOLLOW THE LEADER

discovery.zen.minimum_master_nodes

Trust users?

Scaling up or down?

Three Node Cluster



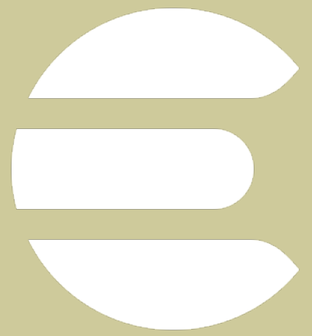
discovery.zen.minimum_master_nodes: ~

```
discovery.zen.minimum_master_nodes: 2
```



```
discovery.zen.minimum_master_nodes: 2
```

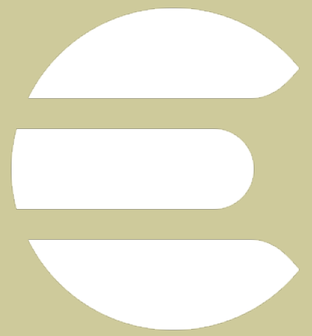




discovery.zen.minimum_master_nodes: 2

discovery.zen.minimum_master_nodes: 2





discovery.zen.minimum_master_nodes: 2

```
discovery.zen.minimum_master_nodes: 2
```

`cluster.initial_master_nodes`

List of node names for the very first election

OK

to set on multiple nodes as long as they are all consistent

Ignored

once node has joined a cluster even if restarted

Unnecessary

when joining new node to existing cluster

Cluster Scaling

Master-ineligible: as before

Adding master-eligible: just do it

Removing master-eligible: just do it

As long as you remove less than half of them at once

Scale down to a single node

```
POST /_cluster/voting_config_exclusions/
elasticsearch1

POST /_cluster/voting_config_exclusions/
elasticsearch2
```

Cluster Rebuild

Empty `cluster.initial_master_nodes`

Log

```
elasticsearch2    | {"type": "server",
                     "timestamp": "2019-05-24T14:02:51,173+0000",
                     "level": "WARN",
                     "component": "o.e.c.c.ClusterFormationFailureHelper",
                     "cluster.name": "docker-cluster",
                     "node.name": "elasticsearch2",
                     "message":
```

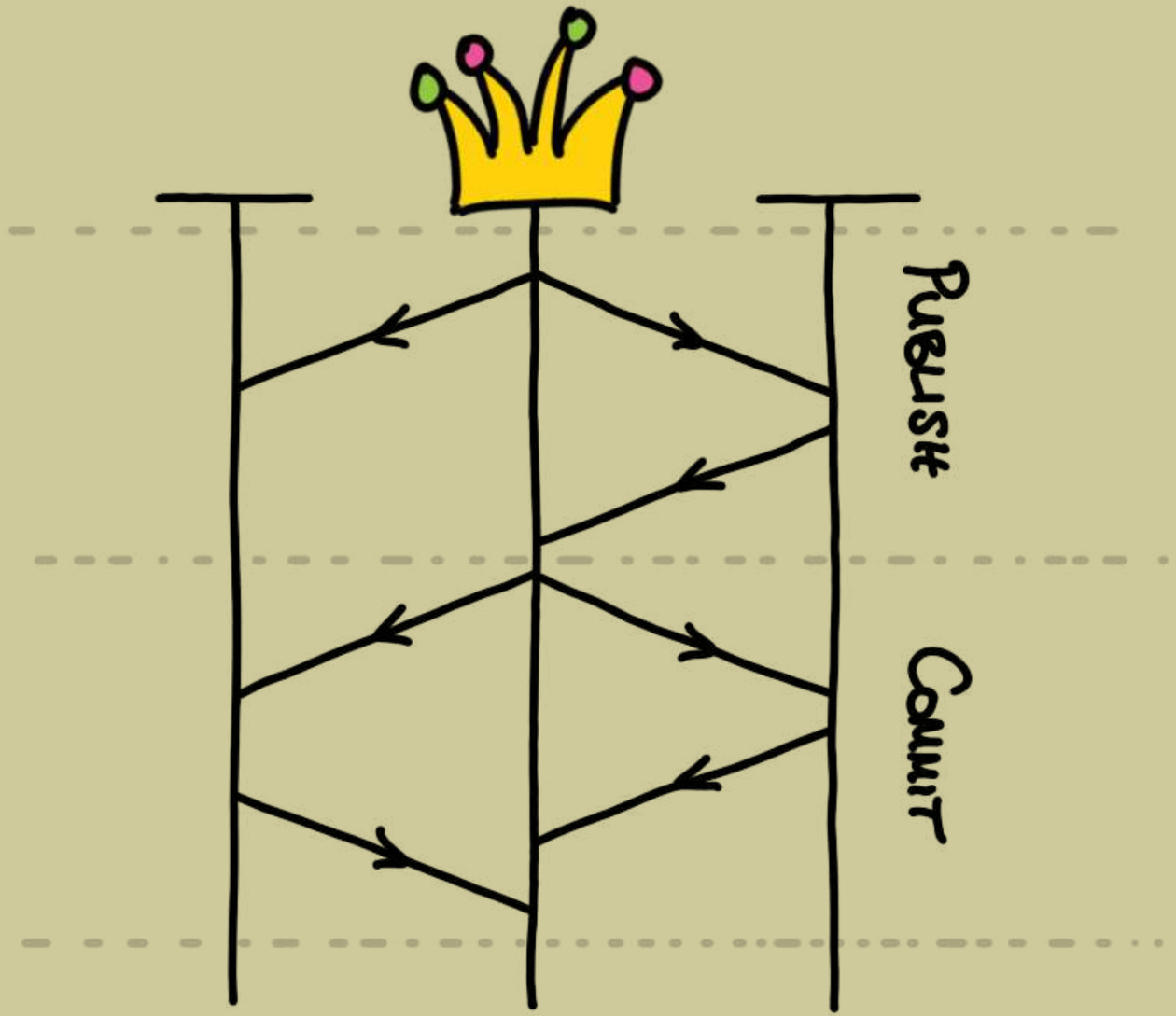
```
"master not discovered yet,
this node has not previously joined a bootstrapped (v7+) cluster,
and [cluster.initial_master_nodes] is empty on this node:
have discovered [
    {elasticsearch1}{pSUJ60tSRWSrcWkRevLfyA}{_jIaabgyTQOHAOjcwUruIQ}
        {192.168.112.3}{192.168.112.3:9300}
        {...},
    {elasticsearch3}{ngaTCze8QHSHydCXsttXyw}{mbIad-A4SLOJvP7Ava5dEw}
        {192.168.112.4}{192.168.112.4:9300}
        {...}
];
```

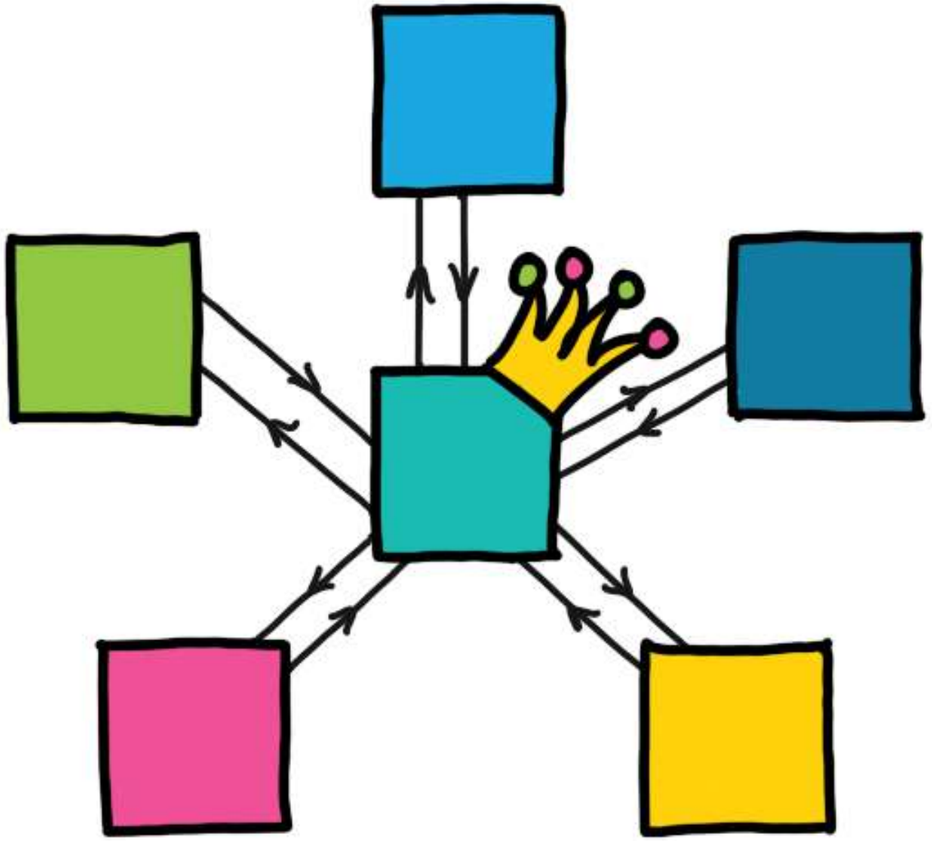
```
discovery will continue using
        [192.168.112.3:9300, 192.168.112.4:9300]
        from hosts providers and [
    {elasticsearch2}{iANt64LESxqjJv8tHV5KKw}{K0bYEuQ2TnamsiOefTUXgQ}
        {192.168.112.2}{192.168.112.2:9300}
        {...}
]
from last-known cluster state;
node term 0, last-accepted version 0 in term 0"
```

Cluster State Publication

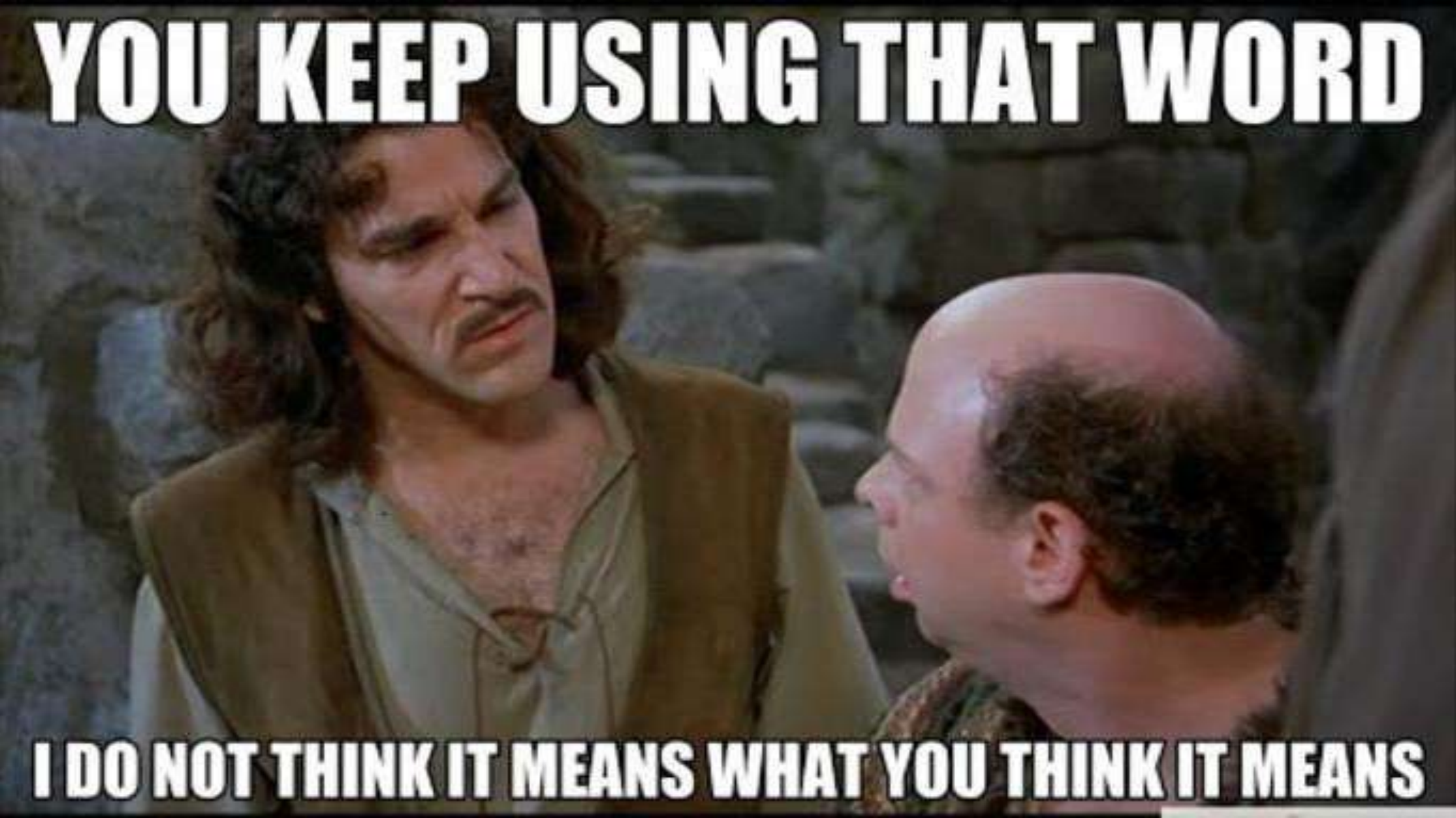
Agree on cluster state updates

Broadcast updates to all nodes


PUBLISH
COMMIT



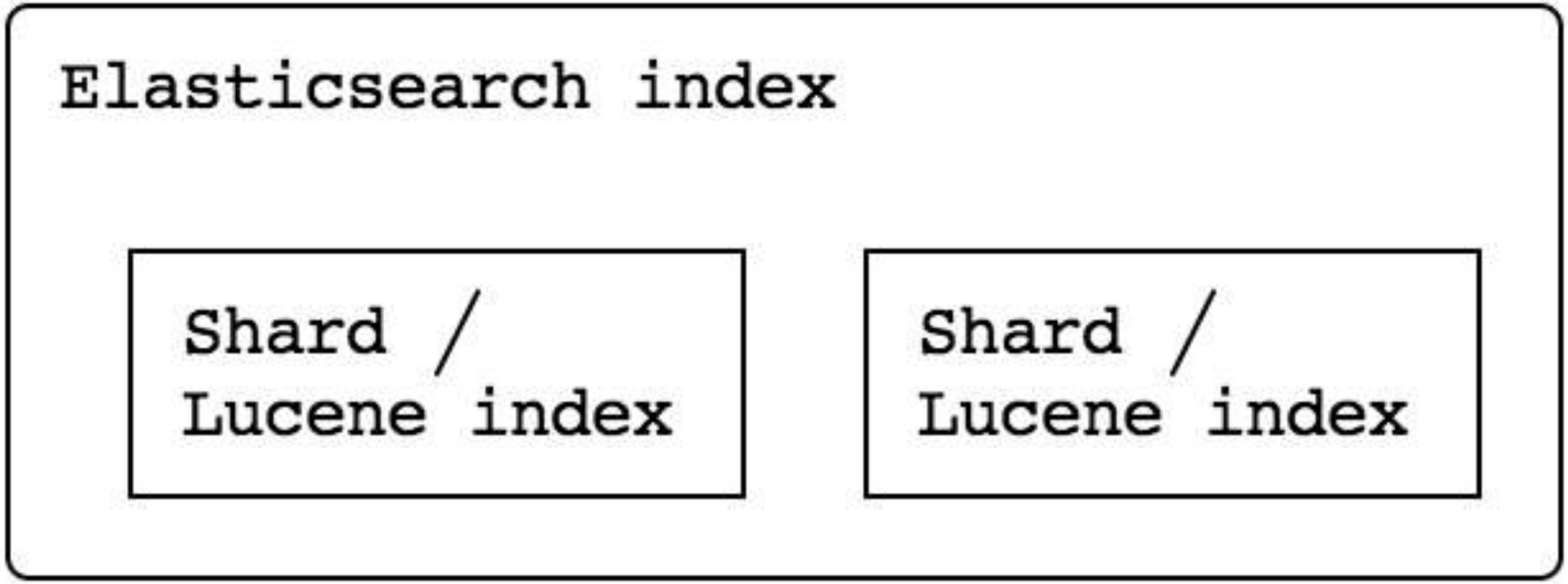
Indexing a Document


YOU KEEP USING THAT WORD
I DO NOT THINK IT MEANS WHAT YOU THINK IT MEANS

Elasticsearch index

Lucene index

To index


Elasticsearch index
Shard / Lucene index
Shard / Lucene index

Document

Unique combination: `_index` `_id`

PS: Types removed

```
POST /databases/_doc
```

vs

```
PUT /databases/_doc/
elasticsearch
```

Autogenerated ID

20 characters, URL-safe, Base64-encoded, GUID strings

`AVfD6ukyeuK3k1LGtSwT`

Consistent Hashing

Before 2.0: djb2

```c
unsigned long
hash(unsigned char *str)
{
    unsigned long hash = 5381;
    int c;

    while (c = *str++)
        hash = ((hash << 5) + hash) + c; /* hash * 33 + c */

    return hash;
}
```

Consistent Hashing

Current default: murmur3

https://github.com/elastic/elasticsearch/blob/5.6/core/src/main/java/org/elasticsearch/common/hash/MurmurHash3.java

Consistent Hashing

Better distribution

100,000 incremental IDs

https://github.com/elastic/elasticsearch/pull/7954

Consistent Hashing

3 shards

murmur3 [33185, 33347, 33468]

djb2 [30100, 30000, 39900]

Consistent Hashing

5 shards

murmur3 [19933, 19964, 19940, 20030, 20133]

djb2 [20000, 20000, 20000, 20000, 20000]

Consistent Hashing

33 shards

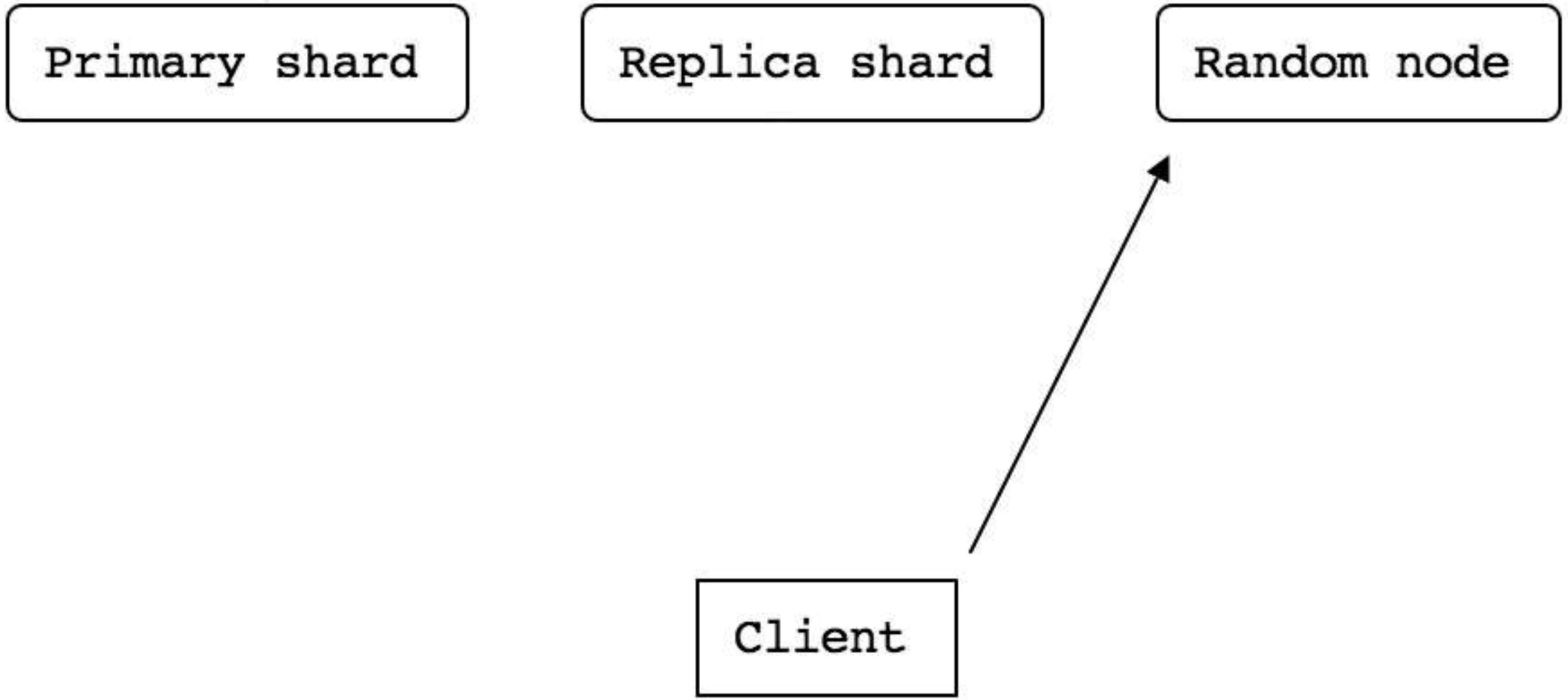
murmur3 [2999, 3096, 2930, 2986, 3070, 3093, 3023, 3052, 3112, 2940, 3036, 2985, 3031, 3048, 3127, 2961, 2901, 3105, 3041, 3130, 3013, 3035, 3031, 3019, 3008, 3022, 3111, 3086, 3016, 2996, 3075, 2945, 2977]

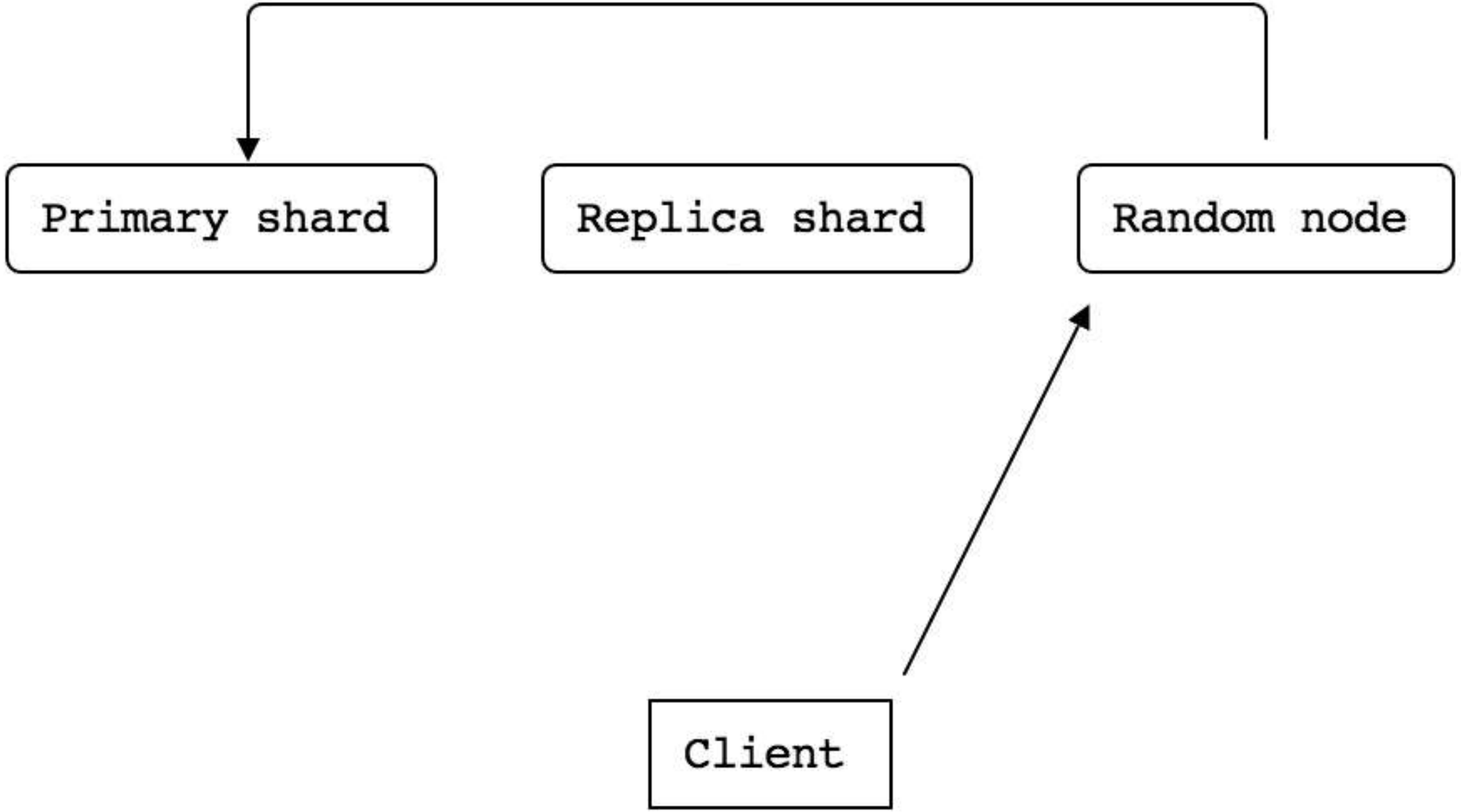
djb2 [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 900, 900, 900, 900, 1000, 1000, 10000, 10000, 10000, 10000, 9100, 9100, 9100, 9100, 9000, 9000, 0, 0, 0, 0, 0, 0]

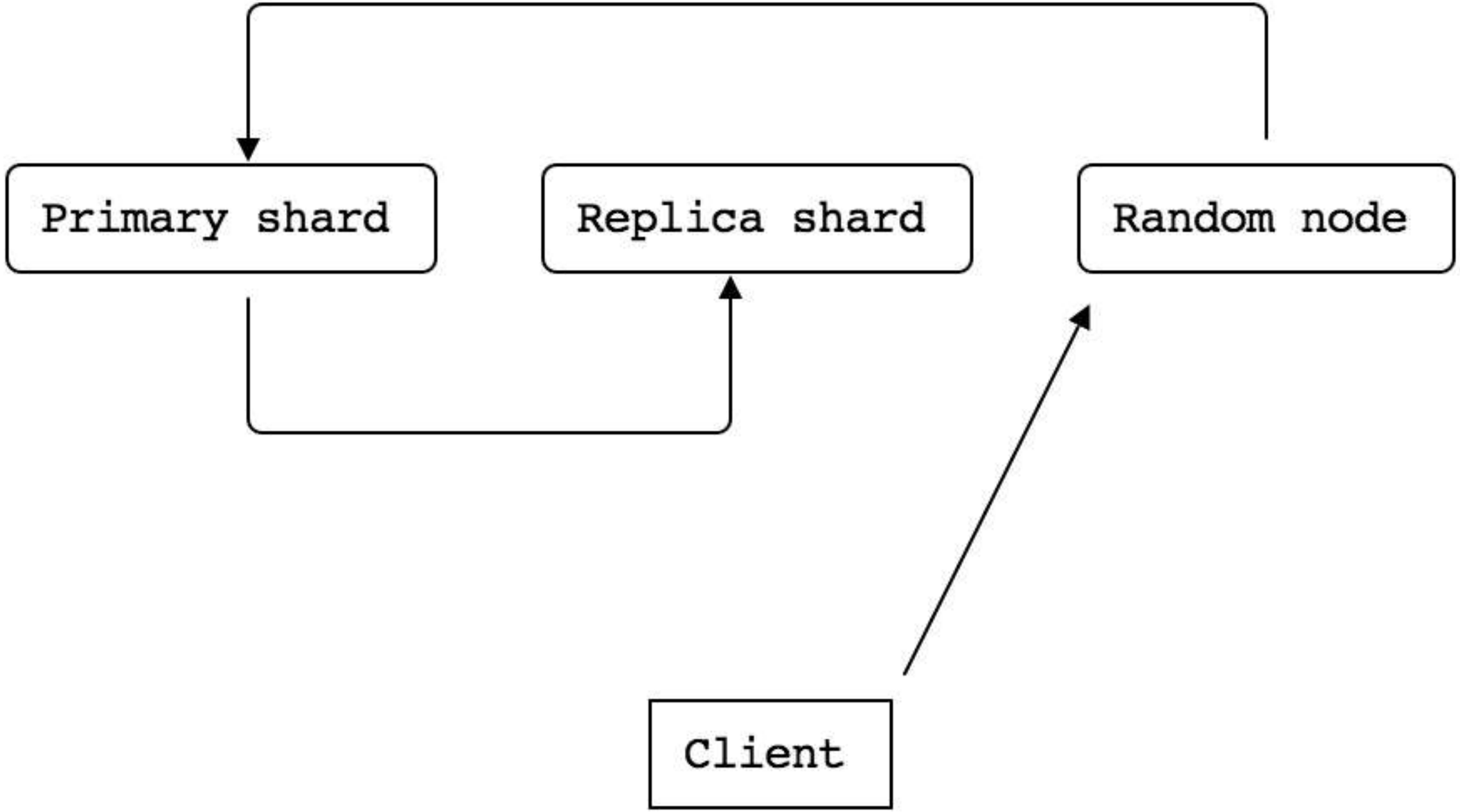
Shard Decision

```
shard = hash(doc_id) %
(num_of_primary_shards)
```

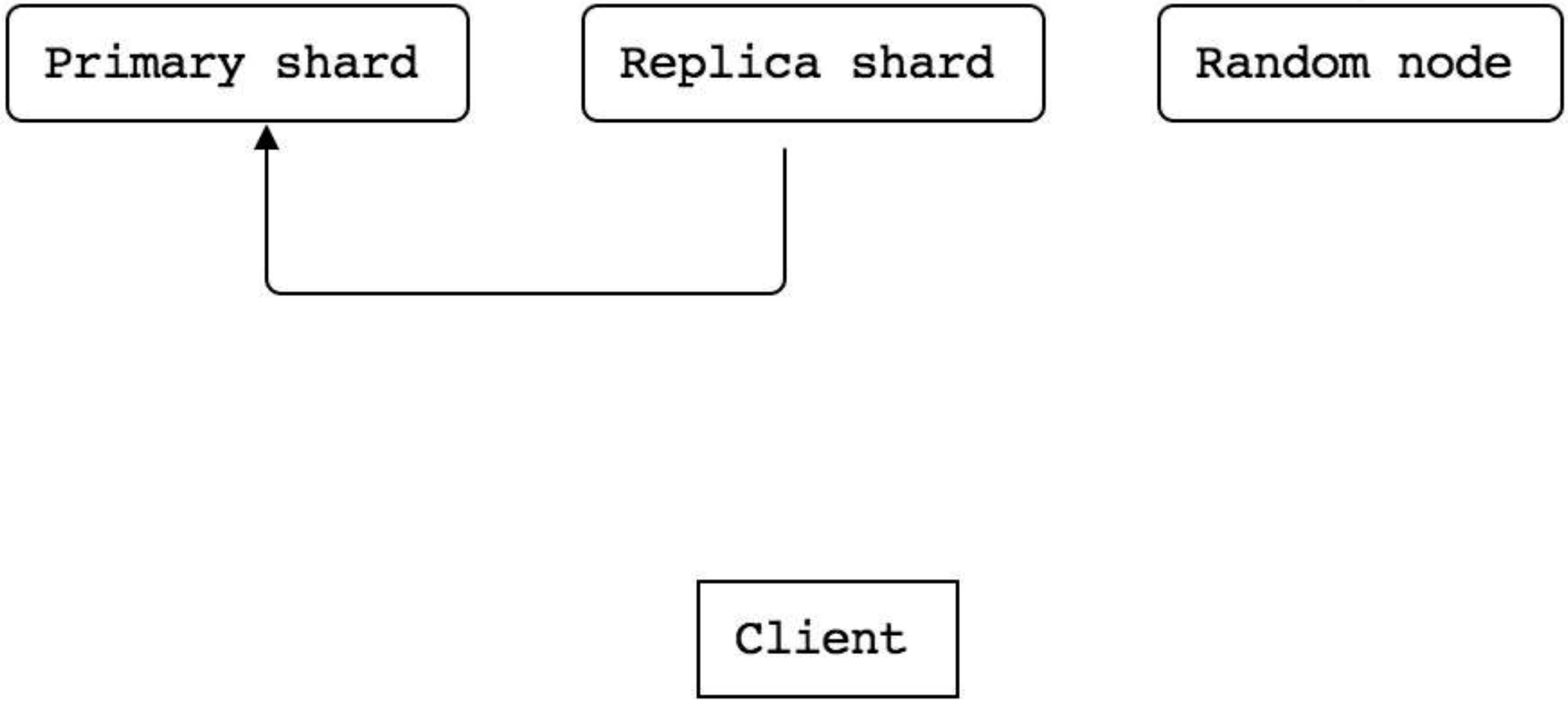
Write

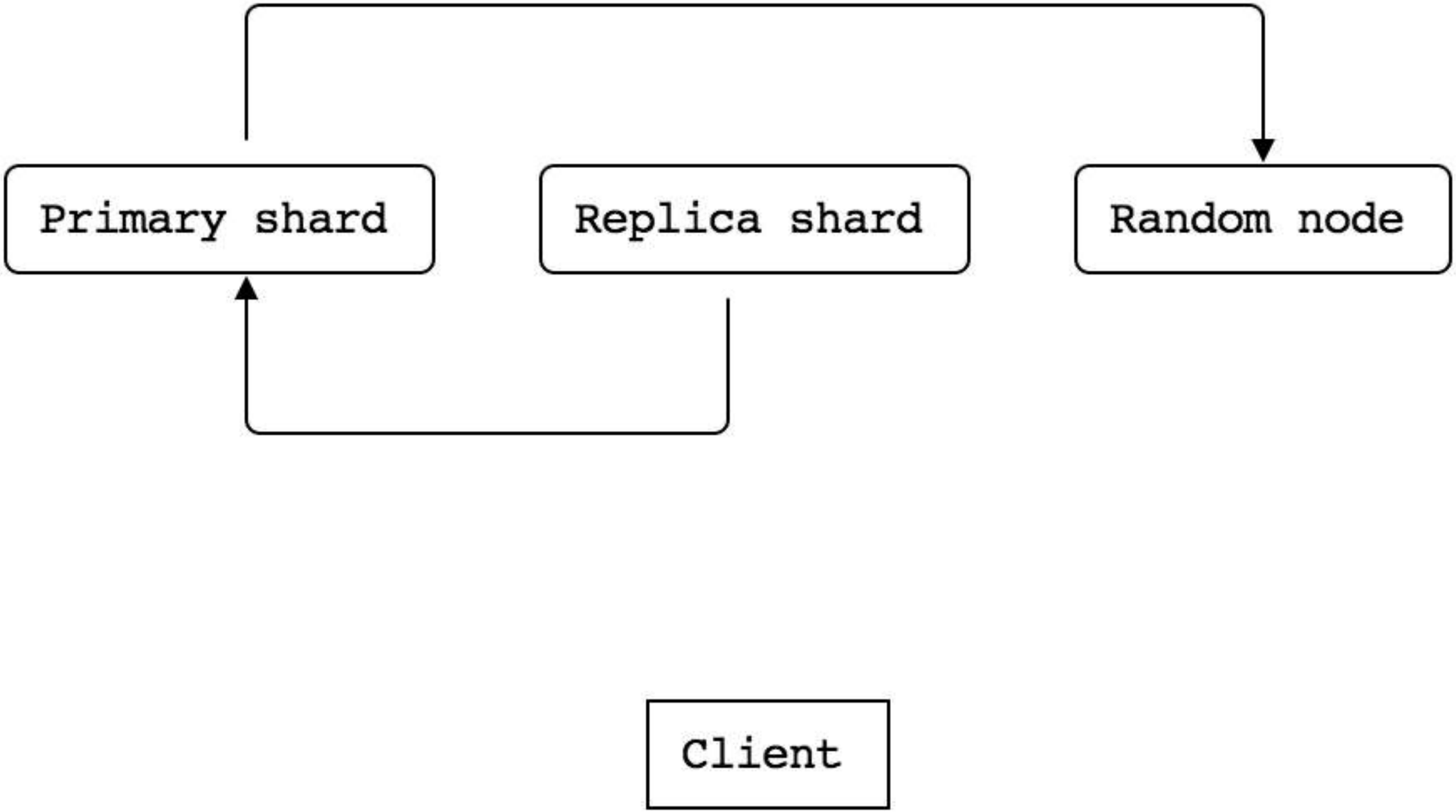

Primary shard
Replica shard
Random node
Client

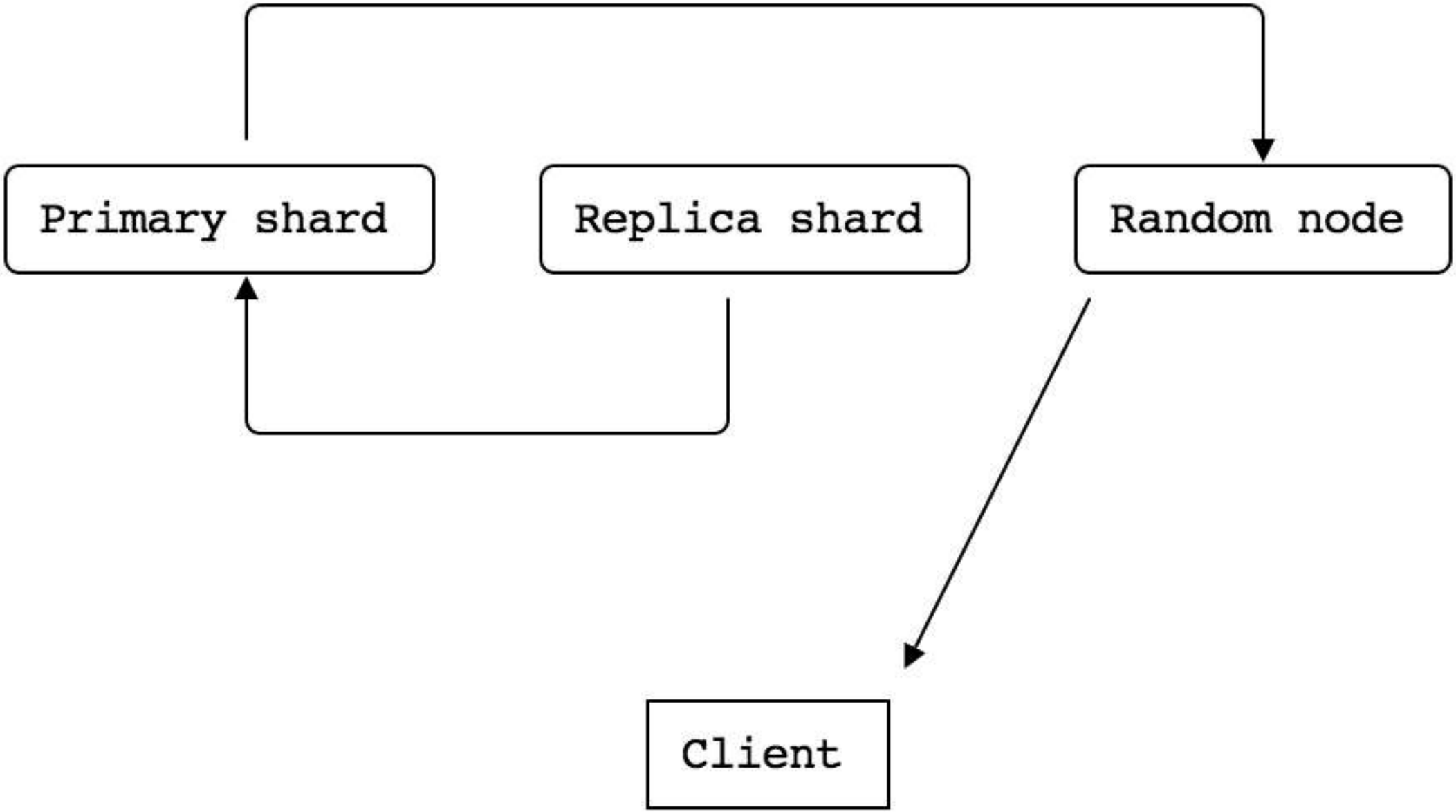

Primary shard
Replica shard
Random node
Client


Primary shard
Replica shard
Random node
Client

Acknowledge


Primary shard
Replica shard
Random node
Client


Primary shard
Replica shard
Random node
Client


Primary shard
Replica shard
Random node
Client

GET /_cluster/health

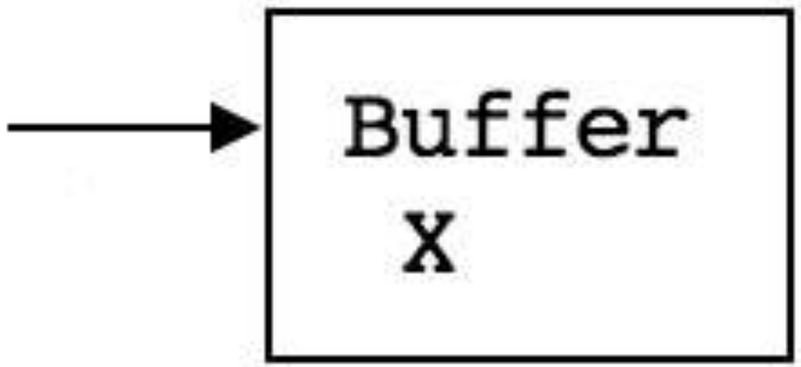
green yellow red

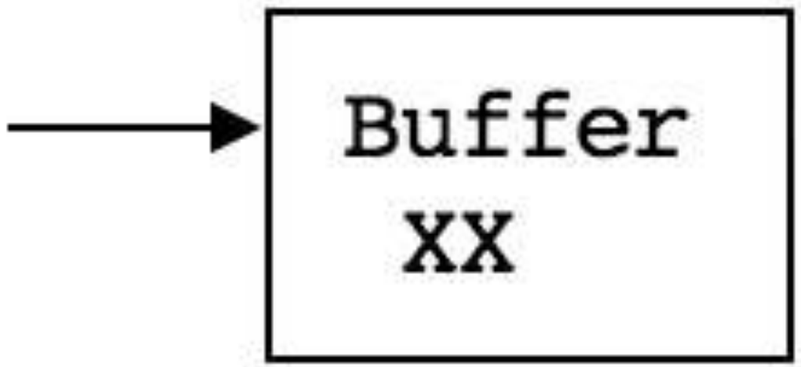
Optimistic Concurrency Control

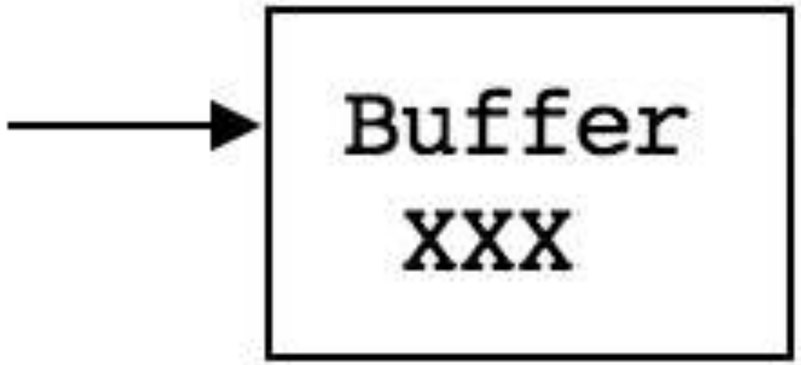
`_version`

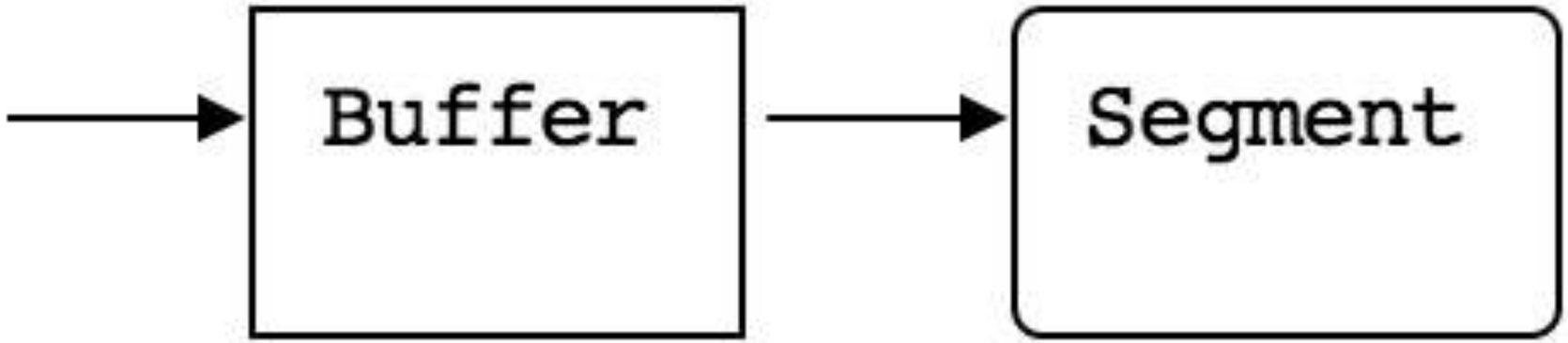
Lucene Segment

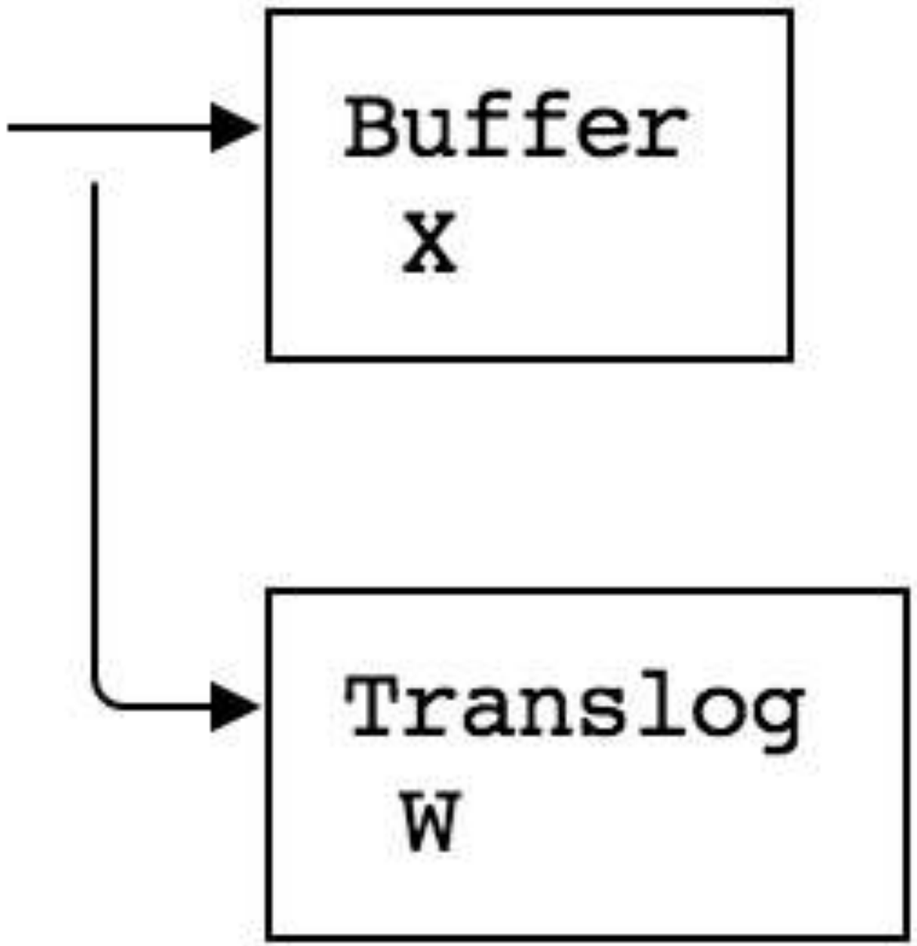
Lucene at Work

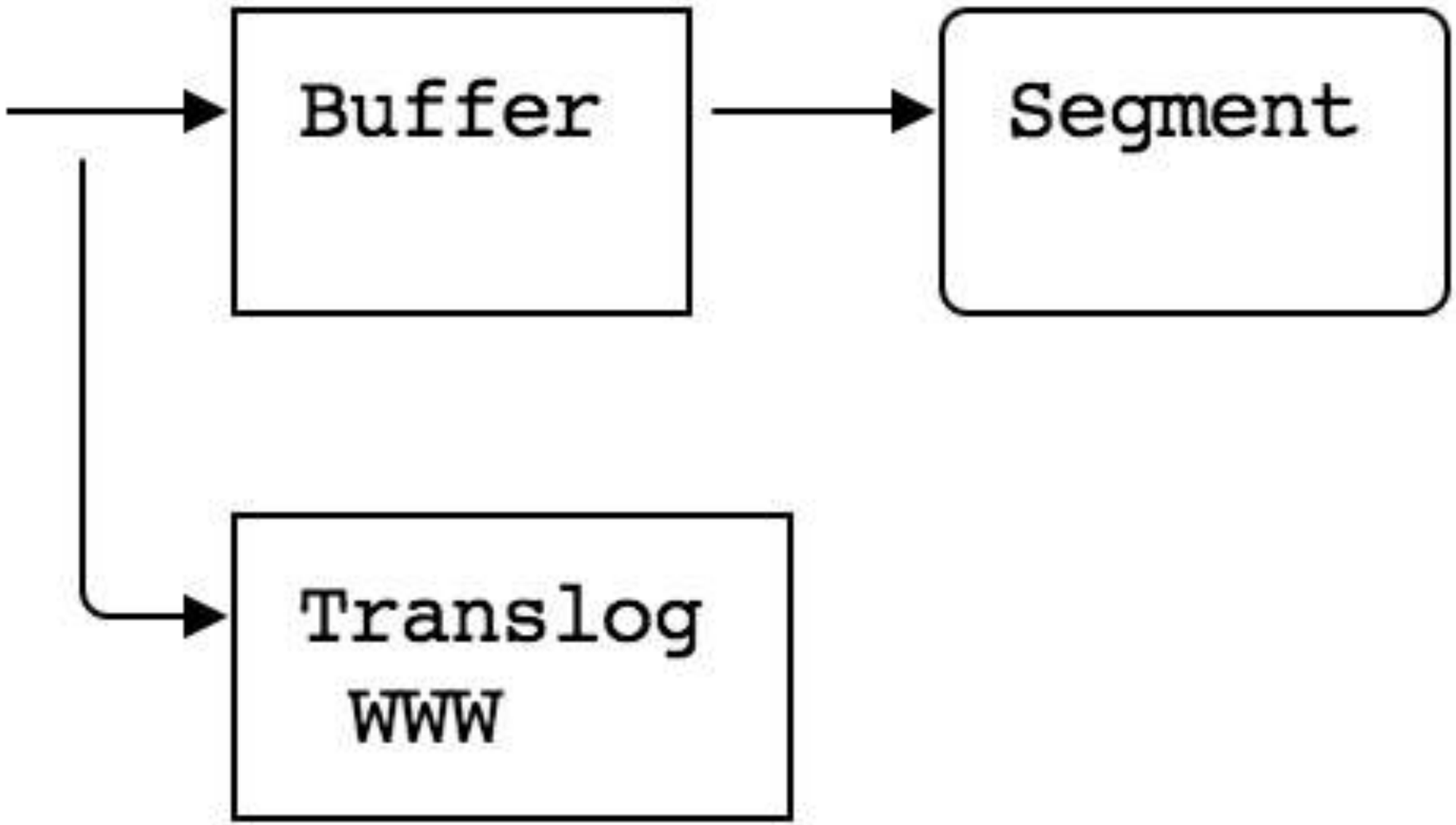

Buffer
X

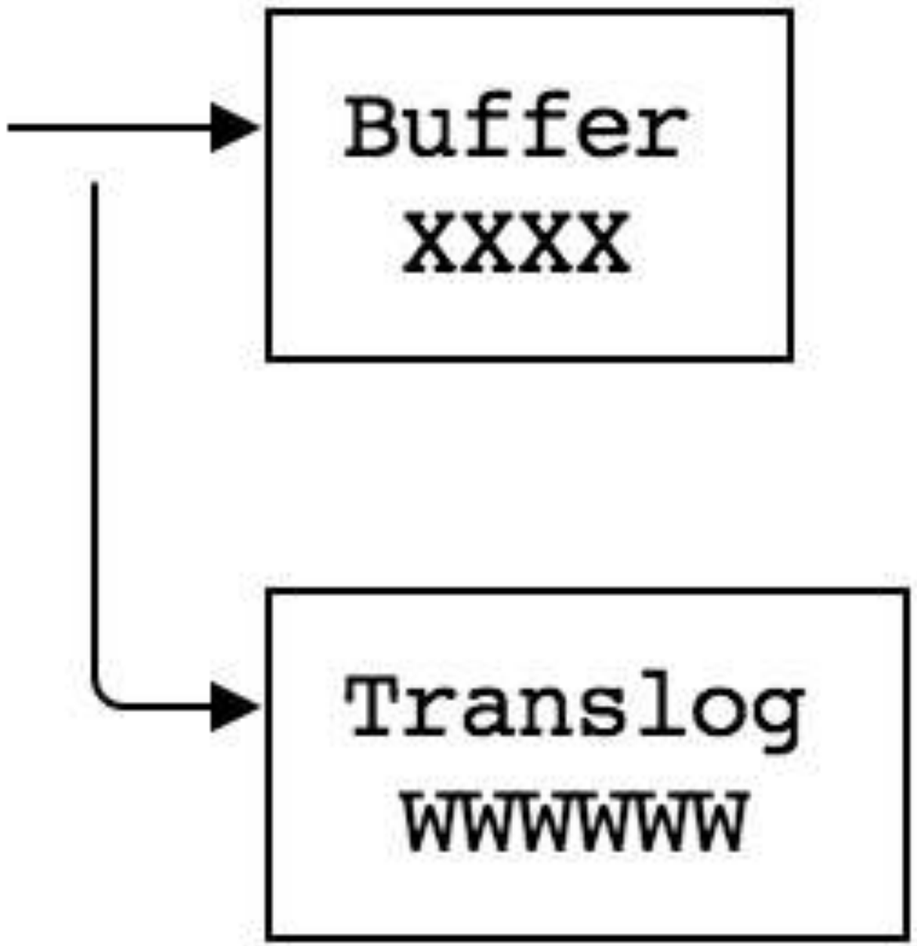

Buffer
XX

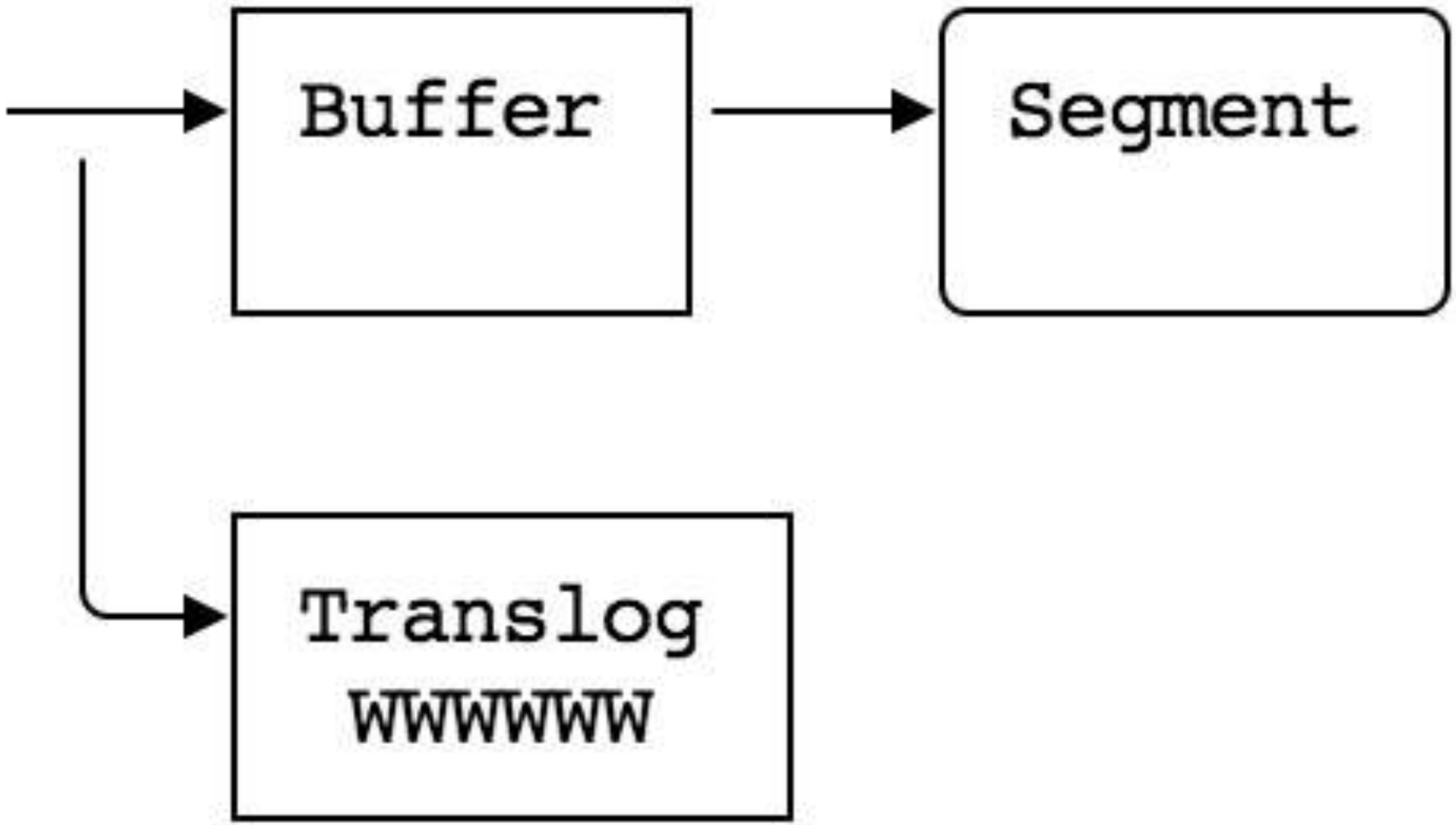

Buffer
XXX

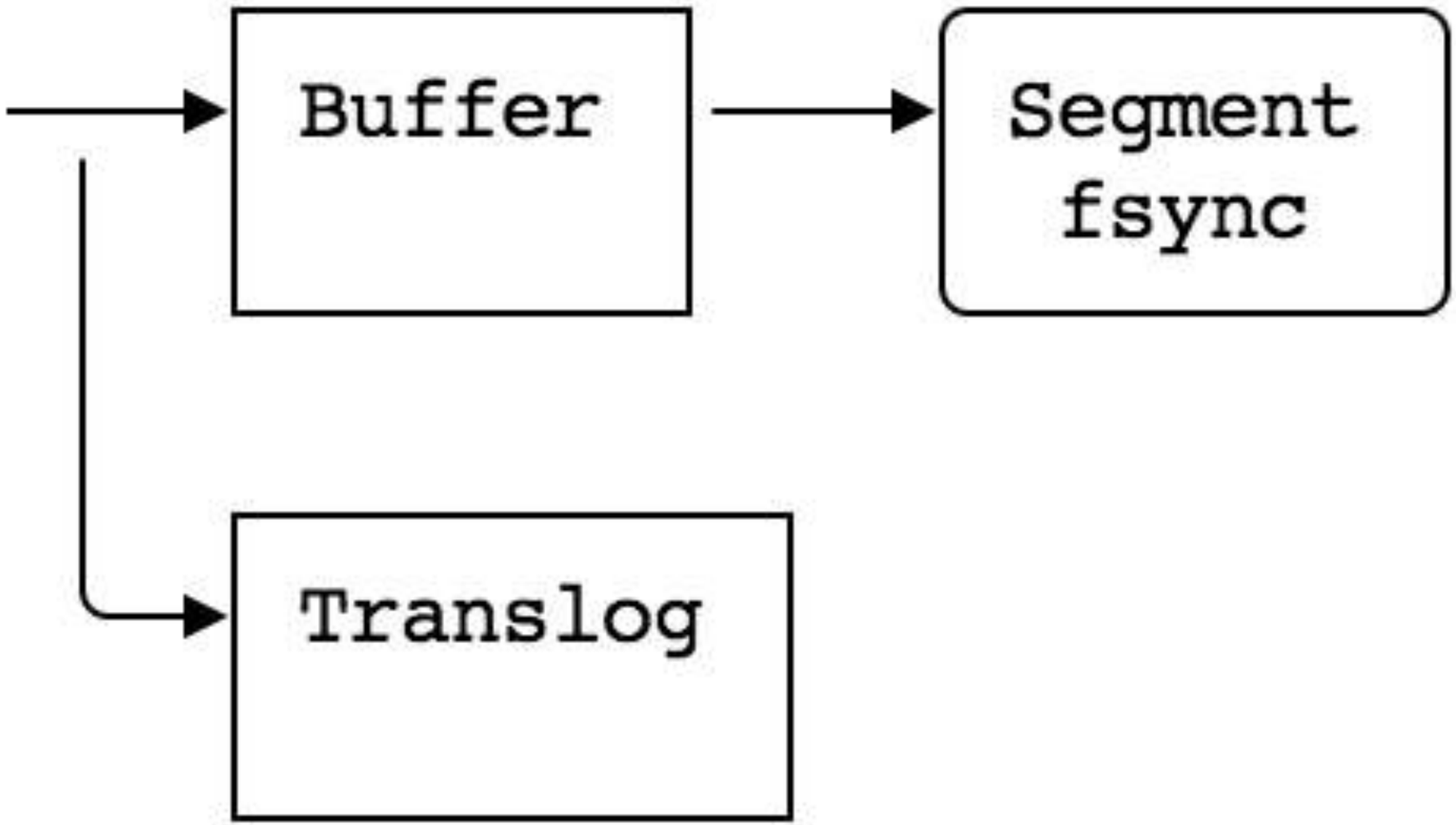

Buffer
Segment


Buffer
X
Translog
W


Buffer
Segment
Translog
WWW


Buffer
XXXX
Translog
WWWWW


Buffer
Segment
Translog
WWWWW


Buffer
Segment
fsync
Translog

Sequency Numbers

Quick recovery in 6.0

@xeraa

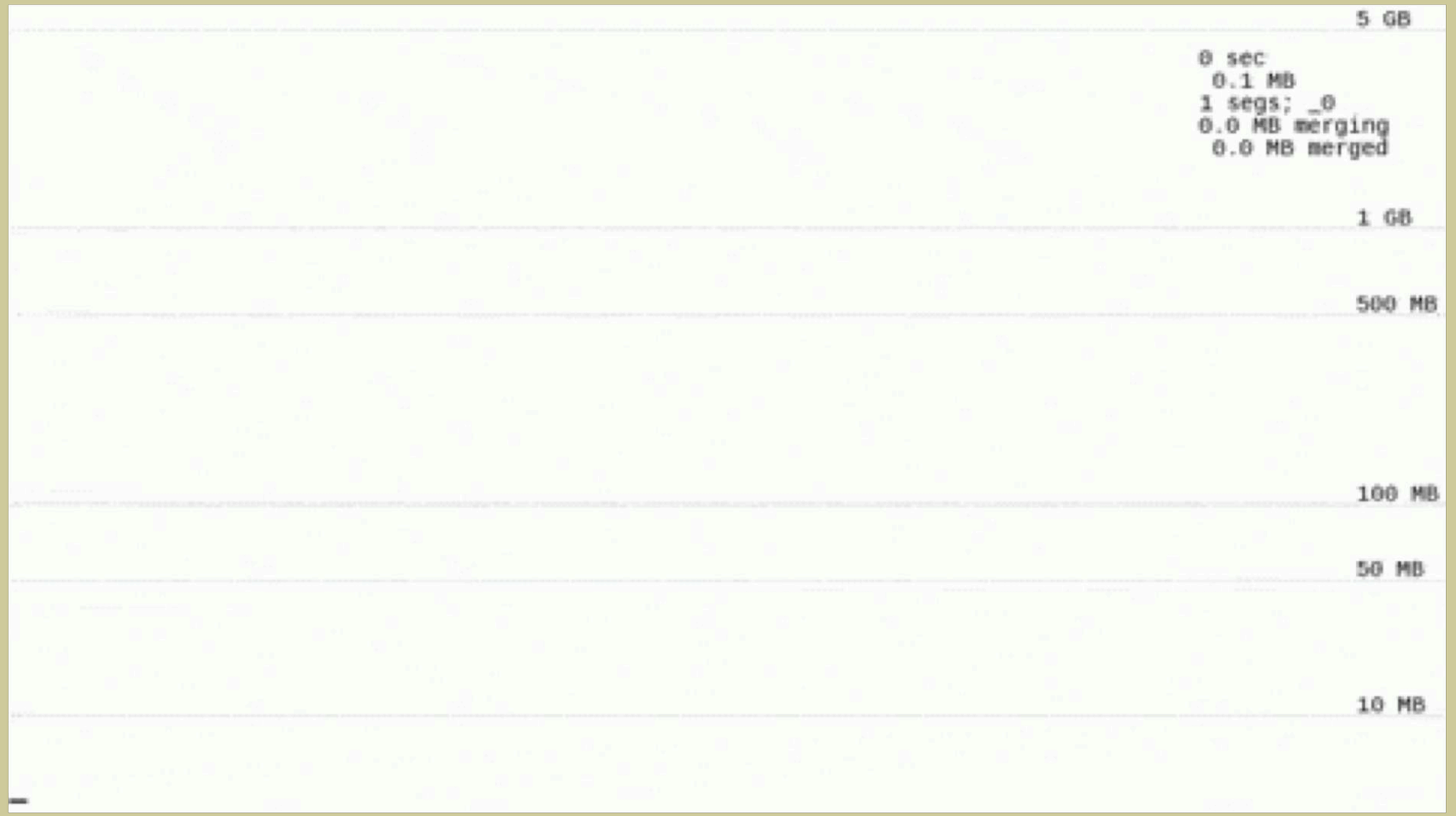
Segments Are Immutable

Merge

Combine (and Clean Up) Segments

Visualize Merges

http://blog.mikemccandless.com/2011/02/visualizing-lucenes-segment-merges.html


5 GB
0 sec
0.1 MB
1 segs; _0
0.0 MB merging
0.0 MB merged
1 GB
500 MB
100 MB
50 MB
10 MB

Refresh Demo

```
PUT databases
{
  "settings": {
    "refresh_interval": "30s"
  }
}
```

```
PUT databases/_doc/1
{
  "name": "Elasticsearch",
  "author": "Shay Banon",
  "stable_version": "7.13.2"
}
```

```
GET databases/_doc/1

GET databases/_search
{
  "query": {
    "match": {
      "name": "elasticsearch"
    }
  }
}
```

```
PUT databases/_doc/1
{
  "name": "Elasticsearch",
  "author": "Shay Banon",
  "stable_version": "7.14.0"
}

GET databases/_doc/1
```

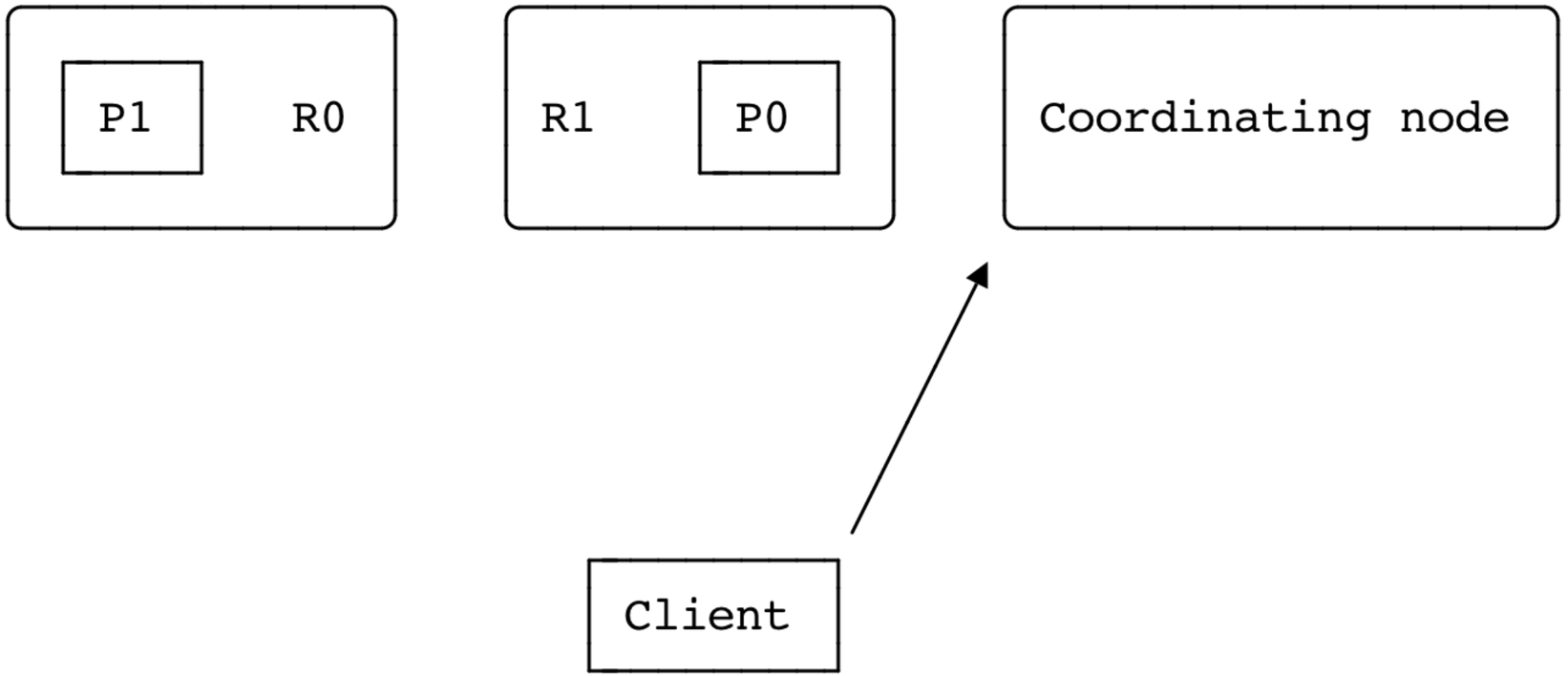
```
PUT databases/_doc/1?refresh=wait_for
{
  "name": "Elasticsearch",
  "author": "Shay Banon",
  "stable_version": "7.14.1"
}
```

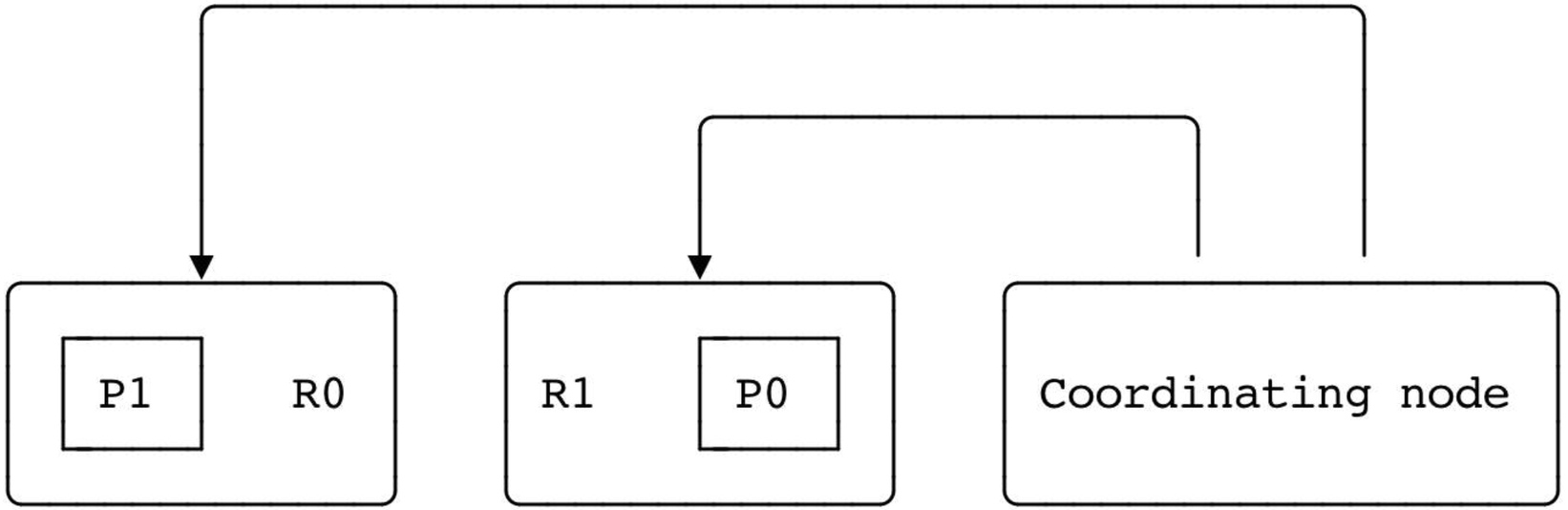
Searching

Coordinating Node

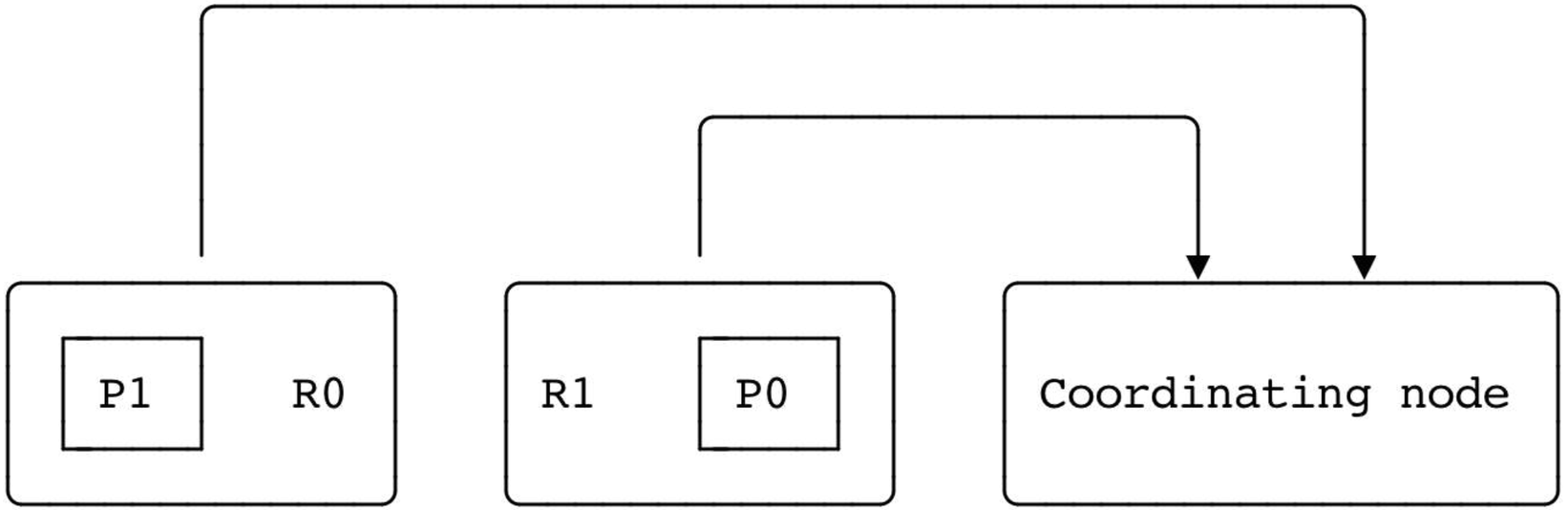
Scatter & Gather

Query Phase


P1
R0
R1
P0
Coordinating node
Client

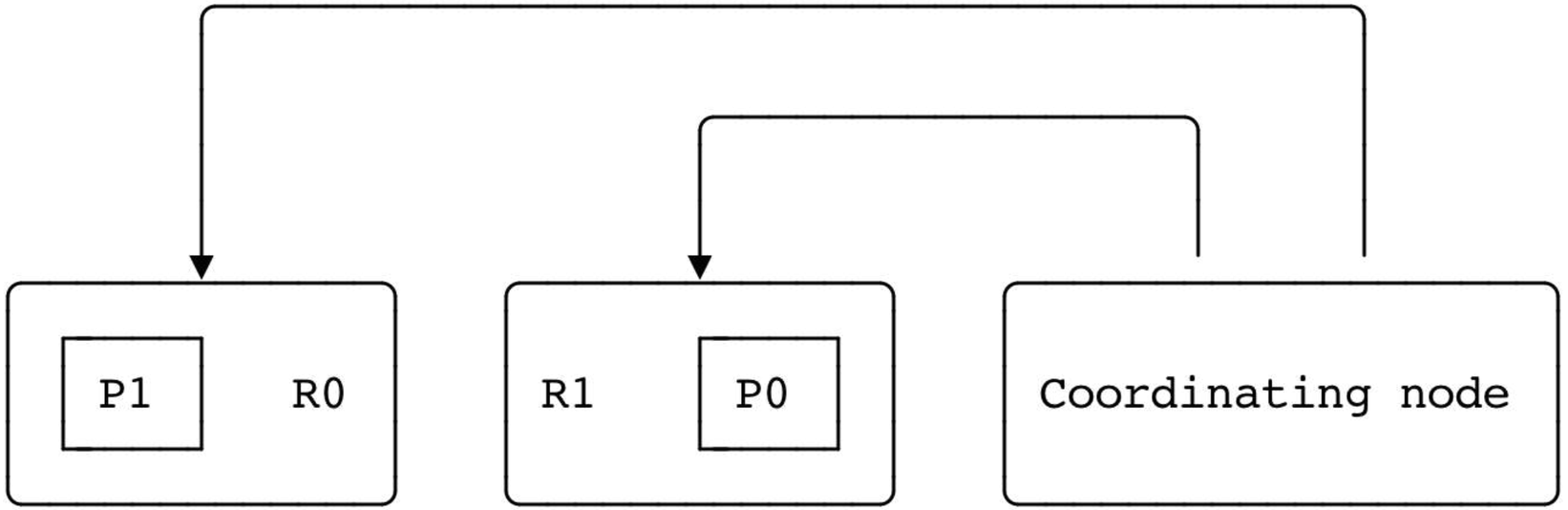


Client

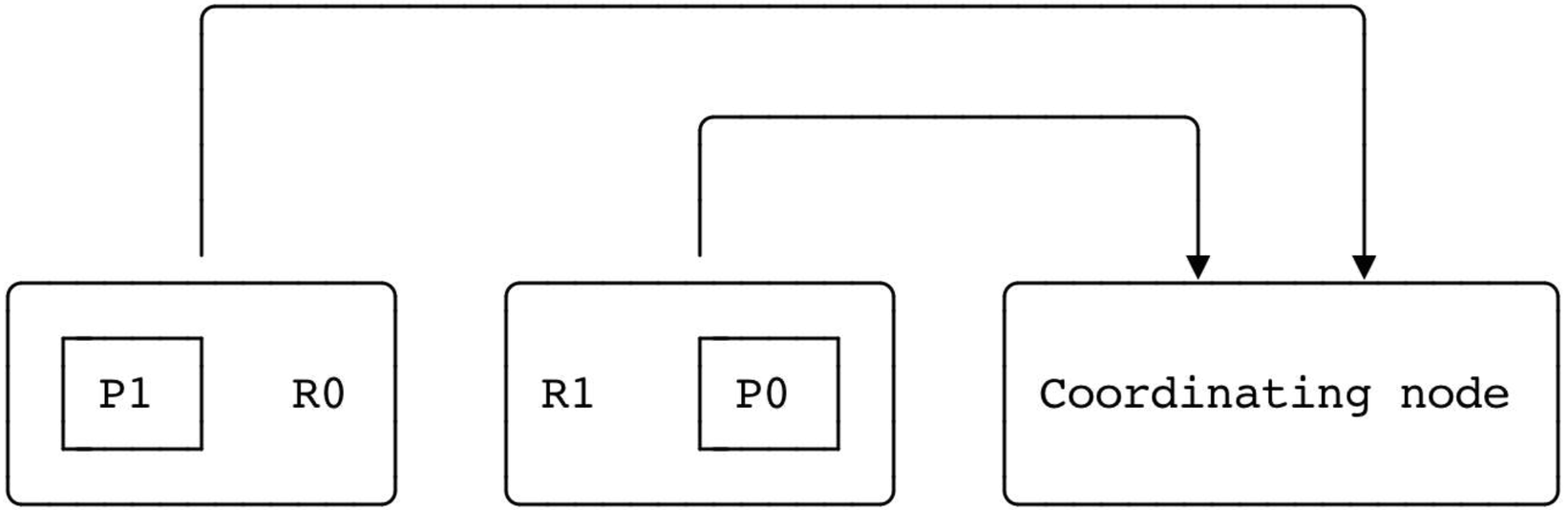


Client

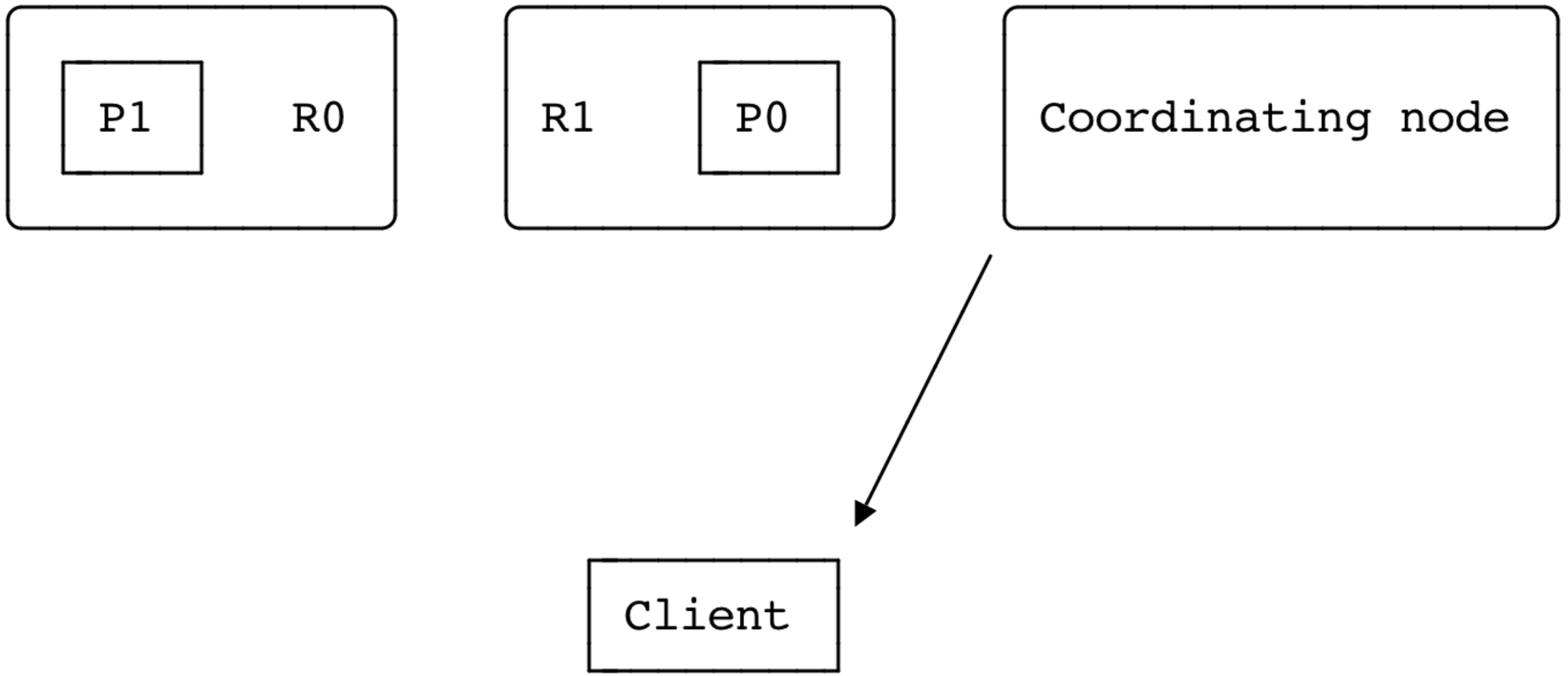
Fetch Phase



Client



Client


P1
R0
R1
P0
Coordinating node
Client

Conclusion

The Meaning of *Index*

Form a Cluster

Store Documents

Search Documents

Questions?

Philipp Krenn **@xeraa**