Real-time analytics with Cassandra

Sylvain Lebresne
sylvain@datastax.com

Berlin Buzzword - 6th June, 2011

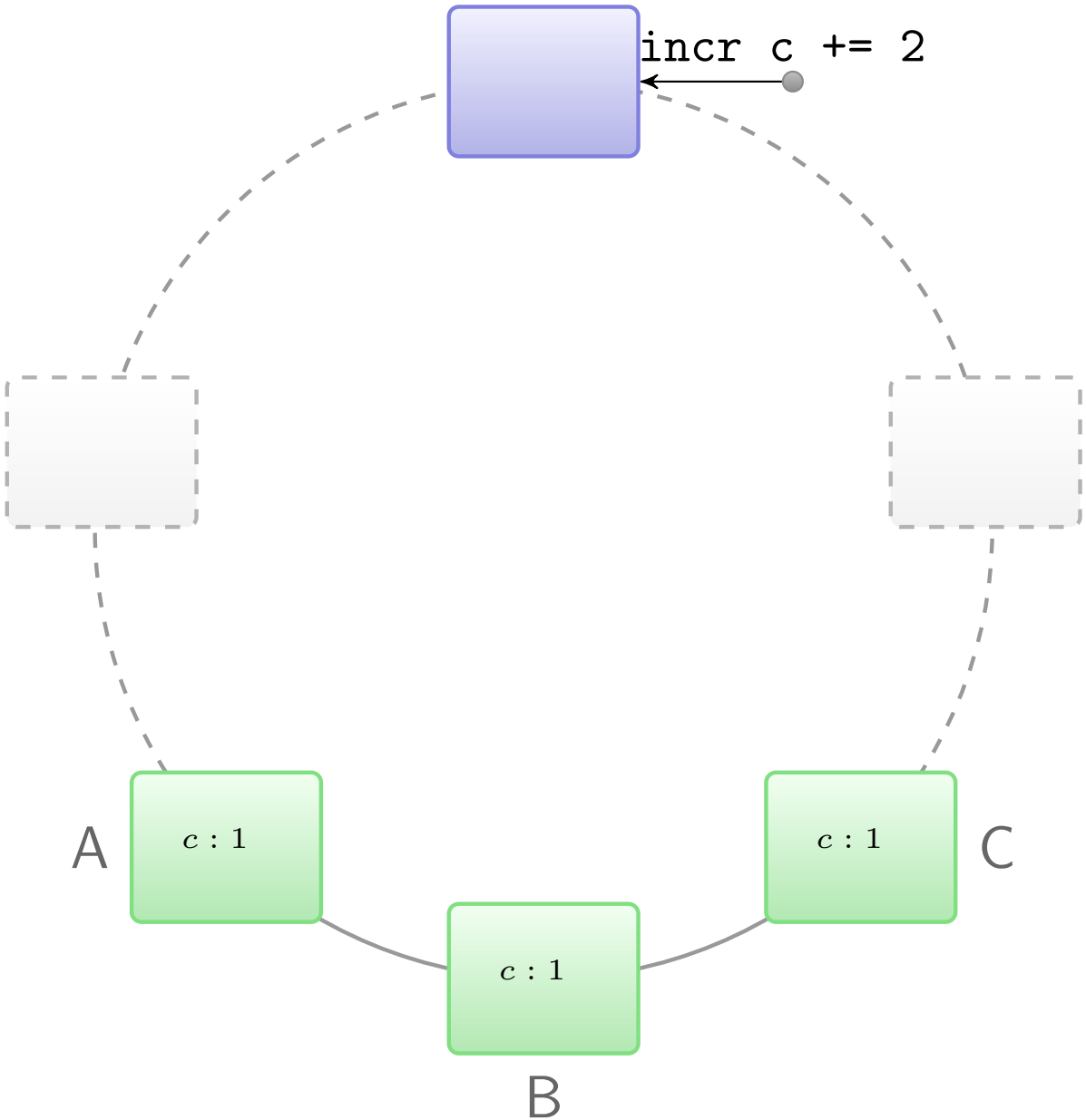
Distributed counters

- Really a talk about the distributed counter implementation in Cassandra.
- Just released in Cassandra 0.8.
- Before, counting (using set/get) required external synchronization (Zookeeper, Cages).
- Patch initiated by Digg, then Twitter (Rainbird).
- Goal: count lots of stuffs very quickly.

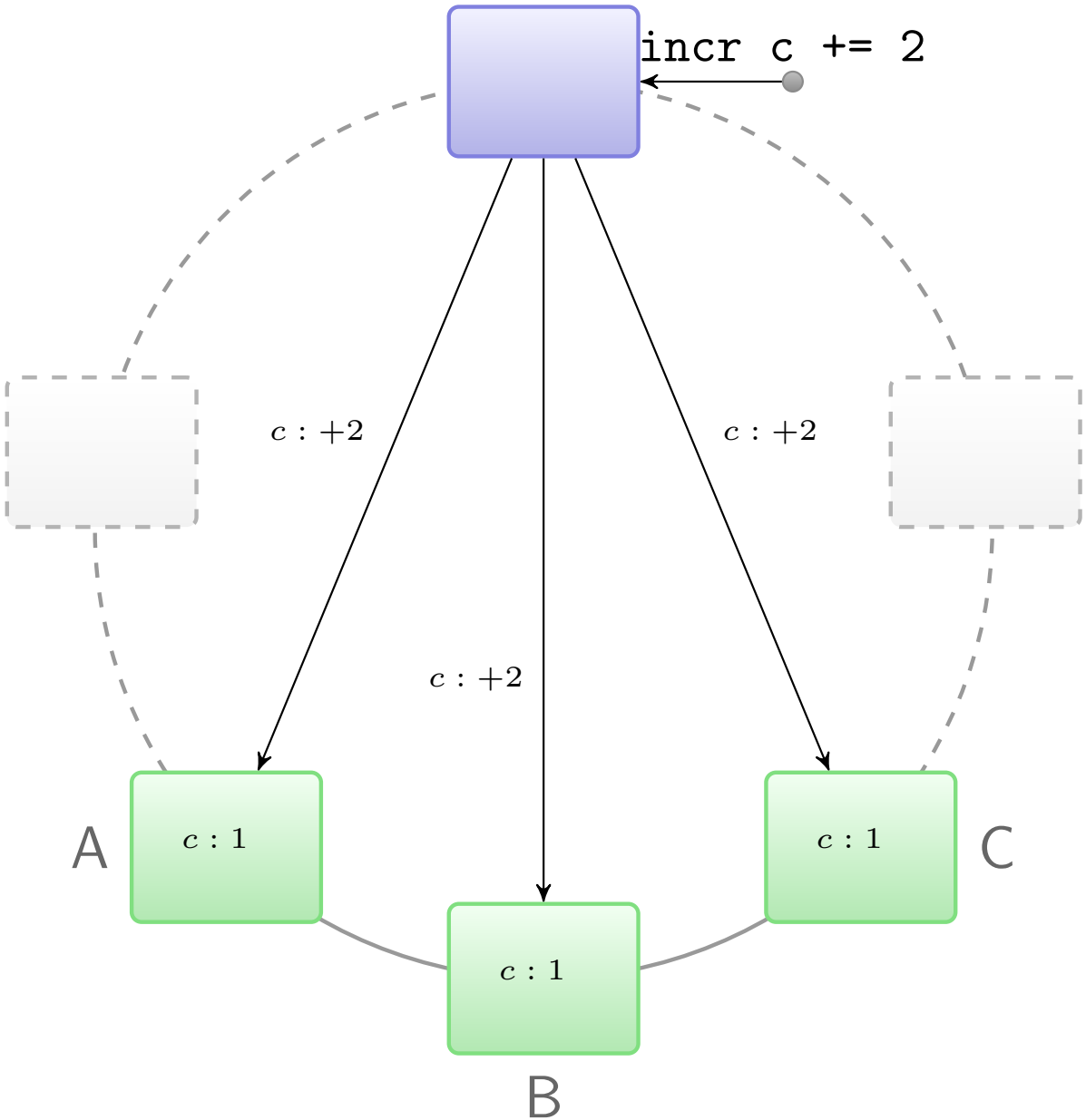
Cassandra

- Distributed & replicated (no SPOF)
- Highly available (partition tolerant)
- Writes are fast
- Multi-datacenter awareness built-in.
- We want to add counters (i.e, an increment operation).

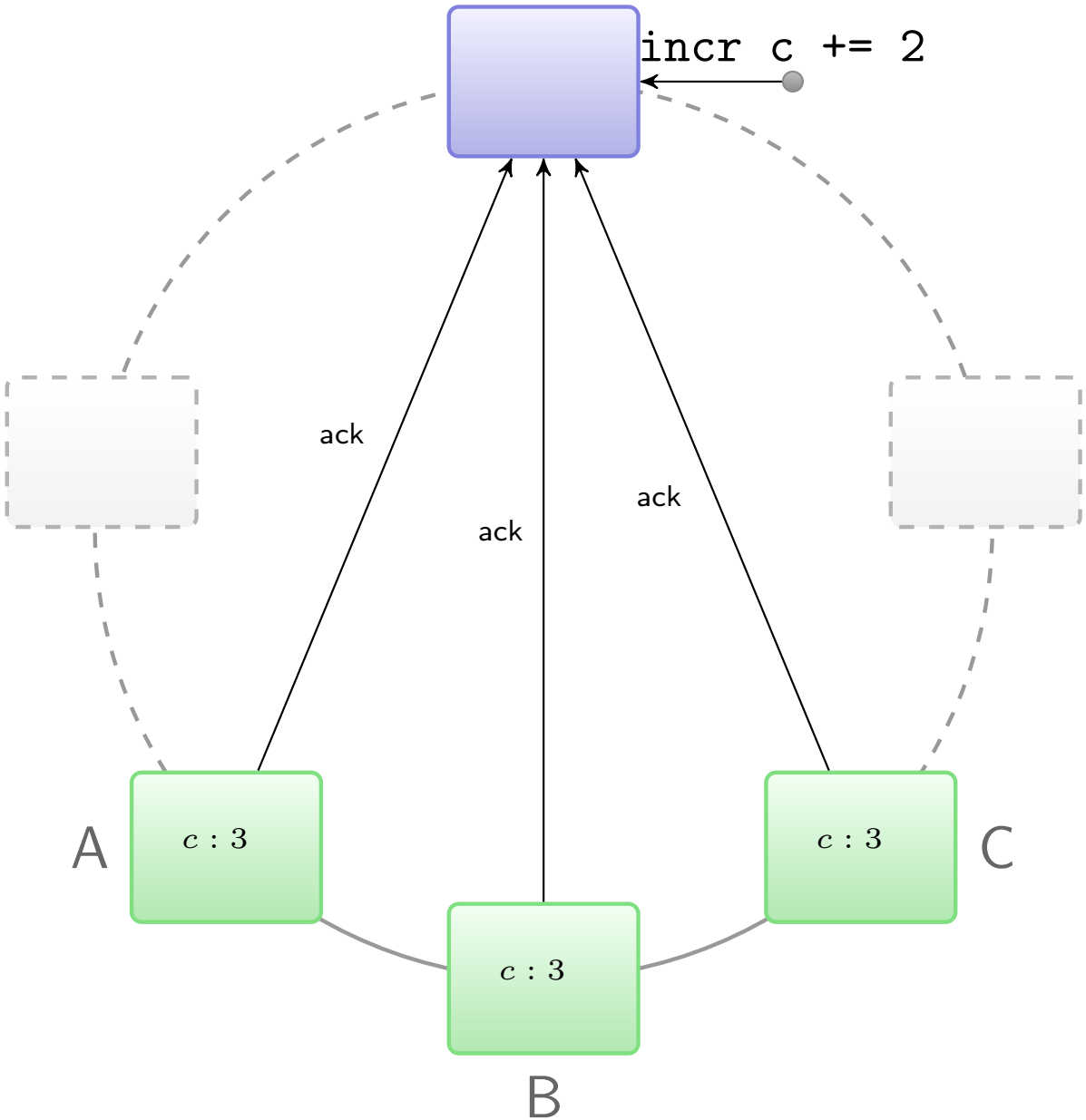
The naive approach



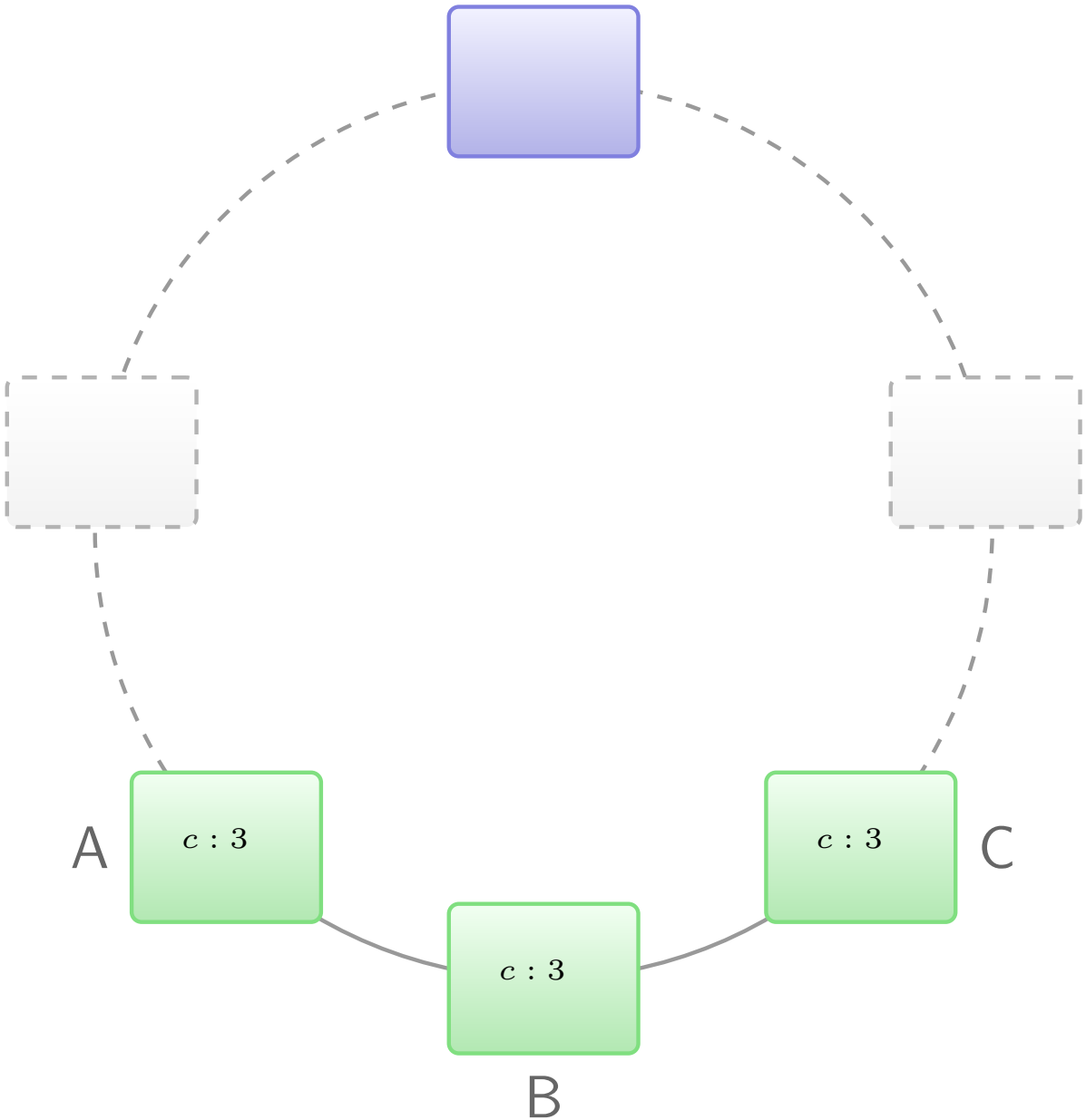
The naive approach



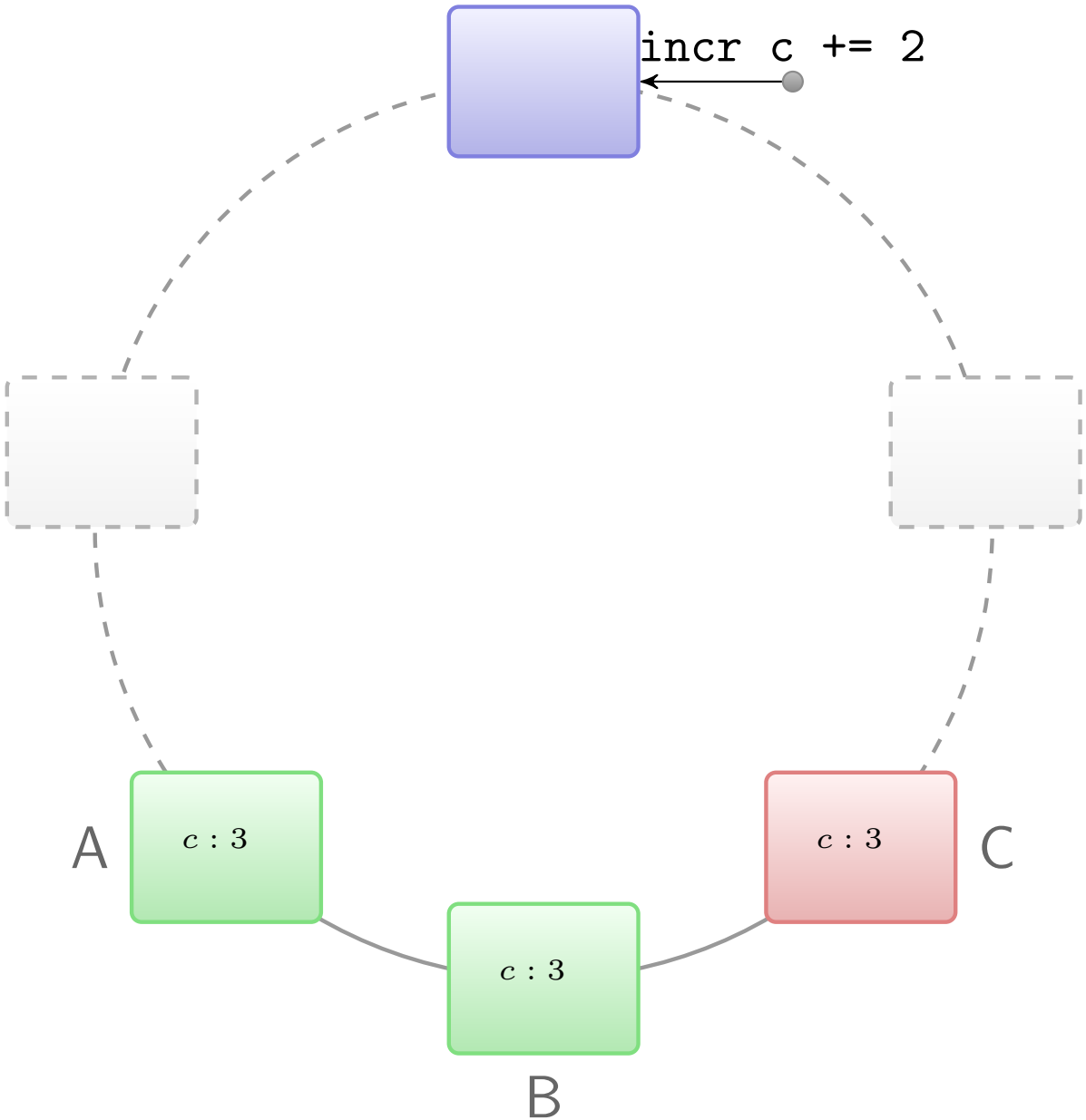
The naive approach



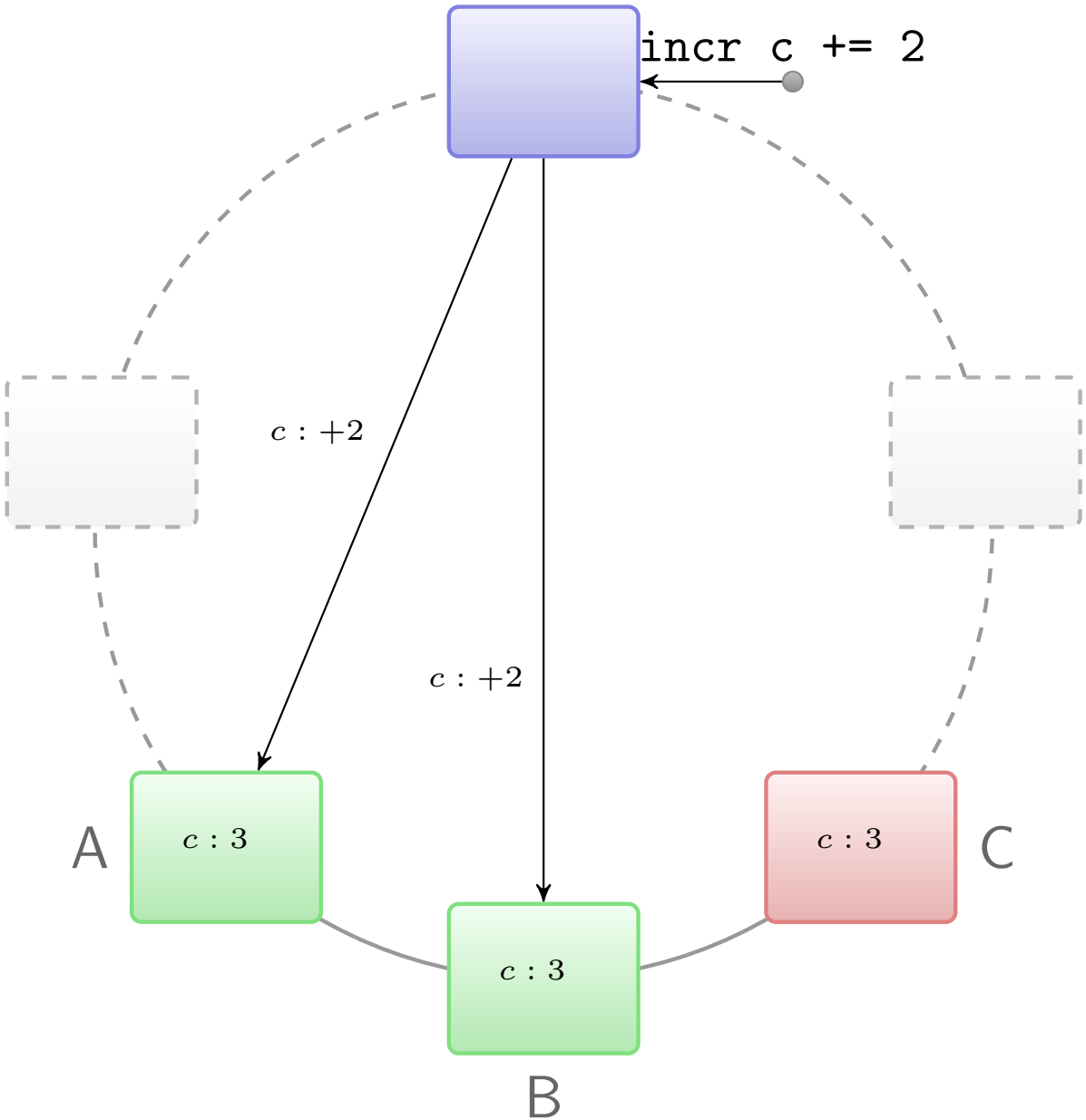
The naive approach



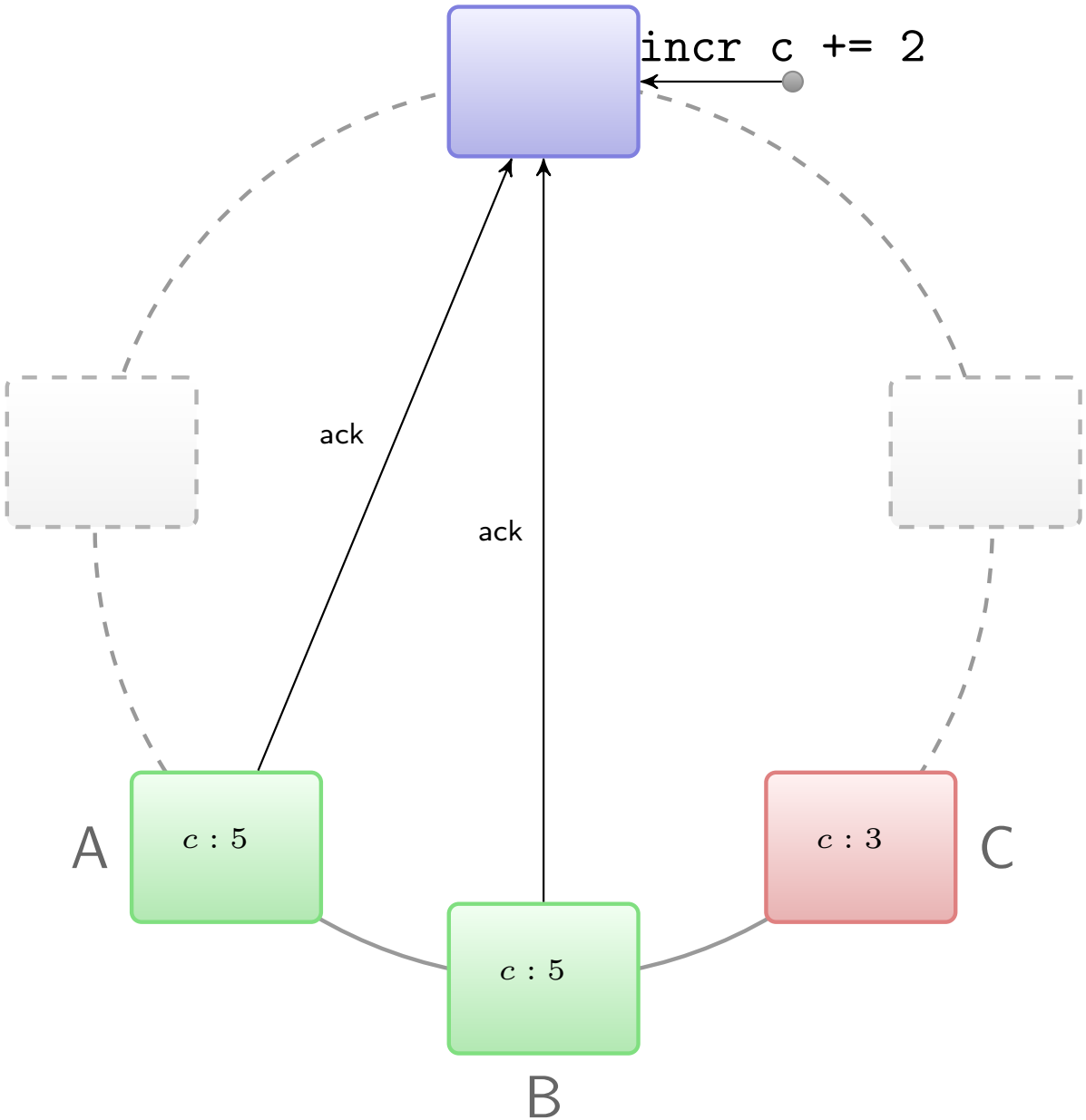
The naive approach



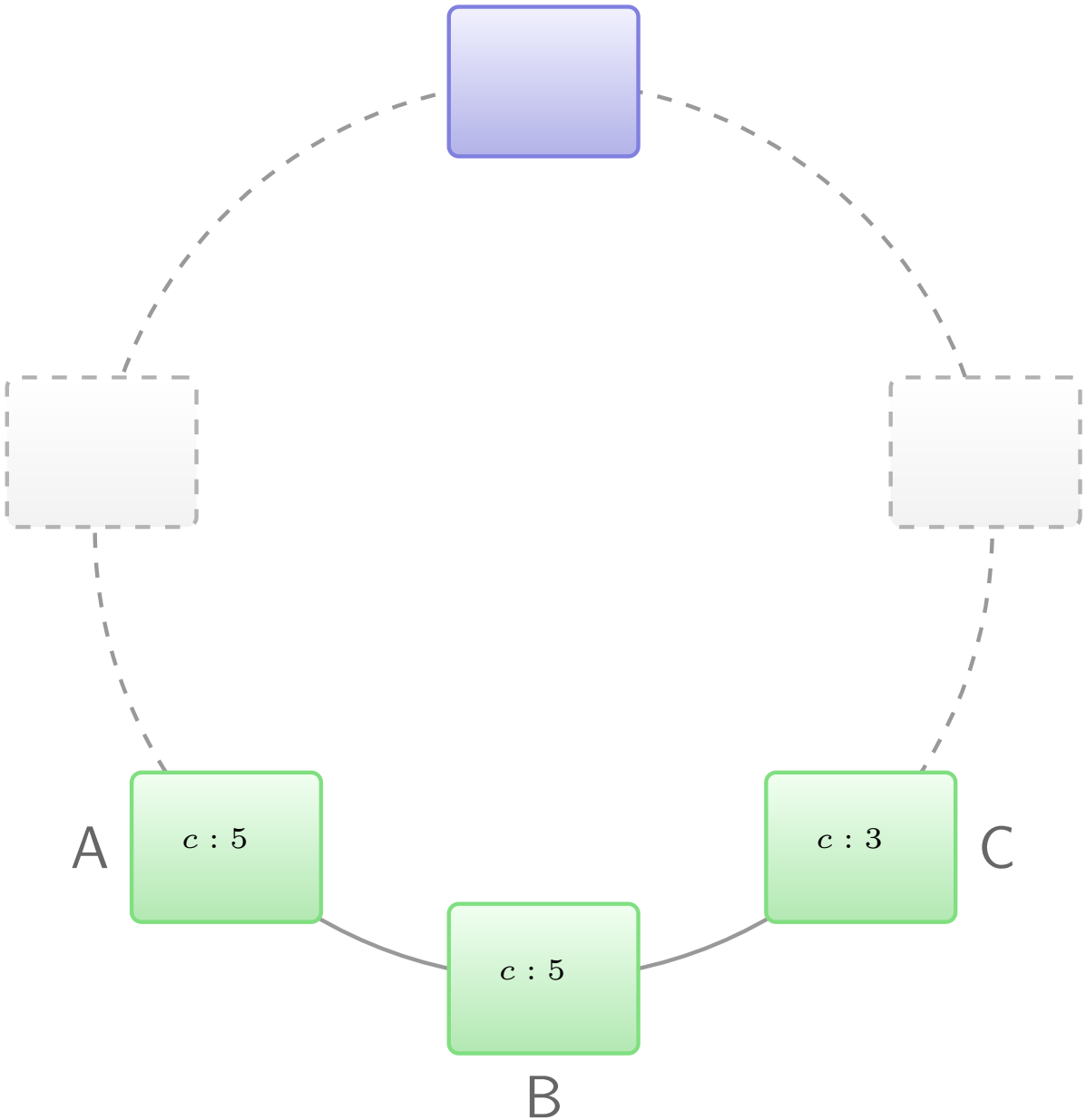
The naive approach



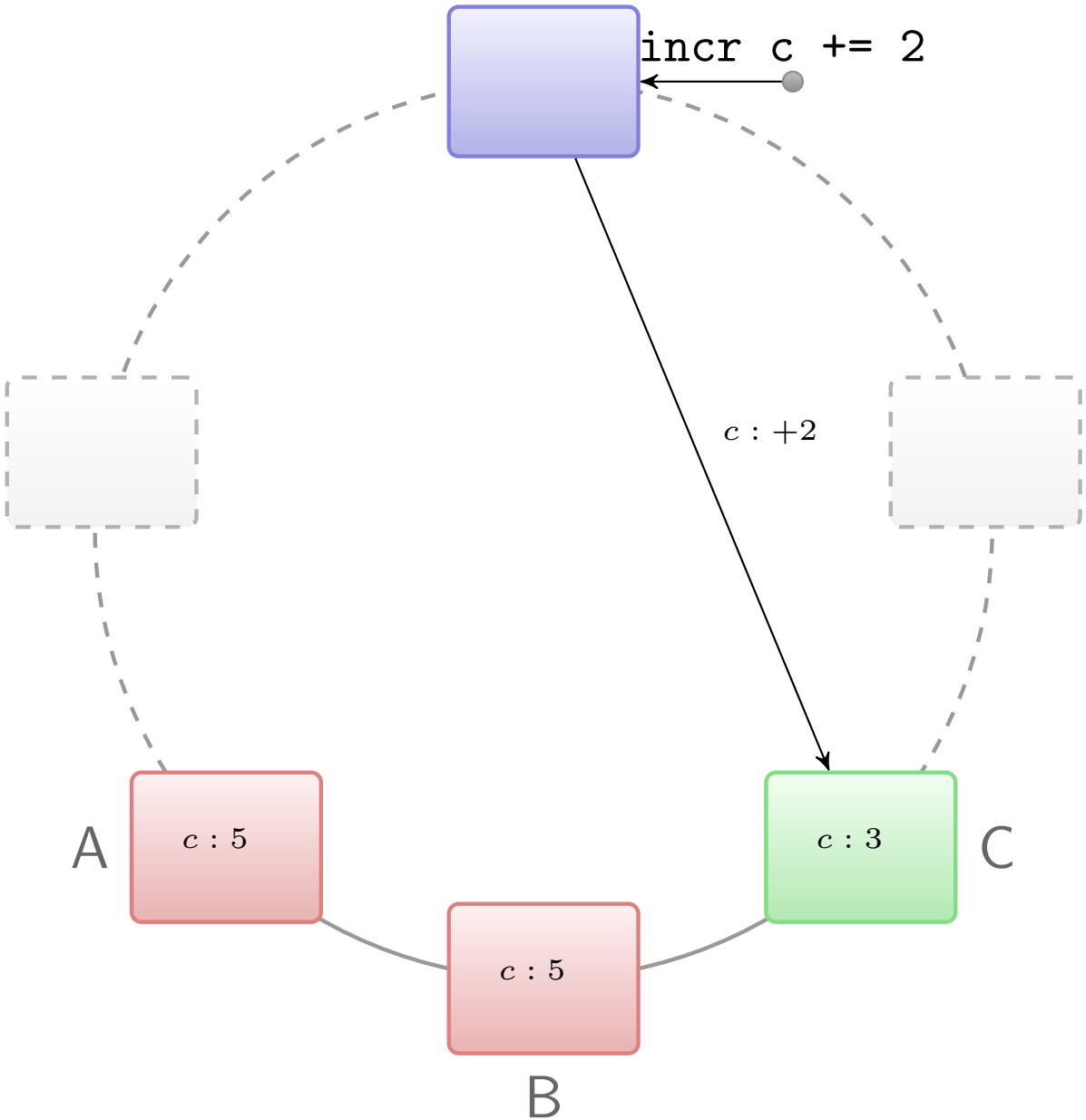
The naive approach



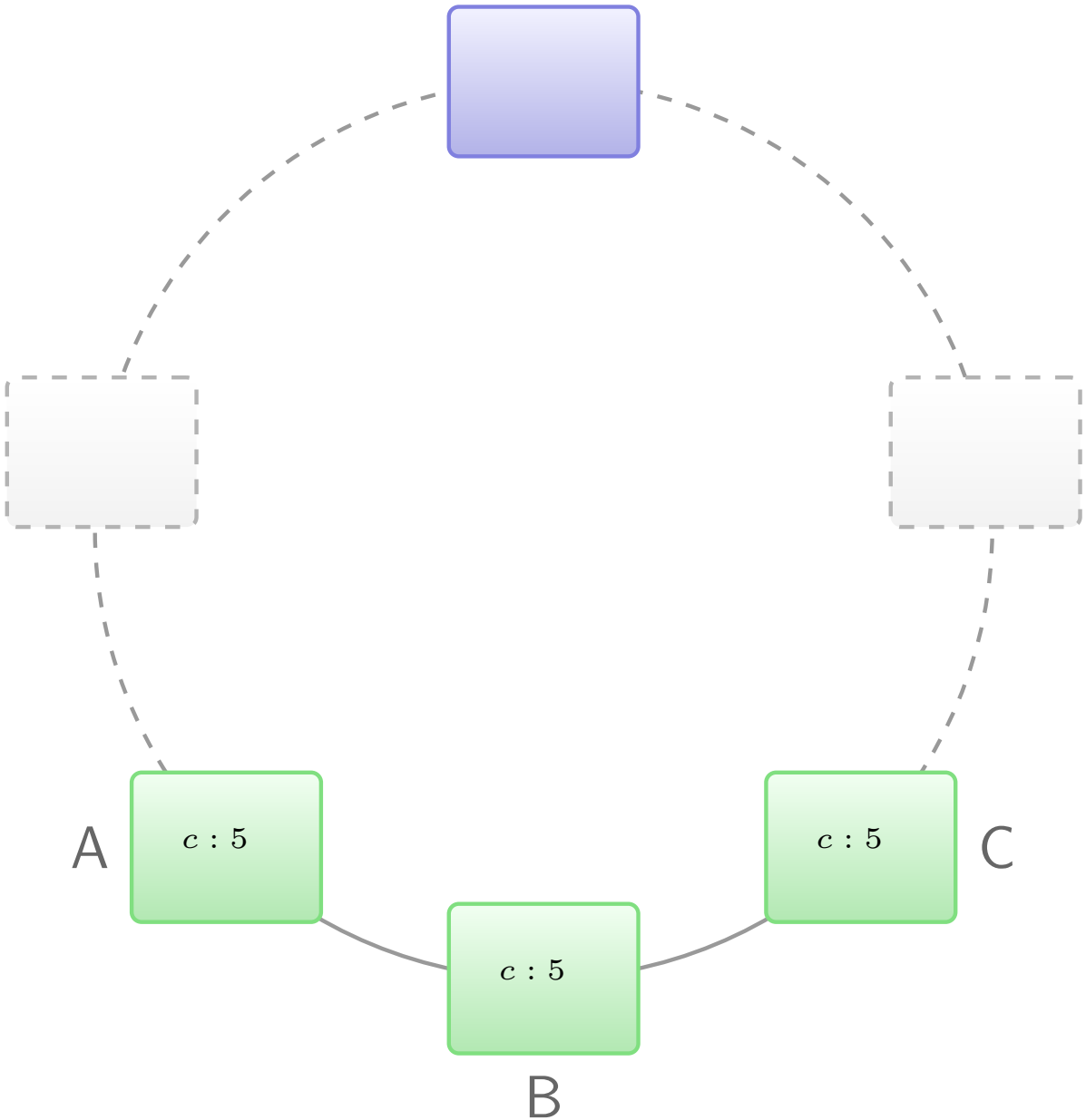
The naive approach



The naive approach



The naive approach

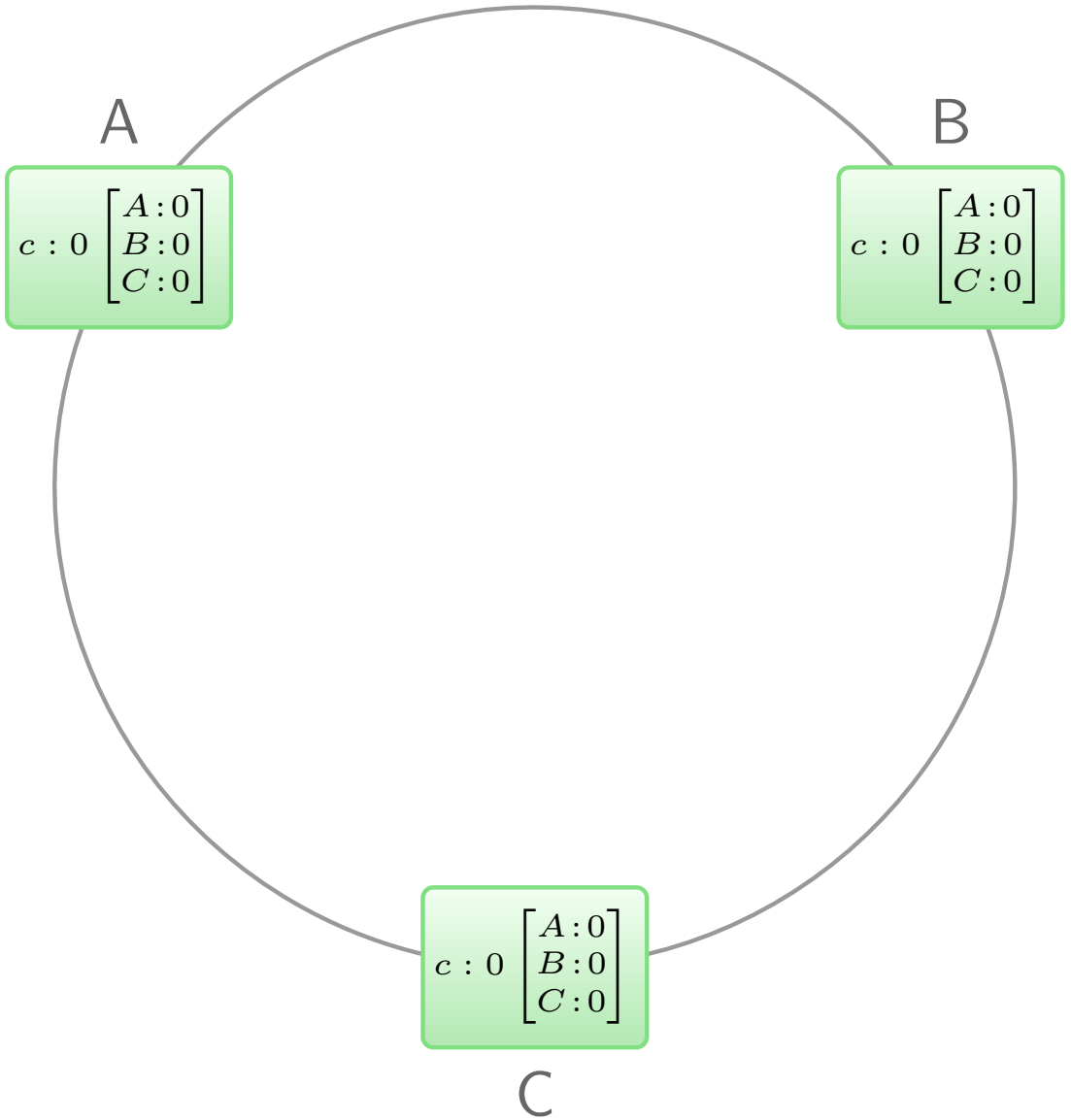


Vector clocks/Version vectors

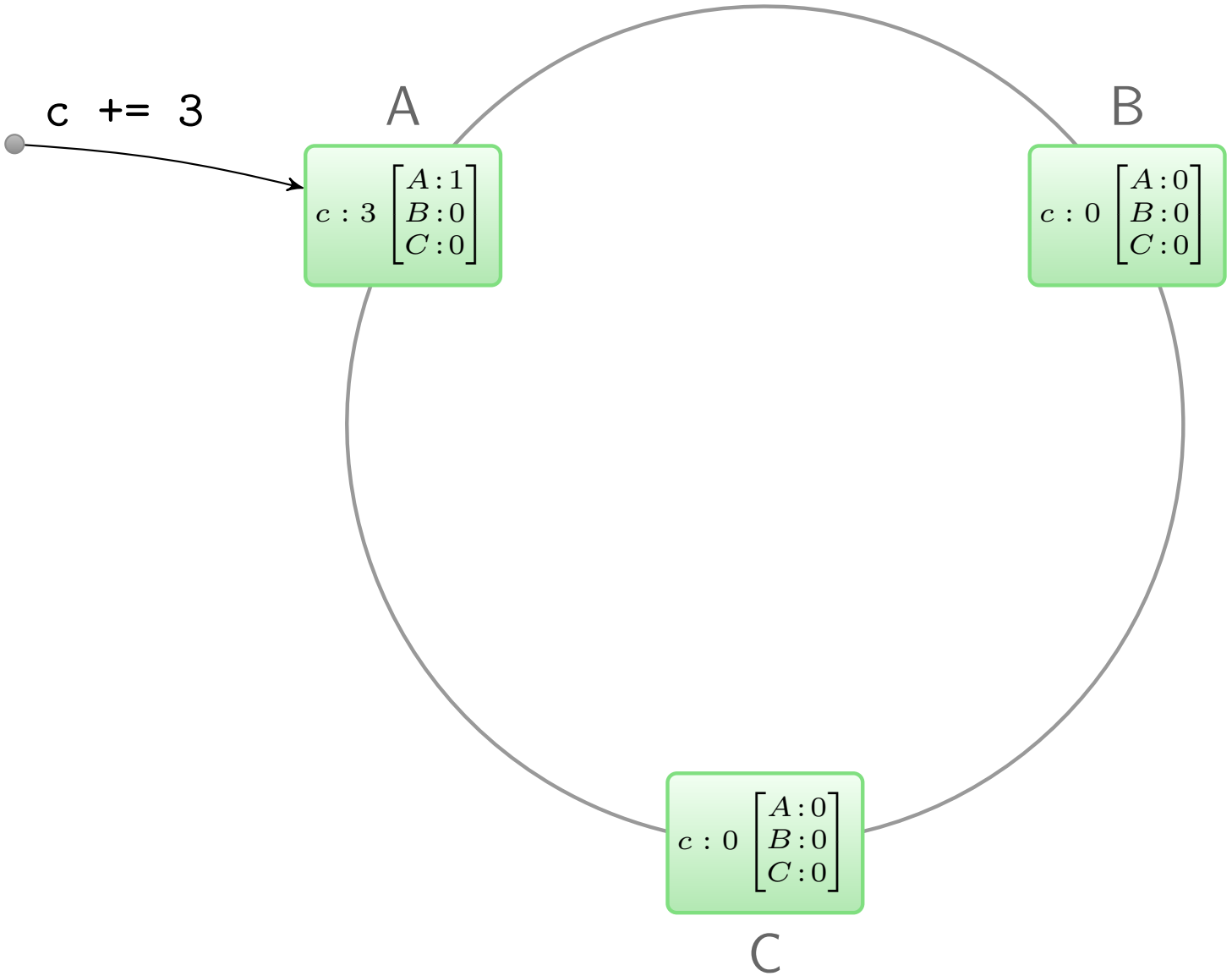
"Vector clocks is an algorithm for generating a partial ordering of events in a distributed system and detecting causality violations"

– Wikipedia

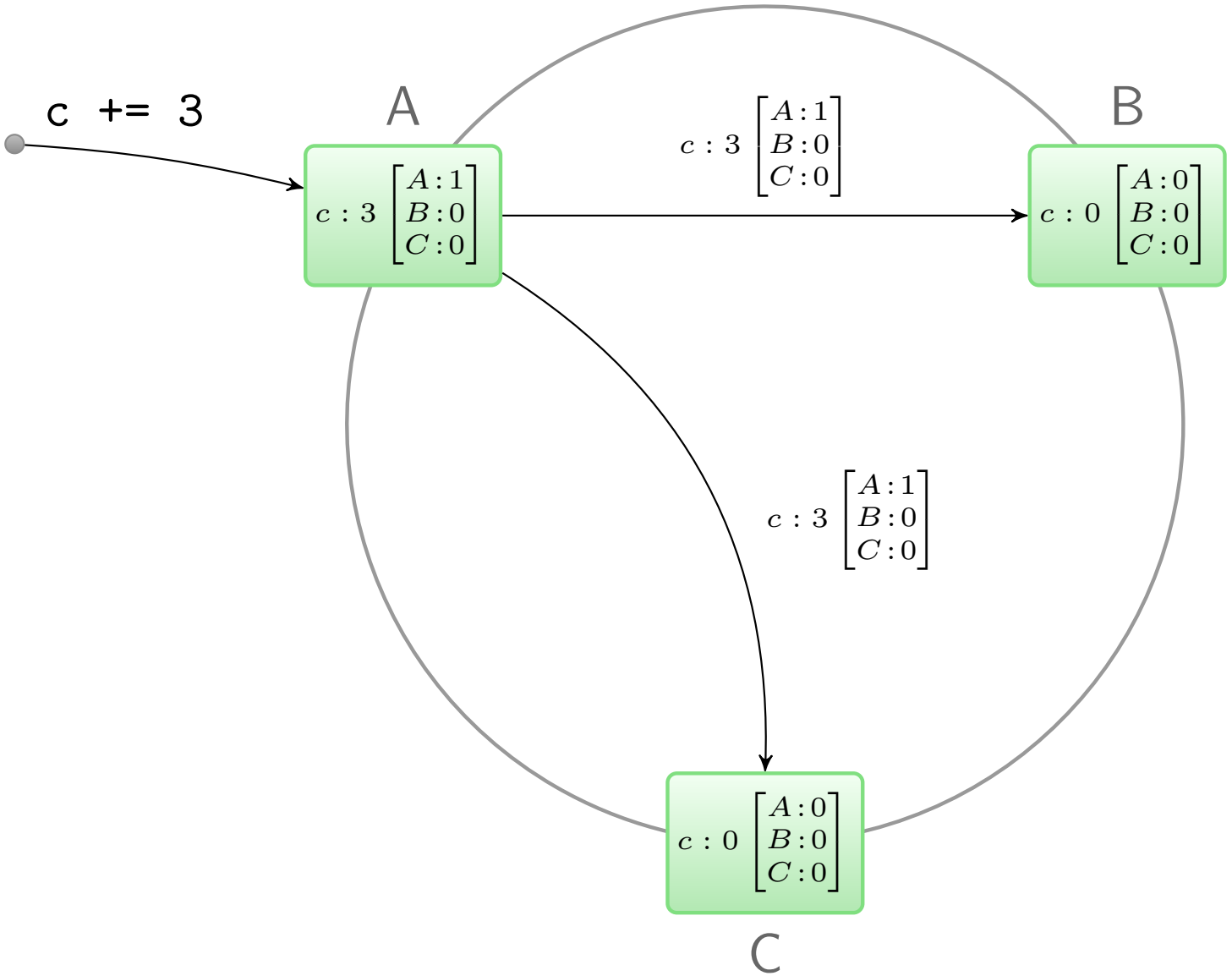
Counting with version vectors



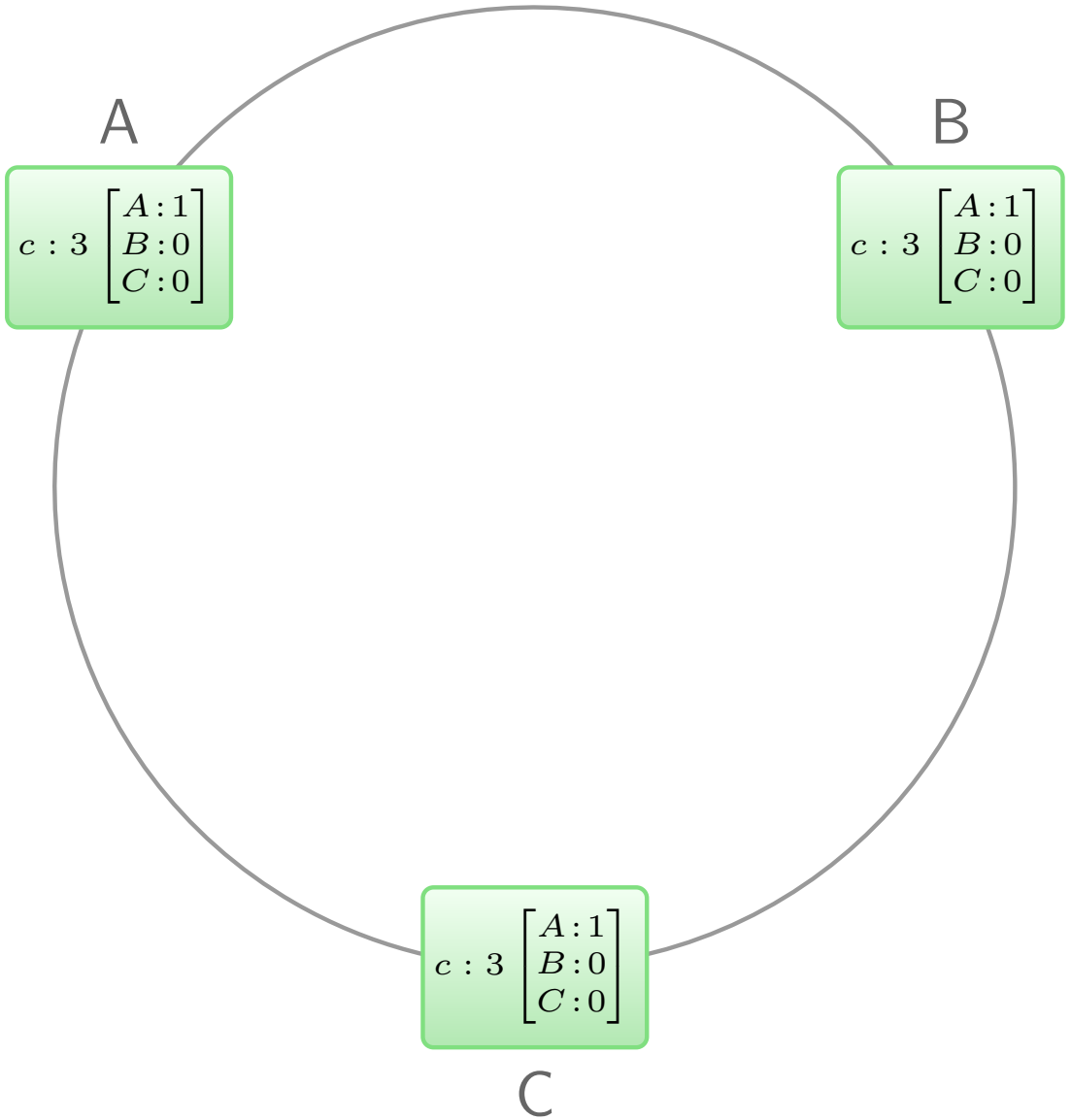
Counting with version vectors



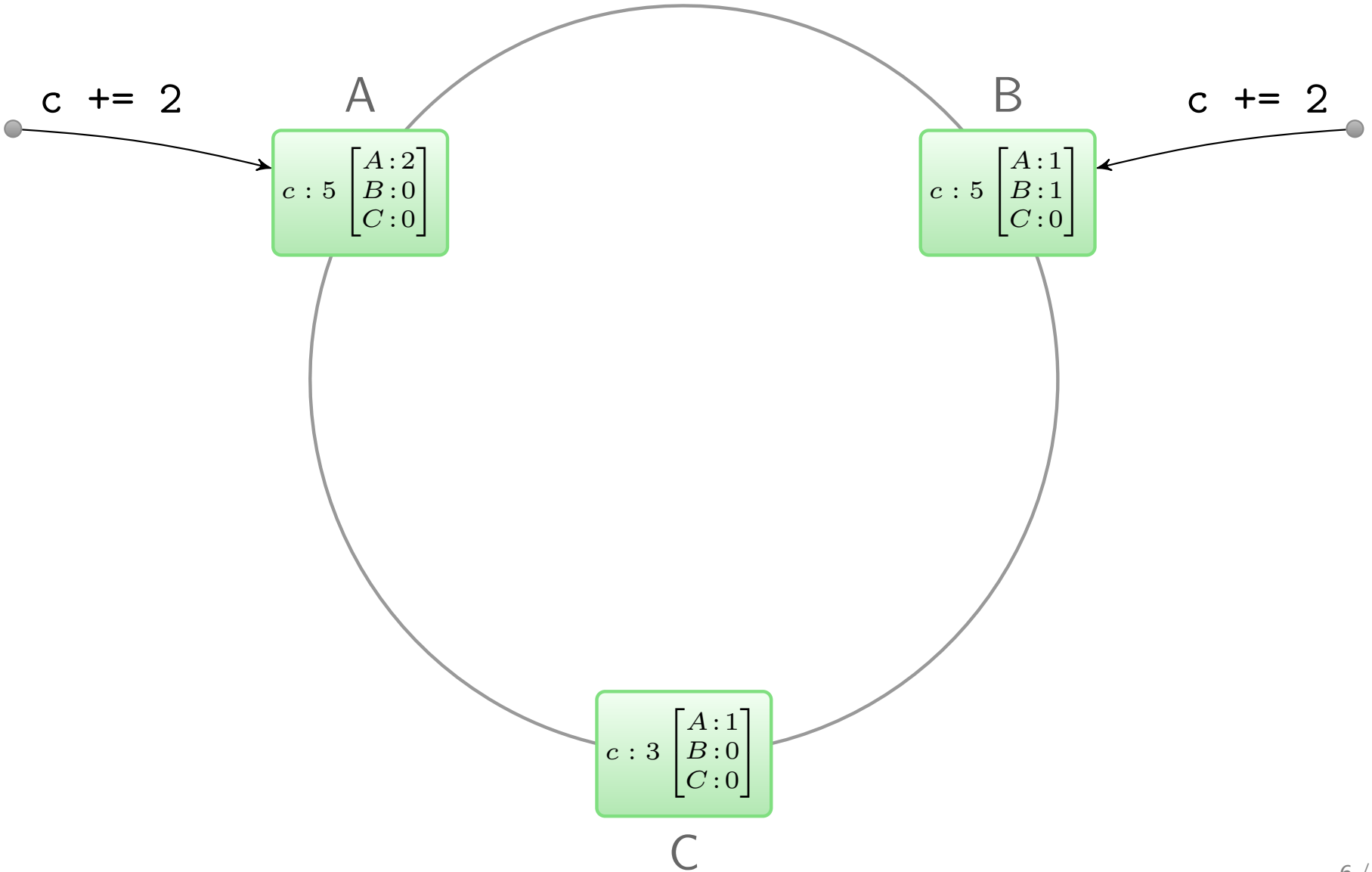
Counting with version vectors



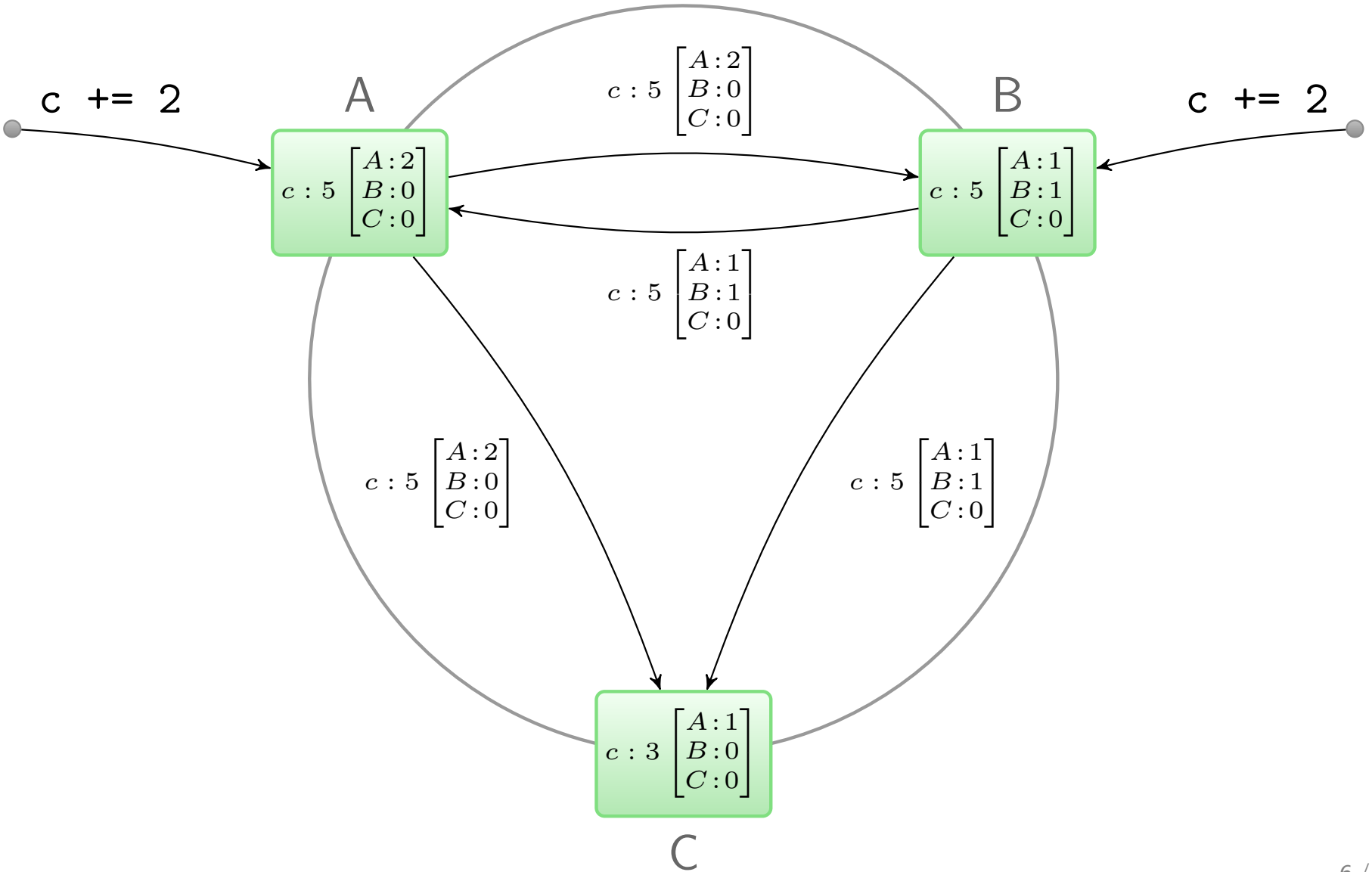
Counting with version vectors



Counting with version vectors



Counting with version vectors



- Vector clocks/version vectors are about detecting conflicting updates ...
- ... but says nothing about how to resolve those conflicts.

- Vector clocks/version vectors are about detecting conflicting updates ...
- ... but says nothing about how to resolve those conflicts.
- Still, the idea of distinguishing what each replica has seen is key.

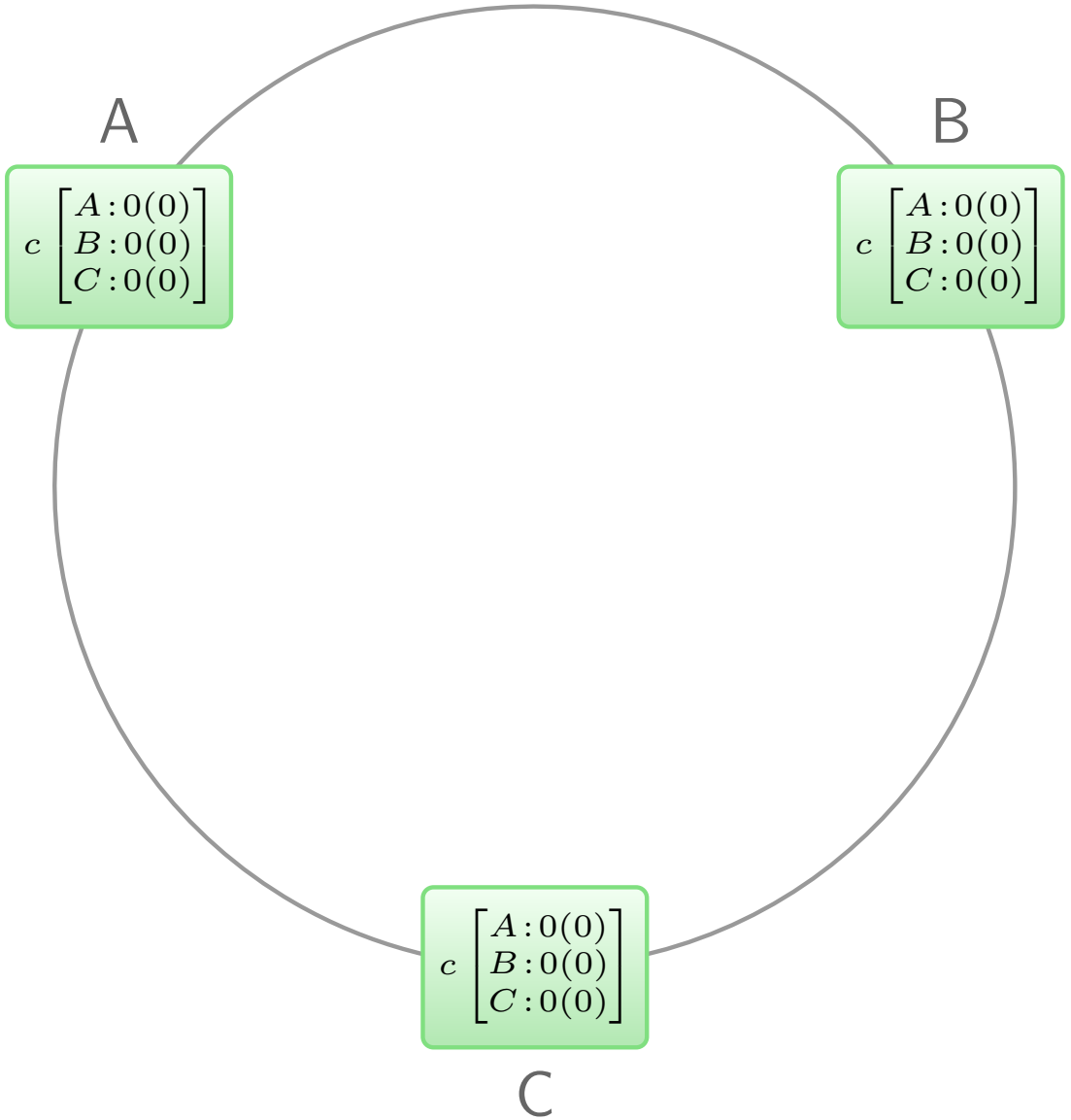
Partitioned counters

A counter is partitioned into one sub-count by replica of the counter, it is a vector of tuple (host id, sub-count, version).

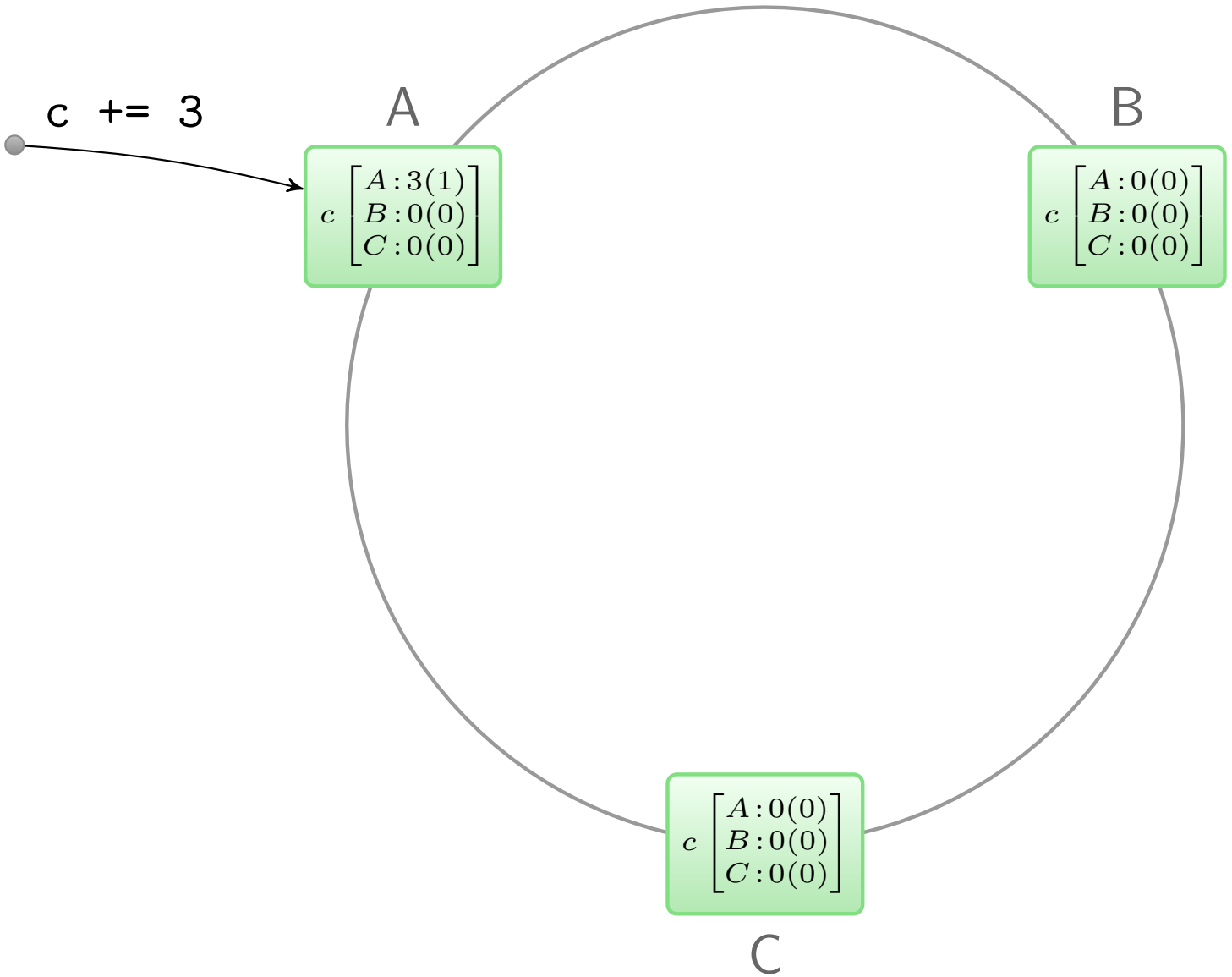
The actual value of the counter is the sum of all the sub-counts. It is only resolved on reads before answering the client.

$$c : \begin{bmatrix} A : 24 \ (2) \\ B : 42 \ (3) \\ C : 17 \ (1) \end{bmatrix} \iff c = 83$$

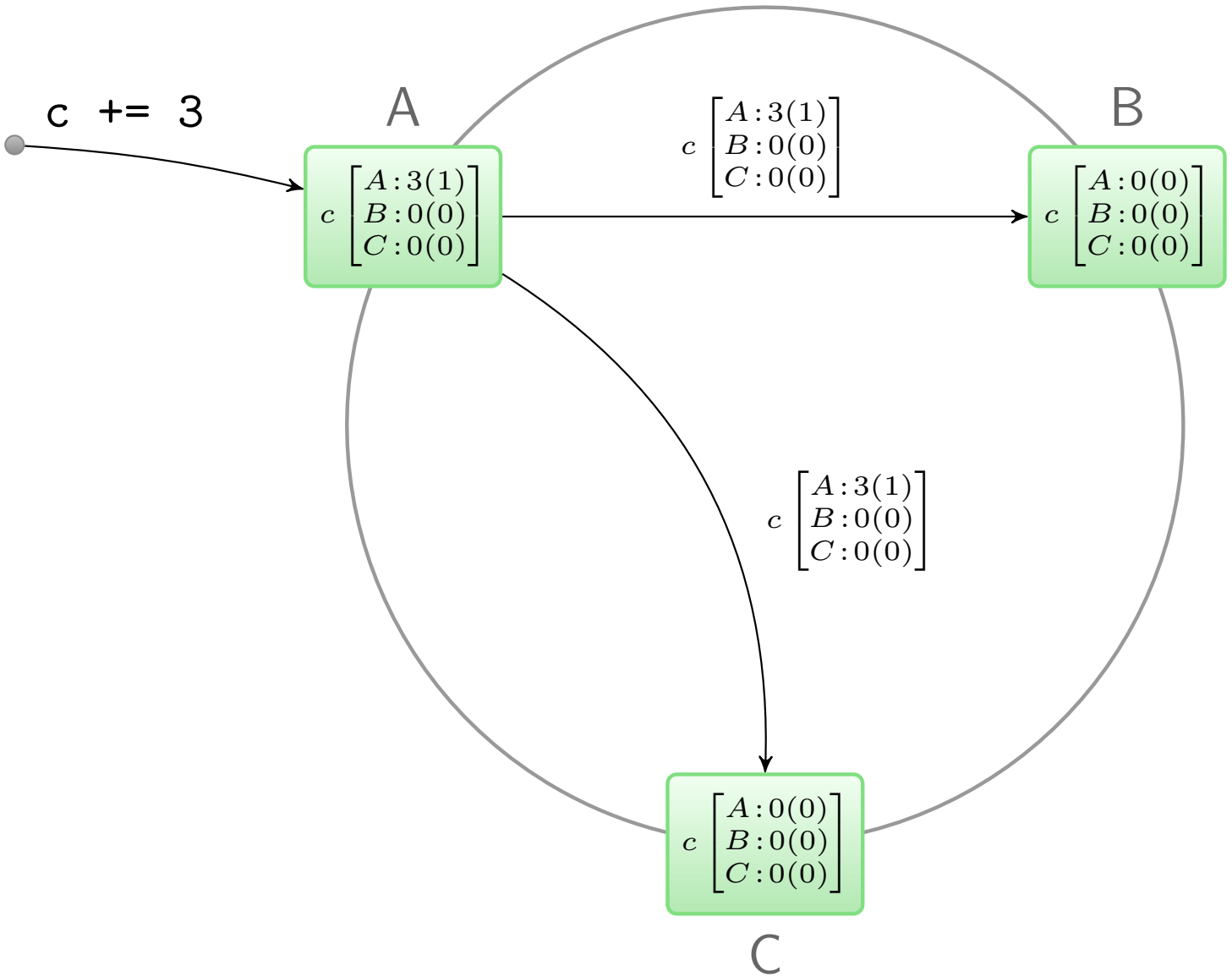
The Cassandra way



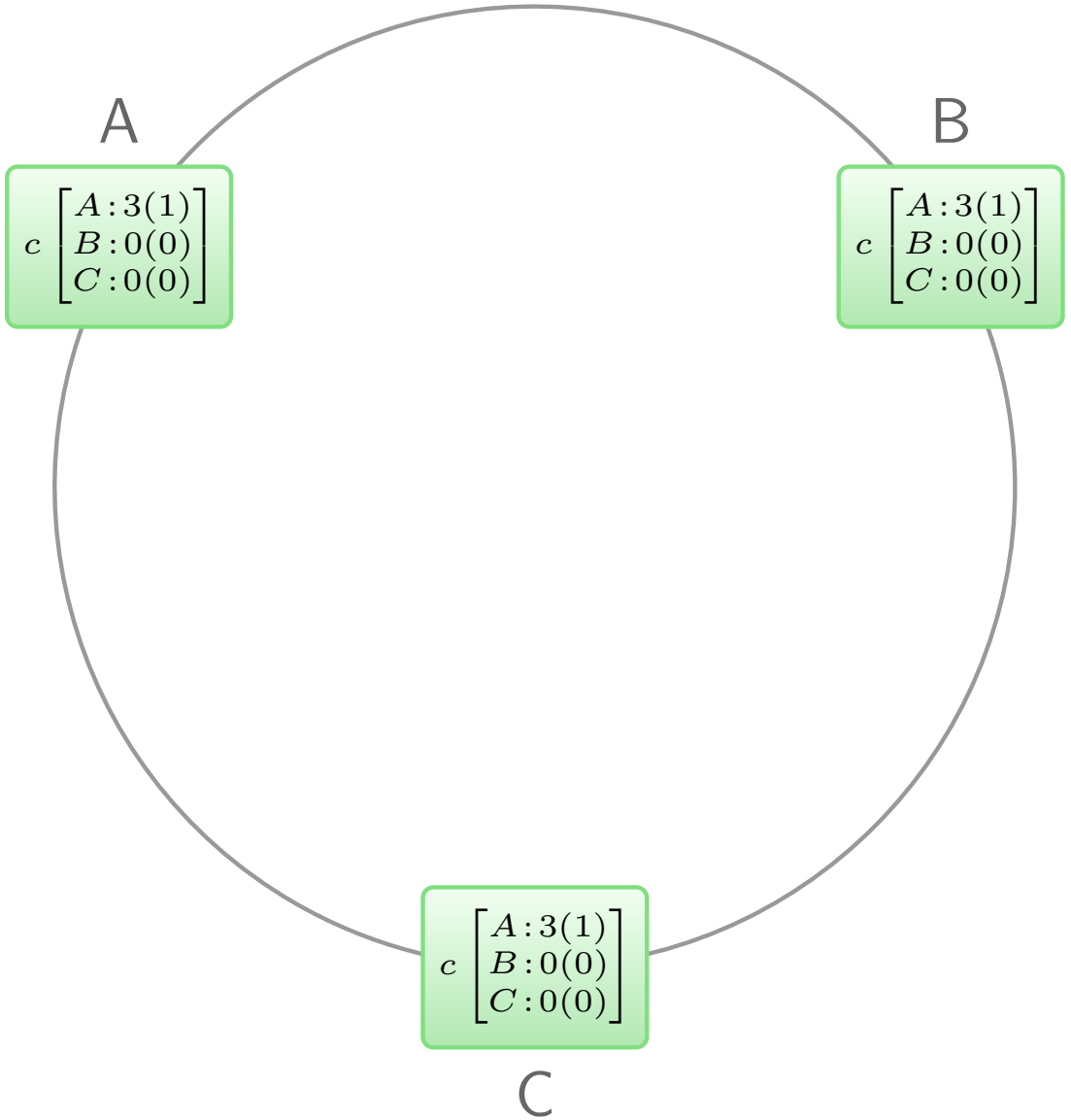
The Cassandra way



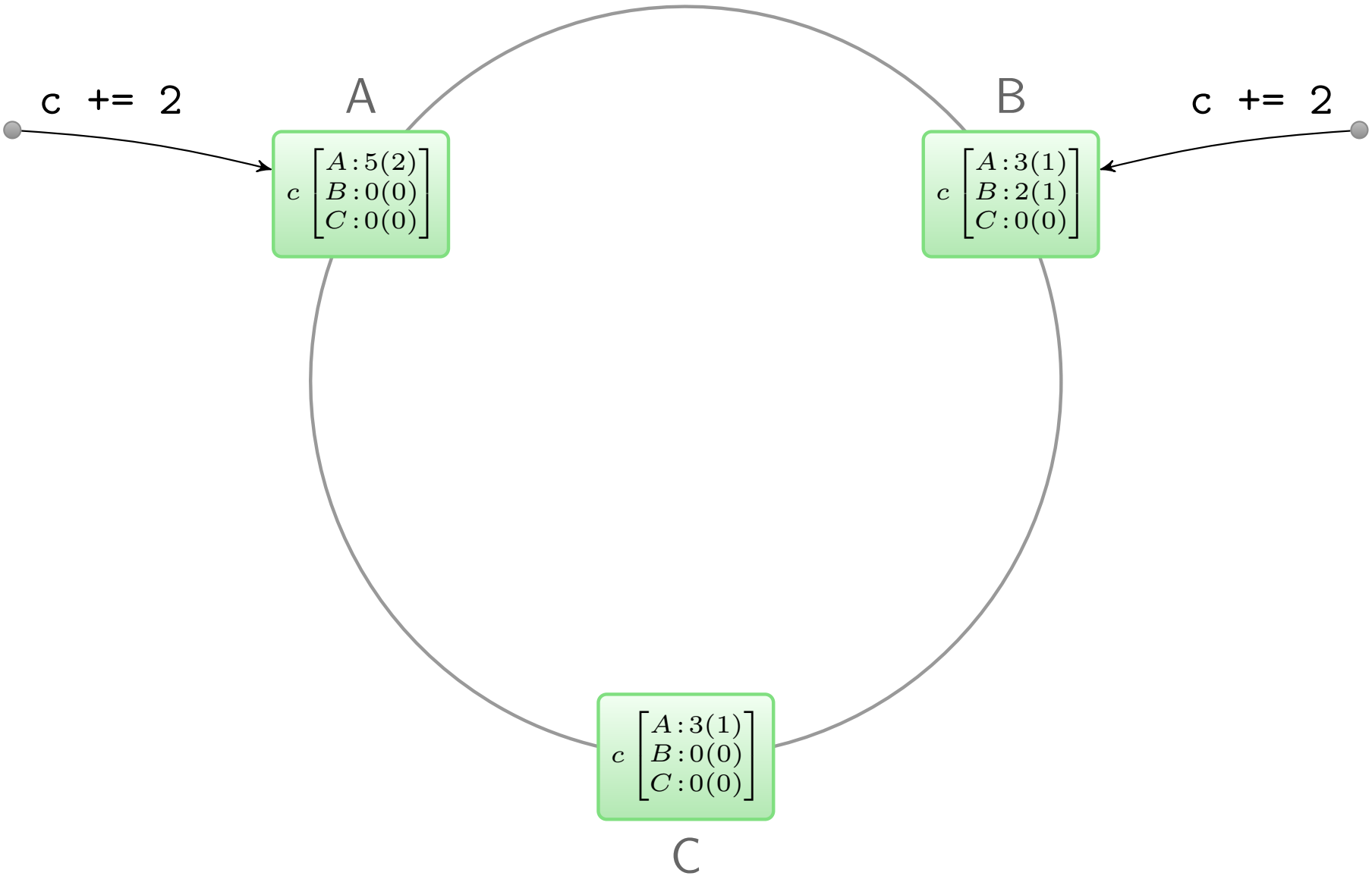
The Cassandra way



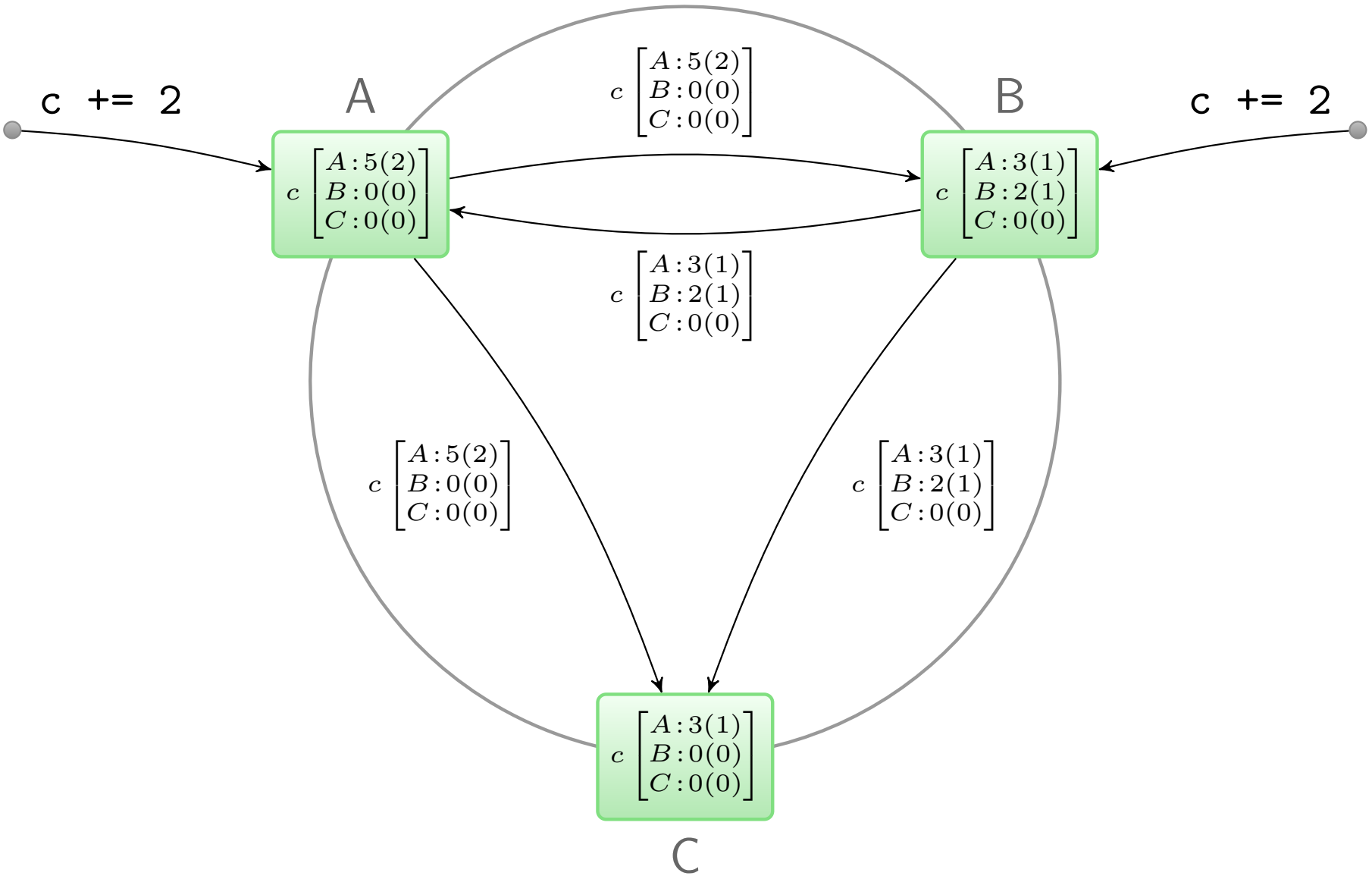
The Cassandra way



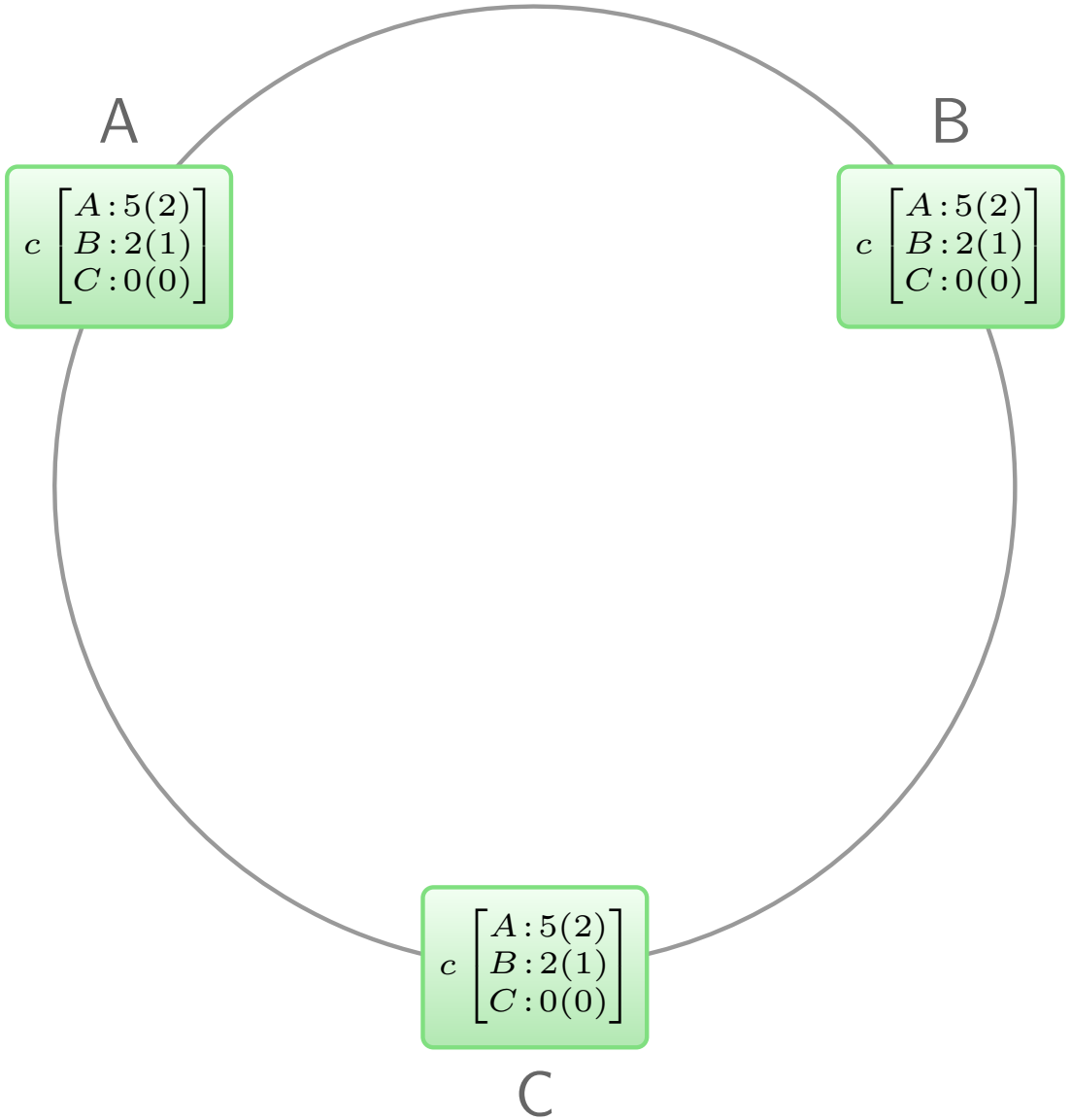
The Cassandra way



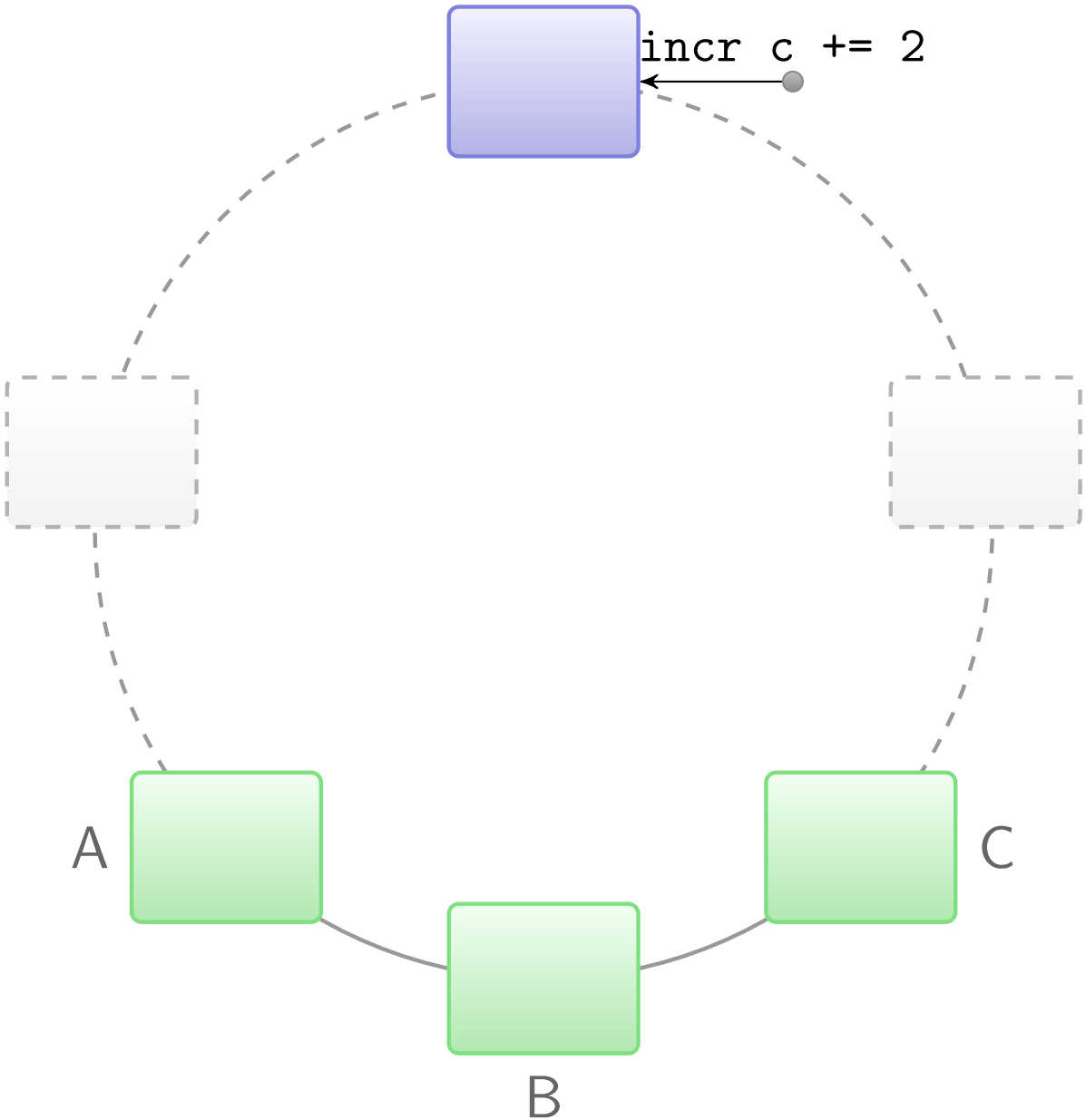
The Cassandra way



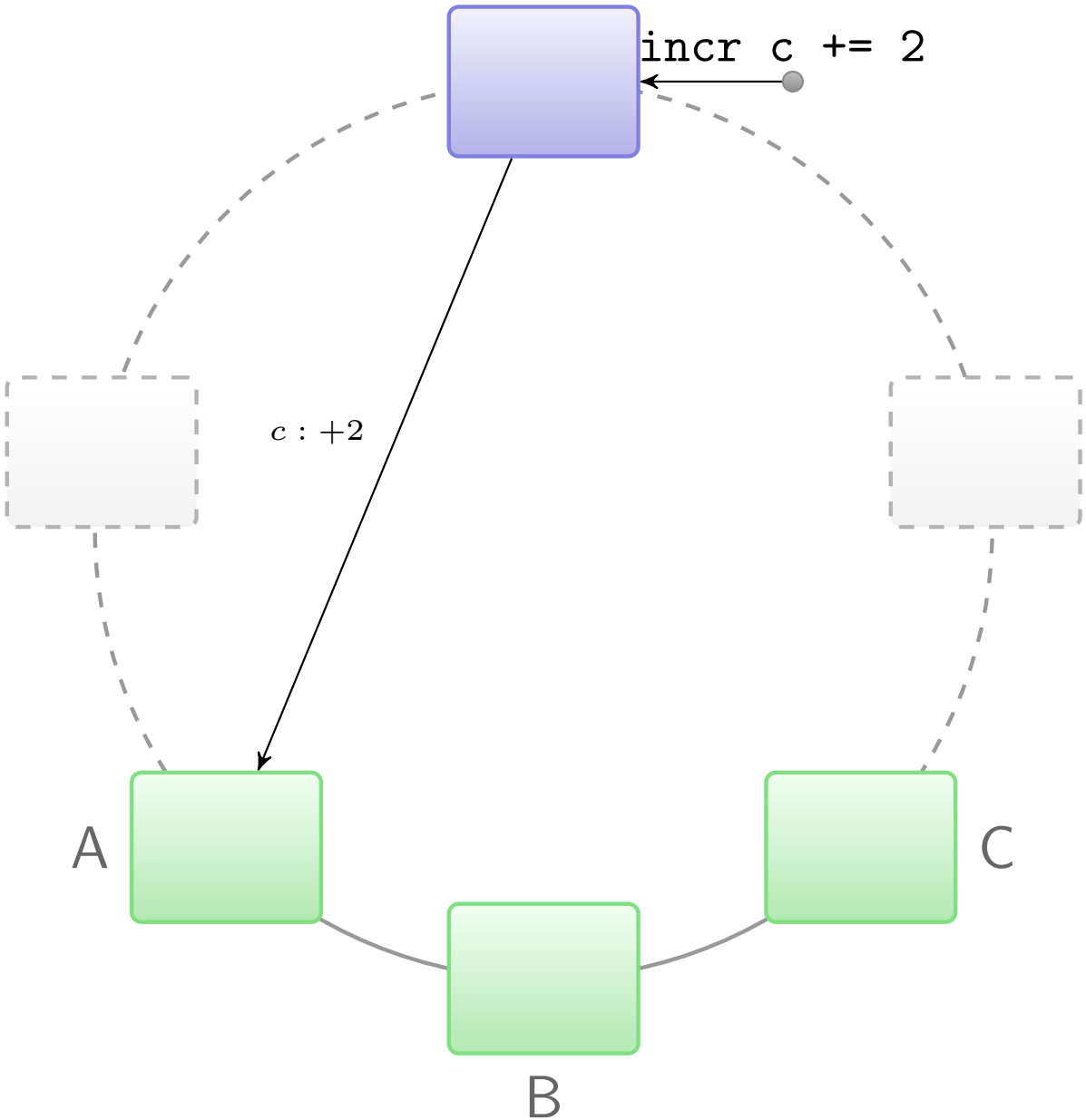
The Cassandra way



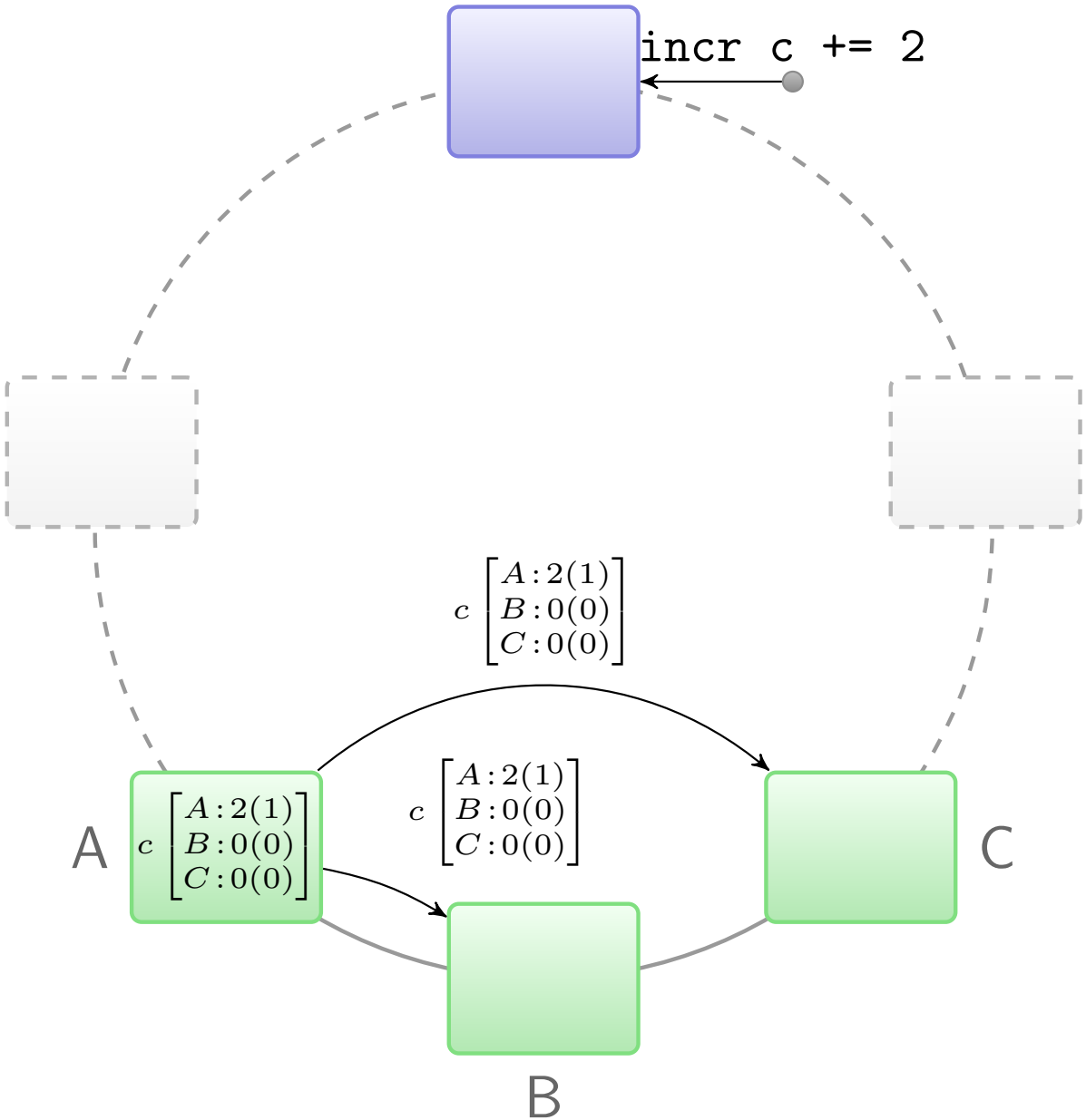
Counter write/read protocol



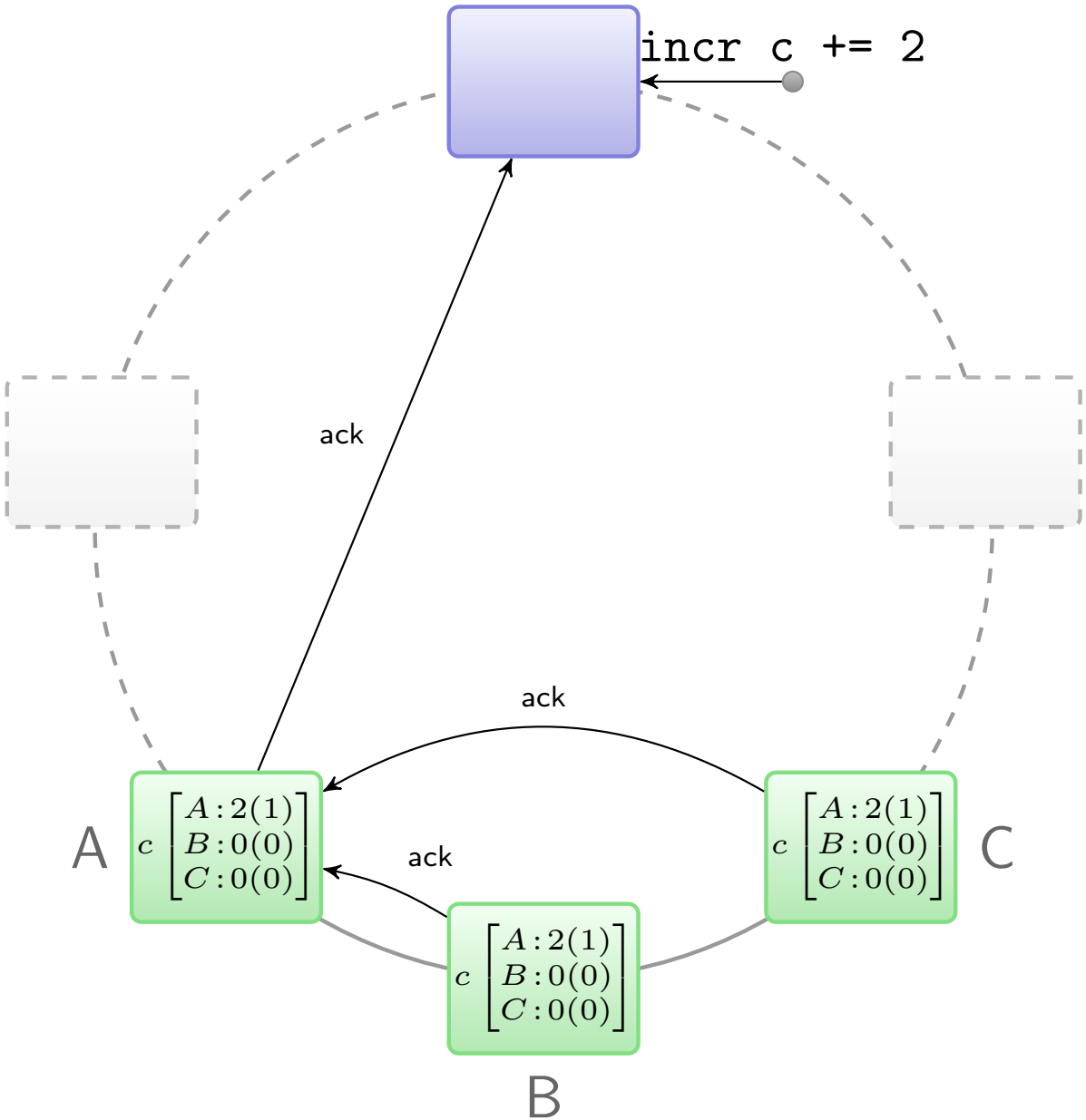
Counter write/read protocol



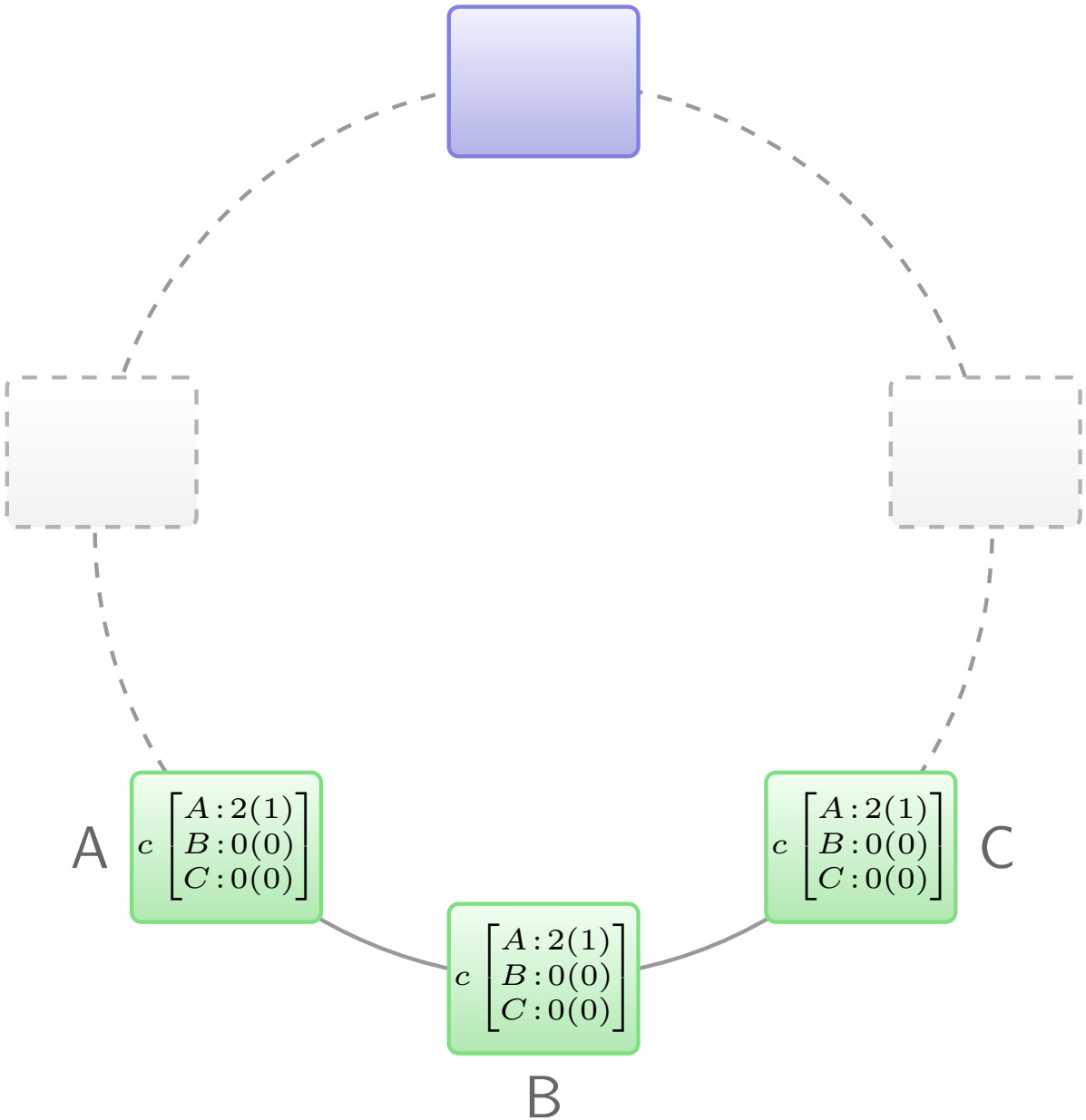
Counter write/read protocol



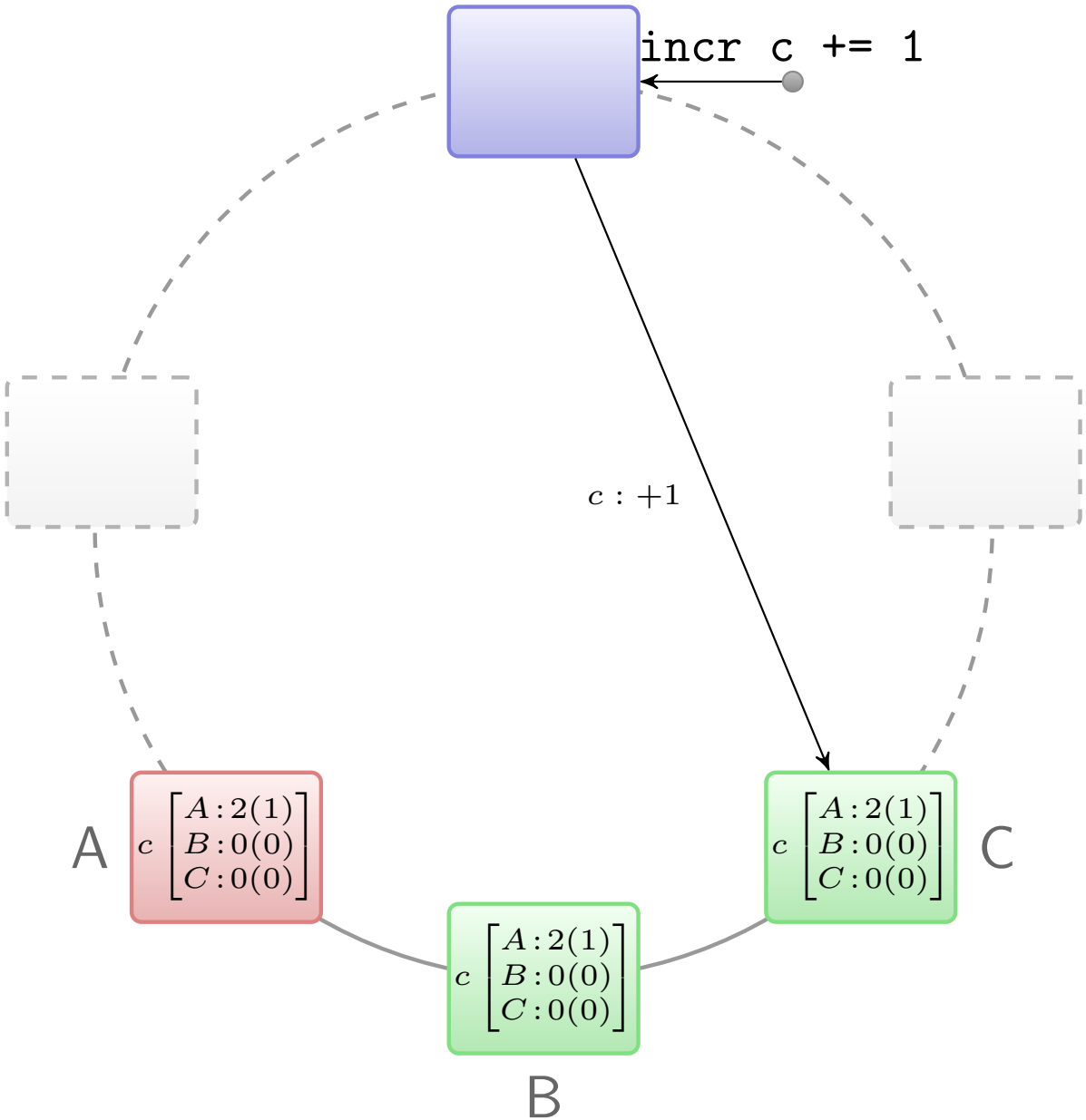
Counter write/read protocol



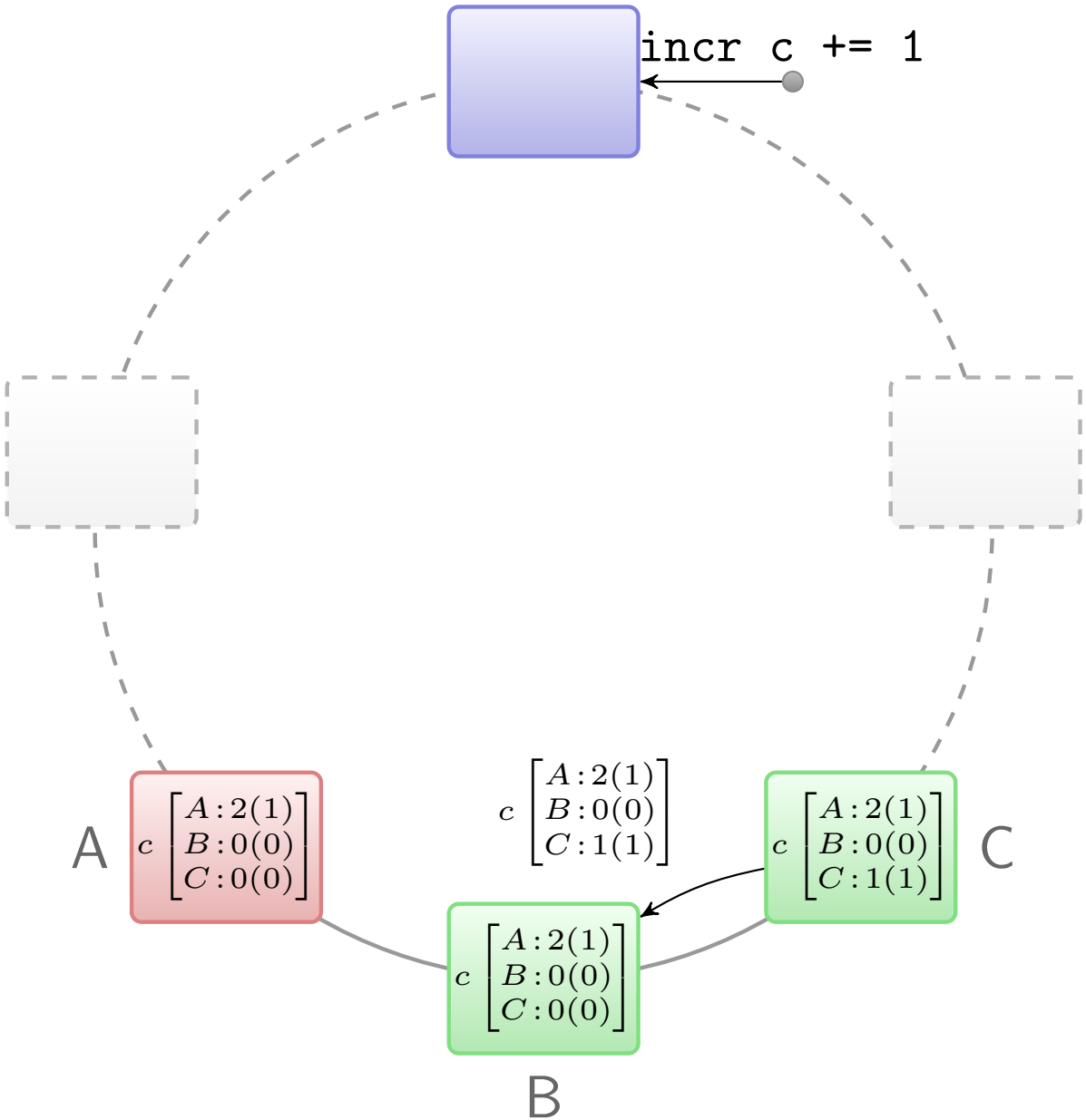
Counter write/read protocol



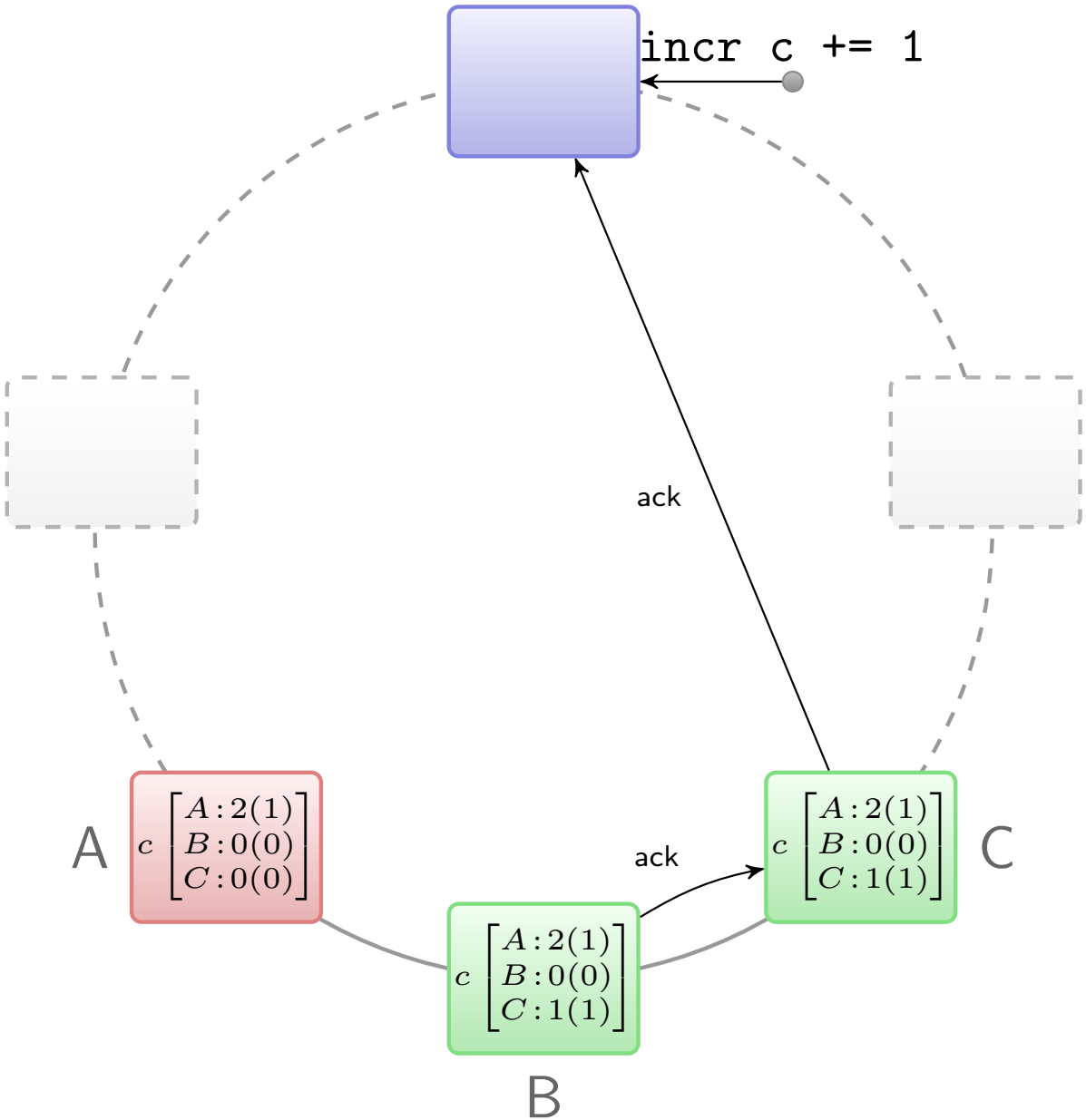
Counter write/read protocol



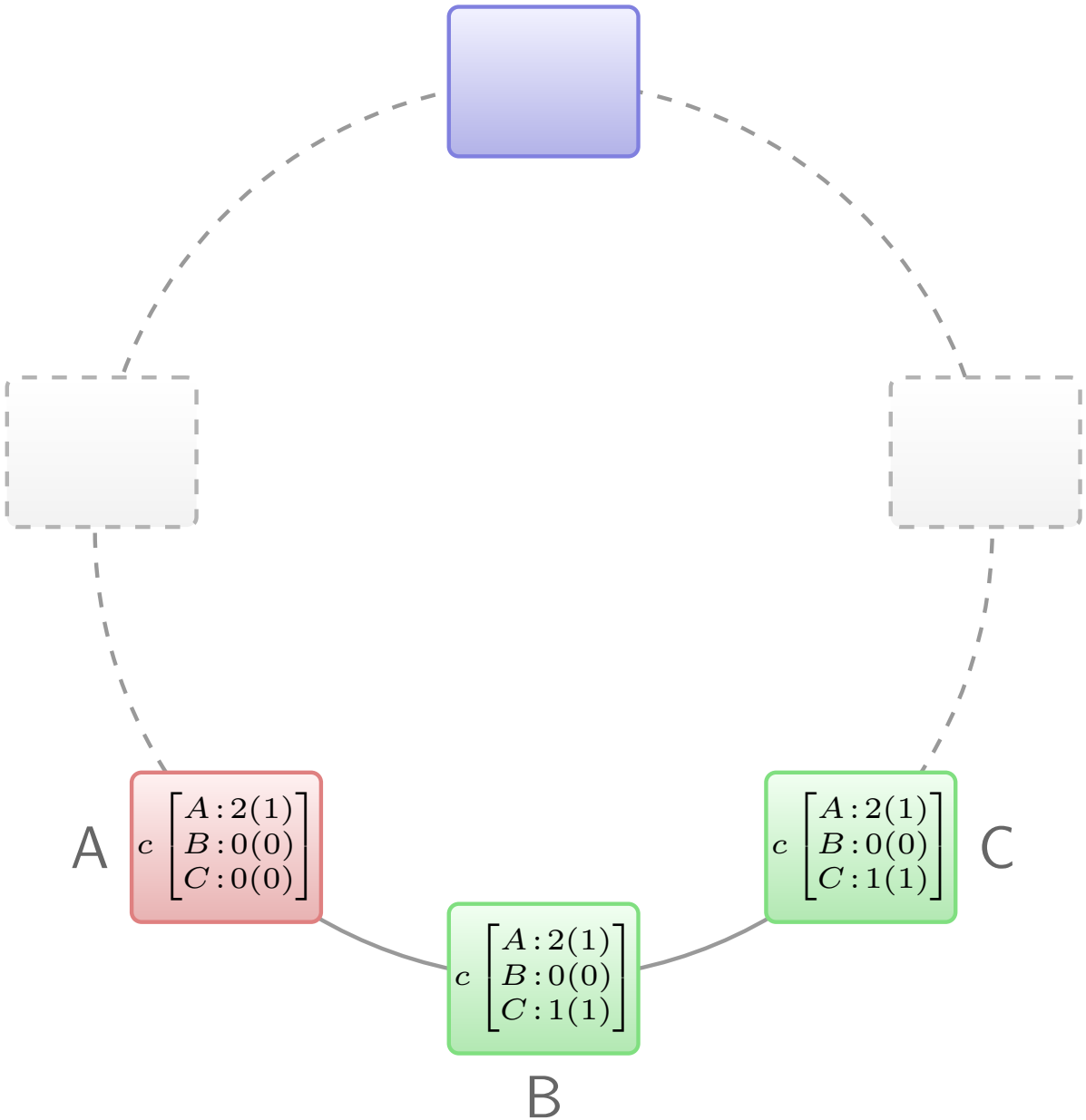
Counter write/read protocol



Counter write/read protocol



Counter write/read protocol



Avoiding excessive synchronization

- On the first replica, we still need to increment the component for that replica.
- We could do a synchronized read-before-write but not too contention-friendly.
- Solution: insert only increments and merge on read/compaction.
- We still need to read after the write to replicate to the other replica. But this need not be synchronized.

Limitation

- In some cases, if the first replica dies in the middle of a write, the client will receive a TimeoutException. In that case, the client will not know for sure that the increment was persisted.
- We're working on improving that.
- But in the meantime, this is better suited for analytics (or you need a manual way to check-and-repair counter).

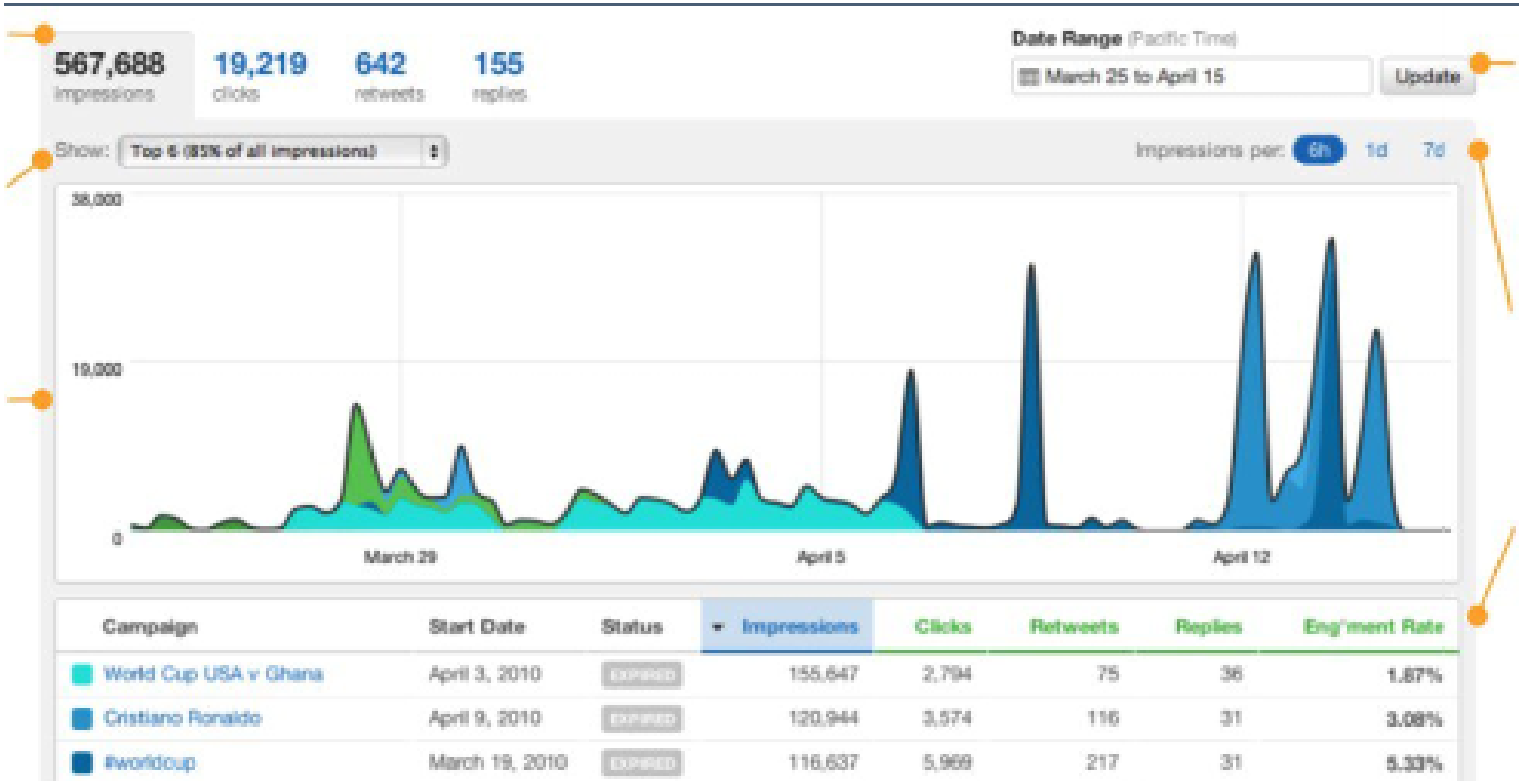
Rainbird (Twitter)

- Twitter's system to count stuffs.
- Needed to scale to 100,000s WPS and 10,000 RPS.
- Layer on top of Cassandra counters.
- Uses in production:
  - Promoted tweets analytics.
  - Internal monitoring and alerting.
  - Tweet button counts.
- They promised to open-source it. But the Cassandra counter implementation has just been released.

Rainbird

- Example: Tracking URL shortener tweets/clicks (t.co).
- Someone click on `http://music.amazon.com/some_long_path`
- Rainbird will automatically increment the count for:
  - [com, amazon, music, some_long_path]
  - [com, amazon, music]
  - [com, amazon]
  - [com]
- And this for different time granularities (avoid big scan at read time).
- Also does mean, standard deviation, etc...

Rainbird


567,688
impressions
19,219
clicks
642
retweets
155
replies
Date Range (Pacific Time)
March 25 to April 15
Update
Show:
Top 6 (85% of all impressions)
Impressions per:
6h
1d
7d
38,000
19,000
0
March 29
April 5
April 12
Campaign
Start Date
Status
Impressions
Clicks
Retweets
Replies
Eng'ment Rate
World Cup USA v Ghana
April 3, 2010
EXPIRED
155,647
2,794
75
36
1.87%
Cristiano Ronaldo
April 9, 2010
EXPIRED
120,944
3,574
116
31
3.08%
#worldcup
March 19, 2010
EXPIRED
116,637
5,969
217
31
5.33%

Counters are in Apache Cassandra 0.8 (released last Thursday).

Counters are in Apache Cassandra 0.8 (released last Thursday).

Questions?