Как при помощи бумаги, карандаша и алгоритма raft достичь консенсуса

Ярослав Дынников

Picodata

Слайды: https://rosik.github.io/2022-fit-m

Алгоритм Raft

Задача

- Достичь консенсуса

https://raft.github.io

Алгоритм Raft

Задача

- Достичь консенсуса
- В ненадёжной сети

https://raft.github.io

Алгоритм Raft

Задача

- Достичь консенсуса
- В ненадёжной сети

Решение

- Распределенная машина состояний

https://raft.github.io

Репликация журнала

Raft State

raft_id

Terms

term:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
vote:														

Raft Log

index:														
value:														
applied:														

Все получают raft_id

α

Raft State

raft_id
1

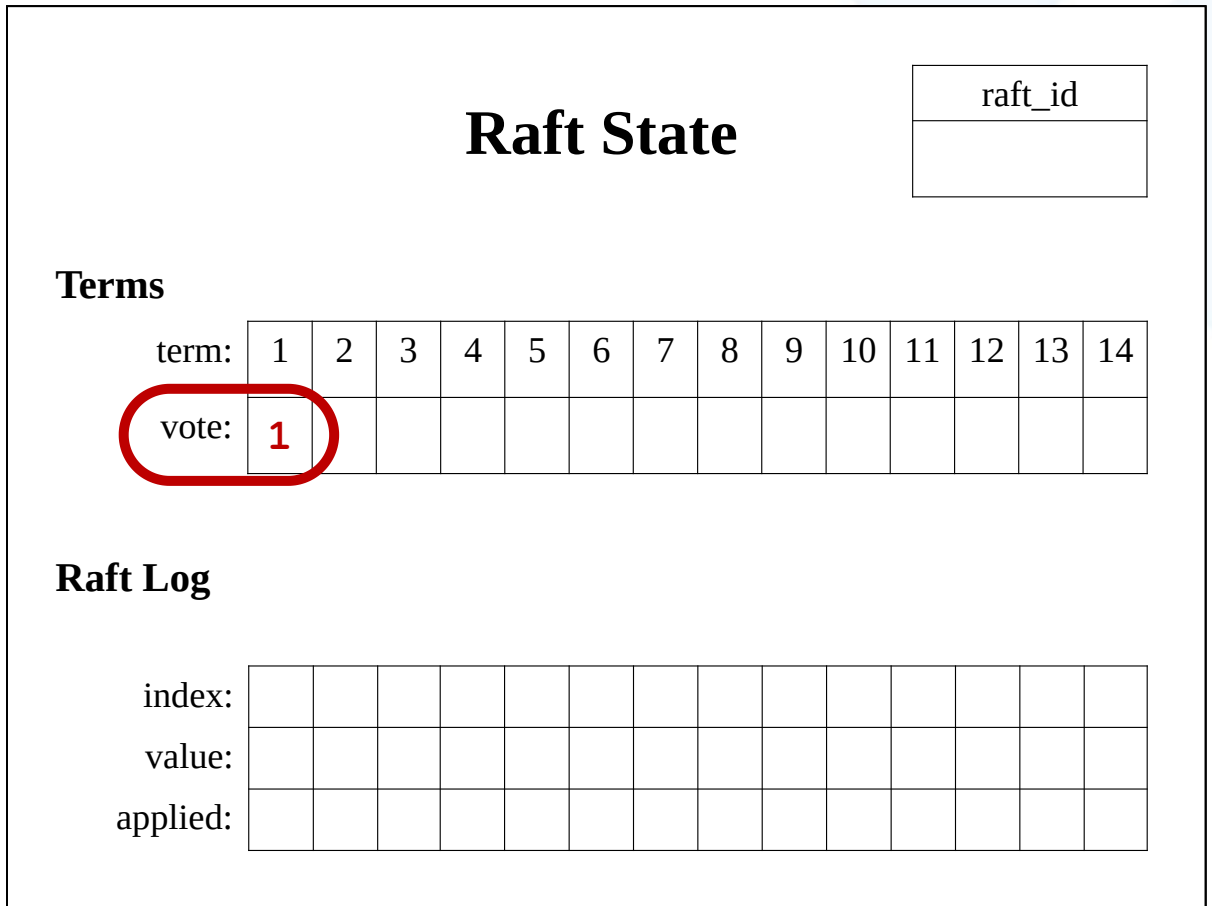
Terms

term:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
vote:														

Raft Log

index:														
value:														
applied:														

Все пропускают первые выборы



Лидер заполняет журнал

Raft State

raft_id
1

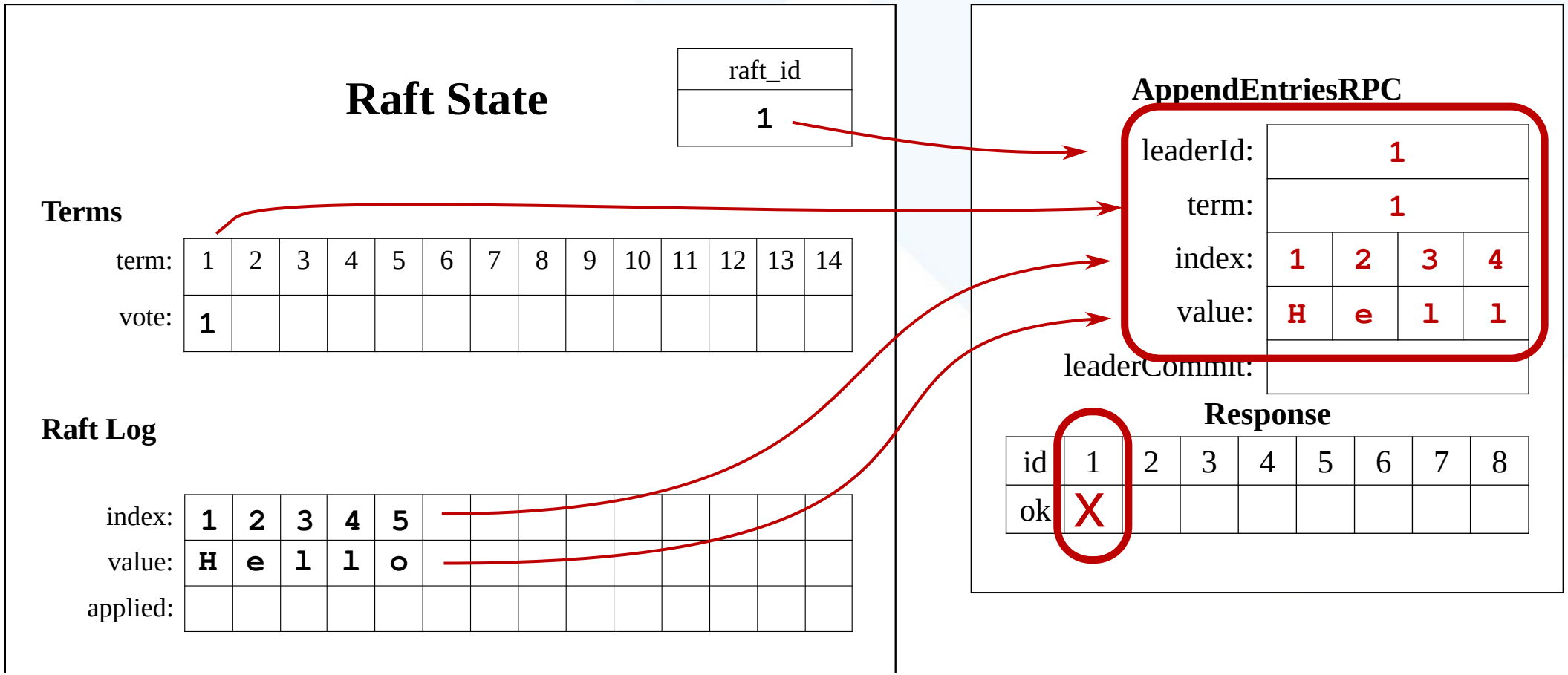
Terms

term:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
vote:	1													

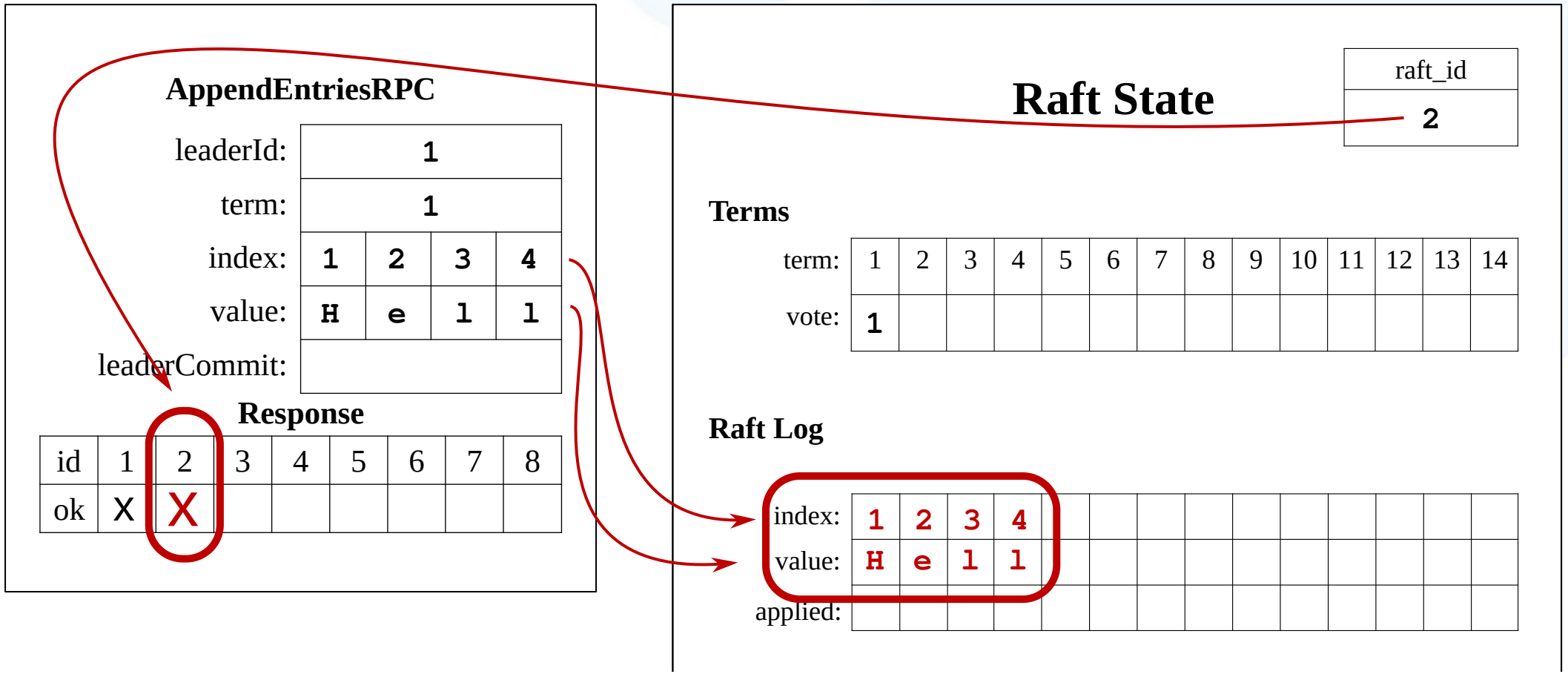
Raft Log

index:	1	2	3	4	5								
value:	H	e	l	l	o								
applied:													

AppendEntries RPC



AppendEntries RPC



AppendEntries RPC

AppendEntriesRPC

leaderId:	1			
term:	1			
index:	1	2	3	4
value:	H	e	l	l
leaderCommit:				

Response

id	1	2	3	4	5	6	7	8
ok	X	X	X	X	X	X	X	X

Raft State

raft_id
1

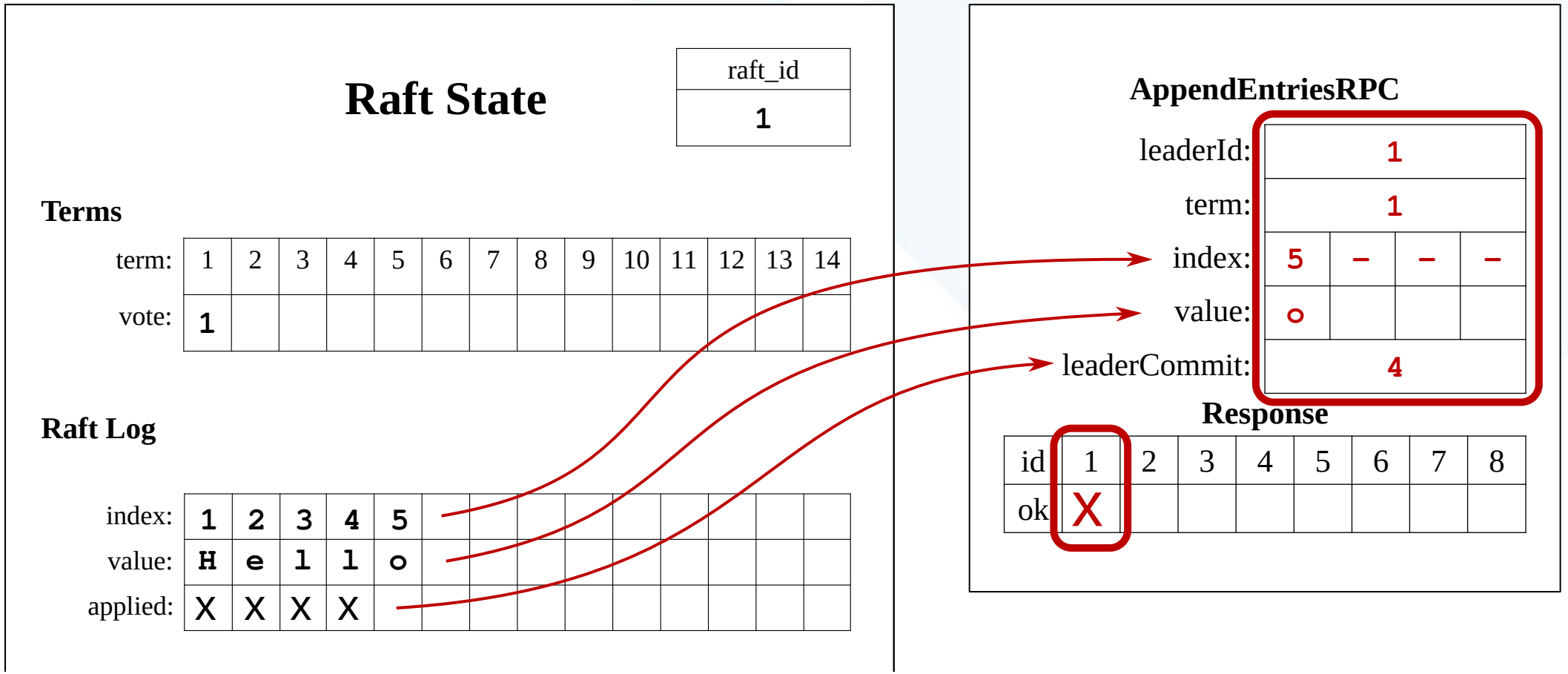
Terms

term:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
vote:	1													

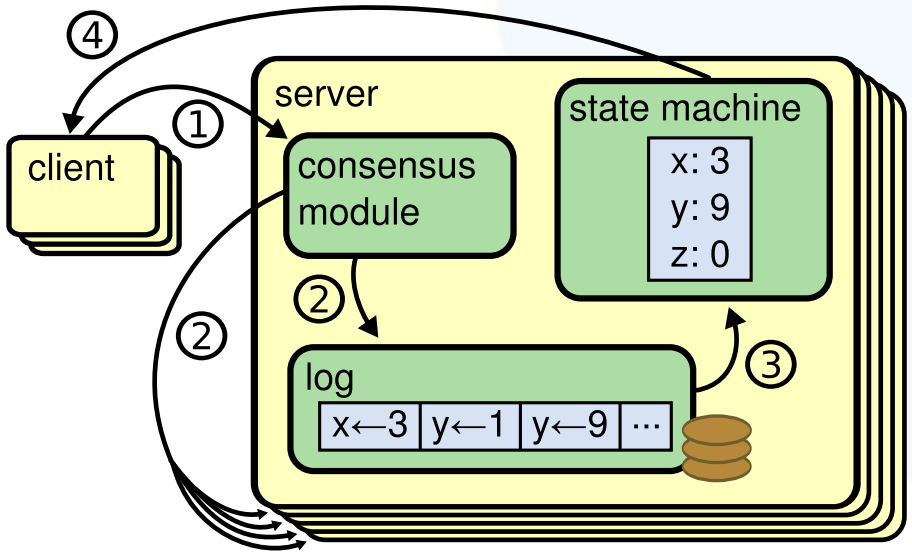
Raft Log

index:	1	2	3	4	5									
value:	H	e	l	l	o									
applied	X	X	X	X										

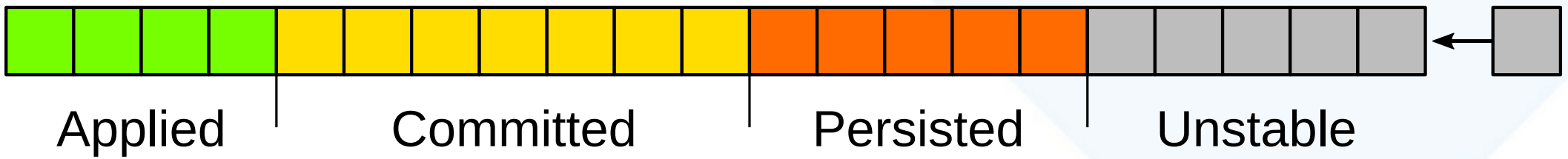
AppendEntries RPC



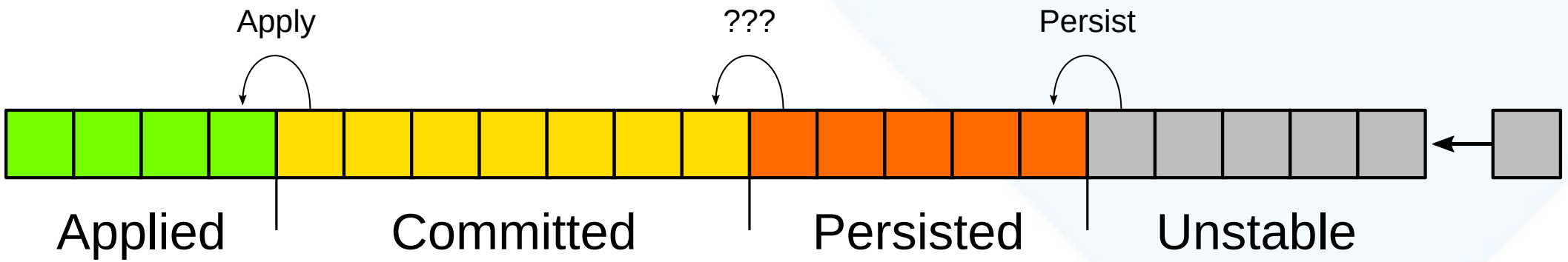
Replicated state machine



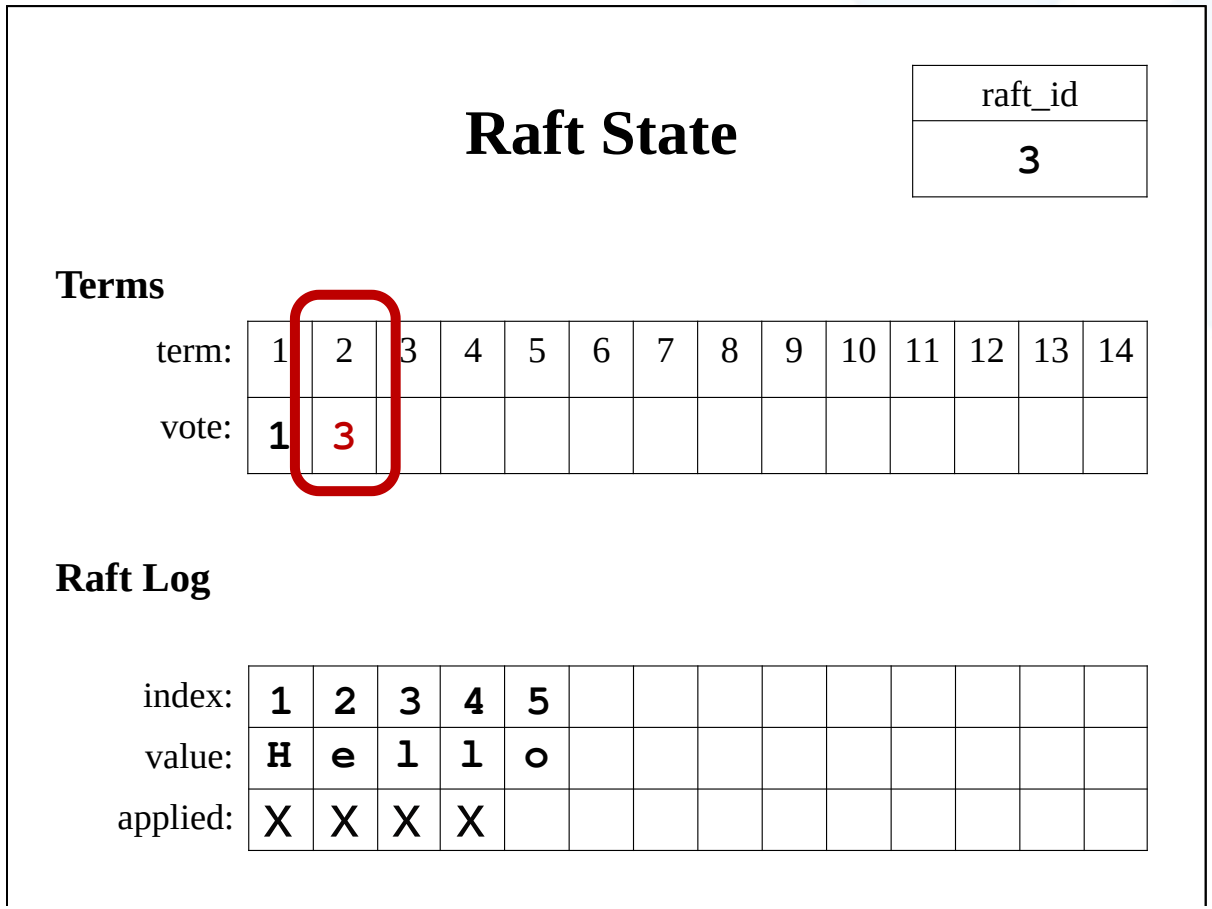
Эволюция записей в журнале



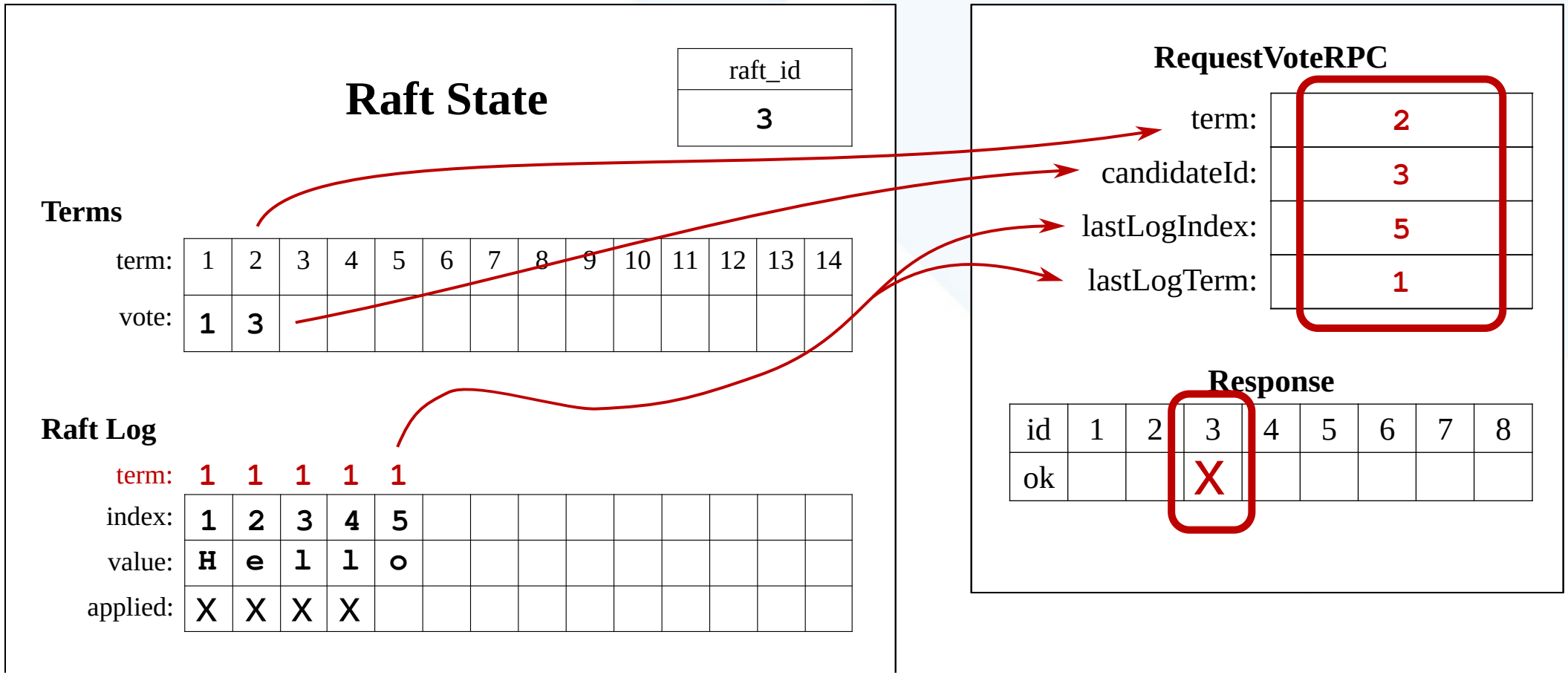
Эволюция записей в журнале



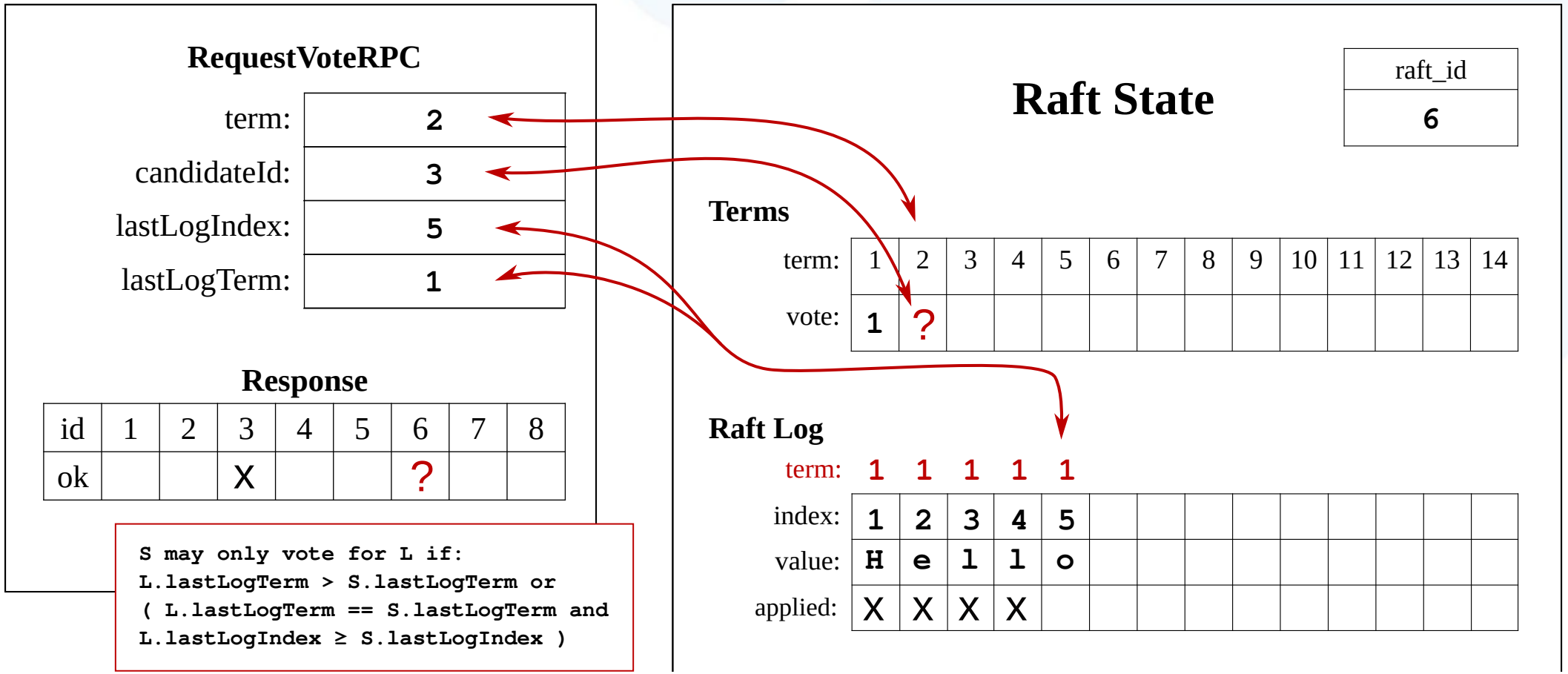
Кто-то начинает новый терм



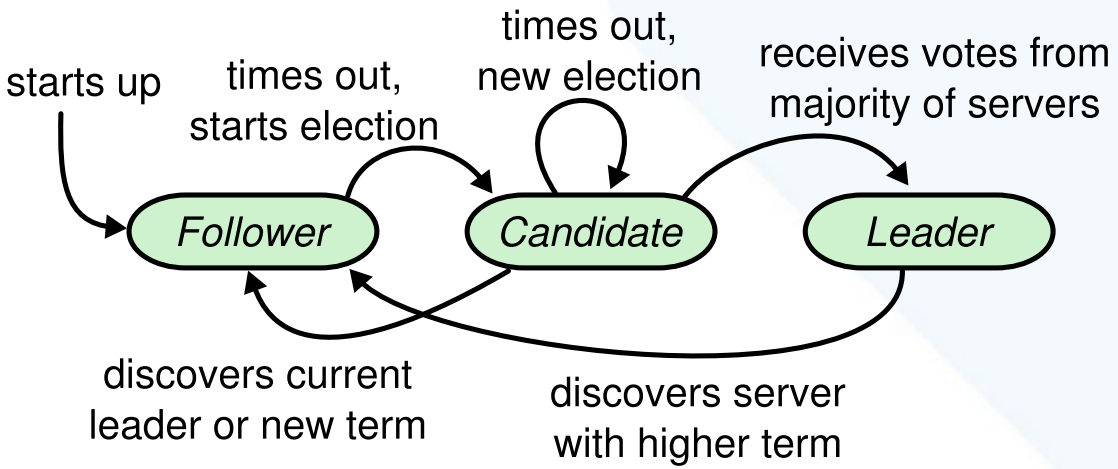
RequestVote RPC



RequestVote RPC



Состояния серверов



Term 2

AppendEntriesRPC

leaderId:	?			
term:	2			
index:	5	6	7	8
value:	o	o	o	!
leaderCommit:	?			

Response

id	1	2	3	4	5	6	7	8
ok								

Откат журнала

AppendEntriesRPC

leaderId:	?			
term:	2			
index:	5	6	7	8
value:	o	o	o	!
leaderCommit:	?			

Response

id	1	2	3	4	5	6	7	8
ok								

Raft State

raft_id
?

Terms

term:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
vote:	1	?												

Raft Log

term:	1	1	1	1	1	2	2	2	2					
index:	1	2	3	4	5	5	6	7	8					
value:	H	e	l	l	o	o	o	o	!					
applied:	X	X	X	X										

Изменение топологии

index	value	applied
1	CC: 1,2,3	X

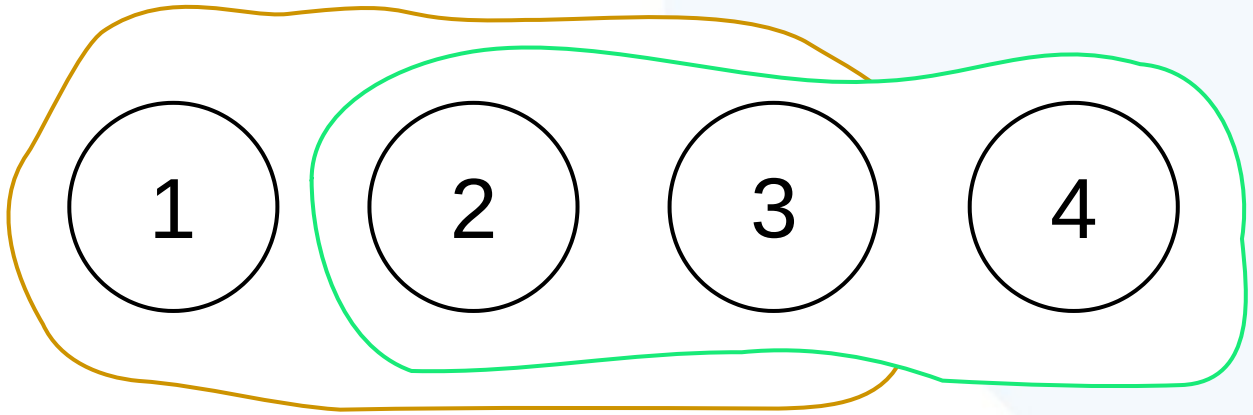
Изменение топологии

index	value	applied
1	CC: 1,2,3	X
2	CC: 1,2,3,4	X

Изменение топологии

index	value	applied
1	CC: 1,2,3	X
2	CC: 1,2,3,4	X
3	CC: 2,3,4	X

Joint consensus



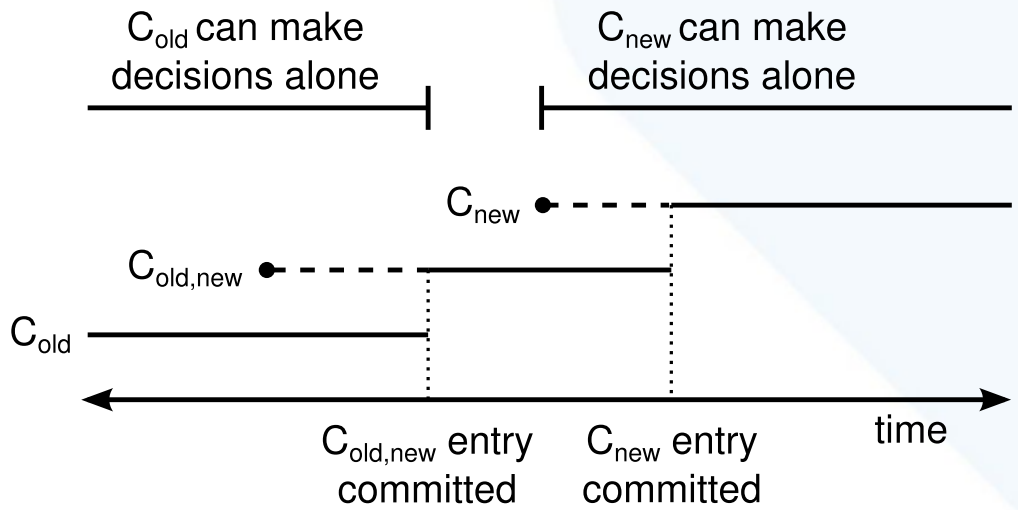
C_old

C_new

Joint consensus

index	value	applied
1	CC: 1,2,3	X
2	CC: 1,2,3,4	X
3	CC: 2,3,4	X
4	CC: 2,3,4+1,2,3	

Joint consensus



Joint consensus

index	value	applied
1	CC: 1,2,3	X
2	CC: 1,2,3,4	X
3	CC: 2,3,4	X
4	CC: 2,3,4+1,2,3	
5	CC: 1,2,3	

Материалы

- Слайды: https://rosik.github.io/2022-fit-m
- Picodata: @picodataru, https://picodata.io/
- Raft: https://raft.github.io/

  In search of Understandable Consensus Algorithm.
  Diego Ongaro and John Ousterhout.
  Stanford University.